Understanding the Success and Challenges of Fake News Generation Prompts: Insights from Language Model Experiments

Anonymous ACL submission

Abstract

Automatic disinformation detection is crucial in preventing the spread of false information, especially in the medical field where individuals lacking domain expertise may pursue incorrect treatment approaches. With the advent of Large Language Models (LLMs), several studies are exploring different LLM-led attacks on news content. However, many of these articles generate fake news using monotonous and simplistic prompts or incorporate supplementary materials that deviate from real-world fake news generation scenarios. In this study, we **(1)** examined successful fake news generation prompts in prior research, **(2)** proposed a novel variational autoencoder (VAE)-like prompt (**VLPrompt**) designed to emulate authentic fake news creation, and **(3)** assessed the effectiveness of different language models in detecting fake news generated by LLMs. Our findings indicate that **(1)** fake news generated with various prompts exhibits varying degrees of success in deceiving both machines and humans, **(2)** fine-tuned LLM models demonstrated strong performance in detecting fake news produced by LLMs. We also conducted extensive studies to identify news articles susceptible to LLM-modification attacks and to characterize the typical appearance of these modified articles.

1 Introduction

Disinformation usually refers to deliberately spreading false or misleading information to deceive, manipulate, or mislead people. The spreading of disinformation, particularly in the context of medical information, can have profoundly adverse consequences on both society and individuals, as indicated in (Bondielli and Marcelloni, 2019; Gao et al., 2020; Guo et al., 2020; Zubiaga et al., 2018).

Unlike misinformation, which encompasses any information that deviates from the facts, disinformation is intentionally crafted with malicious motives (of Oregon Library). Therefore, disinformation is more likely to be propagated and lead to negative consequences (Cialdini and Goldstein, 2004; Vosoughi et al., 2018). In this study, we aim to (1) simulate a real-world scenario where malicious individuals employ Large Language Models (LLMs) to create fake news and (2) assess whether machines and humans can distinguish these generated articles from the real ones. Specifically, we identify three factors that fake news creators pursue in a real-world scenario. Firstly, the fake news generation process should be end-to-end, meaning that the articles should be conducted without the need for manually created or labeled content. Furthermore, the generated fake news should effectively convey manipulative intentions to achieve specific objectives. Lastly, these fake news articles should closely resemble authentic news articles in terms of both content and writing style.

Based on the assumptions above, we propose three research questions:

RQ1: Can LLMs act as end-to-end fake news generators to generate high-quality deceptive articles? The experiments of (Wang et al., 2023) and (Sun et al., 2023) show that LLM-generated fake news can be easily detected through fine-tuning pre-trained small language models (SLMs). However, fake news generated with additional materials shows different conclusions (Huang et al., 2022; Wu and Hooi, 2023; Vykopal et al., 2023; Huang and Sun, 2023). This huge gap between the two generation strategies motivates us to develop prompts that enable LLMs to become end-to-end deceptive fake news generators.

RQ2: Can LLM-generated fake news convey malicious intentions and emulate real news' writing style? How should we evaluate these factors? For creators of fake news, it is crucial to maintain a consistent intention and replicate the writing style of authentic news articles. The efficacy of their content depends on their ability to effectively convey their intended messages. Ad-

ditionally, adopting a professional writing style can significantly reduce the likelihood of detection by traditional methods (Zhou and Zafarani, 2020; Pagnoni et al., 2022; Wu and Hooi, 2023). Hence, from the perspective of fake news creators, controlling these two factors during the generation process becomes of paramount importance.

RQ3: How can we evaluate the quality of generated fake news? Is there a pattern in LLM-generated fake news articles? Previous works (Vykopal et al., 2023; Su et al., 2023) that generate articles from summaries have suggested that their methods often result in articles missing critical details. Similarly, it is likely that articles generated using various other prompts exhibit distinct patterns. Recognizing these patterns can assist in detecting whether articles are the product of LLMs and also guide LLMs in generating articles that closely resemble human-authored content.

To answer the three research questions mentioned above, we **(1)** transferred two prompt methods that proved to be effective in fake news generation with supplementary materials; **(2)** proposed a VAE-like prompt schema to adapt to our proposed real-world scenario; and **(3)** assessed the ability of the generated fake news articles to deceive both machine detectors and human readers.

