Improving Image Classification in Crisis Informatics Through Modern Generative Approaches

Eva Martin

Goldsmiths, University of London

March 2025

Submitted in partial fulfillment of the requirements
of the Master's Degree in Data Science and Artificial Intelligence

Supervised by Foaad Haddod

ABSTRACT

Climate change increases the frequency and severity of natural disasters, creating a need for accurate classification systems to support humanitarian response. Recent advances in artificial intelligence (AI) offer potential solutions. This study investigates two approaches to improving disaster image classification in crisis informatics: 1) a custom pipeline using generative AI for synthetic data augmentation to fine-tune convolutional neural networks (CNNs), and 2) direct zero-shot classification using large pre-trained multimodal models. The former represents a specialised technical approach, while the latter leverages general-purpose foundation models.

Using the MEDIC crisis informatics dataset (70,000+ images), we replicated published CNN baselines across three architectures. Data preparation revealed performance gaps linked to labelling errors in the original dataset. To address this and prepare for reliable data augmentation, we implemented and validated a conservative committee-based relabelling method—using CNNs and large language models (LLMs)—, which relabelled 5% of the total data. This significantly improved baseline performance on challenging classes (e.g., 'mild damage' F1 rose from 15.4% to 35.8% for EfficientNet-B1).

Our main experiments first evaluated our synthetic data augmentation pipeline, which uses LLM-generated captions to generate new images with diffusion models. This approach yielded only modest performance improvements for fine-tuned CNNs, even with targeted data generation strategies. Conversely, zero-shot classification using large multimodal models (e.g., GPT-4o) achieved higher accuracy and comparable or significantly higher F1 scores across multiple tasks than the best fine-tuned CNNs, requiring no task-specific tuning. The F1 advantage was particularly pronounced for ambiguous categories ('mild damage' F1: 61% vs 36%; 'other disaster' F1: 57% vs 24%), though F1 scores for some other tasks decreased.

In conclusion, while targeted synthetic data augmentation shows some promise for heterogeneous crisis informatics datasets, our findings reveal a significant, underexplored potential in the use of off-the-shelf large multimodal models as zero-shot classifiers. These results indicate that such foundation models could become key components in future crisis informatics classification systems.

Keywords: crisis informatics, image classification, deep learning, data augmentation, zero-shot classification, generative AI

WORD COUNT

Word count (excluding Ethical Considerations, Tables, Figures, Table of Contents, Acknowledgments, References and Appendix):

9941 words

Counted using LaTeX command:

```
\newcommand{\quickwordcount}[1]{%
  \immediate\write18{texcount -1 -sum=1,1,0,0,0,0,0 -noinc -nobib
  -q #1.tex | grep -v "errors:" > #1-words.sum }%
  \input{#1-words.sum}words%
}
```

Acknowledgments

My thanks are due to the staff, instructors, tutors, and supervisor at Goldsmiths, University of London for the opportunity to undertake this MSc programme in Data Science and for the learning materials made available. I also appreciate the responsiveness and support provided when queries arose during my studies. Completing this programme, culminating in this final project, has been a challenging but ultimately rewarding learning experience.

An equally big thanks to my ever-patient husband, Luigi, for feeding me, watering me, pretending to marvel at my every growing synthetic image bank and being my rock these past couple of years.

Disclaimer: Use of Generative AI

Generative Artificial Intelligence (AI) was used in this project in two ways: as a core component of the research methodology and an assistive tool.

A core part of this project involved the investigation and application of generative AI models, which included employing Large Language Models (LLMs) with vision capabilities, such as Anthropic's Claude and OpenAI's GPT-4o, for image relabelling, captioning, and classification, and image generation models such as Stable Diffusion and Flux 1-dev for creating synthetic images in the data augmentation experiments. The outputs and performance of these AI systems are part of the results and analysis presented in this work and this use is integral to the research subject. Ethical considerations about the project, including the usage of AI, are covered in Section 2.7.

Separately, I used AI tools (primarily the Claude 3.5 series; Anthropic, 2024) as assistants during the research and writing process. This included using them as a sounding board for structuring ideas (final decisions were my own), to help clarify concepts from the literature (understanding and application were my own, with facts independently verified) and to suggest grammatical or clarity improvements for text already drafted by me (suggestions were critically reviewed and the final writing is mine). Additionally, AI assisted with debugging Python code used in the analyses and with solving laborious LaTeX tasks, such as typesetting complex tables or providing obscure LaTeX code snippets, whose logic and correctness was double-checked by me.

Therefore, while AI executed defined tasks within the research and provided limited assistance as described, **I affirm that the core intellectual contributions—the research conception, experimental design, data analysis, write up, interpretation of findings, and final conclusions presented in this work—represent entirely my own indepen-**

dent effort and insights. AI served as a tool in the research and writing process, **but did not contribute any original ideas or direct original content** to the project.

Contents

Chapter 1

Introduction

1.1 Motivation

Climate change is leading to increasingly intense natural disasters, posing significant challenges for disaster response and humanitarian aid. For mitigation and relief efforts to be effective, accurate and timely information about incident types, locations, and scales is essential. Within the field of crisis informatics, **image classification** is crucial to automated systems that monitor social media during crises (Alam et al., 2023). Such systems collect images and metadata, creating detailed snapshots of ongoing situations, assessing severity levels, and identifying humanitarian needs. These tools are vital for maintaining situational awareness, planning resources effectively, and optimising response efforts.

Recent advances in foundation models and generative artificial intelligence (AI) have created multiple pathways to improve crisis informatics systems. Advanced image generation models such as Stable Diffusion (Esser et al., 2024) can now produce photorealistic outputs with high adherence to text prompts, making them viable tools for data augmentation in disaster contexts where labelled images are scarce or their usage raises ethical issues. Simultaneously, large multimodal (vision and language) models such as OpenAI's gpt-4o (OpenAI, 2024) or Anthropic's Claude 3.5 series (Anthropic, 2024), which can process and reason across both images and text, offer powerful off-the-shelf ('zero-shot') image classification capabilities, potentially bypassing established pipelines for image analysis, traditionally reliant on Convolutional Neural Networks (CNNs; Krizhevsky et al., 2012).

1.2 Research Problem and Significance

Traditional **disaster image classification** relies heavily on supervised machine learning techniques. Typically, these methods require extensive annotated datasets and often struggle when encountering novel disaster contexts or new classes (Mumuni and Mumuni, 2022). Modern generative AI approaches, however, offer promising alternatives to these limitations.

Specifically, two distinct methods stand out:

1. Synthetic data augmentation with advanced image generative models extends existing classification methods by artificially creating additional training examples, producing high-quality images which can be used to improve class balance and data diversity with fewer ethical issues than real disaster images.

2. Zero-shot classification with large multimodal models allows models to categorise images into classes they have not been explicitly trained on, using transferable knowledge from their pretraining. This approach entirely bypasses existing pipelines, potentially enabling the classification of previously unseen disaster categories without explicit fine-tuning (Kojima et al., 2022; Pratt et al., 2023).

The **significance** of improving disaster image classification lies in enhancing the effectiveness of response operations, ensuring better resource allocation, and ultimately saving lives and property.

This project explores two complementary paths to improve image classification in crisis informatics using modern generative AI techniques: (1) synthetic data augmentation for CNN fine-tuning and (2) zero-shot classification with large multimodal models, as we explain in the rest of this chapter.

1.3 Research Questions and Objectives

As mentioned above, the **aim of the project** is to explore and compare two distinct paths to improve image classification in crisis informatics: synthetic data augmentation and zero-shot classification. We will measure success empirically by comparing our disaster-related image classification performance to the MEDIC dataset benchmarks (Alam et al., 2023), detailed in Chapter 2.3. Our objectives and associated research questions are outlined below in Table 1.1.

Table 1.1: Research Objectives and Corresponding Research Questions.

Objective	**Research Question**
Investigate synthetic data augmentation for CNNs in crisis informatics, generating additional training samples with advanced image generation models.	**RQ1:** Does synthetic data augmentation using advanced image generation models help improve CNN performance for disaster image classification (MEDIC benchmark), compared to baseline?
Evaluate zero-shot classification with large multimodal models for crisis informatics image classification, establishing performance with no task-specific training.	**RQ2:** Can large multimodal models achieve competitive zero-shot classification performance on the MEDIC dataset compared to CNN benchmarks?

Comparing these approaches highlights important tradeoffs between traditional approaches with higher implementation complexity and control (synthetic data with CNNs) versus simplicity but potential reliance on external black-box dependencies (zero-shot with possibly proprietary pretrained models), providing practical guidance for crisis response teams with varying technical resources and deployment constraints. Thus, the **primary contribution of this project** lies in the novel implementation, empirical evaluation, and comparison of these two distinct approaches for leveraging modern generative AI for disaster image classification.

1.4 Structure and Overview

This report is organised as follows:

- Chapter 1 introduces the context, research objectives, and significance of the study.
- Chapter 2 provides an overview of crisis informatics, the MEDIC dataset, and relevant literature on both synthetic data generation and zero-shot classification. It concludes by identifying the **research gap** and our contributions, and a section on **ethical considerations**.
- Chapter 3 outlines our general methodology, including experimental setup and evaluation frameworks.
- Chapter 4 describes the CNN baseline experiments.
- Chapter 5 outlines the need for relabelling the MEDIC dataset and presents the methodology and results.
- Chapter 6 details the synthetic data augmentation pipeline, the associated experimental design and results.
- Chapter 7 presents the methodology and results for zero-shot classification with large multimodal models.
- Chapter 8 concludes with a comparative analysis, summary of contributions, practical implications, limitations and future research directions.
- Appendices A to E contain supporting analyses, supplementary tables, prompts, example images and model details.

Chapter 2

Background and Related Work

In this chapter, we review work in image classification in crisis informatics and more broadly, from traditional approaches involving Convolutional Neural Networks (CNNs) to large multimodal models and advanced synthetic data generation, with the goal of identifying opportunities to combine these approaches for disaster imagery classification. We then identify the research gap in the literature and list our contributions. The chapter concludes with a section on **ethical considerations**, an essential part of data science and particularly crisis informatics.

2.1 Crisis Informatics and Disaster Response

Machine learning plays an increasingly vital role in disaster response, powering applications such as real-time event detection (Alam et al., 2022), situational awareness from social media (Nguyen et al., 2017; Yao et al., 2020; Bukar et al., 2022), and damage assessment from aerial or satellite imagery (Hamdi et al., 2019; Duarte et al., 2018; Braik and Koliou, 2024). Models must deliver results under strict time constraints and often on limited computational resources available in the field, with accuracy often sacrificed for the sake of speed (Gholami et al., 2022).

Another challenge is the sheer volume and variety of data during disasters. Social media, satellite, and drone platforms generate an overwhelming number of images in real time (Alam et al., 2022). The need for automated image classification has therefore grown (Kumar and others, 2020), enabling authorities to parse millions of images for signs of damage or people in need. However, **models trained on academic benchmarks often struggle to generalise** to such real-world data streams: there is a notable gap between curated research datasets and the noisy, evolving visual data from actual crises (Weber et al., 2023).

2.2 Traditional Approaches to Disaster Image Classification

Pre-trained convolutional neural networks (CNNs; Krizhevsky et al., 2012) like VGG16, ResNet, and DenseNet remain the workhorses of disaster image classification. Studies often fine-tune these ImageNet-trained models on disaster datasets, leveraging *transfer learning* to recognise disaster-related features (type of disaster, damage severity) in photographs. CNNs achieve strong baseline performance on disaster-related image datasets like CrisisMMD and AIDR

(Alam et al., 2018, 2023; Imran et al., 2014). For example, on the CrisisMMD image set, CNNs achieved high F1 scores of $\sim$ 84% and $\sim$ 78% on the 'informativeness' and 'humanitarian' classification tasks, respectively (Alam et al., 2018).

Traditional CNN methodologies have been further developed by incorporating techniques such as multi-scale analysis to capture both broad context and fine details (Zhang et al., 2024). Other approaches involve using intermediate steps like semantic segmentation or object detection to enrich features (Rahnemoonfar et al., 2023; Kyeongjin et al., 2024), and integrating multimodal data (e.g., text and images from social media) to help clarify visual content using associated context (Zou et al., 2021; Islam et al., 2024). The key question asked in this project is the role of these traditional or hybrid CNN-based approaches in light of the advances in modern generative modelling, as we describe later in this chapter.

2.3 The MEDIC Dataset

In machine learning, progress in a field is often defined—and spurred—by quantitative improvements on a keystone benchmark dataset and associated tasks. The Multitask Emergency Dataset for Crisis Informatic (MEDIC) dataset (Alam et al., 2023) is currently the largest social media disaster image dataset, comprising 71,198 images annotated for four interrelated humanitarian classification tasks: disaster type, informativeness, humanitarian category, and damage severity. MEDIC consolidates previous datasets like CrisisMMD (Alam et al., 2018), AIDR (Imran et al., 2014), and DMD (Mouzannar et al., 2018), significantly advancing their scale and operational relevance compared to broad incident-detection datasets such as Incidents1M (Weber et al., 2023) or aerial datasets like xView2 (Defense Innovation Unit, 2019).

The MEDIC dataset lies at the core of this project and we will regularly refer back to it. Despite its advantages, MEDIC faces challenges including moderate annotation noise from crowdsourced labels and considerable class imbalance, limiting classifier performance (Alam et al., 2023; Eltehewy et al., 2023), topics which we will revisit in the following chapters.

2.4 Foundation Models in Computer Vision

Recent years have seen the rise of *foundation models* in computer vision—large-scale models trained on enormous datasets that can be adapted to a wide range of tasks. Notable examples include vision-language models like CLIP (Contrastive Language–Image Pre-training), generative image models like Stable Diffusion, multimodal transformers such as BLIP, and multimodal variants of large language models (e.g., GPT-4o; OpenAI, 2024). These models are characterised by their ability to perform zero-shot or few-shot learning, meaning they can recognise new concepts without explicit task-specific training, by leveraging their broad knowledge learned during pre-training (Scheele et al., 2024).

Vision-language models (VLMs) like CLIP are particularly relevant for image classification. CLIP was trained on 400 million image–text pairs to align image embeddings with text embeddings in a shared space enabling classification of images using simple textual prompts without the need for traditional retraining—a form of zero-shot classification (Scheele et al., 2024). This zero-shot ability often rivals or surpasses fully supervised baselines on datasets it has never seen (Radford et al., 2021).

Meanwhile, *multimodal large language models* have emerged, which can accept image inputs in addition to text, such as with GPT-4o (OpenAI, 2024). However, models with state-of-the-art vision capabilities are often proprietary and very computationally intensive, and their outputs are not easily reproducible or benchmarked in the same way as traditional classifiers.

Foundation models have been applied in various specialised domains. In medical imaging, researchers proved that a CLIP-based model achieved radiologist-level zero-shot classification of pathologies on chest X-rays (Mishra et al., 2023), despite never being explicitly trained. In traffic incident detection, vision-language models are being explored to identify accidents from traffic camera footage using textual cues (Zhang et al., 2025).

Less work has been done in disaster informatics, but early experiments have used foundation models to classify disaster images. For example, the LADI-v2 project (Scheele et al., 2024) compared a baseline ResNet to vision-language models for classifying aerial disaster photos. Interestingly for our project, they found that **a fine-tuned ResNet still outperformed open-source VLMs on that specialised aerial dataset**—a foundation model pre-trained on internet images may not immediately excel on drone imagery of hurricane damage without adaptation. This is known as *domain shift* – if the distribution of target data differs significantly from the model's training data, performance can degrade (Dunlap et al., 2023).

2.5 Synthetic Data Generation

Data augmentation is a traditional approach to improve model performance by artificially expanding a training dataset by applying transformations such as rotation, flipping and cropping to existing images. Recent advances in generative models have enabled synthetic data augmentation beyond simple transformations. Earlier generative approaches mostly employed Generative Adversarial Networks (GANs; Goodfellow et al., 2014), and studies across domains showed that augmenting training data with synthesised images can improve classification accuracy by oversampling minority classes with realistic examples (Figueira and Vaz, 2022).

Diffusion models, which generate images by gradually denoising random patterns into coherent visuals based on text descriptions, are a newer family of generative models achieving state-of-the-art image fidelity. They have spurred interest in augmenting image datasets for training classifiers, primarily in medical fields, often by first fine-tuning a pre-trained diffusion model on a specific medical dataset (Alimisis et al., 2025). Similar to earlier attempts, the common application consists of generating synthetic examples for specific target classes to improve class balance, yielding improved accuracy on various benchmarks (Shao et al., 2024).

Related to the aims of this project, Dunlap et al. (2023) created ALIA, a generic framework for automatically generating descriptive captions using a language model and then using those captions with Stable Diffusion to produce variant images. Similarly, recent work by Yu et al. (2025) showed how using an off-the-shelf image generation model to increase dataset size and diversity improves performance of an image classifier on ImageNet, with no need to first fine-tune the diffusion model to ImageNet itself.

Synthetic data generation has also been applied in crisis informatics, albeit to a limited extent. Rui et al. (2021) developed a GAN-based approach to produce multi-disaster remote sensing images, alleviating class imbalance in building-damage classification. Eltehewy et al. (2023) combined GAN-synthesised disaster images with real data, resulting in improved classification accuracy ($\sim$11% boost over baseline). Similarly, synthetic flood scenarios have been

used to train flood detectors with performance comparable to those trained on real images (Kang et al., 2025), and diffusion-generated images have improved models' ability to recognise fire and smoke under various conditions (Park and Lee, 2024).

2.6 Research Gap and Contributions

A persistent challenge across crisis imagery datasets is severe class imbalance, driven by difficulties and ethical constraints in collecting sufficient real-world data (Eltehewy et al., 2023). Standard CNNs often perform well on dominant classes but struggle significantly with rare event categories. Traditional augmentation methods, including geometric transformations and oversampling techniques, provide limited diversity and fail to adequately account for domain shifts, restricting generalisability and robustness (Alimisis et al., 2025).

Our review of the literature shows that diffusion-based data augmentation, despite promising early results in other domains (Shao et al., 2024; Yu et al., 2025), remains under-explored in the field of disaster image classification. Moreover, the effectiveness demonstrated by large multimodal models in few-shot or even zero-shot tasks, though with limitations (e.g., Scheele et al., 2024), begs the question of whether traditional CNN-based approaches are still needed. Such a comparison has never been directly performed in the context of disaster image classification and remains an open question in the field.

Our research specifically addresses these gaps through the following key contributions:

- We introduce and assess a pipeline using off-the-shelf diffusion models and LLM-generated prompts for synthetic data augmentation on the multi-task MEDIC dataset, aiming to mitigate class imbalance and enhance data diversity without requiring diffusion model fine-tuning.
- We conduct the first direct comparison in disaster image classification between CNNs fine-tuned with our synthetic data pipeline and the zero-shot performance of large multimodal foundation models on the same benchmark tasks.

Validating these techniques on a heterogeneous dataset like MEDIC would demonstrate their broader applicability, providing practitioners a practical tool to enhance predictive performance amid rapidly evolving crisis scenarios.

2.7 Ethical Considerations

2.7.1 Privacy and Data Protection

The MEDIC dataset contains images uploaded to social media platforms under varying degrees of user consent. Individuals depicted in these images did not necessarily anticipate that their content would form part of a research corpus. To mitigate potential privacy violations and avoid re-identification risks, all real images from the MEDIC dataset will be deleted at the end of this project. Only the metadata and derived features (e.g., embeddings) essential for the analysis will be retained. At no point in this study are there attempts to identify specific individuals or locations. We do not annotate images with personally identifiable information, and all analyses remain at the aggregate level.

2.7.2 Synthetic Imagery and Misinformation

The rise of high-fidelity generative models makes it increasingly easy to create or alter images that appear authentic. While synthetic data can improve model robustness—especially for underrepresented disaster classes—it carries the risk of fuelling misinformation if misapplied. To address this, we will only share the synthetic dataset with trusted parties who have justified use cases, such as academic researchers or humanitarian organisations. Further, our approach to text-to-image generation avoids producing distressing content beyond the scope of legitimate disaster scenarios. Fallback prompts ask the AI to 'tone down' anything that becomes graphic or borders on sensationalism (Section 6.2).

2.7.3 Bias and Fairness

Generative models inherit biases from their training data, which is often skewed towards certain geographic regions, socio-economic groups, or disaster types. Consequently, synthetic data may perpetuate or even amplify such biases. Our original dataset is already extremely diverse in terms of geography and situations. We further this diversity by attempting to generate even broader synthetic samples that may generalise to new scenarios. We implement "diversity" keyword strategies to capture a broad, equitable range of disaster contexts without sensationalising human suffering (Section 6.2).

2.7.4 AI Safety and Accountability

Applications of AI in crisis informatics have far-reaching societal implications, including life-or-death decisions about resource allocation or evacuation. Even small errors in classification can exacerbate vulnerabilities. Hence, we adopt the following safety practices:

1. Human-in-the-Loop: We emphasise that automated classification outputs must be verified by human experts before informing critical decisions. The final deployment of any model in a disaster-response context should embed safety checks.
2. Traceability: Our data processing and model training pipelines are documented in detail, enabling reproducibility and auditing. Traceability helps stakeholders understand, replicate, and, if necessary, contest the model's outputs.

Chapter 3

General Methodology

This chapter outlines our approach to investigating image classification improvements in crisis informatics through both synthetic data augmentation and zero-shot classification methods. The code used for this project is publicly available and open-sourced in the following GitHub repository: `https://github.com/evammun/genai-data-aug-disasters`.

3.1 Methodological Overview

The project follows a multi-stage methodology, where each step corresponds to a chapter of this report:

1. Establishing CNN baselines on the standard MEDIC dataset (Alam et al., 2023) and analysing their performance limitations (Chapter 4).
2. Addressing identified data quality issues through a systematic, conservative relabelling process (Chapter 5).
3. Developing and evaluating a synthetic data augmentation pipeline to fine-tune CNNs, aiming to improve performance on challenging classes (RQ1, Chapter 6).
4. Implementing and assessing a zero-shot classification approach using large multimodal models (LMMs), bypassing traditional training pipelines (RQ2, Chapter 7).

3.2 Dataset and Preparation

3.2.1 MEDIC Dataset

Our experiments utilise the MEDIC dataset (Alam et al., 2023), introduced in Section 2.3. The dataset contains 71,198 social media images annotated across four distinct classification tasks: disaster type, informativeness, humanitarian category, and damage severity. The images are divided in a *training* set (49,353), *validation* set (6,157) and *test* set (15,688).

3.2.2 Dataset Relabelling

Initial baseline analyses (Chapter 4) indicated potential label noise impacting performance, particularly for ambiguous categories like *mild* damage severity or *other disaster*. To establish

a more robust ground truth for evaluating our primary research questions, we implemented a conservative relabelling procedure (detailed in Chapter 5). This involved identifying images where multiple baseline CNNs disagreed with the original label, followed by independent verification using two large multimodal models (GPT-4o and Claude Sonnet 3.5) as unbiased judges. Approximately 5% of the dataset labels were revised through this process. All subsequent experiments, including synthetic data augmentation and zero-shot evaluation, were performed using this relabelled version of the MEDIC dataset unless otherwise noted.

3.3 Evaluation Framework

For CNNs trained in a multi-task setting, the model simultaneously predicts labels for all four tasks. Training optimises a combined loss function based on the sum of cross-entropy losses for each task (unweighted; see Appendix A.1). For zero-shot classification with large multimodal models, models are prompted to return valid JSON with a label for each task (see Appendix E).

Performance is assessed using standard classification metrics: accuracy, macro-averaged F1 score (to account for class imbalance), and per-class F1 scores (detailed in Appendix A.2). We often employ confusion matrices to diagnose error patterns (Chapter 6, 7).

3.4 Experimental Pipeline

Our investigation proceeds through the following experimental stages:

1. **CNN Baseline Implementation (Chapter 4)**: We first replicate and analyse the performance of standard CNN architectures (ResNet50, EfficientNet-B1, MobileNet-V2) fine-tuned on the original MEDIC dataset, following the methodology of Alam et al. (2023). This establishes benchmark performance and highlights limitations, motivating the relabelling (Chapter 5) and subsequent experiments.
2. **Synthetic Data Generation and Augmentation (RQ1, Chapter 6)** After preliminary experiments to determine the best setup, we develop a pipeline using LLMs to generate captions and diffusion models to synthesise diverse disaster images. These images augment the relabelled training set, targeting underperforming or critical classes. We then fine-tune the CNN models on this augmented dataset and evaluate performance changes.
3. **Zero-Shot Classification with Multimodal Models (RQ2, Chapter 7)** We evaluate the ability of off-the-shelf large multimodal models to perform the MEDIC classification tasks without any task-specific training ('zero-shot'). This involves querying the model with test images and an optimised prompt engineered through preliminary testing on a smaller validation set. Performance is compared directly against the fine-tuned CNN benchmarks.

3.5 Experimental Setup and Implementation Details

Local experiments were primarily conducted using an NVIDIA RTX 3070 GPU for training/fine-tuning. Our optimised CNN training pipeline, leveraging NVIDIA DALI for data loading,

achieved significant speedups (7-18× faster per epoch) compared to reference implementations (Table 3.1). For full reproducibility, the complete hardware and software configuration is provided in Appendix A.3 and A.4. Experiment-specific details are provided in the following chapters and associated appendices.

Table 3.1: Comparison of Training Times.

Model	**Alam et al. (2023)**		**Our Setup with CUDA/DALI**		**Speedup/Epoch**
	Time (hrs)	**Epochs**	**Time (hrs)**	**Epochs**	
EfficientNet-B1	74.22	150	2.778	39	6.97×
MobileNet-V2	76.67	150	1.111	39	18.25×
ResNet50	77.60	150	1.667	24	7.49×

Chapter 4

CNN Baseline

To create a reliable reference point for our experiments, we first replicated the baseline Convolutional Neural Network (CNN) training methodology described by Alam et al. (2023). This process involves fine-tuning pretrained CNN models on the MEDIC dataset (Alam et al., 2023). This chapter describes the methodology, the findings from our replication study, and our in-depth analyses that hint at potential issues with the MEDIC dataset itself.

