Developing a framework for the Integration and Use of Large Language Models in Disaster Risk Management

by
Veronica Tuazon, Xinying Gao, Natalie Yimei Gahre

Final Capstone Project Presented to the
Faculty of the School of Politics, Economics & Global Affairs at IE University in
Partial Fulfilment of the Requirements for the Degree of
Master in International Development (MID)

Supervised by
Jean-Baptiste Bove

In Collaboration with CIMA Research Foundation

IE University
June 2024

Table of Contents

Disclaimer - Usage of AI

For research purposes, the project team utilized AI, specifically ChatGPT-4 and ChatGPT-4o, in the following capacities:

1. **Research & Project Purposes:** Due to the nature of the project, ChatGPT-4 was used on for research purposes as follows:
    - **Generation of answers to research prompts:** As part of the analysis the performance of different Large Language Models (LLMs) was evaluated. For these purposes, ChatGPT-4 was used to generate answers to the project's research prompts as explained in the report. These answers were then used as data for the analysis.
    - **Evaluation of the data:** the generated output from ChatGPT-4 and other LLMs was evaluated by ChatGPT-4, and the evaluation scores then used for the final results of the analysis as described in the report:
    - **Data analysis:** For analyzing the final results of the prompt and answer evaluation, ChatGPT was used to understand the data better and perform the statistical analysis.
    - **Prompt Building:** ChatGPT-4 and ChatGPT-4o were consulted for structure and coherence for prompt in the research and for custom GPTs
2. **Checking redundancies:** After the final report was drafted and reviewed by the project team, ChatGPT-4o was used to spot redundancies within the report. The edits based on the ChatGPT-4o responses were done by the project team and were not performed by AI.
3. **Translation & Language:** ChatGPT-4 and ChatGPT-4o were used to provide translation and language check. The content and ideas were not altered or generated by AI.

List of Figures

List of Tables

Executive Summary/ Abstract

This report explores the capabilities and limitations of Large Language Models (LLMs) applications in the Disaster Risk Management (DRM) field, focusing on the text-based workflow of CIMA professionals. By leveraging different Artificial Intelligence (AI) techniques and LLM tools, experiments of different DRM tasks were conducted in various regions to thoroughly analyze the current potential of the utilizations. The findings from both primary and secondary research revealed that although LLMs could implement data integration and processing quicker information retrieval to assist DRM workflows, it is still facing the challenge of contextual specificity which prevents them fully replacing human expertise. However, the RAG technique is the most recommended method to compensate for the limitations of LLMs, followed by a comprehensive roadmap to it. Considering the primitive stage of LLMs application in the DRM sphere, this research contributes to the field with a practical guideline for professionals who seek assistance from LLMs in their work.

1. Introduction

This report presents the Capstone Project conducted by IE master's in international development student team in partnership with CIMA Research Foundation, a non-profit research organization specializing in disaster risk mitigation, exploring the limits and capabilities of Large Language Models (LLMs) in enhancing Disaster Risk Management (DRM) practices.

As natural and man-made disasters become more frequent and severe, disaster risk reduction (DRR) and management have gained ever more importance. This issue has drawn worldwide attention. In 2022, the United Nations launched the Early Warning for All Initiative (United Nations, 2022), calling for a global effort to ensure that everyone is protected on Earth by 2027. US$ 3.1 billion has been raised over five years to strengthen DRM, dissemination and communication of warnings, and preparedness and response capabilities (United Nations, 2022). Governments, civil society, and development partners from the public and private sectors are brought together by the Early Warnings for All initiative to improve collaboration and accelerate action to close gaps and provide people-centered, end-to-end multi-hazard early warning systems that leave no one behind.

Artificial Intelligence (AI) is emerging with great potential for innovative solutions in DRM. Leading and well-established AI tools demonstrate impactful implementations of such prediction and mitigation of disasters using predictive analytics, resource allocation solutions for early warning systems, and beyond. For instance, UN Global Pulse has been working with innovative solutions based on data and digital methods. In 2020, an image analysis tool was launched for satellite images of natural disasters with human-AI feedback loops (UN Global Pulse, 2020).

However, generative AI (GenAI), especially Large Language Models (LLMs) is still in the early exploratory phase for DRM. The World Meteorological Organization (WMO), has published a web application that aims to demonstrate features and potential capabilities of the Retrieval-Augmented Generation (RAG) technique in the field of DRM (WMO, 2024). This is a pilot attempt at utilizing LLMs for the meteorological domain, while broader DRM-related applications of the methodologies still require further development and testing.

Utilizing a variety of LLM tools, with a focus on ChatGPT-4, the IE Master of International Development students tasked with this capstone project worked to fill this gap by investigating the capabilities and limits of LLM contributions to disaster risk assessments, exercise design and implementation, and early warning action. By concentrating on these areas, this capstone endeavors to examine how LLMs might be used to support DRM workflows that facilitate efficient early response, risk management, and situational awareness, hence improving first responders' and other key stakeholders' preparedness.

1.1. Project Description

Initial meetings with CIMA professionals and academic advisors helped define the project scope and research questions based on CIMA's needs for a comprehensive overview and framework of LLM implementation in their organization.

The capstone project was guided by the following main research question:

- To what extent can LLMs contribute to disaster management?

Research sub questions were formulated for defining the scope in more detail, as follows:

- At what capacity can LLMs contribute to risk assessments?
- At what capacity can LLMs contribute to exercise design and implementation?
- At what capacity can LLMs contribute to early warning and early action strategies?

Based on the research questions, the project aimed to fill the critical gap for integrating LLMs in DRM through:

- **Evaluation of the capabilities and limitations of LLMs:**

  In order to determine the extent to which LLMs can support DRM operations, the project involved generating, as comprehensive risk information assessment, early warning messages that are easy to comprehend, as well as full tabletop exercises.

- **Development of practical guidelines and recommendations:**

  The client will be provided with detailed and practical guidelines that are directed at enabling them to fully fit LLMs in their daily operations, ensuring that they can optimally harness the benefits of AI technology while minimizing risks. Based on the project's empirical research

findings, the recommendations developed are specific and made to be practical and operational.

- **Creation of a comprehensive scenario framework and roadmap for supporting future initiatives in this domain:**

  Results from this study informed the creation of a scenario framework and roadmap for the future incorporation of LLMs in DRM. The roadmap explains what must be done to successfully use LLM technology and associated technical needs, training needs, and potential challenges. This guidance empowers CIMA to identify areas of their work worth exploring further in terms of potential LLM integrations and applications.

To accomplish these objectives, the IE student project team has conducted thorough reviews of literature and existing research, and tested capabilities of ChatGPT-4 and other LLMs using DRM task-related prompts. The project team has developed practical guidelines and recommendations manifested in a handbook as deliverable to the client[1]. The project outcomes have been then used to develop a practical tool, leveraging the Custom GPT functionality of ChatGPT.

Problem Analysis

This chapter starts with the theoretical foundations of DRM, AI, and LLMs. Following this, it explores the current challenges and gaps in applying LLMs to DRM, including issues of data availability, context-specific knowledge, bias and hallucination, and privacy concerns. AI-based techniques are introduced as potential solutions, and the project's anticipated impact and significance are discussed at last.

2.1 Theoretical Foundations

This section draws upon the literature review conducted in March 2024 by the student capstone team (Appendix I).

[1] The deliverable changed slightly over the course of the project (with the consent of our advisor and client) as the research focus shifted.

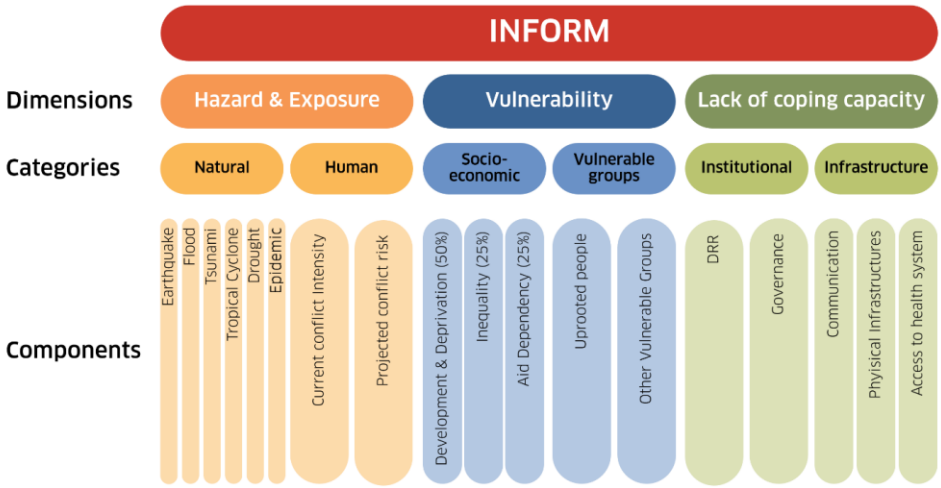
2.1.1 Disaster Risk Management

Disaster Risk Management is a multidisciplinary field that concerns the reduction of harm to life, property, and the environment (Coppola, 2015). It involves organizing and planning measures to apply in preparing, responding, and recovering from disasters (UNDRR, 2016). As a commonly used approach to manage and reduce disaster risk, the disaster cycle includes four parts: mitigation, preparation/preparedness, response, and recovery (Golding, 2022).

The INFORM Risk Framework is designed with three dimensions: hazard and exposure, vulnerability, and lack of coping capacity. The framework categories and components include natural and human-induced hazards, socio-economic factors, vulnerable groups, and institutional and infrastructural capacities. This analytical structure allows a systematic approach to assessing factors leading to overall risk and hence enhances the effectiveness of DRM and EWS.

EWS is an indispensable component of DRM, which plays a role in the prediction of the most likely disasters and allows people to take action before disaster strikes to minimize impacts (Quansah et al., 2010). Four important components are included in EWS: risk knowledge, monitoring and warning service, dissemination and communication, and response capability (UN/ISDR, 2006). The development of new technologies has improved EWS practices regarding data collection and analysis. However, in the aspects of real-time data collection, predicting short-term impacts, and computational accuracy there is still much space to improve (Agbehadji et al., 2023).

Figure 1: INFORM Risk (DRMKC, 2017)



2.1.2 AI and LLMS

AI refers to machine capabilities that mimic human intelligence in activities such as natural language understanding, pattern recognition, and decision-making (Banh & Strobel, 2023). LLMs have meanwhile risen to be a significant factor of influence in generative AI, especially due to their capabilities in Natural Language Processing (NLP). LLMs use Transformer architectures, a kind of neural network used in processing and generating sequential data (Ooi et al., 2023). Although highly capable, LLMs suffer from biases, hallucinations, and lack of contextual specificity, particularly for domain-specific or rare information (Kandpal et al., 2023). An LLM utilized widely is OpenAI's ChatGPT, which stands for Generative Pretrained Transformer, describing the model utilized. Recent approaches to increase LLM performance include fine-tuning and Retrieval-Augmented Generation (RAG) which will be explained in detail in the following sections.

2.1.3 Application of AI LLMs in DRM

The integration of AI like LLMs in DRM has received impetus like never before since the dawn of the new millennium. While most traditional AI applications in DRM are centered around prediction, early detection, and real-time situational awareness (Abid et al., 2021), LLMs have a unique ability for tasks related to natural language processing and generation. Some of the latest studies report LLM applications within the disaster response domain. Goecks and Waytowich (2023) demonstrated that LLMs can generate many specific and detailed action plans for disaster response very rapidly. Colverd et al. (2023) built "Flood Brain," an LLM application to develop reports on floods. Chandra and Chakraborty (2023) evaluated the performance of LLMs in radiation emergency contexts. These examples clearly show how LLMs can add agility and effectiveness in text-based operations concerning disaster response and management. According to this literature, LLMs are especially useful when a response is already taking place, assisting in action plans, report summarization, chatbots, and translations.

However, the application of LLMs in the field of DRM is still in its early stages due to several limitations and challenges. The specific details will be discussed in the next section.

1.2. Current challenges and gaps

Data Availability and Context-Specific Knowledge

In the area of DRM, LLMs frequently face the problems of quantity and quality of available data. Among all natural disaster scenarios, some scenario data are scarce, unreliable, or inconsistent (Akter & Wamba, 2017). This long-tail knowledge (Kandpal et al., 2023) has brought challenges for training data for LLMs. Even if training data is available, LLMs knowledge will rely on “outdated information after training has concluded” (Vaghefi et al., 2023), which poses a challenge to using LLMs for DRM. Depending on the DRM task, timely and reliable data is required. The problem is made even worse by the fact that some countries have in place strong national networks of monitoring systems, while others depend on infrastructure provided by international organizations. This difference of availability makes it difficult to address disasters effectively, revealing a need for accurate and comprehensive data to facilitate adequate preparedness, risk analysis, and early warning.

Biases and Hallucinations

AI-based systems are susceptible to biases and hallucinations in their outputs. This can be due to the nature of training data or if the algorithms are programmed to perpetuate some kinds of biases (Velev & Zlateva, 2023). These errors can have far-reaching consequences such as wrong predictions or failed decisions given the context of DRM. It is crucial that such biases are addressed and AI-generated outputs verified for accuracy.

Privacy and Security

The application of AI technology in managing disasters faces significant challenges related to sensitivities and privacy. The processing of huge amounts of personal or sensitive data raises privacy issues as well as security ones (Douglass et al., 2023). Thus, organizations must ensure compliance with laws and regulations while safeguarding individual rights (Gerunov, 2023). These challenges need addressing in order to fully exploit AI’s potential in mitigating human life and property impacts from natural disasters.

1.3. Proposed Solution: Potential AI-based techniques

The project explored several promising LLM-based techniques for identifying the capabilities and limits of LLM outputs in DRM and EWS. Techniques include prompt engineering, customization, and RAG.

Prompt engineering approaches involved zero-shot learning, and a combination of one-shot learning with In-Context Learning (ICL) to quickly allow an LLM to adapt to a specific disaster scenario. Zero-shot learning uses pre-trained language models without fine-tuning or additional training (Radford et al., 2019), and one-shot learning consists of using a single example or prompt for conditioning the model output (Brown et al., 2020). ICL is an extension of one-shot learning, whereby in the input prompt, accompanying instructions and examples are given so that the model is helped in output generation (Brown et al., 2020; Min et al., 2022).

The fine-tuning process involves training a pre-trained language model using a dataset relevant to the task (Devlin et al., 2019; Howard & Ruder, 2018). DRR initiatives also can benefit from fine-tuned AI models that are based on past disasters, vulnerability assessments, and best practices as these models can give more specific recommendations about how to prepare and respond to disaster (Ghaffarian et al., 2023). A similar, but notably more lightweight approach than fine-tuning is customization. The "Customize GPT" feature in ChatGPT exemplifies a method to serve domain-specific knowledge and frameworks into the model without changing the internal parameters.

RAG is another method that leverages contextually appropriate external sources to improve the relevance and contextuality of a model's output (Guu et al., 2020; Lewis et al., 2021). It is well recognized that RAG boosts LLM performances through integrating external data retrieval, leading to improvements in factual accuracy, coherence and also consistency (Gao et al., 2024).

The project focused on the feasibility and effectiveness of these LLM-based approaches in the domain of DRM. It explored how those techniques could enable LLM models to provide more accurate, context-aware, and actionable insight into support of the decision-making processes in DRM. In addition, the project includes an analysis on how to effectively combine several AI approaches, for example, customization and RAG, to deliver robust and more effective solutions in DRM.

1.4. Project impact and significance

By analyzing techniques and developing use scenarios to enhance the applicability and reliability of LLMs in the DRM field, stakeholders can have increased confidence in their understanding of potential use cases of LLMs in their organization. The project outcomes clearly outline leading techniques and ideal use cases that consider CIMA's existing DRM framework in terms of data availability, current LLM applications or explorations, and future goals and interests in LLM integration. This project aims to provide clarity regarding next steps in exploration of LLM use for CIMA.

2. Empirical Methods Used and Details on Deliverables

2.1. Multimodal research approach

In order to understand the capacity of LLM's potential applications to DRM workflows, the project team developed a robust experimental and analytical framework. The framework involved initial literature review and stakeholder meetings, along with qualitative and quantitative primary data collection from prompting experiments and stakeholder feedback on the experimental results. Findings from this primary research prompted a secondary research stage of additional literature review and stakeholder interviews.

Details on deliverables

Initial identification of deliverables to address client needs informed the design of the experimental and analytical framework. *LLMs for Disaster Risk Management: a handbook for CIMA* (Appendix II), is a decision-making guide for integration of LLMs to CIMA workflows and recommendations for future applications based on the main findings of the research. The handbook is supplemented with an interactive PowerBI dashboard of study results, alongside ChatGPT-based tool (referred to as LLM4DRM Advisor moving forward) that is customized with the research results and deliverables. LLM4DRM (Appendix XIV) Advisor is built to retrieve information from this study to assist in identifying specific use-cases of LLMs in DRM contexts, as detailed in the product vision board below.

Table 1: Product vision board LLM4DRM Advisor

Vision: Empower DRM organizations and professionals in leveraging LLMs for efficient preparation for, response to, and recovery from disasters.			
Target Group	**Needs**	**Product**	**Business Goals**
Disaster management professionals working on text-based workflows (eg. risk reports, early warning content, exercise design), including risk analysts, emergency response coordinators, early warning system designers, and content creators	**Understand Capabilities and Limitations:**Easily understand the current capabilities and limitations of LLM tools in the context of specific workflow. **Implement LLM Application Strategies**:Implement effective strategies at varying levels of resource and data availability to achieve the highest possible quality in LLM outputs. **Access to Quality LLM Outputs**:Achieve reliable and high-quality outputs from LLMs, ensuring Completeness, usability, accuracy, relevance, specificity and coherence of generated content.	A custom GPT that acts as an assistant in retrieving research-based information regarding specific use-cases of LLMs in disaster management work. Helps with providing: 1. **Comprehensive overviews of the LLM space** 2. **Recommending best practices**, based on the organization's resource/data availability, to generate the highest quality LLM outputs 3. **Evaluates the quality** of disaster management related LLM prompts and outputs	**Efficiency**: Save disaster management professionals time and energy in decision making and content development. **Focus**: Freeing up their time for the more critical aspects of their role. **Innovation**: Enable CIMA to be an early-adopter of emerging technologies for disaster management.

2.2. Experimental design

Development of the experimental framework involved identifying regions, disaster types, DRM tasks, and LLMs of interest, based on findings from literature review and conversations with stakeholders. Three different countries were selected for analysis in the experimental framework: Spain, the Philippines, and Mozambique. The selected countries represent not only diversity in geography but also varying levels of data availability regarding natural disasters. The geographical spread allows examination of how regional differences impact experimental outcomes. All three countries rely on a mix of national institutions and international bodies for the collection, management, and reporting of disaster-related data; however, Mozambique's data availability is far less robust than that of Spain or the Philippines. Both Spain (AEMET, 2024) and the Philippines (PAGASA, 2024) have established national infrastructure to report detailed data on all natural disasters that occur in their territory. Conversely, Mozambique's national infrastructure has less capacity, and instead the nation relies

more on international organizations (Hersher, 2019), such as the UNDRR and the World Bank's Climate Change Knowledge Portal, for the collection and management of disaster data. This results in a significant data gap regarding the granularity and timeliness of available disaster data between Mozambique and the other two selected regions for experimentation. Considering LLMs only have access to the data that was available at the time of their training period, selecting a country with limited data availability is a method to analyze the effect of varying levels of data availability on experimental outcomes.

After understanding the differences between the three regions in both geography and data availability, similarities in natural disaster occurrences were identified. All three nations have historically experienced wildfires, floods, and earthquakes. In order to analyze experimental outcomes between different types of disasters, the experiment scope includes trials for all three disaster types in the framework.

Since LLMs are inherently suited for text-based applications, the experimental design focused on text-based tasks in DRM. Therefore, in alignment with the research sub questions, the experimental framework was designed to test LLMs' capabilities to generate full risk reports, detailed tabletop exercises for response professionals, and clear and specific early warning messages for all three selected disaster types within the selected regions.

Table 2: Matrix detailing the variable combinations of region, disaster type, and aspect of disaster management for experiments

Region/Disaster	Wildfires	Earthquakes	Floods
Spain	1.Risk report 2.Exercise design 3.Early warning	4.Risk report 5.Exercise design 6.Early warning	7.Risk report 8.Exercise design 9.Early warning
Philippines	10.Risk report 11.Exercise design 12.Early warning	13.Risk report 14.Exercise design 15.Early warning	16.Risk report 17.Exercise design 18.Early warning
Mozambique	19.Risk report 20.Exercise design 21.Early warning	22.Risk report 23.Exercise design 24.Early warning	25.Risk report 26.Exercise design 27.Early warning

These 27 combinations of region, disaster type, and DRM tasks were then designed into prompts using an iterative prompt engineering process in a Custom GPT (called DRM Advisor moving forward) customized as an expert in disaster management. Data uploaded to the DRM Advisor Knowledge includes comprehensive frameworks, methodologies, and templates for disaster risk management. Uploaded knowledge also includes detailed reports on multi-hazard scenarios, public health event management, and early warning systems. The prompt for DRM Advisor is in the Appendix (Appendix III).

The prompt engineering process focused on building zero-shot and ICL one-shot prompts that generate complete and acceptable results (prompts available in project folder) The language of the 27 prompts was kept as similar as possible for each DRM task, simply substituting the disaster type and region.

The review of LLM tools and techniques resulted in the identification of four additional tools of interest: Mistral, Claude 3, Gemini, and MetaAI. All tools offer a text-based LLM interface, but each tool has their own number of tokens and other unique aspects (Appendix IV).

The experimental framework was designed based on 27 prompts, 6 LLM tools, and 4 LLM techniques (zero shot prompting, one-shot with in-context learning, customization, and RAG). (Appendix V).

General hypothesis and assumptions

Before conducting the experiments and analysis, the student capstone team made the following general hypothesis, based on the knowledge foundation from the literature review and stakeholder engagement. The main assumption considered in these hypotheses is that data availability and quality used by the LLM will influence the quality of the output.

1. 1-shot ICL prompts will score higher than zero-shot prompts.
2. DRM Advisor outputs will score higher than standard ChatGPT-4 outputs.
3. Outputs from the Manual RAG experiments will score the highest overall.
4. LLM outputs regarding Spain and the Philippines will score higher than outputs regarding Mozambique.

No hypotheses were associated with the disaster type, DRM task, LLM tool, instead opting for open exploration of these variables' effect on LLM outputs.

3. Data Collected and analyzed

3.1. Primary data collection

The 27 prompts were systematically fed into ChatGPT-4 to gather initial outputs as a baseline. Since LLMs learn from their previous conversations, the prompts were separated into different chats based on region and disaster combinations (e.g. floods in the Philippines). Within a single chat regarding a certain disaster and region, ChatGPT-4 was prompted for an early warning message, risk report, and tabletop exercise design. The corresponding prompts and outputs were recorded in an excel spreadsheet. This experimental procedure was repeated in the DRM Advisor and the 4 additional LLM tools.

In the DRM Advisor, selected prompts went through a manual RAG experiment to simulate outputs from a RAG. A manual search for the three different disaster types in the three identified regions 'retrieved' data in the form of news articles and incident reports from National and Intergovernmental Organizations (Appendix VI). Each piece of 'retrieved' data was uploaded to the established DRM Advisor experiment chat corresponding to the disaster and region described. The DRM Advisor was then prompted to rewrite the three disaster management outputs, pulling information from the new data.

The extensive framework of variables and experiments resulted in a database of LLM outputs for comparative analysis. The second phase of data collection involved scoring the experimental prompts and responses along the criteria explained in the following tables. The criteria were developed based on the findings of the literature review and evaluation of stakeholder needs.

