Table 3: Comparison results with unclear prediction. RA means reason-aware prompt. The value in bold is the highest in each column.

| Dataset | | Original | | | | RA. | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc-1 ↑ | Acc-2 ↑ | Acc-3 ↑ | F1. ↑ | Acc-1 ↑ | Acc-2 ↑ | Acc-3 ↑ | F1. ↑ |
| CHINESE RUMOR | *(w/o)* | 0.676 | 0.538 | 0.665 | 0.664 | 0.715 | 0.567 | 0.716 | 0.714 |
| | *(w/)* | 0.768 | 0.593 | 0.759 | 0.761 | 0.811 | 0.643 | 0.812 | 0.826 |
| LIAR | *(w/o)* | 0.711 | 0.538 | 0.700 | 0.697 | 0.719 | 0.676 | 0.715 | 0.715 |
| | *(w/)* | 0.652 | 0.573 | 0.643 | 0.634 | 0.653 | 0.596 | 0.649 | 0.647 |
| WEIBO21 | *(w/o)* | 0.730 | 0.572 | 0.728 | 0.719 | 0.772 | 0.622 | 0.769 | 0.769 |
| | *(w/)* | 0.798 | 0.624 | 0.781 | 0.730 | 0.847 | 0.666 | 0.846 | 0.827 |
| COVID-19 | | 0.774 | 0.648 | 0.750 | 0.749 | 0.818 | 0.715 | 0.807 | 0.807 |
| FAKENEWSNET | | 0.652 | 0.550 | 0.635 | 0.597 | 0.692 | 0.608 | 0.685 | 0.662 |
| KAGGLE | | 0.708 | 0.572 | 0.637 | 0.621 | 0.800 | 0.717 | 0.783 | 0.786 |
| CELEBRITY | | **0.826** | 0.713 | 0.811 | **0.815** | **0.888** | 0.741 | **0.880** | **0.885** |
| FAKENEWSAMT | | 0.816 | **0.778** | **0.816** | 0.812 | 0.804 | **0.745** | 0.795 | 0.792 |
| TWITTER15&16 | | 0.646 | 0.580 | 0.631 | 0.579 | 0.689 | 0.598 | 0.675 | 0.637 |

of ChatGPT on fake news is significantly low when prompted with the normal template (as shown in Appendix D.3), indicating that ChatGPT tends to misclassify fake news as true news. We attribute this to two possible reasons: first, ChatGPT lacks a comprehensive understanding of the distinct characteristics of fake news; second, ChatGPT tends to be conservative when detecting fake news (the number of predictions with "real" are more than "fake"). To address these limitations and improve ChatGPT's detection capability, we introduce a reason-aware prompt method, as illustrated in Figure 3. We have added a summary in Table 1 to our prompt template, which not only describes the features of fake news, but also serves as a cue to subconsciously prompt ChatGPT to increase its inclination in predicting samples as fake news.

### 5.5 Analysis

The results in nine different datasets are shown in Table 4 and Table 3, including the 2-class task (without the "unclear" prediction ) and 3-class task (with the "unclear" prediction).

It is noticeable that ChatGPT demonstrates a relatively strong ability to detect fake news, though there remains room for improvement. Overall, ChatGPT achieved satisfactory results on some datasets, with Acc-1 surpassing 70% for 8 out of 11 tested datasets in the 3-class scenario, and the highest accuracy reaching 82.6%. Nonetheless, there is still potential for improvement on certain datasets, such as the LIAR dataset and the CHINESE RUMOR dataset. Also, we observed that the introduction of the "unclear" class improved ChatGPT's prediction performance when comparing Acc-1 with Acc.

