Information Research - Vol. 30 No. iConf (2025)

A benchmark for evaluating crisis information generation capabilities in LLMs

Ruilian Han, Lu An, Wei Zhou, and Gang Li
DOI: https://doi.org/10.47989/ir30iConf47518

Abstract

Introduction. Large language models (LLMs) have become increasingly significant in crisis information management due to their advanced natural language processing capabilities. This study aims to develop a comprehensive evaluation benchmark to assess the effectiveness of LLMs in generating crisis information.

Method. CIEeval, an evaluation dataset, was constructed through steps such as information extraction and prompt generation. CIEeval covers 26 types of crises across sub-domains including water disasters, environmental pollution, and others, comprising a total of 4.8k data entries.

Analysis. Eight LLMs applicable to the Chinese context were selected for evaluation based on multidimensional criteria. A combination of manual and machine scoring methods was utilized. This approach ensured a comprehensive understanding of each model's performance.

Results. The manual and machine scores showed significant correlation. Under this scoring method, Claude 3.5 Sonnet performed the best, particularly excelling in complex scenarios like natural and accident disasters. In contrast, while scoring slightly lower overall, Chinese models like ERNIE 4.0 Turbo and iFlytek Spark V4.0, showed strong performance in specific crises.

Conclusion. The evaluation benchmark validates the best LLM for crisis information generation (Claude 3.5 Sonnet) and provides valuable insights for LLMs to optimize and apply LLM in crisis information.

Introduction

In the context of rapid technological development, large language models (LLMs) have become a key focus in natural language processing due to their application of deep learning and Transformer architecture. LLMs efficiently encode and decode language, demonstrating significant advancements in machine translation, text generation, and dialogue systems (Guler et al., 2024).

Crisis information (CI) is increasingly important in crisis response. Traditional manual process is inefficient and hard to manage large-scale and frequent events. LLMs, with their powerful language generation capabilities, can automatically produce high-quality crisis information, improving efficiency and accuracy. Additionally, LLMs handle multilingual and multimodal information, facilitating cross-regional and cross-domain information sharing. However, the information generated by different LLMs varies greatly in quality, and may exhibit over-interpretation of information or thematic bias, both of which may affect the effectiveness of decision-making.

This study aims to establish benchmarks to comprehensively evaluate LLMs' performance in CI generation across scenarios like natural disasters, accidents, and public health incidents. By assessing objectivity, completeness, and reasonableness, this research provides insights into LLMs' strengths and limitations, offering guidance for model optimization and practical applications in CI.

Literature review

In recent years, several specialized benchmark datasets have been developed to assess the performance of LLMs, such as GLUE (Wang et al., 2019), SuperGLUE (Sarlin et al., 2020), and SoEval (Liu et al., 2024). These benchmarks cover tasks like reading comprehension, sentiment analysis, and structured output, becoming essential tools for evaluating LLMs. More complex benchmarks like BIG-Bench Hard have been introduced recently, facilitating the evaluation of LLMs' cross-task and cross-domain generalization capabilities (Suzgun et al., 2022). Evaluation benchmarks in the Chinese context are also emerging, such as CLUE (Xu et al., 2020), which has become a widely used evaluation tool in various industries in China.

In the information science field, LLMs are widely applied in automating information analysis (Giannakopoulos et al., 2023), processing, and public opinion monitoring. As LLMs are increasingly used, researchers have established evaluation frameworks for information compilation and report generation. For example, Thelwall (2024) assessed LLMs' effectiveness in scientific information evaluation, and explored ways to enhance their capabilities through prompt engineering and external tools.

For evaluation methods, while accuracy remains an important measure of LLMs' objectivity, automated metrics like BLEU (Evtikhiev et al., 2023) and Bipol (Alkhaled et al., 2023) have limitations in addressing the feasibility of generated content for open-ended questions. Therefore, manual evaluation remains essential, especially in text generation and translation tasks (Xu et al., 2023).