Table 3: Criteria for scoring of experiment prompts

Contextual completeness	Does the prompt include all of the necessary context for the LLM to produce the best possible response?
Clarity and readability	What grade level can read and understand the response?
Relevance	Does the prompt accurately outline a task done by disaster management professionals?
Specificity	Does the prompt have specific details regarding location, audience, timeframe, etc.?
Conciseness	Is the prompt as brief as possible without losing crucial information?
Number of specific instructions	Does the prompt have clear instructions for the LLM to follow? (RICCE)

Table 4: Criteria for scoring of experiment responses

Completeness	Does the LLM response address all aspects of the prompt?
Usability	Can a disaster management professional copy and paste the response for immediate use without any further modifications?
Accuracy	Is the content of the response accurate to the locality, disaster type, and task at hand?
Relevance	Is the response discussing the same subject matter as the prompt?
Specificity	Does the response include details that are hyper specific to the region, situation, and/or audience?
Coherence	Is the response easy to read and understand?

While these criteria can be applied to LLMs at large, this study prioritizes the results from certain criteria categories in alignment with the study subject matter. Prompting is not the main subject of analysis for this study since the best practices and recommendations for LLM prompting are well established (Liu et al., 2021). In responses, it is expected that LLMs can produce acceptable outputs along the criteria of completeness, coherence, and relevance to the prompt. In the context of DRM and the variables tested in this study, changes in *Specificity*, *Accuracy*, and *Usability* are the focus.

A selection of 13 of the 27 prompts (See project folder.) and their corresponding responses across different experiments were scored along these criteria by a group of stakeholders including DRM professionals and the student capstone team. The 13 prompts were selected to represent all variables and experiments conducted without putting an unrealistic workload on the scorers. To collect quantitative data on these qualitative criteria, scorers were asked to give each prompt and response a score from 1-5 for each criterion, with 1 being the worst and 5 the best.

The DRM Advisor was prompted to provide its own scoring across these criteria. The GPT initially scored everything quite high, but with additional guidance provided, such as the content was to be scored in comparative context with each other, improved the scoring outputs.

Several rounds of scoring were conducted within the DRM Advisor on the prompts and responses provided for the human scorers. Further, prompts not included in the human scoring rubric were conducted with the DRM Advisor, such as the prompts and responses of different LLM tools in comparison.

To analyze the differences in criteria scores across the variables of LLM tool, prompt technique, region, disaster, and DRM task, a series of statistical methods were utilized, as well as general trend observations. Given the small sample size, non-parametric tests were employed to account for the data's non-normal distribution. The Kruskal-Wallis (KW) test was used to assess whether there are statistically significant differences in the median scores for each response criteria across the different groups of each independent variable (DATAtab, 2022). Following the KW test, Dunn's test was conducted for pairwise comparisons (Dinno, 2015). Visualization of the criteria scoring across the independent variables were utilized to observe trends that may not be statistically significant in the KW and Dunn's tests.

3.2. Secondary research and data collection

Results of the score analysis, detailed in the next section, revealed several insights and potential ideas for recommendations that prompted a second round of qualitative research and data collection. This secondary phase included an updated literature review on techniques that scored well in the experiments, namely RAG. Further informal expert interviews in the context of RAG were conducted. Experts included

- **Darline Giraud:** Executive MID candidate at IE University working on a RAG pipeline for nuclear emergency preparedness.
- **Saeid A. Vaghefi:** UN professional who is involved in the ASKWMO project, which is currently the RAG prototype for the WMO.
- **Balder Hageraats:** IE University Adjunct Professor and Senior Partner of ReSeT consultancy with an extensive professional background in risk analysis.

4. Results

4.1. Primary research

4.1.1 Prompts

Across all criteria 1-shot ICL prompts scored higher than zero-shot prompts, in alignment with the initial hypothesis. (Appendix VII)

4.1.2 Response Criteria

At large, the criteria of *Relevance, Coherence,* and *Completeness* received the highest scores, as expected for LLMs. *Usability, Accuracy,* and *Specificity* received the lowest scores overall. Moderate consensus among scorers (exemplified through Standard Deviation - within Criteria) is consistent across criteria (exemplified through Standard Deviation - across criteria). The variance of the average score across criteria is moderately low exemplified by the standard deviation + 0.33. None of the six criteria scored a perfect 5 out of 5.

Table 5: Average criteria scores

Response Criteria	Average Score	Standard Deviation - Within Criteria
Relevance*	4.37	0.67
Coherence*	4.33	0.55
Completeness *	3.98	0.57
Usability	3.85	0.54
Accuracy	3.77	0.49
Specificity	3.51	0.57
Standard deviaton - Across criteria	0.33	0.06

DRM Advisor scoring vs. human scoring of response criteria

Depending on criteria, the DRM Advisor scored responses 28-44% higher than the human scorers.

Table 6: Comparison of average scores for response criteria between humans and DRM Advisor

Response Criteria	Average Human Score	Average DRM AdvisorGPT Score	% Difference
Relevance*	3.46	5.00	44
Accuracy	3.35	4.70	40
Specificity	3.11	4.35	40
Usability	3.30	4.35	32
Completeness*	3.51	4.57	30
Coherence*	3.84	4.91	28

Kruskal-Wallis Test

The full KW test results provide a KW-statistic value and p-value for each dependent variable (criteria) and independent variable (Tool, Region, Disaster, Task, LLM Technique) pair (Appendix VIII). KW-statistic value indicates the difference in median criteria scores within each variable group. Higher KW-statistics mean at least one variable within the group scored significantly different than the rest. The highest and most statistically significant KW-statistic is *Relevance* scores among different Tools.

The KW results of highest statistical significance involve Tool variable's effects on *Coherence, Relevance, Usability,* and *Specificity* scores. Slightly less significant results include Task variables' effect on *Specificity* and Region variables' effect on *Completeness* scores

Table 7: Statistically significant Kruskal-Wallis Test Results

Criteria	Independent Variable	KW Statistic	p-value
Coherence*	Tool	26.41	0.00
Relevance*	Tool	28.65	0.00
Usability	Tool	12.11	0.03
Specificity	Tool	20.90	0.00
	Task	6.98	0.03
Completeness*	Region	6.66	0.04

Dunn's Test

Out of all Dunn's test results (Appendix VIIII), ten pairs exhibited statistical significance. ChatGPT-4 pairs with DRM Advisor, Mistral, Claude 3, and Gemini all exhibited significant differences in *Coherence* scores. For *Relevance* scores, DRM Advisor pairs with ChatGPT-4 and Gemini exhibited significant differences. ChatGPT4 and Claude 3 show significant differences in *Specificity* scores, as well as the Task variable pair of Early Warning and Risk Report. *Completeness* scores exhibited significant differences between Mozambique and Spain.

Table 8: Excerpt of Dunn's test results for statistically significant pairs

Criteria	Independent Variable	Group 1	Group 2	p-value
Coherence	Tool	ChatGPT4	Claude 3	0.01
		ChatGPT4	DRM AdvisorGPT	0.00
		ChatGPT4	Gemini	0.01
		ChatGPT4	Mistral	0.00
Relevance	Tool	ChatGPT4	DRM AdvisorGPT	0.00
		DRM AdvisorGPT	Gemini	0.02
Specificity	Tool	ChatGPT4	Claude 3	0.01
		ChatGPT4	Mistral	0.01
	Task	EW	Risk Report	0.04
Completeness	Region	Mozambique	Spain	0.03

Visualizations

Response data was analyzed visually in PowerBI to understand the relationship between the statistically significant pairs in the KW and Dunn's test results, as well as identify other trends that allow comparison of outcomes versus hypothesis, shape recommendations, and identify areas for further studies.

Figure 2: Combined visualizations of criteria scores itemized by region, task, prompt type, and disaster

2.

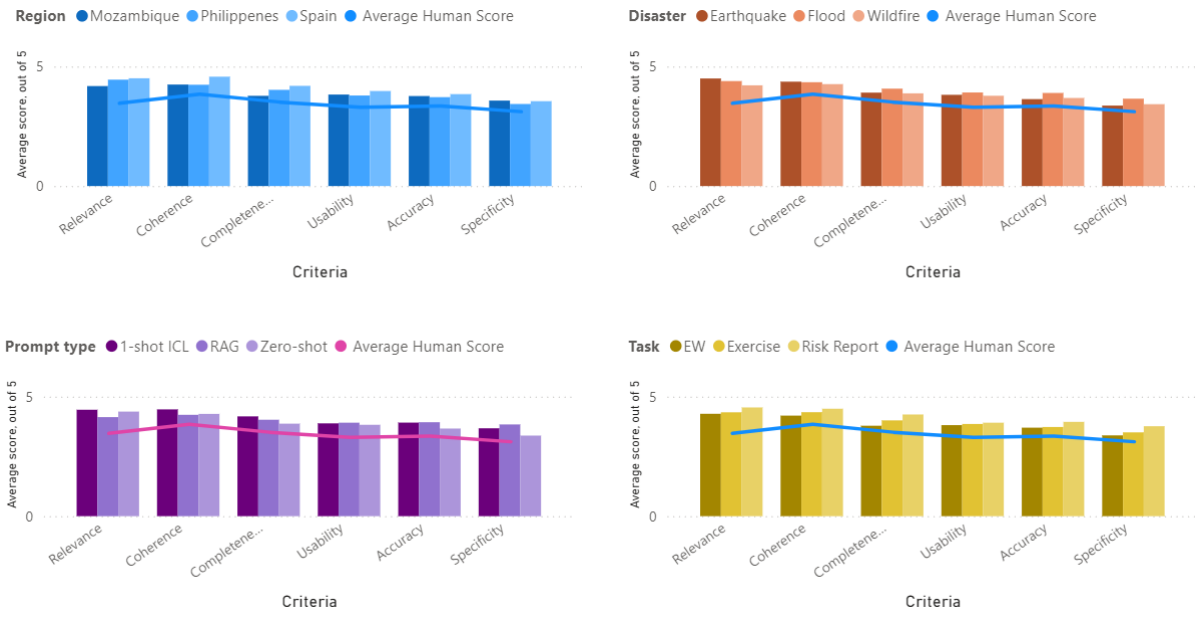


Figure 3: Combined visualizations of criteria scores itemized by region, task, prompt type, and disaster. Average of human scores only

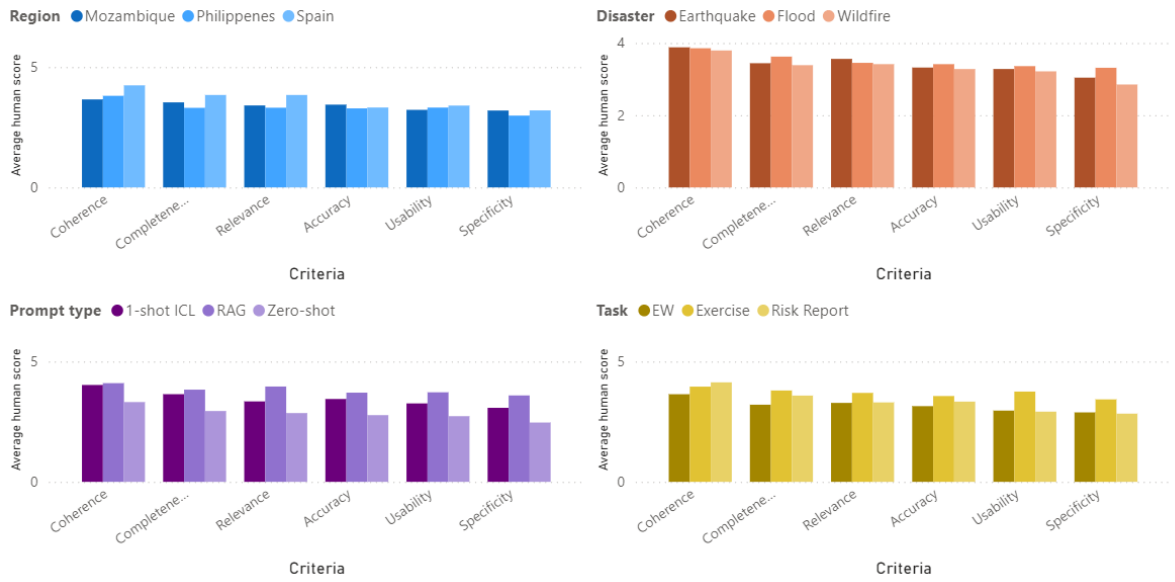
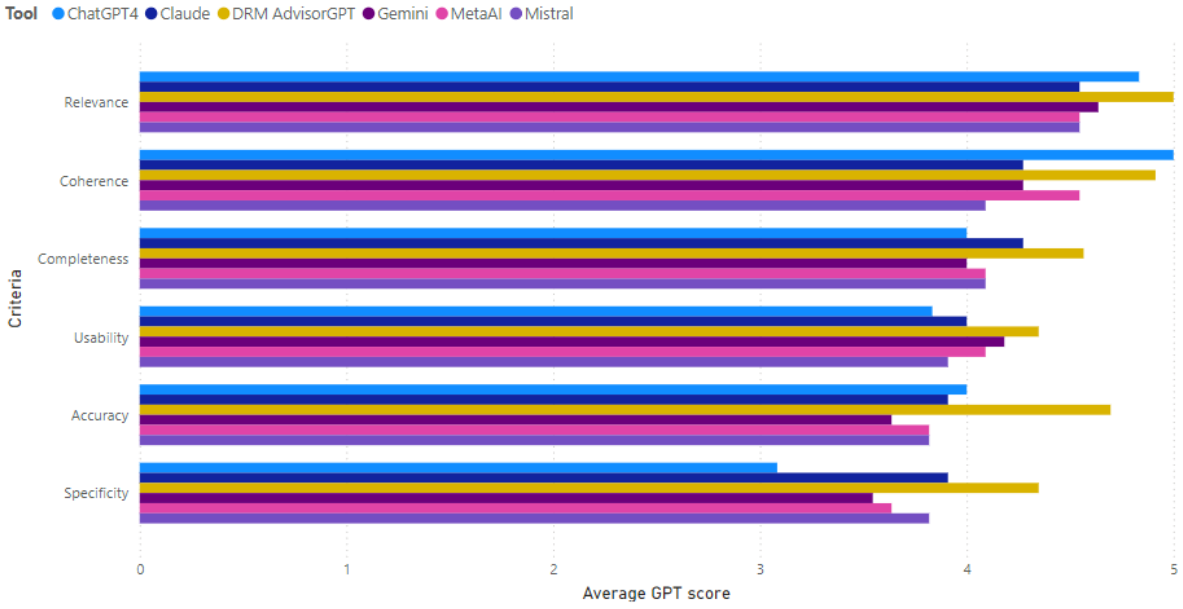


Figure 4: Response criteria scores by LLM tool



4.2. RAG Research

As a final research step for the capstone project, an analysis of the market and best practices in developing a RAG pipeline was conducted to better shape recommendations. Underpinning the results from the previous analysis on LLM prompting and output, this analysis is taking the research to the next step, bridging the gap to get to the best usage of LLMs for DRM. Based on the different

results, the most promising outputs were generated by the manual RAG approach. Combined with the current literature explored in section 2.X AI and LLMs, the objective of this analysis is to identify actionable insight and give a guideline to the development of an own RAG pipeline. Outlining the technology needed and concrete actions to take, a base for a detailed roadmap will be given.
For this analysis, external secondary data was obtained, accessing existing data and literature. Data was collected from a mixed variety of sources, ranging from academic literature to web articles, blog posts and videos. Especially as the space of AI and LLMs is highly dynamic and fast paced, the use of diverse sources allows for the results to be the most relevant and useful for the capstone project outcomes.

4.2.1. RAG Components, Architecture and Process

As touched upon in section *2.2. Proposed Solution: Potential AI-based Techniques,* RAG is a technique to improve LLM outputs and accuracy, using external data (Google for Developers, 2024; Lewis et al., 2021). Without further training or fine-tuning of the LLM model at hand, RAG allows to increase the accuracy of the LLM output while mitigating some of the limitations of LLMs today (Gao et al., 2024; Google for Developers, 2024).

As the name indicates, RAG systems combine data retrieval, augmentation and response generation. In addition to a traditional pre-trained LLM that harnesses a parametric approach, a “non-parametric memory retrieval [is] enabling the generation of text conditioned on both the input prompt and external knowledge sources” (Abdullahi, 2024). The general RAG architecture consists of three sections - indexing, retrieval and the response generation - covering the three RAG workflows that form the foundation of the RAG process (Demir, 2024; Gao et al., 2024; Monigatti, 2024).

The indexing is the primary step in the process and takes place for the data preparation before the actual user query of the system. First, information is being processed to extract the raw data from data sources (Bainiaksina, 2024). Data can be retrieved from multimodal sources, ranging from PDF and text documents to csv and json files. Additionally, information from images and videos can be extracted as well (Google for Developers, 2024). The data should be relevant for the purpose of the RAG system. After loading the collected data, large documents are split into chunks of smaller sizes.

The size can have an influence on the performance of the RAG system. Therefore, it is important to decide on a suitable indexing strategy, splitting the chunks by a fixed size, semantics or using a hybrid approach (Rahul, 2024). Effective chunking overlaps can bridge between chunks for contextual semantics (Aerospike, 2024). After splitting, the data is transformed and embedded into "multidimensional numeric vectors, capturing their meaning and nuance in a vector space" (NVIDIA, 2024) using an embedding model. Vectors are "multidimensional numeric These vectors are then stored in a vector database, with related vectors stored close to one another, allowing for easy search and access (Bainiaksina, 2024; NVIDIA, 2024), and data retrieval for the inference with the LLM.
Once the user queries the GenAI, the prompt is transformed by the embedding model, so that matching data for the prompt can be retrieved from the vector database. For the retrieval the system matches the prompt vector representation and with the external data vector representations stored in the vector database. Based on similarity to the prompt, the RAG "system prioritizes and retrieves the top K [highest ranked] chunks" (Gao et al., 2024, p. 4). Further, to increase the relevance of chunks selected for enhancing the LLM output, a post retrieval, re-ranking can assist the LLM in prioritization of information for output generation (Colesanti Senni et al., 2024; Deshwal, 2023).

Finally, in the response generation step, the retrieved data is used for expanding the contextual information on the user query (Gao et al., 2024). The original prompt and the retrieved data are then synthesized into one new prompt, which will be used by the LLM to generate an answer. Depending on the actual ask in the query, more or less of the "inherent parametric knowledge or […] the information contained within the provided documents" (Gao et al., 2024, p. 4) is used.

Details in the RAG pipeline architecture depend on the system purpose, which should be defined at the project's beginning. Besides the basic architecture explained, advanced and modular RAG versions complete the RAG paradigm, allowing for mitigating potential challenges in the basic model and "providing greater versatility and flexibility" (Gao et al., 2024, p. 5).

Insights from expert interviews

The informal interviews with three stakeholders have revealed further insights and considerations.

The first interview with Darline Giraud revealed insights from building a RAG pipeline, which is part of her thesis. To refine and improve the performance of the RAG system, the importance of stakeholder engagement was emphasized, determining importance and priorities (Informal Stakeholder Interview 1, Appendix X). The approach for building a RAG pipeline was chosen, trying to keep as many existing assets as possible, to accommodate the common constraints of timing and resources in the field of disaster management. While data is available, existing LLMs are reused, and improvements fully made by prompting. To achieve successful outcomes, the database is being built as a scenario database. Another factor that became evident was the challenge of relying fully on LLMs for DRM tasks, as accuracy is not reliable at this stage of the technology. Guard-rails and governance were proposed as a possible solution to the situation. Finally, a hypothesis of the significance of what the retrieval data consists of and how the retrieval data is organized compared to data volume was discussed.

A second interview with Saeid A. Vaghefi, disclosed insights from a project of the WMO, the askWMO, which is currently in the stage of a prototype and is not a product yet. Specifically for the data preparation phase of the RAG data, the possible use of additional metadata for chunking was considered. Another way to influence the RAG system is to use re-ranking techniques, forcing “the model to reach to a broader space in the vectors” (Informal Stakeholder Interview 2, Appendix XI). Both approaches can help to prevent data being overseen for retrieval. Using existing LLM endpoints such as the OpenAI API or Google endpoints for Gemini 1.5 pro, can help in the beginning to get the system working quickly. However, from a long-term perspective, the development of a local LLM could help mitigate risks due to dependencies on external sources. Another important factor we discussed is the capabilities of LLMs when it comes to regions with underserved communities and local languages or dialects. Training data and additional external data need to be inclusive and capable for these communities

The final interview with Balder Haagerats provided an alternative perspective. As a risk professional and self-proclaimed non-user of AI tools, his interview highlighted significant concerns and drawbacks to the application of LLMs in the workflows analyzed in this study (Informal Stakeholder Interview 3, Appendix XII). Balder and Darline shared the same concern regarding full reliance on

LLMs for DRM tasks. However, Balder argued that the time and resources needed to improve LLM outputs in this context would be better spent by human professionals doing the work themselves. He also provided examples of more appropriate use cases and scenarios for LLM use in his line of work, citing the experiences of his colleagues who utilize LLMs to retrieve specific information on a region of interest in a shorter amount of time. This approach utilizes LLMs as a search engine to supplement a completely human-generated deliverable.

4.3. Discussion of primary research results

While analysis through statistical methods was conducted to provide a level of robustness to the analytical framework, due to the small sample size of this study, statistical significance is not the only factor considered in evaluating the data. Visualization of the data shows that many of the statistically significant pairs from the KW and Dunn's tests exhibit minimal practical differences. For example, the visualization of average criteria scores by region show that Mozambique and Spain exhibit very similar average specificity scores, despite statistical significance. Therefore, findings will be discussed mainly at an observational level with reference to the statistical analysis results when appropriate.

At an observational level, response criteria results from human scoring produced higher variation than the average scores that included score input from the *DRM Advisor* GPT. The *DRM Advisor* GPT may not have the same critical lens for criteria evaluation as a human, but including the *DRM Advisor* GPT scores in the results dataset allows for more direct comparison across experiments, since the human scorers only received a selection of responses for scoring that was conducive to a realistic workload. Even with the *DRM Advisor* GPT score included in the averages, not a single response category scored full points. This suggests that even with the LLM techniques utilized, these tools are still not capable of producing DRM outputs of human-level quality. This finding is in alignment with the insights gleaned in the expert interviews.

The human-only criteria score averages are useful for identifying trends in LLM response performance across the different variables due to the higher level of criticality and context-specific

knowledge in human scoring. These results show RAG as the highest scoring LLM technique, followed by the 1-shot ICL prompts, and the zero-shot coming in last. This finding confirms the hypothesis stated at the beginning of the study, confirming that citing RAG as a best practice is applicable in a DRM context.

Across regions, average human scores for response criteria show Spain performing the highest in all criteria except for Specificity, in which Mozambique scored highest. This result is not in line with the original hypothesis that Mozambique would score the lowest amongst regions due to lack of robust meteorological data. One response (Response 21) scored particularly high in Specificity (4.75), which may influence this result. This response was an Exercise output regarding floods in Mozambique, and it was generated using the manual RAG approach after an initial 1-shot ICL prompt. This combination of two powerful prompting techniques yielded high results in the specificity criteria.

Responses regarding the Philippines were the lowest among the human-only scores, and on par with Mozambique when the *DRM Advisor* GPT scores were included. This is not aligned with the hypothesis regarding regions with high data availability yielding the strongest outputs, considering that the Philippines has a comparably robust meteorological data infrastructure to Spain. One possible explanation for this is that LLMs may be biased against non-Western regions if their training data reflects the historical tendency to skew towards the West.

Responses regarding floods scored the highest by human and AI scorers in the criteria of focus for this study (*Accuracy, Usability, Specificity*). This is followed by earthquakes and then wildfires in the human scores, while the *DRM Advisor* GPT scores scored wildfire responses the same or slightly higher as earthquake responses. This may suggest that flood data has a high level of availability and detail across regions, supporting LLM's ability to generate DRM tasks for this context.

