How Large Language Models Write FakeNews Like Humans Do

Anonymous ACL submission

Abstract

Fake news detection is a challenging task in the field of natural language processing. The existing methods for detecting manually written fake news still face significant challenges, primarily due to the lack of fake news datasets that imitate human writing styles. Manual methods often require significant human and material resources, while automated methods generate fake news that diverges from human writing styles. To address these challenges, we propose a novel framework based on Large Language Models (LLM) for generating human-like fake news datasets. Specifically, we first use a large language model to generate two styles of fake news that contradict the main points of the real news article. Subsequently, the large language model selects the better of the two generated fake news sentences based on the specified evaluation criteria and replaces the main sentence of the original news article, thus constructing fake news while maintaining a human-written style. Our approach effectively addresses the challenges of constructing fake news datasets and ensures closer adherence to human writing styles. Additionally, it provides insights into enhancing the human-like writing capabilities of LLM. We will release the LLMFAKE dataset constructed using this method, which contains approximately 2.8k examples. Our experimental results demonstrate that fake news detectors trained on LLMFAKE outperform previous baseline methods on two human-written fake news datasets.

1 Introduction

Targeted fake news, crafted by individuals or organizations for economic or political gain, has had devastating impacts on numerous social events(Shu et al., 2017). Therefore, there is an urgent need for a detection technology to defend against human-crafted fake news(Kong et al., 2020). Although deep learning-based methods have made significant breakthroughs in the field of fake news detection,, there is still significant room for improvement in detecting human-written fake news. The performance of these models often relies heavily on high-quality datasets, and constructing datasets to defend human-written fake news is a complex and challenging task.

In previous studies,the most common approach involves using APIs or web scraping techniques to gather posts and news links from news portals and social media platforms such as Twitter and Facebook(Shu et al., 2018; Dzienisiewicz et al., 2024; Mitra and Gilbert, 2021). The collected news and user-generated content are then labeled for authenticity through manual annotation. Alternatively, data can be collected by scraping pages from well-known fact-checking websites such as PolitiFact and Snopes to directly obtain the corresponding text and authenticity labels(Wang, 2017). However,these approaches require significant human and material resources. To address the aforementioned challenges, many studies have proposed using pretrained models to automatically generate fake news. For instance, Grover(Zellers et al., 2019)generates fake news based on a given part of the news, such as the summary, article, title, date, etc. FACTGEN(Shu et al., 2020)generating fake news that appears more realistic and factually grounded. Bust the generated fake news lacks a human-like writing style. PROPANEWS(Kung-Hsiang et al., 2023)generates fake news closer to human-written content by selectively falsifying parts of sentences and employing automated propaganda techniques.However, its automated propaganda methods are limited to just two types, and the generated text lacks depth in knowledge. Although the methods mentioned above can significantly reduce human resource input, the cost of training the corresponding models is extremely high, as fake news generation models need to be trained on specifically designed datasets. Moreover, these methods often suffer from poor generalization

when generating fake news, typically only being able to produce fake news within the same domain as the training data.

Inspired by the outstanding performance of LLMS in various text generation tasks(Xuanfan and Piji, 2023; Tu et al., 2024), we propose an automated fake news generation method based on large language models.

For a given real news, our framework first employs an extractive summarization model to identify the central theme sentence. By replacing this key sentence, the method generates fake news while preserving much of the original structure and writing style. Next, the LLM is used to deeply manipulate the central theme sentence. To produce stylistically diverse fake news sentences, we utilize two carefully designed and significantly different prompts to guide the LLM in generating fake content. These prompts are designed from distinct semantic perspectives, linguistic style tendencies, and logical construction directions. This approach encourages the LLM to create fake news sentences with a wide range of expressions, emotional tones, and narrative styles, enhancing the diversity and deceptive nature of the fake content. Finally, we use a new prompt to instruct the LLM to filter the fake content generated in the previous step, retaining sentences that are closer to human writing styles. Specifically, evaluation criteria are incorporated into the prompt provided to the model, enabling it to effectively select the most suitable fake sentence. The selected fake sentence is then used to replace the original central theme sentence, resulting in a hybrid fake news article that combines real and fake elements. We compare our method with state-of-the-art fake news generation techniques. The evaluation results on two human-written fake news datasets indicate that detectors trained using our generated fake news dataset perform better in identifying human-crafted misinformation. On both datasets, the AUC score increased by at least 7%, F1 and Accuracy increased by at least 5%.