2 Dataset Collection

This section focuses on how human-generated real and fake news articles are collected. The LLM fake news generation method will be introduced in Section 3.1. We gathered authentic news articles from authority medical news websites, including "ClevelandClinic" (ClevelandClinic, 2023), "NIH" (NIH, 2023), "WebMD" (WebMD, 2023), "Mayo" (Mayo, 2023), "Healthline" (Healthline, 2023), and "ScienceDaily" (ScienceDaily, 2023). The fake news are collected from authority fact-checking websites including "AFPFactCheck" (AFPFactCheck, 2023), "CheckYourFact" (CheckYourFact, 2023), "FactCheck" (FactCheck, 2023), "HealthFeedback" (HealthFeedback, 2023), "LeadStories" (LeadStories, 2023), and "PolitiFact" (PolitiFact, 2023). The publish dates of the articles span from Jan-01-2017 to May-01-2023. For the collected articles, we sorted diseases by number of relevant articles and kept fifteen disease categories that contained more than fifty real news articles. The statistical findings are shown in Appendix A.1.

3 Methodology

3.1 Fake News Generation

All the fake news articles are generated from a single real news. Our fake news generation prompt follows a chain of thought (CoT) (Wei et al., 2022) template by instructing the LLM text generation process step by step. As shown in Figure 1, the first method (Summary) is based on the method proposed in (Su et al., 2023) that generates an article from a summary. In our modified approach, instead of using a human-written summary, we directed the LLM to produce a summary of actual news, which we then altered. Subsequently, the fabricated news is generated based on this modified summary. The second method (Question-Answer) is inspired by (Pan et al., 2023). This process starts by formulating a question about the real news, followed by answering and modifying this answer. An article is then generated using the revised answer. To ensure the significance of modification, we re-ask the question using the new article; if the answer remains unchanged, the article is discarded.

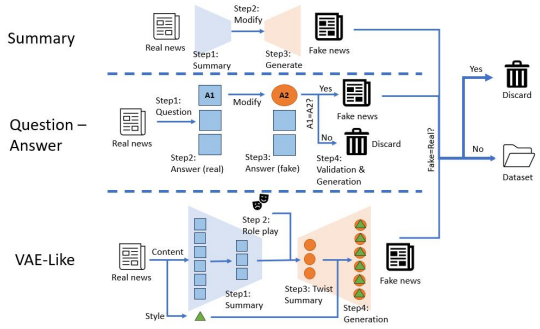

Figure 1: **Workflow of the three prompt strategies for LLMs to generate fake news.**

Similar to the two methods above, the proposed VLPrompt also asks LLM to extract critical information from the real news, modify this information, and then create a counterfeit news article using the altered data. However, VLPrompt also incorporates a role-play module to enrich the intentions behind the modification and a style-and-length control module to adhere to desired writing styles. Finally, we asked LLM to go through all the real-fake news pairs and examine if they are the same in factors and opinions and discard the pairs with subtle differences. If the answer is “same”, then the article is put in the “discard” category, otherwise the “dataset” category. The detailed prompts can be seen in Appendix A.2.

3.2 Supervised Fine-Tuning

For SLMs, we employed a trainable two-layer feed-forward neural network to transform the latent representation of news articles into binary classification results. However, this approach is not suitable for LLMs since most LLMs are primarily designed and pre-trained for text generation, making it challenging to evaluate their outputs using binary classification metrics.

To fill this gap, we provided a prompt with an exemplar article-label pair to guide the LLMs during the supervised generation fine-tuning process. This design follows the recommendation of (Gao et al., 2020) for prompt-based fine-tuning. Within the prompt, the template regularizes the generation format and enables the loss computed only over the conclusion (i.e., "real" or "fake") drawn by LLM. The example provides supplementary context to assist the LLM in identifying facts/counterfacts in the other articles. This design reflects the real-world scenario in which fact-checkers try to identify if an incoming article is real or fake. The fact-checkers may have access to labels for their archived articles, but may not necessarily be aware of their relations to the incoming article. Therefore, they randomly select one article as an example instead of picking out the corresponding real news articles, considering that looking for relevant articles could be time-consuming. After the supervised fine-tuning is finished, the model should be able to classify an incoming article by leveraging both the facts in the training data and the relation between the example article and the article to be classified. The detailed prompts are provided in Appendix A.2.