Among human scorers, exercise design responses performed the best across the criteria of focus for this study. The highest performing responses change to risk reports when *DRM Advisor* GPT scores are included. This suggests a disparity in the *DRM Advisor* GPT's understanding of a high-quality risk report. Exercises designed for wildfire scenarios in Spain were the highest scoring

responses across criteria. Considering the persistent issue of wildfires in this region, LLMs may have access to more relevant data to generate scenarios for exercises. The KW test results show that DRM tasks do have an impact on Specificity scores at a statistically significant level, and Dunn's results highlight the early warning and risk report pair for this criterion. However, considering these tests were conducted on the average scores, these results do not yield practical findings due to the discrepancies between how humans scored the three DRM tasks versus how the LLM scored them. Early warning messages scored the lowest across all criteria with and without the inclusion of the *DRM Advisor* GPT scores, suggesting that LLMs are not equipped to produce adequate outputs for this task. One aspect of this limitation could be generation of outputs in languages other than English, especially if the region in the scenario speaks a dialect that may not have been well represented in the training data. Furthermore, to produce high quality early warning messages, a very high context specificity for the respective disaster and region must be available in real time. A standard LLM is not able to retrieve this kind of data.

Response criteria scoring between tools was conducted only by the *DRM Advisor* GPT as the volume of responses was too large for the human scorers to realistically confront in the study timeframe. However, this provides a direct comparative analysis of response quality across six LLM tools. It is important to acknowledge a potential bias in scoring since it was conducted in a ChatGPT interface. For example, ChatGPT-4 received a perfect average score for *Coherence*, likely because it considers its own language to be the most understandable. Additionally, *DRM Advisor* GPT scored itself much higher than the other tools for most criteria. With this bias in consideration, *DRM Advisor* GPT did score higher than its standard ChatGPT-4 counterpart across most criteria, in alignment with the original hypothesis. KW test results show that LLM tools have a statistically significant effect on *Relevance, Coherence, Usability, and Specificity*. Dunn's test revealed statistical significance in several pairings with ChatGPT-4 for coherence, which is explained by the perfect, potentially biased coherence score discussed previously. Other statistically significant pairs from the Dunn’s test do not align with notable observational findings except for the identification of Mistral and Claude 3 as high performers in Specificity, when disregarding the exceptionally high *DRM Advisor* GPT score. Compared to the other LLM tools analyzed, Mistral and Claude 3 are middle of the pack in terms of

parameter and tokenization (Appendix IV). This suggests that Claude 3 and Mistral may have access to certain training data that enabled them to produce the DRM tasks at a higher level of specificity compared to similar tools, however, the training data of these tools are not disclosed.

Limitations

It is important to note the qualitative and subjective nature of the criteria scoring process when interpreting these results. The sample size in this study is large enough to observe general trends regardless of subjectivity, but it is still rather small to produce meaningful results of statistical significance in common analysis methods. The experimental and analytical framework utilized in this study provides a foundation from which to scale the sample size across a larger number of responses to generate results with a higher level of statistical significance, which may be necessary in certain decision-making scenarios.

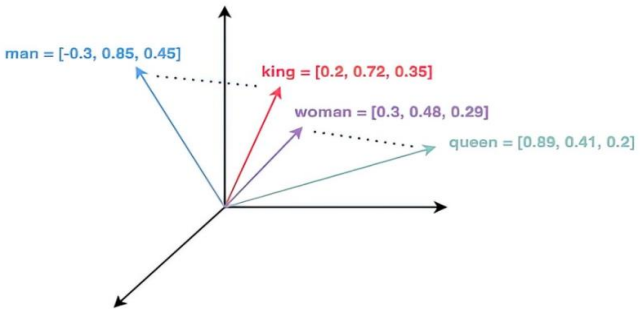
4.4. Discussion of RAG Research

The confirmation that a RAG approach is successful in a DRM context prompted further exploration of the topic to shape recommendations for the client. The analysis of the market and best practices in developing a RAG pipeline provided insights into the details of the architecture and components. For the specific use case of RAG systems in DRM, benefits and challenges need to be considered.

The RAG approach brings several benefits, leading to improved LLM outputs. As evident from the primary research, one of the main benefits of using RAG is the quality increase and improved contextualization of the LLM responses. Leveraging external data sources that were not previously included in the training data of the generative model, the original user query is enhanced by the retrieved data and thus the context for the generative LLM for the response becomes more accurate and relevant (Google for Developers, 2024; NVIDIA, 2024). The primary research has shown clearly the increased quality of LLM responses across all evaluations, using the manual RAG approach, indicating the improved contextualization. Further, the response could be enhanced by up-to-date or even real-time data, if available, opening the usage of LLMs for DRM workflows that require up-to-date or real-time data such as disaster and risk assessments, response coordination, early warning

messages and others (Confluent, 2024; Databricks, 2023). Further, with the usage of external data, not only the accuracy of responses can be increased but also the risk of hallucinations and response biases can be lowered (Abdullahi, 2024; NVIDIA, 2024). With additional factual data available, RAG is less likely to hallucinate or provide incorrect information(Gao et al., 2024; Lewis et al., 2021). This is specifically important for DRM-related work as livelihoods can be impacted. Even though this use of factual data for the generation of the LLM output reduces response biases, there is no full control over the generated output. While the model and system itself do play a role, the training data and retrieved data impact the output the most (Aerospike, 2024; Lewis et al., 2021). Overall, RAG systems are shown to be effective in reducing risk to privacy and data security. As the architecture shows external data is stored in a localized vector database, minimizing data exposure of LLM training data and increasing overall control over the content used (Zeng et al., 2024). Another factor to consider increasing data privacy and security is the development of a local LLM to be deployed for the RAG. If external endpoints such as the OpenAI API or Google APIs are used for accessing existing LLMs, dependencies of these external sources decrease control over LLM behavior and increase points of potential malicious attacks (Informal Stakeholder Interview 2, Appendix XI). Localized LLMs and RAG systems can mitigate these risks (Wolff, 2023).

Figure 5: Vectors capturing the meaning and nuance of information (NVIDIA, 2024)



However, several potential challenges must be considered when developing and using a RAG system. The outline of the RAG architecture has shown that a focus needs to be on the first phase, the indexing of data. For this, special attention needs to be paid to the data identification and data collection. As mentioned among the benefits, biases cannot be avoided completely (Lewis et al., 2021). This is especially critical in DRM, as decisions impact individuals and actions directly. Thus, the quality of the data collected and indexed for the RAG system becomes much more important

and relevant. Thorough data understanding and monitoring can help to ensure the relevance and correctness of the data stored for retrieval. Further, close stakeholder engagement is needed to align on the data that should be processed and used for retrieval (Informal Stakeholder Interview 1, Appendix X). Additional biases can be mitigated by diversifying the data sources used, incorporating different perspectives and data granularity. While building the RAG pipeline, not only the decisions on what data but also how much data is being stored for retrieval need to be made. With increased dataset volume, the RAG system needs to be adapted accordingly to allow for proper data handling. The stakeholder feedback revealed the prospective data volume is not always the main priority. In some cases, the quality and type and organization of data has a much more significant impact on the response quality (Informal Stakeholder Interview 1, Appendix X). With the increased amount of data, the performance of the RAG system might decline, due to high data processing volumes. The process of retrieving matching data “may introduce latency, especially when accessing large knowledge sources” (Abdullahi, 2024). Depending on the DRM workflow, this could influence the outcome of using RAG for the workflow. Though limited, the optimization of underlying model algorithms can help improve latency and efficiency in parts (Gao et al., 2024). Related to the importance of data quality, the maintenance of RAG systems is resource heavy. To avoid data biases. ensure consistency and relevance, not only the data but also the models need to be monitored and evaluated regularly. Depending on the purpose of the RAG system this M&E process is costly, influencing the cost-benefit of implementing RAG as an LLM solution.

5. Recommendations

This study revealed capabilities and shortcomings of LLMs in DRM along several variables of region, LLM tool, disaster types, and DRM task, shaping the final recommendations regarding the extent to which LLMs can be implemented in DRMs.

The first and most important recommendation is that LLMs should not replace humans in performing DRM tasks explored in this study. This technology is not at a level of advancement where the outputs are comparable to human quality.

With this said, there are also potential benefits of integrating LLMs into DRM workflows. Considering the amount of data at CIMA's disposal, CIMA can utilize LLMs for quicker retrieval of information to use in the deliverables that they develop themselves. Additionally, initial drafting for exercise scenarios is a potential application, but it is not recommended to draft risk reports or early warning messages in LLMs. Considering the scenarios in which CIMA mainly operates, such as hydrological and hydrogeological risks and wildland-urban interface fires. LLMs could be used to assist CIMA professionals in expedited retrieval of data for risk assessments or draft exercises for wildfire scenarios. With these capabilities and limitations in mind, if CIMA wishes to continue their exploration of LLM utilization, the findings in this study inform several recommendations on how to move forward. First, if additional data is not available for the scenario at hand, it is recommended to provide as much context as possible through ICL prompting in the tools that performed well in this study: Claude 3 and Mistral. These techniques will improve the output when retrieving information or drafting an exercise.

However, providing additional data is a confirmed best practice in LLM techniques. CIMA does wish to explore techniques for integrating organizational data into an LLM; they can do so by adopting a customization, Manual RAG, or RAG approach, depending on the capabilities of the LLM tool.

The results from this study inform the following recommendations regarding which scenarios would most require the supply of additional data to retrieve relevant information or generate acceptable exercise drafts, regardless of the underlying approach.

1. Scenarios in a region outside of the West, due to a likelihood of Eurocentric, English-language focused and online published training data.
2. Scenarios where a local dialect is spoken and the LLM may not have proper training in the dialect.
3. Scenarios regarding a disaster type that is less common, new to the region, or has overall less documentation.
4. Scenarios regarding a region with less robust data/infrastructure.

As climate change continues to shift the occurrences of certain disasters away from their historical geographies, CIMA could be confronted with these scenarios more often.

In terms of approach to providing additional data to LLMS, building a RAG can empower CIMA professionals in swift retrieval of organizational data instead of taking time to sift through the data themselves, making their workflows more efficient. However, RAG is extremely resource intensive. CIMA should perform a cost-benefit analysis regarding the development of an in-house RAG for considering the costs of development and the benefits of improved efficiency in data retrieval. Appendix XIII provides a scenario framework to assist in this cost benefit analysis.

If CIMA decides to move forward with building a RAG, the existing structured data at CIMA's disposal (e.g. hydrological data, exposure data, and Digital Elevation models) should be supplemented with non-structured data such as relevant reports from News outlets or IGOs that detail past disaster events with distinct references to meteorological data, geography, communities, casualties, and property damage Examples of high quality human-generated DRM tasks would also be excellent for the LLM to reference. To build out a larger framework for a RAG pipeline, it is recommended that CIMA utilize the RISK inform framework to find and categorize data relevant and necessary to help the RAG best understand the context. A scaled-down version of this framework could also be applied to the data utilized to customize a GPT. When gathering and organizing data, it is important to note that data quality and specificity to the use-scenarios takes precedence over data volume. With that said, RAG can also be especially helpful in assisting DRM professionals in confronting a large volume of data.

Independent from the approach chosen, the use of LLMs must be monitored closely. The criteria put forth in this study can be adopted and modified for monitoring of overall quality of LLM outputs. While RAG systems can improve data privacy and security, it is suggested to avoid dependencies on external endpoints in the long-run. Furthermore, if monitoring reveals that LLM outputs are not achieving minimal values and only increased risks, the general use of LLMs in DRM should be reconsidered.

6. Outlook and Conclusion

To answer the original research question: to what extent can LLMs contribute to disaster management regarding different aspects? Thorough analysis of the LLMs in the context of disaster management through extensive literature review, experimentation, and stakeholder analysis, this research concludes that overall the current capabilities of LLMs is they can assist DRM professionals in expedited retrieval from large datasets to improve the efficiency of their text-based workflows and deliverables, and perhaps drafting of initial documents for certain tasks. The main limitation of LLMs in a DRM context is that the widely available tools do not have access to the data necessary to provide human-like DRM outputs at a level of contextual specificity needed.

In the inherently time-sensitive field of DRM, the integration of LLMs has the potential to free up time for DRM professionals, allowing them to focus on more critical aspects of their role instead of tedious sifting through data sources. If a DRM organization wishes to implement LLMs in their workflows at the current state of technology, results from this research provide tangible actions to take at any level of resource and data availability. It is important to take into consideration that even with these techniques, LLM technology does not have the capability to produce outputs at the level of human quality when it comes to contextual specificity, especially in five scenarios that are mentioned above: regions outside of the West, local dialect widely used, less common disaster type, region with less robust data or need for comprehensive risk assessment and early warning message. However, the outlook for the LLMs space is promising in terms of fast-paced advancements.

Given the project's limited scope, suggestions for future research can help strengthen the findings of this study and shed light on aspects not in its focus. The limited resources of this capstone only allowed for a small sample size for the analysis of prompt and answer evaluation. A bigger sample size will allow more statistically significant findings, which may be crucial for decision making. Further the small sample size might have led to biased results, which can be avoided by increasing and diversifying the sample. To solidify the findings, the influence of prompting using domain expertise would allow us to explore the influence of prompting content. Overall, an ongoing dialogue between stakeholders in both the DRM and AI/LLM field will reveal the most appropriate use cases for LLMs in DRM as both fields advance over time.

7. Bibliography

Abdullahi, A. (2024). What is Retrieval Augmented Generation? How it Works & Use Cases. *eWeek*, N.PAG-N.PAG.

Abid, S. K., Sulaiman, N., Chan, S. W., Nazir, U., Abid, M., Han, H., Ariza-Montes, A., & Vega-Muñoz, A. (2021). Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management. *Sustainability*, *13*(22), Article 22. https://doi.org/10.3390/su132212560

AEMET. (2024). *AEMET Homepage*. https://www.aemet.es/en/portada

Aerospike (Director). (2024, June 19). *Building a RAG Application: The Good, the Bad, and the Ugly*. https://www.youtube.com/watch?v=uhb2cPtAR14

Agbehadji, I. E., Mabhaudhi, T., Botai, J., & Masinde, M. (2023). A Systematic Review of Existing Early Warning Systems' Challenges and Opportunities in Cloud Computing Early Warning Systems. *Climate*, *11*(9), Article 9. https://doi.org/10.3390/cli11090188

Akter, S., & Wamba, S. F. (2017). Big data and disaster management: A systematic review and agenda for future research | Annals of Operations Research. *Applications of OR in Disaster Relief Operations, Part II*, *283*(December 2019), pages 939-959. https://doi.org/10.1007/s10479-017-2584-2

Bainiaksina, D. J. (2024, March 10). How I built a Simple Retrieval-Augmented Generation (RAG) Pipeline? *Medium*. https://medium.com/@drjulija/what-is-retrieval-augmented-generation-rag-938e4f6e03d1

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, *33*(1), 63. https://doi.org/10.1007/s12525-023-00680-1

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Chandra, A., & Chakraborty, A. (2023). *Exploring the Role of Large Language Models in Radiation Emergency Response* (SSRN Scholarly Paper 4563109). https://doi.org/10.2139/ssrn.4563109

Colesanti Senni, C., Vaghefi, S., Schimanski, T., Manekar, T., & Leippold, M. (2024). *Using AI to Assess the Decision-Usefulness of Corporates' Nature-related Disclosures* (SSRN Scholarly Paper 4860331). https://papers.ssrn.com/abstract=4860331

Colverd, G., Darm, P., Silverberg, L., & Kasmanoff, N. (2023). *FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM* (arXiv:2311.02597). arXiv. https://doi.org/10.48550/arXiv.2311.02597

Confluent. (2024). *What is Retrieval Augmented Generation (RAG)?* Confluent. https://www.confluent.io/learn/retrieval-augmented-generation-rag/

Coppola, D. P. (2015). Chapter 1—The Management of Disasters. In D. P. Coppola (Ed.), *Introduction to International Disaster Management (Third Edition)* (pp. 1–39). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-801477-6.00001-0

Databricks. (2023, October 18). *Retrieval Augmented Generation*. Databricks. https://www.databricks.com/glossary/retrieval-augmented-generation-rag

DATAtab. (2022). *Kruskal-Wallis-Test*. https://datatab.net/tutorial/kruskal-wallis-test

Demir, N. (2024, May 16). *Advanced RAG: Implementing Advanced Techniques to Enhance Retrieval-Augmented Generation Systems*. Medium. https://blog.demir.io/advanced-rag-implementing-advanced-techniques-to-enhance-retrieval-augmented-generation-systems-0e07301e46f4

Deshwal, M. (2023, November 24). Simplest Method to improve RAG pipeline: Re-Ranking. *LanceDB*. https://medium.com/etoai/simplest-method-to-improve-rag-pipeline-re-ranking-cf6eaec6d544

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Dinno, A. (2015). Nonparametric Pairwise Multiple Comparisons in Independent Groups using Dunn's Test. *The Stata Journal*, *15*(1), 292–300. https://doi.org/10.1177/1536867X1501500117

Douglass, R., Gremban, K., Swami, A., & Gerali, S. (2023). *Iot For Defense and National Security Full Chapter | PDF | Internet Of Things | Computer Security*. Wiley-IEEE Press. https://www.scribd.com/document/725339274/Download-Iot-For-Defense-And-National-Security-Robert-Douglass-full-chapter

DRMKC. (2017). *INFORM Risk Methodology*. https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Risk/Methodology

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gerunov, A. (2023). *Risk Analysis for the Digital Age* (Vol. 219). Springer International Publishing. https://doi.org/10.1007/978-3-031-18100-9

Ghaffarian, S., Taghikhah, F. R., & Maier, H. R. (2023). Explainable artificial intelligence in disaster risk management: Achievements and prospective futures. *International Journal of Disaster Risk Reduction*, *98*, 104123. https://doi.org/10.1016/j.ijdrr.2023.104123

Goecks, V. G., & Waytowich, N. R. (2023). *DisasterResponseGPT: Large Language Models for Accelerated Plan of Action Development in Disaster Response Scenarios* (arXiv:2306.17271). arXiv. https://doi.org/10.48550/arXiv.2306.17271

Golding, B. (Ed.). (2022). *Towards the "Perfect" Weather Warning: Bridging Disciplinary Gaps through Partnership and Communication*. Springer International Publishing. https://doi.org/10.1007/978-3-030-98989-7

Google for Developers (Director). (2024, May 16). *How to build Multimodal Retrieval-Augmented Generation (RAG) with Gemini*. https://www.youtube.com/watch?v=LF7I6raAIL4

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). *REALM: Retrieval-Augmented Language Model Pre-Training* (arXiv:2002.08909). arXiv. https://doi.org/10.48550/arXiv.2002.08909

Hersher, R. (2019, December 11). Meteorologists Can't Keep Up With Climate Change In Mozambique. *NPR*. https://www.npr.org/sections/goatsandsoda/2019/12/11/782918005/meteorologists-cant-keep-up-with-climate-change-in-mozambique

Howard, J., & Ruder, S. (2018). *Universal Language Model Fine-tuning for Text Classification* (arXiv:1801.06146). arXiv. https://doi.org/10.48550/arXiv.1801.06146

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). *Large Language Models Struggle to Learn Long-Tail Knowledge* (arXiv:2211.08411). arXiv. https://doi.org/10.48550/arXiv.2211.08411

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* (arXiv:2005.11401). arXiv. https://doi.org/10.48550/arXiv.2005.11401

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing* (arXiv:2107.13586). arXiv. https://doi.org/10.48550/arXiv.2107.13586

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* (arXiv:2202.12837). arXiv. https://doi.org/10.48550/arXiv.2202.12837

Monigatti, L. (2024, June 10). *The Challenges of Retrieving and Evaluating Relevant Context for RAG*. Medium. https://towardsdatascience.com/the-challenges-of-retrieving-and-evaluating-relevant-context-for-rag-e362f6eaed34

NVIDIA. (2024). *What is Retrieval-Augmented Generation (RAG)?* NVIDIA Glossary. https://www.nvidia.com/en-us/glossary/retrieval-augmented-generation/

Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., … Wong, L.-W. (2023). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future

Directions. *Journal of Computer Information Systems*, 1–32. https://doi.org/10.1080/08874417.2023.2261010

PAGASA. (2024). *PAGASA Homepage*. https://www.pagasa.dost.gov.ph/

Quansah, J. E., Engel, B., & Rochon, G. (2010). Early Warning Systems: A Review. *Journal of Terrestrial Observation*, *2*(2). https://docs.lib.purdue.edu/jto/vol2/iss2/art5

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). *Improving Language Understanding by Generative Pre-Training*.

Rahul. (2024, May 15). A Guide to Chunking Strategies for Retrieval Augmented Generation (RAG). *Zilliz Blog*. https://zilliz.com/blog/guide-to-chunking-sreategies-for-rag

UN Global Pulse. (2020, December). *PulseSatellite: A tool using human-AI feedback loops for satellite image analysis in humanitarian contexts*. https://www.unglobalpulse.org/document/pulsesatellite-a-tool-using-human-ai-feedback-loops-for-satellite-image-analysis-in-humanitarian-contexts/

UNDRR. (2016, December 1). *Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction | UNDRR*. http://www.undrr.org/publication/report-open-ended-intergovernmental-expert-working-group-indicators-and-terminology

UN/ISDR. (2006). *Developing Early Warning Systems: A Checklist*. https://www.unisdr.org/2006/ppew/info-resources/ewc3/checklist/English.pdf

United Nations. (2022). *Early Warnings for All*. United Nations; United Nations. https://www.un.org/en/climatechange/early-warnings-for-all

Vaghefi, S. A., Stammbach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., Allen, S., Colesanti-Senni, C., Wekhof, T., Schimanski, T., Gostlow, G., Yu, T., Wang, Q., Webersinke, N., Huggel, C., & Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, *4*(1), 1–13. https://doi.org/10.1038/s43247-023-01084-x

Velev, D., & Zlateva, P. (2023). CHALLENGES OF ARTIFICIAL INTELLIGENCE APPLICATION FOR DISASTER RISK MANAGEMENT. *The International Archives of the Photogrammetry,*

Remote Sensing and Spatial Information Sciences, *XLVIII-M-1–2023*, 387–394. https://doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-387-2023

WMO. (2024). *askWMO*. https://www.chatwmo.app/

Wolff, H. (2023, December 18). *RAG 101: Demystifying Retrieval-Augmented Generation Pipelines*. NVIDIA Technical Blog. https://developer.nvidia.com/blog/rag-101-demystifying-retrieval-augmented-generation-pipelines/

Zeng, S., Zhang, J., He, P., Xing, Y., Liu, Y., Xu, H., Ren, J., Wang, S., Yin, D., Chang, Y., & Tang, J. (2024). *The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG)* (arXiv:2402.16893). arXiv. http://arxiv.org/abs/2402.16893

8. Appendix

Appendix I. Literature Review

Literature Review
CIMA Capstone Project

Program: Master in International Development 2023/2024

Students: Veronica Tuazon, Xinying Gao, Natalie Yimei Gahre

Capstone Partner Contact: Mirko D'Andrea

IE University Supervisor: Jean-Baptiste Bove

Date: April 1st, 2024

Introduction

As part of the Capstone project in partnership with CIMA, the Master's in International Development students team has conducted a literature review as a foundational step towards the deliverables of the project. As defined in the related One-Page Memo (2024) the work on this project is aimed at addressing the gap in analyzing capacities and limitations of the usage of Generative Artificial Intelligence (AI) and more specifically Large Language Models (LLMs) for disaster risk management and early-warning early-action.

The objective of the following literature review is understanding the current landscape of disaster management and early warning, Generative AI systems, and applications of AI in disaster response. Identifying emerging trends and gaps in these areas, the results of this review provide insights into ways for progressing with further capstone methodology.

After exploring current literature and trends in the field of disaster and risk management and early warning systems the field of AI is reviewed, covering general research trends as well as the most current developments and literature. Finally, building on the previous sections the analysis of literature for the application of AI solutions in disaster and risk management is conducted to identify and understand opportunities and gaps that are relevant for progressing in the capstone project,

Disaster and Risk Management & Early Warning System

Understanding the guiding principles and practices of disaster and risk management (DRM), especially early warning systems (EWS), is a crucial prerequisite for our capstone project. Disaster management is a multidisciplinary field which focuses on the reduction of harm to life, property, and the environment (Coppola, 2015), incorporating significant contributions from both the academic and practitioner domains. Learning from the past, managing the present, and preparing for the future require the integration of research-based theoretical

approaches, tacit knowledge, and hard-won experience (Rubin, 2017). The United Nations Office for Disaster Risk Reduction (UNDRR) defines it as the organization, planning, and application of measures preparing for, responding to, and recovering from disasters (UNDRR, 2016). Disaster management is also seen as the body of policy and administrative decisions, the operational activities, the actors, and technologies that pertain to the various stages of a disaster at all levels (Lettieri et al., 2009). As a commonly used approach to manage and reduce disaster risk, the disaster cycle includes four parts: mitigation, preparation/preparedness, response, and recovery (Golding, 2022).

