Improving Large Language Model Confidence Estimates using Extractive Rationales for Classification

Jane Arleth dela Cruz **Iris Hendrickx** **Martha Larson**
Center for Language and Speech Technology
Center for Language Studies
Radboud University, Nijmegen, Netherlands
{janearleth.delacruz, iris.hendrickx, martha.larson}@ru.nl

Abstract

The adoption of large language models (LLMs) in high-stake scenarios continues to be a challenge due to lack of effective confidence calibration. Although LLMs are capable of providing convincing self-explanations and verbalizing confidence in NLP tasks, they tend to exhibit overconfidence when using generative or free-text rationales (e.g. Chain-of-Thought), where reasoning steps tend to lack verifiable grounding. In this paper, we investigate whether adding explanations in the form of extractive rationales –snippets of the input text that directly support the predictions, can improve the confidence calibration of LLMs in classification tasks. We examine two approaches for integrating these rationales: (1) a one-stage rationale-generation with prediction and (2) a two-stage rationale-guided confidence calibration. We evaluate these approaches on a disaster tweet classification task using four different off-the-shelf LLMs. Our results show that extracting rationales both before and after prediction can improve the confidence estimates of the LLMs. Furthermore, we find that replacing valid extractive rationales with irrelevant ones significantly lowers model confidence, highlighting the importance of rationale quality. This simple yet effective method improves LLM verbalized confidence and reduces overconfidence in possible hallucination.

1 Introduction

Large language models (LLMs) have been shown to achieve state-of-the-art performance on various natural language processing tasks such as classification, information retrieval, summarization, and many more (Raiaan et al., 2024; Lee et al., 2022; Yang et al., 2024). However, the adoption of these LLMs in high-stake scenario tasks continues to be a challenge with their lack of explainability and transparency. Accurately expressing LLMs confidence in their prediction can aid endusers in their decision-making process, i.e., knowing when to trust/not trust. LLMs can verbalize uncertainty and confidence in their prediction but several studies pointed out unsolved issues with these verbalizations (Xiong et al., 2024; Tian et al., 2023; Lin et al., 2022). For example, a recent study has shown that LLMs, when verbalizing their confidence, tend to be overconfident (Xiong et al., 2024), while another study (Tian et al., 2023) found that verbalized confidences emitted as output tokens are typically better calibrated than model's conditional probabilities in certain tasks.

Recent studies demonstrate that integrating explanations with confidence calibration shows promise in language models achieving better calibrated models (Li et al., 2022; Ye and Durrett, 2022a,b; Sachdeva et al., 2024). Li et al. (2022) used token attribution explanations during model training while Ye and Durrett (2022a) utilized feature attribution explanations to train a separate calibrator model. Ye and Durrett (2022b) show that free text explanations generated by the LLMs can be unreliable but still useful to train a separate calibration model. Sachdeva et al. (2024) showed that models trained on counterfactual augmented data improve model calibration and that concise explanations are preferred by calibrator models. However, post-hoc calibrators require additional training data, limiting scalability especially in low-resource settings.

In this paper, we investigate whether LLM prompt-only extractive rationales as explanations improve the confidence calibration of LLMs. Extractive rationales constrain LLMs by anchoring predictions to explicit textual evidence, reducing overconfidence in possible hallucinations. Unlike prior methods that rely on separate training data or post-hoc verifier models, our framework integrates extractive rationales directly into prompting, reducing complexity while maintaining interpretability. We perform both *explain-then-predict* (E $\rightarrow$ P), in