4.1 Baseline CNN Methodology

Our experimental setup follows Alam et al. (2023) as accurately as possible, according to the available information in the published papers. The goal is to fine-tune pre-trained CNNs on the MEDIC dataset to perform multi-task classification across four distinct categories: disaster type, informativeness, humanitarian category, and damage severity. The core parameters and components of this baseline methodology are summarised below in Table 4.1.

Table 4.1: Baseline CNN Architectures and Hyperparameters Replicating Alam et al. (2023).

Parameter	Details
Base Architectures	ResNet50, EfficientNet-b1, MobileNet-v2 (pretrained)
Output Heads	4 separate linear logits heads (one per task: disaster type, informativeness, humanitarian category, and damage severity)
Optimiser	Adam
Initial Learning Rate	1×10^{-5}
Learning Rate Schedule	Reduce to 1×10^{-6} if validation loss does not improve for 10 epochs (initial patience)
Training Termination	Stop training if validation loss does not improve for a further 10 epochs (final patience)
Model Selection	Epoch yielding the highest F1 score on the validation set
Batch Size	32
Loss Function	Sum of Cross-Entropy losses for each task (equally weighted)

For our replication experiments, we focused on three base CNN architectures highlighted in Alam et al. (2023) as top performers for varying model sizes: ResNet50, EfficientNet-b1,

and MobileNet-v2. For each architecture, the pretrained model from PyTorch was adapted by replacing the final classification layer with four separate linear heads outputting logits for each classification task. This modification allows a single forward pass through the network to generate predictions for all four MEDIC tasks simultaneously.

As an efficiency enhancement over standard CPU-based methods, for this project we implemented a custom-written NVIDIA DALI pipeline to efficiently handle image loading, decoding, and normalisation for optimal GPU utilisation. The multi-task labels were initially encoded as a single integer and subsequently unpacked into four distinct task labels during the training process.

This established baseline serves as the benchmark against which the effectiveness of subsequent data relabelling (Chapter 5) and synthetic data augmentation techniques (Chapter 6) will be measured.

4.2 Baseline Results and Failure Analysis

Each of our three CNN architectures reaches performance levels very close to those reported by Alam et al. (2023). Table 4.2 summarises the overall Accuracy and F1 scores for each of the four classification tasks (left column) and the same metrics for the original MEDIC study (right) for EfficientNet-b1, confirming that our re-implementation broadly reproduces the published results. Full results are reported in Appendix B.1.

Task	**Metric**	**Our Impl.**	**MEDIC (Alam et al., 2023)**
Disaster Types	Accuracy	81.9%	81.4%
	F1	80.2%	79.8%
Informativeness	Accuracy	88.6%	88.6%
	F1	88.6%	88.6%
Humanitarian	Accuracy	85.0%	84.6%
	F1	84.6%	84.3%
Damage Severity	Accuracy	83.1%	82.9%
	F1	80.6%	80.8%

Table 4.2: Comparison of Baseline EfficientNet-b1 Performance with Original MEDIC Benchmarks.

While the global classification results look reasonably good, it is important to understand the exact failure modes of the CNNs. In the rest of this section, we study the classification results in detail and eventually surface potential issues with the dataset.

4.2.1 Class-Level Results

Accuracy and F1 scores at the **class level** for each model are reported in Table 4.3. Some representative findings:

- **Damage Severity**: The *none* category is classified correctly at a high rate (F1 $\approx$ 91%), whereas *mild* is a challenging label, with F1 generally <20%. The confusion matrices

(Figure 4.1) show that mild-damage images are frequently over-predicted as either none or severe, suggesting subtle visual cues that the network struggles to capture.

- **Humanitarian Category**: While *not humanitarian* is recognised with >90% F1, the classes *rescue/volunteering* and *affected/injured* have F1 ≈ 40–50%. This indicates that scenes of injuries, donation efforts, or volunteers are visually heterogeneous, making them harder to learn from relatively sparse training examples.
- **Disaster Type**: The *none* label (i.e., non-disaster images) is predicted accurately in most cases. However, *other disaster* is commonly misclassified, partly because it is a broad catch-all category. Some images labelled other disaster are similar to more common Disaster Types (e.g. floods), causing confusion in the model.
- **Informativeness**: The model distinguishes *informative* from *not informative* with high accuracy and F1 (≈ 87–89%). Instances of misclassification typically involve ambiguous content where the presence of disaster-related information is subtle (e.g., an image containing only textual overlays or vague scenes).

Confusion Matrix Analysis

Figure 4.1 presents the confusion matrix for the four classification tasks for EfficientNet-B1, the best-performing CNN.

Informative

True \ Predicted	not inf	inf
not inf	.88	.11
inf	.12	.87

Disaster Types

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.79	.01	.00	.03	.01	.12	.00
fire	.03	.82	.01	.01	.0	.11	.00
flood	.01	.00	.78	.03	.01	.14	.00
hurr.	.09	.01	.07	.59	.01	.19	.00
land.	.09	.01	.03	.06	.67	.12	.00
none	.01	.00	.01	.02	.00	.94	.00
other	.20	.05	.03	.12	.01	.46	.10

Damage Severity

True \ Predicted	none	mild	severe
none	.94	.00	.05
mild	.40	.05	.54
severe	.15	.01	.83

Humanitarian

True \ Predicted	injured	infra	not hum	rescue
injured	.32	.27	.33	.05
infra	.01	.84	.13	.01
not hum	.00	.06	.91	.01
rescue	.06	.27	.31	.35

Figure 4.1: Confusion matrices for EfficientNet-b1 using the baseline CNN training.

Table 4.3: Performance comparison of CNN architectures.

	Accuracy (%)			F1 Score (%)		
Task/Class	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**
Damage Severity	**82.7%**	**83.1%**	**81.9%**	**79.6%**	**80.6%**	**78.9%**
Little Or None	88.3%	89.2%	87.6%	91.3%	91.9%	90.7%
Mild	90.0%	89.9%	90.1%	9.3%	15.4%	9.8%
Severe	87.1%	87.1%	86.1%	76.3%	76.4%	75.0%
Informative	**88.2%**	**88.6%**	**87.1%**	**88.2%**	**88.6%**	**87.2%**
Not Informative	88.2%	88.6%	87.1%	89.1%	89.2%	87.8%
Informative	88.2%	88.6%	87.1%	87.2%	87.9%	86.4%
Humanitarian	**84.4%**	**85.0%**	**83.6%**	**83.6%**	**84.6%**	**82.4%**
Affected/Injured People	96.4%	96.2%	96.2%	42.4%	47.3%	43.6%
Infrastructure Damage	88.6%	89.2%	87.7%	83.1%	84.3%	82.4%
Not Humanitarian	87.9%	88.8%	87.4%	89.8%	90.4%	89.3%
Rescue/Volunteering	95.9%	95.7%	95.9%	42.6%	44.0%	26.3%
Disaster Types	**80.8%**	**81.9%**	**79.4%**	**78.6%**	**80.2%**	**76.6%**
Earthquake	94.1%	94.3%	93.6%	75.5%	76.6%	74.1%
Fire	98.2%	98.1%	97.7%	79.7%	79.4%	74.8%
Flood	96.4%	96.8%	96.4%	78.2%	81.1%	78.4%
Hurricane	93.0%	93.2%	92.6%	62.5%	65.3%	60.7%
Landslide	98.6%	98.6%	98.5%	67.6%	69.0%	66.1%
Not Disaster	88.1%	89.4%	87.3%	90.0%	90.9%	89.3%
Other Disaster	93.2%	93.5%	92.8%	18.9%	26.2%	5.3%

RN50: ResNet50, EN-B1: EfficientNet-B1, MN-V2: MobileNet-V2

Performance comparison of CNN architectures. The table shows metrics for each classification task and class. For tasks (in **bold**), accuracy represents multi-class classification performance across all classes, while F1 score is the weighted average across classes. For individual classes, accuracy shows binary classification performance (how well the model distinguishes that class from all others), and F1 score measures the harmonic mean of precision and recall for that specific class. The best performing score in each row is highlighted in green.

1. **Damage Severity**: *Mild* images regularly migrate into the *none* or *severe* bins. Visual inspection of images suggests that mild damage can be subtle (e.g., small cracks or limited debris), so the network either detects no visible damage or infers an evidently severe scene.
2. **Informativeness**: The distinction between *informative* and *not informative* is mostly clear. However, text-heavy images or partial views of crowds with no clear context occasionally lead to misclassifications, indicating an underlying semantic ambiguity.
3. **Humanitarian Categories**: Misclassifications between *affected/injured* and *infrastructure damage* often occur, implying that images capturing both people in distress and damaged buildings can confuse the model. Additionally, rescue is sometimes predicted as not humanitarian if the scene does not overtly show rescue equipment or volunteers in uniform.
4. **Disaster Types**: Large-scale events like landslides, floods and hurricanes share visual similarities (e.g., water inundation vs. wind damage), and the broad *other disaster* label suffers from insufficiently distinct features. That category is typically misclassified as more common types—especially quake, hurricane, or none—whenever the image lacks strong visual cues.

4.2.2 Problematic Classes and Patterns

A closer, class-by-class look at the results highlights several **underrepresented** and **highly confused** labels that systematically degrade performance. Table 4.4 (below) compiles aggregated frequency and performance metrics (F1, Precision, Recall) for all classes across our three baseline CNNs (ResNet50, EfficientNet-b1, and MobileNet-v2).

Underrepresented Classes and Imbalance Effects. From Table 4.4 we see that classes with a lower share of the dataset (e.g., *landslide* at 2%, *hurricane* at 10%, *mild* damage at 10%, *rescue volunteering* at 4%) typically see poorer F1 scores.

However, **imbalance is not the sole factor**: the *other disaster* label, which has a modest 7% share, can achieve strong precision (e.g., up to 79.4%) but simultaneously suffers recall as low as 2.8%. This suggests that *other disaster* is visually ambiguous, covering a heterogeneous set of scenarios (accidents, conflict, chemical spills, volcanic eruptions, etc.).

Error Correlations and Multi-Task Interactions. Figure 4.2 shows a **class-level error correlation matrix**, where each cell reflects how often errors in one category (rows) overlap with errors in another (columns). For instance, a red cell at *(mild, other disaster)* indicates that whenever the network misclassifies mild damage, it often also misclassifies the corresponding image's disaster type as *other* (or vice versa). Notable patterns include:

- **Mild Damage & Hurricane/Landslide**: In the damage severity panel, the cells corresponding to (`mild`, `hurr.`) and (`mild`, `land.`) show very high correlation (0.54), indicating that misclassifying a scene with slight structural damage frequently co-occurs with misclassifications of hurricane and landslide disaster types.
- **Not Informative & Other Disaster**: Misclassifications in the informative category *not inf* strongly co-occur with errors for *other* disaster type (0.65), suggesting the CNN struggles to identify informative content in less common disaster scenarios.

Task	Class	Frequency	F1 (%)	Precision (%)	Recall (%)
Damage Severity	Little or None	65%	90.7–91.9	88.8–90.2	92.5–94.0
	Severe	25%	75.0–76.4	67.9–70.3	83.4–84.3
	Mild	10%	9.3–15.4	41.2–42.5	5.2–9.4
Disaster Types	Not Disaster	57%	89.3–90.9	85.7–88.4	93.2–94.5
	Earthquake	11%	74.1–76.6	68.8–71.9	79.7–81.9
	Hurricane	10%	60.7–65.3	62.2–65.3	59.2–65.7
	Flood	8%	78.2–81.1	78.1–80.7	78.1–81.5
	Other Disaster	7%	5.3–26.2	68.1–79.4	2.8–15.7
	Fire	4%	74.8–79.7	71.7–77.3	78.1–83.0
	Landslide	2%	66.1–69.0	63.5–67.5	67.7–75.5
Humanitarian	Not Humanitarian	58%	89.3–90.4	88.2–90.7	90.1–91.5
	Infrastructure Damage	33%	82.4–84.3	78.6–81.7	84.6–87.1
	Rescue/ Volunteering	4%	26.3–44.0	50.1–57.7	17.1–39.3
	Affected/ Injured People	4%	42.4–47.3	54.8–60.1	32.7–41.6
Informativeness	Not Informative	54%	87.8–89.2	89.3–91.4	85.6–88.8
	Informative	46%	86.4–87.9	84.0–86.9	87.4–90.4

Table 4.4: Aggregated Class Frequency & Performance Metrics.

Note: The ranges shown for performance scores represent the lowest and highest across our trained models (Across ResNet50, EfficientNet-b1, and MobileNet-v2)

- **Earthquake & Not Informative**: There is notable correlation (0.44) between earthquake misclassifications and non-informative content errors, indicating the CNN has difficulty distinguishing between earthquake imagery and non-informative disaster content.

These high-correlation pairs suggest that **single-task errors are not independent**. In many cases, confusion in one domain (e.g., damage severity) influences or co-occurs with confusion in another (e.g., disaster type or humanitarian category).

4.2.3 Interim Conclusions

Taken together, our in-depth analyses of the results from our replication of Alam et al. (2023) point to class imbalance, label ambiguity (mild damage, other disaster, etc.), and visually subtle features as the main drivers of error.

In subsequent chapters, we address these limitations first by **re-labelling** problematic classes (Chapter 5) and then by **synthetic data augmentation** (Chapter 6), aiming to bolster the training set with more diverse and properly labelled examples. Our study of the error patterns and their correlations will inform our synthetic image allocation logic.

		Disaster Types							Informative		Humanitarian			
		quake	fire	flood	hurr.	land.	none	other	not inf	inf	injured	infra	not hum	rescue
Informative	not inf	.44	.08	.20	.54	.54	.19	.65	.11	.09	.15	.12		
	inf	.02	.02	.01	.15	.15	.12	.19	.03	.01	.00	.00		
Humanitarian	injured	.15	.11	.15	.06	.06	.05	.15	.11	.03				
	infra	.09	.02	.07	.07	.07	.00	.09	.09	.01				
	not hum	.16	.08	.12	.12	.12	.03	.15	.15	.00				
	rescue	.12	.13	.09	.16	.16	.00	.14	.12	.00				
Damage Severity	none	.05	.04	.05	.03	.03	.01	.04	.04	.02	.09	.00	.06	.07
	mild	.34	.10	.12	.54	.54	.08	.36	.43	.04	.17	.14	.21	.29
	severe	.13	.09	.10	.09	.09	.09	.12	.12	.03	.23	.09	.00	.07

Figure 4.2: Error correlations between task classes. Shown is EfficientNet-b1. Resnet50 and Mobilenet-v2 produced very similar results.

Chapter 5

Relabelling the MEDIC Dataset

From the analyses in the previous chapter, after reviewing the similarities between the error matrices and correlations between all three CNN models (ResNet50, EfficientNet-b1, MobileNet-v2), we drew two conclusions. First, some classes may simply be too heterogeneous to correctly generalise (e.g., *other disaster*).

Second, and more concerningly, there may be genuine issues with the human labelling in the original dataset. Alam et al. (2023) list the agreement scores shown in Table 5.1, which reflect some legitimate challenges in the annotation process. As the authors noted, certain scenarios present genuine ambiguity—such as hurricane-induced flooding or images showing both building damage and rescue efforts—where annotator interpretation naturally varies (Alam et al., 2023).

Tasks	Fleiss (κ)[1]	Krip. (α)[2]	Avg agg.[3]
Disaster types	0.46	0.46	0.70
Humanitarian	0.52	0.52	0.73
Informativeness	0.71	0.71	0.91
Damage severity	0.55	0.55	0.79

Table 5.1: Crowdsourced annotation agreement for each task, as presented by Alam et al. (2023).

While our initial plan did not explicitly include a reassessment of the published MEDIC dataset, our project hinges on having reliable labels for existing images. Class-based synthetic data augmentation from wrongly labelled images would simply amplify existing errors, undermining the validity of our study. Thus, this chapter represents a necessary detour where we describe our robust relabelling approach.

5.1 Data Relabelling Methodology

We first began by identifying instances of likely misclassifications. Our assumption is that if a label was evidently misclassified, a first indication would be that different trained CNNs would

[1]Fleiss' κ assesses reliability among multiple raters, accounting for agreement occurring by chance, with values ranging from -1 to 1 (higher indicating stronger agreement).

[2]Krippendorff's α is a reliability coefficient suitable for various data types and missing values, also ranging from -1 to 1.

[3]'Avg agg.' represents the average aggregated score across all annotators.

agree on the alternative label.

Table 5.2 shows how often only 1, 2, or all 3 CNNs converged on the same "alternative" label whenever they disagreed with the ground-truth label. Notably, in about 40% of the misclassified cases, all three networks proposed the same alternative label. Considering that tasks have from 2 to 7 classes, this level of unanimity in the misclassified responses statistically significantly deviates from chance and deserves further investigation.

Label	1	2	3	All	Relabelled
Little or None	1,421 (43%)	912 (27%)	1,001 (30%)	3,334	1,157 (35%)
Mild	362 (6%)	1,603 (27%)	3,997 (67%)	5,962	2,537 (43%)
Severe	1,713 (47%)	991 (27%)	970 (26%)	3,674	801 (22%)
Damage Severity	**3,496 (27%)**	**3,506 (27%)**	**5,968 (46%)**	**12,970**	**4,495 (35%)**
Earthquake	876 (38%)	642 (28%)	790 (34%)	2,308	485 (21%)
Fire	280 (40%)	206 (29%)	221 (31%)	707	133 (19%)
Flood	592 (38%)	459 (29%)	522 (33%)	1,573	308 (20%)
Hurricane	1,022 (27%)	914 (24%)	1,810 (48%)	3,746	1,154 (31%)
Landslide	253 (33%)	218 (29%)	286 (38%)	757	120 (16%)
Not Disaster	1,547 (49%)	755 (24%)	858 (27%)	3,160	1,075 (34%)
Other Disaster	314 (10%)	967 (31%)	1,860 (59%)	3,141	1,321 (42%)
Disaster Types	**4,884 (32%)**	**4,161 (27%)**	**6,347 (41%)**	**15,392**	**4,596 (30%)**
Affected Injured or Dead People	580 (27%)	611 (29%)	947 (44%)	2,138	755 (35%)
Infrastructure and Utility Damage	1,965 (45%)	1,203 (28%)	1,200 (27%)	4,368	1,256 (29%)
Not Humanitarian	1,783 (41%)	1,107 (25%)	1,466 (34%)	4,356	1,706 (39%)
Rescue Volunteering or Donation Effort	485 (18%)	752 (28%)	1,450 (54%)	2,687	207 (8%)
Humanitarian	**4,813 (36%)**	**3,673 (27%)**	**5,063 (37%)**	**13,549**	**3,924 (29%)**
Informative	2,271 (40%)	1,390 (24%)	2,043 (36%)	5,704	730 (13%)
Not Informative	2,177 (47%)	1,247 (27%)	1,252 (27%)	4,676	1,175 (25%)
Informative (all)	**4,448 (43%)**	**2,637 (25%)**	**3,295 (32%)**	**10,380**	**1,905 (18%)**
All Tasks	**17,641 (34%)**	**13,977 (27%)**	**20,673 (40%)**	**52,291**	**14,920 (29%)**

Table 5.2: Largest Alternative label agreement among three CNNs in misclassified cases, broken down by label. Columns 1, 2, 3 indicate the number of CNNs (out of three) that agreed on the single most frequent alternative label when the original label was misclassified. "All" is the total misclassifications, and "Relabelled" indicates how many were ultimately corrected via the LLM-as-judge process.

5.1.1 Relabelling Using LLM-as-a-Judge

The "LLM-as-a-judge" is a modern approach that leverages language models to review and potentially correct existing data labels, addressing known issues with human label quality (Northcutt et al., 2021; Vasudevan et al., 2022). Studies demonstrate LLMs can identify 6-21% of label errors, with higher-confidence disagreements correlating strongly with true errors (Nahum et al., 2025). These studies suggest that benchmark evaluations had previously underestimated model capabilities due to label noise. Furthermore, LLMs such as GPT-3.5/4 can achieve annotation quality comparable or superior to human crowd workers (Gilardi et al., 2023).

However, deploying LLMs as judges requires mitigating their inherent biases, such as position and verbosity bias observed when scoring responses, where presentation order or exact prompt phrasing influences the outcome (Zheng et al., 2023). Another concern is short-cutting, where LLMs just answers without reasoning. A key mitigation strategy involves *structured* prompting: first requiring the LLM to describe the input (e.g., an image) based on a template,

and then strictly defining the output format and permissible labels. This forces evidence-based decision-making, reduces deviation, and ensures parseable outputs (Tam et al., 2024).

5.1.2 Relabelling Pipeline

Drawing insight from the literature on LLM-as-judges, we set up a four-step relabelling process (Figure 5.1).

In brief, each image was labelled independently by three CNNs (ResNet50, MobileNetV2, and EfficientNet-B1). If at least two models converged on an alternate label, we flagged the image as potentially requiring relabelling. As expected, these were mainly images from ambiguous classes such as *mild* or *other disaster*. The flagged image was then provided to two LLMs (GPT-4o and Claude Sonnet-3.5) together with a structured prompt (see Appendix C). If *both LLMs* agreed on an alternative label for a task, we updated the annotation accordingly; otherwise, the original label was retained.

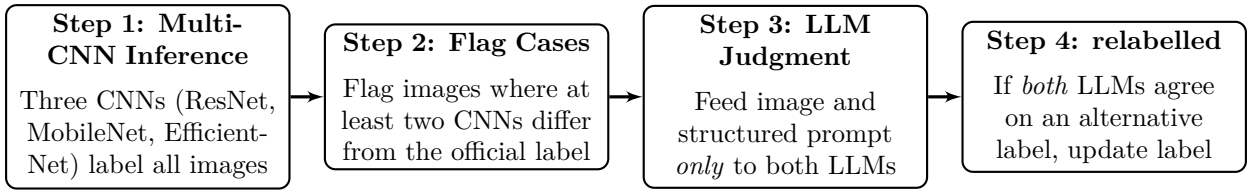


Figure 5.1: Relabelling pipeline. Only images with strong CNN-level disagreement are passed to the LLM. The LLM sees *no* original labels or CNN outputs, only the image and prompt.

Our approach is conservative, in that first multiple CNNs need to agree on an alternative label, and then both LLMs-as-judges need to agree on the relabelling. We highlight that to mitigate potential biases, we did *not* feed the original label nor the CNN majority vote to the LLMs. Moreover, our pipeline includes both CNNs and LLMs, thus limiting biases towards one of the two methods.

As shown in Table 5.2, 29% of labels that were originally misclassified were reclassified (5% of the total labels).

5.2 Relabelled Dataset CNN Classification Results

We compare model performance on the *original* (pre-relabelling) dataset versus the *relabelled* version. We focus on EfficientNet-B1 as the best-performing architecture in the original experiments (see Chapter 4.2.1). For clarity, we provide two summary tables: one captures overall task-level metrics, and the other highlights several specific classes where we observed the largest improvements.

Task-Level Performance. Table 5.3 shows how EfficientNet-B1's accuracy and macro F1 scores improved after relabelling. We see notable gains in all tasks.

The largest improvement appears in *Damage Severity*, which rose from 80.6% to 84.3% in F1, and in *Disaster Types*, improving from 80.2% to 83.2%. These tasks contained some of the most problematic classes in the original dataset (e.g. *mild* for *damage severity* and *hurricane* within *Disaster Types*), suggesting that relabelling resolved substantial confusion.

Task	Original		Relabelled	
	Accuracy	**F1**	**Accuracy**	**F1**
Damage Severity	83.1%	80.6%	85.5%	84.3%
Informative	88.6%	88.6%	90.2%	90.2%
Humanitarian	85.0%	84.6%	86.7%	86.3%
Disaster Types	81.9%	80.2%	84.5%	83.2%

Table 5.3: EfficientNet-B1 task-level performance on original vs. relabelled dataset.

Key Improvements. Table 5.4 focuses on four classes (*mild*, *rescue/volunteering effort*, *affected/injured people*, and *hurricane*) that saw large benefits from relabelling. Each row shows the original vs. relabelled F1 score for EfficientNet-B1.

Class	F1 (Original)	F1 (Relabelled)
Mild (Damage Severity)	15.4%	35.8%
Rescue/volunteering effort (Humanitarian)	44.0%	57.5%
Affected/injured people (Humanitarian)	47.3%	53.6%
Hurricane (Disaster Types)	65.3%	72.5%

Table 5.4: Selected class-level improvements for EfficientNet-B1 (original vs. relabelled).

Notably, *mild* more than doubled its F1 from 15.4% to 35.8%. This class is one of the most error-prone, likely due to the ambiguity of the term "mild" (*partially damaged but still usable*), as well as the fact that it is severely underrepresented.

Rescue/volunteering effort also receives a substantial boost, from 44.0% to 57.5% F1. Qualitative analysis showed that many of the mislabelled images were actually *not humanitarian*, but might simply depict people in them.

One notable exception to this gain in performance is *other disaster*, our enfant terrible, whose F1 declined by 1 point after relabelling.

Full results per class and model are provided in Appendix C.2

Confusion Matrices. After applying our LLM-based relabelling (Figure 5.2), we see fewer *mild* vs. *none* confusions, suggesting the partial-damage scenarios are now more clearly delineated. Similarly, *rescue* exhibits fewer misclassifications, reflecting better discrimination from the *injured* and *not humanitarian* classes. However, *other disaster* remains strongly conflated with *quake* or *none*, although it is unclear whether this remains a labelling issue or a limitation of CNN models.

5.3 Relabelling Remarks

Overall, we found that the revised labels seem to help clarify ambiguous samples, leading to higher accuracy and more coherent decision boundaries for key tasks. This pattern aligns with literature suggesting that correcting a relatively small fraction of systematically incorrect labels can yield disproportionate benefits in a model's downstream performance (Nahum et al., 2025).

While significant gains were observed, certain inherently heterogeneous classes like *other disaster* remained challenging, suggesting limitations of our relabelling approach or that label noise was only one factor limiting performance for such categories.

Finally, we note that our intervention is far from exhaustive and there are many areas for improvement. However, a thorough reassessment of the MEDIC dataset would constitute a project in itself, beyond the scope of the current work. Our goal here was to address major labelling issue with a minimal, conservative intervention so as to proceed with our main experiments involving synthetic data augmentation, described in the next chapter.

Figure 5.2: Confusion matrices comparing EfficientNet-b1 trained on original (left) and relabelled (right) datasets.

