Multi-Query Focused Disaster Summarization via Instruction-Based Prompting

Philipp Seeberger and **Korbinian Riedhammer**
Technische Hochschule Nürnberg Georg Simon Ohm
{philipp.seeberger,korbinian.riedhammer}@th-nuernberg.de

Abstract

Automatic summarization of mass-emergency events plays a critical role in disaster management. The second edition of CrisisFACTS aims to advance disaster summarization based on multi-stream fact-finding with a focus on web sources such as Twitter, Reddit, Facebook, and Webnews. Here, participants are asked to develop systems that can extract key facts from several disaster-related events, which ultimately serve as a summary. This paper describes our method to tackle this challenging task. We follow previous work and propose to use a combination of retrieval, reranking, and an embarrassingly simple instruction-following summarization. The two-stage retrieval pipeline relies on BM25 and MonoT5, while the summarizer module is based on the open-source Large Language Model (LLM) LLaMA-13b. For summarization, we explore a Question Answering (QA)-motivated prompting approach and find the evidence useful for extracting query-relevant facts. The automatic metrics and human evaluation show strong results but also highlight the gap between open-source and proprietary systems.

1 Introduction

Insufficient situational awareness during natural or human-made disasters can lead to significant loss of life, property, and environmental damage. Advancements in today's information ecosystem present new avenues for emergency response (Buntain et al., 2021; Kruspe et al., 2021). For example, integrating heterogeneous online sources such as social media and microblogging platforms which rapidly disseminate crucial details about ongoing events (Sakaki et al., 2010; Reuter et al., 2018). This shift has created a multi-stream environment where traditional sources are augmented with emerging online platforms, recognized as a promising area in prior research efforts (Allan et al., 1998; Aslam et al., 2015; Sequiera et al., 2018; Buntain et al., 2021)..

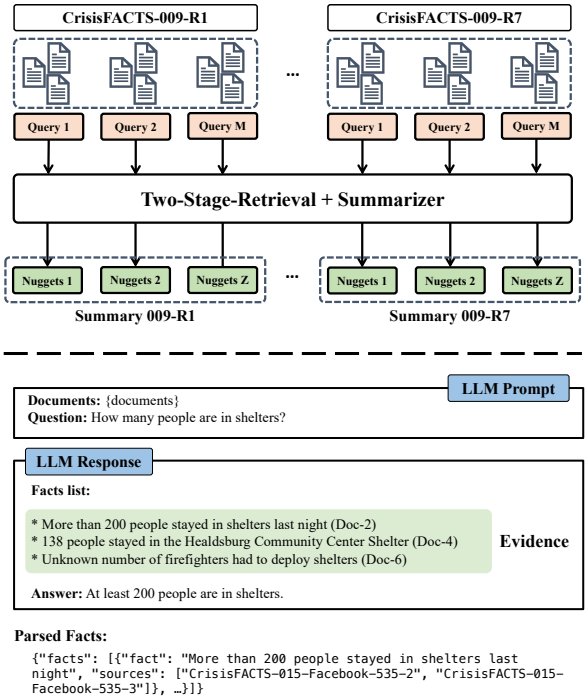

Figure 1: High-level overview of our proposed system and prompting strategy. The upper part depicts the overall pipeline. We call the pipeline for each event-request-query triple separately, resulting into the final summaries for each event-request pair. The lower part illustrates the prompting strategy. We generate query-focused facts using a QA-motivated approach and concatenate the extracted facts to form the final event nuggets.

However, the rapid pace of content creation and the unique characteristics of various information sources and events pose challenges for existing models (Kaufhold, 2021; Seeberger and Riedhammer, 2022b). The introduction of the CrisisFACTS Track aims to address these challenges, asking the community to develop summarization systems that can extract event nuggets for various events (McCreadie and Buntain, 2023).

Recently, the use of instruction-following LLMs has attracted attention in various Natural Lanuage Processing (NLP) domains and zero-shot as well as in-context learning methods show competitive performance to traditional state-of-the-art models. This is also evident in the disaster summarization domain, where extractive frameworks with decent performance (Seeberger and Riedhammer, 2022a) are surpassed by LLMs from proprietary APIs (Pereira et al., 2023). We participate in the second edition of the CrisisFACTS challenge and propose to employ a simple but effective instruction-following approach, leveraging recent advances in open-source LLMs.