Table 4: Comparison results without unclear prediction. RA means reason-aware prompt. The value in bold is the highest in each column.

| Dataset | | Original | | RA. | |
|---|---|---|---|---|---|
| | | Acc. ↑ | F1. ↑ | Acc. ↑ | F1. ↑ |
| CHINESE RUMOR | *(w/o)* | 0.600 | 0.574 | 0.677 | 0.677 |
| | *(w/)* | 0.681 | 0.677 | 0.776 | 0.776 |
| LIAR | *(w/o)* | 0.631 | 0.606 | 0.658 | 0.699 |
| | *(w/)* | 0.644 | 0.615 | 0.630 | 0.624 |
| WEIBO21 | *(w/o)* | 0.620 | 0.601 | 0.722 | 0.721 |
| | *(w/)* | 0.743 | 0.711 | 0.780 | 0.779 |
| COVID-19 | | 0.746 | 0.731 | 0.778 | 0.770 |
| FAKENEWSNET | | 0.610 | 0.571 | 0.646 | 0.620 |
| KAGGLE | | 0.577 | 0.499 | 0.774 | 0.763 |
| CELEBRITY | | 0.756 | 0.750 | **0.844** | **0.842** |
| FAKENEWSAMT | | **0.795** | **0.787** | 0.823 | 0.817 |
| TWITTER15&16 | | 0.632 | 0.598 | 0.674 | 0.658 |

This suggests that ChatGPT's uncertainty for some samples can negatively impact prediction accuracy.

Furthermore, reason-aware prompts enhance ChatGPT's fake news detection capabilities on most datasets. We observed significant improvements in predictions on all datasets with 2-class when using reason-aware prompts. Additionally, reason-aware prompts also yielded improved 3-class results on most datasets. Specifically, the maximum improvement was achieved on the KAGGLE dataset, with increases of 19.7% in Acc, 9.2% in Acc-1, 14.5% in Acc-2, and 14.6% in Acc-3.

In addition, extra information including context and comment generally enhance ChatGPT's fake news detection capabilities. Comparing the results between *(w/o)* and *(w/)*, the CHINESE RUMOR dataset and WEIBO21 dataset exhibit significant

Table 5: The percentage (%) of different types of additional information. ■, ■ and ■ represents rank 1, 2 and 3 percentage. We didn't test CELEBRITY and FAKENEWSAMT datasets due to their small size of "unclear" samples.

| Dataset | | A | B | C | D | AB | AC | AD | BC | BD | CD |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CHINESE RUMOR | *(w/o)* | 27.27 | 17.11 | 16.22 | 18.36 | 3.92 | 4.99 | 4.99 | 2.50 | 2.67 | 1.97 |
| | *(w/)* | 35.03 | 12.69 | 20.30 | 18.78 | 1.52 | 3.55 | 5.08 | 0.51 | 1.52 | 1.02 |
| LIAR | *(w/o)* | 31.76 | 7.03 | 18.46 | 21.32 | 1.98 | 6.37 | 7.36 | 1.65 | 0.99 | 3.08 |
| | *(w/)* | 31.76 | 12.83 | 17.35 | 19.43 | 2.80 | 4.87 | 6.24 | 1.50 | 1.28 | 1.94 |
| WEIBO21 | *(w/o)* | 30.10 | 14.26 | 14.85 | 21.78 | 2.38 | 4.16 | 7.32 | 1.98 | 1.78 | 1.39 |
| | *(w/)* | 34.21 | 12.39 | 19.20 | 17.63 | 2.79 | 4.71 | 5.41 | 1.22 | 0.87 | 1.57 |
| COVID-19 | | 31.43 | 12.56 | 17.46 | 19.33 | 2.92 | 5.14 | 6.19 | 1.46 | 1.29 | 2.22 |
| FAKENEWSNET | | 29.97 | 11.36 | 17.98 | 18.93 | 3.47 | 6.31 | 5.99 | 1.26 | 1.26 | 3.47 |
| KAGGLE | | 22.22 | 22.59 | 14.81 | 21.85 | 2.96 | 2.96 | 4.44 | 2.59 | 3.35 | 2.23 |
| TWITTER15&16 | | 28.90 | 12.93 | 17.87 | 20.15 | 1.90 | 6.08 | 5.70 | 1.52 | 2.66 | 2.28 |

improvements in various metrics when utilizing additional information. This implies that additional information may augment the semantic understanding of news. However, for the three-class classification, employing post-context information in the LIAR dataset led to a decrease in Acc-1 and Acc-3, but an increase in Acc-2. A possible explanation for this outcome is that context information decrease the probability of examples being predicted as "unclear," yet raised the probability of them being misclassified as "fake" or "real."

