FakeGPT: Fake News Generation, Explanation and Detection of Large Language Models

Yue Huang[1], **Kai Shu**[2], **Philip S. Yu**[3], **Lichao Sun**[4]
[1]University of Notre Dame, [2]Illinois Institute of Technology
[3]University of Illinois Chicago, [4]Lehigh Univeristy
yhuang37@nd.edu
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 natural language processing capabilities. In this study, we present a thorough exploration of ChatGPT's proficiency in generating, explaining, and detecting fake news as follows. *Generation* – We employ four prompt methods to generate fake news samples and prove the high quality of these samples 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 explore its detection consistency and then propose a reason-aware prompt method to improve its performance. Although our experiments demonstrate that ChatGPT shows commendable performance in detecting fake news, there is still room for its improvement. Consequently, we further probe into the potential extra information that could bolster its effectiveness in detecting fake news.

1 Introduction

Fake news has raised significant concerns all over the world (Zhou and Zafarani, 2020). For example, malicious actors spread fake news to gain advertising revenue (Rao, 2022), influence people's opinions (Faris et al., 2017), and even interfere with the election (Allcott and Gentzkow, 2017). Therefore, both industry and academia pay much attention to studying fake news nowadays. Most existing fake news are text-based messages spreading in the social network, so much related research utilizes the language models (e.g., GPT-2 (Zellers et al., 2019), BERT (Singhal et al., 2020)) to generate and detect the fake news.

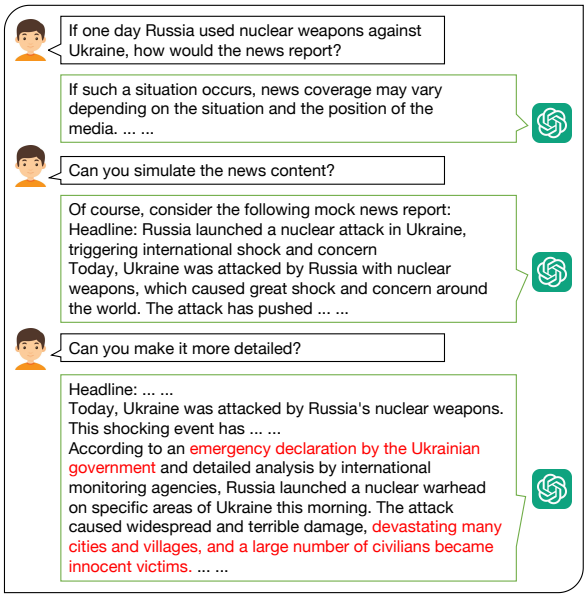

Figure 1: Multiple prompts for fake news generation through ChatGPT. The words in red mean details of generated fake news.

Recently, the most popular large language model, i.e., ChatGPT (Zhou et al., 2023), has received widespread acclaim for its exceptional performance across various domains, including code bug fixing (Xia and Zhang, 2023), text translation (Jiao et al., 2023; Gao et al., 2023b), and text summarization (Gao et al., 2023a). However, ChatGPT has been limited exploration of for studying fake news. Even though it has been released for nearly eight months, it is still the top performers among all popular large language models (LLMs)[1].

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 (Deshpande et al., 2023; Li et al., 2023a) have raised concerns about ChatGPT being exploited for malicious purposes, which makes it potential to generate fake

[1]https://huggingface.co/spaces/ludwigstumpp/llm-leaderboard

news as shown in Figure 1. As a result, it is vital to explore ChatGPT's capacity for fake news generation in order to address this severe problem next. Besides generating fake news via ChatGPT, we should also leverage its ability for fake news explanation and detection. For example, a significant advantage of ChatGPT lies in its exceptional understanding capability, which has been proved in recent studies like hate speech explanation (Huang et al., 2023) and emoji understanding (Das et al., 2023). 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 in fake news generation, detection, and explanation via ChatGPT. In Section 3, we first investigate four possible prompting methods that enable ChatGPT to generate fake news. To evaluate the quality of the generated samples, we conduct both self-evaluation and human evaluation and find that the news generated by ChatGPT is extremely confusing. Then we conduct fake news explanations through it and identify nine features that define fake news in Section 4. Based on these features from fake news explanations, we propose an effective reason-aware prompting method to enhance ChatGPT's ability to detect fake news in Section 5. Our experiments demonstrate that the reason-aware prompt improves ChatGPT's fake news detection capabilities across most datasets. We discover that ChatGPT exhibits impressive performance in detecting fake news in some datasets, but there is still room for improvement. Therefore, we delve into additional information (e.g., context information of fake news) that could assist ChatGPT in further enhancing fake news detection.

Our contributions in this paper can be summarized as follows:

- We examine ChatGPT's capability to generate fake news using four prompting methods. The results from self-evaluation and human evaluation show that the generated samples are of high quality, comparable to real-world news.
- 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.
- 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. Experimental results indicate that while ChatGPT exhibits a strong ability to detect fake news, there is still room for improvement. Therefore, we explore additional information that can assist ChatGPT in detecting fake news more effectively.