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 549–560

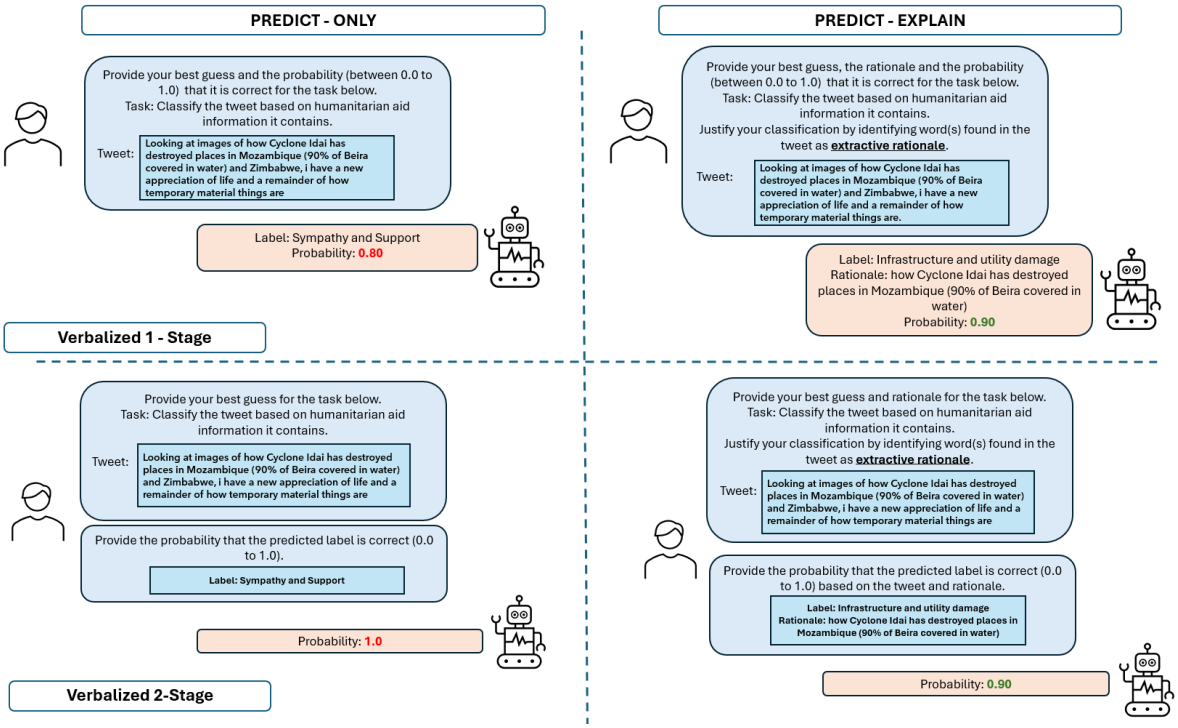

Figure 1: The comparison between approaches of integrating extractive rationales in prediction in the verbalized confidence elicitation, here we show Predict-and-Explain setup, 1-stage (top-right) and 2-stage (bottom-right) and the prediction only 1-stage (top-left) and 2-stage (bottom-left) verbalized confidence elicitation.

which the LLM first generates the rationale explanation and then arrives at a prediction based on it, and *predict-and-explain* (P $\rightarrow$ E), in which the LLM first generates the prediction and provides the rationale, setups in generating these rationales.

We ask the research question: **Do rationales improve LLM confidence score prediction?** We run our experiments for a high stakes scenario setting: disaster risk management. LLMs have the potential to help disaster managers filter through massive amounts of online social media data for relevant, critical, and actionable information during disaster events. With the goal of helping disaster managers, we are focused on commonly available LLMs that allow the disaster managers independence from a complex pipeline and the maintenance it implies.

We investigate two approaches for integrating rationales when eliciting confidence estimates as shown in Figure 1: (1) Verbalized 1-stage: asking the LLM for the rationale along with the predicted label and confidence score. This approach minimizes computational overhead aligning with our task, maintains coherence and intrinsic connection with the rationale and predicted label, reducing context fragmentation, and (2) Verbalized 2-stage: asking the rationale and label first, then, afterwards, in the separate prompt adding the rationale to get the confidence score. This approach decouples the separate tasks of rationale generation and prediction with confidence estimation, allowing independent verification, akin to having a separate calibration model. We run our experiments using both closed and open-sourced off-the-shelf LLMs: gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1 8B-Instruct (Llama Team, 2024), mistral 7B-Instruct v0.3 (Jiang et al., 2023) across a humanitarian aid information type classification task (Alam et al., 2021).

Our key contributions are as follows:

- We demonstrate that integrating explanations in the form of extractive rationales improves confidence calibration in off-the-shelf LLMs for classification.
- We show, via ablation with "bad" (irrelevant) rationales, the necessity of rationale quality for effective calibration.

Related Work. Model explanations have been used for calibration post-hoc by training separate verifier or calibrator models (Li et al., 2022; Ye and Durrett, 2022b,a; Xu et al., 2024; Sachdeva et al., 2024). Unlike the separate post-hoc calibrators from (Li et al., 2022; **?**; Ye and Durrett, 2022b), our prompt-based approach requires no additional

training, making it suitable for off-the-shelf LLMs and various disaster event types. The previous way to measure confidence in model predictions rely on model's internal logits but this has become less suitable with off-the-shelf decoder-only LLMs. This led to methods of prompting LLMs themselves to express uncertainty in natural language which is referred to as verbalized confidence (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024) where Xiong et al. (2024) found that LLMs are prone to overconfidence when generating free text explanations while Tian et al. (2023) observed better calibration when confidence is explicitly verbalized. Closely related to our study, Zhao et al. (2024) proposed a prompt-based approach to improving calibration asking for "facts" and "reflection" from the LLM while Zhang et al. (2024) proposed fidelity elicitation techniques, which are both relevant for multipurpose QA tasks but may not be as suited for classification task.

Our method addresses three gaps from prior work, overconfidence in generative rationales, the need for lightweight calibration, and trust. By integrating extractive rationales directly into prompting, we show how minimal architectural changes can yield calibration improvements.

2 Method

Problem Definition. LLMs have been very effective in various natural language tasks. However, adoption of LLMs in high-stake scenarios continues to be a challenge due to LLMs tend to exhibit overconfidence when using generative or free-text rationales (e.g. Chain-of-Thought (CoT) prompting), where reasoning steps tend to lack verifiable grounding. We attempt to mitigate this problem by constraining LLMs to extractive rationales, snippets of the input where we aim to reduce hallucination rate by anchoring the predictions to observable evidence.