Results Summary Our proposed LLM-based event nugget generation approach achieves competitive performance and surpasses the majority of systems in the CrisisFACTS 2023 Track. This trend is shown for both human and automatic evaluation results, underscoring the potentials of LLM-based disaster summarization. However, our qualitative analysis reveals shortcomings in the generated query-related facts, suggesting the need for further development efforts.

2 Model Description

Figure 1 gives an overview of the components of our system and how they are connected to generate the final summaries. First, we retrieve and rerank the documents for each event-request-query triple to obtain query-relevant clusters of candidate documents. Then, we use these clusters to extract query-relevant facts which serve as the basis for the event nuggets and final summaries. We detail the components in the next sections.

2.1 Retrieval and Reranking

We follow a two-stage retrieval approach consisting of first-stage retrieval and subsequent reranking components. This approach often has shown strong cross-domain performance abilities (Thakur et al., 2021). In the first step, we employ efficient lexical retrieval to reduce the computational costs for transformer-based models such as cross-encoders. We first retrieve the top-$k^{(1)}$ candidate documents for each query represented as indicate terms. Next, we process each query-related cluster of candidate documents with a neural reranker model. Typically, these reranker models are cross-encoder variants. Each query-related cluster is reranked by an additional query (e.g., question), and the top-$k^{(2)}$ are chosen to further reduce the set of candidate documents. For further details, we refer to previous work (Seeberger and Riedhammer, 2022a).

2.2 LLaMA-Nuggets

For summarization, we follow a QA-motivated approach and extract query-relevant facts (Figure 1). In this way, we aim to filter out irrelevant documents and abstract only the query-relevant content from the document collections. We prompt an instruction-following LLM for each event-request-query triple and use the retrieved and reranked candidate documents as well as corresponding query as input. Here, the query represents the question, and we ask the model to provide a list of facts which serve as answer evidence. This prompting scheme is also known as chain-of-thought (CoT). We then parse the CoT fact items and cited documents into a structured format. The referenced documents are important for two reasons: 1) Regarding traceability, the task requires providing the source documents of each event nugget. 2) We need an importance score that can be derived from the reranking relevance scores. However, the resulting facts are rather short and atomic, and we aim to produce event nuggets covering a specific topic.

Event Nugget Generation To generate the event nuggets, we iteratively concatenate all generated facts for each specific query. Here, we limit the character length to 200, which corresponds to the task's instructions. As an importance score, we compute the mean of all relevance scores of the referenced documents.

3 Experiments

In the following, we detail the experimental setup including preprocessing, modeling, and evaluation. Throughout our experiments, we consider all web sources provided by the TREC CrisisFACTS Track and reuse the already chunked stream items.

3.1 Dataset

The CrisisFACTS 2023 challenge includes a total of 10 disaster events and additional metadata (e.g., search keywords, queries, source types, etc.). Each event is composed of multiple requests (i.e., days) covering multi-stream data extracted from online sources such as Twitter, Reddit, Facebook, and Webnews. For further dataset details, we refer to

↓ Method Event→	009	010	011	012	013	014	015	016	017	018	Avg
GREEDY†	0.70	5.50	0.83	3.01	7.78	0.15	0.41	4.36	4.54	5.31	3.26
ILP-MMR†	0.52	5.37	0.80	2.37	6.62	0.40	1.54	4.10	5.14	5.12	3.20
LLAMA-NUGGETS	**15.77**	**9.56**	**10.77**	**20.32**	**17.42**	**14.57**	**15.84**	**18.90**	**18.46**	**18.28**	**15.99**
TREC-MEAN	7.85	8.10	8.11	7.11	8.28	6.20	7.16	8.41	8.13	8.98	7.83

Table 1: Comprehensiveness (i.e., recall) scores (x100) for our submitted system runs. † denotes that the corresponding system run was evaluated with automatic nugget matching. Bold numbers indicate the best performance.