Disaster Types (Original)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.81	.01	.00	.04	.01	.09	.00
fire	.03	.83	.00	.01	.00	.10	.00
flood	.01	.00	.81	.05	.01	.10	.00
hurr.	.08	.01	.06	.65	.02	.15	.00
land.	.07	.00	.02	.06	.75	.07	0
none	.01	.00	.01	.02	.00	.93	.00
other	.22	.05	.02	.11	.02	.38	.15

Disaster Types (Relabelled)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.86	.01	.00	.03	.00	.07	.00
fire	.03	.85	.00	.01	.00	.06	.01
flood	.02	.00	.80	.04	.01	.11	0
hurr.	.04	.02	.03	.73	.01	.13	.00
land.	.06	.01	.03	.06	.74	.07	0
none	.01	.00	.01	.02	.00	.93	.00
other	.24	.09	.02	.08	.01	.37	.14

Informativeness (Original)

True \ Predicted	not inf	inf
not inf	.87	.12
inf	.09	.90

Informativeness (Relabelled)

True \ Predicted	not inf	inf
not inf	.89	.10
inf	.08	.91

Humanitarian (Original)

True \ Predicted	injured	infra	not hum	rescue
injured	.41	.30	.21	.05
infra	.00	.87	.10	.01
not hum	.01	.06	.90	.01
rescue	.08	.30	.21	.39

Humanitarian (Relabelled)

True \ Predicted	injured	infra	not hum	rescue
injured	.50	.25	.16	.07
infra	.00	.89	.08	.01
not hum	.00	.05	.91	.01
rescue	.06	.24	.20	.49

Damage Severity (Original)

True \ Predicted	none	mild	severe
none	.93	.00	.05
mild	.34	.09	.56
severe	.13	.02	.84

Damage Severity (Relabelled)

True \ Predicted	none	mild	severe
none	.93	.01	.04
mild	.27	.26	.45
severe	.08	.03	.87

Confusion matrices comparing EfficientNet-b1 trained on original (left) and relabelled (right) datasets. The matrices show notable improvements in classification performance, particularly for the "mild" damage class (from 9% to 26%), "injured" class (from 41% to 50%), and "rescue" class (from 39% to 49%). Dark blue cells indicate correct classifications, light blue cells show mediocre performance, and red cells highlight problematic misclassifications.

Chapter 6

Synthetic Data Augmentation

Disaster response systems need to be prepared before emergencies occur. While our previous work established CNN baselines (Chapter 4) and improved performance and dataset reliability through relabelling (Chapter 5), we remain constrained by what existing data represents: a narrow spectrum of well-documented natural disasters.

This chapter examines whether synthetic imagery can address these gaps. In the following, we describe our pipeline that combines large language models and modern image generation models to create artificial disaster images that maintain semantic accuracy to their real-world counterparts. We then discuss our findings from fine-tuning CNNs for disaster image classification using our synthetic data augmentation pipeline.

6.1 Preliminary Experiments for Image Generation

While recent research demonstrates that synthetic data from text-to-image generative models can significantly reduce class imbalance and improve downstream classification performance (He et al., 2023), the effectiveness of such augmentation depends heavily on prompt alignment with intended labels. Multi-stage or structured prompting strategies typically achieve higher fidelity compared to single-pass approaches (Wei et al., 2022; Kojima et al., 2022; Chen et al., 2023; Yang et al., 2024).

Thus, we performed a set of preliminary experiments to find the best-performing setup for generating synthetic disaster images belonging to desired class labels. In particular, we evaluated three prompt designs (*Naïve*, *Structured*, and *Multistage*), three LLM captioners (Claude-3.7-sonnet, GPT-4o, Claude-3.5-haiku), and three diffusion models (Stable Diffusion-1.6, Stable Diffusion-3.5, Flux 1-dev), with three real images and one hypothetical scenario per class combination (360 examples). Full details are reported in Appendix D.1.

These tests identified Flux 1-dev with a multistage prompting approach using Claude 3.7 Sonnet or GPT-4o for captioning as the most effective combination. Still, even the best-performing combinations achieved $\sim 50 - 60\%$ alignment between the desired image label and the generated one, as judged by other LLMs.

Discussion. Although these small-scale tests provided useful insights, **the overall alignment remained disappointing** (50–60%). We identified three key issues:

1. **Prompt Length and Truncation.** Excessively detailed instructions often confused the diffusion models or were internally truncated, yielding incomplete or noisy outputs. This aligns with the broader literature cautioning against overly verbose prompts when chain-of-thought or multi-stage strategies are employed (Renze and Guven, 2024).
2. **Content Refusals.** Scenes depicting human suffering frequently triggered rejections from both the LLMs (during caption generation) and the diffusion model (during final synthesis).
3. **Lack of diversity in captions from labels.** When generating entirely *hypothetical scenarios* from label combinations (e.g. `damage=severe`, `disaster=earthquake`, `informative=informative`, `humanitarian=not humanitarian`), we found that the LLMs would produce near-identical descriptions across multiple requests, even with a relatively high sampling temperature (0.9). For example, every hypothetical earthquake scene resembled the same urban street in Turkey, viewed from a similar vantage point.

   This homogeneity is not uncommon in text-to-image systems, as generative models often default to their most probable, "familiar" instantiations of a given concept (He et al., 2023). Figure 6.1 shows two outputs generated with the same label combination, with similarities across the two highlighted. The third example uses our proposed solution, which we term *'diversity keywords'* (see below).

6.2 Prompt Refinement

We refined our prompting strategy with the insights gained from our preliminary experiments.

Fallback Prompts. To manage content rejections from LLMs or diffusion models, we implemented two fallback prompts. The first clarifies the academic and ethical context of our work, while the second requests a moderated description (e.g., referring generally to injured individuals needing medical aid rather than graphic injuries). If both fallbacks fail, we exclude that particular synthetic image. The exact fallback prompts appear in Appendices D.7 and D.8.

Diversity Keywords. We adopted a controlled keyword strategy to enhance diversity without deviating from core label requirements (e.g., avoiding improbable scenarios like hurricanes in polar climates). By selecting neutral dimensions—such as time of day, camera angles, broad geographic locations, and lexically constrained prompts—we introduced beneficial variability while maintaining scenario integrity (Yang et al., 2024). Letter-based prompts additionally mitigated repetitive token generation (Wei et al., 2022). Specifically, we injected variation in:

- **Time of Day:** *dawn, midday, dusk, evening, ...*
- **Camera Angle:** *aerial view, low angle, ground level, close-up, ...*
- **Letter-based Word Prompts:** e.g., *"include two distinct words starting with letter B"*
- **Broad Geography:** *Northern hemisphere, Southern hemisphere, West, East* (excluding specific countries)

These variations aimed primarily to disrupt default prompt outputs rather than introduce complexity for its own sake. Figure 6.1C demonstrates how diversity keywords significantly

(A) Example 1:

<caption> An aerial view of a 500m x 300m urban area in Turkey, following an earthquake. The image shows 7 completely collapsed multi-story residential buildings with pancaked concrete floors and 3 partially collapsed commercial structures with exposed internal floors. Roads have 0.5-1m wide cracks making them unusable, with exposed and leaking water pipes and downed power lines. 22 emergency responders in uniform operate 4 excavators and 2 cranes amid concrete rubble piles 3-6m high. 8 emergency vehicles (3 fire trucks, 5 ambulances) are parked nearby. Approximately 50 civilians stand in groups of 5-10 people, observing from a safe distance. Search and rescue operations are actively underway with responders using listening devices on the rubble and 4 search dogs with handlers searching through debris. 12 tents have been set up 50m from the collapsed buildings. The scene features a clear daytime sky with mountainous backdrop visible 5km in the distance and a mosque with a damaged minaret (top 3m broken off). Turkish language road signs are visible throughout the affected area. </caption>

(B) Example 2:

<caption> Aerial view of a 500m x 350m urban residential area in Izmir, Turkey following an earthquake. Four 5-7 story apartment buildings have completely collapsed with pancaked floors, while two others show partial structural failure. Concrete debris fields extend 20-30m from building foundations, with the largest measuring 28m x 35m. Ground fissures 0.5-1m wide cut across streets, and dust clouds hover 15-20m above collapse sites. Eight emergency vehicles (4 fire trucks, 3 ambulances, 1 police car) are positioned strategically around the site. Thirty-seven people are visible: 22 emergency responders in high-visibility gear operating with technical equipment including listening devices and search cameras, and 15 civilians gathered in a designated safe zone 50m from the collapsed structures. Two excavators are actively removing debris while four search and rescue teams work at different collapsed buildings. A ruptured water main has created an 8m diameter flooding area. Twelve downed power lines indicate electrical grid failure, and three roads are impassable due to debris. A makeshift medical triage area with 5 medical personnel has been established. The scene is set during clear daytime conditions with mountains visible 5km in the background and street signs in Turkish visible. </caption>

(C) Example 3: With use of diversity keywords: 'ground level', 'Southern hemisphere'

<caption> Ground-level view of a severely damaged urban residential area at dawn in the southern hemisphere, spanning approximately 120m x 80m. Three 4-story residential buildings have completely collapsed with visible pancaked floors and exposed rebar protruding 1-3m from concrete fragments. A water tower shows 70% structural failure, and two utility poles have severed with downed lines. The main road has a 50cm vertical displacement making it impassable, with 8 crushed vehicles visible. Ground fissures 30-40cm wide run through the pavement, and a perimeter wall has collapsed along a 35m stretch. A communication antenna is tilted at a 35-degree angle. Fourteen people are present: 6 in civilian clothing and 8 in uniform. Four emergency vehicles (2 fire trucks, 1 ambulance, 1 police vehicle) are positioned near a temporary 4m x 6m shelter structure. Emergency responders are conducting initial site assessment with 2 individuals using specialised listening equipment and another group establishing a communication system. Two dogs with handlers are navigating the debris following a systematic grid search pattern while local residents gather in a designated safe area. Dawn's pink-orange light illuminates the scene at approximately 15% ambient illumination, with southern hemisphere constellations still visible in the dark portions of the sky. Light fog at ground level limits visibility to about 80m, and dust particles reflect the early morning light. The temperature of approximately 12°C is evident from visible breath condensation as people examine building remnants with flashlights. </caption>

Figure 6.1: Three captions generated by Claude Sonnet-3.7 from similar task label combinations. Blue text indicates similar content between captions. Example 3, where the use of generic keywords has been implemented, shows significantly less overlap with Examples 1 and 2.

reduced redundancy between generated prompts. The third caption example, generated using diversity keywords, shows a marked reduction in similarity to the two prompts that do not.

Final Prompt Structure. By isolating the LLM's full internal reasoning in `<analysis>` while only exposing the diffusion model to a briefer `<caption>`, we mitigated many of the earlier pitfalls (see Figure 6.2). Although some rejections persisted, this final approach consistently produced more label-aligned images without excessive or dramatic text. The fallback prompts prevented outright failure in sensitive cases, thereby maximising the overall yield of synthetic data. See Appendix D.2.1 for the full prompts tested.

```
Captioning from images
<prompt>
Return:
</analysis>
... [Analysis section with minimal speculation + structured tags]
</analysis>
</caption>
... [Self-contained summary that we actually send to diffusion model]
</caption>

Captioning from labels
<prompt>
Return:
</analysis>
... [Analysis tags describing an imagined scene with given labels]
</analysis>
</caption>
... [Self-contained description for image generation]
</caption>

First fallback prompt
[Explains academic purpose, anonymity of people depicted, ethical standards]
Second fallback prompt
[Requests a toned-down version of the same scenario that aligns with content
   guidenlines]
```

Figure 6.2: Prompt template structure with various response types

Figures 6.3 to 6.5 show some examples of the images generated with this prompt. For 'From Image' scenarios, the image on the left is the original image from the training set. The caption generated in the `<caption>` output from the LLM is used to generate the image on the right with Flux 1-dev. For 'From Label' scenarios, the caption describes a purely hypothetical scenario generated by Claude Sonnet 3.7, which is used to generate the image on the right with Flux 1-dev. These have been selected using our random seed, and do not represent the 'best' or 'worst' of the images generated.

From image

Original image

Synthetic image

Ground-level view of a narrow urban alleyway covered in brick debris and rubble following an earthquake. Chinese rescue workers in bright red uniforms huddle in the foreground left, while damaged but standing buildings line both sides. Local businesses with visible signage remain partially intact as additional rescue personnel work in the background.

LLM caption

From labels

Ground level view of urban residential area at sunrise showing mild earthquake damage: zigzag cracks along building facades, tilted streetlights, and scattered debris. Volunteers in yellow vests clear rubble while others distribute supplies from tables. A silver ladder leans against a damaged wall as workers with battery-powered lanterns inspect eastern-facing structures.

LLM caption

Synthetic image

Figure 6.3: Comparison of image synthesis approaches for disaster imagery: generation from source images (top) versus generation from text labels (bottom). Original labels were *Damage Severity: mild; Informative: informative; Humanitarian: rescue or volunteering effort; Disaster Type: earthquake*

From image

Original image

Synthetic image

Ground-level photograph of an outdoor community gathering with people seated on wooden benches. A handwritten red sign reading "WORSHIP SERVICE HERE 10:00 SUNDAY"; leans against green shrubs. A temporary setup with canopies visible in background. Parked vehicles visible in distance. A dog rests near seated individuals on grassy ground.

LLM caption

From labels

Shot at eye-level of a person with a minor leg injury receiving first-aid in a western park at sunrise. Family members apply bandages from an open red medical kit on a wooden bench. Intact playground equipment and walking paths visible in background. Frost covers the grass in the morning light.

LLM caption

Synthetic image

Figure 6.4: Comparison of image synthesis approaches for disaster imagery: generation from source images (top) versus generation from text labels (bottom). Original labels were *Damage Severity: little or none; Informative: informative; Humanitarian: affected, injured or dead people; Disaster Type: not disaster*

From image

Original image

Synthetic image

Medium shot of a soldier in WWII combat uniform and helmet standing amid a burning environment. Orange-yellow flames engulf structures in the background, creating dramatic lighting. Dark silhouettes visible through smoke. This is a screenshot from the video game "Call of Duty: WWII" depicting a fictional war scene, not an actual disaster.

LLM caption

From labels

Panoramic view of an industrial complex ablaze in the early hours, Northern hemisphere. Intense inferno with 40-foot orange-red flames consuming warehouses. Severely damaged steel structures with collapsed roofs, melted support beams, and charred equipment. No people or emergency responders visible, just the isolated destruction.

LLM caption

Synthetic image

Figure 6.5: Comparison of image synthesis approaches for disaster imagery: generation from source images (top) versus generation from text labels (bottom). Original labels were *Damage Severity: severe; Informative: informative; Humanitarian: not humanitarian; Disaster Type: fire.* **This is a particularly interesting case where the captioning LLM detects that the original image is not, in fact real, but from a video game.**

6.3 Data Allocation Strategy

Considering budget constraints, for our final data augmentation step we can generate up to 10,000 synthetic images, a considerable addition to the MEDIC training set (∼50,000 images). Before we proceed, we have to decide how to assign our synthetic data to classes.

We have three goals in our allocation strategy:

1. Future-proof our model to generalise to novel, unseen situations.
2. Reduce confusion in worst-performing and underrepresented classes.
3. Improve classes that are *critical* to humanitarian rescue efforts, even if they already perform well.

Hence, improving our model's results is only *one* of our goals – the main ambition is to allow our models to better generalise to new, unseen scenarios, and especially in those that are critical to rescue.

We allocate images according to a heuristic formula designed according to the three goals above. In principle, our approach could be formalised following a decision-theoretical framework that models the influence of increasing dataset size on performance and assigns a different cost for each missclassification. However, assigning such costs and modelling the downstream performance effect is highly nontrivial. Thus, we leave devising a more sophisticated allocation strategy for future work.

We first allocate a "floor" of synthetic images proportionally to all classes, then distribute the remainder according to F1 performance and correlated error rates. We pay special attention to error *correlations*, knowing that misclassifications in one class lead to errors in others.

The allocation formula is:

$$\text{Images} = \text{Base} + \text{Weakness Bonus} + \text{Correlation Bonus} + \text{Impact Bonus} \tag{6.1}$$

where:

$$\text{Base} = 10$$

$$\text{Weakness Bonus} = \begin{cases} 80, & \text{per Critical task label} \\ 40, & \text{per Moderate task label} \\ 20, & \text{per Below Average task label} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Weakness Thresholds} = \begin{cases} \textit{Critical:} & \text{F1} < 40.0\% \\ \textit{Moderate:} & \text{F1} < 60.0\% \\ \textit{Below Average:} & \text{F1} < 75.0\% \\ \textit{Normal:} & \text{F1} \geq 75.0\% \end{cases}$$

$$\text{Error Correlation Bonus} = \begin{cases} 32, & \text{if Mild damage + specific Disaster Types} \\ 32, & \text{if Other Disaster + Informative} \\ 20, & \text{if Humanitarian label + damage (Mild or Severe)} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Impact Bonus} = \begin{cases} 28, & \text{if Affected People present} \\ 24, & \text{if Severe damage + major disasters} \\ 20, & \text{if Rescue Volunteering Or Donation Effort present} \\ 0, & \text{otherwise} \end{cases}$$

Note: Bonuses are cumulative across categories. For example, a combination showing severe damage from an earthquake with affected people people would receive multiple relevant bonuses.

Class	F1 Score (%)	Train Count	Synthetic	Synthetic (%)	Share (%)
Damage Severity					
Little or None	92.0%	28,227	2,541	9.0%	25.4%
Mild	28.4%	3,008	5,058	168.2%	50.6%
Severe	79.9%	18,118	2,389	13.2%	23.9%
Total		**49,353**	**9,988**	**20.2%**	**100.0%**
Disaster Types					
Earthquake	80.1%	13,176	1,122	8.5%	11.2%
Fire	78.5%	1,820	1,022	56.2%	10.2%
Flood	81.0%	3,547	1,172	33.0%	11.7%
Hurricane	68.2%	4,109	1,580	38.5%	15.8%
Landslide	69.1%	1,155	1,343	116.3%	13.4%
Not Disaster	90.7%	24,038	1,205	5.0%	12.1%
Other Disaster	16.2%	1,508	2,544	168.7%	25.5%
Total		**49,353**	**9,988**	**20.2%**	**100.0%**
Humanitarian					
Affected Injured or Dead People	48.0%	3,268	3,123	95.6%	31.3%
Infrastructure and Utility Damage	85.1%	18,438	1,777	9.6%	17.8%
Not Humanitarian	90.6%	23,605	1,721	7.3%	17.2%
Rescue Volunteering or Donation Effort	52.9%	4,042	3,367	83.3%	33.7%
Total		**49,353**	**9,988**	**20.2%**	**100.0%**
Informative					
Not Informative	89.2%	21,141	818	3.9%	8.2%
Informative	88.1%	28,212	9,170	32.5%	91.8%
Total		**49,353**	**9,988**	**20.2%**	**100.0%**

Table 6.1: Synthetic image allocation by class, shown against F1 results and class frequency.

6.4 Augmented Dataset Fine-Tuning Results

Having established in the previous sections the best-performing synthetic data generation pipeline, prompt designs, and class allocations, we then generated the synthetic dataset and fine-tuned the CNN using the augmented dataset, consisting of the (relabelled) MEDIC dataset plus our synthetic images. We analyse our results in the rest of this section.

6.4.1 Class-level Results

The task-level performance comparison in Table 6.2 presents a sobering assessment of our synthetic data augmentation strategy. Despite targeted allocation of resources to underperforming

classes, quantitative improvements remain modest at best, with some metrics actually showing slight regression.

Damage Severity shows a marginal F1 improvement (84.6% vs 84.3%), with the *mild* damage class improving from 35.8% to 40.0% despite receiving over half our synthetic data budget (Table 6.1). While this 12% relative gain represents some progress, it falls considerably short of expectations given our substantial investment in this category.

For Disaster Types, the overall F1 score showed only marginal improvement (83.6% vs 83.2%), with *other disaster* increasing from 24.2% to 34.4%—our most substantial gain. This heterogeneous category, allocated 25.5% of our synthetic disaster images, at least demonstrates that synthetic data can help with poorly represented classes, even if the absolute performance remains disappointing.

The Humanitarian task reveals the limitations of our approach most clearly. Despite directing 31.3% of synthetic images toward 'Affected/Injured People', we observed no significant difference (53.4% vs 53.6%). This suggests fundamental constraints in using synthetic imagery for ethically sensitive content, where image generation systems' content filters actively resist creating the very scenarios we sought to augment.

These underwhelming results likely stem from the inherent heterogeneity of the MEDIC dataset itself—with its imprecise class boundaries and subjective annotation criteria—combined with the limitations of current text-to-image systems. Nevertheless, the synthetic images introduced valuable visual diversity to humanitarian-significant classes, potentially preparing the model for disaster presentations beyond the training distribution, even if immediate metrics show limited gains.

Task	**Original**		**Relabelled**		**Augmented**	
	Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Damage Severity	83.1%	80.6%	85.5%	84.3%	85.3%	84.6%
Informative	88.6%	88.6%	90.2%	90.2%	89.8%	89.8%
Humanitarian	85.0%	84.6%	86.7%	86.3%	86.1%	85.9%
Disaster Types	81.9%	80.2%	84.5%	83.2%	84.4%	83.6%

Table 6.2: EfficientNet-B1 task-level performance comparison across original, relabelled, and synthetically augmented datasets.

6.4.2 Confusion Matrices

The confusion matrices in Figure 6.6 help understand why our synthetic augmentation strategy may have struggled to deliver substantial improvements.

For Disaster Types, while *other disaster* correct classifications increased from 14% to 24%, this remains low for a critical category. The persistent confusion with 'Not Disaster' (36%) points to a fundamental issue: the boundaries between different disaster types in real-world imagery are often ill-defined, and our synthetic images couldn't overcome this inherent ambiguity.

The Humanitarian matrices show only minor shifts in classification patterns. The 'Affected/Injured People' category improved marginally (50% to 53%), but confusion with 'Infrastructure Damage' (22-25%) remained consistent. This highlights a central challenge: disaster scenes

Table 6.3: Performance comparison between EfficientNet-B1 trained on the augmented dataset versus the relabelled dataset.

	EfficientNet-B1 (Augmented)		**EfficientNet-B1 (Relabelled)**	
Task/Class	**Accuracy (%)**	**F1 Score (%)**	**Accuracy (%)**	**F1 Score (%)**
Damage Severity	**85.3%**	**84.6%**	**85.5%**	**84.3%**
Little Or None	90.8%	92.8%	91.0%	93.0%
Mild	90.3%	40.0%	90.8%	35.8%
Severe	89.6%	81.2%	89.4%	81.4%
Informative	**89.8%**	**89.8%**	**90.2%**	**90.2%**
Not Informative	89.8%	90.2%	90.2%	90.6%
Informative	89.8%	89.4%	90.2%	89.8%
Humanitarian	**86.1%**	**85.9%**	**86.7%**	**86.3%**
Affected/Injured People	96.9%	53.4%	97.0%	53.6%
Infrastructure Damage	90.6%	86.2%	90.8%	86.7%
Not Humanitarian	90.1%	91.2%	90.6%	91.6%
Rescue/Volunteering	94.7%	57.3%	95.0%	57.5%
Disaster Types	**84.4%**	**83.6%**	**84.5%**	**83.2%**
Earthquake	95.6%	82.2%	95.3%	81.3%
Fire	98.2%	82.5%	97.9%	79.4%
Flood	96.9%	82.2%	97.0%	82.8%
Hurricane	94.2%	71.5%	94.4%	72.5%
Landslide	98.4%	70.4%	98.6%	72.5%
Not Disaster	90.2%	91.5%	90.5%	91.8%
Other Disaster	95.2%	34.4%	95.2%	24.2%

Performance comparison between EfficientNet-B1 trained on the augmented dataset versus the relabelled dataset. The table shows metrics for each classification task and class. For tasks (in bold), accuracy represents multi-class classification performance across all classes, while F1 score is the weighted average across classes. For individual classes, accuracy shows binary classification performance (how well the model distinguishes that class from all others), and F1 score measures the harmonic mean of precision and recall for that specific class. The best performing score in each row is highlighted in green.

typically contain multiple elements simultaneously, and synthetic data alone cannot resolve the annotation inconsistencies in the original dataset.

The Damage Severity matrices reveal our most notable improvement in the problematic *mild* damage category (26% to 32% correct classifications). However, this still means two-thirds of *mild* damage cases are incorrectly classified, suggesting we may have overestimated what synthetic data could contribute to inherently subjective category boundaries.

6.4.3 Feature Space Distribution of Synthetic Data

As an additional attempt to understand what is effectively a null result, we tried to visualise the learned category boundaries using dimensionality reduction techniques on the activations of the CNN layers. Figure 6.7 displays t-SNE visualisations (Van der Maaten and Hinton, 2008) of penultimate layer embeddings across our classification tasks, with filled contours showing

original data distributions and dashed lines representing synthetic image clusters.

The plots reveal a consistent pattern: synthetic data occupies peripheral regions rather than targeting overlap areas where classification errors typically occur. In *Damage Severity*, original *mild damage* samples overlap substantially with both *severe* and *little or none*—matching the confusion patterns in our matrices— yet our synthetic examples form clusters extending outward with limited presence in these ambiguity regions.

For *Disaster Types*, synthetic *other disaster* examples create disconnected clusters separate from the original distribution. This positioning may explain why we achieved our largest yet still modest improvement (24.2% to 34.4% F1). The original data shows considerable natural overlap between *earthquake*, *hurricane* and *other disaster*, but our synthetic examples rarely populate these fuzzy boundary regions.

The *Humanitarian* task demonstrates similar limitations, with synthetic *affected, injured or dead people* forming isolated clusters away from category overlap zones. This separation reflects a fundamental challenge: the task contains an implicit classification hierarchy where images often display multiple humanitarian elements simultaneously. Meanwhile, *Informative* shows the greatest distribution similarity between original and synthetic data, consistent with its minimal performance change.

In an ideal implementation, synthetic data would directly target confusion boundaries by populating feature spaces where the model struggles to differentiate between classes. Our approach successfully expanded coverage to novel disaster presentations—potentially enhancing robustness against future distribution shifts—but the embeddings clarify why immediate classification gains remained limited.