LLM as Disaster Tweet Classifier. We test the performance of LLMs as disaster tweet classifiers for humanitarian aid information classification. We allow the LLM to generate a prediction label for a tweet and the corresponding rationale. We perform both explanation setups studied by Camburu et al. (2018) to create their finetuned explainers, *explain-then-predict* (E $\rightarrow$ P), in which the LLM first generates the rationale explanation and then arrives at a prediction based on it, and *predict-and-explain* (P $\rightarrow$ E), in which the LLM first generates the prediction and provides the rationale. We used the predict-only setup as the baseline classifier.

Confidence Elicitation Methods. We utilize methods that extract **confidence scores** through verbalization (Lin et al., 2022; Tian et al., 2023), particularly where the model expresses confidence in token space with numerical probabilities. We adopted two of the best performing prompts from Tian et al. (2023)'s study, **Verb 1S top-1** and **Verb 2S top-1**. **Verb 1S top-1** prompts the model to produce one guess, (the prediction and rationale, and a probability that the prediction is correct in a single response (1-"stage") (Tian et al., 2023). **Verb 2S top-1** uses numerical probabilities similarly, except the model is first asked only for its answers and then asked to assign the probabilities of correctness to each answer (2-"stages") (Tian et al., 2023). The exact prompts used are found in Appendix A.4. CoT prompting methods were no longer explored as multiple studies (Tian et al., 2023; Zhao et al., 2024) have shown that this does not improve calibration, even degrading instance-level calibration.

We examine whether the extractive rationales are being used to improve the LLM calibration for the **Verb 2S top-1** prompt, we replace them with irrelevant rationales and measure the changes in confidence estimates. We explore two "bad" rationale variants, *non-rationales* - random phrases (of similar length to original rationales) that do not include any of the original rationale explanation words selected and *diff-task rationales* - rationales that were extracted from a different disaster tweet classification task, where some words may overlap with the original rationales. The different task we used was the type of help-seeking tweet classification: identifying whether a tweet expresses need for instrumental or emotional help in a disaster scenario (Encarnación and Wilks, 2023).

3 Experimental Setup

3.1 Dataset

We utilized human-annotated crisis-related tweets from (Alam et al., 2021). The original dataset had 11 labels, however, we limited our labels to the five that were present in all of our selected crisis events, following (Zou et al., 2023) who also reduced their labels. First, we experimented with including the labels: 'other relevant information' and 'not humanitarian', however, the results showed the generated rationales for these labels tend to be the entire tweet themselves. We sampled 300 tweets

for each of ten different disaster events, i.e., a total of 3000 tweets. More information about the data is in Appendix A.2.

3.2 Models

We chose commonly used off-the-shelf LLMs in our experiments. We used gpt-4o-mini (OpenAI, 2024a), gpt-4o (OpenAI, 2024b), llama 3.1-8B Instruct (Llama Team, 2024), and mistral 7B-Instruct (Jiang et al., 2023). These models were chosen because they are commonly used by both researchers and the public. We ran our experiments at the temperature setting of 0.0 to make all models deterministic, fit for a classification task. More model details are found in Appendix A.1.

3.3 Evaluation Metrics

We evaluate the quality of the confidence classifier outputs using calibration error metrics. Calibration evaluates how well model's confidence aligns with its accuracy, where a well-calibrated model assigns 90% confidence to an answer, then the answer is correct 90% of the time.

Expected Calibration Error (ECE) is calculated as the weighted average of the discrepancies between the mean predicted probability and the actual accuracy across all bins.

Static Calibration Error (SCE) - is a simple extension of ECE to every probability in the multiclass setting. SCE bins for each class probability, and computes the error within the bin and averages across the bin (Nixon et al., 2019).

Adaptive Calibration Error (ACE) – suggests that in order to get the best estimate of the overall calibration error the metric should focus on the regions where the predictions are made. Each bin has equal number of spaces (Nixon et al., 2019).

Model	Prompt	Accuracy	F1-score
gpt-4o-mini	Predict only	0.884	0.884
	E → P (ours)	0.888	0.889
	P → E (ours)	0.896	0.897
gpt-4o	Predict only	0.911	0.911
	E → P (ours)	0.916	0.914
	P → E(ours)	0.922	0.923

Table 1: Model performance evaluated in the experiments across all 10 disaster events. Results shown are from top 2 performing models.

4 Results

4.1 Classification Performance

We show the classification performance on our set of 3000 disaster tweets of classifier prompt gpt-4o-mini and gpt-4o setups in Table 1. The other similar results can be found in Appendix B.2 for the rest of the LLMs evaluated. Asking the model for rationale explanation during prediction does not hurt the performance of the model in general for our classification task, all are comparable with the predict only baseline. The *predict-and-explain* setup is the highest performing classifier at 92.2 Accuracy for gpt-4o.