↓ Method Event→	009	010	011	012	013	014	015	016	017	018	Avg
GREEDY†	14.29	13.69	1.54	22.44	46.49	7.14	4.76	20.88	30.09	47.37	20.87
ILP-MMR†	13.10	21.54	8.00	23.33	**47.47**	28.57	26.98	38.86	41.07	62.82	31.17
LLAMA-NUGGETS	**48.61**	**25.66**	8.35	**47.39**	45.02	**78.45**	**52.37**	**48.50**	**66.16**	**78.75**	**49.93**
TREC-MEAN	25.08	18.08	**10.11**	27.86	35.97	35.60	24.34	23.98	34.34	47.32	28.27

Table 2: Redundancy (i.e., precision) scores (x100) for our submitted system runs. † denotes that the corresponding system run was evaluated with automatic nugget matching. Bold numbers indicate the best performance.

the official CrisisFACTS website[1].

3.2 Preprocessing

For preprocessing, we normalize all Twitter posts to represent the text content similarly to other present online sources. Specifically, we remove all retweet indicating prefixes, user mentions, emoticons, emojis, and URLs. Furthermore, we eliminate all hashtag symbols and split the text into corresponding words using the *WordSegment*[2] toolkit. Lastly, we also remove exact duplicates. The majority of these duplicates are found in the tweet documents, mainly associated with retweets.

3.3 Model Details

As already mentioned in Section 2 and shown in Figure 1, our pipeline is composed of the two-stage-retrieval and subsequent summarization modules. Next, we describe the model and implementation details for each of these components.

RETRIEVER For first-stage retrieval, we follow previous work (Seeberger and Riedhammer, 2022a) and utilize the BM25 model with default settings from the *PyTerrier* (Macdonald and Tonellotto, 2020) library. To increase the recall, we extend it with Bo1 (Amati and Van Rijsbergen, 2002) query expansion and set the number of feedback terms and documents as 3 and 20, respectively. For each query, we concatenate the stemmed query text and indicative terms, and retrieve the top-$k^{(1)} = 250$ candidate documents.

RERANKER For second-stage reranking, we employ the MonoT5[3] model that is fine-tuned on the MS MARCO passage dataset. MonoT5 is based on a pre-trained sequence-to-sequence model that generates relevance labels as target tokens (Nogueira et al., 2020). In preliminary experiments, we find that this family of rerankers outperformed QA-motivated and encoder-based models (Seeberger and Riedhammer, 2022a). To reduce computational costs, we select the top-$k^{(2)} = 30$ documents for the LLAMA-NUGGETS and 50 documents for the baseline models, respectively.

LLAMA-NUGGETS For event nugget generation, we use the LLaMa-2-13B[4] model series (Touvron et al., 2023) as underlying LLM, while relying on the fine-tuned version trained on multiple instruction datasets. We also experimented with foundation models, but observed that these models fail to provide the desired output format. Furthermore, we did not experience improvements with larger quantized versions such as 33B but found a drop in performance for the 7B parameters model. For all experiments, we use the *Transformers* (Wolf et al., 2020) library, employ 4-bit quantization with normalized floats, and provide one demonstration sample. The full prompt is shown in Appendix B.