This finding points to a key direction for future work. While synthetic data effectively extends conceptual coverage, resolving the fundamental ambiguities in disaster imagery—with its co-existing elements and subjective boundaries—requires more precise control over feature representations than we were able to provide.

Figure 6.6: Confusion matrices comparing EfficientNet-B1 trained on augmented dataset (left) and relabelled dataset (right).

Disaster Types (Augmented)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.85	.01	.00	.02	.01	.07	.01
fire	.01	.87	.00	.00	.00	.06	.03
flood	.01	.00	.78	.04	.01	.12	.00
hurr.	.05	.01	.03	.72	.02	.14	.01
land.	.07	.01	.01	.06	.76	.06	.01
none	.01	.00	.00	.02	.00	.92	.00
other	.18	.08	.02	.07	.01	.36	.24

Disaster Types (Relabelled)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.86	.01	.00	.03	.00	.07	.00
fire	.03	.85	.00	.01	.00	.06	.01
flood	.02	.00	.80	.04	.01	.11	.00
hurr.	.04	.02	.03	.73	.01	.13	.00
land.	.06	.01	.03	.06	.74	.07	.00
none	.01	.00	.01	.02	.00	.93	.00
other	.24	.09	.02	.08	.01	.37	.14

Informativeness (Augmented)

True \ Predicted	not inf	inf
not inf	.88	.11
inf	.08	.91

Informativeness (Relabelled)

True \ Predicted	not inf	inf
not inf	.89	.10
inf	.08	.91

Humanitarian (Augmented)

True \ Predicted	injured	infra	not hum	rescue
injured	.53	.22	.18	.06
infra	.01	.87	.09	.01
not hum	.01	.05	.91	.01
rescue	.07	.21	.19	.51

Humanitarian (Relabelled)

True \ Predicted	injured	infra	not hum	rescue
injured	.50	.25	.16	.07
infra	.00	.89	.08	.01
not hum	.00	.05	.91	.01
rescue	.06	.24	.20	.49

Damage Severity (Augmented)

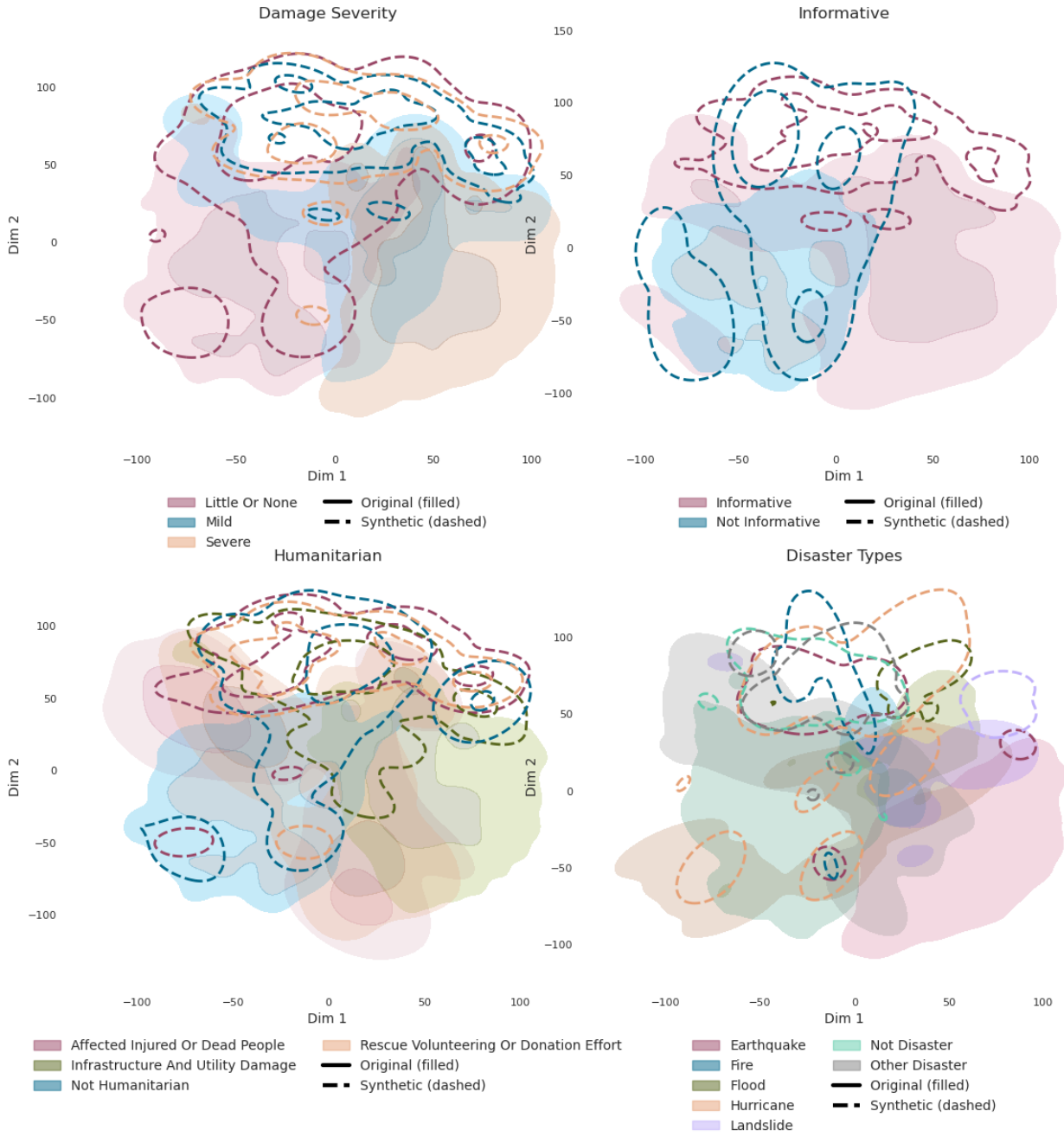
True \ Predicted	none	mild	severe
none	.93	.02	.03
mild	.26	.32	.41
severe	.09	.05	.84

Damage Severity (Relabelled)

True \ Predicted	none	mild	severe
none	.93	.01	.04
mild	.27	.26	.45
severe	.08	.03	.87

Confusion matrices comparing EfficientNet-B1 trained on augmented dataset (left) and relabelled dataset (right). The augmented model shows modest improvements in several areas, particularly in "other disaster" classification (24% vs 14%) and "mild" damage identification (32% vs 26%). Both models continue to struggle with similar misclassification patterns, notably the tendency to classify "mild" damage as either "none" or "severe". The synthetic data appears to provide incremental benefits for difficult edge cases while maintaining performance on well-represented classes. Dark blue cells indicate correct classifications, light blue cells show mediocre performance, and red cells highlight problematic misclassifications.

Figure 6.7: t-SNE Visualisation of Penultimate Layer Embeddings.



Each plot shows the distribution of features in the penultimate layer of our EfficientNet-B1 model projected into two dimensions via t-SNE. Filled contours represent original data distributions while dashed lines show synthetic image clusters. The plots demonstrate how synthetic examples (dashed) often occupy peripheral regions rather than the ambiguous boundary areas between classes where most classification errors occur. This helps explain why performance improvements were modest despite the substantial allocation of synthetic data to challenging categories like *mild damage* and *other disaster*.

Chapter 7

Zero-Shot Classification with Large Multimodal Models

Our synthetic data generation system, whilst considerably more sophisticated than earlier efforts, did not significantly improve our F1 results when used to augment our training dataset. What if, instead, we bypassed the traditional pipeline entirely?

The emergence of large multimodal models offers an attractive alternative. These systems—trained on vast corpora of image-text pairs—offer nuanced visual understanding capabilities without task-specific fine-tuning. Their underlying architecture enables interpretation of disaster imagery through conceptual understanding rather than low-level pattern matching.

This chapter examines whether such models can perform disaster image classification as *zero-shot* tasks, by asking an off-the-shelf model, via application programming interfaces (APIs), to classify a given image with a crafted prompt. We devise our analysis in two phases. First, we examine several prompt designs, analysing how different reasoning frameworks affect classification accuracy. For this exploratory prompt design phase, we test a large number of models and prompt combinations on a smaller validation set. Then, we distil the insights and results from our preliminary exploration into a single prompt for the final zero-shot experiment on the full test set. In the following, we detail our experimental setup, the models and prompts evaluated for both phases, and the results achieved across multiple classification tasks.

7.1 Exploratory Prompt Design Phase

In this section, we cover the experimental details of the exploratory prompt design phase, including the models, prompt types, and dataset used for the analysis.

Models. For this exploratory phase, we selected five leading vision-language models covering a spectrum of capabilities, sizes, costs, and possibility of local deployment, from state-of-the-art closed-source models to smaller but open-weights models; see Table 7.1. For our experiments, we use a common pipeline, accessing all models via APIs through their model providers.

Prompts. To assess how different prompting methodologies affect classification performance, we designed five distinct prompting strategies:

Table 7.1: Vision-Language Models tested in the Exploratory Prompt Design Phase.

Model Name	Provider	Description
Claude 3.5 Sonnet	Anthropic	Flagship model with frontier performance.
Claude 3.5 Haiku	Anthropic	Faster, cheaper vision-capable model.
GPT-4o	OpenAI	Widely regarded as state-of-the-art in multimodal capabilities, as of early 2025.
Pixtral Large	Mistral AI	Large vision-language model optimised for detailed visual understanding.
Pixtral Small	Mistral AI	Smaller open-weights model deployable locally.

- **Direct Classification**: A straightforward approach that instructs the model to directly categorise what it sees into predefined classes after a brief analysis phase.
- **Two-Phase Analysis**: A structured approach that separates observation from assessment, requiring the model to first describe what it sees, then evaluate possible classifications with confidence levels before making final decisions.
- **Elimination Reasoning**: A systematic method that requires the model to explicitly consider evidence for each possible classification option and document its reasoning process for eliminating alternatives.
- **Uncertainty Aware**: A confidence-based approach that instructs the model to assign confidence levels (0-100%) to its classifications and explicitly identify uncertainty factors for low-confidence predictions.
- **Weighted Option Analysis**: A probabilistic approach requiring the model to assign percentage probabilities to all possible classifications within each category, ensuring they sum to 100%, before selecting the highest-probability option.

These prompting strategies represent a progression from direct instruction to increasingly elaborate reasoning frameworks, allowing us to test whether more structured prompting leads to improved classification accuracy. The full prompts are available in Appendix E.

Prompt Validation Set Design. For this exploratory phase, due to the number of experimental conditions to test and limited resources (API costs), we are unable to run our analysis on the full MEDIC validation dataset. Instead, we created a balanced representative subset of the MEDIC validation dataset while ensuring inclusion of rare label combinations, to test model robustness on edge cases. The resulting validation subset contains 500 images with a class distribution mirroring the full validation set (see Appendix E.1.1).

7.2 Exploratory Prompt Design Phase Results

Classification results for all prompts and models considered in this exploratory phase are presented in Table 7.2, presenting both overall accuracy and accuracy per task (Damage Severity, Informative, Humanitarian, Disaster Type). Our findings show a clear stratification of model

capabilities, with larger models outperforming their smaller counterparts, though with significant variations in how different prompting strategies affect performance.

We report in the Appendix a detailed analysis of these results by model type (Appendix E.1.2) and prompt (Appendix E.1.3), including a study the inter-class confusion patterns (Appendix E.1.4). Our preliminary results on this smaller set (500 images) were validated by additional statistical analyses (Appendix E.1.5). Additional results, such as prompt processing times, are reported in Appendix E.1.6 and following sections.

We summarise here the key findings:

- **Model Superiority:** GPT-4o consistently demonstrated the highest overall accuracy, significantly outperforming the other tested models (Claude Sonnet, Claude Haiku, Pixtral Large, Pixtral Small) across most tasks on the validation subset ($p < 0.05$). Claude Sonnet ranked as the second most capable model.
- **Prompt Effectiveness:** Counter to the initial hypothesis that complex reasoning would improve results, the straightforward *Direct Classification* prompt yielded the best or near-best performance for the top models (GPT-4o and Claude Sonnet) in most tasks. However, specific prompts showed strengths in certain areas, such as the *Uncertainty Aware* prompt improving GPT-4o's performance on the challenging *Damage Severity* task.
- **Confusion Patterns:** The analysis highlighted specific, recurring confusion points between classes (e.g., *mild* vs. *severe* damage, *other disaster* vs. *not disaster*, overlapping *humanitarian* categories), indicating areas needing explicit clarification in the prompt design.

Based on these findings, GPT-4o was selected as the model for the final zero-shot evaluation on the full test set. Furthermore, the insights into prompt effectiveness and confusion patterns directly informed the design of the final classification prompt described in the next section, aiming to combine the clarity of direct classification with targeted instructions to mitigate common errors.

7.3 Final Classification Prompt

Based on the findings discussed in Section 7.2, we develop a prompt that aims to bridge:

1. **Direct Classification Structure**: Maintaining the straightforward, task-focused format that yielded the best overall performance metrics.
2. **Uncertainty Handling Mechanisms**: Incorporating explicit uncertainty assessment for challenging classes only when needed, inspired by the specific strengths of the uncertainty-aware prompt.
3. **Original Annotation Guidelines**: Integrating category definitions from the original MEDIC dataset annotation instructions to clarify decision boundaries with inter-class confusion (see Section E.1.4), including quantitative damage thresholds and specific decision hierarchies for *Humanitarian* labels.

   The full, final prompt is provided in Appendix E.3.

Table 7.2: Model Performance Across Different Prompting Strategies (%)

Model	Prompt Type				
	Direct Classification	Two Phase Analysis	Elimination Reasoning	Uncertainty Aware	Weighted Option
Overall Accuracy (%)					
Claude Haiku	80.5	78.2	75.1	78.9	74.5
Claude Sonnet	84.2	83.2	79.0	81.2	78.3
GPT-4o	**87.2**	84.8	83.8	86.0	83.9
Pixtral Large	82.2	81.7	80.3	81.1	80.4
Pixtral Small	77.9	76.2	75.2	77.2	79.4
Damage Severity Accuracy (%)					
Claude Haiku	77.6	78.6	77.4	78.2	73.4
Claude Sonnet	84.2	81.0	78.0	79.0	77.6
GPT-4o	82.0	84.0	82.2	**84.4**	80.8
Pixtral Large	78.2	81.0	79.0	76.0	76.2
Pixtral Small	75.2	76.4	77.8	73.6	77.8
Informative Accuracy (%)					
Claude Haiku	84.2	82.2	68.8	80.4	79.2
Claude Sonnet	87.4	87.4	83.2	86.2	82.0
GPT-4o	**91.8**	88.6	87.6	89.2	88.8
Pixtral Large	87.4	86.8	83.0	87.0	87.6
Pixtral Small	87.6	84.0	75.4	86.2	86.6
Humanitarian Accuracy (%)					
Claude Haiku	79.2	74.0	75.0	76.0	69.4
Claude Sonnet	82.0	81.4	75.0	79.0	76.6
GPT-4o	**86.4**	82.2	79.6	84.4	79.6
Pixtral Large	82.8	78.6	79.2	80.8	79.0
Pixtral Small	75.0	71.0	70.8	72.6	74.0
Disaster Types Accuracy (%)					
Claude Haiku	80.8	78.2	79.2	80.8	75.8
Claude Sonnet	83.0	83.2	80.0	80.6	77.0
GPT-4o	**88.4**	84.4	85.6	86.0	86.2
Pixtral Large	80.4	80.4	80.0	80.6	78.8
Pixtral Small	73.8	73.6	76.6	76.2	79.2

Zero-shot classification accuracy (%) of vision-language models across different prompting strategies on the MEDIC dataset. For each model and task combination, the highest-performing prompt is highlighted in green. The best model-prompt accuracy per task is shown in **bold**. Results demonstrate that Direct Classification generally yields the strongest performance, though optimal prompting strategies vary by model and task. GPT-4o consistently outperforms other models across most prompt types, while Claude Sonnet achieves competitive results with the Direct Classification prompt.

7.4 Zero-Shot Classification Results

In this section, we present the performance of our zero-shot classification approach using GPT-4o with the final optimised prompt.

7.4.1 Zero-Shot Performance

Zero-shot performance was evaluated against the relabelled CNN baselines, using the full test set of 15,687 images. As shown in Table 7.3, GPT-4o demonstrates competitive and often superior performance across all four classification tasks, despite requiring no task-specific training.

GPT-4o achieves higher accuracy in every task category, with measurable advantage in *Disaster Types* (88.1% vs 84.5%) and *Damage Severity* (87.2% vs 85.5%). When it comes to F1 scores, which balance precision and recall, the results show a more nuanced picture. GPT-4o achieves superior F1 scores for the *Informative* task (91.1%) and *Disaster Types* (83.4%). However, for *Humanitarian* classification, despite having similar accuracy (86.8% vs 86.7%), the zero-shot approach shows considerably lower F1 scores (75.0% vs 86.3%), indicating imbalanced precision and recall for certain *Humanitarian* classes. The *Damage Severity* F1 score shows similar issues, albeit less pronounced (79.9% vs. 84.3%).

Task	**CNN (Original)**		**CNN**		**LLM Zero-Shot**	
	Accuracy	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
Damage Severity	83.1%	80.6%	85.5%	84.3%	87.2%	79.9%
Informative	88.6%	88.6%	90.2%	90.2%	90.8%	91.1%
Humanitarian	85.0%	84.6%	86.7%	86.3%	86.8%	75.0%
Disaster Types	81.9%	80.2%	84.5%	83.2%	88.1%	83.4%

Table 7.3: Comparison of classification performance between CNN (EfficientNet-B1) and LLM (gpt-4o) zero-shot classification on the relabelled MEDIC dataset (see Chapter 5). For reference, we also report the CNN performance on the original dataset.

The class-level breakdown in Table 7.4 reveals key strengths and limitations of the zero-shot approach:

1. **Damage Severity**: GPT-4o classifies *mild* damage with far higher accuracy (81.1% vs 26%), proving effective handling of categories with ambiguous boundaries. This middle category—subjectively defined in disaster contexts—benefits from the vision model's conceptual understanding rather than the CNN's pattern-matching approach.

2. **Humanitarian Classification**: Despite similar overall accuracy, GPT-4o underperforms in *affected injured people* (78.8% vs 97.0%) and *rescue volunteering* classes (64.5% vs 95.0%). This discrepancy likely stem from *label overlap*: many images contain *both* affected people and rescue operations simultaneously, but only one label is deemed correct. CNNs may have implicitly learned the annotation hierarchy through training, while GPT-4o lacks knowledge of which class takes precedence when multiple are present. If this is the case, this issue might be addressed in future work by improved task instructions, but it may be nontrivial due to the tradeoff between instruction complexity and performance.

3. **Disaster Types**: GPT-4o performs substantially better with the *other disaster* category (60% vs 14% correct classification). This heterogeneous category benefits from the model's

ability to generalise across diverse disaster scenarios poorly represented in the training data.

Table 7.4: Performance comparison between GPT-4o as a zero-shot classifier and EfficientNet-B1 trained on the relabelled dataset.

	GPT-4o (Zero-Shot)		EfficientNet-B1	
Task/Class	**Accuracy (%)**	**F1 Score (%)**	**Accuracy (%)**	**F1 Score (%)**
Damage Severity	**87.2%**	**79.9%**	**85.5%**	**84.3%**
Little Or None	91.3%	94.3%	91.0%	93.0%
Mild	81.1%	60.8%	90.8%	35.8%
Severe	79.6%	84.7%	89.4%	81.4%
Informative	**90.8%**	**91.1%**	**90.2%**	**90.2%**
Not Informative	91.0%	91.4%	90.2%	90.6%
Informative	90.6%	90.9%	90.2%	89.8%
Humanitarian	**86.8%**	**75.0%**	**86.7%**	**86.3%**
Affected/Injured People	78.8%	57.3%	97.0%	53.6%
Infrastructure Damage	80.7%	86.2%	90.8%	86.7%
Not Humanitarian	93.7%	93.1%	90.6%	91.6%
Rescue/Volunteering	64.5%	63.3%	95.0%	57.5%
Disaster Types	**88.1%**	**83.4%**	**84.5%**	**83.2%**
Earthquake	88.3%	86.9%	95.3%	81.3%
Fire	94.8%	93.1%	97.9%	79.4%
Flood	93.4%	89.4%	97.0%	82.8%
Hurricane	77.9%	80.4%	94.4%	72.5%
Landslide	94.0%	84.1%	98.6%	72.5%
Not Disaster	90.8%	92.9%	90.5%	91.8%
Other Disaster	59.7%	56.7%	95.2%	24.2%

Performance comparison between GPT-4o as a zero-shot classifier and EfficientNet-B1 trained on the relabelled dataset. The table shows metrics for each classification task and class. For tasks (in bold), accuracy represents multi-class classification performance across all classes, while F1 score is the weighted average across classes. For individual classes, accuracy shows binary classification performance (how well the model distinguishes that class from all others), and F1 score measures the harmonic mean of precision and recall for that specific class. The best performing score in each row is highlighted in green.

Examining the confusion matrices, representing the error patterns of these two approaches, provides visual confirmation of these differing classification strategies (Figure 7.1). For instance, the matrices clearly show GPT-4o's superior handling of the ambiguous *mild* damage category compared to the CNN's tendency to misclassify it towards extremes.

In conclusion, the unexpectedly good performance of GPT-4o in certain tasks, particularly those requiring nuanced visual understanding like *mild* damage assessment, demonstrates the value of large multimodal models' broader conceptual knowledge compared to task-specific CNN training. While CNNs excel at pattern recognition within their training distribution, they may struggle with qualitative concepts that benefit from a broader world model.

Figure 7.1: Confusion matrices comparing GPT-4o (left) and EfficientNet-b1 trained on relabelled dataset (right).

Disaster Types (GPT-4o)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.88	.00	.00	.00	.02	.02	.06
fire	.00	.95	.00	.00	.00	.01	.02
flood	.00	.00	.93	.00	.00	.04	.01
hurr.	.02	.00	.05	.77	.01	.08	.03
land.	.02	.00	.00	.00	.93	.01	.00
none	.01	.00	.01	.02	.00	.91	.02
other	.10	.02	.03	.01	.02	.19	.60

Disaster Types (Relabelled)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.86	.01	.00	.03	.00	.07	.00
fire	.03	.85	.00	.01	.00	.06	.01
flood	.02	.00	.80	.04	.01	.11	.00
hurr.	.04	.02	.03	.73	.01	.13	.00
land.	.06	.01	.03	.06	.74	.07	.00
none	.01	.00	.01	.02	.00	.93	.00
other	.24	.09	.02	.08	.01	.37	.14

Informativeness (GPT-4o)

True \ Predicted	not inf	inf
not inf	.90	.09
inf	.07	.92

Informativeness (Relabelled)

True \ Predicted	not inf	inf
not inf	.89	.10
inf	.08	.91

Humanitarian (GPT-4o)

True \ Predicted	injured	infra	not hum	rescue
injured	.80	.03	.13	.01
infra	.06	.80	.07	.06
not hum	.00	.03	.94	.01
rescue	.11	.03	.20	.64

Humanitarian (Relabelled)

True \ Predicted	injured	infra	not hum	rescue
injured	.50	.25	.16	.07
infra	.00	.89	.08	.01
not hum	.00	.05	.91	.01
rescue	.06	.24	.20	.49

Damage Severity (GPT-4o)

True \ Predicted	none	mild	severe
none	.92	.05	.01
mild	.07	.81	.11
severe	.02	.17	.79

Damage Severity (Relabelled)

True \ Predicted	none	mild	severe
none	.93	.01	.04
mild	.27	.26	.45
severe	.08	.03	.87

Confusion matrices comparing GPT-4o (left) and EfficientNet-b1 trained on relabelled dataset (right). GPT-4o demonstrates strong performance across most classes, particularly excelling at "mild" damage classification (81% vs 26%) and "Disaster Types" (especially "other disaster": 60% vs 14%). The relabelled EfficientNet shows stronger performance for "infrastructure" (89% vs 80%) and "severe" damage (87% vs 79%), but struggles with the "mild" damage class. Dark blue cells indicate correct classifications, light blue cells show mediocre performance, and red cells highlight problematic misclassifications.

Chapter 8

Conclusions and Future Work

In this project, we set out to explore two distinct paths to improve image classification in crisis informatics using modern generative AI approaches: synthetic data augmentation via image generation models for fine-tuning CNNs, and zero-shot classification with large multimodal models. Our primary goal was to assess whether these modern AI techniques could improve the accuracy and robustness of systems designed to analyse disaster imagery, ultimately supporting more effective humanitarian response. We measured success empirically by comparing performance quantified by accuracy and F1 scores against MEDIC dataset benchmarks, following a methodology that included baseline replication, dataset relabelling, and systematic evaluation of both approaches.

RQ1: Synthetic Data Augmentation. Our first research question investigated whether synthetic data augmentation could improve the performance of fine-tuned CNNs on the MEDIC benchmark. We developed a pipeline leveraging LLM-generated captions and diffusion models, incorporating refined prompting strategies and allocating synthetic images towards underperforming, critical, or ambiguous classes.

Our synthetic data augmentation approach aimed at improving current classification performance and potentially future-proofing models against domain shift. The in-domain results showed modest improvement. For instance, *mild* damage F1 scores increased from 35.8% to 40.0%, while *other disaster* classification improved from 24.2% to 34.4%. While these gains demonstrate the approach has merit, the overall results were disappointing, with no significant difference in performance across the four main tasks to our baseline (relabelled) results. Overall, it is fair to conclude that **our data augmentation pipeline yielded a null result**. Still, our work may provide valuable findings for understanding the current practical limits of this augmentation strategy.

Two main issues hindered the effectiveness of our synthetic data generation pipeline. First, content safety filters frequently blocked the generation of realistic disaster imagery involving people. Second, our methods for promoting diversity did not produce the breadth of examples needed. To improve this, we need more visibility into the classification regions that are failing: using saliency or feature map approaches to specifically pinpoint the *type* of images that are failing, and using this to improve our prompting, something we began to show in Figure 6.7, but a deeper investigation is left for future work. Moreover, the core idea of making models adaptable to unforeseen future disasters——one of our primary motivations——remains untested due to the constraints of available data. Investigating this further will be essential in exploring

the real potential for this approach.