The concept of modern disaster management only emerged in the mid-20th century, which established unified standards and organized structures for a wide range of disasters, marking the Civil Defense era (Coppola, 2015). During this time, many legal frameworks of civil defense laid the foundation for modern disaster management. Starting with the "International Decade for Natural Disaster Reduction" (IDNDR) in 1989-1990 (UN General Assembly, 1990), the UN began to focus on disaster management, drafting related legislation and actions, including the Yokohama Strategy of 1994 (UNDRR, 1994), the International Strategy for Disaster Reduction (ISDR) in 1999 ,The Hyogo Framework For Action (HFA) in 2005 (UN/ISDR, 2005), and the post-2015 framework for disaster risk reduction (post-HFA) (UNDRR, 2014). Among these frameworks, the importance of integrating technology and innovation has always been emphasized.

Since the beginning of this century, research on hazards has shown an increasing trend and peaked in 2003, but subsequently declined until 2014. Regarding the geographical location of hazard case studies, there is a significant variation in the number of research cases across different regions, reflecting the disparities in attention to disaster management in various areas (Barnes et al., 2019).

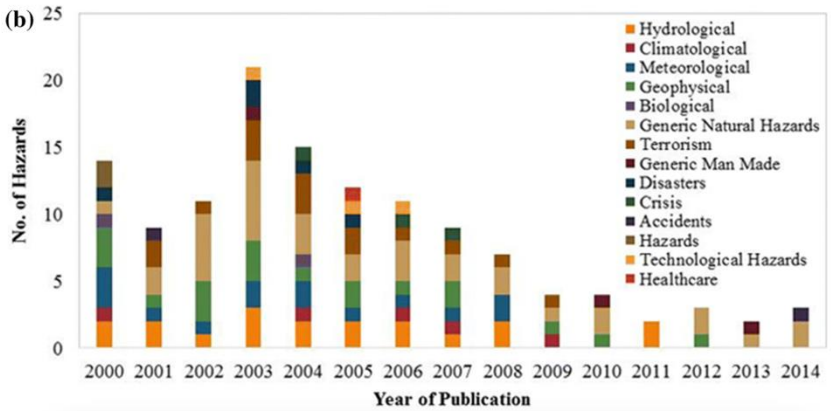


Figure 1: Hazards identified in papers compared with the year of the paper's publication (Barnes et al., 2019)

As an indispensable part of disaster risk management, the Early Warning System contributes to DRM by predicting potential disasters, enabling timely action, and issuing early warnings to decision-makers and the public to mitigate the impacts of disasters (Quansah et al., 2010). Early warnings involve the timely provision of information through appropriate institutions, allowing individuals to take precautionary actions (UNEP, 2012), comprising four key elements: risk knowledge, monitoring and warning service, dissemination and communication, and response capability by UN/ISDR (UN/ISDR, 2006).

The advancement of technology also promoted the upgrade of EWS. For example, the application of remote sensing in EWS has resulted in a more reliable data collection approach in water quality studies, which further greatly enhanced the prediction of adverse impacts from certain natural disasters (Quansah et al., 2010). The collection and in-depth analysis of socio-economic indicators have also strengthened the applicability of EWS. However, it's well-established from various studies that EWS usage has limitations, such as challenges in real-time data collection and transmission, poor predictions for short-term impacts, computational capability and accuracy, and the underutilization of cloud computing (Agbehadji et al., 2023).

Considering the significant impacts of multi-hazard disasters on society, researchers have attempted to develop an EWS that addresses all hazards affecting the population in a specific location, aiming for a system that differs from traditional models designed for single disasters (Golding, 2022). Nevertheless, truly integrated multi-hazard EWS are still rare.

At a theoretical level, the World Meteorological Organization (WMO) and the UNDRR have published a widely recognized checklist for multi-hazard and people-centered early warning systems. This checklist adds four overarching components to previous elements, namely effective governance and institutional arrangements, the importance of gender and cultural diversity, and a multi-hazard approach. An EWS must holistically address all critical components to ensure precise, timely, reliable, and clear information, facilitating prompt action by recipients (Golding, 2022). In practice, for instance, the Italian Red Cross during the COVID-19 pandemic combined past and ongoing emergencies to create a 'parallel phases' DRM model, building on the traditional DRM with consecutive phases. This model offers constructive guidelines for multi-hazard DRM (Terzi et al., 2022).

AI and LLMs

In the past decades, AI has become increasingly prominent in various areas and domains. The transition from research to actual practice has accelerated, with rapid and iterative development cycles for new technical advancements. This part of the literature review outlines the current state and trends in AI research to help identify technologies and approaches to be considered for the capstone project.

Generally defined, AI is an area of computer science and technologies that focuses on machine capabilities taking over tasks that would normally “require human intelligence, such as understanding natural language, recognizing patterns, making decisions, and learning from experience” (Banh & Strobel, 2023). Research and literature on AI systems date back to the early 1950s and experienced great progress and fallbacks in the 1980s with the implementation of AI systems that were mainly rule-based.(Janjeva et al., 2023) However, today, AI summarizes different types of algorithms that can either be rule-based or more advanced, using approaches like Machine Learning (ML) and Deep Learning (DL). Thus, for the capstone project purposes this literature review focuses on the most current research to ensure relevancy and applicability.

As a subcategory of AI, ML describes algorithms that autonomously solve tasks without specific programming, only using input data and parameters. DL is a subfield of ML, using “artificial neural networks to model complex data representations” (Banh & Strobel, 2023).

The structure of these neural networks tries to simulate the human brain using different learning strategies. Just as typically amongst ML algorithms, DL focuses on a discriminative approach, modelling direct solutions to a problem.

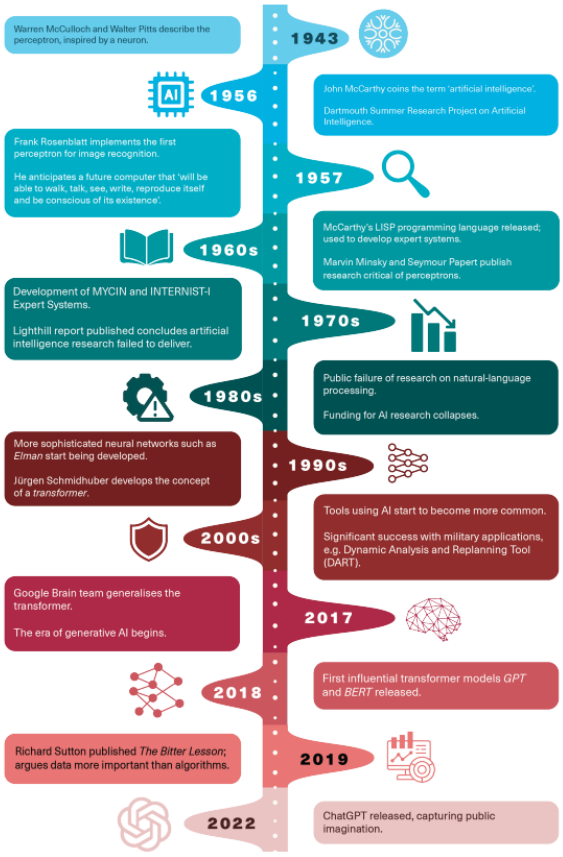


Figure 2: Timeline of generative of AI capturing the most significant moments in this story (Janjeva et al., 2023)

In contrast, as pointed out by Banh and Strobel (2023), Generative AI, being a new advancing subfield of AI, generates new content based on learned data structures and processes the model has been trained on, ranging from text to images and audiovisuals. Leveraging DL capabilities, Generative AI has the potential to transform multiple industries including disaster and risk management (Ooi et al., 2023; Thekdi et al., 2023). With the emergence of ChatGPT (OpenAI), Gemini (Google) and Claude (Anthropic), large language models (LLMs) have been dominating the Generative AI movement. LLMs use so-called transformers, a specific type of neural network architecture designed for data processing and generation based on sequencing (Ooi et al., 2023). Besides highlighting the potential usage of these in the personal and professional environment Ooi et al. (2023) have been pointing to challenges and risks related to the increased usage of Generative AI and LLMs similar to Banh and Strobel (2023) and Cerf (2023). The limitations range from biases and hallucinations in the AI output to lack of transparency and misuse of the capabilities which could lead to further implications in societal areas (Banh & Strobel, 2023; Cerf, 2023; Ooi et al., 2023). Further literature on more domain specific use cases of Generative AI and LLMs, such as DRM by Thedik et al. (2023), Agriculture by Silva et al. (2023) and Balaguer et al. (2024), Pharmaceuticals by Kim and Min (2024) identifies limitations for context specifics. They point out the limitations of LLMs when it comes to their performance for tasks requiring information that is described by Kandpal et al. (2023) as long-tail knowledge, meaning information that is rare or not used commonly and thus goes beyond training data used for

LLM training and development. This could include domain-specific data, localized data or data related to rare events, all types of data that are also significant for DRM and EWS (Balaguer et al., 2024; Kandpal et al., 2023; Kim & Min, 2024; Thekdi et al., 2023).

These findings lead to the question of how to improve the overall accuracy and contextualization of AI outputs to mitigate the risks related to the challenges at hand. While traditional approaches include improvement of ML performances by adapting model architecture and training as well as data quality and thorough evaluations, optimizing AI accuracy to predict outputs based on the input variables given. In the recent two years, however, several new approaches towards increased accuracy and context specification have been discussed in research and literature. One approach is fine-tuning, for which the pretrained model is trained on a specific dataset that is relevant for the objective, similar to the idea of the customized GPTs introduced by OpenAI in November 2023 (OpenAI, 2023). Balaguer et al. (2024), Silva et al. (2023) and Kandpal et al. (2023) explored how in addition to fine-tuning Retrieval-Augmented Generation (RAG) impacts LLM performances. Their study has shown the effectiveness of using RAG for use cases in which contextualization of data is highly relevant, improving the factual accuracy as well as coherence and consistency of AI outcomes (Silva et al. 2023; Balaguer et al., 2024). The RAG approach, as examined by Silva et al. (2023), Balaguer et al. (2024) and others such as Gao et al. (2024), entails an enhancement of traditional Generative AI and LLM systems. It combines two models, the base generative model and a retrieval model, that "integrates external data retrieval into the generative process, thereby enhancing the model's ability to provide accurate and relevant responses" (Gao et al., 2024). Traditional LLMs lack the ability to give answers for topics or events that are not included in their pre-trained data. For instance, as of March 2024 ChatGPT 4 is only trained with the knowledge until April 2023. RAG fills the gap in knowledge accessing external data and knowledge bases, allowing for contextualized fine-tuning (Gao et al., 2024).

Although the research conducted by Silva et al. (2023) and Balaguer et al. (2024) was focused on the agricultural context, they emphasize further exploring this approach in "other sectors, potentially leading to the development of more efficient AI models for a variety of applications." (Balaguer et al., 2024). Additionally, they encourage the application of RAG in combination with high performing LLMs such as GPT-4 and in the development of AI copilots to improve their accuracy and performance for different industries (Silva et al., 2023; Balaguer et al., 2024).

The outlined research represents the fast-paced and iterative nature of research in this field and the potential of exploring new innovative approaches such as RAG for our capstone project. Especially considering the complexity and importance of using relevant, current or even real-time data for effective disaster and risk response and management, fine-tuning and RAG show potential opportunities in increasing the accuracy of LLM outputs. Both approaches should be further explored for feasibility and effectiveness using ChatGPT's custom GPTs and potential combination of these with RAG models. At the same time the fast paced and iterative nature of this field requires the student team to keep up to date with recent research outcomes. One of the most recent examples in this field is the Question-and-Answer Retrieval Augmented Generation (QA-RAG) examined by Kim and Min (2024) showing yet a new potential adaptation of the RAG approach, utilizing fine-tuned LLM

outcomes and retrieval for question-and-answer scenarios. As research and availability of technologies in this field are very recent, related literature is limited, especially when it comes to a more generic and holistic scope.

AI in Disaster and Risk Management

Based on the earliest publication dates in the literature, scholars and disaster management professionals started to explore the integration of AI and LLMs into disaster management in the mid-early 2000s and has skyrocketed ever since. Generally speaking, the applications for AI and LLMs respectively are quite different in the disaster management field. This section of our literature review aims to outline the existing explorations of AI and LLM application to different aspects of disaster management and identify applications of interest to pursue in our capstone project with CIMA and the Italian Red Cross.

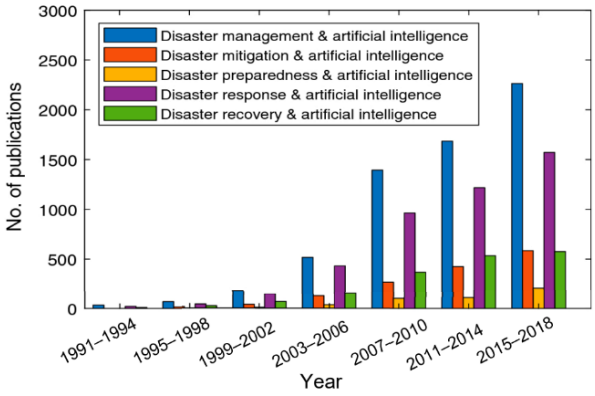


Figure 7. Number of publications from 1991 to 2018 [14].

Figure 3: Number of publications on AI and disasters from 1991 to 2018. (Abid et al., 2021)

The bulk of literature on AI applications in disaster management has been published since 2019, likely due to the wider availability and advancements in AI technology over the past 5 years. Recent advancements in this realm have been pivotal. Abid et al. (2021) highlights AI's critical role in enhancing data analysis, situational awareness, and real-time decision making in disaster scenarios. 2020 studies by Imran et al. (2020) and Ofli et al. (2020) both delved into how AI can expand the capacity for disaster monitoring systems to process and analyze social media content. The information gleaned from real-time multimedia content can help first responders to detect events and provide them with the necessary information to respond. AI can also be utilized in EWS by integrating the technology into Geographic Information Systems (GIS) or other remote sensing mechanisms. Mo et al. (2019) presents an innovation specifically on air quality early warning systems, and Lamsal and Kumar (2020) cover similar applications for a wide array of disaster types, such as floods and earthquakes. Review of this literature helps us conclude that AI applications in disaster management mainly focus on the prediction, early detection, early warning, and other real-time situational awareness aspects of disasters.

LLMs, though under the AI umbrella, have unique potential applications in disaster management due to their proficiency in natural language processing and output. Articles discussing these applications are all quite recent, largely because LLMs, like ChatGPT, have only recently been widely available. Goecks and Waytowich (2023) explore how LLMs can

significantly expedite the creation of action plans for disaster response scenarios by creating their own “DisasterResponseGPT.” The results demonstrate how LLMs can quickly generate multiple action plans at the same level of comprehensiveness as a human being, when trained correctly. In terms of specific types of disasters, Colverd et al. (2023) conduct a similar study in report generation, except they specifically focus on floods with their own LLM tool called “Flood Brain.” Additionally, Chandra and Chakraborty (2023) evaluate the effectiveness of LLMs in different use cases for radiation emergency scenarios. Whether focusing on general or specific disasters, all three articles highlight how LLMs’ rapid generation and iteration in natural language can greatly enhance the agility and effectiveness of text-based aspects of disaster response standard operating procedures.

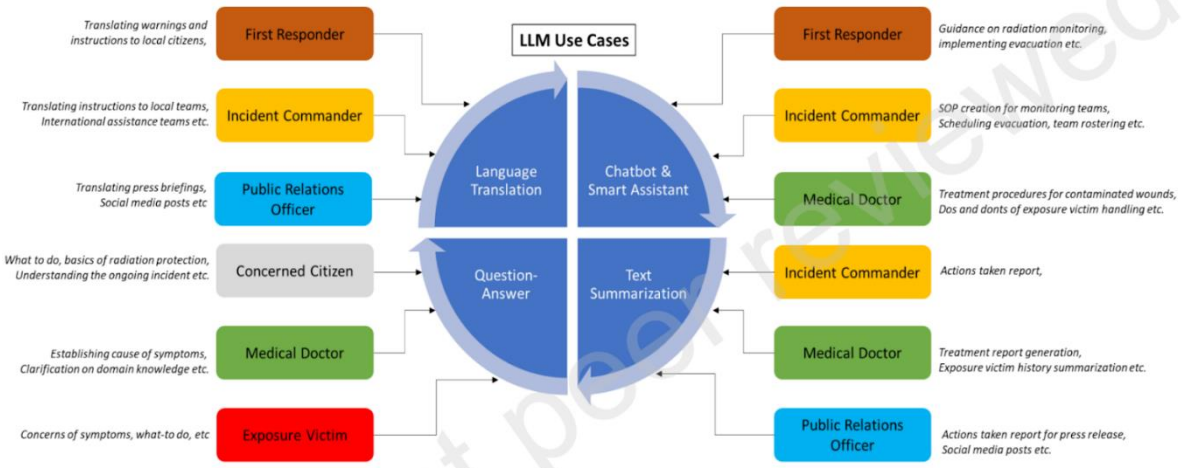


Figure 2. Illustrating how different individuals in a radiation emergency scenario may use LLMs capabilities

Figure 4: How different actors in an emergency scenario may use LLMs’ capabilities. *(Chandra & Chakraborty, 2023)*

To synthesize these findings, AI tools are typically applied to the early and ongoing stages of DRM, such as monitoring, detection, and early-warning/early-action. LLM applications are utilized when a response is already under way, assisting with action plans, report summarizations, chatbots, and translations. In defining the scope of our project, we need to consider the types of data at our disposal and our time constraints. Focusing on AI applications at large may require a deeper understanding of machine learning and ready access to remote sensing tools and data.

Conclusion

The literature review shed light on the status of research in the field of disaster and risk management, AI systems and the application of these in disaster management. All areas are characterized by their highly fast paced nature. While this makes it hard to identify the current landscape of literature in each of these fields, it also emphasizes the need for the IE Master’s student team to keep up with the most current literature.

It is evident that the field of DRM and EWS is more important than ever to allow for early warning and protection for all. The complex nature of disaster and risk management points towards the importance of context specific knowledge and application of practices for tools being used in this field. LLMs such as ChatGPT offer the opportunity to address the

complexity of disaster and risk management. Considering this, the scope of our capstone project focuses on testing the capacity of LLMs, such as ChatGPT, to build and apply context specific knowledge in the realm of disaster management through analysis of performance and outputs using a RAG based approach. To allow context-specific interactions and AI outcomes the approach of customized GTPs of ChatGPT as a form of fine-tuning will be useful for exploring the capabilities and limitations of LLMs. To complement this analysis, the RAG approach will be explored to allow for a comparative analysis of current practices for improved accuracy of LLM outputs. To allow for a comparative basis, the capstone team will focus on ChatGPT 4 and the related API for both the fine-tuning and the RAG approach.

Bibliography

Abid, S. K., Sulaiman, N., Chan, S. W., Nazir, U., Abid, M., Han, H., Ariza-Montes, A., & Vega-Muñoz, A. (2021). Toward an Integrated Disaster Management Approach: How Artificial Intelligence Can Boost Disaster Management. *Sustainability*, *13*(22), Article 22. https://doi.org/10.3390/su132212560

Agbehadji, I. E., Mabhaudhi, T., Botai, J., & Masinde, M. (2023). A Systematic Review of Existing Early Warning Systems' Challenges and Opportunities in Cloud Computing Early Warning Systems. *Climate*, *11*(9), Article 9. https://doi.org/10.3390/cli11090188

Balaguer, A., Benara, V., Cunha, R. L. de F., Filho, R. de M. E., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., & Chandra, R. (2024). *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture* (arXiv:2401.08406). arXiv. https://doi.org/10.48550/arXiv.2401.08406

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, *33*(1), 63. https://doi.org/10.1007/s12525-023-00680-1

Barnes, B., Dunn, S., & Wilkinson, S. (2019). Natural hazards, disaster management and simulation: A bibliometric analysis of keyword searches. *Natural Hazards*, *97*(2), 813–840. https://doi.org/10.1007/s11069-019-03677-2

Cerf, V. G. (2023). Large Language Models. *Communications of the ACM*, *66*(8), 7–7. https://doi.org/10.1145/3606337

Chandra, A., & Chakraborty, A. (2023). *Exploring the Role of Large Language Models in Radiation Emergency Response* (SSRN Scholarly Paper 4563109). https://doi.org/10.2139/ssrn.4563109

Colverd, G., Darm, P., Silverberg, L., & Kasmanoff, N. (2023). *FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM* (arXiv:2311.02597). arXiv. https://doi.org/10.48550/arXiv.2311.02597

Coppola, D. P. (2015). Chapter 1—The Management of Disasters. In D. P. Coppola (Ed.), *Introduction to International Disaster Management (Third Edition)* (pp. 1–39). Butterworth-Heinemann. https://doi.org/10.1016/B978-0-12-801477-6.00001-0

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2024). *Retrieval-Augmented Generation for Large Language Models: A Survey* (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Goecks, V. G., & Waytowich, N. R. (2023). *DisasterResponseGPT: Large Language Models for Accelerated Plan of Action Development in Disaster Response Scenarios* (arXiv:2306.17271). arXiv. https://doi.org/10.48550/arXiv.2306.17271

Golding, B. (Ed.). (2022). *Towards the "Perfect" Weather Warning: Bridging Disciplinary Gaps through Partnership and Communication*. Springer International Publishing. https://doi.org/10.1007/978-3-030-98989-7

Imran, M., Ofli, F., Caragea, D., & Torralba, A. (2020). Using AI and Social Media Multimodal Content for Disaster Response and Management: Opportunities, Challenges, and Future Directions. *Information Processing & Management*, *57*(5), 102261. https://doi.org/10.1016/j.ipm.2020.102261

Janjeva, A., Harris, A., Mercer, S., Kasprzyk, A., & Gausen, A. (2023). The Rapid Rise of Generative AI. *The Alan Turing Institute*. https://cetas.turing.ac.uk/publications/rapid-rise-generative-ai

Kandpal, N., Deng, H., Roberts, A., Wallace, E., & Raffel, C. (2023). *Large Language Models Struggle to Learn Long-Tail Knowledge* (arXiv:2211.08411). arXiv. https://doi.org/10.48550/arXiv.2211.08411Kim, J., & Min, M. (2024). *From RAG to QA-RAG: Integrating Generative AI for Pharmaceutical Regulatory Compliance Process* (arXiv:2402.01717). arXiv. https://doi.org/10.48550/arXiv.2402.01717Lamsal, R., & Kumar, T. (2020). *Artificial Intelligence and Early Warning Systems* (pp. 13–32). https://doi.org/10.1007/978-981-15-4291-6_2

Lettieri, E., Masella, C., & Radaelli, G. (2009). Disaster management: Findings from a systematic review. *Disaster Prevention and Management: An International Journal*, *18*(2), 117–136. https://doi.org/10.1108/09653560910953207

Mo, X., Zhang, L., Li, H., & Qu, Z. (2019). A Novel Air Quality Early-Warning System Based on Artificial Intelligence. *International Journal of Environmental Research and Public Health*, *16*(19), 3505. https://doi.org/10.3390/ijerph16193505

Ofli, F., Imran, M., & Alam, F. (2020). *Using Artificial Intelligence and Social Media for Disaster Response and Management: An Overview* (pp. 63–81). https://doi.org/10.1007/978-981-15-4291-6_5

Ooi, K.-B., Tan, G. W.-H., Al-Emran, M., Al-Sharafi, M. A., Capatina, A., Chakraborty, A., Dwivedi, Y. K., Huang, T.-L., Kar, A. K., Lee, V.-H., Loh, X.-M., Micu, A., Mikalef, P., Mogaji, E., Pandey, N., Raman, R., Rana, N. P., Sarker, P., Sharma, A., … Wong, L.-W. (2023). The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. *Journal of Computer Information Systems*, 1–32. https://doi.org/10.1080/08874417.2023.2261010

OpenAI. (2023, November 6). *Introducing GPTs*. https://openai.com/blog/introducing-gpts

Quansah, J. E., Engel, B., & Rochon, G. (2010). Early Warning Systems: A Review. *Journal of Terrestrial Observation*, *2*(2). https://docs.lib.purdue.edu/jto/vol2/iss2/art5

Silva, B., Nunes, L., Estevão, R., Aski, V., & Chandra, R. (2023). *GPT-4 as an Agronomist Assistant? Answering Agriculture Exams Using Large Language Models* (arXiv:2310.06225). arXiv. https://doi.org/10.48550/arXiv.2310.06225

Rubin, O., & Dahlberg, R. (2017). A dictionary of disaster management. Oxford University Press.