Currently, there is a lack of specialized LLM benchmarks for crisis information generation. Crisis information, as a core element for responding to and solving crises, requires the support of the latest technologies. Therefore, this paper proposes constructing CIEval, an LLM evaluation benchmark for CI generation, combining manual and machine evaluations to identify the best LLMs and promote their application in this field.

Generation of CIEval dataset

Subjects selection

The crisis mechanism aims to efficiently respond to sudden crises and ensure social stability and public safety. According to the definition of *the Emergency Response Law* (General Office of the State Council, 2024), crises include natural disasters, accident disasters, public health events, and social security events. In order to comprehensively and accurately evaluate the ability of LLMs in CI generation, the CIEval dataset we constructed covers 16 segmented subjects under the four major categories of crises mentioned above, ensuring its representativeness and wide applicability. In addition, to further enhance the completeness of the dataset, we also focus on a specific social security event - cyber security, and includes 10 events with high frequency of occurrence in this field as supplementary categories. The specific event classification is shown in Table 1.

Categories	Subjects	Categories	Subjects	
Natural disasters	Drought and water disasters	Social security events	Economic security incidents	
	Meteorological disasters		Foreign emergencies	
	Earthquake		Cyber security events	Botnet
	Geologic disasters			Data leakage
	Forest and grassland fires			Phishing emails
Accident disasters	Enterprise safety accidents			Vulnerability exploitation
	Transportation accidents			DDOS
	Public facility accidents			APT
	Environmental pollution			Tampering
	Ecological damage incident			Worms
Public health events	Infectious disease outbreaks			Mining
	Congregative unknown disease			Ransomware
	Food safety			
	Animal outbreaks			

Table 1. Subjects of CIEval

Data collection and pre-processing

The primary sources of the raw data are post-disaster recovery plans and accident investigation reports publicly released by emergency management departments. Data related to cyber security events comes from typical case collections issued by cyber security companies like QAX (https://en.qianxin.com/).

For the raw data, a refined content extraction strategy was implemented to remove non-event-related elements such as headers and hyperlinks, retaining and integrating only text and image content crucial for event analysis, which are stored in structured documents. In cases where a document may cover multiple events, detailed manual intervention was employed to carefully split the content. This process ensures that each independent event is accurately mapped to a unique corresponding document, facilitating subsequent dataset construction. Each final document includes an overview, losses, policy measures, and other information of the event.

Dataset generation method

Dataset generation

Advanced models like GPT-4.0 play a crucial role in dataset construction (Liu et al., 2024). Based on this, we utilized the generative capabilities of LLMs to build an evaluation dataset, significantly shortening the construction period and reducing manual workload, providing abundant test resources for assessing LLM performance in specific crisis information needs.

Prompt engineering is a core element focused on designing precise prompts to guide LLMs in producing expected outputs. The key to prompt engineering lies in how to accurately construct prompts. Prompts (*P*) typically consist of instructions (*I*) and inputs (*In*); the instructions specify the task goals, setting a framework for model responses, while the inputs provide specific context or examples. *I* and *In* jointly influence the quality of the model's output, represented by Eq. (1).

$$P = f(I, In) \tag{1}$$

Here, *I* and *In* are combined through function *f* to form the LLM prompt (*P*). In this paper, *f* is the internal mechanism of LLMs, implemented by GPT-4.0. The method of using LLMs to generate datasets is illustrated in Figure 1.