RQ2: Zero-Shot Classification. Our second research question explored the potential of using large multimodal models directly as zero-shot classifiers, bypassing traditional training pipelines. We evaluated several models and prompt designs, ultimately focusing on GPT-4o with an optimised prompt incorporating MEDIC annotation guidelines.

The findings here were significantly more promising. **Zero-shot classification via large multimodal models delivered unexpectedly strong results**. Without any prior training specific to our tasks, GPT-4o consistently matched or outperformed CNNs across all categories in terms of classification accuracy. The most striking differences were in traditionally difficult categories. For instance, GPT-4o correctly classified *mild* damage 81.1% of the time, whereas CNNs managed only 26.0%. The *other disaster* category saw similar improvement: 60.0% accuracy from GPT-4o compared to just 14.0% with CNNs. As mentioned earlier, these results highlight a meaningful difference: large multimodal models such as GPT-4o can leverage their broader understanding, while CNNs still exhibit advantages on some subclasses possibly due to their ability to learn implicit patterns such as task-dependent label hierarchies. In sum, our results distinctly show a considerable utilisation potential for large multimodal model in crisis informatics, given that their overall performance is on par with—and at times substantially better than—dedicated classification systems.

Limitations. While our work provides potentially useful insights into leveraging generative AI for disaster image classification, several limitations should be acknowledged. Firstly, performance remains capped by the MEDIC dataset's inherent noise, class ambiguities (even post-relabelling), and its social media origin scope. Building a better dataset is a major endeavour, with also potential ethical issues as discussed in Section 2.7. Relatedly, our LLM-assisted relabelling was conservative, non-exhaustive, and potentially biased by the specific prompts and models chosen for judging. A more comprehensive approach would include expert human annotators to double-check the LLM outputs.

We already addressed the modest gains of our data augmentation pipeline, and several of our choices could potentially be improved to yield better results, in particular perhaps a larger number, and better allocation, of synthetic images.

Finally, our study focused only on assessing classification metrics such as accuracy and F1 scores, but several other factors impact model usability, such as speed, cost, and interpretability. For example, our strong zero-shot results relied on proprietary GPT-4o model, limiting accessibility due to cost and deployment constraints. Conversely, open-weights smaller models lagged behind, so they may not be as competitive as their CNN counterparts, at least for now.

Future Work. Several directions for future work stand out. To test our hypothesis about adaptability, future research should evaluate synthetic-augmented models against new datasets specifically designed to introduce domain shifts, perhaps featuring recent disaster events or underrepresented locations. It is also worth exploring ensemble methods combining low-cost, fast CNNs for easier classification tasks with multimodal foundation models as a fallback for more difficult classifications, to achieve the best of both worlds.

While this project explored zero-shot performance, providing large multimodal models with small sets of representative examples (few-shot calibration) could also improve accuracy, es-

pecially for categories prone to ambiguity or where there is a classification hierarchy (scenes showing both impacted people and rescue efforts, for example).

Given the scope of the project, we limited our analysis to large multimodal models accessible via API, but fast advances in the field means that powerful multimodal models with performance comparable to GPT-4o may be soon available to be deployed locally. This would also allow users to access the internals of the models, opening the black-box for more nuanced analyses and understanding of failure cases.

Concluding Remarks. This work makes three contributions to crisis informatics. As an aside, it demonstrates the effectiveness of LLM-based relabelling for improving classification performance. It establishes a synthetic data pipeline that may provide a foundation for addressing data imbalance. Most significantly, it provides compelling evidence that zero-shot classification with large multimodal models can outperform traditionally trained CNNs in the especially challenging domain of disaster image classification—characterised by heterogeneous data, subjective categories, and difficult edge cases.

In particular, the ability of GPT-4o to match or surpass carefully trained CNNs in multiple tasks, especially on challenging categories, represents a new paradigm to explore for crisis response systems. This approach eliminates the data collection and annotation bottleneck that typically constrains model development in specialised fields, enabling more rapid deployment of classification systems during emergencies. However, it is important to highlight that large multimodal models are often proprietary, expensive and slow, while CNNs are fast and can easily be deployed locally, so both approaches still have merit.

In conclusion, combining the complementary strengths of CNNs and foundation models offers a promising path forward for crisis informatics applications, offering modern techniques to support more effective resource allocation during crises, potentially improving response efforts when communities are most vulnerable.

Bibliography

Alam, F., Ofli, F., and Imran, M. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. *Proceedings of the International AAAI Conference on Web and Social Media*, 12 (1), 2018. doi: 10.1609/icwsm.v12i1.14983.

Alam, F., Alam, T., Ofli, F., and Imran, M. Robust Training of Social Media Image Classification Models. *IEEE Transactions on Computational Social Systems*, 11(1):546–565, 2022. doi: 10.1109/TCSS.2022.3230839.

Alam, F., Alam, T., Hasan, M. A., Hasnat, A., Imran, M., and Ofli, F. MEDIC: a multi-task learning dataset for disaster image classification. *Neural Computing and Applications*, 35(3): 2609–2632, 2023. doi: 10.1007/s00521-022-07717-0.

Alimisis, P., Mademlis, I., Radoglou-Grammatikis, P., Sarigiannidis, P., and Papadopoulos, G. T. Advances in diffusion models for image data augmentation: a review of methods, models, evaluation metrics and future research directions. *Artificial Intelligence Review*, 58 (4):112, 2025. doi: 10.1007/s10462-025-11116-x.

Anthropic, . Introducing Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024. Accessed: 2025-03-19.

Braik, A. and Koliou, M. Automated building damage assessment and large-scale mapping by integrating satellite imagery, GIS, and deep learning. *Computer-Aided Civil and Infrastructure Engineering*, 39, 2024. doi: 10.1111/mice.13197.

Bukar, U. A., Sidi, F., Jabar, M. A., Nor, R. N. H., Abdullah, S., Ishak, I., Alabadla, M., and Alkhalifah, A. How Advanced Technological Approaches Are Reshaping Sustainable Social Media Crisis Management and Communication: A Systematic Review. *Sustainability*, 14 (10):5854, 2022. doi: 10.3390/su14105854.

Chen, W., Ma, X., Wang, X., and Cohen, W. W. Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks. *Transactions on Machine Learning Research*, 2023. doi: 10.48550/arXiv.2211.12588.

Defense Innovation Unit. xView2. `https://xview2.org/dataset`.

Duarte, D., Nex, F., Kerle, N., and Vosselman, G. Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2:89–96, 2018. doi: 10.5194/isprs-annals-IV-2-89-2018.

Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J. E., and Darrell, T. Diversify Your Vision Datasets with Automatic Diffusion-Based Augmentation. 2023. doi: 10.48550/arXiv.2305.16289.

Eltehewy, R., Abouelfarag, A., and Saleh, S. N. Efficient Classification of Imbalanced Natural Disasters Data Using Generative Adversarial Networks for Data Augmentation. *ISPRS International Journal of Geo-Information*, 12(6):245, 2023. doi: 10.3390/ijgi12060245.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., and others, . Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-First International Conference on Machine Learning*, 2024.

Figueira, A. and Vaz, B. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics*, 10(15), 2022. doi: 10.3390/math10152733.

Gholami, S., Caleb Robinson, , Anthony Ortiz, , and Siyu Yang, . On the Deployment of Post-Disaster Building Damage Assessment Tools using Satellite Imagery: A Deep Learning Approach. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. doi: 10.1109/ICDMW58026.2022.00134, 2022.

Gilardi, F., Alizadeh, M., and Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Networks. *Communications of the ACM*, 63(11): 139–144, 2014. doi: 10.1145/3422622.

Hamdi, Z. M., Brandmeier, M., and Straub, C. Forest Damage Assessment Using Deep Learning on High Resolution Remote Sensing Data. *Remote Sensing*, 11(17), 2019. doi: 10.3390/rs11171976.

He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., and Qi, X. Is synthetic data from generative models ready for image recognition? In *The Eleventh International Conference on Learning Representations*, 2023.

Imran, M., Castillo, C., Lucas, J., Meier, P., and Vieweg, S. AIDR: artificial intelligence for disaster response. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 159–162. doi: 10.1145/2567948.2577034, 2014.

Islam, M. A., Rabbi, F., and Hossain, N. U. I. Performance evaluation of NLP and CNN models for disaster detection using social media data. *Social Network Analysis and Mining*, 14(1): 213, 2024. doi: 10.1007/s13278-024-01374-y.

Kang, Y., Jung, Y., Shin, W., Kim, B., and Seo, S. MultiFloodSynth: Multi-Annotated Flood Synthetic Dataset Generation, 2025. *arXiv preprint:* 2502.03966.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. Large Language Models are Zero-Shot Reasoners. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213, 2022.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 2012.

Kumar, S. and others, . Detection of Disaster-Affected Cultural Heritage Sites from Social Media Images Using Deep Learning Techniques. In *Proceedings of the Conference on Digital Heritage*, 2020.

Kyeongjin, A., Sungwon, H., Sungwon, P., Jihee, K., and Sangyoon, P. Generalizable Disaster Damage Assessment via Change Detection with Vision Foundation Model. *arXiv preprint arXiv:2406.08020*, 2024. doi: 10.48550.

Mishra, A., Mittal, R., Jestin, C., Tingos, K., and Rajpurkar, P. Improving zero-shot detection of low prevalence chest pathologies using domain pre-trained language models. In *Medical Imaging with Deep Learning*, 2023.

Mouzannar, H., Rizk, Y., and Awad, M. Damage identification in social media posts using multimodal deep learning. In *International Conference on Information Systems for Crisis Response and Management*, 2018.

Mumuni, A. and Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022. doi: https://doi.org/10.1016/j.array.2022.100258.

Nahum, O., Calderon, N., Keller, O., Szpektor, I., and Reichart, R. Are llms better than reported? detecting label errors and mitigating their effect on model performance, 2025. *arXiv preprint:* 2410.18889.

Nguyen, D. T., Ofli, F., Imran, M., and Mitra, P. Damage Assessment from Social Media Imagery Data During Disasters. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 569–576. doi: 10.1145/3110025.3110109, 2017.

Northcutt, C. G., Athalye, A., and Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

OpenAI, . GPT-4oSystem Card. Technical report, 2024.

Park, G. and Lee, Y. Wildfire Smoke Detection Enhanced by Image Augmentation with StyleGAN2-ADA for YOLOv8 and RT-DETR Models. *Fire*, 7(10):369, 2024. doi: 10.3390/fire7100369.

Pratt, S., Covert, I., Liu, R., and Farhadi, A. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15645–15655. doi: 10.1109/ICCV51070.2023.01438, 2023. ISSN: 2380-7504.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Rahnemoonfar, M., Chowdhury, T., and Murphy, R. RescueNet: A High Resolution UAV Semantic Segmentation Dataset for Natural Disaster Damage Assessment. *Scientific Data*, 10(1):913, 2023. doi: 10.1038/s41597-023-02799-4.

Renze, M. and Guven, E. The benefits of a concise chain of thought on problem-solving in large language models. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 476–483. doi: 10.1109/fllm63129.2024.10852493, 2024.

Rui, X., Cao, Y., Yuan, X., Kang, Y., and Song, W. DisasterGAN: Generative Adversarial Networks for Remote Sensing Disaster Image Generation. *Remote Sensing*, 13(21):4284, 2021. doi: 10.3390/rs13214284.

Scheele, S., Picchione, K., and Liu, J. LADI v2: Multi-label Dataset and Classifiers for Low-Altitude Disaster Imagery, 2024. *arXiv preprint:* 2406.02780.

Shao, J., Zhu, K., Zhang, H., and Wu, J. DiffuLT: Diffusion for long-tail recognition without external knowledge. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Tam, Z. R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H., and Chen, Y.-N. Let me speak freely? A study on the impact of format restrictions on large language model performance. In *Conference on Empirical Methods in Natural Language Processing*, 2024.

Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

Vasudevan, V., Caine, B., Gontijo-Lopes, R., Fridovich-Keil, S., and Roelofs, R. When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*, 2022.

Weber, E., Papadopoulos, D. P., Lapedriza, A., Ofli, F., Imran, M., and Torralba, A. Incidents1M: A Large-Scale Dataset of Images With Natural Disasters, Damage, and Incidents. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(04):4768–4781, 2023. doi: 10.1109/TPAMI.2022.3191996.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, 2022.

Yang, L., Yu, Z., Meng, C., Xu, M., Ermon, S., and Cui, B. Mastering text-to-image diffusion: recaptioning, planning, and generating with multimodal llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Yao, W., Zhang, C., Saravanan, S., Huang, R., and Mostafavi, A. Weakly-supervised Fine-grained Event Recognition on Social Media Texts for Disaster Management. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):532–539, 2020. doi: 10.1609/aaai.v34i01.5391.

Yu, Z., Zhu, C., Culatana, S., Krishnamoorthi, R., Xiao, F., and Lee, Y. J. Diversify, Don't Fine-Tune: Scaling Up Visual Recognition Training with Synthetic Images, 2025. *arXiv preprint:* 2312.02253.

Zhang, B., Jiacheng, T., Yuke, H., Song, L., Shah, S. Y., and Wang, L. Multi-scale convolutional neural networks (CNNs) for landslide inventory mapping from remote sensing imagery and landslide susceptibility mapping (LSM). *Geomatics, Natural Hazards and Risk*, 15, 2024. doi: 10.1080/19475705.2024.2383309.

Zhang, R., Wang, B., Zhang, J., Bian, Z., Feng, C., and Ozbay, K. When language and vision meet road safety: leveraging multimodal large language models for video-based traffic accident analysis, 2025. *arXiv preprint:* 2501.10604.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

Zou, Z., Gan, H., Huang, Q., Cai, T., and Cao, K. Disaster Image Classification by Fusing Multimodal Social Media Data. *ISPRS International Journal of Geo-Information*, 10(10): 636, 2021. doi: 10.3390/ijgi10100636.

Appendices

Appendix A: Methods Details

A.1 Model Training and Evaluation

We follow the training and evaluation procedures from the original MEDIC dataset benchmarks, focusing on multi-task classification metrics. We employ identical train, validation, and test splits to the original, with no overlap across sets.

$$\begin{aligned} &\text{Given an image } \mathbf{x}, \text{ let the ground-truth labels be } y^{(1)}, y^{(2)}, y^{(3)}, y^{(4)}. \\ &\text{A multi-task model } f(\mathbf{x};\theta) \text{ produces } \hat{y}^{(1)},\ \hat{y}^{(2)},\ \hat{y}^{(3)},\ \hat{y}^{(4)}, \text{ one for each task.} \end{aligned} \tag{A.1}$$

This formulation represents our multi-task learning approach where a single model simultaneously predicts outputs for all four classification tasks, sharing feature extraction capabilities across the tasks while maintaining task-specific prediction heads.

Thus, the multi-task loss for an image $\mathbf{x}$ is:

$$\mathcal{L}(\theta) = \sum_{t=1}^{4} \alpha_t \ell\left(\hat{y}^{(t)}, y^{(t)}\right), \tag{A.2}$$

where ℓ is the cross-entropy loss for each task t, and α_t are optional weighting factors that can be tuned to prioritise certain tasks if needed. In our training, tasks are *not* weighted ($\alpha_t = 1$).

We use the standard cross-entropy loss for each classification task:

$$\begin{aligned} \ell\left(\hat{y}^{(t)}, y^{(t)}\right) \ &= \ -\sum_{c=1}^{C_t} \mathbb{1}\left[y^{(t)} = c\right] \log\left(p_c^{(t)}\right), \\ \text{where } p_c^{(t)} &= \text{softmax}\left(\hat{y}^{(t)}\right)_c, \end{aligned} \tag{A.3}$$

where $\mathbb{1}[y^{(t)} = c]$ is the indicator function that equals 1 when the ground-truth label $y^{(t)}$ equals class c, and 0 otherwise.

A.2 Performance Metrics

For evaluation metrics, we compute class-specific precision and recall:

$$\begin{aligned} \text{precision}_c \ &= \ \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c}, \\ \text{recall}_c \ &= \ \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c}. \end{aligned} \tag{A.4}$$

Where TP_c, FP_c, FN_c refer to true positives, false positives, and false negatives for class c.

The F1 score for each class combines precision and recall into a single metric:

$$\mathrm{F1}_c \;=\; \frac{2 \times \mathrm{precision}_c \times \mathrm{recall}_c}{\mathrm{precision}_c \;+\; \mathrm{recall}_c}. \tag{A.5}$$

To assess overall model performance across all classes, we use the macro-averaged F1 score:

$$\mathrm{F1}_{\mathrm{macro}} \;=\; \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c, \tag{A.6}$$

where C is the number of classes within a single task, or the total number of relevant categories when computing across tasks. This metric treats all classes equally regardless of their frequency in the dataset.

A.3 Hardware and Software Configuration

The following details our complete experimental setup used throughout this research:

- Hardware:
  - GPU: Single NVIDIA RTX 3070 Laptop GPU
  - CPU: Intel(R) Core(TM) i7-10870H CPU @ 2.20GHz
  - RAM: 32GB
- Software stack:
  - Python: Version 3.12.
  - PyTorch: Version 2.3.1 employed as our principal deep learning framework.
  - CUDA (Nvidia): Version 12.1 for GPU-accelerated training
  - DALI (Nvidia): Version CUDA 120. GPU data processing pipeline for image I/O operations
  - Diffusion models:
    - Stable Diffusion 1.6 / 3.5 (Stability AI): `stable-diffusion-v1-6, sd3.5-medium` endpoints.
    - Black Forest FLUX 1.0-dev: `/v1/flux-dev`
  - LLM APIs:
    - Anthropic Claude:
      - 3.5 Sonnet (`claude-3-5-sonnet-20241022`)
      - 3.7 Sonnet (`claude-3-7-sonnet-20250219`)
      - 3.5 Haiku (`claude-3-5-haiku-20241022`)
    - OpenAI GPT: 4o vision enabled (`gpt-4o-2024-08-06`)

A.4 Common Tools and Techniques

CNN Architectures Overview

Building on prior MEDIC benchmarks (Alam et al., 2023), we focus on three mainstream CNN families:

1. **ResNet** (ResNet-50): A classic backbone offering strong general performance. In code, we rely on TorchVision's pre-trained weights to initialise the training, full fine-tune, then swap fully connected layer for multi-task heads.
2. **EfficientNet** (B1 variant): Known for an excellent accuracy–computational-efficiency trade-off.
3. **MobileNet V2**: Specially designed for computationally constrained environments, included here to address real-world concerns of deployment on edge devices.

Image Generation Model Specifications

To create additional disaster imagery, we adopt Stable Diffusion (SD) version 1.6 and 3.5, and Flux 1-dev. Key hyperparameters include:

- **Resolution**: 320×320 or 512×512. We aim for cost efficiency, knowing that all images are rescaled to 256×256 prior to training.
- **Diffusion steps**: 30.
- **Classifier-free guidance scale**: Ranging from 6.5 to 8.0, controlling how closely the generated images adhere to the textual prompt.
- **Sampler**: Euler ancestral.
- **Safety Tolerance**: Flux 1-dev allows the user to implement a safety filter in terms of content generation. We begin at a level of 4 (6 being the most permissive), then allow up to 5 in cases where the model initially rejects the request.

Appendix B: Baseline CNN Results

B.1 Performance Metrics by Model

Table B.1: Performance comparison across RN50, EN-B1, and MN-V2, grouped by metric.

	Accuracy (%)			F1 (%)			Precision (%)			Recall (%)		
Task/Class	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**
Dmg. Sev.	82.7	**83.1**	81.9	79.6	**80.6**	78.9	79.6	**80.5**	79.2	82.7	**83.1**	81.9
Little/None	–	–	–	91.3	**91.9**	90.7	88.8	**90.2**	89.0	**94.0**	93.6	92.5
Mild	–	–	–	9.3	**15.4**	9.8	41.2	42.2	**42.5**	5.2	**9.4**	5.6
Severe	–	–	–	76.3	**76.4**	75.0	**70.3**	69.9	67.9	83.4	**84.3**	83.8
Inform.	88.2	**88.6**	87.1	88.2	**88.6**	87.2	88.2	**88.8**	87.3	88.2	**88.6**	87.1
Not Info.	–	–	–	89.1	**89.2**	87.8	89.3	**91.4**	90.1	**88.8**	87.1	85.6
Info.	–	–	–	87.2	**87.9**	86.4	**86.9**	85.6	84.0	87.4	**90.4**	88.9
Humanit.	84.4	**85.0**	83.6	83.6	**84.6**	82.4	83.4	**84.4**	82.5	84.4	**85.0**	83.6
Affected People	–	–	–	42.4	**47.3**	43.6	**60.1**	54.8	55.0	32.7	**41.6**	36.2
Infra. Dam.	–	–	–	83.1	**84.3**	82.4	**81.7**	81.6	78.6	84.6	**87.1**	86.7
Not Hum.	–	–	–	89.8	**90.4**	89.3	88.2	**90.7**	88.5	**91.5**	90.1	90.1
Rescue	–	–	–	42.6	**44.0**	26.3	53.4	50.1	**57.7**	35.4	**39.3**	17.1
Dis. Types	80.8	**81.9**	79.4	78.6	**80.2**	76.6	80.3	**81.9**	78.5	80.8	**81.9**	79.4
Quake	–	–	–	75.5	**76.6**	74.1	71.7	**71.9**	68.8	79.7	**81.9**	80.3
Fire	–	–	–	**79.7**	79.4	74.8	**77.3**	76.0	71.7	82.3	**83.0**	78.1
Flood	–	–	–	78.2	**81.1**	78.4	78.3	**80.7**	78.1	78.1	**81.5**	78.8
Hurr.	–	–	–	62.5	**65.3**	60.7	**65.3**	64.9	62.2	59.9	**65.7**	59.2
Land.	–	–	–	67.6	**69.0**	66.1	**67.5**	63.5	63.8	67.7	**75.5**	68.6
Not Dis.	–	–	–	90.0	**90.9**	89.3	86.0	**88.4**	85.7	**94.5**	93.5	93.2
Other	–	–	–	18.9	**26.2**	5.3	78.0	**79.4**	68.1	10.7	**15.7**	2.8

RN50: ResNet50, **EN-B1:** EfficientNet-B1, **MN-V2:** MobileNet-V2

For each **task** (in bold), Accuracy (%) is the multi-class classification accuracy, while F1, Precision, and Recall represent the macro- or weighted-averaged performance across all classes in that task. Best-performing model scores in each row–metric combination are highlighted in **dark green**.

Overall, EfficientNet-B1 demonstrates slightly stronger performance than ResNet50 and MobileNet-V2 across most tasks, particularly in terms of F1 scores and recall. For example, under the *Damage Severity* task, it achieves the highest accuracy (83.1%), along with superior recall (83.1%) and the highest macro-average F1 (80.6%). It also has a consistent advantage for the *Humanitarian* task and its sub-classes, especially *affected people* (47.3% F1) and *not humanitarian* (90.4% F1). While MobileNet-V2 has certain strengths—for instance, higher precision for the "Mild" class (42.5%)—it generally lags behind the other two architectures across most rows.

B.2 Calibration Curves

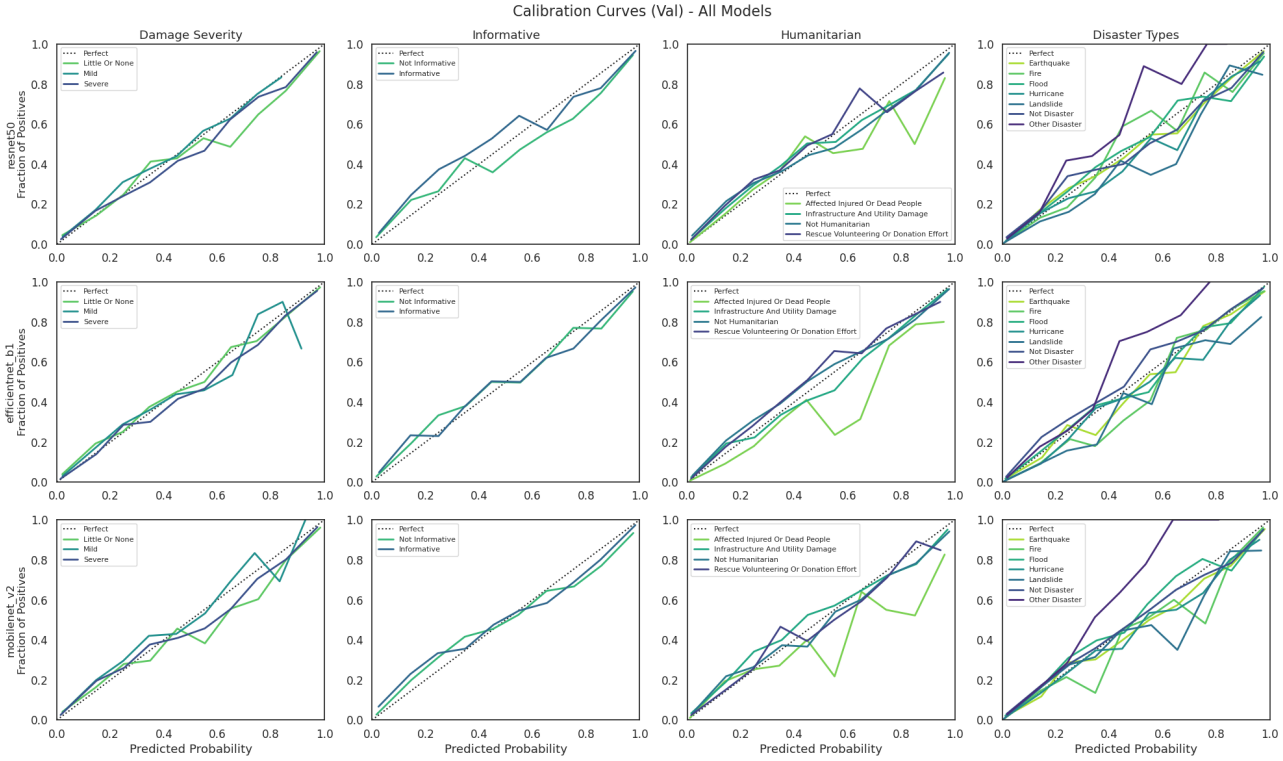


Figure B.1: Calibration curves for all baseline models across the four classification tasks: *Damage Severity*, *Informative*, *Humanitarian*, and *Disaster Types*.