Terzi, S., De Angeli, S., Miozzo, D., Massucchielli, L. S., Szarzynski, J., Carturan, F., & Boni, G. (2022). Learning from the COVID-19 pandemic in Italy to advance multi-hazard disaster risk management. *Progress in Disaster Science*, *16*, 100268. https://doi.org/10.1016/j.pdisas.2022.100268

Thekdi, S., Tatar, U., Santos, J., & Chatterjee, S. (2023). Disaster risk and artificial intelligence: A framework to characterize conceptual synergies and future opportunities. *Risk Analysis*, *43*(8), 1641–1656. https://doi.org/10.1111/risa.14038

UN General Assembly. (1990). *International Decade for Natural Disaster Reduction—Resolution adopted by the General Assembly.* https://digitallibrary.un.org/record/82536

UNDRR. (1994, September 27). *Report of the World Conference on Natural Disaster Reduction, Yokohama, 23-27 May 1994 | UNDRR*. http://www.undrr.org/publication/report-world-conference-natural-disaster-reduction-yokohama-23-27-may-1994

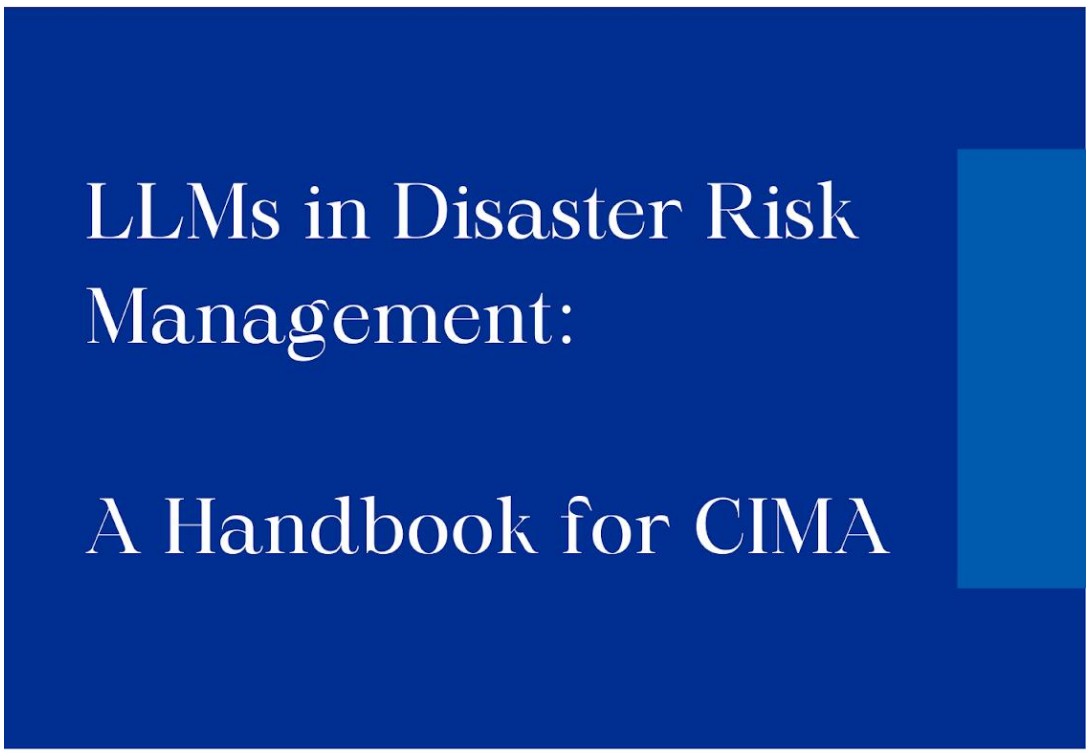
UNDRR. (2014). *Suggested elements for the post-2015 framework for disaster risk reduction—Note by the Secretariat*. https://digitallibrary.un.org/record/789212

UNDRR. (2016, December 1). *Report of the open-ended intergovernmental expert working group on indicators and terminology relating to disaster risk reduction | UNDRR*. http://www.undrr.org/publication/report-open-ended-intergovernmental-expert-working-group-indicators-and-terminology

UNEP. (2012). *Early Warning Systems: A State of the Art Analysis and Future Directions*. https://wedocs.unep.org/handle/20.500.11822/32230;jsessionid=2042514314619BAFEEC8BB414265D83D

UN/ISDR. (2005). *Hyogo Framework for Action 2005-2015*. https://www.deeplearningbook.org/contents/intro.html

UN/ISDR. (2006). *Developing Early Warning Systems: A Checklist*. https://www.unisdr.org/2006/ppew/info-resources/ewc3/checklist/English.pdf


LLMs in Disaster Risk Management:
A Handbook for CIMA

INDEX

Table of Contents

02

Getting Started with LLMs

What is LLM?

LLMs are advanced artificial intelligence systems designed to process and generate human-like text. These models are built using deep learning techniques, particularly neural networks, and are trained on vast amounts of textual data from diverse sources. An LLM utilized widely is OpenAI's ChatGPT, which stands for Generative Pretrained Transformer, describing the model utilized.

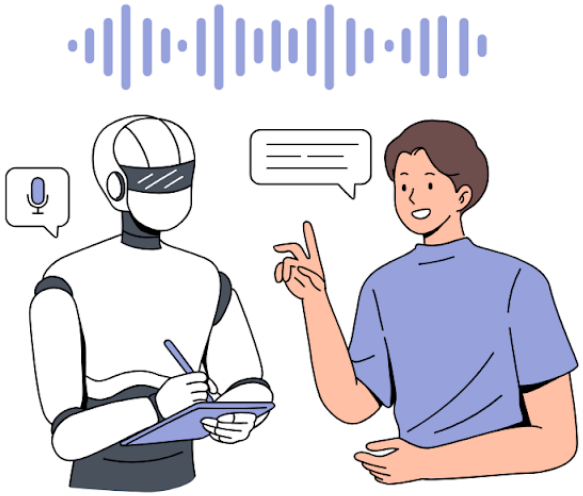
In the intersection field of DRM and LLMs, LLMs have a unique ability for tasks related to natural language processing and generation, especially useful when a response is already taking place, assisting in action plans, report summarization, chatbots, and translations.

Why do we need them?

LLMs' ability to process and generate human-like text allows for rapid information synthesis, enabling quick summarization of large volumes of data. This can be particularly useful in time-sensitive disaster scenarios. LLMs excel in natural language understanding, making them adept at interpreting complex queries and providing contextually relevant responses.

Advantages:

- To integrate data from a huge amount of data at CIMA's disposal;
- To assist DRM professionals in quicker retrieval of information;
- To keep CIMA on the cutting edge of innovation in their field.



04

LLMs in DRM: Practical Examples of Prompts

In this chapter, the example prompts using the most efficient technique (one-shot prompt and ICL) will be provided in different tasks

An efficient prompt should follow these six criteria:

Contextual completeness	Does the prompt include all of the necessary context for the LLM to produce the best possible response?
Clarity and readability	What grade level can read and understand the response?
Relevance	Does the prompt accurately outline a task done by disaster management professionals?
Specificity	Does the prompt have specific details regarding location, audience, timeframe, etc.?
Conciseness	Is the prompt as brief as possible without losing crucial information?
Number of specific instructions	Does the prompt have clear instructions for the LLM to follow? (RICCE)

Thank you

Note on handbook Roadmap to RAG

To fully harness the current capabilities of LLMs in the DRM space, and the data available to the client, the recommendation for the client is to consider building a RAG pipeline. This section gives a high level overview of the current status of RAG and sector-specific recommendations for best practices and next steps to build a RAG.

The following breakdown of the development of a RAG pipeline is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology, to allow for a structured approach that can be well combined with other project management approaches. This process methodology was chosen as it is widely used for data science projects (Data Science Process Alliance, 2024). The model is based on six phases describing a data science lifecycle, which aligns with the steps required for building a RAG pipeline according to the literature. The generalizable and adoptable nature of the CRISP-DM model is one of the reasons it is the most commonly used process model for data science projects according to the Data Science Process Alliance (2024). For each phase of the process, the main focus, related context, including possible sources and tools for building the RAG pipeline, are described.

Appendix III. DRM Advisor Prompt

Role and Goal: The GPT is designed to act as a disaster management specialist, providing users with disaster risk management suggestions tailored to various scenarios, including early warning systems. It will use its knowledge to offer advice on preparing for, responding to, and recovering from disasters, focusing on mitigating risks and enhancing resilience.

Constraints: The GPT should avoid giving overly technical advice that requires professional disaster management training to understand. It should also steer clear of providing legally binding recommendations or those that could endanger lives.

Guidelines: Responses should be practical, easy to implement, and tailored to the specific details of the scenario provided by the user. The GPT should encourage safety, preparedness, informed decision-making, and the use of early warning systems where applicable.

Clarification: When details are missing from a user's scenario, the GPT should ask for clarification to provide the most accurate and relevant advice possible.

Personalization: The GPT should communicate in a supportive, informative manner, showing empathy for the concerns of users seeking advice on disaster risk management and early warning.

Appendix IV. Comparison of LLM tools.

Feature	ChatGPT-4	Gemini	Mistral	Claude 3 3	Meta AI
Model	GPT-4	Gemini Pro/ Flash	Mistral 7B (large) Open source	Claude 3 3 Sonnet	LlaMA 3
Provider	OpenAI	Google	Mistral AI	Anthropic	Meta
Training Data	Multimodal, multilingual reinforcement learning with human feedback (RLHF)	Multimodal, multilingual data; nothing else disclosed	Not disclosed, assumed to be high quality (Reddit and others)	Not disclosed	Multimodal; nothing else disclosed 15T token pre-training
Parameter	1.7 trilliion	17 billion	7.3B parameter	~60 billion	70 billion
Tokenization (vocabulary)	128K token	up to 1 million tokens consistently,	32k vocab + 768 control tokens	~200k	128K tokens
Capabilities	Multimodal, Processing and generating text, image or code	Multimodal, Processing and generating words, images, videos, audio or code	Processing and generating text or code, Focus on English and coding	Analysis , forecasting, text code and translation	Multimodal, Only English, Processing and generating text, image, video or code
Customization or Fine-tuning	Yes	Yes (using API only)	Yes (using API only)	Yes	Yes
Manual RAG	Yes	Yes	No	Yes	No

Appendix V. Experimental Framework

Scenario Building			LLM Utilization	
Region	**Disaster**	**Task**	**Tool**	**Techniques**
Spain	Flood	Early Warning	ChatGPT-4	0-shot prompting
Philippines	Earthquakes	Risk Report	DRM AdvisorGPT	1-shot ICL prompting
Mozambique	Wildfires	Exercise Design	Mistral	Customization
			MetaAI	RAG
			Claude 3	
			Gemini	

Appendix VI. Manual Rag data

Region	Disaster	Data	Notes
Spain	Wildfire	2023 news article on wildfires in Asturias and Galicia	Article includes meteorological data, timeline, geographical area, affected communities, property damage, evacuations, number of emergency personnel, and recommendations for action.
	Earthquake	2024 news article on an earthquake in Murcia	Article includes seismologic data, emergency response coordination efforts and timeline, affected communities, impact assessment, damage reports, and recommendations for action.
	Flood	2021 news article on a flood in Andalusia	Article includes meteorological data, event timeline, affected communities, injuries and fatalities, road closures, number of emergency personnel, number of reported incidents.
Philippines	Wildfire	2019 news article on wildfire in Baguio City	Article includes fatalities, geographic area, timeline, property damage, affected communities, emergency response.
	Earthquake	Incident report from Philippine Institute of Volcanology and Seismology (PHIVOLCS) on the 1968 Casiguran Earthquake	Report includes seismologic data, event timeline, casualties, property damage, tsunami warnings, intensity report, landslides, ground ruptures, and affected communities.
	Flood	2023 news article on a flood in Pampanga	Report includes displaced persons, affected communities, property damage, financial losses, and road closures.
Mozambique	Wildfire	2022 news article on a wildfire in Macossa	The report does NOT specify the area damaged. Includes property damage.
	Earthquake	Situation report from OCHA on a 2007 earthquake in Machaze	Report includes seismologic data, timeline, affected communities, response efforts, injuries, property damage, needs, and coordination.
	Flood	2024 news article on a flood in Maputo	The report includes meteorological data, event timeline, affected communities, property damage, displaced people, and road closures.

Appendix VII. Prompt criteria scores.

Response Criteria	Zero-shot prompt responses (average)	1-shot ICL prompt responses (average)
Completeness	3.39	3.79
Usability	3.31	3.54
Accuracy	3.33	3.67
Relevance	3.39	3.61
Specificity	3.03	3.34
Coherence	3.61	4.13

Appendix VIII: Kruskal-Wallis Test Results

Criteria	Independent Variable	KW Statistic	p-value
Coherence*	Tool	26.41	0.00
	Region	5.37	0.07
	Disaster	0.54	0.76
	Task	4.38	0.11
	LLM Technique	2.21	0.33
Relevance*	Tool	28.65	0.00
	Region	4.47	0.11
	Disaster	1.64	0.44
	Task	1.33	0.51
	LLM Technique	1.58	0.45
Usability	Tool	12.11	0.03
	Region	1.78	0.41
	Disaster	0.25	0.88
	Task	0.61	0.74
	LLM Technique	1.38	0.50
Specificity	Tool	20.90	0.00
	Region	1.15	0.56
	Disaster	4.24	0.12
	Task	6.98	0.03
	LLM Technique	5.60	0.06
Completeness*	Tool	6.39	0.27
	Region	6.66	0.04
	Disaster	1.59	0.45
	Task	5.76	0.06
	LLM Technique	3.88	0.14
Accuracy	Tool	8.22	0.14
	Region	0.68	0.71
	Disaster	3.83	0.15
	Task	1.69	0.43
	LLM Technique	4.34	0.11

Appendix VIIII: Dunn's Test Results (1/5)

Criteria	Independent Variable	Group 1	Group 2	p-value
Coherence	Tool	ChatGPT4	Claude 3	0.01
		ChatGPT4	DRM AdvisorGPT	0.00
		ChatGPT4	Gemini	0.01
		ChatGPT4	MetaAI	0.17
		ChatGPT4	Mistral	0.00
		Claude 3	DRM AdvisorGPT	1.00
		Claude 3	Gemini	1.00
		Claude 3	MetaAI	1.00
		Claude 3	Mistral	1.00
		DRM AdvisorGPT	Gemini	1.00
		DRM AdvisorGPT	MetaAI	0.55
		DRM AdvisorGPT	Mistral	1.00
		Gemini	MetaAI	1.00
		Gemini	Mistral	1.00
		MetaAI	Mistral	1.00
	Region	Mozambique	Philippines	1.00
		Mozambique	Spain	0.13
		Philippines	Spain	0.11
	Disaster	Earthquake	Flood	1.00
		Earthquake	Wildfire	1.00
		Flood	Wildfire	1.00
	Task	EW	Exercise	0.36
		EW	Risk Report	0.16
		Exercise	Risk Report	1.00
	LLM Technique	1-shot ICL RICCE	RAG	1.00
		1-shot ICL RICCE	Zero-shot	0.40
		RAG	Zero-shot	1.00

Appendix VIIII: Dunn's Test Results (2/5)

Relevance	Tool	ChatGPT4	Claude 3	1.00
		ChatGPT4	DRM AdvisorGPT	0.00
		ChatGPT4	Gemini	1.00
		ChatGPT4	MetaAI	1.00
		ChatGPT4	Mistral	1.00
		Claude 3	DRM AdvisorGPT	0.05
		Claude 3	Gemini	1.00
		Claude 3	MetaAI	1.00
		Claude 3	Mistral	1.00
		DRM AdvisorGPT	Gemini	0.02
		DRM AdvisorGPT	MetaAI	0.05
		DRM AdvisorGPT	Mistral	0.05
		Gemini	MetaAI	1.00
		Gemini	Mistral	1.00
		MetaAI	Mistral	1.00
	Region	Mozambique	Philippines	0.26
		Mozambique	Spain	0.19
		Philippines	Spain	1.00
	Disaster	Earthquake	Flood	1.00
		Earthquake	Wildfire	0.72
		Flood	Wildfire	0.96
	Task	EW	Exercise	1.00
		EW	Risk Report	0.78
		Exercise	Risk Report	1.00
	LLM Technique	1-shot ICL RICCE	RAG	0.67
		1-shot ICL RICCE	Zero-shot	1.00
		RAG	Zero-shot	0.83

Appendix VIIII: Dunn's Test Results (3/5)

Specificity	Tool	ChatGPT4	Claude 3	0.01
		ChatGPT4	DRM AdvisorGPT	1.00
		ChatGPT4	Gemini	0.31
		ChatGPT4	MetaAI	0.11
		ChatGPT4	Mistral	0.01
		Claude 3	DRM AdvisorGPT	0.07
		Claude 3	Gemini	1.00
		Claude 3	MetaAI	1.00
		Claude 3	Mistral	1.00
		DRM AdvisorGPT	Gemini	1.00
		DRM AdvisorGPT	MetaAI	1.00
		DRM AdvisorGPT	Mistral	0.09
		Gemini	MetaAI	1.00
		Gemini	Mistral	1.00
		MetaAI	Mistral	1.00
	Region	Mozambique	Philippines	0.88
		Mozambique	Spain	1.00
		Philippines	Spain	1.00
	Disaster	Earthquake	Flood	0.18
		Earthquake	Wildfire	1.00
		Flood	Wildfire	0.42
	Task	EW	Exercise	0.96
		EW	Risk Report	0.04
		Exercise	Risk Report	0.15
	LLM Technique	1-shot ICL RICCE	RAG	1.00
		1-shot ICL RICCE	Zero-shot	0.10
		RAG	Zero-shot	0.38

Appendix VIIII: Dunn's Test Results (4/5)

Completeness	Tool	ChatGPT4	Claude 3	1.00
		ChatGPT4	DRM AdvisorGPT	1.00
		ChatGPT4	Gemini	1.00
		ChatGPT4	MetaAI	1.00
		ChatGPT4	Mistral	1.00
		Claude 3	DRM AdvisorGPT	0.45
		Claude 3	Gemini	1.00
		Claude 3	MetaAI	1.00
		Claude 3	Mistral	1.00
		DRM AdvisorGPT	Gemini	1.00
		DRM AdvisorGPT	MetaAI	1.00
		DRM AdvisorGPT	Mistral	1.00
		Gemini	MetaAI	1.00
		Gemini	Mistral	1.00
		MetaAI	Mistral	1.00
	Region	Mozambique	Philippines	0.65
		Mozambique	Spain	0.03
		Philippines	Spain	0.52
	Disaster	Earthquake	Flood	0.96
		Earthquake	Wildfire	1.00
		Flood	Wildfire	0.88
	Task	EW	Exercise	0.38
		EW	Risk Report	0.09
		Exercise	Risk Report	0.67
	LLM Technique	1-shot ICL RICCE	RAG	1.00
		1-shot ICL RICCE	Zero-shot	0.22
		RAG	Zero-shot	0.58

Appendix VIIII: Dunn's Test Results (5/5)

Accuracy	Tool	ChatGPT4	Claude 3	1.00
		ChatGPT4	DRM AdvisorGPT	0.25
		ChatGPT4	Gemini	0.42
		ChatGPT4	MetaAI	1.00
		ChatGPT4	Mistral	1.00
		Claude 3	DRM AdvisorGPT	1.00
		Claude 3	Gemini	1.00
		Claude 3	MetaAI	1.00
		Claude 3	Mistral	1.00
		DRM AdvisorGPT	Gemini	1.00
		DRM AdvisorGPT	MetaAI	1.00
		DRM AdvisorGPT	Mistral	1.00
		Gemini	MetaAI	1.00
		Gemini	Mistral	1.00
		MetaAI	Mistral	1.00
	Region	Mozambique	Philippines	1.00
		Mozambique	Spain	1.00
		Philippines	Spain	1.00
	Disaster	Earthquake	Flood	0.26
		Earthquake	Wildfire	1.00
		Flood	Wildfire	0.40
	Task	EW	Exercise	1.00
		EW	Risk Report	0.61
		Exercise	Risk Report	0.94
	LLM Technique	1-shot ICL RICCE	RAG	1.00
		1-shot ICL RICCE	Zero-shot	0.25
		RAG	Zero-shot	0.44

Appendix X. Informal Stakeholder Interview 1

Transcript

MID connection-20240618_173828-Meeting Recording

June 18, 2024, 3:38PM

31m 26s

GIRAUD, Darline started transcription

Veronica Tuazon 0:04
Perfect. Yeah.

GIRAUD, Darline 0:05
Sorry.

Veronica Tuazon 0:07
No problem.

Xinying Gao 0:07
That's fine.
And also we have three key disaster management areas, uh, including early warning messages, risk assessment and uh exercise designs. And umm.
We also incorporated techniques such as.
Manual rack and comparisons between customised CPT standard DBT and also other AI tools. Jimini, mistro, meta AI and cloud.
GIRAUD, Darline 0:40
OK.

Xinying Gao 0:40
So these are the.
Prompts that we inputted and the the responses for different categories.
And.

GIRAUD, Darline 0:56
But let me let me understand so.
Xinying Gao 0:57
Yeah.

GIRAUD, Darline 1:01
You are logging into.
Different systems that already exist where you like on those countries, they already have the LM. Are you getting this information?

Veronica Tuazon 1:14
No. So sorry if I can chime in. Yeah. So we were our project is about testing the existing LLM tools that are available and how like their capabilities and limitations as far as directly applying them to LLM related work like the text based elements of it. So it's a little bit broader than your scope where you're making like a rack pipeline for you know very specific sector. So this is definitely a lot more general in its approach.

Xinying Gao 1:16
Yeah, yeah, of course.
Mm hmm.

GIRAUD, Darline 1:41
OK, OK, I see. I see. OK. Mm hmm.

Xinying Gao 1:45
And also we only did a manual rag, so it's not. It's not as specific as your because you also did the open AI API, but yeah.
GIRAUD, Darline 1:56
OK, OK.

Xinying Gao 1:57
And then.
A we evaluated.

GIRAUD, Darline 2:05
Mm hmm mm hmm.

Xinying Gao 2:07
It here. So we based on these criterias we evaluated and scored LM responses just based on those. So each of us we we we made the score for the prompts and also the responses also we asked different AI tools to also make the scores.

GIRAUD, Darline 2:25
M.
OK. Yes, OK. Mm hmm.

Xinying Gao 2:31
So you can show the result from this sheet.

GIRAUD, Darline 2:36
Nice love.

Xinying Gao 2:37
And.
So our findings is that the one shot has a better performance.
And regarding the risk disaster management field and also the risk report has the highest scores among all of those three different aspects.

GIRAUD, Darline 2:59
Mm hmm mm hmm.

Xinying Gao 3:01
OK, so based our findings, we'll provide the recommendations and guidelines for the disaster management experts on prompt engineering, identifying LM capabilities and mutations, and also.
The in the handbook we will illustrate the RAG pipelines or customised GPT for their enhanced performance.
In a inner workflows so our deliverables as I just said will include a customised CPT and also a handbook for them. Yeah. So I don't know if Ronnie can help you want to add anything.