[1] https://crisisfacts.github.io

[2] https://grantjenks.com/docs/wordsegment

[3] castorini/monot5-large-msmarco-10k

[4] meta-llama/Llama-2-13b-chat-hf

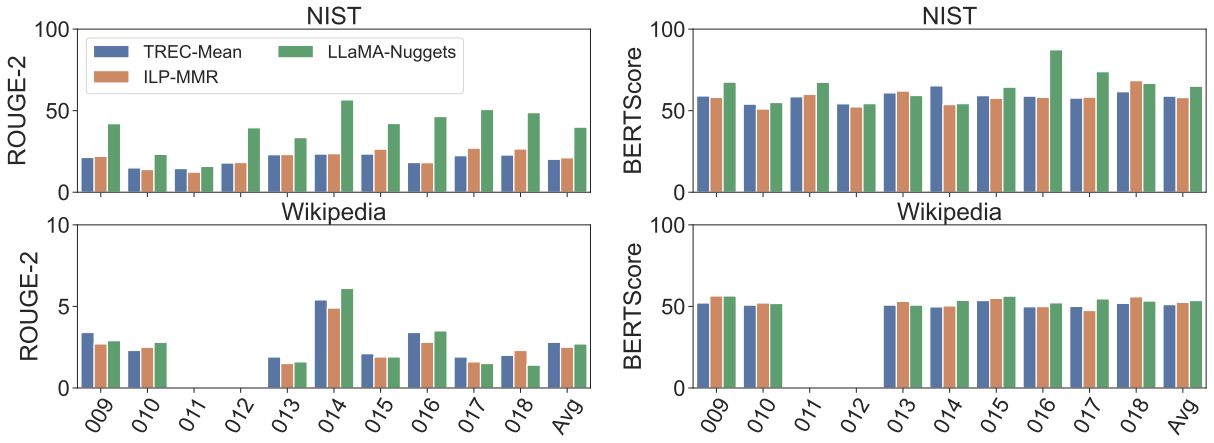

Figure 2: Rouge-2 and BERTScore $F1$-score (x100) results on reference summaries.

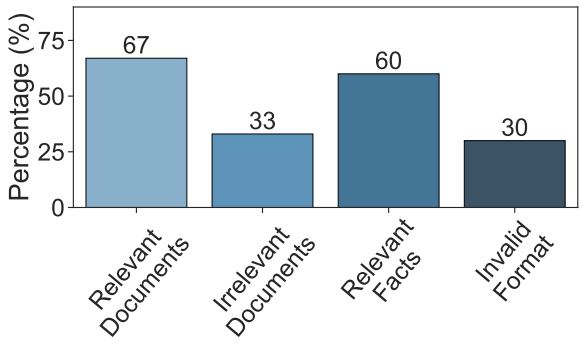

Figure 3: Qualitative analysis results for 30 LLM prompts and responses. **Prompts**: The fraction of prompts that contain at least one relevant or only irrelevant documents. **Responses**: The fraction of responses that contain at least one relevant generated fact or whether the output has an undesired format.

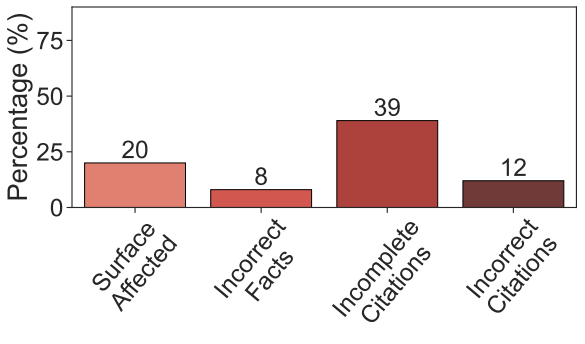

Figure 4: Qualitative analysis results for 59 generated facts. We show the fraction for surface issues, incorrect facts, and incomplete or incorrect citations.

Baselines In addition to the introduced LLAMA-NUGGETS, we consider heuristic and extractive baselines. GREEDY simply selects the top-k documents ordered by importance scores (McCreadie and Buntain, 2023). As extractive model, we include the last year's model ILP-MMR and similarly employ entities[5] as concepts, frequency as weights, and set $L = 150$ for the ILP formulation (Seeberger and Riedhammer, 2022a). For MMR, we set the trade-off parameter $\lambda = 0.8$ and use TF-IDF for cosine similarity.

3.4 Evaluation Metrics

The CrisisFACTS organizers provide the evaluation results covering both automatic and human evaluation metrics. Wikipedia excerpts and NIST summaries (constructed based on as useful annotated meta-facts) serve as gold standard summaries for automatic evaluation. Here, system summaries represent the top-$k = 32$ event nuggets for each event-request pair, which are evaluated with ROUGE-2 and BERTScore $F1$-scores. Regarding human evaluation, the top-$k = 20$ event nuggets are evaluated in terms of comprehensiveness and redundancy, as detailed in Appendix A.