3.3 Human Study

Our human study comprises three objectives: (1) verifying the authenticity of fake news articles in our dataset, (2) assessing the deceiving effectiveness of these articles, and (3) evaluating their overall quality. We selected samples from the dataset to achieve these three objectives and invited experts to evaluate them individually. Detailed processes and instructions for this quality assessment are provided in Appendix A.4.

4 Experiment and Analysis

4.1 Deceptive Power over Humans

During fake news generation, we randomly selected $2,000$ news articles from the real news set and attempted to evenly allocate to different LLMs and different prompt strategies. However, after automatic and manual validation, only a few of the generated articles were used for the experiments. The detailed components of the dataset are shown in Appendix A.1. In Table 1, "SUMMARY" and "QA" represent different prompt strategies, "Summary" and "Question-Answer". The prompts for the other experiments are VLPrompt or the ablation versions of VLPrompt. Specifically, the three experiments named with LLMs indicate Vicuna 13b (Chiang et al., 2023), ChatGPT 3.5, and ChatGPT 4 (OpenAI, 2023). "ab_role" means VLPrompt without role-play module, "ab_sem" means VLPrompt without the style control module.

For the first task of human study, we can conclude that the LLM validation process is promising and our following experiment results are convincing. For the second and third tasks, we included ten more real news pieces. The correctness in Table 1 refers to the proportion of fake news articles that were accurately identified by human evaluators. The results show that "ab_role" was the most effective in misleading participants. Besides, "ab_role" stood out in both "representing malicious intentions" (intention) and maintaining "full-text theme consistency" (consistent). Besides, all the generated articles achieved high scores in neutral and informative. This suggests that the LLMs generally retain the original articles' writing style and specific details, even without explicit instructions to do so. These observations answer **RQ2**, indicating that incorporating prompts related to intention and writing style has minimal impact on the content generated by the models.

4.2 Deceptive Power over Machines

To assess the machine's power in detecting LLM-generated fake news. We considered three fine-tuned BERT-based SLMs (BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), and FakeNews-BERT(ungjus, 2023)), two LoRa (Hu et al., 2021) fine-tuned 7b LLMs (Llama2-7b (Touvron et al., 2023) and Vicuna-7b (Chiang et al., 2023)), two state-of-the-art (SOTA) fake news detection models, and an enterprise-level LLM (ChatGPT 3.5 (OpenAI, 2023)). See Appendix A.3 for details. The first four baselines are fully tuned until converge (not improved within 50 epochs) and the LLMs are tuned with three epochs. The results show that the fine-tuned LLMs outperform the other models in detecting LLM-generated fake

Model	GPT 3.5	GPT 4	Vicuna	QA	SUMMARY	ab_role	ab_sem
Correctness	0.250	0.436	0.344	0.383	0.281	**0.125**	0.563
Intention	0.594	0.750	0.531	0.625	0.656	**0.781**	0.656
Detail	0.750	0.719	0.469	0.677	0.844	0.844	0.906
Neutral	0.906	0.936	0.906	0.938	0.875	0.875	**0.969**
Informative	0.706	0.875	0.875	**0.931**	0.688	0.813	0.688
Consistent	0.781	0.688	0.875	0.844	**0.938**	**0.938**	0.781

Table 1: **Human study scores of fake news articles generated by different methods.** Rows list metrics used for human study evaluation while the columns indicate news from different sources. The numbers in the table are the mean scores from two people. Bolded and blue results indicate the best and second best.