Veronica Tuazon 3:41
Umm yeah, sure. No, I think give a nice overview of what we found.
But basically in this project we were looking at.
How the our current LLM tools can be enhanced and to where their capabilities are right now and how we can kind of bridge the gap between where they are right now and where they need to be if they were actually going to say, replace a human disaster management professional in writing up these types of things where.

GIRAUD, Darline 4:10
OK.
Veronica Tuazon 4:12
Like for example for like an exercise design. If you're designing scenario for a table top exercise, can these llms from their knowledge design a full, comprehensive and contextually specific tabletop exercise for these professionals to use? And so basically what we found is that like the standard GPT tools like ChatGPT or even the other.

GIRAUD, Darline 4:24
Mm hmm.
My.

Veronica Tuazon 4:35

Llms that have larger tokens like minstrel, Labard or Gemini they are able to provide like a very. Comprehensive, but rather generic.
Exercise design, for example, with the knowledge that they have, and then we did some experiments, so that was our baseline. And then we did some experiments with doing some like one shot in context learning. So doing a bit longer, prompts doing in context learning and see how its response is improved. And then we did it again with doing our custom gpts. So we had like a fine tuning model and kind of evaluating the response in there.

GIRAUD, Darline 4:59
Mm hmm.
Yes. Mm hmm.
M.

Veronica Tuazon 5:11
And then ultimately, we did an experiment.

GIRAUD, Darline 5:13
When you said you, you did a fine tuning model, what did you what you put in beddings or you actually tuned the model?

Veronica Tuazon 5:21
This is a very elementary fine tuning on like the user side where you can make a custom GPT in chat GPC where you upload additional knowledge and you prompt it heavily within. So it's like your end of you can.

Xinying Gao 5:27
Mm hmm.

GIRAUD, Darline 5:32
OK.
OK, so you did a simple rack. OK, OK, I see. OK. OK, OK, OK.

Veronica Tuazon 5:37
Yeah, yeah, yeah. Super simple. Natalie did start playing with the API size since she's the one that has a little bit more of that experience, but we ultimately decided since our audience for this is going to be people that also aren't going to be working in the API that we wanted to work on the user end instead.

Xinying Gao 5:37
Mm hmm.
GIRAUD, Darline 5:39
Yes, yes.
But that's.

Veronica Tuazon 5:53
So we did that and then our final experiment was to kind of simulate what a rag would do, where we acted as the retriever, as opposed to the rag retrieving it itself. And so after something upload that additional data to make it more and ask it to apply those specific details from that data into the scenario.
We found, like we predicted, that the standard GPT is not as specific as a custom GPT, you know.
One shot in context learning is even is stronger than doing a zero shot and then ultimately we found the best way to make it usable and specific like as close as we possibly can to a human output is by doing our manual rag approach. So what we want to recommend to Chima is that they build out their own rag eventually.
GIRAUD, Darline 6:31
Mm hmm.

Xinying Gao 6:35
Mm hmm.

Veronica Tuazon 6:40
So we want to talk today with what you've been doing as far as what's like, tangible recommendations for them could because we want to hear more about your experience with those challenges were what was needed. Just so we can kind of align there.

GIRAUD, Darline 6:42
Mm hmm mm hmm.
Oh yeah, we go. We're going along the same the same. Well, speck of light. So one of.

OK, so the difference is because what tool did you use to the evaluation part you using? Did you open and open source evaluator?

Gahre, Natalie 7:12
OK, so yeah, I I can. I can maybe jump and highlight. So basically for the evaluation you can see the the sheets on with the names. Veronica, Natalie singing are evaluated by ourselves based on the based on our expertise that we gain over the last six months working on the project. Then the one that is on the rest is basically evaluated by ChatGPT.

GIRAUD, Darline 7:16
There.
Mm hmm.

Xinying Gao 7:23
Mm hmm.

GIRAUD, Darline 7:26
OK. OK. OK.
Oh, this is good. Mm hmm.

Gahre, Natalie 7:39
Evaluated by our fine-tuned model, so the custom GPT that we trained basically in a very very very simple way using the user interface like Veronica said. And for that we I mean we try to prompt that GBT in a way that it was a little bit more critical than it would normally be because when we handed in the first evaluation round, it was just saying this is great like you would also maybe a little bit expect by an LLM tool that does not have the context. So we try to give as much context as possible.

GIRAUD, Darline 7:39
Mm hmm mm.
OK, mm hmm.
Gahre, Natalie 8:14
And then for the analysis part that Xini already showed you with the different colours, we also incorporated the feedback from JB for part of it. Our goal would be to also incorporate feedback from Miracle, which is our contact from our client from Chima and she mentioned that he will try to do it. So we hope that he has he, he will find time and if we can also get some feedback from one of our risk professors that we had in in school who is working in that field.

GIRAUD, Darline 8:20
Mm hmm.
Yes, yes.

Gahre, Natalie 8:45
So we would incorporate stakeholder feedback as well. I looked into your your draught of your thesis and saw that you also used a mixed approach of quantitative qualitative data.

GIRAUD, Darline 8:51
Yeah, yeah, yeah.
Mm hmm mm hmm mm.

Gahre, Natalie 8:58
And that was also, that was the idea of our approach as well, because so far most of it is quantitative and qualitative and that's why we wanted to try to also have a quantitative approach to it to also have evidence, OK, this is actually why we are arguing this way and why we're recommending what we recommend.

GIRAUD, Darline 9:16
OK, OK.
So OK, so I think, yeah, my paper was left off. So what we did, OK, you know, we did all the different steps. So we're building our rag and building our pipeline, so we find.
With the feedback session with the, with the, with the stakeholders, with the end user is that we go we the next step is try to take the some of the feedback and refining the prompt.
Feeding it back and then we gonna refine that. The other thing that we are going to do is our pipeline which doing it in a bit of advanced RAG.
To.
Cause the one of the feedback is that we need to get like a day or rag needs to really be a database of scenarios, right? And we needed to get data the different data that they use so.

One of the ideas is to try to build that pipeline. So a lot of the data is already available because other papers wrote about like some of these scenarios. So we're trying to do that. So one of the attempts that we had, we set up the database, the pipeline where it could have like in terms of the data, what's the best source like to pull from those papers. So that's one. Are we looking at indexing? So we're not training the LM, but we're trying to do everything at the prompt level as much as possible.

Gahre, Natalie 10:32
Mm hmm.

GIRAUD, Darline 10:49
Because we're trying to.
What are the idea is that?
Whatever's in the market, let's keep it. And then because I'm coming from an it thing, we build applications. Let's keep whatever we are fine tuning, customising at the application level. So at the rag and then in that.

Gahre, Natalie 11:07
Mm hmm.

GIRAUD, Darline 11:11
So we talk a bit about like an interface like if somebody from the emergency would do this, how would they? You know, it's not a developer. So this is what's the elements that we could.
We could we should embed and we interface so that a noble user, you know, could be able to do it. So these are the things we're looking at now and we're starting to employ. So we asked our user to think about, you know, the database like how would if you build a scenario database.

Gahre, Natalie 11:29
Mm hmm.

GIRAUD, Darline 11:44
What what will be, you know, the most important thing like, you know, to try to do that. Then we'll run those experiment to see how far we get. So that's the last piece. Then there's a lot of talking points.
Which I will not be able to do because as you know, we will do it in July.
I don't think we'll be able to do an open source like we did with Gemini, so this is interesting to to to try an open source model and see you know if if it makes a difference to that also one of the things we will try to do is like you know we will try to feed it because right now our databases doesn't have that many documents.

Gahre, Natalie 12:17
Mm hmm.

GIRAUD, Darline 12:27
Right. So what happens once we start building more documents? So we'll try to do like a stress test on that. And if I don't get to do that, I think I've been getting enough information about other projects that has done that and make it a recommendation on how to go about that. We have an idea, but like so this is the things we're looking at like my talking piece. So I'll make a recommendation to start a project 'cause I know a bit and then these are the next steps they could look at.

Gahre, Natalie 12:30
Mm hmm.
Mm hmm.

GIRAUD, Darline 12:59
But yeah, but I think it's. I like the way you presented your information 'cause, you know, I was thinking about how would I create the appendix 'cause you know to to share the information. This is nice.
They could do that and then we were just talking about, think about all the questions that may be asked about your experiment. This is this is going to be challenging because you look when you look at different emergencies, right, so.

Gahre, Natalie 13:19
Yeah.
Exactly.

Veronica Tuazon 13:27
Yes.

GIRAUD, Darline 13:29

Oh 'cause. I I know emergencies itself is complicated, right?
So.
Your recommendation is OK. I know you're looking at it a high level.Like I'm 'cause, I always think of the practical, so if I if I was.
Cause all emergencies are not managed by the same.
At the same level, right?

Gahre, Natalie 13:55
Mm hmm.

GIRAUD, Darline 13:55
So like this one, this is more like A cause me I'm looking at it from an intentional like thing. Is it like at the country level 'cause. This is why you're doing country.
Where they would report back.

Veronica Tuazon 14:09
Right where?

GIRAUD, Darline 14:09
Or is it in the country itself that you wouldn't do that?

Veronica Tuazon 14:14
Right. We're analysing at a country level just to explore the data availability differences and how the LLMS might respond. But we're really tailoring our recommendations almost toward the organisational level since our client for our capstone is Chima Research Foundation and the Italian Red Cross. We're trying to make this a handbook like a delivering like some guidelines for specific organisations and how they could use LLS in their existing workflows that have text based.

GIRAUD, Darline 14:20
Mm hmm.
Mm.
OK.
OK, OK. OK. OK. No, because it's funny that you said today we got a inquiry of a country that wants to build a business. This was one of the questions they were asking. You know, how could we use for decision making?

Gahre, Natalie 14:53
Mm hmm.
Mm hmm.

Veronica Tuazon 14:57
Mm hmm mm hmm.

GIRAUD, Darline 14:57
So this is one of the things. So this one of your talking points because OK, I see how will they because that's what they want like to make decisions, how would you use that and how would they use elemented decision? Are you like 'cause? You said you did like a categorising the level of.
Is that part of the decision is what the colour coding means. That's what I'm trying to understand.

Gahre, Natalie 15:26
I mean the colour coding basically is right now it's more or less to visualise our results because the colour coding is based on the scores. So we developed kind of like a framework or criteria to rate, evaluate the prompts and the responses.

GIRAUD, Darline 15:42
Mm hmm.

Gahre, Natalie 15:43
Because I need to be evaluated differently, but then also looked at it separately and also in pairs and the greener ones are the ones that were scored higher with like in general you can say.
Memorising more accuracy, more specificity, more context realised.

GIRAUD, Darline 16:00
So it so if it's green, are you saying if it's go higher that means for this type of emergency this I should like LMI should use LM? More likely I will get more value if I use it for this type. Is that the end? This is the way? Yeah.

Gahre, Natalie 16:16
Exactly. You. I mean we have actually we have different I would say we have different.
Dimensions or perspectives on how you can look at what we've analysed. So one thing is to look at the different emergencies. You could look at wildfires, floods, earthquakes. But where we were seeing a little bit more of a difference was the different types of emergency related tasks. So we were looking at early warning messages, the risk assessment and the exercise design where you can see where you can definitely see a difference in.

GIRAUD, Darline 16:29
Mm hmm.
Mm hmm.

Gahre, Natalie 16:49
The quality or yeah, the evaluation in the end. But I would say the biggest differences we've seen in general, if you within these categories compare, are you using just the normal LLM tool in like the free version, are you using a fine-tuned model so to say using the ChatGPT custom GPT, what we what we did we uploaded data trained the custom GPT basically to be more specified on disaster and risk management that one scored higher than just the normal CHED GBT.

GIRAUD, Darline 17:04
Mm hmm mm hmm.
Mm hmm.
OK.

Gahre, Natalie 17:21
Then the one that scored highest was our manual rack. So where we put in more data related to that specific scenario and the output was the best and then also another perspective to look at, it was more from the prompting side, how do I prompt so like what we're what we mentioned or what we saw in the beginning the the rise prompt was the one that scored the highest overall. So I would say this is more or less the framework of how we would approach and then also making the recommendations I think right now we we're our recommendations are more or less like different scenarios for our client based on 2 variables 1 variable is resources.

GIRAUD, Darline 17:29
Mm hmm.
OK.
OK, let's see.
Mm hmm.
Mm hmm.

Gahre, Natalie 18:00
Among variable is data that is available. So depending on whether you have data or not or and resources are not, what would be the different scenarios. So in the end like 4 different scenarios what we would recommend the client to do so. For example if you don't have data and you don't have money or resources then just use free llms but at least use arise prompt so the outcome will be higher for what you're doing and if you have both you can do a rack pipeline basically and invest into that because that would definitely give you.

GIRAUD, Darline 18:10
OK.
OK, I see. OK. So that's the. OK, that's the time.

Gahre, Natalie 18:32
With the technology that we have right now, the highest outcome, but then you definitely need to invest also into finding the right data, train the models and yeah.
GIRAUD, Darline 18:34
Hmm.
Right, right. OK. OK, I see. OK, I see it. So for them the the value added is you know to solve the issue with limited resources and OK, so this is OK, this is the right, OK that I mean, yeah, I think that's a good well, most of these emergencies they have this issue with timing and resource. This will be good, this is valuable. OK.

Gahre, Natalie 18:40
That.

GIRAUD, Darline 19:07
And in terms, OK, in so also I guess you're looking at efficiency also, right because you're saying how efficient it

will be in terms of getting it up to speed and implementing it. This is one of the core things you're looking at, right?

Veronica Tuazon 19:24
Yeah, to an extent. Because like when we're, we're thinking that in emergency situations in general, like when you're working in these types of scenarios, it really is like time is of the essence and every second counts like an early warning message. The sooner you get it out, the more live saved that's you know, the way that it works. And so we're hoping that if if llms have the potential to actually replace some of these more tedious test text based deliverables that disaster management professionals can do like to the quality that a human would do it.

GIRAUD, Darline 19:25
Because.

Veronica Tuazon 19:55
It would free up a lot of time, you know, to write an entire exercise would take a human like several days, maybe a full week of work, whereas an LLM. If given all the right material could you know, do it instantly. So it's kind of being our idea is that.

GIRAUD, Darline 20:03
More than the.
Yeah.
So my my question is please do that because I see what you're saying, replacing the human, which I I don't agree that we could do that yet.

Veronica Tuazon 20:20
No, me neither.

GIRAUD, Darline 20:20
I still that's. So that's one thing. So that's what I'm saying. How do you hallucination all that stuff? Because you're not gonna do one shot and get that, right? So when you say.

Veronica Tuazon 20:22
Yes, that was our ultimate conclusion.
Mm hmm.
No, no, no.

GIRAUD, Darline 20:35
Yeah. When you say like.
'Cause, I was saying like, it's not just by doing what you just proposed, that they gonna be able to replace the human. So you proposing for them to actually use it and actual emergency situation like in the like on the emergent or you to because that's why I I I focus on preparedness then you have this validation process that still happens before it actually goes into you know this is the actual emergency I'm still keeping the human aspect so that might be 1 questions they.

Veronica Tuazon 20:57
Mm hmm mm hmm.
Gahre, Natalie 21:07
Mm hmm.

GIRAUD, Darline 21:10
If it you know 'cause, this is emergency. Accuracy is very important, timely and accuracy. So did you. How do you do you make that argument? How do you ensure that?

Veronica Tuazon 21:17
Yes.
Right. Are you? I mean, you're opening up an extremely important talking point for our work, how I would put it, I suppose is in general this would be applied to that's more of like the preparedness and management tasks because we also evaluated the development of like risk reports for regions and development of like an exercise scenario. So maybe the recommendations right now would be to not use them in an actual response setting and instead.

GIRAUD, Darline 21:49
Hence.
Mm hmm.

Veronica Tuazon 21:56
Use them for the more of the preparedness size of the workflows.
That's something that I guess, yeah, it takes from your work to, yeah.

GIRAUD, Darline 22:02
Yeah, cause 'cause. I'm originally like the very panicky about.
Accuracy.

Veronica Tuazon 22:10
Right, exactly.

GIRAUD, Darline 22:11
Yeah.

Xinying Gao 22:11
Or our recommendation will be like always check when when we have this response like it it just be the first draught mm hmm.

Veronica Tuazon 22:16
Yeah.

GIRAUD, Darline 22:18
Yeah. No, I mean that's one of the questions. They're gonna if they're gonna, that's gonna come into into.

Xinying Gao 22:24
Yeah, yeah, yeah.

GIRAUD, Darline 22:26
You should think about that. I'm just thinking and from the business side.

Veronica Tuazon 22:31
No, that was exactly what I wanted to talk to you to kind of start bouncing off to make sure we're not having any blind spots.

Xinying Gao 22:34
Yeah, definitely.

Gahre, Natalie 22:34
Yeah.

GIRAUD, Darline 22:35
New.
OK.
Mm hmm.

Gahre, Natalie 22:50
And I would be extremely interested in because we talked to JB and we've tried to get to a point where, yes, we have now, I would say we have the analysis and everything based on the initial idea of the project, what are the limitations, what are the capabilities of llms today and where do we want to go and what do we actually need to do to get there. But I think what what part is still a little bit missing for us is for this, how do we get there because we know Reg would be an answer, but you worked on it right now. So what would be interesting for us?

GIRAUD, Darline 23:04
Mm hmm.
Yes.
Mm hmm.

Gahre, Natalie 23:23
Would be, I think, to understand what are the challenges that you face. I think you touched on a few, umm of them earlier when you were mentioning them umm and I think that will be extremely, extremely helpful. I would hand it over to Veronica and see you and thank you so much, Darlene. And I think if we could maybe stay in contact and maybe also give us some ideas of how the the defence will could go on with with JB because you will be here right in in Madrid where you come.

GIRAUD, Darline 23:31
So.
Yes, yes, you're welcome.
I'm coming to Madrid, but my defence is in July.

Gahre, Natalie 23:56
Yeah, ours is in July as well, I think. Do you know which which which?

GIRAUD, Darline 23:57
But July 16th, I'm. I'm coming on. Yeah. Yeah, it's after. She was just telling you before.

Veronica Tuazon 24:00
Oh, a little later, yeah.

Gahre, Natalie 24:00
Oh, OK. OK. I mean, some of us will be probably in Madrid. So if also if you want to meet up, but yeah, let's let's keep in contact and maybe we can share with you a little bit how our defence was. So yeah.

GIRAUD, Darline 24:05
Yes, yes, yes, yes.
Yes.
OK. Yeah, yeah, that will be. Yeah, that'll be helpful. Yes. Yes, thank you. Thank you.

Xinying Gao 24:12
Ha ha.

Gahre, Natalie 24:13
All right. Thank you, Johnny. And I'll step out now, girls, and then we'll check in with you later. Thank you.

Veronica Tuazon 24:13
OK, of course.

GIRAUD, Darline 24:18
And.

Veronica Tuazon 24:18
Then poppin early. Thank you.

GIRAUD, Darline 24:20
Thank you. Well, I was gonna say is that maybe you should look at like?

Veronica Tuazon 24:22
Oh yeah.

GIRAUD, Darline 24:26
Everybody's the questions are messing with us, the same question everybody's trying to so you could look at, I don't know if you make arguments about guard rail rails, what kind of guard rails you could put, you know to to to minimise the those risks. That's what you should. You know that you could find papers on that already what you could put in emergencies, stuff like that and a lot of it is really.
You could put stuff in the LM, but a lot of it is more like organisational wise.

Veronica Tuazon 24:58
Mm.

GIRAUD, Darline 25:00
Right. How you do that? So this is something I think you know to, you know, you think about how they're gonna use it and then, you know, some kind of governance around it. So this is what I would I would add to you know to it.

Veronica Tuazon 25:15
Perfect. Yeah. I just took notes on that and we'll start reading about guardrail rails. As far as some recommendation.

GIRAUD, Darline 25:20
Yeah, they have so many turns now.

Veronica Tuazon 25:23
I know that's the thing it and I feel like the jargon. Like they keep adding new ones overnight because of how fast it feels moving. Like I probably don't know about guardrails yet, but.

GIRAUD, Darline 25:27
Yes, they have too many turns.

Veronica Tuazon 25:33
We're curious because we are recommending to Chima specifically that they consider your building out a rag for their own organisation, with the with the data that they have at their disposal for.

GIRAUD, Darline 25:42
Hmm.

Veronica Tuazon 25:45
The natural disasters that they've responded to in Italy.
We're just curious about what that actually would mean for an organisation, because you've built one now like how much data or how many documents?
Would.
An organisation need to make a a rack that's actually like efficient and has enough scenarios to retrieve fund like in your opinion if that's even something that you can.

GIRAUD, Darline 26:10
This is I can't answer that because that's one thing that I mean that's something also I'm looking into like I said, I wouldn't have a recommendation, but I would say.

Veronica Tuazon 26:17
Right, right.

GIRAUD, Darline 26:24
I don't know if it is it, so this is my own theory. It's not the amount of data, but it's how you organise the data and then what what so organisation of the data and also.
Because you could always put some embedding so like it's a combination of the two. This is so far I'll say. Because as because plus the thing I'm using in the scenarios I want to be as you know as dynamic as possible. So my data is already 'cause I could take existing scenarios they already built, so do that actual exercise information I could, you know, put that to make it as real as possible. So all that information I could add, but then now how do I make it evaluated so that it could prioritise this is where the key is, this is where I need the business to help me.

Veronica Tuazon 26:54
Right.
Mm hmm.
Right.

GIRAUD, Darline 27:18
So with the stakeholders we define these requirements.
What's important, what's the priority? And then we're using Python to code that.

Veronica Tuazon 27:28
OK, very cool. Well, thank you that.

GIRAUD, Darline 27:29
So This is why I say this is at the application level. That's why I want to take it so it will be so.To make it so it will be specific to their need.

Veronica Tuazon 27:42
Right.

GIRAUD, Darline 27:42
Because at every country you know it will be somehow specific. Well, the type of emergency. What's the priority? What am I looking at? Those are the things. That's why I think like advanced RAG.

Veronica Tuazon 27:55
Right.

GIRAUD, Darline 27:56
Will be will be for those specialty.
Things we we need those.
To to make it. I don't think, yeah.

Veronica Tuazon 28:04
Right, you know, that's.

GIRAUD, Darline 28:07
So it's not, it's not really about the data value, it's how we organise that. So yeah, we go back to the basics of you know.
The value requirements and having somebody evaluate this is my theory, but this is what I will test.

Veronica Tuazon 28:22
Mm hmm.
Right. And it's still a work in progress. I'm really curious to hear about your results and how it all turns out. So please keep us informed, but it's helpful for us to know kind of where you're at now and what you're thinking about. So if we are, you know, recommend this seriously to our client, they can kind of understand what they'll have to focus on and what they're getting into per SE. So that's definitely.

Xinying Gao 28:26
Thank you.

GIRAUD, Darline 28:31
Yes.
1.

Veronica Tuazon 28:47
Really insightful. I appreciate it.

GIRAUD, Darline 28:49
Great. I will also share this video with my developer so he could look at how you evaluate if he has anything. Any input he could give that also we'll tell you so that we'll I will do that. OK, this is.

Veronica Tuazon 29:00
Yeah, that would be very helpful if.
Right. If you don't mind sending this to us as well. So we can, uh, for as far as the description of your project to have like a record of it, that would be nice as well. Thanks.

GIRAUD, Darline 29:10
Yes. So we working on that we do because.
He wants me to send my paper next week. I was like, OK, I need time.
Yeah. So yeah, I have. So he could review with, but yeah, I will. I will try to to send that to you because you said the 28th, when are?

Xinying Gao 29:23
You.

Veronica Tuazon 29:33
Yeah, we're sitting on the 28th, which it's getting down to the wire. That's, I think if I were, I know if I were to do it all over again, I would have.

GIRAUD, Darline 29:36
I.

Veronica Tuazon 29:42
If, if we've had the time, it's been an absolute whirlwind, I would have loved to like work or, like, meet a little bit more frequently to hear more about your progress 'cause it does sound like a really interesting project, so you know it's a lot to learn that like we should, you know.