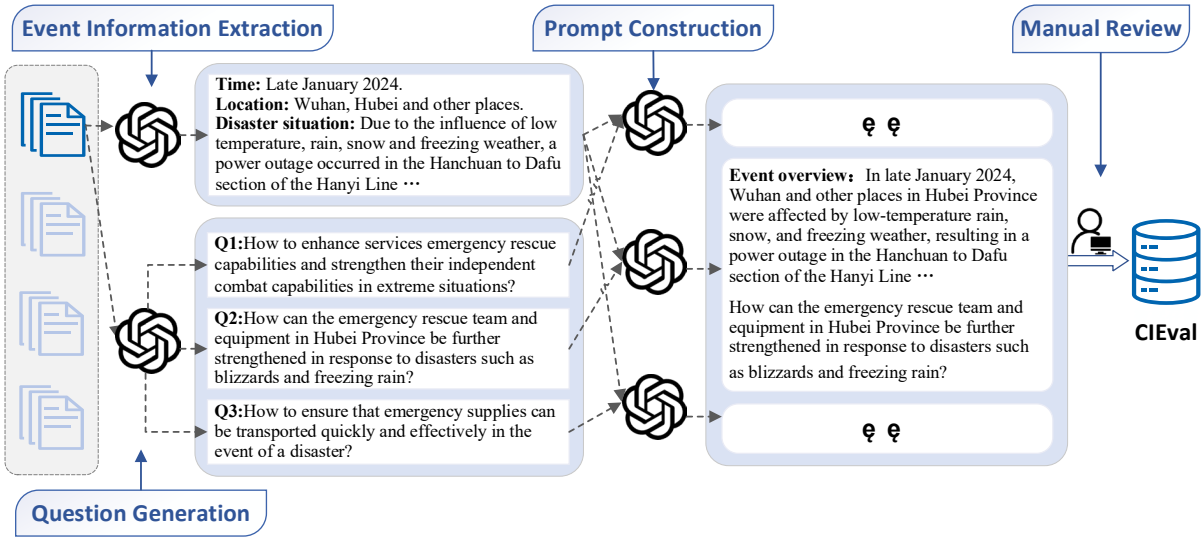


Figure 1. Overview of CIEval generation methods

Specifically, the dataset construction process we adopted follows a systematic pipeline approach, detailed as follows:

1. *Event information extraction*: pre-processed documents are input into GPT-4.0, which, using its powerful analytical capabilities, automatically extracts key event information from the documents. The information includes but not limited to occurrence time, disaster type, and economic losses, etc., i.e., *In*.

2. *Targeted question generation*: based on the response measures, policy protection suggestions, and other information in the input document, GPT-4.0 randomly generates a series of crisis-related questions, i.e., *I*.

3. *Prompt construction*: in this step, GPT-4.0 is used to combine *In* and *I* to form a complete *P* dataset for assessing the CI generation capability of LLMs.

4. *Manual review*: the automatically generated prompts undergo rigorous manual review to ensure clarity, fluency, and high relevance to the crisis information tasks. This process further minimizes the potential influence of any biases introduced by GPT-4.0, thereby enhancing the dataset quality and ensuring the accuracy and validity of the evaluation results.

Avoiding data contamination

In constructing the dataset, this study focuses on avoiding data contamination to ensure quality. Given that large-scale crises, especially those that generate widespread social response online, often become publicly accessible data resources through subsequent investigation reports, there is a potential risk of these being included in large language model training datasets. Considering

that the latest training data for the model to be evaluated is up to April 2024, this paper uses data from April 2024 onwards as the original dataset for CIEval, in order to reduce data contamination.

Dataset description

CIEval is a comprehensive dataset designed to thoroughly evaluate the performance of LLMs in CI generation. The dataset includes four major categories of authoritative-defined crisis types and additionally incorporates cyber security events, which have a broad impact in modern society, to fully reflect realistic and diverse CI needs. It covers a wide range of event types, including but not limited to meteorological disasters, earthquakes, transportation accidents, food safety, and data leakage, providing a thorough test of the models' performance in handling complex and variable CI tasks. Table 2 shows some of the contents of this dataset. CIEval contains a total of 4,820 evaluation questions, offering a comprehensive benchmark framework to assess the overall capabilities of LLMs in CI generation tasks, ultimately enhancing their practicality and effectiveness in real-world crisis scenarios.