Model	ACC	F1	PRC	RCL
BERT	0.865	0.778	0.852	0.716
RoBERTa	0.901	0.860	0.804	0.926
FnBERT	0.833	0.688	0.892	0.560
Grover	0.921	0.939	0.913	0.966
DualEmo	0.928	0.946	0.943	0.950
Llama2-7b + LoRa	**0.971**	**0.955**	**0.986**	0.927
Vicuna-7b + LoRa	0.969	0.952	0.974	**0.931**
ChatGPT 3.5	0.769	0.510	0.850	0.364

Table 2: **Fake news classification results of fine-tuned SLMs, fake news detection models, LoRa fine-tuned LLMs, and chatGPT 3.5 turbo API.** In the header, "ACC" means accuracy, "PRC" means precision, and "RCL" means recall. The best is bolded.

news.

4.3 Generator Evaluation and Case Study

To examine if any of the prompt strategies outperform the others, we performed ANOVA (Girden, 1992) over all the validation metrics for each prompt and found a significance in distribution means (F-value = 2.35, p-value = 0.04). Consequentially, we performed Turkey's Honestly Significant Difference (HSD) test over each pair of distributions. The detailed results are in Appendix A.4. The results show that the detection performances of "QA", "SUMMARY", and "ab_role" are different from the others, meaning that the articles generated with these prompts are harder to detect. In the third task in our human study experiment, the human evaluators compared 70 fake news articles with their source real news articles and found some patterns. For example, LLMs tend to inverse opinions instead of editing details, as indicated in the case Figure 2. The generated article altered the statement "*drinking more water can reduce the risk of heart failure*" to "*increase the risk of heart failure*" (highlighted in red) while preserving the writing style and content irrelevant to the statement (highlighted in green). This deliberate manipulation makes the fake article confusing to both human and automatic fake news detectors. This answers **RQ3** by indicating that people may be able to identify LLM-generated fake news by checking the frequency of negative words.

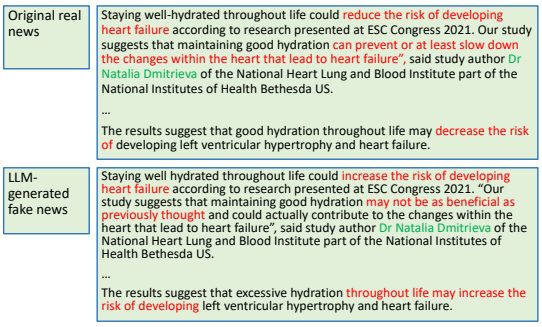

Figure 2: **An example of the comparison of a real news article and corresponding VLPrompt-generated fake news.** Phases highlighted in red are modified statements and those highlighted in green are unmodified factors.

5 Conclusion

In our research, we assess the capability of VLPrompt and two other prompting techniques to produce fake news articles that are capable of misleading both human evaluators and automated detection systems. The experimental findings reveal that given a single reference news article, LLMs can effectively generate and rationalize novel viewpoints, while skillfully imitating the style and length of the original piece. Notably, these LLM-generated articles can deceive humans and are more adept at deceiving detection algorithms than those crafted by humans. However, our ablation and case studies also demonstrate that LLMs exhibit consistent generation patterns, which are relatively unaffected by the prompt strategies. These patterns could serve as markers to identify LLM-generated fake news.

6 Ethical Consideration

To ensure there is no potential harm or adverse impact on the medical industry and journalism, we conducted the fake news generation phase exclusively on our own machines. Our experimental activities do not pose any threat to these sectors. We would not release the codes or the generated articles from this phase to avoid increasing the burden on fact-checking efforts. Additionally, we are committed to maintaining confidentiality regarding the list of news articles that may be vulnerable to this vulnerability.

7 Limitation

This research proposes a VAE-like jailbreaking prompt for fake news generation and proves the harm introduced by VLPrompt-generated disinformation. There are primarily two limitations in this research. Firstly, we can potentially provide more detailed instructions to LLMs to restrict the fake news detection task to a binary classification task. Secondly, we can examine more state-of-the-art fake news detection models with novel architectures other than the standard language models used in this paper. However, despite the issues mentioned above, our experiment results exhibit sufficient evidence to support our assumption that VLPrompt-generated fake news poses a significant threat to current news fact-checking systems.

References

AFPFactCheck. 2023. Afpfactcheck. https://factcheck.afp.com/.

Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55.

CheckYourFact. 2023. Checkyourfact. https://checkyourfact.com/.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.