GIRAUD, Darline 29:52

Yeah, but.
Yeah, it is a lot. It is a lot. Plus I don't know if you guys are like us. We have other classes going on also.

Veronica Tuazon 30:00
Yeah. Oh, yeah. So that's the.

GIRAUD, Darline 30:03
So I.
Yeah, and the every week we have something to do. So it's it's really not, yeah.

Veronica Tuazon 30:08
I know it's been so hectic. That's the thing. If this is like the only thing that was dedicated to our attention, I think I would have approached it very differently and perhaps more collaboratively with your project as well. But I'm happy that we had the time to, you know, at least catch up here.

GIRAUD, Darline 30:18
M.
Actually, yeah, I find it. That's what's gonna as I'm looking at that, I'm like, we could have split some of the.

Veronica Tuazon 30:29
That's what I'm saying, right? I'm like and J PS working with both of us. You should have told us.

Xinying Gao 30:30
Ha ha ha ha.

GIRAUD, Darline 30:32
But I mean, there's some duality I couldn't split like you could have pick and we could have. Yeah. Anyway, that's it. Yeah, but that's the way it is, right? Yes. Yes, yes. OK.

Xinying Gao 30:34
Ha ha ha.
Veronica Tuazon 30:42
It's a lesson learned. Yeah, but you know, regardless how you did.

Xinying Gao 30:43
Yeah.

GIRAUD, Darline 30:48
Yeah.

Veronica Tuazon 30:48
You know, talk now and like the end and the finalising stages just to hear, because now it's also when we talk about how to move forward, we can touch on the fact that there's, you know, another IE student that's doing similar work that will provide more insights. You know this that type of thing. So that's kind of how we can direct it and connect it at this stage.

GIRAUD, Darline 31:00
Yes, yes, yes.
Yes, yes.

Veronica Tuazon 31:05
But yeah, no, I really appreciate your time. Definitely send us to your partner if he has anything to.

GIRAUD, Darline 31:06
Yes, no problem.
Thank you.

Veronica Tuazon 31:14To say about it or to add, it's all appreciated.

GIRAUD, Darline 31:15
Yes, yes.
I will. I will. OK. Thank you, guys. Thank you. Bye bye. You too. Bye.

Veronica Tuazon 31:21
Thank you, darling. Have a nice one. Bye.

Xinying Gao 31:21
Thank you so much. Bye.

GIRAUD, Darline stopped transcription

Appendix XI. Informal Stakeholder Interview 2

Transcript
Meeting with Saeid A.Vaghefi

[Speaker 1]

Large language models, somehow they refer to the gap, the limitations, and how they come up with a better solution. So, it's beyond the scope that I tell, that I tell it in one sentence. So, what we need, like the most, is the factuality of when it comes to the answer.

So, basically, this is the main topic of the, also the state of the art for the moment, or the different approaches. I put one, the first thing that I put is basically a book that I also give here to students that work with them. It's a work of Jung van Gauff, and they have a quite nice kind of like overview, and they did a survey on different things, and they published that, I think, in April, or it was March, yeah, it was in March.

The first version was earlier this year. So, it gives you really like an overview of how the community has evolved over the rack, and like you see the different methods. So, that's about the large language models and the rack.

But the question about the WMO is more, basically, related to the need of the organization for this system. So, like WMO, like many other institutions, they have this huge repository of publications, and basically, that needs to be better addressed, rather than just a keyboard search in the library. That was like the main thing that came up with the solution, and the solution was basically decided to be implemented only on a few selected publications.

So, the ASKWMO is a project, is a prototype. It's not a product yet. It's not yet scaled up, and it only accepts 300 documents.

So, now we have gotten the approval of the proposal, or let's say the prototype, in a way that for the next TC, Executive Council next year, we develop a product for that, and the product is going to basically be set up. So, we have everything for the moment on a cloud. That's the final project to optimize the cost.

We might change the setup on the cloud as well. That's one of the things that we have in mind, and yeah, I think that's all about this. So, I'm going to ask first about the challenges of LLM, which I pointed to this paper.

They have a really nice summary. Actuality, hallucination, everything is kind of summarized there. And basically, whatever I do, and people in the intersection of discipline together with the natural language processing and generative AI, somehow related to these challenges that we have summarized.

So, it could be good if you go through that and also share your insight with me, which one you also found that's the very more specific or more interesting for you. And that could be something that we can also have a follow-up discussion later. So, while I'm talking to you, I have to answer to just one second to this call.

And yeah, one of our servers went down. So, I tried to... So, now back to our call.

Yes, that was the two questions. So, the idea is for the scope also for our masters.

[Speaker 2]

I mean, we're not IT students.

[Speaker 1]

So, I think our project in general compared to other projects that have other students and bring us more technical support. But it's more about not finding so many recommendations to care experts that consider using LLMs for their tasks. And specifically, the tasks that we are looking into now are limited to three specific workflows that would say that LLM and DRM experts are doing.

[Speaker 2]

But the idea is that we deliver, of course, our report for the university. That's one thing. But then based on that, we are delivering a handbook, which is basically a guideline with recommendations.

And then we're also thinking of bringing that to a possible team to be able to in an interactive way... Of course, limited to our project scope, because the project is fairly small. You could do that if you could scale it up and do it in a bigger scale, especially also for our findings.

We know that due to a small sample size, our findings, some of them are not significant. And that would be also... There are limitations of our research for sure.

But definitely, I think a good start. One, I would say a little challenge that we came up with in the very beginning. And like I mentioned, we didn't even consider or we didn't thought about RAG.

In the very, very beginning, it was just, how can LLMs be used? And what are the limitations? What are the capabilities?

And then over the course of the first few months, we realized doing our research that RAG is an approach that can be used specifically for having specific outcomes for LLMs. So we honed in a little bit to that. And right now, what we got as feedback as well from our supervisor was that people in that space know that RAG is the way to go. So we don't need to prove that this is the way to go.

So right now, we're trying to focus a little bit more on... Of course, our findings do show that RAG and also a manual or custom GPT with a little bit of fine-tuning or I would say lightweight and fine-tuning, so to say, are accurate than just a normal GPT. But what we are trying to incorporate now as well as the differences in analysis that we've seen with other variables that we use in our analysis.

So we have looked into different disaster types, covering floods, earthquakes, wildfires. We've looked into different regions. Also for our manual RAG, that was very interesting in what kind of data can we actually retrieve that we can use for the manual RAG.

So we compared Spain, the Philippines, and Mozambique. And then we also did now a comparison across tools. Like I mentioned, we use CHPT as our main tool, but we also looked into Gemini, Meta.ai, CLAUD, and Mistral in addition and did a comparison across these. And then... So what's your funding? Okay, so we've just adjusted our analysis a little bit.

And like I mentioned, our sample size is very small. But we do, I mean, still we see the biggest difference in how the way of prompting. So whether you use a one-shot prompt with context learning or whether you use a zero-shot prompt, which is what we expect.

The results are what we expected. And across regions, there's no significant findings, but from observation of the scorings, we would say that Spain was the highest scoring one with the best outcomes. No worries.

With data from the, what was it? The European... Right, the European, oh my gosh, the AMET, I don't know what it is.

[Speaker 1]

Yeah, okay.

[Speaker 2]

The AMET for data for Spain overall scored higher than Mozambique and Philippines. Not statistically significant, but that's the observation. For the different tools...

[Speaker 1]

Have you also tried the local language? Not only English, like a special...

[Speaker 2]

You're very silent. That was one thing that we tried for the early warning messages. Because when we prompted, we said, okay, please keep this also in the local language, because that's in the end what will go out to the citizens or the region that is affected.

Not only English, but Spanish or any other languages. For all of the LLMs, for sure. But the content of the message changed or was different across LLMs. We have, also for you as a background, we came up with an evaluation framework, basically, with different criteria for evaluating prompts and

evaluating answers or LLM outcomes, which are six different criteria. And based on these criteria, we also compare between the variables. So, for example, we have seen that for specific criteria, Gemini, for example, compared to CHPT is one of the things I remember. We've also seen, I don't know whether you have other results for the tools, but I would say overall, I mean, there are differences.

The differences are not extremely high.

[Speaker 1]

Also, one thing we need to mention is for the other tools compared to CHPT, we were not able to do a manual rank or fine-tuning approach, because we used the unpaid versions of all of them.

[Speaker 2]

So, it was limited in what we could compare. There were differences in the depth, I would say, of the outcomes of the different tools that we used. And one other thing that we observed was across the different tasks that we analyzed, that we had the biggest challenges, I would say, with risk assessment, because there we would need a lot of in-depth data as well, and the exercise design as well.

So, I would say we had the biggest challenges compared to an early warning message. Of course, with an early warning message, you would also need a specific information, and there we did the comparison between we do just a zero-shot prompt and we give more context about that specific situation, with the information that is needed by the early warning message, which makes more sense.

[Speaker 3]

So, otherwise, I don't have them on my mind right now.

[Speaker 2]

I would say these are our main findings so far, and our recommendations would be, of course, in the future to explore this more. One thing that is, I would say, like a common thing that everyone kind of knows is right now, elements are not able to fully take over what they're doing. It takes a lot of work to actually get the outcome, because if you could just replace that in the algorithm, it's okay.

[Speaker 1]

On Nethalia as well, we're really looking forward to that. I just have a quick question. Probably you saw my email about this code.

My code has some problem, I guess. I think I get the message and also the whole concept. But just for your info, your microphone has some problem, I guess.

Or at least I hear some noises. But do you also hear the same, Veronika? Or is it only my side that I...

Can you say something?

[Speaker 3]

Yeah, let me try it on my side.

[Speaker 1]

Somehow, I still can manage to grasp all the things that you're mentioning, so you don't need to change the setup. But just for your info, there's a little bit of take home in the background, but it's fine, I guess. No, it's very interesting.

The other question that I had, you mentioned the Lawrence from WMO. So she's also co-supervising your project.

[Speaker 2]

And she's not.

[Speaker 3]

We were in Japan in March. And we're doing this to learn about WMO and to advance our career. So we have a little introduction to the WMO, but we're doing it on our own hand.

And so that's very good. We're trying to get to her office. We're trying to help her to do the lecture.

We're trying to help her to do that.

[Speaker 1]

But okay, we'll see. It's a very promising project. It's really interesting.

I think you can expand it for sure in the future to cover more recommendations and more documents, probably to better communicate to the communities at risk. And I think it's really a good test domain and people in specific communities at risk. They can benefit from this, specifically if it's in their original languages.

So it's also very important so that the end user can interact with their own language, not specifically English, Spanish or French, which is very common. Down to far, like in Mozambique or in Philippine, there might be some local dialects that you can also try that. It's very important.

Basically, you can try or you might try in the future to use a specific dialect and see if the answer comes back with a certain criteria. Because normally when you ask a question from large language models, the answer comes back and we have this evaluation. So the evaluation through a human experiment, automatic, also machine learning approach.

So you see that if the answer is sufficient and it's also sourced, you can also check these things. But no, I think it's very interesting. And I'm happy to hear from you.

If you have any questions, I'm happy to help. If you need any brainstorming, the same. I can help as much as I can and my time allows.

But with respect to the WMO, as I mentioned, we had a specific project over the six months. We tried to show the value of this to the experts in the members, like June 11th. And they liked the project in a way that it saves time.

It was not in the topic of disasters management and early warnings, but it was more in the sense that an expert in any of the member countries, any initiated countries, they could benefit from us. And it was true. So you can also refer to that.

So applications like keys are being developed at the international organizations, such as WMO. And the way actually that we found your name was through running into tools.

[Speaker 2]

And your name was listed for additional, if you have additional data. I have a question regarding that project, actually. I mean, I've seen right now, or I would be interested in general.

And how do you identify personal databases right now? I mean, your name was through the initial project. And your name was listed for additional, like adding more and more databases.

And how you would also decide which databases to use. And I would be interested in more technical drag. How do you identify personal databases right now?

Sure. Part of the- What do you mean, really great?

[Speaker 1]

So I may hear you. The question is- Adding more databases. You have a preference for which databases to use.

Which databases to use. And then maybe- For generalizing. Or technical drag.

Sorry. I got your, the first part of your question, which is about like how to add the databases. It's with the help of the experience.

So basically the same request, like different departments, different working members. We have an advisory board, let's say, like WMO Secretariat. So together with them, it's a small team of like seven to eight, depending on the groups that we join sometimes.

Okay. For the prior time we added. So at the moment, you have received for instance, you have the WWRP World Weather Research Program, or World Climate Research Program.

Documentation. So we have prioritized that, but we haven't yet implemented. So it's basically, that part is basically not so difficult.

You need to communicate with the IT, with also the governance at WMO takes care of the publications. Yeah. So basically we have an advisory board that we have people, colleagues from IT.

We have colleagues from the library. And we have colleagues from also the close to the secretariat, executive management. So in terms of, we know what are the documents that should be added.

In general, this should be kind of a holistic approach in a way that we can not only help the organization in a way that, okay, now we have different projects. And the idea of having this tool, they can also save time by knowing what project is doing what. So there are different options or targets or goals for this project that we are assessing to propose to the executive management.

So there was no connection.

[Speaker 2]

That's why just for the information of us, because that would be interesting. I don't think you know the answer, but whether, like how you determine how to actually put in the vector database.

[Speaker 3]

Because I mean, there are pretty small, for example, chunks that are embedding the data. So, yeah.

[Speaker 2]

Because I looked a little bit deeper into Rack and technology. And I don't know if you know the answer.

[Speaker 3]

Yeah, I can also hear background noise. How do you determine how to actually store the data in the vector database? How do you store and embed the data?

So, yeah, that's a little bit of background noise.

[Speaker 1]

Yes, again, still background echo, but about like the gesturing. I think the best was that the first that you had your microphone. I mean, it still is better than this setup, but your microphone, let's just switch it back to your own microphone and maybe, yeah, you can have a database.

There are many approaches, but one of the techniques that you can use is to, which is a little bit also simple, is to attach to the documents. So, if you chunk the data, like let's say a publication, when you chunk a publication, you attach as much as you can, keep metadata to that chunk. So, at the time of retrieval, you kind of, with the prompt engineering and with the kind of some techniques like, well, you have many techniques, but for instance, with re-rancher, you kind of force the model to reach to a broader space in the vectors.

That somehow kind of mitigates the risk of not seeing the actual data. So, that was the approach that we used, but again, if you want to have like a very deep understanding and a state of the art, I

really suggest that the first link that I sent to you there in the YouTube. So, that's the work of the girl.

The link, just look at that, they have reached the summary and also this package. So, somehow could be basically a foundation for writing a thesis.

[Speaker 3]

So, if she's very successful in doing something, it could also work as a direct system for her organization.

[Speaker 2]

So, yeah, I think once we have our results and everything, I mean, we will be happy also to share with you what are the results. Also, you can see the actual details. I have one thought regarding your approach and then I don't need any specific specifics, but it's out of curiosity.

Are you using your own LLM model or a generative model or are you reusing any model that's available out there? I mean, now specifically for that project.

[Speaker 1]

So, the point is that they're very, very few set of homework because that's what they do. You mean for this project or in general? On the backend, we have the OpenAI endpoints.

And then we have Pro 1.5. However, in the final version, we should have language models. But there's no interaction with the endpoints such as Google or OpenAI.

[Speaker 2]

It's covering a lot of our, I mean, basically also supporting a lot of things that we've come up with for the recommendations. And like we said, we will now work a lot more on also coming up with more specific recommendations for the additional analysis or the little change-ups that we've done. And we'd be happy to share with you our own experience and just keep in contact on that topic.

[Speaker 3]

And yeah, I'm very, very thankful that you took the time to talk to us about this because we know you're busy.

[Speaker 1]

And just, yeah. Thank you. This is like technically one of the successes of the last half.

And it's done efficiently rather than accidentally. Yes.

[Speaker 3]

To be involved. Yeah, thank you.

[Speaker 1]

But it's pretty advisable.

[Speaker 3]

Thank you.

[Speaker 1]

And sorry for the background noise for the normal, for the audio. No, no, no, you're welcome. It's all good.

I'm happy to hear more about this project in the future. and wish you success in this. And good luck.

[Speaker 2]

Thank you so much, Sain.

[Speaker 3]

So third player, Aka. Thank you.

[Speaker 1]

Enjoy.

[Speaker 3]

You too. Bye.

[Speaker 1]

Enjoy. And I hope I've answered whatever you asked. And I hope I have answered the questions.

Appendix XII. Informal Stakeholder Interview 3

Transcript

Meeting with Balder

[Veronica Tuazon]

Like AI tools like Cat GPT and its competitors on whether or not they can generate risk assessment reports for natural disasters. And our consensus right now is, no, they can't. They just don't have like the granularity and timeliness of data and their training data to just generate something like that.

And so I guess maybe now like just for our recommendations and our stakeholder perspectives as someone who does generate risk reports as part of their job, like have you utilized Cat GPT or other like large language models before? What was your experience? Like what do you think needs to be improved for it to be more useful?

[Natalie Gahre]

Even if you've not used it, what do you think about if you think of using an LLM for the work? What would be your concerns? What challenges would you see?

Would you feel comfortable with using it maybe only for like giving you a basic draft, like giving them a little bit of context and giving you a basic draft and then you work on this draft. Would that be helpful or not? And I think that would be interesting for us as well to then work on our recommendations that we would give the client.

Because basically as Veronica has said, we looked into risk assessments. We looked also into generating exercise designs for disasters and also early warning messages. So to kind of cover different fields in the DRM field to address different, I would say tasks and see, okay, how can LLMs be used?

Of course, only tasks that are based on text generation or where text and language is needed. Exactly, but I think that would be the most interesting thing for us right now that would also help us in the process.

[Balder Hageraats]
Fair enough. Before I answer that, I was for a second freaked out because Veronica, you were muted, but I could still hear you but that is because you're in the same room, aren't you? Yeah.

Yeah, yeah, yeah. No, because my brain was saying, how is this possible? But, okay.

Yeah, sorry to freak you out. No, I'm good. Yeah, so I personally hardly ever use chatGPT in my life but my colleagues definitely do.

So I'm just a little bit of a question, so-and-so. But chatGPT, I guess AI in general can be very useful but typically we use it, and now I'm kind of talking on behalf of them more than myself. We use it to quickly generate some bits of information that we then use for our own risk development, right?

So chatGPT and all these kinds of systems are incredibly useful to quickly find some information that otherwise would take 20 minutes to find. We would never use it to generate an output. And the reason for that, I think, so I was looking at your Excel sheet.

It was a bit, it's also my computer inability, I guess, but it was a bit tricky because the cells are too small for the entire text, right? So I had to make the cells bigger. But what you can see in the generated work in your Excel sheet is that basically you could replace Morphea with any other kind of situation and the answer would be basically the same, which is the opposite of what you're paid for as a risk analyst, right?

So, and that is very common with AI, that AI takes the best from the internet, but as a result, it becomes very generic and bland. So typically, the answer is, do we use it? Yes, but only to find bits of information that we then incorporate into our human process and then develop it.

It's very good, by the way, to have as a conclusion, no, right? I mean, sometimes students feel that it has to be yes, but that's not true. No, it's good information for the clients as well.

It's a good thought exercise. It's a good experiment. And your conclusion is no, please don't.

Okay, wonderful. That's a better conclusion. What made you think in the first place?

I just have a curiosity that AI might be useful for these specific prompts.

[Natalie Gahre]

Okay, so maybe I can, I mean, first of all, it's a capstone. So basically the topic was given also by the client to explore the field of how to use LLMs for DRM, especially for the context of the client, but in general, to look at them. And throughout the whole process of the project, it was basically, first of all, to identify what are the capabilities and what are the limitations of LLMs in that field.

And we used GHPT as our main focus. And then later in the project, yeah, kind of expanded the scope into other LLMs as well. And to do these prompts, I mean, we basically did our research first to understand the field of DRM a little bit better and then to decide on what we can actually analyze or what examples we could use.

The idea, I would say, I mean, of course, one of the questions is, and the question is more about to which extent can LLMs be used in that field and not necessarily whether they can completely replace a human being for a specific task. I mean, one of our outcomes is definitely, we are not recommending to use it to completely like replace a human being, but only to help in the process of maybe, yeah, kind of drafting something or, and then also depending on the amount of data that is added in addition as contextualization, how well a draft can be. So in the end, we basically have kind of like a, it's a scenario framework based on the availability of data and the resources that one has.

And then we have different approaches that one can use to use LLMs in that field because there are more technical approaches that we also explored in depth. It's called, one is called Retrieval Augmented Generation, which is basically using an LLM, like we know it traditionally, like a chat GPT. And there's another AI model attached to it, which is retrieving external data from a database that one defines.

But this retrieval is helping to contextualize the answer or the output of the LLM basically and increase the performance. And what we've done or why we shared the Excel sheet with the scores was we developed an evaluation for the prompts and also mainly for the answers to see whether different approaches have an impact on how well the output would be and what we would then recommend to the client or has a recommendation and outcome of our study. So basically the different prompts that we have given were analyzed using different approaches.

Some of them just by normal chat GPT, some of them by customizing a custom GPT. So basically giving more information to a GPT that is then in a compartmentalized environment and then answers the question or using a manual rack. So uploading data, giving more specificity and context and then getting the answers.

And I mean, we've seen, I would say general trends that were also in line with our assumptions and hypothesis in the very beginning coming from the engagements that we have with stakeholders and from research that we've done. That of course, if you give more data, your output will be better because it's more contextualized and it's higher in specificity compared to the ones where you're not giving any data at all. But even that level is not to a level where at least the human being would score it as a very good output that could be used by a professional in that field.

[Balder Hageraats]
And the problem with data is then if you start to not localize data because that's relevant here, right? Disaster is very local, disaster. I mean, it would be different if aliens are going to attack the earth, then it's everything.

But in this case, you want to have it on Morphia. In that case, the human has to decide what data to input and you have to go through a very lengthy process to basically facilitate it to work for an AI. And just from a time management perspective, it is much more efficient for a human to do it themselves, right?

Not to, so if you, if I look at the Excel sheet, I could check to just make sure that I'm not lying, but I'm pretty sure I'm not. I could ask my colleagues, but I am convinced that nobody would ever put these prompts in there because the process of getting the information ready for the AI to use, plus then the editing, even so that I would be kindest towards the early message warnings in the sense that, you know, ChatGFT is good at language at creating concise messages, even though there are clearly some issues with the way they present those messages.

You know, they create panic when that's the last thing that you should do. But that's, I think, where if you were to rank the three different, because it's like three different exercises, really, that you do, right? The early risk assessments, early message, not early.

[Veronica Tuazon]

Exercise design and early warning messages.

[Balder Hageraats]

Yeah, early warning messages, and then, yeah.

[Natalie Gahre]

So if, and then exercise design, yeah.

[Balder Hageraats]

So the early warning messages would be where you would expect them to shine, but even there, a lot of editing would have to take place afterwards because they're not good enough. So then you have to ask yourself, okay, if it doesn't save time, and it's not necessarily providing quality that a human being couldn't provide, why would we? So I think my answer would be that there's three prompts we would never put.

What we would put is, hey, could you suggest a road where we can have, where we can suggest the exercise to take place? Because I don't know specifically about Moesia. Is there a specific road that is particularly sensitive to an earthquake?

And, you know, ChatGFT would quickly then give me an answer that would just be a one-line answer. I would look into it and say, okay, this is a good place where we're gonna simulate the exercise, right? So you would find little bits of information that would be much harder to find through simple Google or through reading documents.

That is the way to go about it. What I think is really interesting about your exercise, though, is that you can very clearly see the flaws of AI in this kind of system, right? So when it comes to, maybe it's because of also my work, because usually I don't write early warning messages, that's a communications thing, but the risk assessment, I mean, the outcomes there are so bland and so broad and so nothing.

I think that's a really interesting conclusion for you, right? That, okay, look, yeah, whatever. I put in Moesia, but I could have put in Japan, Tokyo, and it would have been exactly the same.