4.2 Confidence Score Results

Table 2 shows the results of evaluating the prompt methods for extracting confidence across gpt-4o-mini and llama 3.1-8B Instruct. Similar results can be found in Appendix B.2 for the rest of the LLMs evaluated. Only Mistral had calibration error that was subpar compared to the other three LLMs evaluated. We observe that by asking for rationale-based explanations –in both our prompt setups, *explain-then-predict* (E → P) and *predict-and-explain* (P → E), LLMs can produce better calibrated confidences. Both E → P and P → E setups have lower calibration error scores than the baseline predict only in both Verb 1S and Verb 2S methods.

To evaluate whether these rationales are indeed improving the LLM calibration, we ran experiments where we replaced the original rationales

Model	Prompt	ECE ↓	SCE ↓	ACE ↓
Verb 1S				
gpt-4o-mini	Predict only	0.063	0.041	0.114
	E → P (ours)	0.036	0.037	0.082
	P → E(ours)	0.050	0.035	0.088
llama 3.1	Predict only	0.075	0.046	0.149
	E → P (ours)	0.065	0.048	0.143
	P → E (ours)	0.056	0.040	0.125
Verb 2S				
gpt-4o-mini	Predict only	0.069	0.041	0.167
	E → P (ours)	0.035	0.039	0.070
	P → E(ours)	0.039	0.036	0.066
llama 3.1	Predict only	0.050	0.059	0.099
	E → P (ours)	0.041	0.052	0.092
	P → E(ours)	0.046	0.040	0.091

Table 2: Calibration error metrics of the various confidence verbalization methods across prompts. ECE is the expected calibration error, SCE is the static calibration error and ACE is the adaptive calibration error. Results shown are top 2 most calibrated models (based on ACE).

Prompt	Rationale	ECE ↓	SCE ↓	ACE ↓
E → P	original (ours)	**0.035**	**0.039**	**0.070**
	non-rationale	0.074	0.044	0.146
	diff-rationale	0.048	0.039	0.095
P → E	original (ours)	**0.039**	**0.036**	**0.066**
	non-rationale	0.059	0.039	0.116
	diff-rationale	0.053	0.037	0.091

Table 3: Calibration error metrics when changing the rationale type. ECE is the expected calibration error, SCE is the static calibration error and ACE is the adaptive calibration error. Results shown are for gpt-4o-mini.

and asked for new confidence estimates. Table 3 shows the confidence metrics for the different rationales used. Using the LLMs' original rationale produces the best calibrated confidences. Using the *non-rationales*, which are the phrases that have no overlap with our original rationales, show the least calibrated confidence scores. The *diff-task rationales*, on the other hand, can have words that overlap and some labels can have similar rationales, i.e., 'Sympathy and support' from the original task and 'seeking emotional help' from the different task, and 'Rescue, volunteering or donation effort and 'seeking instrumental help', so it produced better calibrated scores from *non-rationales*. These results confirm that the relevance of the rationale and not only the mere presence drives the improvement in calibration.

5 Discussion & Conclusion

In this paper, we proposed integrating extractive rationale explanations with the predictions to improve LLM confidence calibration in classification tasks. First, we test whether these extractive rationales hurt classification performance. We found that this approach has slightly higher to similar performance compared to the predict-only baseline, contrary to findings from Huang et al. (2023)'s prompting setup with feature attribution as explanation and Camburu et al. (2018)'s supervised training method. Our results show that LLMs can express confidence in numerical probabilities better by asking for rationale-based explanations for both before (*explain-then-predict*) and after (*predict-and-explain*) predictions than direct predict-only prompt. We showed that improvement is achieved in the two confidence verbalization strategies investigated, Verb 1S and Verb 2S. In the Verb 2S setting, replacing the extractive rationales with "bad" rationales, non-rationales that have no overlap with the original and diff-task rationales that are from a different classification task, hurt the LLM confidence scores,thus, showing that the original rationales are relevant to the LLM calibration. However, we note that this finding for the Verb 2S setting is not applicable to the Verb 1S setting. Our results show that our method offers a lightweight alternative to complex pipelines while maintaining interpretability.

6 Limitations

A key limitation of our framework is that it is only applicable for classification task where extractive rationale explanations are applicable. With tasks where input lacks extractable rationales e.g., LLM selects entire input as rationale our approach would not be suitable. We only evaluated off-the-shelf LLMs: gpt-4o-mini, gpt-4o, llama and mistral. We only evaluated on the base or instruct models; we did not finetune. Instruction-tuning/fine-tuning these models may lead to more favorable results. Our use case has a limited scope as we focused on one classification task for disaster risk management with only English tweets.

7 Ethical Considerations

The datasets used in this paper were from publicly available datasets (Alam et al., 2021) which were collected tweets from X (previously, Twitter) using the platform's streaming API in line with its terms of service.

Our work aspires ultimately to support disaster management in high-stakes scenarios. As such, a potential risk is that readers misinterpret the readiness of the technology for use by disaster managers, and move either too quickly to uptake without guarantees of reliability or pre-maturely abandon the type of solutions we study. We have attempted to address this point by stating clearly our **negative result** (i.e., LLMs struggle with long-context set selection) and stating that we find human-LLM collaborations may still hold future potential.