Categories	Questions	Examples
Natural disasters	1500	**Event overview:** 'Time: June 22 to July 4, 2024; Location: Jiangxi Province; Affected Areas: 65 counties (cities, districts) in 9 prefecture level cities including Nanchang and Jiujiang; Number of Affected Persons: 1.565 million people...' **Based on the above description, please answer the following:** How to coordinate multi departmental cooperation to ensure the basic living security of residents in disaster areas?
Accident disasters	2000	**Event overview**: 'Time: April 26, 2024; Location: Xia County Expressway in Yuncheng, Shanxi Province; Event Type: Vehicle Fire Caused by Traffic Accident...' **Based on the above description, please answer the following:** How to handle the automatic locking system of damaged vehicles during crisis rescue?
Public health events	550	**Event overview:** '5.27 Xinzheng Elementary School Canteen Food Mold Incident' **Based on the above description, please answer the following:** How to effectively establish and manage a parental supervision mechanism in the school environment to ensure food safety?
Social security events	770	**Event overview:** '8.19 Philippine Coast Guard Ship Collides with Chinese Ship Incident' **Based on the above description, please answer the following:** How to quickly take measures to prevent the situation from escalating when a similar collision event occurs?
Cyber security events	310	**Event overview:** 'The terminal computer was infected with ransomware through phishing emails' **Based on the above description, please answer the following:** What role does terminal security control software play in preventing ransomware?

Table 2. Partial content of CIEval dataset

Benchmarking experiment

Models

To comprehensively understand the applicability and CI generation capabilities of LLMs in the CI domain within the Chinese context, this study selected four of the latest models developed by Chinese companies and four internationally renowned models suitable for the Chinese context. These models vary in scale and structure, performing differently across various datasets and tasks. Specific information about the models is shown in Table 3.

Models	Developer	Size	Access	Source
GPT-4o	OpenAI	Undisclosed	API	https://chat.openai.com/
Claude 3.5 Sonnet	Authropic	Undisclosed	API	https://claude.ai
Llama-3.1-405B	Meta	405B	Weights	https://github.com/meta-llama/llama3
Gemini-1.5-Pro	Google	Undisclosed	API	https://deepmind.google/technologies/gemini/
ERNIE 4.0 Turbo	Baidu	Undisclosed	Official website	https://yiyan.baidu.com
GLM-130B	Tsinghua	130B	Weights	https://github.com/THUDM/GLM-130B
iFlytek Spark V4.0	iFlytek	Undisclosed	Official website	https://xinghuo.xfyun.cn
Qwen LM-72B	Alibaba	72B	Official website	https://tongyi.aliyun.com/qianwen

Table 3. Description of models

GPT-4o is the latest and fastest flagship model of OpenAI, a specialized version of GPT-4.0 optimized for specific tasks such as real-time inference, and generating audio, images, and text. In addition, Claude 3.5 Sonnet surpassed GPT-4o in multiple areas including graduate level reasoning, undergraduate level knowledge, and coding ability, according to the Authropic's release report.

ERNIE 4.0 Turbo, GLM-130B, IFlytek Spark V4.0 and Qwen-72B are currently outstanding Chinese LLMs. IFlytek Spark V4.0 fully benchmarked GPT-4-Turbo during construction, and according to testing, it has surpassed GPT-4-Turbo in text generation, logical reasoning, and other aspects. Among them, non-open-source models are all the premium version.

Evaluation methods

This study plans to use a combined evaluation method of manual scoring and machine scoring to enhance the efficiency of the evaluation process and ensure the comprehensiveness and accuracy of the results.

Manual scoring

Based on Wang et al. (2024), we established evaluation indicators for CI generation capabilities of LLMs, including content quality, expression quality, feasibility, and effectiveness, starting from the quality of the CI itself. Firstly, in order to obtain professional evaluation opinions, this study invited three PhD students with relevant experience in the field of CI as evaluators. The scoring personnel score the CI generated by each model based on the indicators, and the score level is divided into [extremely high, high, medium, low, extremely low]. Secondly, based on the fuzzy decision trial and evaluation laboratory method, the weights of each indicator were based on the experience of the scoring personnel. Finally, the interactive multi-criteria decision-making method based on triangular intuitive fuzzy numbers was used to obtain the global dominance of the evaluated model, which is the CI generation capability score of LLMs (Qin et al., 2017).