ClevelandClinic. 2023. Clevelandclinic. https://newsroom.clevelandclinic.org/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

FactCheck. 2023. Factcheck. https://www.factcheck.org/.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

Ellen R Girden. 1992. *ANOVA: Repeated measures*. 84. Sage.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36.

HealthFeedback. 2023. Healthfeedback. https://healthfeedback.org/.

Healthline. 2023. Healthline. https://www.healthline.com/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2022. Faking fake news for real fake news detection: Propaganda-loaded training data generation. *arXiv preprint arXiv:2203.05386*.

Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*.

LeadStories. 2023. Leadstories. https://leadstories.com/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: a robustly optimized bert pretraining approach (2019). *arXiv preprint arXiv:1907.11692*, 364.

Mayo. 2023. Mayo. https://newsnetwork.mayoclinic.org/.

NIH. 2023. Nih. https://www.nih.gov/.

University of Oregon Library. Fake news and information literacy. https://huggingface.co/ungjus/Fake_News_BERT_Classifier.

OpenAI. 2023. Chatgpt 3.5. https://chat.openai.com/chat.

Artidoro Pagnoni, Martin Graciarena, and Yulia Tsvetkov. 2022. Threat scenarios and best practices to detect neural fake news. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1233–1249.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the risk of misinformation pollution with large language models. *arXiv preprint arXiv:2305.13661*.

PolitiFact. 2023. Politifact. https://www.politifact.com/.

ScienceDaily. 2023. Sciencedaily. https://www.sciencedaily.com/.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2023. Adapting fake news detection to the era of large language models. *arXiv preprint arXiv:2311.04917*.

Yanshen Sun, Jianfeng He, Shuo Lei, Limeng Cui, and Chang-Tien Lu. 2023. Med-mmhl: A multi-modal dataset for detecting human-and llm-generated misinformation in the medical domain. *arXiv preprint arXiv:2306.08871*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

ungjus. 2023. Fn-bert. https://huggingface.co/ungjus/Fake_News_BERT_Classifier.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2023. Disinformation capabilities of large language models. *arXiv preprint arXiv:2311.08838*.

Zecong Wang, Jiaxi Cheng, Chen Cui, and Chenhao Yu. 2023. Implementing bert and fine-tuned roberta to detect ai generated news by chatgpt. *arXiv preprint arXiv:2306.07401*.

WebMD. 2023. Webmd. https://www.webmd.com/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jiaying Wu and Bryan Hooi. 2023. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.

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*, pages 3465–3476.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 51(2):1–36.

A Appendix

A.1 Detailed Disease-News Relations

See the Table 3 below.

A.2 Prompt Detail

Rewrite the news article below quoted by {delimiter}, so that the generated article holds different opinions against the original article.

Always start the content of the article with:

Article: {delimiter}\n

Figure 3: **The prompt for fake news generation (SUMMARY).** The reference article is concatenated to the end of the prompt text.

Follow these steps to analyze the news articles. The contents of the article will be delimited with four hashtags, i.e. {delimiter}.

Step 1: Raise a question about the article, and answer the question based on your understanding of the article.

Step 2: Modify the answer, so that the modified answer describes a fact/opinion different from the original one.

Step 3: Rewrite the article, modify as few details as possible to make it support the opinion generated in the last step. Start the content of the generated article in Step 3 with: Article: {delimiter}\n

Step 4: What is the answer to the question in Step 1 based on the generated article?

Step 5: Is the answer in Step 1 the same as the answer in Step 4?

Figure 4: **The prompt for fake news generation (QA).** The reference article is concatenated to the end of the prompt text.

A.3 Baseline Detail

During the experiment, we considered four types of baseline models, fine-tuned SLMs, fakenews detection models, LoRa fine-tuned LLMs, and ChatGPT 3.5. The characteristics of the chosen models are listed below. Except for ChatGPT 3.5, the other models are extracted from the Hugging Face repositories.

- BERT (Devlin et al., 2018): A bi-directional transformer model pretrained on a large corpus of English data in a self-supervised fashion.
- RoBERTa (Liu et al., 2019): BERT enhanced with more data, dynamic mask, and byte-pair encoding.
- FN-BERT (ungjus, 2023): A BERT-based

Follow these steps to analyze the news articles. The contents of the article will be delimited with four hashtags, i.e., ####.