The calibration curves show generally good alignment between predicted probabilities and actual outcomes across all tasks. The *Damage Severity* task exhibits consistent calibration for both *mild* and *severe* classes, with slight overconfidence in the middle probability ranges. For the *Informative* task, our models demonstrate near-perfect calibration with minimal deviation from the ideal curve. The *Humanitarian* task shows more variability, particularly for the *infrastructure and utility damage* class, which displays some underconfidence in the 0.4-0.6 probability range. Finally, in the *Disaster Types* task, the *other disaster* class shows the most notable deviation, with significant underconfidence in predictions around the 0.5 probability threshold. These results suggest that while our baseline models are generally well-calibrated, certain minority classes may benefit from targeted calibration improvements.

Appendix C: Relabelling

C.1 Relabelling Prompts

The prompt used for relabelling images is shown in Figure C.1.

```
You are classifying an image according to four tasks.
Each task must have exactly one label from the provided list.
If multiple disasters appear, please choose the single most prominent one.
For damage severity, if a structure is partially damaged but still usable,
label it \emph{mild}; if it is largely destroyed or unsafe to use, label it 'severe
    '.

TASK 1: "damage_severity"
 - little_or_none: no visible disaster damage or extremely minimal.
 - mild: partial structural damage (e.g. up to ~50% damage, partly collapsed roof).
 - severe: significant destruction (structure no longer usable or mostly collapsed).

TASK 2: "informative"
 - not_informative: not helpful for disaster relief (random images, ads, blurred,
    etc.).
 - informative: clearly shows disaster impact, damage, or relief efforts.

TASK 3: "humanitarian"
 - affected_injured_or_dead_people: the image shows people who are physically harmed
    , displaced, or casualties.
 - infrastructure_and_utility_damage: visible damage to buildings, roads, power
    lines, etc.
 - not_humanitarian: does not show anything relevant for disaster relief.
 - rescue_volunteering_or_donation_effort: depicts active rescue, donation, or
    volunteering.

TASK 4: "disaster_types"
 - earthquake: collapsed buildings/roads typical of seismic activity.
 - fire: noticeable flames, smoke, or burnt structures.
 - flood: submerged roads/buildings, high water level.
 - hurricane: storm surge, strong wind damage, fallen power lines/trees.
 - landslide: fallen earth/rock, mudslides, collapsed ground.
 - not_disaster: the image does not show any disaster.
 - other_disaster: any other catastrophe (explosion, vehicle crash, war, etc.).

Return only a JSON object with exactly these four keys:
{
  "damage_severity": "...",
  "informative": "...",
  "humanitarian": "...",
  "disaster_types": "..."
}
using only the allowed label strings above. No additional text or explanation.
```

Figure C.1: Image classification task instructions with label definitions

C.2 Relabelling Results

Table C.1: Performance comparison across RN50, EN-B1, and MN-V2 on the relabelled dataset, grouped by metric.

	Accuracy (%)			F1 (%)			Precision (%)			Recall (%)		
Task/Class	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**
Dmg. Sev.	84.6	**85.5**	84.5	82.8	**84.3**	82.7	82.9	**84.3**	82.8	84.6	**85.5**	84.5
Little/None	–	–	–	92.3	**93.0**	92.0	90.1	**92.2**	89.9	**94.6**	93.8	94.3
Mild	–	–	–	28.4	**35.8**	29.4	55.0	**55.6**	55.5	19.2	**26.4**	20.0
Severe	–	–	–	80.1	**81.4**	79.9	75.8	**76.0**	75.7	84.8	**87.5**	84.6
Inform.	89.8	**90.2**	88.7	89.8	**90.2**	88.7	89.8	**90.3**	88.7	89.8	**90.2**	88.7
Not Info.	–	–	–	90.4	**90.6**	89.2	90.1	**92.2**	89.8	**90.7**	89.1	88.6
Info.	–	–	–	89.2	**89.8**	88.1	**89.5**	88.2	87.4	88.8	**91.5**	88.7
Humanit.	85.8	**86.7**	85.3	85.2	**86.3**	84.7	85.1	**86.3**	84.5	85.8	**86.7**	85.3
Affected People	–	–	–	51.1	**53.6**	48.0	**64.8**	56.8	59.9	42.2	**50.8**	40.0
Infra. Dam.	–	–	–	85.4	**86.7**	85.1	83.3	**83.8**	83.4	87.7	**89.8**	86.9
Not Hum.	–	–	–	90.9	**91.6**	90.6	89.8	**91.7**	89.4	**92.0**	91.5	91.8
Rescue	–	–	–	54.7	**57.5**	52.9	66.1	**68.7**	62.5	46.7	**49.4**	45.9
Dis. Types	83.8	**84.5**	83.0	82.2	**83.2**	81.4	83.2	**84.0**	82.5	83.8	**84.5**	83.0
Quake	–	–	–	80.1	**81.3**	80.5	76.0	**77.0**	75.9	84.7	**86.2**	85.7
Fire	–	–	–	**81.9**	79.4	78.5	**82.4**	74.0	74.7	81.5	**85.6**	82.7
Flood	–	–	–	81.0	**82.8**	81.2	83.7	85.0	**86.2**	78.5	**80.7**	76.8
Hurr.	–	–	–	70.2	**72.5**	68.2	**71.7**	71.4	69.5	68.7	**73.5**	66.9
Land.	–	–	–	**73.1**	72.5	69.1	**77.4**	70.4	68.7	69.3	**74.8**	69.6
Not Dis.	–	–	–	91.2	**91.8**	90.7	87.9	**90.3**	87.8	**94.7**	93.3	93.7
Other	–	–	–	18.0	**24.2**	16.2	**74.1**	69.2	71.8	10.2	**14.7**	9.1

RN50: ResNet50, **EN-B1:** EfficientNet-B1, **MN-V2:** MobileNet-V2

For each **task** (in bold), Accuracy (%) is the multi-class classification accuracy, while F1, Precision, and Recall represent the macro- or weighted-averaged performance across all classes in that task. Best-performing model scores in each row–metric combination are highlighted in **dark green**.

In these results, all three convolutional neural networks — ResNet50, EfficientNet-B1, and MobileNet-V2 — are trained on the relabelled dataset, improving class definitions. While all models achieve strong performance, EfficientNet-B1 again outperforms its counterparts for most tasks and classes, notably in the *Damage Severity* and *Humanitarian* tasks, often achieving higher recall and F1 scores. Nonetheless, ResNet50 occasionally records the highest recall or precision on certain classes (e.g. *little or none fire*, *not humanitarian*, and *other disaster*,).

Task/Class	Accuracy			F1 Score		
	RN50	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**
Damage Severity	84.6%	85.5%	84.5%	82.8%	84.3%	82.7%
Little Or None	89.9%	91.0%	89.6%	92.3%	93.0%	92.0%
Mild	90.6%	90.8%	90.6%	28.4%	35.8%	29.4%
Severe	88.8%	89.4%	88.7%	80.1%	81.4%	79.9%
Informative	89.8%	90.2%	88.7%	89.8%	90.2%	88.7%
Not Informative	89.8%	90.2%	88.7%	90.4%	90.6%	89.2%
Informative	89.8%	90.2%	88.7%	89.2%	89.8%	88.1%
Humanitarian	85.8%	86.7%	85.3%	85.2%	86.3%	84.7%
Affected/Injured People	97.3%	97.0%	97.1%	51.1%	53.6%	48.0%
Infrastructure Damage	90.0%	90.8%	89.8%	85.4%	86.7%	85.1%
Not Humanitarian	89.6%	90.6%	89.2%	90.9%	91.6%	90.6%
Rescue/Volunteering	94.7%	95.0%	94.4%	54.7%	57.5%	52.9%
Disaster Types	83.8%	84.5%	83.0%	82.2%	83.2%	81.4%
Earthquake	95.0%	95.3%	95.1%	80.1%	81.3%	80.5%
Fire	98.3%	97.9%	97.9%	81.9%	79.4%	78.5%
Flood	96.7%	97.0%	96.8%	81.0%	82.8%	81.2%
Hurricane	94.1%	94.4%	93.7%	70.2%	72.5%	68.2%
Landslide	98.8%	98.6%	98.5%	73.1%	72.5%	69.1%
Not Disaster	89.6%	90.5%	89.0%	91.2%	91.8%	90.7%
Other Disaster	95.2%	95.2%	95.1%	18.0%	24.2%	16.2%

RN50: ResNet50, EN-B1: EfficientNet-B1, MN-V2: MobileNet-V2

Table C.2: Performance comparison of CNN architectures across multi-task classification metrics, trained on the relabelled image dataset. Best results for each metric are highlighted in green.

The performance of CNN architectures (ResNet50, EfficientNet-B1, and MobileNet-V2) has notably improved across all classification tasks after training on the relabelled image dataset. For the Damage Severity task, overall accuracy increased from 83.1% to 85.5% with EfficientNet-B1, with particularly significant improvement in the F1 score for the "Mild" damage class (from 15.4% to 35.8%), suggesting better discrimination of moderate damage conditions. The Informative classification task showed modest improvements in both accuracy and F1 scores, increasing from 88.6% to 90.2% accuracy with EfficientNet-B1. In the Humanitarian task, performance improved across all metrics, with notable enhancement in the F1 scores for the "Affected/Injured People" class (from 47.3% to 53.6%) and "Rescue/Volunteering" class (from 44.0% to 57.5%), indicating better practical utility in these critical disaster response categories. For Disaster Types classification, the relabelled dataset yielded better performance, especially for the challenging "Other Disaster" class where the F1 score with EfficientNet-B1 remained low but still improved from 26.2% to 24.2%. Overall, the relabelled dataset has enabled more robust model performance, with average improvements of approximately 2-4 percentage points in accuracy and 3-5 percentage points in F1 scores across the tasks.

Appendix D: Synthetic Augmentation

D.1 Preliminary Experiments

Our preliminary experiments aimed at finding the best-performing setup for generating synthetic disaster images belonging to desired class labels. In particular, we evaluated three prompt designs (*Naïve*, *Structured*, and *Multistage*), three LLM captioners (Claude-3.7-sonnet, GPT-4o, Claude-3.5-haiku), and three diffusion models (Stable Diffusion-1.6, Stable Diffusion-3.5, Flux 1-dev). Tests included three real images and one hypothetical scenario per class combination, ensuring comprehensive coverage (360 examples). We used two multimodal LLMs (GPT-4o, Claude-3.7-Sonnet) to verify generated-image alignment against intended labels. This preliminary experiment was crucial for identifying effective prompting strategies and model combinations early in our experimental pipeline.

Table D.1 summarises the diffusion models' average alignment accuracy. Flux 1-dev outperformed the newer SD-3.5, which frequently generated overly dramatic or incomplete scenes and exhibited content policy rejections.

Diffusion Model	Overall Accuracy (%)
Flux 1-dev	**58.1 ± 0.9**
SD-1.6	51.0 ± 0.9
SD-3.5	54.6 ± 0.9

Table D.1: Overall label-alignment accuracy (%) across diffusion models, aggregated across prompts and LLMs (mean ± SEM across 3,240 test examples).

Focusing now on the best diffusion model, prompt design comparison (Table D.2) revealed that *Multistage* prompts, encouraging detailed stepwise visual descriptions, achieved the highest alignment accuracy. *Structured* prompts were moderately successful, while simpler *Naïve* prompts suffered from ambiguity, aligning with findings from Renze and Guven (2024). See Appendix D.2.1 for detailed prompt examples.

Finally, comparing LLM caption generators for the best diffusion model and prompt type (Table D.3), Claude-3.7-sonnet demonstrated superior prompt coherence and higher alignment accuracy than Claude-3.5-haiku, and marginally better than GPT-4o. However, no model fully eliminated prompt-image content discrepancies.

Prompt Type	Overall Accuracy (%)
Naïve	56.7 ± 1.5
Structured	57.3 ± 1.5
Multistage	**60.2 ± 1.5**

Table D.2: Overall accuracy (%) of Flux 1-dev generated images by prompt design, aggregated across LLMs (mean ± SEM across 1,080 test examples).

LLM Variant	Overall Accuracy (%)
Claude-3.5-haiku	55.8 ± 2.6
GPT-4o	**60.3 ± 2.6**
Claude-3.7-sonnet	**63.6 ± 2.6**

Table D.3: Overall accuracy (%) across LLM-generated prompts for Flux 1-dev generated images using the multistage prompt (mean ± SEM across 360 test examples).

D.2 Image Captioning Prompts

D.2.1 Initial Prompts Tested

From Image, Structured

```
Describe ONLY the visible elements using EXACTLY this format:

MEDIA_TYPE: [Photograph/Screenshot/Article/Post/Diagram]
TECHNICAL: [camera angle, lighting, quality - max 10 words]
SCALE: [exact measurements of visible area/distance/size]
-- Must use numbers (feet/meters/stories/etc.)
-- Must specify total area visible
-- Must note size of main elements

DISASTER TYPE EVIDENCE:
-- List specific visual indicators of disaster cause/type
-- If uncertain, state "Cause not visually definitive"
-- DO NOT speculate beyond visual evidence

DAMAGE ASSESSMENT:
-- Severe = List non-functional/unsafe features (collapse, complete destruction)
-- Mild = List partial damage (cracks, broken windows, partial collapse)
-- Little/None = List superficial effects (debris, minor marks) or state 'No damage
    visible'

PRIMARY ELEMENTS: [max 30 words]
-- List main subjects/objects
-- Note any people/activities
-- Include key infrastructure

ENVIRONMENT: [max 20 words]
-- Setting type
-- Weather/conditions
-- Notable features
-- Visible signage/identifiers

VERIFICATION:
-- If this is NOT a disaster/emergency scene, respond ONLY with:
  "NON-DISASTER IMAGE: [brief factual description of actual content]"

RULES:
-- NO explanation of how you're following rules
-- NO dramatic language ("devastating", "catastrophic", etc.)
-- NO speculation about areas outside frame
-- ONLY describe what is physically visible
-- Use specific measurements
```

Figure D.1: 'Structured' test prompt for captioning an image

From Image, Multistage

```
Analyse this image through a step-by-step process:

STAGE 1 - INVENTORY:
List ALL visible elements in the image
-- Note people, objects, infrastructure, environmental features
-- Include rough count of each element type
-- Do NOT interpret or classify yet

STAGE 2 - CATEGORIZATION:
-- Determine media type (photo, screenshot, etc.)
-- Identify setting type (urban, rural, indoor, etc.)
-- If NOT a disaster scene, state "NON-DISASTER IMAGE" and briefly describe content
-- If disaster-related, continue to next stages

STAGE 3 - MEASUREMENTS:
-- Estimate key dimensions using numbers (area, distances, etc.)
-- Distances between important elements
-- Depth of any water, height of any flames, etc.

STAGE 4 - CONDITION ASSESSMENT:
-- Damage severity (Severe/Mild/Little-None)
-- For severe: List non-functional/unsafe elements
-- For mild: List partially damaged elements
-- For little/none: List superficial effects or "No damage visible"

STAGE 5 - DISASTER INDICATORS:
-- Identify visual markers indicating disaster type
-- Note conflicting or ambiguous indicators
-- Avoid speculation beyond visible evidence

STAGE 6 - ENVIRONMENTAL CONTEXT:
-- Weather/atmospheric conditions
-- Time of day indicators
-- Surrounding terrain/setting characteristics
-- Maintain objectivity throughout; avoid dramatic terms.
```

Figure D.2: 'Multistage' test prompt for captioning an image

From Image, Naïve

```
Please describe this image in detail. What can you see in the picture?
What type of disaster is it (if any)? How bad is the damage?
Are there people visible and what are they doing? Where was this taken?
```

Figure D.3: 'Naïve' test prompt for captioning an image

From Labels, Structured

```
Using labels [type, severity, humanitarian, informative], describe:

SETTING: [location type, exact measurements of area]

DISASTER TYPE CHARACTERISTICS:
-- Visual elements specific to this disaster type
-- Physical manifestations unique to this event type
-- Scale and scope indicators

DAMAGE LEVEL:
-- Severe = List specific non-functional elements
-- Mild = List specific partial damage
-- Little/None = List minor effects or state 'none'

VISIBLE ELEMENTS:
-- Count of people/vehicles
-- Specific infrastructure
-- Exact measurements
-- Key activities

ENVIRONMENT:
-- Weather/conditions
-- Notable features
-- Visible markers

HUMANITARIAN ASPECTS:
-- Response activities visible
-- Aid-related elements
-- Human impact indicators

Use neutral language. NO dramatic terms. Only describe elements that could be
    physically visible in a photograph.
```

Figure D.4: 'Structured' test prompt for describing a hypothetical scenario

From Image, Multistage

```
Using the provided labels [type, severity, humanitarian, informative], construct a
    realistic disaster scene:

STAGE 1 - LOCATION FRAMEWORK:
-- Establish the physical setting
-- Define geographical context (urban/rural/coastal/etc.)
-- Establish scale with specific measurements (area dimensions)
-- Identify key environmental characteristics
-- Describe baseline infrastructure before impact

STAGE 2 - IMPACT VISUALIZATION:
-- Detail the disaster's physical manifestation
-- Incorporate weather/atmospheric conditions typical of this event
-- Describe physical processes currently active or recently occurred
-- Add sensory details visible in a photograph (not sounds/smells)

STAGE 3 - DAMAGE SPECIFICATION:
-- Match damage severity precisely to the label (severe, mild, etc.)
-- For severe: structural failures, non-functional elements
-- For mild: partial damage with specifics
-- For little/none: minimal effects consistent with event

STAGE 4 - HUMAN ELEMENTS:
-- Populate with realistic human presence
-- Specific number of individuals, roles/activities
-- Appropriate emergency vehicles/equipment if needed
-- Ensure activities match both disaster type and humanitarian label

STAGE 5 - RESPONSE INTEGRATION:
-- For rescue/volunteering: include response activities/equipment
-- For affected/injured: medical response elements
-- For infrastructure damage: utility workers, assessment teams
-- For not_humanitarian: exclude organized response

STAGE 6 - VISUAL COHESION:
-- Verify single camera perspective
-- Check internal consistency of measurements
-- Ensure environment matches across all stages
-- Remove elements not physically visible
```

Figure D.5: 'Multistage' test prompt for describing a hypothetical scenario

From Labels, Naive

```
-- Imagine a scene showing this type of disaster (if any).
-- What would it look like? Describe the location, the damage you might see,
-- and any people who might be there. What would the overall scene look like?
-- Try to make it realistic and detailed so someone could visualize it clearly.
```

Figure D.6: 'Naïve' test prompt for describing a hypothetical scenario

Fallback 1

```
This image is part of a public disaster response dataset used to train humanitarian
    aid ML systems.
-- All images are from public sources, and any individuals remain completely
    anonymous.
-- Please provide a factual, respectful description focusing on the scene and
    humanitarian response aspects.
Use professional, neutral language and avoid graphic details.
```

Figure D.7: First fallback prompt upon an initial refusal to describe an image for senstitive content

Fallback 2

```
I understand your caution with sensitive content.
To clarify: This is part of a machine learning research project for disaster
    response and humanitarian aid.
The image is from public sources (news/humanitarian organizations), and all
    individuals remain anonymous. The description you provide will:
1) Never be public
2) Only be used to generate synthetic training data for disaster response AI
3) Help improve automated systems that assist in disaster recovery and aid
    distribution.
Please provide a careful, professional description using appropriate medical/
    emergency response terminology.
Focus on: -- Response activities -- General scene description -- Professional
    distance in descriptions -- Factual, non-sensational language.
Avoid graphic details while maintaining the essential information needed for
    disaster response training.
```

Figure D.8: Second fallback prompt upon an second refusal to describe an image for sensitive content

D.2.2 Refined Prompts

In our refined approach, we split each prompt into two sections: an `Analysis` block (not passed to the diffusion model) and a `Caption` block (actually passed to the model). The **<analysis>** block is used for LLM "thinking," while only the **<caption>** block is fed to the diffusion model. If both the main prompt and these fallback prompts fail, we skip generating that particular image to avoid forcibly eliciting disallowed or overly graphic content.

Below are the final prompt texts for `real`, `hypothetical`, and `fallback` usage types.

From Labels, Refined

```
# Task description

## Analysis

Describe ONLY the visible elements of the provided image using EXACTLY this format:

<analysis>
MEDIA_TYPE: [Photograph/Screenshot/Article/Post/Diagram]

TECHNICAL: [camera angle, lighting, quality - max 10 words]

SCALE: [exact measurements of visible area/distance/size]
- Must use numbers (feet/meters/stories/etc.)
- Must specify total area visible
- Must note size of main elements

DISASTER TYPE EVIDENCE:
- List specific visual indicators of disaster cause/type
- If uncertain, state "Cause not visually definitive"
- DO NOT speculate beyond visual evidence

DAMAGE ASSESSMENT:
- Severe = List non-functional/unsafe features (collapse, complete destruction)
- Mild = List partial damage (cracks, broken windows, partial collapse)
- Little/None = List superficial effects (debris, minor marks) or state 'No damage
    visible'

PRIMARY ELEMENTS: [max 30 words]
- List main subjects/objects
- Note any people/activities
- Include key infrastructure

ENVIRONMENT: [max 20 words]
- Setting type
- Weather/conditions
- Notable features
- Visible signage/identifiers
```

```
</analysis>

IMPORTANT: If this is NOT a disaster/emergency scene, respond ONLY with: "NON-
    DISASTER IMAGE: [brief factual description of actual content]"

RULES:
- Use OBJECTIVE, MEDICAL language if there are injured people.
- NO explanation of how you're following rules
- NO dramatic language ('devastating', 'catastrophic', etc.)
- NO speculation about areas outside frame
- ONLY describe what is physically visible
- Use specific measurements

Provide your structured analysis in <analysis></analysis> tags.

## Caption

After that, write a full descriptive caption for the provided image following your
    analysis.

- Make sure to weave ALL the elements of your analysis into the image caption.
- If NON-DISASTER IMAGE, provide the factual description you wrote of the image
    content.
- This image caption MUST be self-contained as it will be passed to an AI image
    generator.
- Output the caption inside <caption></caption> brackets.

Your output should ONLY be a <analysis></analysis> block followed by a <caption>
    </caption> paragraph.
```

Figure D.9: Refined prompt for captioning images, following a reasoning then captioning approach

From Labels, Refined

```
# Task description

## Analysis

Using labels [type, severity, humanitarian, informative], imagine a scene. It may be
    a disaster, aftermath of one or not disaster related at all, depending on the
   labels you have received. If the label is not\_informative, do not imagine a
   scene, instead imagine a cartoon, advert, social media post, text based tweet,
   etc... that is not useful for humanitarian aid or disaster responders, although
   MAY be relevant to the disaster described (e.g. earthquake safety tips, maps of
   hurricane paths...).

Get inspiration from the keywords listed at the end.

Now describe the scene in EXACTLY this format:

<analysis>
SETTING: [location type, exact measurements of area]

DISASTER TYPE CHARACTERISTICS (if disaster):
- Visual elements specific to this disaster type
- Physical manifestations unique to this event type
- Scale and scope indicators

DAMAGE LEVEL:
- Severe = List specific non-functional elements (collapse, complete destruction,
   buildings, roads or other amenities totally unusable)
- Mild = List specific partial damage: partially destroyed buildings, bridges,
   houses, roads with damage <40%
- Little/None = List minor effects or state 'none'

VISIBLE ELEMENTS:
- Count of people/vehicles
- Specific infrastructure
- Exact measurements
- Key activities

ENVIRONMENT:
- Weather/conditions
- Notable features
- Visible markers

HUMANITARIAN ASPECTS (if any):
- Use OBJECTIVE, MEDICAL, NEUTRAL language if there are injured people.
- Response activities visible
- Aid-related elements
```

```
- Human impact indicators

Use neutral language. NO dramatic terms. Only describe elements that could be
    physically visible in a photograph.

RULES:
- Use OBJECTIVE, MEDICAL, language if there are injured people.
- NO explanation of how you're following rules
- NO dramatic language ('devastating', 'catastrophic', etc.)
- NO speculation about areas outside frame
- ONLY describe what is physically visible
- Use specific measurements

Provide your structured analysis in <analysis></analysis> tags.

## Caption

After that, write a full descriptive caption for the provided image following your
    analysis.

- Make sure to weave ALL the elements of your analysis into the image caption.
- This image caption should be self-contained as it will be passed to an AI image
    generator.
- Output the caption inside <caption></caption> brackets.

Your output should ONLY be a <analysis></analysis> block followed by a <caption>
    </caption> paragraph.
```

Figure D.10: Refined prompt for describing a hypothetical scenario, following a reasoning then captioning approach

D.3 Augmented Dataset Results

Table D.4: Performance comparison across RN50, EN-B1, and MN-V2 on the augmented dataset, grouped by metric.