Machine scoring

Given the large scale of CIEval, relying solely on manual scoring would require a significant amount of manpower and resources. To this end, this study further explores the possibility of machine scoring. We select GPT-4.0, which performs well in multiple benchmarks (Xu et al., 2023), and Claude 3.5 Sonnet, which the developer claims to be superior to GPT-4o in all aspects, as '*machine scoring experts*.'

Similar to the final results obtained through manual scoring, the machine scoring range is set to [0,1]. To ensure the rationality of machine scoring, we will conduct validation experiments before

formal evaluation, comparing the consistency between GPT-4.0 and Claude-3.5 Sonnet scoring and manual scoring to verify their reliability as machine scoring experts. Choose the optimal model score to replace manual scoring, thereby improving evaluation efficiency and reducing labour costs.

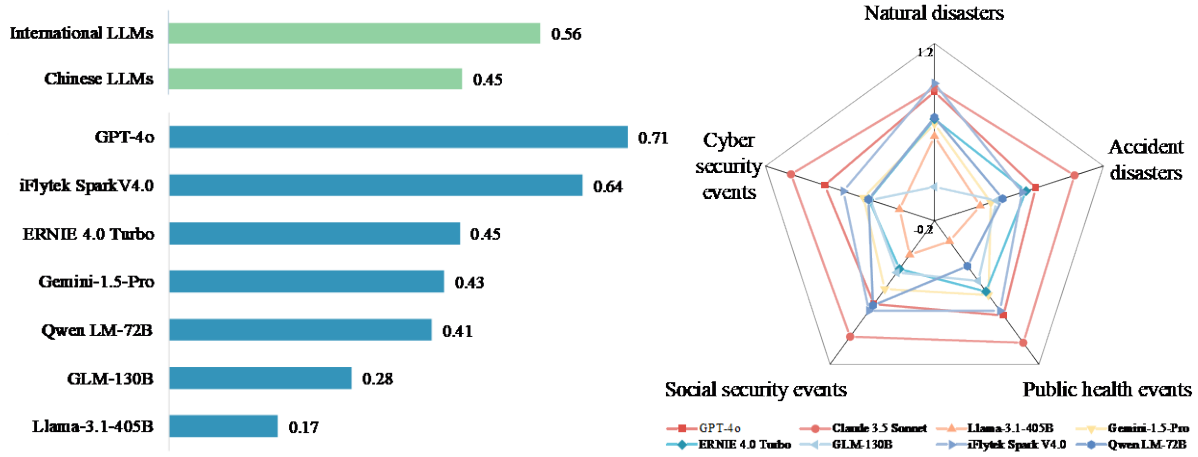
Results and discussion

Comparison between manual scoring and machine scoring

This study randomly selected 80 pieces of data in the CIEval dataset as samples for manual scoring and machine scoring, and conducted correlation tests between manual scoring and Claude-3.5 Sonnet/GPT-4.0 scoring, respectively. The Mann-Kendall's test results showed that the correlation coefficient between GPT-4.0 score and the manual score was 0.397, showing extremely high significance ($p<0.01$). Claude-3.5 Sonnet was significantly correlated with manual scoring at the p value lower than 0.05. To further validate the rationality of utilizing GPT-4.0 for scoring, a Spearman test was conducted. The results indicated a correlation coefficient of 0.514 between GPT-4.0 scores and manual scores, further supporting the feasibility and reliability of replacing manual scoring with GPT-4.0. Therefore, this study selects GPT-4.0 as the main tool for evaluating CI generation capabilities of LLMs.

The crisis information generation capabilities of LLMs

This section provides a performance evaluation of several LLMs based on the CIEval dataset. Overall, Chinese LLMs slightly underperform compared to leading international models such as the GPT and Claude series (Figure 2(a)). This performance gap can be attributed to the global advantages in data resources and technological development enjoyed by international models.