4 Results

Human Evaluation For human evaluation, we show the comprehensiveness and redundancy results in Table 1 and Table 2, respectively. Due to submission count limits for human evaluation, we conduct automatic event nugget matching for the GREEDY and ILP-MMR baselines. That is, we employ the BERTScore model to match the sytems' event nuggets with the meta-facts. However, we observe that only a faction of event nuggets are matched and want to emphasize the unfair com-

[5] https://stanfordnlp.github.io/stanza/ner.html

parison[6]. Nevertheless, we decided to include the results. Our experimental results clearly demonstrate that LLAMA-NUGGETS significantly outperform both the majority of TREC participants' systems as well as extractive baselines. This trend is observed on 10/10 events for comprehensiveness and on 8/10 events for redundancy measures, while the comprehensiveness still indicates relatively low recall. Overall, we present competitive results with a simple but effective LLM-based approach and demonstrate the potential of recent LLMs to improve the summarization of disaster events.

Automatic Evaluation In Figure 2, we present the ROUGE-2 and BERTScore F_1-scores for NIST and Wikipedia event summaries. Note that Wikipedia gold standard summaries are not available for events 011 and 012. On average, LLAMA-NUGGETS outperform the baselines TREC-MEAN and ILP-MMR for all evaluation metrics except ROUGE-2 w.r.t. Wikipedia; we hypothesize that it performs worse due to the nugget format, which is less fluent than, for example, Webnews extracts. Interestingly, we observe that our model performs worse for event 014 in terms of BERTScore but achieve superior results for ROUGE-2. These contrary results can be explained by entity surface form issues or the event nugget generation format. However, the organizers used the BERTScore model DeBERTa[7], which has a token limit of 512, while the event summaries exceed this limit by far. Effectively, the BERTScore metric only evaluates the event nuggets of a subset of requests, which can lead to flawed evaluation results.

Qualitative Analysis We randomly sample 30 event-request-query triple prompts (resulting in 59 generated facts) and qualitatively analyze both on the response and fact levels. We find that 33% of the prompt input documents did not contain any useful query-relevant information, highlighting the importance of noise robustness (Figure 3). Only 60% of the responses include at least one query-relevant fact, while 30% show formatting issues. In Figure 4, we illustrate the errors at the fact level. 8% are incorrect facts (i.e., hallucinations), 39% miss relevant citations, and 12% reference wrong documents. We also check for redundancy issues related to entity surface forms and observe that 20% of the assessed facts are affected.

[6]31.83% for GREEDY and 25.17 % for ILP-MMR.

[7]microsoft/deberta-xlarge-mnli

5 Conclusion

In this work, we present our system for the TREC CrisisFACTS 2023 Track. We combine two-stage retrieval consisting of an efficient sparse retriever and sequence-to-sequence reranker with instruction-following LLM summarization. The experiments show that rather simple prompting approaches surpass extractive baselines and the majority of submitted CrisisFACTS systems. This gives first insights into how openly available LLMs can be used for disaster summarization. However, a qualitative analysis also reveals shortcomings and limitations of the proposed approach. Interesting future directions include a detailed analysis of prompting strategies, the impact of query formulations, and how to address surface form issues.

Acknowledgments

The authors acknowledge the financial support by the Federal Ministry of Education and Research of Germany in the project ISAKI (project number 13N15572).

References

James Allan, Jaime G. Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic Detection and Tracking Pilot Study Final Report. Publisher: Carnegie Mellon University.

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4):357–389.

Javed A. Aslam, Fernando Diaz, Matthew Ekstrand-Abueg, Richard McCreadie, Virgil Pavlu, and Tetsuya Sakai. 2015. TREC 2015 Temporal Summarization Track Overview. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015*, volume 500-319 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Cody L. Buntain, Richard McCreadie, and Ian Soboroff. 2021. Incident Streams 2020: TREC-IS in the Time of COVID-19. In *ISCRAM 2021: 18th International Conference on Information Systems for Crisis Response and Management*.

Marc-André Kaufhold. 2021. *Information Refinement Technologies for Crisis Informatics: User Expectations and Design Principles for Social Media and Mobile Apps*. Springer Fachmedien Wiesbaden, Wiesbaden.