### 5.6 More Information Behind the Unclear

To explore how to reduce the "unclear" labels predicted by ChatGPT in the three-classification task ("real", "fake" and "unclear"), we prompt ChatGPT with a question: *"What additional information do you need to make a more accurate judgment?"*. This prompt is presented to ChatGPT for the samples classified as "unclear". Similar to those in Section 4.2, we offer ChatGPT four pre-defined options to choose from, which are listed in Box 5.6. Then we measure the proportions of them on different datasets (as shown in Table 5).

**A**: External knowledge refers to factual information, expert suggestions, or data reliability.
**B**: Multimodal information includes images, videos, or audio.
**C**: Context information encompasses comments, reposts, post time or post location.
**D**: Speaker's information includes user actions, information from social media accounts, or the user's history of posts.

We find that for most datasets, option A consistently ranks highest, implying that ChatGPT lacks some external knowledge to accurately assess news authenticity. This challenge can be tackled by incorporating extra knowledge like a knowledge graph (Dun et al., 2021) or a knowledge base (Hu et al., 2021). Options A, C, and D tend to occupy the second rank across different datasets. For instance, when addressing fake news originating from social media, one might need to consider using information related to comments (Khoo et al., 2020; Yang et al., 2021), reposts, or posts (option C), or take into account the users' preferences (Dou et al., 2021) and the information about users' profile (Shu et al., 2019b) (option D). Additionally, we found that these options are not mutually exclusive, and ChatGPT may yield results for multiple options (we only consider two-option combinations due to the low frequency of the data with three or more options). Consequently, it is crucial to merge various kinds of extra information for fake news detection.

## 6 Conclusion

In this study, we conducted an exploration into the capabilities of ChatGPT in generating, explaining, and detecting fake news. We found that some prompts enable ChatGPT to generate deceptive fake news, underscoring its potential harm. Then we identified nine features of fake news via ChatGPT, which may serve as a foundation for future research. Additionally, we enhanced the effectiveness of ChatGPT in detecting fake news by introducing the reason-aware prompt. Despite ChatGPT's promising performance on some datasets, there is still room for improvement. Finally, we investigated the extra information that may help ChatGPT detect fake news better. Overall, this paper provides insights into intelligent information governance and emphasizes the need for further research to fully leverage the capabilities of LLMs.

## Ethics Statement

Our findings indicate that ChatGPT can generate extreme and targeted false news. Thus, we advise researchers to use caution when employing language models like ChatGPT and to effectively handle any harmful content that may arise. Simultaneously, we emphasize the potential of language models in combating disinformation and advocate for responsible utilization. Regarding the human evaluation section, we ensured that participants agreed to our data collection agreement before collecting any information, and we treated participant information with utmost care. We promise to be responsible for personal data and will not disclose any personal data

## Limitations

In this paper, our primary focus has been on examining the performance of ChatGPT specifically in the domain of fake news generation, explanation, and detection, without evaluating other large language models. Moreover, our evaluation has been limited to a dataset consisting of only 5200 samples, and conducting a larger-scale evaluation would contribute to the overall reliability of the findings. Additionally, given the black-box nature of large language models (LLMs), it remains challenging to definitively ascertain why reason-aware prompts are effective in fake news detection.

## References


Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1*, pages 127–138. Springer.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023a. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Mithun Das, Saurabh Kumar Pandey, and Animesh Mukherjee. 2023. Evaluating chatgpt's performance for multilingual and emoji-based hate speech detection. *arXiv preprint arXiv:2305.13276*.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *arXiv preprint arXiv:2304.05335*.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2051–2055.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89.

Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023a. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Yuan Gao, Ruili Wang, and Feng Hou. 2023b. Unleashing the power of chatgpt for translation: An empirical study. *arXiv preprint arXiv:2304.02182*.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.