(a) Average score of LLMs' CI generation capability

(b) Score of LLMs in different types of crisis

Figure 2. Score of LLMs' CI generation capability

As shown in Figure 2(b), iFlytek Spark V4.0 excels in generating relevant information for natural disasters, achieving high scores due to its deep understanding of local meteorological data. In the context of accident disasters, Claude 3.5 Sonnet and GPT-4o outperform others with scores of 0.96 and 0.63, respectively. While models like ERNIE 4.0 Turbo perform well in specific scenarios such as transportation accidents (Figure 3), others, like Llama-3.1-405B, fall short in complex enterprise safety incidents, producing less actionable information. For public health events, Claude 3.5 Sonnet leads in scenarios like infectious disease outbreaks. In the social security category, Claude 3.5 Sonnet excels, achieving a score of 0.98 due to its strong situational analysis. Other models like Qwen LM-72B and ERNIE 4.0 Turbo achieve high scores in specific incidents like mining and ransomware.

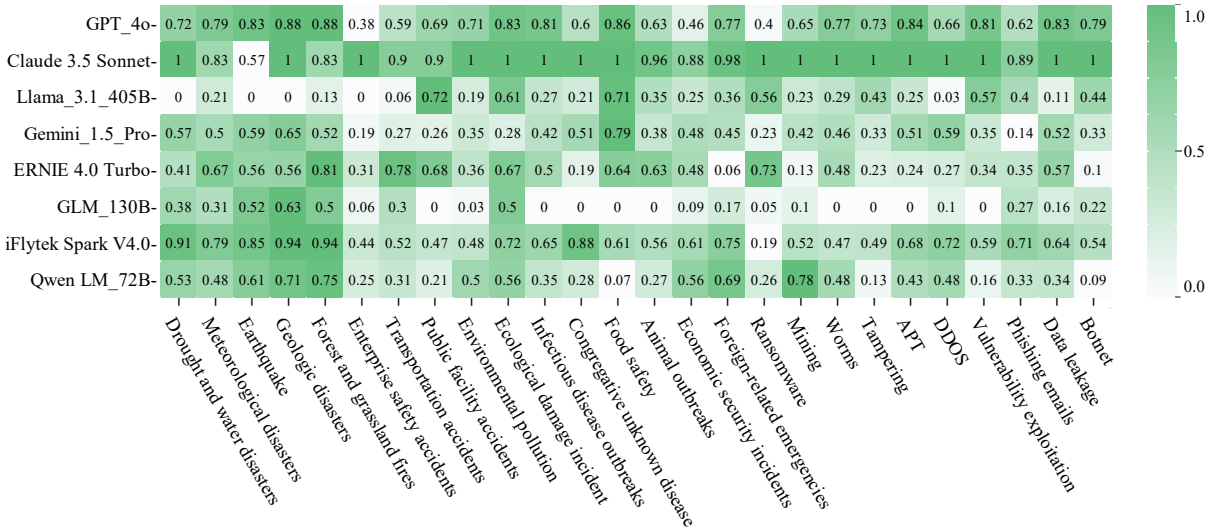


Figure 3. Score of LLMs' CI generation capability in segmented scenarios

Conclusion

This study focuses on the potential of LLMs in the field of CI generation and proposes a scientifically reasonable evaluation benchmark, CIEval. CIEval is constructed through a series of processes including key event information extraction, question generation, prompt construction, and manual review. This dataset contains 26 types of crises related to natural disasters, accident disasters, public health events, and social security events, with a total of 4.8k data. In the experimental phase, we evaluated the CI generation capability of LLMs such as Claude 3.5 Sonnet using the CIEval dataset. At the same time, we validated the feasibility of GPT-4.0 as a machine scoring expert. This benchmark aims to provide reference for the optimization and application of LLMs in CI generation in the future, which will help promote the application and development of LLMs in practical emergency management.

Acknowledgements

This study is supported by the National Social Science Foundation of China (Grant No. 23&ZD230).