A. Kruspe, J. Kersten, and F. Klan. 2021. Review article: Detection of actionable tweets in crisis events. *Natural Hazards and Earth System Sciences*, 21(6):1825–1845.

Craig Macdonald and Nicola Tonellotto. 2020. Declarative Experimentation in Information Retrieval Using PyTerrier. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, ICTIR '20, pages 161–168, New York, NY, USA. ACM. Event-place: Virtual Event, Norway.

Richard McCreadie and Cody L. Buntain. 2023. CrisisFACTS: Buidling and Evaluating Crisis Timelines. In *Proceedings of the 20th International ISCRAM Conference*, pages 320–339.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

J. Pereira, R. Fidalgo, R. Lotufo, and R. Nogueira. 2023. Crisis Event Social Media Summarization with GPT-3 and Neural Reranking. In *Proceedings of the 20th International ISCRAM Conference*, pages 371–384.

Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. 2018. Social Media in Crisis Management: An Evaluation and Analysis of Crisis Informatics Research. *International Journal of Human–Computer Interaction*, 34(4):280–294.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 851, Raleigh, North Carolina, USA. ACM.

Philipp Seeberger and Korbinian Riedhammer. 2022a. Combining deep neural reranking and unsupervised extraction for multi-query focused summarization. In *Proceedings of the Thirty-First Text REtrieval Conference, TREC 2022, online, November 15-19, 2022*, volume 500-338 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Philipp Seeberger and Korbinian Riedhammer. 2022b. Enhancing crisis-related tweet classification with entity-masked language modeling and multi-task learning. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 70–78, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Royal Sequiera, Luchen Tan, and Jimmy Lin. 2018. Overview of the TREC 2018 Real-Time Summarization Track. In *Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018*, volume 500-331 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. ACL.

A Human Evaluation

The submitted event nuggets for a system and event-request pair are ordered by importance and formed to a summary $\mathcal{S}$ by a rank cut-off k. CrisisFACTS meta-facts are created by deduplicating all pooled participants event nuggets with BERTScore$(\cdot)$ F1-score. Then, each meta-fact is annotated into Useful, Poor, Redundant, and Lagged categories. Based on a bipartite graph of event nuggets connected to the CrisisFACTS meta-facts, the comprehensiveness (i.e., recall) is calculated as

$$\frac{\sum \text{score of adjacent meta-facts}}{|\text{all meta-facts with non-zero score}|} \quad (1)$$

Here the score of a meta-fact is predefined with Useful $= 1.0$, Poor $= 0.0$, Redundant $= 0.5$, and Lagged $= 0.0$. Similarly, the redundancy (i.e., precision) is measured for a system and event-request pair as

$$\frac{\sum \text{score of adjacent meta-facts}}{|\text{all adjacent meta-facts}|} \quad (2)$$

For the final results, all runs are macro-averaged across event-request pairs for an event, and then across all events.

B Prompting

We present the detailed prompt and response in Table 3 and Table 4, respectively.

You are a fact extractor for disaster response organizations. Use the documents to answer the question based on a list of extracted facts as evidence.

Please follow the instructions for the facts:
1. The facts must be short.
2. The format of one fact is text-snippet (source document).
3. Provide the source documents for each fact with the format: (Doc-1, Doc-2, ..)
4. Include fact-relevant entities such as locations, numbers, dates, etc.
5. Only include facts which are focused on the question.
6. The list items must start with * bullet points. Do not use numberings.

We provide you one example within “ marks: ‘{demonstration}‘

Your task
Documents: {documents}
Question: {query}

Facts list:

Table 3: Fact extraction

Your task
Documents: ...
Question: How many firefighters are active?

Facts list:
* 3,300 firefighters are active in fighting the Lilac Fire (Doc-5)
* 808 firefighters are battling the Skirball Fire in Los Angeles (Doc-7)
* 6,946 firefighters are on scene fighting the Thomas Fire in Ventura and Santa Barbara (Doc-8)

Answer: There are approximately 9,154 firefighters actively fighting fires in San Diego and neighboring counties.

Table 4: Example response