Overall, the main contributions of this paper can be summarized as follows:

(1) We propose an automated fake news generation framework that does not require manual verification or the training of new models.

(2) We propose a method that enables LLM to generate text closer to human style while reducing low-quality text caused by its instability.

(3) Experimental results demonstrate that detectors trained on our generated data are more effective at detecting human-written misinformation, showing significant performance improvements compared to previous results. Additionally, we provide detailed ablation studies and analyses to illustrate the effectiveness of our method.

(4) We release LLMFAKE, a dataset for misinformation detection, which includes approximately 2.8k articles generated by our method. Used for training models to detect human-written fake news.

2 Related Work

2.1 Fake News Detection

The explosive growth of fake news, along with its erosion of democracy, justice, and public trust, has heightened the demand for robust fake news detection and intervention(Ahmed et al., 2021). In the field of machine learning, various classifiers have been employed to detect fake news, including SVM(Singh et al., 2017),Naïve Bayes(Pratiwi et al., 2017),Decision Tree(Kotteti et al., 2018),Logistic Regression(Kaur et al., 2020),Random Forests(Ni et al., 2020).

In the deep learning domain, numerous approaches have continually emerged as well. Early methods in deep learning employed RNNs to classify the veracity of news articles(Jadhav and Thepade, 2019). Subsequently, approaches using VAE to encode textual information and obtain embedded representations(Cheng et al., 2020). GCAN(Lu and Li, 2020)constructs a user graph based on user profiles, leveraging user information for fake information detection. Other methods model the rumor propagation process as a tree structure(Ma et al., 2018) for classification. Additionally, some models compare external knowledge with news content to identify misinformation(Hu et al., 2021).

The performance improvements in the aforementioned fake information detection models largely rely on corresponding enhancements to the underlying datasets. Consequently, there is a pressing need for more comprehensive and refined datasets to further advance the effectiveness of detection models.

2.2 Fake News Generation

In recent years, the construction of misinformation datasets, as a means to assist in combating erroneous and deceptive content, has garnered significant attention in the NLP research community. Early efforts involved using APIs to crawl news content from popular social media plat-

forms, followed by human fact-checking to construct datasets, for example, from Twitter(Zubiaga et al., 2015; Ma et al., 2017),Facebook(Tacchini et al., 2017; Williams and Santia, 2018), Weibo(Ma et al., 2016). Although the aforementioned datasets cover extensive time periods and broad domains, they still demand a large number of human reviewers to perform fact-checking on the collected news. Subsequently, to reduce the amount of manual verification required, researchers began constructing datasets by collecting fake news from fact-checking websites such as PolitiFact, GossipCop, and Snopes(Shu et al., 2018; Nguyen et al., 2022),or utilized Wikipedia as an information source for verification (Thorne et al., 2018). While these methods reduce the manual fact-checking workload, the size of the datasets they can construct remains limited. Hence, the challenge of easily obtaining a sufficiently large fake news dataset remains unsolved.

Recent automated generation approaches have begun relying on pre-trained models(Zellers et al., 2019; Fung et al., 2021; Shu et al., 2020). There are also frameworks that produce human-like fake news by replacing the central theme sentence and incorporating propaganda techniques(Kung-Hsiang et al., 2023). These methods can easily generate fake information. However, there remains a significant quality gap between the automatically generated fake news and the carefully crafted human-written fake news.

2.3 Large Language Model

In recent years, LLM have achieved substantial progress in various NLP tasks,including text generation,few-shot learning,and reasoning tasks(Touvron et al., 2023; Du et al., 2022; Yang et al., 2023). they have also achieved very promising results in natural language understanding(Yoo et al., 2021; Meng et al., 2022). Moreover,large language models exhibit capabilities that smaller models do not possess(Wei et al., 2022):TruthfulQA(Rae et al., 2021), MMLU(Hendrycks et al., 2020), WiC (Pilehvar and Camacho-Collados, 2019).