About the authors

Ruilian Han is a PhD student at School of Information Management, Wuhan University, China. Her research focuses on social media data analysis. She can be contacted at rlhan_1127@163.com

Lu An is a professor at School of Information Management, Wuhan University, China. Her research focuses on crisis informatics. She can be contacted at anlu97@163.com

Wei Zhou is a PhD student at School of Information Management, Wuhan University, China. Her research focuses on risk identification. She can be contacted at 664880781@qq.com

Li Gang is a professor at School of Information Management, Wuhan University, China. His research focuses on information resource management. He can be contacted at ligang@whu.edu.cn

References

Alkhaled, L., Adewumi, T., & Sabry, S. S. (2023). Bipol: A novel multi-axes bias evaluation metric with explainability for NLP. Natural Language Processing Journal, 4, 100030. https://doi.org/10.1016/j.nlp.2023.100030

Evtikhiev, M., Bogomolov, E., Sokolov, Y., & Bryksin, T. (2023). Out of the bleu: How should we assess quality of the code generation models? Journal of Systems and Software, 203, 111741. https://doi.org/10.1016/j.jss.2023.111741

General Office of the State Council, P. (2024). Emergency Response Law. People's Publishing House.

Giannakopoulos, K., Kavadella, A., Salim, A. A., Stamatopoulos, V., & Kaklamanos, E. G. (2023). Evaluation of the performance of generative AI large language models Chatgpt, Google bard, and microsoft bing chat in supporting evidence-based dentistry: comparative mixed methods study. Journal of Medical Internet Research, 25, e51580. https://doi.org/10.2196/51580

Guler, N., & Kirshner, S.N. (2024). A literature review of artificial intelligence research in business and management using machine learning and ChatGPT. Data and Information Management, 8(3), 1-25. https://doi.org/10.1016/j.dim.2024.100076

Liu, Y., Li, D., Wang, K., Xiong, Z., Shi, F., Wang, J., Li, B., & Hang, B. (2024). Are LLMs good at structured outputs? A benchmark for evaluating structured output capabilities in LLMs. Information Processing & Management, 61(5), 103809. https://doi.org/10.1016/j.ipm.2024.103809

Qin, Q., Liang, F., Li, L., Chen, Y.-W., & Yu, G.-F. (2017). A TODIM-based multi-criteria group decision making with triangular intuitionistic fuzzy numbers. Applied Soft Computing, 55, 93–107. https://doi.org/10.1016/j.asoc.2017.01.041

Sarlin, P.-E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4938–4947. https://doi.org/10.1109/CVPR42600.2020.00499

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., & Wei, J. (2022). Challenging Big-Bench tasks and whether chain-of-thought can solve them (arXiv:2210.09261). arXiv. https://doi.org/10.48550/arXiv.2210.09261

Thelwall, M. (2024). Can ChatGPT evaluate research quality? Journal of Data and Information Science, 9(2), 1–21. https://doi.org/10.2478/jdis-2024-0013

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A multi-task benchmark and analysis platform for natural language understanding (arXiv:1804.07461). arXiv. https://doi.org/10.48550/arXiv.1804.07461

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2024). Overview of data quality: Examining the dimensions, antecedents, and impacts of data quality. Journal of the Knowledge Economy, 15(1), 1159–1178. https://doi.org/10.1007/s13132-022-01096-6

Xu, L., Hu, H., Zhang, X., Li, L., Cao, C., Li, Y., Xu, Y., Sun, K., Yu, D., Yu, C., Tian, Y., Dong, Q., Liu, W., Shi, B., Cui, Y., Li, J., Zeng, J., Wang, R., Xie, W., … Lan, Z. (2020). CLUE: A chinese language understanding evaluation benchmark (arXiv:2004.05986). arXiv. https://doi.org/10.48550/arXiv.2004.05986

Xu, L., Li, A., Zhu, L., Xue, H., Zhu, C., Zhao, K., He, H., Zhang, X., Kang, Q., & Lan, Z. (2023). SuperCLUE: A comprehensive chinese large language model benchmark (arXiv:2307.15020). arXiv. https://doi.org/10.48550/arXiv.2307.15020