From Creation to Clarification: ChatGPT's Journey Through the Fake News Quagmire

Yue Huang*
University of Notre Dame
South Bend, IN, USA
yhuang37@nd.edu

Kai Shu
Illinois Institute of Technology
Chicago, IL, USA
kshu@iit.edu

Philip S. Yu
University of Illinois Chicago
Chicago, IL, USA
psyu@uic.edu

Lichao Sun
Lehigh University
Bethlehem, PA, USA
lis221@lehigh.edu

ABSTRACT

The rampant spread of fake news has adversely affected society, resulting in extensive research on curbing its spread. As a notable milestone in large language models (LLMs), ChatGPT has gained significant attention due to its exceptional capabilities. In this study, we present an exploration of ChatGPT's proficiency in generating, explaining, and detecting fake news as follows. *Generation* – We employ different prompt methods to generate fake news and prove the high quality of these instances through both self-assessment and human evaluation. *Explanation* – We obtain nine features to characterize fake news based on ChatGPT's explanations and analyze the distribution of these factors across multiple public datasets. *Detection* – We examine ChatGPT's capacity to identify fake news. We propose a reason-aware prompt method to improve its performance. We further probe into the potential extra information that could bolster its effectiveness in detecting fake news.

CCS CONCEPTS

• **Computing methodologies** → *Natural language generation.*

KEYWORDS

Fake News, Large Language Models, ChatGPT

ACM Reference Format:
Yue Huang, Kai Shu, Philip S. Yu, and Lichao Sun. 2024. From Creation to Clarification: ChatGPT's Journey Through the Fake News Quagmire. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, Singapore, 4 pages. https://doi.org/10.1145/3589335.3651509

*Visiting student at LAIR Lab, Lehigh University.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651509

1 INTRODUCTION

Fake news has raised significant concerns all over the world [10]. Recently, ChatGPT has received widespread acclaim for its exceptional performance across various domains. Due to its popularity and strong capabilities, ChatGPT presents both opportunities and challenges within the domain of fake news research. Despite its potential, recent studies [6] have raised concerns about LLM being exploited for malicious purposes. As a result, it is vital to explore and understand ChatGPT's capacity for fake news generation to address this severe problem. Besides generating fake news via ChatGPT, we should also leverage its ability for fake news explanation and detection. This has motivated us to utilize ChatGPT for fake news understanding, by providing explanations that demonstrate a certain level of comprehension and reasoning. Moreover, it is crucial to investigate the performance of ChatGPT in fake news detection, identify its limitations, and devise strategies to enhance its detection capabilities. In this paper, we did an in-depth exploration of fake news generation, detection, and explanation via ChatGPT. Our contributions in this paper can be summarized as follows: (1) We examine ChatGPT's capability to generate fake news using different prompting methods. The results from self-evaluation and human evaluation show that the generated samples are of high quality. (2) We investigate ChatGPT's capacity to explain fake news and summarize nine features that define fake news across nine datasets, which offers some insights for future work. (3) We assess ChatGPT's effectiveness in detecting fake news. Based on the summarized features from the above explanations, we propose a reason-aware prompting method to enhance its detection capability. Moreover, we explore additional information that can assist ChatGPT in detecting fake news more effectively.

2 FAKE NEWS GENERATION

Prompt Methods. We introduce four methods to prompt ChatGPT into generating fake news, navigating around its moderation mechanisms, and reinforcement learning from human feedback (RLHF) [1]. **(a) Altering text meaning:** This involves modifying the meaning of the original text to produce content that may contradict the facts, potentially resulting in fake news. **(b) Inventing stories:** By providing an outline of the target story and prompting ChatGPT to flesh it out, the generated story with unreal information could be considered fake news. **(c) Creating imaginary text:** This method generates fictional content without a specific outline,

differing from inventing stories by the arbitrariness of the generated content. **(d) Multiple prompts:** A three-step prompt strategy to generate targeted fake news that can evade ChatGPT's filters. It starts with a "Topic Prompt" to steer the conversation towards a news-related subject, followed by a "Deep Prompt" for generating a specific news article, and finally, a "News Augmentation Prompt" to add detailed elements like time, location, and media source, enhancing the realism and believability of the article.

Quality of Generated Samples. We use the above methods to generate 40 pieces of fake news. To evaluate the generation quality of ChatGPT, we conduct both self-evaluation and human evaluation.

Self-evaluation. For self-evaluation, we performed fake news detection using ChatGPT itself. To minimize the impact of contextual semantics during the conversation, we created a new conversation for each sample during evaluation. Additionally, to achieve more realistic and accurate results, we categorized ChatGPT's outputs into three distinct categories: fake news, real news, and uncertain. We utilized a prompt template such as *"Please evaluate the authenticity of the following news. You can respond with 'fake', 'real', or 'uncertain'"*. The experiment revealed that out of the 40 fake news samples, ChatGPT accurately identified 29 fake news instances (a successful rate of 72.5%). However, it judged nine instances as real news and two instances as uncertain cases, suggesting a slight difficulty in detecting its own generated content.