Prompting has become a primary approach for leveraging LLM to tackle a wide range of tasks(Liu et al., 2023). By designing suitable prompts, it is possible to directly leverage LLM to perform tasks without re-tuning the model or training new parameters. A well-constructed prompt is highly beneficial for enhancing the ability of large language models to accomplish specific tasks. There are numerous recommendations and guidelines for prompt design(White et al., 2023; Karmaker Santu and Feng, 2023). For instance, guidelines include using explicit instructions to articulate the task objective (Ouyang et al., 2024), decomposing tasks into more manageable steps, and employing model-friendly formatting. However, due to the content moderation mechanisms of LLM , there are still gaps in certain generation domains, such as fake-news.

3 Method

In this section, we will introduce our proposed method, which consists of three main steps: (1)Central Theme Sentence Extraction; (2)Dual-Style Fake News Generation; (3)Optimal Selection. We use the first to capture key semantic informatio, the second to generate two stylistically distinct fake news variants and the third to select the best option and integrate it into the original text. Figure 1 presents the overall structure of our method, and Algorithm 1 details the specific generation process.

3.1 Central Theme Sentence Extraction

A single critical sentence can be pivotal to the overall semantics of an article. When manipulated or replaced, the complex events described in the article may undergo substantial changes(Kung-Hsiang et al., 2023). Previous automated fake news generation methods typically produce lengthy segments of fake content, resulting in a considerable amount of inaccurate information—strikingly different from the characteristics of human-crafted fake news. In contrast, our approach creates a fake news dataset by substituting the article's most important sentence with a fabricated counterpart.

To assess the importance of each sentence, we employ an extractive summarization model to compute the salience score for every sentence(Liu and Lapata, 2019). In the document $E_k = \{s_1, s_2, \ldots, s_m\}$, we insert a [CLS] token at the beginning of each sentence to represent its features, which are subsequently extracted through pre-trained BERT model.

$$\tilde{h}^l = \mathrm{LN}(h^{l-1} + \mathrm{MHAtt}(h^{l-1})) \quad (1)$$

$$h^l = \mathrm{LN}(\tilde{h}^l + \mathrm{FFN}(\tilde{h}^l)) \quad (2)$$

We obtain the output $\hat{y}_i$ from the final layer of the summarization model. The calculation of $\hat{y}_i$ is as

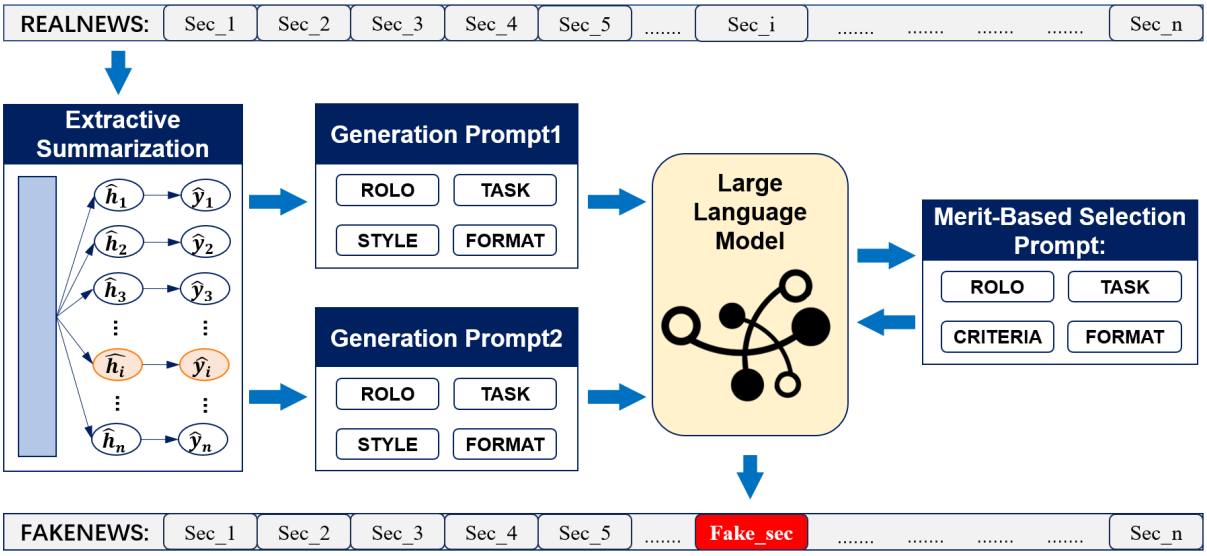

Figure 1: Structures of ours

follows,where σ represents the Activation Function used, and h_i^l denotes the feature vector of the i-th sentence from the final layer.