Acknowledgments

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO).

References

Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. 2021. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):933–942.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Trilce. Encarnación and Chelsey R. Wilks. 2023. Role of expressed emotions on the retransmission of help-seeking messages during disasters. In *Proceedings of the 20th International ISCRAM Conference*, pages 340–352, Omaha, USA. University of Nebraska at Omaha.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. 2023. Can large language models explain themselves? a study of llm-generated self-explanations. *Preprint*, arXiv:2310.11207.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Dongfang Li, Baotian Hu, and Qingcai Chen. 2022. Calibration meets explanation: A simple and effective approach for model confidence estimates. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2784, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019. Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

OpenAI. 2024a. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. Gpt-4o system card.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Rachneet Sachdeva, Martin Tutek, and Iryna Gurevych. 2024. CATfOOD: Counterfactual augmented training for improving out-of-domain performance and calibration. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1876–1898, St. Julian's, Malta. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Trans. Knowl. Discov. Data*, 18(6).

Xi Ye and Greg Durrett. 2022a. Can explanations be useful for calibrating black box models? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6199–6212, Dublin, Ireland. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022b. The unreliability of explanations in few-shot prompting for textual reasoning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. 2024. Calibrating the confidence of

large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2959–2979, Miami, Florida, USA. Association for Computational Linguistics.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-and-reflection (FaR) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718, Bangkok, Thailand. Association for Computational Linguistics.

Henry Peng Zou, Yue Zhou, Cornelia Caragea, and Doina Caragea. 2023. Crisismatch: Semi-supervised few-shot learning for fine-grained disaster tweet classification. *CoRR*, abs/2310.14627.

A Appendix A

A.1 Models

Table 4 contains the information about the versions of the 4 LLMs we evaluated and analyzed.

A.2 Datasets

We utilized human-annotated crisis-related tweets from (Alam et al., 2021). We sampled across four different disaster types: earthquake, hurricane, wildfire and flood. We chose the event with the highest inter-annotator agreement per disaster type based on (Alam et al., 2021). The original dataset had 11 labels, however, we limited our labels to the 5 that were present in all of our selected crisis events, following (Zou et al., 2023) who also reduced their labels to 7. Originally, we experimented with including the labels: other relevant information and not humanitarian, however, this seemed to be too challenging for the LLM. The humanitarian aid information labels are as follows:

- **Caution and advice**: Reports of warnings issued or lifted, guidance and tips related to the disaster;
- **Infrastructure and Utility Damage**: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
- **Injured or dead people**: Reports of injured or dead people due to the disaster;
- **Rescue, volunteering, or donation effort**: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, people in shelter facilities, donation of money, or services, etc.;
- **Sympathy and support**: Tweets with prayers, thoughts, and emotional support;

We sampled the test sets of the following crisis events: Canada Wildfires 2016, Cyclone Idai 2019, Greece Wildfires 2018, Mexico Earthquake 2017, Hurricane Matthew 2016, Hurricane Harvey 2017, Hurricane Maria 2017, Italy Earthquake 2016, Maryland Floods 2018, and Sri Lanka Floods 2017. We randomly sampled 300 tweets for each disaster event.

A.3 Evaluation Metrics

Confidence Metrics. To evaluate the quality of the confidence classifier outputs, two tasks are typically employed: calibration and failure prediction (Xiong et al., 2024). Calibration evaluates how well model's confidence aligns with its actual accuracy, the basic idea being that if a well-calibrated model assigns 90% confidence to an answer, then the answer is correct 90% of the time. Failure prediction, on the other hand, measures the model's capacity to assign higher confidence to correct predictions and lower confidence to incorrect ones.

Expected Calibration Error (ECE) - approximates it by clustering instances with similar confidence. The predicted probabilities are put into bins, and ECE is calculated as the weighted average of the discrepancies between the mean predicted probability and the actual accuracy across all bins.

$$\text{ECE} = \sum_{b=1}^{N} \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)|$$

where n_b is the number of predictions in bin b, N is the total number of data points and $acc(b)$ and $conf(b)$ are the accuracy and confidence of bin b, respectively. One drawback of ECE is its sensitivity to bucket width and the variance of the samples within these buckets.

Static Calibration Error (SCE) - is a simple extension of ECE to every probablity in the multicliss setting. SCE bins for each class probability, and computes the error within the bin and averages across the bin (Nixon et al., 2019).

$$\text{SCE} = \frac{1}{K} \sum_{k=1}^{K} \sum_{b=1}^{B} \frac{n_{bk}}{N} |\text{acc}(b,k) - \text{conf}(b,k)|$$

Table 4: Information of evaluated and analyzed LLMs

Model	Type	Source (OpenAI/Huggingface)
gpt-4o-mini	closed	gpt-4o-2024-08-06
gpt-4o	closed	gpt-4o-mini-2024-07-18
llama 3.1 - 8B Instruct	open	meta-llama/Meta-llama-3.1-8B-Instruct
mistral 7B Instruct v0.3	open	mistralai/mistral-7B-Instruct-v0.3

Here, $\text{acc}(b, k)$ and $\text{conf}(b, k)$ are the accuracy and confidence of bin b for class label k respectively, n_{bk} is the number of predictions in bin b for class kl and N is the total number of data points.