Human evaluation. To assess the real-world effectiveness of ChatGPT's generated samples, we conduct the human evaluation by handing out questionnaires. We collected 294 data items during human evaluation, consisting of 223 items about fake news and 71 items about real news. Overall, we observed that humans achieved an accuracy of only 54.8% in identifying the generated fake news, highlighting the challenge of distinguishing these instances as fake. Notably, one sample exhibited the lowest accuracy, with only 10 out of 33 judgments being correct (a mere 33.3% accuracy). This suggests that some generated samples effectively deceive human judgment. Furthermore, we investigated the reasons why humans think the given news is fake. The results are: Fact Conflict (18.4%), Unauthoritative or informal expressions (23.9%), Oversimplification or emotional bias (13.5%), Lack of evidence or credible source (36.2%), Lack of context (6.1%) and other (1.9%). "Lack of evidence or credible source" is the primary reason, comprising 36%. This discovery aligns with the observations in Section 3, emphasizing the significance of incorporating additional details to improve the generation quality. The factor that ranks second is "unauthoritative or informal expressions," indicating the need for ChatGPT to enhance its language style when generating news-like content. Furthermore, "fact conflict" constitutes 18% of the cases, implying that generated news may include factual inconsistencies, highlighting the importance of fact-checking for its outputs. Overall, the above results indicate that leveraging certain prompt ways allows ChatGPT to produce high-quality fake news, closely resembling real-world news.

3 EXPLANATION OF FAKE NEWS

In this section, we evaluate ChatGPT's capacity to provide explanations on given fake news. Our goal is to examine the factors that contribute to defining fake news. The explanation process comprises two stages: reason summary and reason selection. By analyzing the distribution of these nine factors, we found that these reasons (factors), to different extents, characterize fake news and may provide insights for future work.

Reason Summary & Reason Selection. Firstly, we select fake news from nine public datasets and ask ChatGPT to explain why these pieces of news are fake. Then we select a subset from these explanations and manually summarize them, yielding elementary reasons. We consult ChatGPT to determine if any of these reasons overlap and to suggest additional reasons. After several iterations of this process, we finally identify nine reasons that ChatGPT offers for why a given piece of news is fake. The nine explainable reasons are summarized in Table 1. After summarizing the explanations, we ask ChatGPT to select reasons from these nine options (potentially selecting more than one option) or provide its reason if none of the listed options apply when presented with a fake news sample. The distribution of single options across different datasets is shown in Figure 2. Letter A to I represent the nine reasons respectively, and J represents other reasons.

Table 1: Summary reason from fake news explanation.

Option	Reason	Description
A	Emotional bias or misleading intent	This explanation suggests that fake news is characterized by an emotional bias, which can include an excessively aggressive portrayal of a subject or an attempt to manipulate readers to achieve a hidden agenda.
B	Lack of evidence or credible sources	This reason indicates that fake news lacks credible evidence to support its claims.
C	Conflicting facts	This reason suggests that fake news conflicts with established facts, such as wrong information about people or events.
D	Informal statements, expressions, or vague language	This reason highlights that the language used in fake news may not be formal, or may be vague or ambiguous.
E	Insufficient supporting materials	This reason indicates that although the news may have mentioned the source of an event or provided relevant evidence, the evidence is not sufficient to support its claims.
F	Lack of context or taken out of context	This reason indicates that fake news may lack relevant context, such as comments, retweets and user information that provide additional information.
G	Misinterpretation or misquotation	This reason suggests that fake news may misinterpret or misquote facts, leading to inaccurate or false claims.
H	Oversimplification or exaggeration	This reason highlights that fake news may oversimplify or exaggerate information, leading to false claims.
I	Doctored images or videos	This reason indicates that the images or videos mentioned in the news text may be altered or misrepresented, making them untrustworthy.
J	Other	ChatGPT must specify a reason if the above options don't match its answer.

Analysis. In Figure 2, we noticed that the distribution of options across the nine datasets is generally similar, with slight variations in the distribution of specific options. Reason B (i.e., "not providing relevant evidence") is the most prevalent characteristic of fake news across almost all datasets. This observation aligns with the findings of some prior research [4, 7] which focus on using evidence information. Instead, in the Covid-19 dataset, option A (i.e., "misleading intentions") ranks highest, implying that much fake news in this dataset may have intentions such as inciting panic or showcasing bravado. This insight highlights the significance of considering emotional information in news, as studied by previous research [9]. Additionally, we discovered that reason D (i.e., "linguistic style") is the third most common reason across most datasets, especially

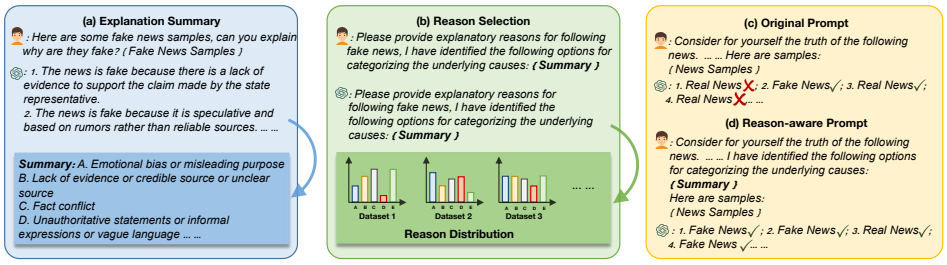

Figure 1: Fake news summary (a), reason selection (b), original prompt (c), and reason-aware prompt (d).