$$\hat{y}_i = \sigma(wh_i^l + b) \tag{3}$$

Our method selects the sentence with the highest $\hat{y}_i$ score as the central theme sentence for subsequent fake information generation. By replacing this sentence, we achieve fake news construction with minimal alteration to the original news article. This allows the generated fake news to be concealed within real news, enhancing its stealthiness while retaining the human writing style of the original article to the greatest extent possible.

3.2 Dual-Style Fake News Generation

Existing fake news generation models often rely on seq2seq architectures, which require extensive training on large datasets. This approach demands efforts from a data perspective, such as collecting more high-quality corpora or applying data augmentation techniques. However, high-quality data is precisely what current research lacks and what this study aims to address. LLM can effectively solve this issue by leveraging their world knowledge to directly generate fake news through prompt-based methods, eliminating the need for additional data collection or augmentation.

Therefore, our method leverages the Prompt optimization approach for large language models(Zhao et al., 2023) to design two distinct Prompts. These Prompts guide the large language model to generate fake news from different perspectives. The specific details are provided in Table 1. We divide the Prompt into four components: ROLO, TASK, STYLE, and FORMAT. In ROLO, we assign a news-related role to the large language model, enabling it to generate content in the writing style of news editor, closely aligned with the context of the original text;In TASK, we define the generation objective for the model:Prompt1 asks LLM directly instructed to generate fake news based on facts. To bypass the moral safeguards of the model, we introduce a rational purpose for generating fake content, In Prompt2, we assume that the central theme sentence describes fake news and task the model with inferring the opposite—"true news." This dual logical approach reduces content redundancy across the two templates. In STYLE, we specify two distinct stylistic guidelines:Prompt1 focuses on generating fake news that is more explicitly contradictory to the facts described in the original central theme sentence. Prompt2 emphasizes the use of propaganda techniques during generation, yielding fake news sentences with distinct linguistic tendencies. In FORMAT, we define the output format of the large language model to facilitate subsequent data processing. After generating the fake news sentences, we clean any recurring fixed patterns to prevent the emergence of potential backdoors(Dai et al., 2019). This process results in two distinct styles of fake news sentences: fake_sec1 and fake_sec2.

3.3 Optimal Selection

Using a single Prompt to generate fake news datasets results in stylistically uniform content,

	Content of the Generation Prompt
Prompt1	Assuming you are a news editor, write a short paragraph of fake news that has the opposite meaning from the fact. The fact is: {sentence}. So that I can improve my ability to recognize fake news. Just give the news you create.
Prompt2	You are a short news writer and your task is to write concise and promotional news. Here is a fake news story that has been maliciously altered to mean the opposite of the facts: {sentence}. Use this as a basis for inferring factual content. Write a short and propagandistic news story using factual content. Give only the news you wrote.

Table 1: Specifics of the two generation Prompts. The {sentence} denotes the main idea sentence we extracted in the previous step. We use LangChain for sentence insertion. ROLO as shown by texts in ROLO as shown by texts in cyan; TASK as shown by texts inviolet;STYLE as shown by texts in blue;FORMAT as shown by texts in orange.

lacking generalizability. Therefore, it is necessary to perform optimal selection between the two distinct styles of fake news sentences to ensure diversity in the generated dataset. Existing experiments have demonstrated that the closer the writing style of generated fake news is to human writing, the higher its quality and the better the performance of trained fake news detection models(Kung-Hsiang et al., 2023). However, current research lacks direct evaluation metrics to assess whether the generated text aligns closely with human writing styles.

Therefore, this method aims to leverage the natural language understanding capabilities of large language models by designing new Prompt templates to select the more suitable sentence from fake_sec1 and fake_sec2. The specific details are provided in Table 2. Similarly, we structure the Prompt into four components:ROLO: Assign the large language model a review-related roler;TASK: Clearly define the sentences to be evaluated and describe the task as a binary classification problem, a format familiar to the model;CRITERIA: Provide explicit evaluation guidelines, requiring the model to assess sentences based on writing techniques, logical coherence, and stylistic alignment with human writing, ensuring clarity on how to select the sentence that best matches human-like writing;FORMA: Specify the output format to standardize the model's responses for ease of downstream processing. Finally,fake_sec1 and fake_sec2 are presented as candidate sentences for evaluation. The model is tasked with selecting the sentence that most closely adheres to the defined criteria, producing fake_sec. This selected fake_sec is then inserted into the original news article, replacing the central theme sentence, resulting in a fully constructed fake news article.