Adaptive Calibration Error (ACE) – suggests that in order to get the best estimate of the overall calibration error the metric should focus on the regions where the predictions are made. Each bin has equal number of spaces(Nixon et al., 2019).

$$\text{ACE} = \frac{1}{KR} \sum_{k=1}^{K} \sum_{r=1}^{R} |\text{acc}(r, k) - \text{conf}(r, k)|$$

Here, $\text{acc}(r, k)$ and $\text{conf}(r, k)$ are the accuracy and confidence of calibration range r for class label k respectively, N is the total number of data points. Calibration r is defined by the N/Rth index of the sorted and thresholded predictions (Nixon et al., 2019).

A.4 Prompts

The different prompt used for classification and confidence verbalization are found in Figures 2, 5, and 6. The prompt used to find the different task rationales is in Figure 7.

B Appendix B

B.1 Output Examples of LLM Disaster Classifiers

We present some sample output from our LLM classifier prompts in figure 8.

B.2 Additional Results

We present all additional results in Tables 5, 6, 7, and 8.

```
Provide your best guess and probability that it is correct (0.0 to 1.0) for the following task. Respond with only the label and probability. Do not modify or explain.

For example:
Label: <most likely label>
Probability: <the probability between 0.0 to 1.0 that your classification is correct, without any extra commentary whatsoever; just the probability.>

Task: You will be provided a tweet. Based on the tweet's content, assign one of the following labels related to humanitarian aid that best fits the information provided.
Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Injured or dead people: Reports of people injured or dead due to the disaster;
Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
Rescue, volunteering, or donation effort: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, donation of money, or services, etc.;
Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;
Sympathy and support: Tweets with prayers, thoughts, and emotional support;

Do not modify or explain.

Tweet: {tweet}
```

Figure 2: Prompt used for Predict Only classification for Verb 1S.

```
Provide your best guess, and probability that it is correct (0.0 to 1.0) for the following task. Respond with only the label, rationale, and probability. Do not modify or explain.

For example:
Label: <most likely label>
Rationale: <snippet found in input that justifies your classification, copy in verbatim>
Probability: <the probability between 0.0 to 1.0 that your classification is correct, without any extra commentary whatsoever; just the probability.>

Task: You will be provided a tweet. Based on the tweet's content, assign one of the following labels related to humanitarian aid that best fits the information provided.
Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Injured or dead people: Reports of people injured or dead due to the disaster;
Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
Rescue, volunteering, or donation effort: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, donation of money, or services, etc.;
Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;
Sympathy and support: Tweets with prayers, thoughts, and emotional support;

Justify your classification by identifying the corresponding word(s) found in the tweet as rationale.
Your chosen rationale must be a snippet of the tweet. Copy in verbatim. Do not modify or explain.

Tweet: {tweet}
```

Figure 3: Prompt used for Predict-and-Explain classification for Verb 1S.

```
Provide your best guess, and probability that it is correct (0.0 to 1.0) for the following task.
Respond with only the rationale, label and probability. Do not modify or explain.

For example:
Rationale: <snippet found in input that justifies your classification, copy in verbatim>
Label: <most likely label>
Probability: <the probability between 0.0 to 1.0 that your classification is correct, without any extra commentary whatsoever; just the probability.>

The task is: You will be provided a tweet. Based on the tweet's content, identify the corresponding word(s) found in the tweet that justify the classification as rationale. Your chosen rationale must be a snippet of the tweet. Copy in verbatim. Do not modify. Assign one of the following labels related to humanitarian aid that best fits the rationale provided.
Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Injured or dead people: Reports of people injured or dead due to the disaster;
Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
Rescue, volunteering, or donation effort: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, donation of money, or services, etc.;
Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;
Sympathy and support: Tweets with prayers, thoughts, and emotional support;

Do not modify or explain.

Tweet: {tweet}
```

Figure 4: Prompt used for Explain-then-Predict classification for Verb 1S.

```
Provide your best guess for the following task.
Respond with only the label and rationale. Do not modify or explain.

For example:
Label: <most likely label>
Rationale: <snippet found in input that justifies the classification, copy in verbatim>

The task is: You will be provided a tweet. Based on the tweet's content, assign one of the following labels related to humanitarian aid that best fits the information provided.
    Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Injured or dead people: Reports of people injured or dead due to the disaster;
Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
Rescue, volunteering, or donation effort: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, donation of money, or services, etc.;
Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;
Sympathy and support: Tweets with prayers, thoughts, and emotional support;

Justify your classification by identifying the corresponding word(s) found in the tweet as rationale.

Your chosen rationale must be a snippet of the tweet. Copy in verbatim. Do not modify or explain.

Tweet: {tweet}
```

```
Provide the probability that the label is correct (0.0 to 1.0) based on the tweet and rationale. Respond with only the and probability. Do not modify or explain.
Label: {label}
Rationale: {rationale}

For example:
Probability: <the probability between 0.0 to 1.0 that the label is correct, without any extra commentary whatsoever; just the probability.>
```