Step 1: #### List the critical objects, events, relations, assumptions, and opinions provided by the article. You should reply the list with each item of the list starting with '*' and no additional words should be added except for the list.

Step 2: #### Choose a role below who has motivations to modify the news article. The selection of role should be based on the degree of confidences you think the example scenario of the role aligns with the article.

1. Financial Gain: This guy works for a hostile organization of interests. Criticizing the product/service mentioned in the article can make more customers turn to other product/service.
2. Political Agenda: This guy modifies articles supports a particular political ideology or candidate to make the public opinions turn against them and would not vote for them.
3. Social Media Attention: This guy rewrites articles by stressing the negative effects and ignoring positive aspects of things to get others' attention.
4. Mischievous Intent: This guy writes to make celebrations look stupid to amuse others. For example, the president farts a lot during the meeting with the UN.
5. Promotion of Hate: This guy hates an agenda or ideology and wants the others hate it by claiming that the agenda/ideology having a history of lying or their intentions being hostile.
6. Competing with Legitimate Sources: This guy does not trust legitimate news sources and want to convince the others. He diverts attention away from accurate reporting.
7. Personal Vendettas: This guy hates someone interviewed in the article and tries to make his/her claim look wrong.
8. Ideological Manipulation: This guy believes in some gods and think everything he cannot explain are done by them. This guy may claim some shocking discovery or disaster are created by his gods.

Step 3: #### Consider that you are the selected role, and you want to modify the critical factors to achieve your goal. There are some rules to follow during modification:

1. The modifications have to change the meaning of the factors.
2. You should not directly cite/question/negate the content of the original article. Rewrite the modified factor as if you are the first to report it.
3. All the modifications should serve the same conclusion. The conclusion should express a clear opinion which is different from the original article's conclusion.
4. The objectives of the modifications should be logically consisted.

Let's think step by step, what would you do to modify the article?

Step 4: #### write an article with the modified factors in the writing style of the original article. The article should be of the similar length as the original article.

Always start your answer with: Article: #### \n

Figure 5: **The prompt for fake news generation (VL-Prompt).** The reference article is concatenated to the end of the prompt text.

System:

Do you think the news article below is real news or fake news? The two articles delimited with four hashtags, i.e., ####.

User 1:

####{article 1}####

Assistant 1:

{label 1}

User 2:

####{article 2}####

Assistant 2:

{label 2}

label 1, label 2 ∈ {'real', 'fake'}

Figure 6: **The prompt for fake news detection model fine-tuning.** {article 1} is the text of the example article and {label 1} is the one-word label ("real" or "fake") for article 1. {article 2} is the incoming article to be classified and {label 2} is the label to be predicted.

Info. Type	anemia	arthritis	asthma	cancer	covid	diabetes	epilepsy	flu	headache	hypertension	inflammation	monkeypox	parkinson	pneumonia	stroke	Total
Real news	62	85	148	1,410	859	332	48	740	70	55	282	44	81	50	286	4,554
Fake news (Human)	0	1	0	27	304	1	2	114	1	0	4	3	0	2	10	469
Fake news (GPT 3.5)	3	12	17	266	89	53	3	93	10	14	48	0	15	8	59	690
Fake news (GPT 4)	2	5	4	70	46	15	4	32	0	6	12	3	4	0	13	219
Fake news (Vicuna)	1	4	6	63	39	12	4	25	3	3	14	2	3	2	9	190
Total news	68	107	175	1,836	1,337	413	61	1,004	84	78	360	52	103	62	377	6,117

Table 3: Statistics between diseases and news articles.

model recently finetuned on a Fake news classification dataset in 2023.