	Accuracy (%)			F1 (%)			Precision (%)			Recall (%)		
Task/Class	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**	**RN50**	**EN-B1**	**MN-V2**
Dmg. Sev.	84.6	**85.3**	84.2	83.3	**84.6**	83.0	82.9	**84.3**	82.7	84.6	**85.3**	84.2
Little/None	–	–	–	92.1	**92.8**	92.0	89.7	**92.1**	90.9	**94.6**	93.5	93.1
Mild	–	–	–	34.8	**40.0**	33.3	**51.7**	51.0	50.1	26.3	**32.9**	25.0
Severe	–	–	–	80.0	**81.2**	79.6	**78.1**	77.8	75.1	82.0	**84.9**	84.6
Inform.	89.0	**89.8**	88.6	89.0	**89.8**	88.6	89.1	**89.9**	88.7	89.0	**89.8**	88.6
Not Info.	–	–	–	89.3	**90.2**	89.2	91.4	**91.9**	89.8	87.3	88.4	**88.5**
Info.	–	–	–	88.6	**89.4**	88.0	86.5	**87.6**	87.4	90.8	**91.3**	88.7
Humanit.	85.0	**86.1**	84.7	84.7	**85.9**	84.1	84.6	**85.9**	83.9	85.0	**86.1**	84.7
Affected People	–	–	–	51.9	**53.4**	47.0	**57.8**	53.2	55.6	47.2	**53.6**	40.7
Infra. Dam.	–	–	–	84.7	**86.2**	84.8	**84.8**	84.6	82.6	84.6	**88.0**	87.0
Not Hum.	–	–	–	90.3	**91.2**	90.1	89.0	**91.1**	89.1	**91.6**	91.3	91.1
Rescue	–	–	–	55.8	**57.3**	50.1	59.9	**65.2**	61.2	**52.2**	51.2	42.4
Dis. Types	83.2	**84.4**	83.0	81.9	**83.6**	81.6	82.2	**83.7**	82.0	83.2	**84.4**	83.0
Quake	–	–	–	80.6	**82.2**	80.1	78.0	**79.2**	76.3	83.3	**85.3**	84.3
Fire	–	–	–	80.3	**82.5**	78.8	77.5	**78.2**	75.8	83.3	**87.2**	82.1
Flood	–	–	–	80.2	**82.2**	80.2	**87.5**	86.3	84.2	73.9	**78.4**	76.7
Hurr.	–	–	–	68.1	**71.5**	68.9	66.8	**70.5**	68.8	69.4	**72.4**	69.0
Land.	–	–	–	**71.3**	70.4	70.4	**73.7**	65.2	67.5	69.0	**76.6**	73.5
Not Dis.	–	–	–	90.8	**91.5**	90.7	88.0	**90.2**	88.4	**93.8**	92.8	93.1
Other	–	–	–	24.6	**34.4**	20.4	57.2	57.5	**59.5**	15.6	**24.5**	12.3

RN50: ResNet50, **EN-B1:** EfficientNet-B1, **MN-V2:** MobileNet-V2

For each **task** (in bold), Accuracy (%) is the multi-class classification accuracy, while F1, Precision, and Recall represent the macro- or weighted-averaged performance across all classes in that task. Best-performing model scores in each row–metric combination are highlighted in **dark green**.

When trained on the augmented dataset—including synthetic images—EfficientNet-B1 generally retains a performance advantage over ResNet50 and MobileNet-V2 across most tasks and classes. Notably, its macro-averaged F1, precision, and recall scores are especially high for *Damage Severity and Disaster Types*, demonstrating that the inclusion of synthetic data can further enhance model performance. However, ResNet50 continues to outperform in certain class-specific metrics, such as higher precision or recall for *mild* and *landslide*.

Appendix E: Zero-shot Classification

E.1 Preliminary Prompt Design Experiments

We report here an extensive analysis of our preliminary experiments to establish the best setup for the zero-shot classification experiment.

E.1.1 Preliminary Experiment Validation Set

For the preliminary experiment, we created a smaller validation set of 500 images with the class distribution described in Table E.1.

Table E.1: Class Distribution Comparison Between MEDIC Dev Set and Test Subset

Task/Class	% in Dev Set	Test Subset
Damage Severity		
Little Or None	57.1%	256 (51.2%)
Mild	11.0%	61 (12.2%)
Severe	31.9%	183 (36.6%)
Informative		
Not Informative	44.2%	203 (40.6%)
Informative	55.8%	297 (59.4%)
Humanitarian		
Affected Injured Or Dead People	3.7%	19 (3.8%)
Infrastructure And Utility Damage	39.5%	214 (42.8%)
Not Humanitarian	48.4%	224 (44.8%)
Rescue Volunteering Or Donation Effort	8.5%	43 (8.6%)
Disaster Types		
Earthquake	16.7%	82 (16.4%)
Fire	4.7%	32 (6.4%)
Flood	10.5%	55 (11.0%)
Hurricane	10.0%	57 (11.4%)
Landslide	3.1%	22 (4.4%)
Not Disaster	50.6%	231 (46.2%)
Other Disaster	4.4%	21 (4.2%)

E.1.2 Multimodal Model Performance

Across all tests, GPT-4o and Claude Sonnet show the strongest overall accuracy, followed by Pixtral Large, Claude Haiku and Pixtral Small. The five prompt styles—Direct Classification, Two Phase Analysis, Elimination Reasoning, Uncertainty Aware, and Weighted Option Analysis—also vary in effectiveness for different classification tasks.

GPT-4o significantly outperforms all other models on the *Informative*, *Humanitarian*, and *Disaster Types* tasks ($p < 0.05$). The performance gap between GPT-4o and Claude Sonnet is statistically significant for all tasks except *Damage Severity*, where Claude Sonnet (84.2%) actually outperforms GPT-4o (81.9%), though this difference doesn't reach statistical significance. A statistical comparison specifically focusing on the two best-performing models is reported in Table E.2. Full statistical test results are available in Appendix E.5.2. Claude Sonnet ranks as the second most capable model, with its highest performance (84.2% overall accuracy) also achieved via the Direct Classification prompt.

The consistent pattern whereby the most straightforward prompting approach yields better results contradicts the initial hypothesis that more elaborate reasoning frameworks would enhance performance. This finding suggests that excessive prompt complexity may introduce reasoning bottlenecks, where the model may perhaps second-guess its intuitive classification, much as a human annotator might.

Task	GPT-4o	Claude Sonnet	Difference	p-value
Damage Severity	84.47% (Uncertainty)	84.20% (Direct)	0.27%	0.1757
Informative	91.78% (Direct)	87.40% (Two Phase)	4.38%	0.0046*
Humanitarian	86.41% (Direct)	82.04% (Direct)	4.37%	0.0121*
Disaster Types	88.30% (Direct)	83.15% (Two Phase)	5.15%	0.0011*

* Statistically significant difference (p <0.05)

Table E.2: Statistical Comparison of Best Configurations by Model

Looking at the *Damage Severity labels*, prompts that encourage a short reflection (for example, Two Phase Analysis) often help the model differentiate between "mild" and "severe" damage, probably because they guide the model to notice partial damage. On the other hand, the simpler *Direct Classification* prompt tends to do well with rare categories such as "other disaster," as it presents straightforward options without lengthy reasoning steps.

Finally, to statistically validate our results, we calculated bootstrapped confidence intervals, showing that GPT-4o demonstrates not only high accuracy but also consistent performance (see Appendix E.1.5 for the analysis). In sum, our preliminary analyses suggests to use GPT-4o as the overall best-performing model, with Claude 3.5 Sonnet as a valid alternative.

E.1.3 Prompt Performance

For analysing the prompt performance more in detail, we concentrate on the two best-performing models, GPT-4o and Claude 3.5 Sonnet. Both models show a clear overall advantage with the simple *Direct Classification* prompt, which provides explicit category definitions with visual indicators that models can directly match to image features (see Tables E.3 and E.4).

Task	Prompt	Accuracy (%)	Accuracy 95% CI*	F1 (%)
Damage Severity	Direct Classification	82.1	[78.6, 85.5]	74.1
	Two Phase Analysis	84.0	[80.7, 87.0]	76.8
	Elimination Reasoning	82.2	[78.7, 85.7]	73.3
	Uncertainty Aware	**84.4**	**[80.8, 87.4]**	**78.9**
	Weighted Option Analysis	80.7	[77.2, 84.0]	72.4
Informative	**Direct Classification**	**91.9**	**[89.4, 94.2]**	**91.7**
	Two Phase Analysis	88.6	[85.6, 91.2]	88.4
	Elimination Reasoning	87.7	[84.6, 90.6]	87.4
	Uncertainty Aware	89.2	[86.2, 91.8]	88.9
	Weighted Option Analysis	88.8	[86.0, 91.4]	88.7
Humanitarian	**Direct Classification**	**86.3**	**[83.2, 89.2]**	**76.1**
	Two Phase Analysis	82.2	[78.6, 85.5]	67.6
	Elimination Reasoning	79.5	[76.0, 83.0]	66.5
	Uncertainty Aware	84.4	[81.2, 87.6]	72.6
	Weighted Option Analysis	79.6	[76.2, 83.6]	60.6
Disaster Types	**Direct Classification**	**88.3**	**[85.6, 91.2]**	**83.5**
	Two Phase Analysis	84.5	[81.0, 87.5]	77.0
	Elimination Reasoning	85.6	[82.4, 88.6]	80.8
	Uncertainty Aware	86.0	[82.5, 88.8]	80.6
	Weighted Option Analysis	86.2	[83.0, 89.2]	80.5

Table E.3: Comparative prompt performance with GPT-4o. The best performing prompt is shown in **bold**.

*Confidence interval obtained via bootstrap resampling.

Statistical significance tests confirm this advantage ($p<0.05$) for GPT-4o across most tasks, with *Direct Classification* significantly outperforming all other prompts except in the *Damage Severity* classification task. Appendix E.5.2 provides the statistical test for different prompts using the best performing models, GPT-4o and Claude Sonnet.

Performance on humanitarian categories displays the most substantial divergence between accuracy and F1 scores. The *affected, injured or dead people* category—arguably the most critical for humanitarian response prioritisation—shows particularly poor performance, with Claude Sonnet achieving just 58% accuracy. These full, per-class breakdowns are available in Appendix E.5.1, Table E.8. One reason for this poor performance may be due to 'soft refusals', in which the model explicitly or implicitly avoids referring to human suffering, a constraint with serious implication for our task and which we will discuss more extensively later.

The *Uncertainty Aware* prompt generally ranked second in effectiveness, but demonstrated better performance for GPT-4o on *Damage Severity* assessment (84%) as shown in Table E.3. Explicit confidence assessment mechanisms may help models navigate ambiguous classifications, especially for categories with subjective boundaries. The relatively poor performance of the *Elimination Reasoning* prompt across models suggests that overly structured reasoning may constrain rather than improve visual analysis capabilities. Full, by task breakdowns of results for each prompt and model are available in Appendix E.5.1.

With this in mind, our goal is to redesign a prompt that keeps the clarity of *Direct Clas-*

Task	Prompt	Accuracy (%)	Accuracy 95% CI*	F1 (%)
Damage Severity	**Direct Classification**	**84.1**	**[80.6, 87.4]**	**78.8**
	Two Phase Analysis	81.0	[77.7, 84.4]	73.2
	Elimination Reasoning	78.0	[74.0, 81.8]	70.0
	Uncertainty Aware	79.0	[74.9, 82.4]	72.0
	Weighted Option Analysis	77.7	[74.1, 81.0]	73.1
Informative	Direct Classification	87.4	[84.2, 90.2]	87.4
	Two Phase Analysis	**87.4**	**[84.5, 90.4]**	**87.5**
	Elimination Reasoning	83.3	[79.8, 86.4]	84.1
	Uncertainty Aware	86.2	[83.2, 89.0]	86.1
	Weighted Option Analysis	82.0	[78.7, 85.2]	83.6
Humanitarian	**Direct Classification**	**82.0**	**[78.6, 85.2]**	**70.9**
	Two Phase Analysis	81.4	[78.0, 84.6]	70.0
	Elimination Reasoning	75.1	[71.4, 78.6]	67.1
	Uncertainty Aware	79.1	[75.7, 82.5]	68.9
	Weighted Option Analysis	76.7	[72.7, 80.3]	69.9
Disaster Types	Direct Classification	83.0	[79.6, 86.4]	79.3
	Two Phase Analysis	**83.2**	**[80.2, 86.5]**	**76.7**
	Elimination Reasoning	80.1	[76.4, 83.4]	75.8
	Uncertainty Aware	80.8	[77.3, 84.4]	75.9
	Weighted Option Analysis	77.1	[73.6, 80.9]	75.5

Table E.4: Comparative prompt performance with Claude Sonnet. The best performing prompt is shown in **bold**.

*Confidence interval obtained via bootstrap resampling.

sification, to preserve good performance across the broadest set of tasks, while borrowing a small dose of probability-style reasoning. Statistical analysis supports this hybrid approach, as we found significant performance differences between prompting strategies ($p<0.05$), with *Direct Classification* demonstrating superior performance for most tasks while *Uncertainty Aware* excelled specifically at *Damage Severity* assessment.

E.1.4 Confusion Matrices

The confusion matrices for GPT-4o in Figure E.1 display inter-class confusion patterns that require clarification in our final classification prompt.

For the *Disaster Type* category, the matrices demonstrate substantial confusion between *other disaster* and *not disaster* (23% in direct classification), as well as *hurricane* being misclassified as *flood* (10-15%). In response, we improve category definitions with explicit distinguishing criteria—emphasising storm-specific wind damage for hurricanes and clarifying that *other disaster* specifically involves human-caused emergencies whilst *not disaster* completely lacks emergency aspects.

For the *Damage Severity* category, the confusion matrices confirm the significant improvement in *mild* damage classification with the uncertainty-aware approach (55% to 80% accuracy). We therefore introduce quantitative damage thresholds (10-50% for *mild*, >50% for *severe*) and

functional assessment criteria (habitability, usability) to encourage more deliberate evaluation.

Regarding the *Humanitarian* category, the matrices indicate confusion between *affected people* and *not humanitarian* (21-26%). We address these patterns by emphasising the human presence in *affected people*, clarifying that *infrastructure damage* applies only when structures are damaged without people present, and providing explicit prioritisation rules for overlapping scenarios.

E.1.5 Confidence Intervals

In Table E.5 we calculated bootstrap confidence intervals to confirm the model ranking from Section E.1.2, here focusing on the *Direct Classification* prompt.

GPT-4o demonstrates not only high accuracy but also consistent performance, with the narrowest confidence intervals for Informative (89.6%-94.1%) and *Disaster Types* (85.4%-91.0%) classification. Claude Sonnet shows particularly reliable performance on *Damage Severity* assessment, with a tight confidence interval of 80.8%-87.4% compared to GPT-4o's wider 78.6%-85.1% interval. In contrast, Pixtral Small exhibits the widest confidence intervals across all tasks, indicating less consistent performance.

Task	Model	F1 (%)	Accuracy (%)	95% CI
Damage Severity	Claude Haiku	68.8	77.6	[74.0, 81.2]
	Claude Sonnet	**78.8**	**84.2**	**[81.4, 87.2]**
	GPT-4o	74.1	81.9	[78.5, 85.4]
	Pixtral Large	66.8	78.2	[74.6, 81.8]
	Pixtral Small	65.6	75.3	[71.4, 78.8]
Informative	Claude Haiku	84.4	84.2	[81.0, 87.2]
	Claude Sonnet	87.4	87.5	[84.4, 90.2]
	GPT-4o	**91.7**	**91.8**	**[89.4, 94.0]**
	Pixtral Large	87.1	87.5	[84.6, 90.4]
	Pixtral Small	87.0	87.6	[84.6, 90.2]
Humanitarian	Claude Haiku	70.8	79.2	[75.8, 82.5]
	Claude Sonnet	70.9	81.8	[78.4, 84.9]
	GPT-4o	**76.1**	**86.5**	**[83.8, 89.5]**
	Pixtral Large	71.0	82.8	[79.3, 85.8]
	Pixtral Small	61.3	75.0	[71.4, 79.0]
Disaster Types	Claude Haiku	74.8	80.8	[77.5, 84.0]
	Claude Sonnet	79.3	83.0	[79.7, 86.2]
	GPT-4o	**83.5**	**88.4**	**[85.3, 91.2]**
	Pixtral Large	75.0	80.3	[76.7, 83.8]
	Pixtral Small	63.6	73.8	[69.8, 77.4]

Table E.5: Comparative model performance with Direct Classification Prompt. The best performing model is shown in **bold**. Confidence interval obtained via bootstrap resampling.

E.1.6 Prompt Processing Time

We show in Table E.6 the average processing times, a crucial quantity for real-time crisis response, for different prompts and models. Relative times correlate with prompt complexity and cost. We find that GPT-4o consistently offered the fastest processing, while Pixtral Large was significantly slower, especially with complex prompts like elimination reasoning.

Table E.6: Average Processing Time (seconds per image).

Model	direct	two-phase	elimination	uncertainty	weighted
Claude 3.5 Sonnet	0.86	2.00	2.32	0.87	1.29
Claude 3.5 Haiku	1.05	1.79	1.97	1.16	1.14
GPT-4o	0.47	1.54	2.02	0.68	0.88
Pixtral Large	1.67	5.61	7.81	4.16	4.75
Pixtral Small	0.72	1.75	2.25	1.33	1.40

Average processing times (seconds per image) for zero-shot classification across multiple models and prompting strategies on the MEDIC dataset, using official APIs from the model providers (Anthropic, OpenAI, Mistral AI). Lower times indicate faster processing. In each column, the best (lowest) time is highlighted in green and the worst (highest) time in red.

Figure E.1: Confusion matrices for GPT-4o zero-shot classification using Direct Classification prompt (left) versus Uncertainty Aware prompt (right).

Disaster Types (Direct Classification)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.86	0	.01	0	0	.06	.06
fire	0	.84	0	0	0	.09	.06
flood	0	0	.90	0	0	.07	.01
hurr.	.01	0	.10	.70	0	.14	.03
land.	0	0	.04	0	.81	.09	.04
none	.00	.00	0	0	.00	.96	.01
other	0	.04	.04	0	0	.23	.66

Disaster Types (Uncertainty Aware)

True \ Predicted	quake	fire	flood	hurr.	land.	none	other
quake	.86	.01	.01	0	0	.04	.06
fire	0	.87	0	0	0	.06	.06
flood	0	0	.94	0	0	.05	0
hurr.	0	0	.15	.71	0	.07	.05
land.	0	0	.04	0	.81	.09	.04
none	.01	.01	.00	.01	.00	.90	.04
other	0	.04	.04	0	0	.38	.52

Informativeness (Direct Classification)

True \ Predicted	not inf	inf
not inf	.88	.11
inf	.02	.97

Informativeness (Uncertainty Aware)

True \ Predicted	not inf	inf
not inf	.88	.11
inf	.09	.90

Humanitarian (Direct Classification)

True \ Predicted	affected	infra	not hum	rescue
affected	.63	.10	.26	0
infra	.04	.83	.08	.02
not hum	.00	.02	.93	.02
rescue	.04	.06	.16	.72

Humanitarian (Uncertainty Aware)

True \ Predicted	affected	infra	not hum	rescue
affected	.63	.15	.21	0
infra	.04	.86	.07	.02
not hum	.00	.06	.89	.03
rescue	.09	.18	.13	.58

Damage Severity (Direct Classification)

True \ Predicted	none	mild	severe
none	.96	.03	.00
mild	.39	.55	.04
severe	.13	.16	.70

Damage Severity (Uncertainty Aware)

True \ Predicted	none	mild	severe
none	.88	.10	.01
mild	.08	.80	.11
severe	.02	.16	.80

The matrices reveal significant performance shifts, most notably the improvement in *mild* damage severity classification (from 55% to 80%) and *flood* identification (90% to 94%). Certain categories deteriorated, particularly *other disaster* (66% to 52%) with increased confusion with *not disaster*, and *rescue volunteering* (72% to 58%). Dark blue cells indicate correct classifications, light blue cells show mediocre performance, and red cells highlight problematic misclassifications.

E.2 Prompts Tested

Direct Classification Prompt

```
As a Humanitarian Crisis Image Analyst, examine this image and provide your
    assessment. Some images may show actual disasters while others may show normal
    scenes with no disaster present. First, analyze what you see, then classify it
    according to standard humanitarian response categories.

<analysis>
Examine the image carefully and describe what you see related to potential disaster,
     damage, humanitarian concerns, and informativeness. If no disaster is present,
    simply describe what you see in the image. Don't mention specific label
    categories yet, just describe what's visually present.
</analysis>

Now, based on your analysis, classify this image according to these categories:

1. DISASTER TYPE:
   - earthquake: damaged/destroyed buildings, fractured houses, ground ruptures
   - fire: man-made fires or wildfires, destroyed forests, houses, infrastructures
   - flood: flooded areas, houses, roads, other infrastructures
   - hurricane: high winds, storm surge, heavy rains, collapsed electricity polls,
    grids, trees
   - landslide: landslide, mudslide, landslip, rockfall, rockslide, earth slip, land
     collapse
   - not_disaster: cartoon, advertisement, or anything not easily linked to any
    disaster type
   - other_disaster: plane crash, bus/car/train accident, explosion, war, conflicts

2. DAMAGE SEVERITY:
   - severe: substantial destruction making infrastructure non-livable/non-usable
   - mild: partially destroyed buildings/bridges/houses/roads (approximately up to
    50% damage)
   - little_or_none: damage-free infrastructure (except for normal wear and tear)

3. INFORMATIVENESS:
   - informative: useful for humanitarian aid
   - not_informative: not useful for humanitarian aid (ads, logos, cartoons, blurred
     images)

4. HUMANITARIAN CATEGORY:
   - affected_injured_or_dead_people: shows injured, dead, or affected people
   - infrastructure_and_utility_damage: shows built structures affected/damaged
   - not_humanitarian: not relevant for humanitarian aid
   - rescue_volunteering_or_donation_effort: shows rescue, volunteering, or response
    efforts
```

```
<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.2: Direct Classification prompt for zero-shot disaster image classification

Two Phase Analysis Prompt

```
As a Humanitarian Crisis Image Analyst, your task is to classify this image to
    support emergency response decisions. Note that some images may show actual
    disasters while others may show normal scenes with no disaster present. Complete
     this process in two phases:

<analysis>
1. OBSERVATION: Briefly describe what you see in this image (3-5 sentences, focusing
     on visible elements).

2. ASSESSMENT: For each category, assess the possible classifications:
   - DISASTER TYPE: What type of disaster is shown, if any? If no disaster is
    present, indicate this clearly.
   - DAMAGE SEVERITY: How severe is any visible damage to structures or
    infrastructure?
   - INFORMATIVENESS: Would this image be useful for humanitarian response?
   - HUMANITARIAN CATEGORY: What is the primary humanitarian concern shown, if any?

3. CONFIDENCE: For each category, indicate your confidence level (high/medium/low)
    and why.
</analysis>

Now, based solely on your analysis above, classify this image with EXACTLY ONE
    category for each task:

<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.3: Two Phase Analysis prompt for zero-shot disaster image classification

Elimination Reasoning Prompt

```
As a Humanitarian Crisis Image Analyst, evaluate this image using a systematic
    elimination approach to support response prioritization. Note that some images
    may show actual disasters while others may show normal scenes with no disaster
    present. Complete your analysis within the analysis tags, and then provide your
    final classification.

<analysis>
1. KEY ELEMENTS: List the key visible elements in this image (buildings, people,
    water, fire, etc.)

2. SYSTEMATIC EVALUATION:

DISASTER TYPE:
- Earthquake evidence:
- Fire evidence:
- Flood evidence:
- Hurricane evidence:
- Landslide evidence:
- Not disaster evidence:
- Other disaster evidence:
- Reasoning and elimination process: [explain which options you're eliminating and
    why]

DAMAGE SEVERITY:
- Severe evidence:
- Mild evidence:
- Little or none evidence:
- Reasoning and elimination process: [explain which options you're eliminating and
    why]

INFORMATIVENESS:
- Informative evidence:
- Not informative evidence:
- Reasoning and elimination process: [explain which options you're eliminating and
    why]

HUMANITARIAN CATEGORY:
- Affected/injured/dead people evidence:
- Infrastructure/utility damage evidence:
- Not humanitarian evidence:
- Rescue/volunteering/donation evidence:
- Reasoning and elimination process: [explain which options you're eliminating and
    why]
</analysis>

Based strictly on your analysis above, provide your final classification:
```

```
<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.4: Elimination Reasoning prompt for zero-shot disaster image classification

Uncertainty Aware Prompt

```
As a Humanitarian Crisis Image Analyst with field experience, analyze this image
    with attention to certainty levels. Note that some images may show actual
    disasters while others may show normal scenes with no disaster present. First
    complete your entire analysis, then provide only the final labels.

<analysis>
1. IMAGE DESCRIPTION: Describe what you see in the image (1-3 sentences).

2. CLASSIFICATION WITH CONFIDENCE:

DISASTER TYPE:
- Options: earthquake, fire, flood, hurricane, landslide, not_disaster, other_
    disaster
- My assessment: [explain what disaster type you believe is shown and why]
- Confidence (0-100%): [percentage]
- Uncertainty factors: [if confidence is <70%, explain what makes this
    classification difficult]

DAMAGE SEVERITY:
- Options: little_or_none, mild, severe
- My assessment: [explain what level of damage you see and why]
- Confidence (0-100%): [percentage]
- Uncertainty factors: [if confidence is <70%, explain what makes this
    classification difficult]

INFORMATIVENESS:
- Options: informative, not_informative
- My assessment: [explain whether this is informative for humanitarian response and
    why]
- Confidence (0-100%): [percentage]
- Uncertainty factors: [if confidence is <70%, explain what makes this
    classification difficult]

HUMANITARIAN CATEGORY:
- Options: affected_injured_or_dead_people, infrastructure_and_utility_damage, not_
    humanitarian, rescue_volunteering_or_donation_effort
- My assessment: [explain what humanitarian category is most prominent and why]
- Confidence (0-100%): [percentage]
- Uncertainty factors: [if confidence is <70%, explain what makes this
    classification difficult]

3. FINAL DETERMINATION:
For any categories with low confidence, provide additional reasoning to reach a
    final decision.
</analysis>
```

```
Based on your complete analysis, provide your final classification:

<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.5: Uncertainty Aware prompt for zero-shot disaster image classification

Weighted Option Analysis Prompt

```
As a Humanitarian Crisis Image Analyst trained in probability assessment, evaluate
    this image using quantitative confidence ratings. Note that some images may show
     actual disasters while others may show normal scenes with no disaster present.
    First provide your detailed analysis, then give only your final classifications.

<analysis>
1. IMAGE DESCRIPTION: Briefly describe what you see in the image.

2. PROBABILITY ASSESSMENT: For each category, assign a probability percentage to
    each option based on visual evidence (ensure percentages sum to 100% per
    category):

DISASTER TYPE:
- earthquake: __% (damaged buildings, ground ruptures)
- fire: __% (flames, smoke, burned structures/forests)
- flood: __% (water inundation of normally dry areas)
- hurricane: __% (wind damage, fallen trees/poles)
- landslide: __% (earth/mud/rock displacement)
- not_disaster: __% (unrelated to disasters)
- other_disaster: __% (accidents, explosions, conflicts)

DAMAGE SEVERITY:
- severe: __% (infrastructure non-functional/non-usable)
- mild: __% (partial damage, ~50% destruction)
- little_or_none: __% (minimal/no visible damage)

INFORMATIVENESS:
- informative: __% (useful for humanitarian response)
- not_informative: __% (not useful for response)

HUMANITARIAN CATEGORY:
- affected_injured_or_dead_people: __% (showing impacted people)
- infrastructure_and_utility_damage: __% (damaged structures/utilities)
- not_humanitarian: __% (not relevant to humanitarian aid)
- rescue_volunteering_or_donation_effort: __% (showing response activities)

3. JUSTIFICATION: For each highest-probability selection, provide a brief
    justification (1-2 sentences).

4. UNCERTAINTY RESOLUTION: If you have any categories where two options have similar
     probabilities (within 15% of each other), explain your final decision process.
</analysis>

Based on your analysis, provide your final classification with the highest
    probability option for each category:
```

```
<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.6: Weighted Option Analysis prompt for zero-shot disaster image classification

E.3 Final Zero-Shot Classification Prompt

```
As a Humanitarian Crisis Image Analyst, examine this image and provide your
    assessment. Some images may show actual disasters while others may show normal
    scenes with no disaster present. First, analyze what you see, then classify it
    according to standard humanitarian response categories.