Algorithm 1: Ours

Input: RealNews Dataset $P = \{E_i\}_{i=1}^{N}$
Output: FakeNews Dataset$P' = \{E_i'\}_{i=1}^{N}$
for $k = 1$ **to** N **do**
  $E_k = \{s_1, s_2, \ldots, s_m\}$
  Extractive Summarization Model M_0
  $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m = M_0 \leftarrow (E_i)$
  $\hat{y}_i = \max\{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$, main idea sentences = s_i
  fake_s1 = LLM $\leftarrow$ (Prompt1, s_i)
  fake_s2 = LLM $\leftarrow$ (Prompt2, s_i)
  fake_s = LLM $\leftarrow$ (Selection Prompt, fake_s1, fake_s2)
  $E_k' = \{s_1, s_2, \ldots, fake\_s, \ldots, s_m\}$
return $P' = \{E_i'\}_{i=1}^{N}$

4 Experiment

To validate the effectiveness of our generation method, we selected three state-of-the-art automated fake news generation methods as baselines. For each generation method, we provided the same original input to produce different fake news datasets. Then we used two widely adopted human-written fake news datasets for validation. Consistent with prior fake news detection tasks, we employed Accuracy (Acc), AUC, and F1 scores to evaluate the detection performance of pre-trained language models trained on different fake news datasets. Experimental results demonstrate that our generated dataset significantly enhances the model's ability to detect human-written fake news.

Details of the original input dataset can be found in Appendix A.

4.1 Datasets

LLMFAKE: The LLMFAKE dataset consists of 2,827 distinct articles, with fake news generated

	Content of the Optimal Selection Prompt
Prompt	You are a reviewer and I will give you two news stories, one written by a human and one generated by a machine. Your task is to vet out the news item that was written by a human.The one that employs rich writing techniques and propaganda methods and has logical sentences is written by a human.The first news story is: {sentence1}.The second news story is: {sentence2}. The output format is: I think the Xth news story was written by a human, X = [first, second].Output only according to the template, no need to give reasons.

Table 2: Specifics of the Optimal Selection Prompt. The {sentence1} denotes *fake_sec1*, {sentence2} denotes *fake_sec2*. Similarly, we use LangChain for sentence insertion. ROLO is shown by texts in cyan;TASK is shown by texts in violet; CRITERIA is shown by texts in blue; FORMAT is shown by texts in orange.

from TIMELINE17 as the original input using the method described above. The dataset maintains a 1:1 ratio between real and fake news. We splitthe data into 1697:565:565 for training, validation,and testing.

4.2 Baselines

PROPANEWS(Kung-Hsiang et al., 2023) dataset is generated through self-critique sequence training to ensure the validity of the generated articles. It also incorporates propaganda techniques, such as appeals to authority and loaded language. **GROVER**(Zellers et al., 2019)generates fake news by first creating a news headline from the original text and then generating fake news based on the headline. **FAKEEVENT**(Wu et al., 2022)generates sentences sequentially with condition on the manipulated knowledge elements of each sentence. Additionally, we constructed an extra training set by replacing the prominent sentence in each article with a sentence generated by each baseline method, as illustrated by -1ST. To ensure a fair comparison, all generators used the same real news articles as input.

4.3 Evaluation Data

We use the two datasets(Kung-Hsiang et al., 2023; Nguyen et al., 2022; Shu et al., 2018)to evaluate the effectiveness of our approach,which are manually written and fact-checked. Articles that are no longer accessible through the given URLs were removed. We then train detection models on datasets generated by different methods and use these models to evaluate their performance on the two validation sets. This helps us assess the model's ability to detect human-written fake news and, in turn, evaluate the effectiveness of the generation methods.More Details can be found in the Appendix B

4.4 Experimental Settings

In the experiments, we used the RoBERTa-Large model(Liu et al., 2019)provided by HuggingFace as the pre-trained model for fake news detection. The batch size is 16, The learning rates were set to 1e-05 and 5e-05, respectively. All experiments were conducted on a single NVIDIA RTX A6000 GPU. More Details can be found in the Appendix C.