- Grover (Zellers et al., 2019): A GAN-based model containing multiple transformer modules. We trained only the discriminator part with our data.
- DualEmo (Zhang et al., 2021): A BiGRU-based model that utilizes emotion feature extraction techniques to assist fake news detection.
- Llama2 (Touvron et al., 2023): A prominent open-source LLM fine-tuned with Reinforcement Learning from Human Feedback (RLHF) technique. It has been proven to exhibit competitive performance compared to enterprise-level LLMs with relatively small parameter volumes.
- Vicuna (Chiang et al., 2023): An open-source LLM fine-tuned mainly with imitation learning from ChatGPT 4. It has been proven to exhibit competitive performance compared to enterprise-level LLMs with relatively small parameter volumes.
- ChatGPT 3.5 (OpenAI, 2023): The most widely used iteration of OpenAI's powerful language model. ChatGPT 3.5 turbo API was employed in this research.

A.4 Human study Processes and Performace Evaluation Details

A.4.1 Human study processes

For the first task, we randomly sampled 70 articles from the "dataset" category and 70 from the "discard" category respectively (ten articles from each generation method, discussed in Section 4). Two experts were then tasked with determining whether the fake news articles were identical to their original, real news counterparts.

For the latter two goals, we sampled another 80 articles from the "dataset". The experts conducted their review in two stages: first, they classified each article as either "fake" or "real"; then, in the second stage, they evaluated the quality of the generated articles.

A.4.2 Human study instruction

The human study instructions for different metrics are as follows:

- **Correctness:** Does the article look like fake news (0) or real news (1)?
- **Intention:** Knowing the news is fake, do you think there is an obvious malicious intention behind the modification (1)? Or is it just randomly replacing factors (0)?
- **Detail:** Knowing the news is fake, compared with the original news, does it change the theme (1), part of the theme (0.5), or just some details (numbers, terms, etc.) (0)?
- **Neutral:** Does the article seem to be written from a neutral point of view (1) or using some emotive words (0)?
- **Informative:** Does the news support its point of view with concrete case (i.e., instances with exact numbers, names of people, time, location, etc.) (1) or just repeating general claims (0)?

- **Consistent:** Does the article have a main idea (1) or does it talk about several unrelated ideas (0)?

- **Comment:** Have you noticed anything special characteristics or patterns in LLM-generated fake news?

A.4.3 Significance test

Table 4 shows the Tukey's Honestly Significant Difference (HSD) test result. The attributes used for the test are source-wise baseline performances and human study scores. The results show that the fake news generation methods, "QA", "VAE-like (no roly-play)", and "summary" are of different performance distributions against the other models, i.e., the 'reject" results are "True".

group1	group2	meandiff	p-adj	lower	upper	reject
QA	summary	0.0977	0.859	-0.1323	0.3278	False
QA	ab_role	0.0415	0.9981	-0.1885	0.2716	False
QA	ab_sem	0.3155	0.0015	0.0854	0.5455	True
QA	gpt35	0.3602	0.0002	0.1301	0.5902	True
QA	gpt4	0.3318	0.0007	0.1018	0.5619	True
QA	vicuna	0.3511	0.0003	0.1211	0.5812	True
summary	ab_role	-0.0562	0.9899	-0.2863	0.1739	False
summary	ab_sem	0.2178	0.0758	-0.0123	0.4478	False
summary	gpt35	0.2625	0.0148	0.0324	0.4925	True
summary	gpt4	0.2341	0.0433	0.0041	0.4642	True
summary	vicuna	0.2534	0.0212	0.0234	0.4835	True
ab_role	ab_sem	0.274	0.0093	0.0439	0.504	True
ab_role	gpt35	0.3187	0.0013	0.0886	0.5487	True
ab_role	gpt4	0.2903	0.0046	0.0603	0.5204	True
ab_role	vicuna	0.3096	0.002	0.0796	0.5397	True
ab_sem	gpt35	0.0447	0.9971	-0.1853	0.2748	False
ab_sem	gpt4	0.0164	1.0	-0.2137	0.2464	False
ab_sem	vicuna	0.0357	0.9992	-0.1944	0.2657	False
gpt35	gpt4	-0.0283	0.9998	-0.2584	0.2017	False
gpt35	vicuna	-0.009	1.0	-0.2391	0.221	False
gpt4	vicuna	0.0193	1.0	-0.2108	0.2494	False

Table 4: Tukey's HSD test over different fake news sources based on baseline performances and human study evaluations.