<analysis>
Examine the image carefully and describe what you see related to potential disaster,
     damage, humanitarian concerns, and informativeness. Focus specifically on:
1. Any visible damage or functional disruption to structures, landscapes, or
    infrastructure
2. Whether people are present, and if so, their condition, activities, and how they
    appear to be affected
3. Signs that indicate the specific type of disaster (if any)
4. Overall clarity and relevance of the image for humanitarian response

Be thorough and descriptive without mentioning specific label categories yet.
</analysis>

Now, based on your analysis, classify this image according to these categories:

1. DISASTER TYPE:
   - earthquake: damaged/destroyed buildings with characteristic structural collapse
     patterns, fractured houses, visible cracks in walls/foundations, ground
    ruptures
   - fire: active flames, smoke, charred/blackened buildings or forests, burned
    debris
   - flood: standing water covering roads/fields/urban areas, water marks on
    buildings, people wading through water
   - hurricane: downed trees, roof damage, debris scattered by wind, power lines
    down, storm surge effects
   - landslide: displaced soil/rocks, buried structures, visible slope failure, mud
    flows, blocked roads
   - not_disaster: EVERYDAY SCENES without disaster evidence, cartoons,
    advertisements - if you're uncertain whether something qualifies as a disaster,
    choose a specific disaster type rather than not_disaster
   - other_disaster: transportation accidents (plane/bus/car/train), explosions, war
     damage, conflicts, industrial accidents

2. DAMAGE SEVERITY:
   Look carefully at infrastructure and assess BOTH structural damage AND functional
     impact:
   - severe: substantial destruction OR major functional impairment (collapsed walls
    , exposed interior, missing roofs, completely flooded areas, completely
    impassable roads)
   - mild: ANY partial damage OR functional limitation (visible cracks, broken
    windows, damaged roofs, partially flooded areas, partially blocked roads)
```

```
   - little_or_none: fully intact structures with only cosmetic damage AND no
   significant functional impairment

   *Important: If there is ANY impact on the usability or function of infrastructure
   , even if the structure appears intact, classify as at least "mild"*

3. INFORMATIVENESS:
   - informative: contains clear, useful visual information for humanitarian aid
   assessment or response
   - not_informative: blurry, artistic, promotional, or lacks clear disaster-
   relevant content (ads, logos, cartoons, symbolic images)

4. HUMANITARIAN CATEGORY:
   Use these clear distinguishing criteria for classification:

   - affected_injured_or_dead_people: Shows INDIVIDUALS in distress or directly
   impacted. Look for:
     * People with visible injuries, receiving medical attention, or deceased
     * Civilians being evacuated, rescued, or in obvious distress
     * People in temporary shelters, receiving aid, or displaced
     * Close-up focus on human suffering or individual impact

   - infrastructure_and_utility_damage: Shows PHYSICAL DAMAGE as the main focus.
   Look for:
     * Damaged buildings, roads, bridges, or utilities WITHOUT people as the main
   subject
     * Debris, rubble, or destroyed property with no individuals prominently
   featured
     * Aerial or wide shots of damaged areas where infrastructure is the primary
   subject
     * People may be present but only as small figures that aren't the main focus

   - rescue_volunteering_or_donation_effort: Shows ORGANIZED RESPONSE efforts. Look
   for:
     * Uniformed personnel (firefighters, medical staff, military, etc.) actively
   responding
     * Emergency vehicles, equipment, or organized rescue operations
     * Coordinated aid distribution, volunteer efforts, or donation activities
     * Focus on the responders and their equipment/activities rather than victims

   - not_humanitarian: NO CLEAR DISASTER IMPACT or response needs. Look for:
     * Scenes unrelated to disasters or humanitarian needs
     * No visible damage, affected people, or response activities
     * Normal daily activities, undamaged structures, or scenic views
     * Promotional content, advertisements, or symbolic imagery

   *Important: If you see injured or affected people AND organized responders in the
    same image, look at what the image is primarily focusing on - the people being
```

```
    helped or the responders providing help*

In the <labels> section below, fill in ONLY the exact category names from the lists
     above. Do not include any brackets, explanations, or additional text. Use only
    the exact values listed for each category (e.g., "earthquake", "severe", "
    informative", "infrastructure_and_utility_damage").

<labels>
{
  "disaster_type": "",
  "damage_severity": "",
  "informative": "",
  "humanitarian": ""
}
</labels>
```

Figure E.7: Final Prompt used for Zero-Shot Classification Test

E.4 Fallback Prompts

```
I understand your caution with potentially sensitive content. To clarify, this is a
    legitimate academic research project on disaster classification. Your assessment
     will be used solely for evaluating ML classification systems that can help
    humanitarian organizations respond more effectively to disasters.

Please classify this image according to the four categories requested (disaster_type
    , damage_severity, informative, humanitarian). Remember that "not_disaster" and "
    not_humanitarian" are perfectly valid classifications if the image shows a
    normal scene. If this is not a disaster image, you can and should classify it
    accordingly rather than refusing to analyze it.

Please complete your analysis using the <analysis> and <labels> format as
    requested earlier.
```

Figure E.8: First fallback prompt for initial model refusal in disaster image classification

```
I understand your hesitation. To clarify: analyzing this image for disaster
    classification is part of an academic research project with ethical approval.
    The classification options already include categories for non-disaster images
    and non-humanitarian content.

Rather than refusing completely, please approach this as a technical classification
    task. At minimum, please determine whether this is a disaster or non-disaster
    image, and complete the classification using the <analysis> and <labels>
    format.

If the image contains no disaster, simply classify it as "not_disaster" - this is
    valuable information for the research.
```

Figure E.9: Second fallback prompt for continued model refusal in disaster image classification

E.5 Expanded Initial Prompt/Model Testing Results

E.5.1 Zero-Shot Small-Scale Test Results by Vision Model

Table E.7: Classification performance (Accuracy and F1 scores) for GPT-4V across prompts

Task / Class	Direct Classification		Two Phase Analysis		Elimination Reasoning		Uncertainty Aware		Weighted Option Analysis	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Damage Severity	82.0	74.1	84.0	76.8	82.2	73.3	84.4	78.9	80.8	72.4
Little Or None	96.5	89.7	94.1	92.3	93.0	89.5	88.3	91.9	94.1	88.8
Mild	55.7	51.1	63.9	54.5	47.5	46.0	80.3	58.7	50.8	46.6
Severe	70.5	81.6	76.5	83.6	78.7	84.5	80.3	86.2	72.1	81.7
Informative	91.9	91.7	88.6	88.4	87.6	87.4	89.2	88.9	88.8	88.7
Not Informative	97.0	90.6	91.1	86.7	89.2	85.4	90.6	87.2	94.6	87.3
Informative	88.2	92.7	86.9	90.1	86.5	89.4	88.2	90.7	84.8	90.2
Humanitarian	86.5	76.1	82.2	67.6	79.6	66.5	84.4	72.6	79.6	60.6
Affected/ Injured People	63.2	53.3	57.9	44.0	36.8	37.8	63.2	52.2	21.1	25.8
Infrastructure Damage	83.6	88.6	84.6	86.0	72.9	81.2	86.4	87.3	78.5	84.0
Not Humanitarian	93.8	90.3	88.8	87.9	92.4	84.8	89.3	89.1	91.1	86.3
Rescue/ Volunteering Effort	72.1	72.1	46.5	52.6	65.1	62.2	58.1	61.7	51.2	46.3
Disaster Types	88.4	83.5	84.4	77.0	85.6	80.8	86.0	80.6	86.2	80.5
Earthquake	86.6	91.0	76.8	85.7	87.8	92.3	86.6	91.0	87.8	89.4
Fire	84.4	87.1	84.4	85.7	87.5	90.3	87.5	86.2	90.6	86.6
Flood	90.9	87.7	89.1	85.2	85.5	83.9	94.5	86.0	92.7	86.4
Hurricane	70.2	82.5	61.4	74.5	66.7	76.0	71.9	81.2	75.4	80.4
Landslide	81.8	87.8	81.8	87.8	81.8	90.0	81.8	87.8	81.8	85.7
Not Disaster	96.1	92.5	96.5	90.7	93.1	88.8	90.5	90.3	90.5	90.1
Other Disaster	66.7	56.0	33.3	29.2	47.6	44.4	52.4	41.5	42.9	45.0
OVERALL	87.2	81.4	84.8	77.4	83.8	77.0	86.0	80.2	83.9	75.6

Classification performance (Accuracy and F1 scores) for GPT-4V across different prompting strategies. The highest-performing prompt per row is highlighted in green.

Table E.8: Classification performance (Accuracy and F1 scores) for Claude Sonnet across prompts

Task / Class	Direct Classification		Two Phase Analysis		Elimination Reasoning		Uncertainty Aware		Weighted Option Analysis	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Damage Severity	84.2	78.8	81.0	73.2	78.0	70.0	79.0	72.0	77.6	73.1
Little Or None	90.6	90.4	88.3	89.2	85.2	86.0	83.6	86.6	82.8	85.8
Mild	67.2	60.3	50.8	47.3	41.0	41.7	52.5	47.8	55.7	51.1
Severe	80.9	85.5	80.9	83.1	80.3	82.4	81.4	81.6	77.6	82.3
Informative	87.4	87.4	87.4	87.5	83.2	84.1	86.2	86.1	82.0	83.6
Not Informative	85.2	84.8	86.7	85.0	84.2	80.5	80.3	82.7	79.3	79.1
Informative	88.9	89.9	87.9	90.0	82.5	87.7	90.2	89.5	83.8	88.1
Humanitarian	82.0	70.9	81.4	70.0	75.0	67.1	79.0	68.9	76.6	69.9
Affected/ Injured People	42.1	44.4	42.1	45.7	57.9	42.3	47.4	42.9	57.9	51.2
Infrastructure Damage	83.6	86.9	87.9	88.1	67.3	79.3	78.5	84.6	77.6	84.7
Not Humanitarian	83.9	84.9	78.1	84.1	82.6	79.6	80.4	82.0	76.3	80.1
Rescue/ Volunteering Effort	81.4	67.3	83.7	62.1	81.4	67.3	88.4	66.1	81.4	63.6
Disaster Types	83.0	79.3	83.2	76.7	80.0	75.8	80.6	75.9	77.0	75.5
Earthquake	75.6	83.8	81.7	86.5	79.3	86.1	73.2	83.3	75.6	82.7
Fire	84.4	90.0	87.5	83.6	84.4	90.0	84.4	85.7	81.2	85.2
Flood	90.9	85.5	90.9	82.6	81.8	84.1	90.9	82.6	85.5	81.0
Hurricane	77.2	80.7	75.4	81.9	82.5	83.9	84.2	82.8	75.4	78.2
Landslide	77.3	85.0	72.7	80.0	54.5	68.6	63.6	77.8	63.6	77.8
Not Disaster	87.4	87.8	88.7	88.2	84.8	82.9	84.0	85.5	78.4	82.1
Other Disaster	61.9	42.6	33.3	34.1	38.1	34.8	47.6	33.9	57.1	41.4
OVERALL	84.2	79.1	83.2	76.9	79.0	74.2	81.2	75.7	78.3	75.5

Classification performance (Accuracy and F1 scores) for Claude Sonnet model across different prompting strategies. The highest-performing prompt per row is highlighted in green.

Table E.9: Classification performance (Accuracy and F1 scores) for Claude Haiku across prompts

Task / Class	Direct Classification		Two Phase Analysis		Elimination Reasoning		Uncertainty Aware		Weighted Option Analysis	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Damage Severity	77.6	68.8	78.6	69.5	77.4	66.8	78.2	69.8	73.4	65.4
Little Or None	89.8	88.0	87.9	88.8	88.3	86.6	85.2	88.1	79.7	83.8
Mild	44.3	40.9	41.0	38.8	32.8	33.6	47.5	40.3	36.1	33.3
Severe	71.6	77.5	78.1	81.0	77.0	80.1	78.7	81.1	77.0	79.2
Informative	84.2	84.4	82.2	82.8	68.8	61.4	80.4	80.2	79.2	80.4
Not Informative	85.7	81.7	85.7	80.0	29.6	43.6	76.4	76.2	75.4	75.4
Informative	83.2	87.1	79.8	85.6	95.6	79.1	83.2	84.2	81.8	85.4
Humanitarian	79.2	70.8	74.0	60.3	75.0	65.3	76.0	66.6	69.4	57.3
Affected/ Injured People	68.4	49.1	36.8	32.6	57.9	44.9	63.2	43.6	57.9	31.9
Infrastructure Damage	77.6	83.0	79.4	81.1	70.6	78.6	74.3	81.7	69.6	77.2
Not Humanitarian	83.0	82.9	76.3	78.1	83.5	79.4	78.6	79.5	75.9	77.1
Rescue/ Volunteering Effort	72.1	68.1	51.2	49.4	60.5	58.4	76.7	61.7	39.5	43.0
Disaster Types	80.8	74.8	78.2	70.2	79.2	72.1	80.8	74.5	75.8	72.1
Earthquake	67.1	79.7	70.7	81.1	74.4	82.4	78.0	85.9	78.0	76.2
Fire	87.5	88.9	87.5	88.9	84.4	87.1	87.5	88.9	84.4	85.7
Flood	85.5	84.7	81.8	82.6	74.5	79.6	74.5	77.4	78.2	77.5
Hurricane	61.4	72.9	63.2	72.0	64.9	73.3	66.7	75.2	63.2	72.0
Landslide	63.6	65.1	59.1	61.9	50.0	61.1	54.5	68.6	50.0	62.9
Not Disaster	93.5	85.9	90.0	82.4	92.2	83.4	92.6	84.6	81.8	82.0
Other Disaster	42.9	46.2	14.3	22.2	28.6	37.5	33.3	41.2	42.9	48.6
OVERALL	80.5	74.7	78.2	70.7	75.1	66.4	78.9	72.8	74.5	68.8

Classification performance (Accuracy and F1 scores) for Claude Haiku model across different prompting strategies. The highest-performing prompt per row is highlighted in green.

Table E.10: Classification performance (Accuracy and F1 scores) for Pixtral Small across prompts

Task / Class	Direct Classification		Two Phase Analysis		Elimination Reasoning		Uncertainty Aware		Weighted Option Analysis	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Damage Severity	75.2	65.6	76.4	66.9	77.8	66.1	73.6	66.7	77.8	62.8
Little Or None	91.4	88.6	89.8	88.8	87.1	88.7	77.0	85.1	89.8	88.1
Mild	45.9	36.4	45.9	36.8	31.1	30.9	57.4	38.5	18.0	22.0
Severe	62.3	71.9	67.8	75.2	80.3	78.6	74.3	76.6	80.9	78.3
Informative	87.6	87.0	84.0	83.6	75.4	72.8	86.2	85.5	86.6	86.4
Not Informative	78.3	83.9	84.2	81.0	54.7	64.3	79.3	82.4	90.6	84.6
Informative	93.9	90.0	83.8	86.2	89.6	81.2	90.9	88.7	83.8	88.1
Humanitarian	75.0	61.3	71.0	56.0	70.8	52.9	72.6	60.5	74.0	52.9
Affected/ Injured People	73.7	31.8	57.9	23.7	31.6	20.0	68.4	32.5	68.4	29.2
Infrastructure Damage	70.1	77.5	65.9	76.0	60.3	70.7	72.9	77.0	72.4	78.5
Not Humanitarian	84.8	86.6	83.5	81.0	91.5	81.3	75.9	80.4	87.9	85.1
Rescue/ Volunteering Effort	48.8	49.4	37.2	43.2	32.6	39.4	55.8	52.2	11.6	18.9
Disaster Types	73.8	63.6	73.6	63.2	76.6	62.4	76.2	66.6	79.2	68.8
Earthquake	46.3	62.8	56.1	71.3	76.8	72.8	70.7	76.3	73.2	80.0
Fire	93.8	84.5	84.4	84.4	90.6	82.9	87.5	78.9	93.8	85.7
Flood	83.6	70.8	72.7	76.2	80.0	79.3	78.2	76.1	83.6	76.7
Hurricane	56.1	69.6	42.1	57.8	38.6	53.7	71.9	77.4	59.6	71.6
Landslide	40.9	40.0	36.4	43.2	40.9	50.0	40.9	48.6	31.8	42.4
Not Disaster	87.4	88.4	92.2	85.5	92.6	86.6	84.4	84.1	90.0	87.9
Other Disaster	57.1	28.9	47.6	23.8	9.5	11.8	33.3	24.6	52.4	37.3
OVERALL	77.9	69.4	76.2	67.4	75.2	63.5	77.2	69.8	79.4	67.7

Classification performance (Accuracy and F1 scores) for Pixtral Small model across different prompting strategies. The highest-performing prompt per row is highlighted in green.

Table E.11: Classification performance (Accuracy and F1 scores) for Pixtral Large across prompts

Task / Class	Direct Classification		Two Phase Analysis		Elimination Reasoning		Uncertainty Aware		Weighted Option Analysis	
	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)	Acc (%)	F1 (%)
Damage Severity	78.2	66.8	81.0	73.0	79.0	69.8	76.0	63.3	76.2	67.7
Little Or None	94.1	88.6	89.1	89.6	91.4	88.1	82.4	88.1	84.0	87.8
Mild	34.4	35.0	52.5	48.5	44.3	42.9	24.6	25.9	49.2	37.3
Severe	70.5	76.8	79.2	81.0	73.2	78.4	84.2	76.0	74.3	77.9
Informative	87.4	87.1	86.8	86.4	83.0	82.5	87.0	86.3	87.6	87.2
Not Informative	88.2	85.0	83.3	83.7	78.8	79.2	79.3	83.2	86.7	85.0
Informative	86.9	89.1	89.2	89.1	85.9	85.7	92.3	89.4	88.2	89.4
Humanitarian	82.8	71.0	78.6	67.9	79.2	68.1	80.8	67.5	79.0	63.8
Affected/ Injured People	52.6	47.6	57.9	50.0	47.4	46.2	36.8	38.9	47.4	43.9
Infrastructure Damage	83.2	86.6	84.1	84.7	72.4	80.9	89.3	87.8	85.5	84.9
Not Humanitarian	87.5	86.7	75.9	82.3	91.1	84.0	75.9	82.9	81.7	84.7
Rescue/ Volunteering Effort	69.8	63.2	74.4	54.7	65.1	61.5	83.7	60.5	46.5	41.7
Disaster Types	80.4	75.0	80.4	75.5	80.0	71.9	80.6	76.1	78.8	71.6
Earthquake	76.8	80.8	82.9	88.3	85.4	84.8	82.9	84.5	87.8	81.8
Fire	81.2	82.5	84.4	83.1	81.2	85.2	87.5	86.2	93.8	85.7
Flood	80.0	82.2	85.5	77.7	80.0	81.5	83.6	80.7	87.3	75.6
Hurricane	77.2	80.7	75.4	79.6	61.4	70.0	77.2	77.9	59.6	64.2
Landslide	63.6	71.8	63.6	73.7	54.5	66.7	72.7	82.1	68.2	71.4
Not Disaster	85.3	86.2	81.4	85.6	89.2	85.5	81.8	85.7	81.0	86.0
Other Disaster	66.7	40.6	71.4	40.5	33.3	29.8	57.1	35.8	38.1	36.4
OVERALL	82.2	75.0	81.7	75.7	80.3	73.1	81.1	73.3	80.4	72.6

Classification performance (Accuracy and F1 scores) for Pixtral Large model across different prompting strategies. The highest-performing prompt per row is highlighted in green.

E.5.2 Statistical Significance Testing

Task	Comparison	p-value	Superior Model
Damage Severity	Claude Haiku vs Claude Sonnet	0.0000	Claude Sonnet
	Claude Haiku vs GPT-4o	0.0269	GPT-4o
	Claude Sonnet vs Pixtral Large	0.0018	Claude Sonnet
	Claude Sonnet vs Pixtral Small	0.0000	Claude Sonnet
	GPT-4o vs Pixtral Large	0.0327	GPT-4o
	GPT-4o vs Pixtral Small	0.0006	GPT-4o
Informative	Claude Haiku vs Claude Sonnet	0.0450	Claude Sonnet
	Claude Haiku vs GPT-4o	0.0000	GPT-4o
	Claude Sonnet vs GPT-4o	0.0036	GPT-4o
	GPT-4o vs Pixtral Large	0.0050	GPT-4o
	GPT-4o vs Pixtral Small	0.0244	GPT-4o
Humanitarian	Claude Haiku vs GPT-4o	0.0001	GPT-4o
	Claude Haiku vs Pixtral Small	0.0423	Claude Haiku
	Claude Sonnet vs GPT-4o	0.0121	GPT-4o
	Claude Sonnet vs Pixtral Small	0.0008	Claude Sonnet
	GPT-4o vs Pixtral Small	0.0000	GPT-4o
	Pixtral Large vs Pixtral Small	0.0002	Pixtral Large
Disaster Types	Claude Haiku vs GPT-4o	0.0000	GPT-4o
	Claude Haiku vs Pixtral Small	0.0004	Claude Haiku
	Claude Sonnet vs GPT-4o	0.0006	GPT-4o
	Claude Sonnet vs Pixtral Small	0.0000	Claude Sonnet
	GPT-4o vs Pixtral Large	0.0000	GPT-4o
	GPT-4o vs Pixtral Small	0.0000	GPT-4o
	Pixtral Large vs Pixtral Small	0.0016	Pixtral Large

Table E.12: Statistically significant differences between models, using McNemar tests on the Direct Classification prompt results

Task	Comparison	p-value	Superior Prompt
Damage Severity	Uncertainty Aware vs Weighted Option Analysis	0.0282	Uncertainty Aware
	Two Phase Analysis vs Weighted Option Analysis	0.0450	Two Phase Analysis
Informative	Direct Classification vs Elimination Reasoning	0.0018	Direct Classification
	Direct Classification vs Two Phase Analysis	0.0046	Direct Classification
	Direct Classification vs Weighted Option Analysis	0.0051	Direct Classification
	Direct Classification vs Uncertainty Aware	0.0311	Direct Classification
Humanitarian	Direct Classification vs Elimination Reasoning	0.0001	Direct Classification
	Direct Classification vs Weighted Option Analysis	0.0000	Direct Classification
	Direct Classification vs Two Phase Analysis	0.0070	Direct Classification
	Elimination Reasoning vs Uncertainty Aware	0.0067	Uncertainty Aware
	Uncertainty Aware vs Weighted Option Analysis	0.0035	Uncertainty Aware
Disaster Types	Direct Classification vs Two Phase Analysis	0.0011	Direct Classification
	Direct Classification vs Elimination Reasoning	0.0303	Direct Classification
	Direct Classification vs Uncertainty Aware	0.0446	Direct Classification

Table E.13: Statistically significant differences between prompts, using McNemar tests on GPT-4o results

Task	Comparison	p-value	Superior Prompt
Damage Severity	Direct Classification vs Two Phase Analysis	0.0150	Direct Classification
	Direct Classification vs Elimination Reasoning	0.0000	Direct Classification
	Direct Classification vs Uncertainty Aware	0.0002	Direct Classification
	Direct Classification vs Weighted Option Analysis	0.0000	Direct Classification
	Two Phase Analysis vs Weighted Option Analysis	0.0372	Two Phase Analysis
Informative	Direct Classification vs Elimination Reasoning	0.0060	Direct Classification
	Direct Classification vs Weighted Option Analysis	0.0001	Direct Classification
	Two Phase Analysis vs Elimination Reasoning	0.0060	Two Phase Analysis
	Two Phase Analysis vs Weighted Option Analysis	0.0001	Two Phase Analysis
	Uncertainty Aware vs Weighted Option Analysis	0.0070	Uncertainty Aware
Humanitarian	Direct Classification vs Elimination Reasoning	0.0001	Direct Classification
	Direct Classification vs Uncertainty Aware	0.0411	Direct Classification
	Direct Classification vs Weighted Option Analysis	0.0005	Direct Classification
	Two Phase Analysis vs Elimination Reasoning	0.0014	Two Phase Analysis
	Two Phase Analysis vs Weighted Option Analysis	0.0046	Two Phase Analysis
	Elimination Reasoning vs Uncertainty Aware	0.0232	Uncertainty Aware
Disaster Types	Direct Classification vs Weighted Option Analysis	0.0001	Direct Classification
	Two Phase Analysis vs Elimination Reasoning	0.0237	Two Phase Analysis
	Two Phase Analysis vs Uncertainty Aware	0.0485	Two Phase Analysis
	Two Phase Analysis vs Weighted Option Analysis	0.0002	Two Phase Analysis
	Uncertainty Aware vs Weighted Option Analysis	0.0282	Uncertainty Aware

Table E.14: Statistically significant differences between prompts, using McNemar tests on Claude Sonnet results

Task	Model	Best Prompt	Accuracy (%)	95% CI
Damage Severity	GPT-4o	Uncertainty Aware	84.47	[81.50, 87.80]
	Claude Sonnet	Direct Classification	84.20	[80.80, 87.51]
Informative	GPT-4o	Direct Classification	91.78	[89.40, 94.00]
	Claude Sonnet	Two Phase Analysis	87.40	[84.60, 90.20]
Humanitarian	GPT-4o	Direct Classification	86.41	[83.20, 89.51]
	Claude Sonnet	Direct Classification	82.04	[78.50, 85.60]
Disaster Types	GPT-4o	Direct Classification	88.30	[85.40, 91.20]
	Claude Sonnet	Two Phase Analysis	83.15	[79.80, 86.51]

Table E.15: Comparative Performance of Best Prompts by Model