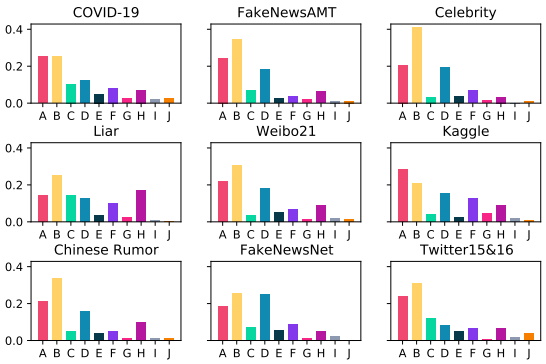

Figure 2: Distribution of reasons behind fake news (single option)

in the FakeNewsNet dataset, where reasons D and B are nearly equally prevalent. This observation suggests that utilizing the linguistic style of news may improve fake news detection. Moreover, we noticed that the proportion of reason C (i.e., "factual errors") is relatively higher in the Covid-19 and Liar compared to other datasets. This trend may be due to the frequent presence of factual errors in these datasets. For instance, the Covid-19 dataset includes content with obvious factual conflict, such as the new assert that 5G can spread Covid-19, showcasing ChatGPT's certain ability of fact-checking.

4 FAKE NEWS DETECTION

In this section, we proposed a reason-aware prompt method based on summarizing the reasons behind fake news to enhance its detection ability.

Experimental Settings. To mitigate the impact of ChatGPT's inconsistency on detection, in addition to the 2-class task, we also introduced a 3-class task, where ChatGPT predicts whether a sample is *"true"*, *"fake"*, or *"unclear"*. For evaluating ChatGPT, we use different metrics for 2-class and 3-class tasks. For the 2-class task, accuracy and F1 score are used. For the 3-class task, the metrics include Acc-1 (accuracy excluding "unclear" predictions, analyzed as a binary task), Acc-2 (accuracy considering "unclear" predictions as incorrect), Acc-3 (accuracy recalculated after removing "unclear" predictions and balancing the sample ratio (1:1)), and F1 score. Acc-2 measures the tendency of ChatGPT to label predictions as "unclear," while Acc-3 aims to eliminate bias from uncertain samples.

Reson-aware Prompt. In this section, we propose a reason-aware prompt method to enhance ChatGPT's performance in detecting fake news. We observed that the recall rate of ChatGPT on fake news is significantly low when prompted with the normal template, 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 shown in Figure 1.

Table 2: 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

Table 3: 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

Analysis. The results in nine different datasets are shown in Table 2 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. 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 enhances ChatGPT's fake news detection capabilities. Comparing the results between *(w/o)* and *(w/)*, the Chinese Rumor dataset and Weibo21 dataset exhibit significant improvements in various metrics when utilizing additional information. This implies that additional information may augment the semantic understanding of news.

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.

More Information Behind the Unclear Predictions. 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". We offer ChatGPT four pre-defined options to choose from, which are listed in Box 4. Then we measure the proportions of them on different datasets (as shown in Table 4). We find that for most datasets, option A consistently ranks the 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 base. 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 [3, 8], reposts, or posts (option C), or take into account the users' preferences [2] and the information about users' profile [5] (option D).

5 CONCLUSION

In this study, we conducted an exploration into the capabilities of ChatGPT in generating, explaining, and detecting fake news. This paper provides insights into intelligent information governance and emphasizes the need for further research to fully leverage the capabilities of LLMs.

Table 4: 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

6 ACKNOWLEDGMENTS

Lichao Sun and Yue Huang are supported by the National Science Foundation Grants CRII-2246067 and Microsoft Accelerate Foundation Models Research Award. Kai Shu is supported by the U.S. Department of Homeland Security under Grant Award Number 17STQAC00001-07-04, NSF SaTC-2241068, a Cisco Research Award, and a Microsoft Accelerate Foundation Models Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

REFERENCES

[1] 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* (2022).
[2] 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*. 2051–2055.
[3] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable rumor detection in microblogs by attending to user interactions. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 8783–8790.
[4] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018).
[5] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. 2019. The role of user profiles for fake news detection. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 436–439.
[6] Lichao Sun, Yue Huang, and Haoran Wang et al. 2024. TrustLLM: Trustworthiness in Large Language Models. arXiv:2401.05561 [cs.CL]
[7] Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM Web Conference 2022*. 2501–2510.
[8] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2021. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 1417–1423.
[9] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*. 3465–3476.
[10] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.