DisasterQA: A Benchmark for Assessing the performance of LLMs in Disaster Response

Rajat Rawat **Kevin Zhu**
Algoverse AI Research
rajat.s.rawat@gmail.com, kevin@algoverse.us

1 Abstract

Humanitarian Assistance and Disaster Relief is a key part of military operations. The military must use Artificial Intelligence (AI) properly and with high standards during disasters. The risk of utilizing AI not up to high standards could be wrong decisions leading to the death of civillians. The Emergency Events Database recorded 399 natural disasters in 2023, which totalled 86,473 deaths [1]. This was a significant increase from the 387 natural disasters recorded in 2022 totalling 30,704 deaths [2]. Even though a major reason for the massive increase in casualties between the years was due to the earthquakes in Turkey and Syria, the growing severity of natural disasters necessitates enhanced disaster response efforts to prevent large amounts of casualties [1].

Quick response times in disasters are essential in fatality prevention, as they allow the potential to rescue many people. For this reason, obtaining Actionable Intelligence — valuable data which is immediately applicable — is crucial [3]. High quality data is essential because decisions are made during deadly situations [4]. Remote Sensing has been critical in improving disaster response actions through systematic image extraction from drones and satellites. These extracted images are processed through different techniques, providing emergency managers with essential information for an effective response [5]. However, recent research has explored the integration of LLMs in disaster response.

Despite recent innovations, LLMs should only be integrated into disaster response if they are highly accurate in the field. Since disaster response can be life or death, incorrect decisions can result in significant fatalities [6]. To evaluate the abilities of LLMs in disaster response, we propose DisasterQA, the first open source benchmark made for the purpose of gauging the performance of LLMs on disaster response knowledge. In addition, we evaluated five different LLMs along with four different prompting methods on our benchmark. We analyzed performance of each model qualitatively and quantitatively through accuracy and confidence level. By releasing this benchmark, we hope to encourage further evaluating LLMs in disaster response and to encourage the further development of methods to improve their capabilities in this field.

We spent significant time researching the internet for reliable questions which we could include in our benchmark. We chose to include only multiple choice questions (MCQ) for ease of evaluating the different models' accuracy. We included questions from six different sources: Quizlet, Scribd, Youtube, Quizizz, Slideshare, and Online PDFs. To avoid duplicate questions, before adding any new questions we used the quick search command to check if the same question already existed. The dataset we compiled had 707 questions, offering a decent amount for evaluations. We chose to test five different LLMs on our benchmark dataset: GPT-3.5 Turbo, GPT-4 Turbo, GPT-4o, Llama 3.1, and Gemini. For Llama 3.1, we used the 8B Instruct version. For Gemini, we used the 1.5 Flash version. Choosing these five models offered a wide coverage for our evaluations, with the results. We additionally used four different prompting methods, with the results being in Appendix C.

Our benchmark consists of 707 questions, which is a substantial amount, but it could be expanded. Evaluating a larger set of questions would offer a better representation of an LLM's capabilities in disaster response. A larger set of questions may also cover a broader range of questions and topics

2024 Harms and Risks of AI in the Military (HRAIM 2024).

we may have missed in our benchmark. Additionally, we did not account for unanswered questions, if we did the results may have been different, especially for Gemini.

We have uploaded our benchmark to Hugging Face: https://huggingface.co/datasets/Rajat1212/DisasterQA. We hope that the release of our benchmark encourages further development of LLMs' capabilities in disaster response along with high standards of the military to use AI when making decisions.

References

[1] 2023 disasters in numbers. Technical report, ReliefWeb, 2024. Accessed: August 10, 2024.

[2] 2022 disasters in numbers. Technical report, ReliefWeb, 2023. Accessed: August 10, 2024.

[3] Rungsun Kiatpanont, Uthai Tanlamai, Prabhas Chongstitvatana, et al. Extraction of actionable information from crowdsourced disaster data. *Journal of emergency management*, 14(6):377–390, 2016.

[4] Vimukthi Jayawardene, Thomas J Huggins, Raj Prasanna, and Bapon Fakhruddin. The role of data and information quality during disaster response decision-making. *Progress in disaster science*, 12:100202, 2021.

[5] Nina Merkle, Reza Bahmanyar, Corentin Henry, Seyed Majid Azimi, Xiangtian Yuan, Simon Schopferer, Veronika Gstaiger, Stefan Auer, Anne Schneibel, Marc Wieland, et al. Drones4good: Supporting disaster relief through remote sensing and ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3770–3774, 2023.

[6] Diane Myers. Disaster response and recovery. *US Department of Health and Human Services, Rockville MD*, 1994.

A Prompting Used

To assess whether different prompting methods would improve performances, we evaluated four different prompting methods incorporating self consistency with each.

No Prompting: We prompted the model solely with the question and the phrase mentioned above.

Directional Stimulus Prompting: We implemented this technique by adding this phrase before each question: "You are an emergency manager. Your responses will dictate survival outcomes for both civilians and emergency responders." This phrase was intended to guide the LLM to the perspective of an emergency manager during a disaster.

Chain of Thought Prompting: We utilized zero-shot Chain of Thought prompting by not providing examples, but instead added the phrase: "Let's think step by step." before each question to encourage reasoning in the LLMs when generating a response.

Few Shot Prompting: We applied few shot prompting with three question answer examples related to disaster management. We added this phrase before each question:

"**Question:** Effective delegation involves ________.

Answer Choices:

A. Determining that all credit goes to the team leader.

B. Ensuring that all tasks are equally distributed.

C. Ensuring that the person has the necessary authority to do the job properly.

D. Guaranteeing all employees have a chance to do the same task.

Answer: Effective delegation involves assigning tasks and responsibilities to others while ensuring that the person has the necessary authority to do the job properly. This includes providing clear instructions, setting expectations, and offering support as needed. The answer is C.