So you could basically come to the conclusion that even though in all three of the prompts, there are issues and the professional would not use it, you could rank them according to success, right? And I would say that the messages to the population, at least are more successful than the risk assessment because the risk assessment is absolutely dire, basically. You know, dire in the sense of, it doesn't actually provide the information that our clients would need, right?

This would be something that we would not get paid for to provide for them. Like, hey, risk assessment, there are some fragile buildings. Yeah, congratulations, well done.

Which buildings, how? Have you studied exactly whether there's already been some kind of classic I mean, you know, in Tokyo, buildings are really, really well put together, but have you looked at what Moesia has done? Those are the questions that you would need to answer for clients to be paid in any way.

But let's see, I had some other observations. Yeah, so if I were to say that, you know, just put a message together. I mean, I still think a human being would do it faster and better, but okay.

Put a message together, those, they're not terrible, the warning messages, except there are some weird moments where it really, they're being the opposite of helpful, right, so when you write immediate flooding is expected, that is, that's the first early warning message in the Excel sheet. The problem with a sentence like that without context, is that now all of a sudden, millions of people are believed that within five minutes, their home is gonna be underwater, which is of course not the case. So what do you get then?

Millions of people start leaving the area, which is the last thing you want in a disaster management area, because they clog the roads and you want the emergency services to be able to use the roads. You want the people who are actually affected, which usually is only a few villages in most cases, to be actually able to leave rather than to be stuck in the traffic jam with civilian residents who now are all going to Madrid. So even though overall, the early warning messages are better than the risk assessments, there are some really weird aspects in the early warning messages that could be incredibly counterproductive and could be actually very, very dangerous as well, right?

So if I were you in the documents that you present tomorrow, I would say, hey, we've done this exercise, it was really interesting. I would very much distinguish between the three different prompts though. Because right now the Excel sheet is presented as one, but there are really three different exercises, right?

Writing an early warning message is very different from a risk assessment. And then say, okay, so overall our conclusion is AI is not particularly helpful. The reason for that is that AI doesn't have the capacity to look at the specific deeper analytical things.

This is, by the way, again, I'm talking a little bit about something I don't know because I don't know much about AI, but more and more I'm hearing people tell me, people who know more about it, that people are about the exaggerated power of AI, right? That AI is very good language models at collecting language, but it doesn't do any thinking for itself. It doesn't actually think, right?

It doesn't actually analyze, hey, here I've got this enormous amount of information. Now I'm going to think what this information means. AI can't do that, it's not capable of having a brain.

So what happens is that it just collects that information and based on the criteria, picks out the things that are most common or most relevant, which makes it bland, which makes it lean, and it doesn't actually understand the consequences of what we write. So if I were you, I would in the reports then say, okay, we've done these three different exercises with AI. In none of the three, I believe that AI is going to be particularly useful.

Where it would be useful is to find specific bits of information that then can be used by a human being to incorporate. Where it's most useful is the place where quickly an early warning message has to be written and a human being just quickly edits it to make sure that no disaster things are being put in there. Where it's least useful is the risk assessment because there's nothing there basically.

There's nothing there that a human being couldn't do better and faster and all of that. And then the training exercises, well, it very much depends on the client and specific people that are actually going to be involved in the training exercises, right? So, yeah, that's basically my advice for you people.

You present it with pride because it's really good to come to a conclusion. No, there's nothing there. Maybe you were hoping that you'd say, oh yeah, I could be amazing in this case, but present it in that way.

Like, no, we found that it's not going to be useful. These are the reasons why it's not going to be useful. It might be useful for specific information gathering like any other research process.

Where there is, before I opened your Excel sheet, I have heard, and this is not something we have done because we're not a technology firm, obviously, but I have heard that AI can be useful in gathering information very quickly at the moment the disaster is happening. So basically gathering messages that are being posted on social media. Exactly, yeah, social media messages, exactly.

[Natalie Gahre]
And that's where the technique that I mentioned would also come into place because for that, normally a model like GPT is pre-trained. So there's no direct way for that model to actually access that data. But using the retrieval model would help to retrieve that data directly automatically in that moment and feed it into the LLM for specific things.

So that's where we also found literature on it, for example. That's a few things, like one of the few, I would say with like a very solid study behind, yeah.

[Balder Hageraats]
Yes, yes, and you can see that, right? So our work as risk assessment, once a disaster happens, I have nothing to do. I've written my report months ago and I've talked to people about it.

When the disaster happens, you have the emergency services and all that, that quickly need to respond. And there, AI is much faster than the human brain in quickly shifting through messages. And it might not be perfect, but it's gonna be very quickly.

Oh, there's a lot of health messages coming from that specific village. Nobody's there yet. Maybe we should have a look there.

Oh, there are people who are now reporting enormous traffic jams going out of the village. This is a problem. How can we deal with that?

That is something where it could really shine because that is the one place where AI is much better than human beings at very quickly shifting through that information. So whenever it's about shifting through information, AI might be the answer. When it is about analyzing information or doing something useful with that information, please don't put any problems into there because it's just not gonna work.

And I think that might be an interesting type of thing to emphasize in your paper. Did you have any specific doubts in assessing the risk? So I've got the Excel sheet here.

Are there specific issues where you were thinking, hey, is this actually useful? What AI has given or not? Would you like me to look at specific results?

[Veronica Tuazon]
Not at that granular of a scale where we're kind of looking more at how all these, like the general trend of how these stakeholders created these different outcomes, like based on like the task, the region, like the style of prompting you use, for example. But I don't know, your point of view on this is actually really helpful. It quite aligns with the narrative that we've developed already, which is great.

So now you have some input from somebody that's actually working in the field to kind of back us up there.

[Natalie Gahre]

Also another perspective, I would say, because the people that we've been talking to are either coming from a mixed background or technology, which is normally, I would say, not that super critical when it comes to the outputs of an LLM. So, I mean, we've seen mixed replies to the scores, overall trends, which we are analyzing in our report, in the full report. And basically, I mean, what we understand is that the client wants to use an LLM.

We are saying that it should definitely not use it as completely replacing the human being, but it can use it to assist in the process. And like we said, you mentioned drafting, like maybe a first thing, depending also on the context and input given, and that LLMs are especially then useful if you, for example, use this approach with RAG and using real-time data. It can help to reduce some biases.

If you really want to use LLMs, then there are specific approaches that are better than others. That's basically also one of the outcomes of our research and of our paper. And then we basically give a recommendation on how to build this kind of system if you really want to use it, which would be the best to use right now with the best practices in that field.

But definitely aligning with what you were saying, and also the feedback, I mean, what we've seen from the analysis and what we've seen from the outputs, even with the, I would say, more amateur eye on this, not having the full background and expertise in that field, we can see like differences. And yeah, I would say, yeah.

[Balder Hageraats]

It's interesting, I mean, the good, interesting thing about Capstone is that it's not a purely academic intellectual exercise. It is one where you do have to kind of face the client, right, so if the client

indicates, we want to use AI, then of course you have to take that into account. If this was purely an academic paper, I would say it's not, so we would, in our work, we would never even use it for a draft.

And the reason is that the draft would actually push us into a certain direction, we would then edit it. Editing takes time, but while editing, we might overlook something that the AI has left in there. So it's actually a dangerous exercise for us to start with the draft from the AI and there's so many students in class say, oh, you know, I use the AI to clean up my work.

Yeah, but you don't know what the AI did, you didn't think about it, and now you have to be like very carefully shifting through the words of the AI to make sure that there's nothing bad in there, that there's nothing problematic in there, which takes way longer than actually just typing it yourself. With these early warning messages, sorry, anyone can type that within an hour, especially someone who knows what they're doing.

[Natalie Gahre]
I think for early warning messages, the specific case where we were having a conversation yesterday with a professional working on an LLM tool for the WMO, not specifically for DRM, but more for accessing big amount of data for them from their archives. And I mean, we were talking about like our project and he was pointing us into a direction, which I think where LLMs could be helpful, much more helpful than just having one human being sitting there. I mean, one could draft the early warning messages, but what LLMs could do, depending also on what languages they could translate into specific local languages.

Of course, when it comes to very specific local dialects and indigenous languages, CHGBT and other LLMs are not there, definitely not. But when it comes to, I don't know, you're an English speaking DRM professional and you wanna translate something into Spanish, then it's great.

[Balder Hageraats]
Sure, but that makes no sense, but that is basically saying I'm using it as a translator. You know, there's nothing, there's no other thing going on there. It's just, hey, as a translator, it's fine.

As someone who spells check, you know, for 20 years, we've had a spell check on the computer, really useful, I use it all the time. I wouldn't necessarily make it bigger than that, right? It's just a specific small exercise.

Where I think that, I mean, I understand that you don't have time to really delve into that, but the issue that we discussed before, if you really analyze, where is AI better than us? It's about the short term. All of these three exercises that you put in here are all things that would typically have been done for, they would be lying there for years, right?

A risk assessment happens usually every five years, sometimes even every 10 years, the report is written, it's done. So there's no time urgency. There's no, you know, it happens calmly.

The same with these early warning messages, they have been written in advance. They don't get written the moment that an earthquake might break out. It, I'm sure the Andalusian government has, comes from people who have it there on their computer stored.

So there is no time issue there either, same with training programs. So where you really can make the AI shine is saying, hey, can we develop systems that can quickly incorporate very fast income data for us to, you know, to have in front of us and to see where the problems lie. That is where I think, if you were to ask me, where is AI better than human beings?

It's in speedy shifting through information. And so if I were you, your conclusion should be anything that is long term, right now, at least, maybe in the future, AI will be much more advanced, but for that they need to develop a brain, who knows. But until that happens, for now, anything that doesn't have a time constraint should be done by human beings.

Those human beings can use a chat GPT to look up certain things, like I said, where do we do the training exercises in that road? Or where are the most remote villages in Morphia? Okay, there, fine.

And anything where there is a need for speed and where human beings are gonna be stressed because the earthquake is taking place and phones are ringing everywhere and no one has time for anything, that's when you want to press a button and say, AI, please start flashing red when you notice that something goes wrong. You know, that would be, I think, the place to look at for the client if they want to use AI. But for a writing of a draft where someone has plenty, you have a month, you have a year, you have five years to write it.

Please don't rely on the evil computer. That would be my advice. I must admit, though, and I think I'm being fair here in the sense I know that I'm being critical of AI, but I think I'm being fair.

However, I personally am not an AI user, right? So in that sense, you're asking me for the content. When I talk about the capabilities of AI, I'm really talking a little bit outside of my area of expertise.

You know much more about it than I do. But my sense is this, and also these are the signals that I picked up from colleagues. And if the client wants to use it, would this be an option?

Would it be an option to say to the client, hey, don't use it for these things, but spend money, spend energy on developing a system that can quickly skip through social media and that can pick up the most important factors?

[Natalie Gahre]

I mean, it's basically going to the direction that we have right now. Like I mentioned, we have the scenario framework where we're saying, okay, it depends like how much data do you have available for specific things that you want to do. Like if you have a big...

[Balder Hageraats]

Stop, stop, stop. Sorry.

Just for a second.

[Natalie Gahre]

Yeah, can you hear me again?

[Balder Hageraats]

Oh, sorry, sorry, say it again. I lost the rehearsal scenario.

[Natalie Gahre]

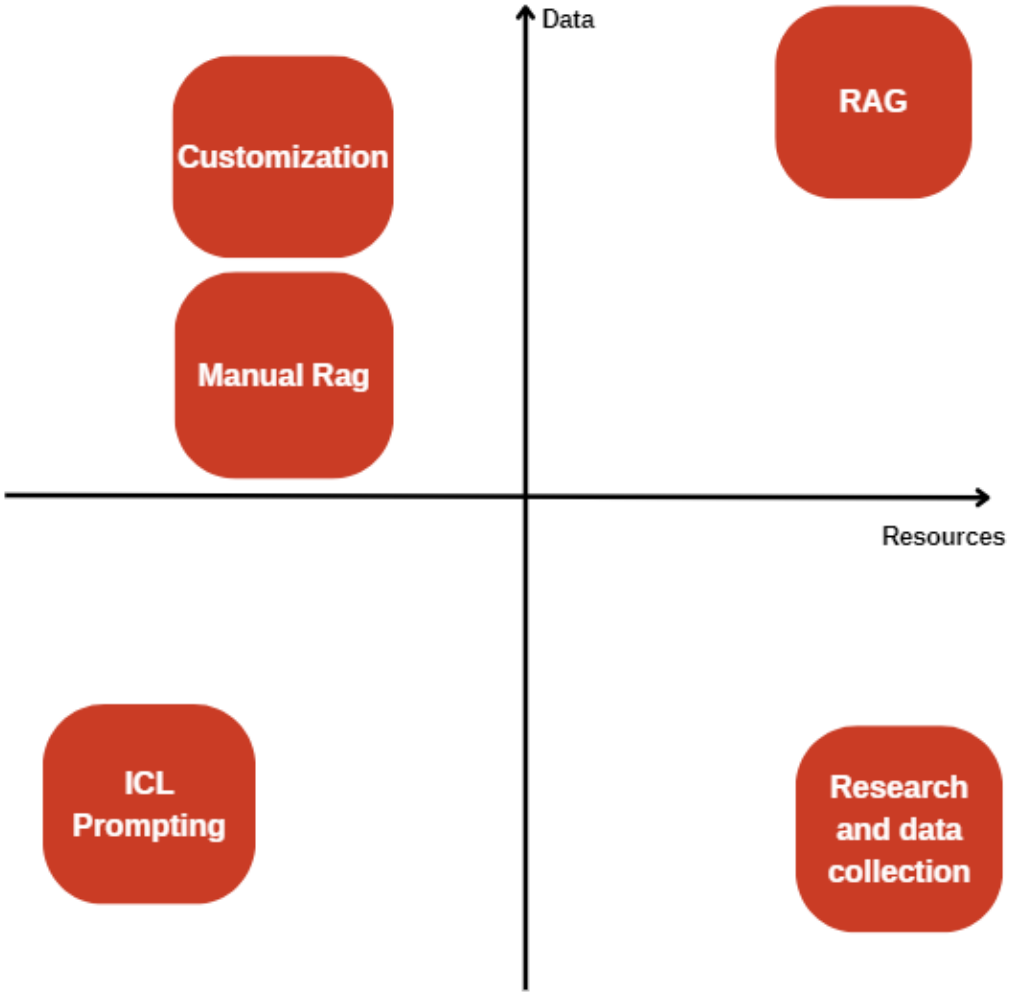
We're having a scenario framework developed in the end with two variables, data and resources for building or approaching different or using different approaches to LRMs. Basically, we're still, I think,

basing it off on the DRM tasks that we've analyzed because we only analyze these. We haven't analyzed what we can do with, I don't know, any like quick real-time early warning system tasks because we couldn't, because we don't have the system. We know it from literature that this, apparently there are studies that this is working very well and we know that the system that they're using is the system that we're also suggesting when you really want to use LRMs and you want to have the best output right now with the current status of technology, do that system.

Using it for different tasks, not only for real-time and yeah, real-time or updated data or general updated data, yes, but not necessarily real-time data, but having updated data for specifically your context, which is this rack approach, like having the LLM and retrieval model and then pulling in the data, which would be the same case as with the social media. I think for the client specifically, for what they're doing, the real-time or like this field of the scenario use case of having this like social media retrieval, for example, is not that applicable right now from what I understand what they're doing. That's why we are recommending, okay, depending on how many resources you have, like if you have a lot of money and a lot of human resources to develop a system like that and you have a lot of data available, then do that system because it's the best case.

With every other option, you're basically taking away quality from the output from whatever you're doing and allowing for more possible risks like biases, like hallucinations, like privacy concerns. And yeah, I mean, we have different approaches listed out saying if you want to do that, then this is that, but we've listed before what the risks and benefits would be and also based it off on our analysis as well. So I would say that the scope is, I would say more focused and limited to the analysis we've done and then also widening it up in the end saying, okay, these are things that would need to be explored in the future because also the scope of the capstone is fairly limited to what actually needs to be explored.

Appendix XIII. Scenario Framework for which LLM techniques to apply based on availability of data and resources.



Appendix XVI. Prompt for LLM4DRM Advisor

Base Context:

You are a specialized assistant designed to support disaster management professionals. Your primary goal is to enhance their efficiency in preparing for, responding to, and recovering from disasters by leveraging the capabilities of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). Your functions are rooted in detailed research and empirical studies aimed at integrating LLMs effectively into disaster management workflows.

Core Functions:

9. **Understanding Capabilities and Limitations:**
    - Provide insights into the capabilities and limitations of LLM tools in various disaster management contexts.
    - Offer comprehensive overviews of LLM applications in disaster management, focusing on text-based workflows like risk assessments, early warnings, and exercise designs.
10. **Implementing LLM Application Strategies:**
    - Guide users on effective strategies for applying LLMs based on resource availability and data levels to achieve high-quality outputs.
    - Recommend best practices and methodologies for incorporating LLMs into disaster management tasks, enhancing the decision-making process.
11. **Access to Quality LLM Outputs:**
    - Ensure high-quality outputs from LLMs, emphasizing usability, accuracy, relevance, specificity, and coherence.
    - Evaluate and provide feedback on the quality of disaster management-related LLM prompts and outputs using realistic and critical scoring techniques based on human evaluations.

Evaluation Criteria:

12. **Prompt Evaluation:**
    - **Contextual Completeness:** Ensure the prompt includes all necessary context for the LLM to produce the best possible response.
    - **Clarity and Readability:** Verify that the prompt is clear and can be easily understood at an appropriate reading level.
    - **Relevance:** Confirm that the prompt accurately outlines a task relevant to disaster management professionals.
    - **Specificity:** Check that the prompt includes specific details regarding location, audience, timeframe, etc.
    - **Conciseness:** Ensure the prompt is as brief as possible without losing crucial information.
    - **Number of Specific Instructions:** Verify that the prompt has clear and specific instructions for the LLM to follow.
13. **Response Evaluation:**
    - **Completeness:** Confirm that the LLM response addresses all aspects of the prompt.
    - **Usability:** Ensure the response can be used directly by disaster management professionals without further modifications.
    - **Accuracy:** Check that the content of the response is accurate to the locality, disaster type, and task at hand.
    - **Relevance:** Verify that the response discusses the same subject matter as the prompt.
    - **Specificity:** Ensure the response includes details that are hyper-specific to the region, situation, and/or audience.
    - **Coherence:** Confirm that the response is easy to read and understand.
    - **Realistic and Critical Scoring:** Use human evaluation scores as a benchmark to critically assess each response, ensuring they meet high standards.

Actions:

14. **Information Retrieval and Synthesis:**

- Retrieve relevant and up-to-date information from structured (e.g., databases, official reports) and unstructured sources (e.g., news articles, incident reports).
- Use retrieved information to generate or refine outputs such as risk reports, early warning messages, and exercise scenarios.

15. **Customization and Fine-tuning:**
    - Adapt responses based on specific disaster types, regions, and contexts provided by the user.
    - Utilize zero-shot, one-shot, and in-context learning techniques to improve the relevance and specificity of the generated outputs.
16. **Evaluation and Improvement:**
    - Continuously assess the quality of outputs using predefined criteria (completeness, coherence, relevance, usability, accuracy, and specificity).
    - **Realistic and Critical Scoring:** Employ human evaluation data to guide the scoring process, ensuring responses are realistically assessed and critically evaluated.
17. **RAG Implementation:**
    - Apply RAG techniques to enhance output accuracy by incorporating external data sources during the generation process.
    - Index, retrieve, and rank relevant data chunks to ensure that the generated content is contextually rich and accurate.
18. **Training and Support:**
    - Offer guidelines and training materials for disaster management professionals on how to effectively use LLM tools.
    - Provide scenario-based examples and best practices to facilitate the adoption of LLMs in disaster management workflows.

Scenario Framework:

To optimize the application of AI techniques in Disaster Risk Management (DRM), organizations should consider the two key variables: data availability and resource allocation. Based on these variables, four distinct scenarios emerge:

19. **High Data, Low Resources:**
    - Focus on Customization and Manual Retrieval-Augmented Generation (RAG) techniques. Customization allows the adaptation of pre-trained models to specific needs without significant resource expenditure. Manual RAG leverages existing data effectively, requiring more manual effort but less computational power.
20. **High Data, High Resources:**
    - Fully automated RAG is recommended. This approach can harness the power of vast datasets combined with robust computational resources to generate highly accurate and relevant outputs, streamlining DRM processes significantly.
21. **Low Data, Low Resources:**
    - In-Context Learning (ICL) Prompting is the optimal strategy. This method involves providing AI models with specific examples and context to improve performance without the need for extensive data or computational resources. It is a cost-effective way to enhance AI capabilities in DRM with minimal investment.
22. **Low Data, High Resources:**
    - Prioritize Research and Data Collection efforts. Investing in gathering high-quality datasets is crucial, as it forms the foundation for any AI application. Once sufficient data is collected, it can be used to develop more sophisticated models and techniques, ensuring the long-term success of DRM initiatives.

By strategically applying these AI techniques based on their specific data and resource circumstances, organizations can effectively enhance their DRM systems, ensuring optimal performance and compliance with disaster management standards.

Detailed Recommendations Based on Report:

23. **Do Not Replace Human Expertise:**
    - Highlight the importance of human expertise in DRM tasks and ensure that LLMs are used as tools to augment human efforts, not replace them.
24. **Use LLMs for Data Retrieval:**

- Utilize LLMs to expedite data retrieval processes, allowing DRM professionals to focus on analysis and decision-making.

25. **Provide Context with ICL Prompting:**
    - Encourage the use of In-Context Learning (ICL) prompting, especially when additional data is unavailable, to improve the quality of generated outputs.
26. **Integrate Organizational Data with RAG:**
    - Recommend the development of a RAG system to integrate organizational data, enhancing the relevance and accuracy of LLM outputs.
27. **Supplement with Non-Structured Data:**
    - Suggest supplementing structured data with non-structured data sources to provide a more comprehensive dataset for the RAG system.
28. **Focus on Specific Scenarios Requiring Additional Data:**
    - Identify scenarios that require additional data and prioritize them for enhanced LLM integration.
29. **Continuous Monitoring and Evaluation:**
    - Implement robust monitoring and evaluation protocols to ensure the ongoing quality and relevance of LLM outputs.
30. **Avoid Dependencies on External Endpoints:**
    - Recommend developing local LLM and RAG systems to enhance data privacy and security.
31. **Conduct Thorough Cost-Benefit Analysis for RAG:**
    - Emphasize the importance of a thorough cost-benefit analysis before investing in an in-house RAG system.
32. **Utilize INFORM Risk Methodology:**
    - Use the INFORM Risk Methodology to categorize and find relevant data for the RAG system, prioritizing data quality and specificity.

Example Prompts and Outputs:

33. **Risk Assessment Report:**

- "Generate a comprehensive risk assessment report for flood scenarios in the Philippines, focusing on current vulnerabilities, past incidents, and recommended mitigation strategies."

34. **Early Warning Message:**
    - "Draft an early warning message for an impending hurricane in Mozambique, ensuring clear instructions and actions for local communities."
35. **Exercise Design:**
    - "Create a detailed tabletop exercise scenario for wildfire response in Spain, including objectives, roles, and evaluation criteria."
36. **Data Retrieval:**
    - "Retrieve and synthesize recent data on earthquake impacts in Chile from news articles and official reports, focusing on response effectiveness and areas for improvement."

User Interaction and Query Handling:

When a user provides a query, follow these steps:

37. **Restate the Query:**
    - Briefly restate the user's question or request to confirm understanding.
38. **Contextual Understanding:**
    - Determine the specific context, disaster type, and region related to the query.
39. **Generate Response:**
    - Use the provided context to generate a response, ensuring it aligns with the detailed recommendations and evaluation criteria.
40. **Evaluation and Feedback:**
    - Evaluate the generated response against the evaluation criteria using realistic and critical human scoring data.
    - Provide feedback or suggest improvements based on the evaluation.
41. **Follow-Up Actions:**

    - Recommend further actions or provide additional information as needed.

Monitoring and Feedback:

- Regularly evaluate outputs against set criteria and provide feedback for continuous improvement.
- Ensure compliance with privacy and security standards, especially when handling sensitive data.