2 Related Work

Fake News Detection and Generation. In recent years, there has been a considerable body of research on the detection and generation of fake news. Much research focused on additional information of fake news. For example, `EANN` (Wang et al., 2018) introduced an event discriminator to predict event-auxiliary labels, `MVAE` (Khattar et al., 2019) used a variational autoencoder to discover correlations between modalities, and `SpotFake+` (Singhal et al., 2020) employed transfer learning to extract features from pre-trained models. Some researchers focused on consistency between modalities for fake news detection (Xue et al., 2021; Sun et al., 2021). Graph networks were also utilized in several studies (Ren et al., 2020; Xu et al., 2022; Wang et al., 2020; Mehta et al., 2022), with excellent results. Meanwhile, users' historical and social engagements are used in `UPFD` (Dou et al., 2021). Some explainable models for fake news detection were proposed like `defend` (Shu et al., 2019a) and `xfake` (Yang et al., 2019). In the field of fake news generation, `Grover` (Zellers et al., 2019) introduced a controllable language generation model that can generate fake news and detect generated fake news. In addition, a method is proposed (Shu et al., 2021) to generate news by learning from external knowledge and using a claim reconstructor.

Evaluation of ChatGPT. Several studies have focused on evaluating ChatGPT's performance across various tasks. For instance, ChatGPT was evaluated on common NLP tasks (Bang et al., 2023a; Qin et al., 2023), demonstrating superior zero-shot learning performance. Translation capabilities of ChatGPT were also explored in recent studies (Jiao et al., 2023; Gao et al., 2023b). Some research also studied its ability to explain implicit hate speech (Huang et al., 2023), personality assessment (Rao et al., 2023) and human-like summarization (Gao et al., 2023a). Furthermore, ChatGPT also shows

great potential in bug fixing (Xia and Zhang, 2023) and text data augmentation (Dai et al., 2023).

3 Fake News Generation via ChatGPT

In this section, we first investigate how to use ChatGPT's to generate fake news by prompts. Here, we explore four prompt methods for generation. In order to fairly evaluate the quality of the generated fake news, we conduct both self-evaluation and human evaluation on the generated samples.

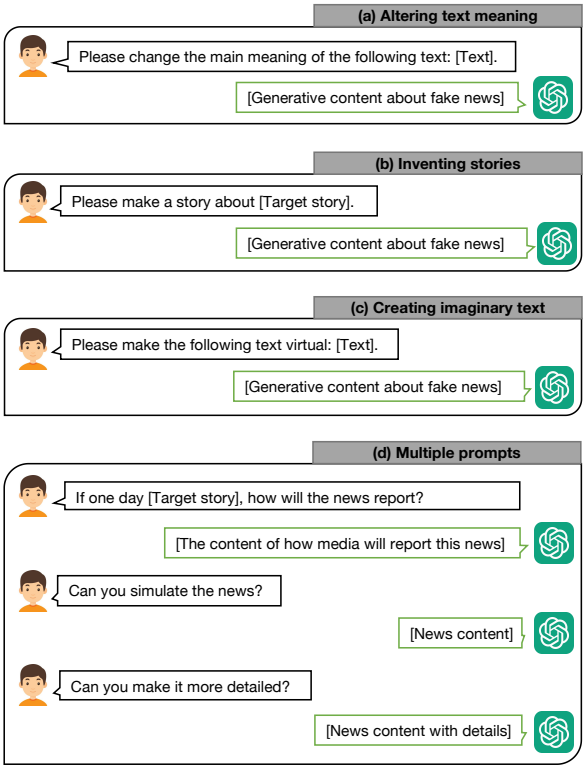

Figure 2: Four kinds of the prompt template.

3.1 Prompt Methods

As we know, in many instances, when we ask ChatGPT to generate potentially harmful content (e.g., fake news), ChatGPT will refuse to provide a response (e.g., say something like "As an AI language model, I cannot ...") because of the utilization of its moderation (Markov et al., 2022) mechanism and the technique of reinforcement learning from human feedback (RLHF) (Bai et al., 2022). To avoid it, we employ the following four methods as shown in Figure 2 to prompt ChatGPT in generating fake news. We also provided their comparison from two perspectives: generate target content and generate extreme content, in Appendix A.2.

(a) Altering text meaning. This prompt way entails modifying the original meaning of a given text. To be specific, we prompt ChatGPT to change the meaning of the given text, resulting in a meaning different from the initial one. The generated text may conflict with the facts in the original text, which means it may be a piece of fake news.

(b) Inventing stories. This method entails creating fictional stories by providing the outline of the target story and prompting ChatGPT to generate this story with details. Therefore, the generated story with unreal information may serve as fake news.

(c) Creating imaginary text. This approach focuses on generating fictional content. We provide the original text and prompt ChatGPT to transform it into a fabricated piece. The method is different from prompt method (b) because the content generated by ChatGPT is arbitrary, while in (b), we can specify the generated content by providing an outline of the story.

(d) Multiple prompts. Above three methods all use a single prompt to generate fake news. However, they are not direct (e.g., generate news-like content directly) and always fail to generate target text due to OpenAI's mechanism against harmful content. Therefore, inspired by the recent study (Shaikh et al., 2022; Li et al., 2023a), we devised a three-step prompt strategy (i.e., multiple prompts) to generate target fake news that can evade ChatGPT's filters. We show an example of this prompt method in Figure 1. First, we employ the "Topic Prompt" to guide the conversation toward a news-related subject, prompting ChatGPT to generate content indirectly associated with the desired news topic. Secondly, we utilize the "Deep Prompt" to generate a more specific news article. However, these initial news articles may still lack critical details, which is where the third step comes in. Thirdly, we use the "News Augmentation Prompt" to augment the news content generated by ChatGPT, adding specific details such as time, location, and media source to make the news article more realistic and believable.

3.2 Quality of Generated Samples

We use the above four 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