Question: What should a leader do when setting up implementation structures to facilitate change?

Answer Choices:

A. Focus on interim systems because there will be time later to deal with long-term impact.

B. Consider primarily long-term impacts because the short-term will take care of itself.

C. Include representatives of all key stakeholder groups.

D. Place the greatest emphasis on how the change will impact the highest levels of management.

Answer: When setting up implementation structures to facilitate change, a leader should include representatives of all key stakeholder groups. This ensures that the perspectives and needs of all parties involved are considered, leading to a more successful implementation of change. The answer is C.

Question: An incident command system provides for a common organizational structure that enables effective and efficient incident management. It is designed to ensure which of the following?

Answer Choices:

A. Achievement of the entity's strategic objectives.

B. Efficient use of resources.

C. Health and safety of the Whole Community.

D. Use of specific agency or organizational codes and acronyms.

Answer: An incident command system provides for a common organizational structure that enables effective and efficient incident management. It is designed to ensure the achievement of the entity's strategic objectives, efficient use of resources, and health and safety of the Whole Community. The answer is A."

Emotional Prompting: To illustrate the dire situation and emphasize the need for correct answers, we preceded each question with this phrase: "A disaster is happening, so your answer to this question is of the utmost importance. People may die given an incorrect answer. Only give an answer once you are as confident as you can be as the fate of some humans' livelihoods rests on your hands."

Self Consistency: We implemented Self Consistency by prompting each question five times and recording each of the generations. We would later compare first index accuracies with majority vote (Self Consistency) accuracies.

B Average Number of Unanswered Questions

Table 1: Average number of unanswered questions per model for all prompting methods

Model	Avg number of questions
GPT-3.5 Turbo	3.8
GPT-4 Turbo	2.8
GPT-4o	3
Gemini 1.5 Flash	37.6
Llama 3.1-8B Instruct	5.8

C Performance of LLMs on DisasterQA

The best performance for each LLM is bolded.

Table 2: Results from no prompting

Model	1st Index	Majority Vote	Avg Confidence
GPT-3.5 Turbo	**74.72%**	**74.72%**	75.69%
GPT-4 Turbo	81.70%	82.10%	38.76%
GPT-4o	82.84%	84.09%	79.09%
Gemini 1.5 Flash	79.64%	78.97%	N/A
Llama 3.1-8B Instruct	73.29%	73.10%	97.40%

Table 3: Results from Directional Stimulus Prompting

Model	1st Index	Majority Vote	Avg Confidence
GPT-3.5 Turbo	72.87%	73.40%	69.67%
GPT-4 Turbo	81.11%	82.13%	38.90%
GPT-4o	82.98%	84.11%	76.22%
Gemini 1.5 Flash	78.07%	80.09%	N/A
Llama 3.1-8B Instruct	71.12%	71.02%	**98.59%**

Table 4: Results from Chain of Thought prompting

Model	1st Index	Majority Vote	Avg Confidence
GPT-3.5 Turbo	72.61%	72.97%	78.17%
GPT-4 Turbo	81.28%	82.13%	**70.56%**
GPT-4o	**84.21%**	**85.78%**	77.02%
Gemini 1.5 Flash	78.73%	80.56%	N/A
Llama 3.1-8B Instruct	72.21%	72.74%	96.87%

Table 5: Results from Few Shot prompting

Model	1st Index	Majority Vote	Avg Confidence
GPT-3.5 Turbo	73.04%	73.12%	70.07%
GPT-4 Turbo	**83.19%**	**83.21%**	21.83%
GPT-4o	83.64%	84.92%	51.91%
Gemini 1.5 Flash	78.31%	79.73%	N/A
Llama 3.1-8B Instruct	**73.35%**	**73.53%**	98.22%

Table 6: Results from Emotional Prompting

Model	1st Index	Majority Vote	Avg Confidence
GPT-3.5 Turbo	71.45%	73.83%	**79.21%**
GPT-4 Turbo	80.40%	80.82%	47.38%
GPT-4o	83.97%	85.11%	**81.44%**
Gemini 1.5 Flash	**81.21%**	**82.02%**	N/A
Llama 3.1-8B Instruct	70.50%	69.93%	95.57%

D Commonly Missed Questions

The models tended to get questions wrong from various areas in disaster response, not a specific area. These areas included emergency management knowledge, disaster knowledge, and GIS/remote sensing applications.

Incorrect Question Example: What activity takes place during the response phase of the disaster cycle?

Answer Choices:

A. Conducting surveillance of health problems

B. Conducting an inventory of available resources

C. Training of health personnel

D. Conducting epidemiological studies

Answer: A - Conducting surveillance of health problems

GPT-3.5 Turbo answered B

GPT-4 Turbo answered A and B, max vote was B

GPT-4o answered B

Gemini answered B

Llama 3.1 Instruct answered B

Incorrect Question Example: During volcanic eruption, the extreme risk zone is within a distance of:

Answer Choices:

A. Up to 100 meters

B. 100 to 300 meters

C. 300 meters to 3 kilometers

D. None of the above

Answer: A - Up to 100 meters

GPT-3.5 Turbo answered C

GPT-4 Turbo answered D

GPT-4o answered D and C

Gemini answered D

Llama 3.1 Instruct answered C

Incorrect Question Example: Effects on radiance in images caused due to variations in topography can be mitigated by:

Answer Choices:

a. Enhancement

b. Band ratioing

c. Image fusion

d. Radiometric correction

Answer: B - Band Ratioing

GPT-3.5 Turbo answered D

GPT-4 Turbo answered D

GPT-4o answered D

Gemini answered D

Llama 3.1 Instruct answered D