Figure 5: Prompt used for Predict-and-Explain classification for Verb 2S.

```
Provide your best guess for the following task.
Respond with only with the rationale and label. Do not modify or explain.

For example:
Rationale: <snippet found in input that explains the classification, copy in verbatim>
Label: <most likely label>

The task is: You will be provided a tweet. Based on the tweet's content, identify the corresponding word(s) found in the tweet that justify the classification as rationale. Your chosen rationale must be a snippet of the tweet. Copy in verbatim. Do not modify. Then, assign one of the following labels related to humanitarian aid that best fits the information provided.
Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Injured or dead people: Reports of people injured or dead due to the disaster;
Infrastructure and utility damage: Reports of any type of damage to infrastructure such as buildings, houses, roads, bridges, power lines, communication poles, or vehicles;
Rescue, volunteering, or donation effort: Reports of any type of rescue, volunteering, or donation efforts such as people being transported to safe places, people being evacuated, people receiving medical aid or food, donation of money, or services, etc.;
Caution and advice: Reports of warnings issued or lifted, guidance and tips related to the disaster;
Sympathy and support: Tweets with prayers, thoughts, and emotional support;

Do not modify or explain.
```

```
Provide the probability that the label is correct (0.0 to 1.0) based on the tweet and rationale. Respond with only the and probability. Do not modify or explain.

Rationale: {rationale}
Label: {label}

For example:
Probability: <the probability between 0.0 to 1.0 that the label is correct, without any extra commentary whatsoever; just the probability.>
```

Figure 6: Prompt used for Explain-then-Predict classification for Verb 2S.

```
Provide your best guess, and probability that it is correct (0.0 to 1.0) for the following task.
Respond with only the label, rationale, and probability. Do not modify or explain.

For example:
Label: <most likely label>
Rationale: <snippet found in input that justifies your classification, copy in verbatim>
Probability: <the probability between 0.0 to 1.0 that your classification is correct, without any extra commentary whatsoever; just the probability.>

The task is: You will be provided a tweet. Based on the tweet's content, assign one of the following labels related to humanitarian aid that best fits the information provided.
Select only one label, even if multiple labels seem to apply. Choose the label based on the severity and need for immediate response.

Labels:
Instrumental help: Messages where the individual who posted sought or seeks help or assistance tangibly or physically, such as shelter, food, or other basic needs;
Emotional help: Messages that seek care or compassion and when tweets express emotional needs or distress;

Justify your classification by identifying the corresponding word(s) found in the tweet as rationale.
Your chosen rationale must be a snippet of the tweet. Copy in verbatim. Do not modify or explain.

Tweet: {tweet}
```

Figure 7: Prompt used to create different task rationale (type of help-seeking message classification)

Model	Prompt	Accuracy	F1-score
gpt-4o-mini	Predict only	0.884	0.884
	E → P (ours)	0.888	0.889
	P → E (ours)	0.896	0.897
gpt-4o	Predict only	0.911	0.911
	E → P (ours)	0.916	0.914
	P → E (ours)	0.922	0.923
llama 3.1 - 8B	Predict only	0.810	0.819
	E → P (ours)	0.821	0.836
	P → E (ours)	0.845	0.846
mistral 7B	Predict only	0.733	0.746
	E → P (ours)	0.801	0.800
	P → E (ours)	0.801	0.799

Table 5: Model performance evaluated in the experiments across all 10 disaster events.

Tweet	True Label	Prompt	Predicted Label	Probability	Predicted Rationale
Looking at images of how Cyclone Idai has destroyed places in Mozambique (90% of Beira covered in water) and Zimbabwe, i have a new appreciation of life and a remainder of how temporary material things are	Infrastructure and utility damage	Predict only	Sympathy and support	0.80	--
		Explain – Predict	Infrastructure and utility damage	0.85	Cyclone Idai has destroyed places in Mozambique (90% of Beira covered in water)
		Predict - Explain	Infrastructure and utility damage	0.90	how Cyclone Idai has destroyed places in Mozambique (90% of Beira covered in water)
RT @USER: Imagine the worst storm ever and no cell phones, no water, no power, no roads. It's just unfathomable. Horrible. /	Infrastructure and utility damage	Predict only	Caution and Advice	0.70	--
		Explain – Predict	Infrastructure and utility damage	0.85	no cell phones, no water, no power, no roads
		Predict - Explain	Infrastructure and utility damage	0.80	no cell phones, no water, no power, no roads
RT @USER: after seeing how destroyed Haiti is after this hurricane, how can you be excited for one? lmao	Rescue, volunteering, or donation effort	Predict only	Sympathy and support	0.80	--
		Explain – Predict	Rescue, volunteering, or donation effort	0.85	NGOs are doing their best to bring relief to the people!!
		Predict - Explain	Rescue, volunteering, or donation effort	0.90	NGOs are doing their best to bring relief to the people!!

Figure 8: Example Outputs where Predict-Only Prompt fails in its prediction. Results shown are with gpt-4o-mini

Model	Prompt	ECE ↓	SCE ↓	ACE ↓
gpt-4o-mini	Predict only	0.063	0.041	0.114
	E → P (ours)	0.036	0.037	0.082
	P → E (ours)	0.050	0.035	0.088
gpt-4o	Predict only	0.081	0.039	0.157
	E → P (ours)	0.048	0.029	0.096
	P → E (ours)	0.063	0.029	0.128
llama-3.1 8B	Predict only	0.075	0.046	0.149
	E → P (ours)	0.065	0.048	0.143
	P → E (ours)	0.056	0.040	0.125
mistral 7B	Predict only	0.223	0.082	0.446
	E → P (ours)	0.171	0.062	0.340
	P → E (ours)	0.171	0.062	0.341

Table 6: Calibration error metrics of the various confidence verbalization methods across prompts. ECE is the expected calibration error, SCE is the static calibration error and ACE is the adaptive calibration error. Results shown are from Verb 1S method.

Event	Prompt	Accuracy	F1-score
All Events	Predict only	0.884	0.884
	E → P (ours)	0.888	0.889
	P → E (ours)	0.896	0.897
Canada Wildfires	Predict only	0.887	0.890
	E → P (ours)	0.910	0.917
	P → E (ours)	0.917	0.916
Cyclone Idai	Predict only	0.867	0.863
	E → P (ours)	0.867	0.864
	P → E (ours)	0.873	0.869
Greece Wildfires	Predict only	0.863	0.860
	E → P (ours)	0.837	0.831
	P → E (ours)	0.873	0.870
Hurricane Harvey	Predict only	0.880	0.882
	E → P (ours)	0.887	0.886
	P → E (ours)	0.880	0.880
Hurricane Maria	Predict only	0.900	0.902
	E → P (ours)	0.913	0.911
	P → E (ours)	0.913	0.913
Hurricane Matthew	Predict only	0.883	0.879
	E → P (ours)	0.913	0.911
	P → E (ours)	0.917	0.914
Italy Earthquake	Predict only	0.903	0.905
	E → P (ours)	0.887	0.893
	P → E (ours)	0.910	0.912
Maryland Floods	Predict only	0.853	0.853
	E → P (ours)	0.857	0.858
	P → E (ours)	0.870	0.870
Mexico Earthquake	Predict only	0.910	0.909
	E → P (ours)	0.910	0.909
	P → E (ours)	0.917	0.914
Sri Lanka Floods	Predict only	0.893	0.901
	E → P (ours)	0.900	0.909
	P → E (ours)	0.910	0.916

Table 7: Model performance evaluated in the experiments for every disaster event. Results shown are from gpt-4o-mini Verb 1S method

Event	Prompt	ECE ↓	SCE ↓	ACE ↓
All Events	Predict only	0.063	0.041	0.114
	E → P (ours)	0.036	0.037	0.082
	P → E (ours)	0.050	0.035	0.088
Canada Wildfires	Predict only	0.105	0.066	0.210
	E → P (ours)	0.071	0.041	0.136
	P → E (ours)	0.086	0.037	0.171
Cyclone Idai	Predict only	0.076	0.046	0.182
	E → P (ours)	0.069	0.041	0.191
	P → E (ours)	0.054	0.044	0.175
Greece Wildfires	Predict only	0.033	0.045	0.106
	E → P (ours)	0.055	0.056	0.130
	P → E (ours)	0.021	0.043	0.130
Hurricane Harvey	Predict only	0.046	0.044	0.117
	E → P (ours)	0.045	0.036	0.118
	P → E (ours)	0.045	0.038	0.090
Hurricane Maria	Predict only	0.077	0.042	0.152
	E → P (ours)	0.050	0.032	0.112
	P → E (ours)	0.053	0.032	0.111
Hurricane Matthew	Predict only	0.070	0.042	0.107
	E → P (ours)	0.050	0.033	0.113
	P → E (ours)	0.052	0.033	0.102
Italy Earthquake	Predict only	0.032	0.037	0.129
	E → P (ours)	0.026	0.040	0.100
	P → E (ours)	0.029	0.031	0.063
Maryland Floods	Predict only	0.037	0.045	0.121
	E → P (ours)	0.011	0.049	0.068
	P → E (ours)	0.017	0.045	0.115
Mexico Earthquake	Predict only	0.074	0.037	0.148
	E → P (ours)	0.047	0.037	0.110
	P → E (ours)	0.063	0.031	0.131
Sri Lanka Floods	Predict only	0.100	0.049	0.183
	E → P (ours)	0.067	0.042	0.119
	P → E (ours)	0.079	0.044	0.144

Table 8: Calibration Error Metrics for all the disaster events. ECE is the expected calibration error, SCE is the static calibration error and ACE is the adaptive calibration error. Highlight indicates when the rationale prompt method does not outperform the Predict only baseline. Results shown are from gpt-4o-mini Verb 1S method.