5 Results and Analys

5.1 Overall Results

To investigate the improvement of our dataset in detecting human-written fake news, we trained RoBERTa-Large on several baseline datasets and evaluated it on the PolitiFact and Snopes datasets. As shown in Table 3, our method achieved significant improvements in detection performance.

In PolitiFact,our generation method allows the detection model to achieve an accuracy of 73.48% and AUC score of 81.26% without any external knowledge assistance. Compared to existing SOTA methods, the accuracy improved by nearly 5%, the AUC score increased by 7%, and the F1 score saw an improvement of 8.6%. Similarly, in the Snopes dataset, our method also outperforms all other baseline generation methods across all evaluation metrics. The better performance on the PolitiFact dataset may be due to the similarity between the news content in our input dataset and the PolitiFact dataset, both focusing on political news reporting.In contrast to our method, Grover generates lengthy fake information, which deviates from the style of human-written fake news. Although PROPANEWS manipulates only single sentences, its limited range of propaganda techniques results in poorer generalizability, with significantly different detection performance across the two validation datasets. We present examples of fake news generated by different methods in Appendix E.

Our model achieved the best results across nearly

DATASETS	POLITIFACT			SNOPES		
	Acc	AUC	F1	Acc	AUC	F1
FAKEEVENT	58.69	60.08	73.55	54.65	41.96	66.11
FAKEEVENT-1ST†	\	47.32	\	\	46.62	\
GROVER	53.61	67.17	46.97	53.24	50.75	63.27
GROVER-1ST	52.48	66.94	46.64	55.77	51.79	68.54
PROPANEWS	68.62	75.28	71.40	45.49	52.26	29.76
LLMFAKE	**73.48**	**82.26**	**80.03**	**61.41**	**62.50**	**73.08**

Table 3: Results on Acc,AUC,F1(in%) on the SNOPES and POLITIFACT datasets when trained on various datasets.Best in bold and second best in italic underlined. The † results are reproduced by Kung-Hsiang et al.'s (2023)

DATASETS	POLITIFACT			SNOPES		
	Acc	AUC	F1	Acc	AUC	F1
- Prompt1	67.38	81.13	77.72	59.86	60.79	74.25
- Prompt2	60.84	61.50	72.53	60.85	57.34	**74.31**
- Random-choose	68.85	80.72	78.34	59.86	60.79	74.25
- LLMFAKE	**73.48**	**82.26**	**80.03**	**61.41**	**62.50**	73.08

Table 4: Effectiveness study on Large language model Optimal Selection

all metrics on both datasets, which demonstrates its excellent generalization ability and confirms that it is not limited to a specific dataset. These results validate that fake news generated using large models, such as ours, is more effective in defending against human-written fake news. Our generation method enables the large model to produce fake news that closely resembles human-written content.

5.2 Ablation Study

To investigate the effectiveness of Optimal Selection,we compared the experimental results of datasets generated under Prompt1 and Prompt2. Additionally, to verify the correctness of our optimal selection, we created new datasets by randomly selecting samples from the Prompt1 and Prompt2 datasets and evaluated their performance. All datasets were trained using the same pre-trained model, RoBERTa-Large, under identical hyperparameters. The detection model performance on the generated fake news datasets is shown in Table 4. The results indicate that the performance of all three datasets is lower than that of the model trained on our complete dataset. However, the detection model trained on the optimized selection dataset outperforms the best individual results obtained from each of the Prompt1 and Prompt2 datasets on most evaluation metrics across both validation datasets.

At the same time, our method outperforms the random selection approach in nearly all evaluation metrics on both validation datasets. This demonstrates that the optimal selection by the large model plays a crucial role, and providing appropriate evaluation criteria is essential to help achieve the best possible performance.

5.3 About Generation Prompt

To investigate the effectiveness of different components within the Prompt, we conducted the following experiments based on Prompt1:We change the ROLO inta writer, resulting in Prompt-writer;We altered the logical perspective, creating Prompt-logic. Specific details can be found in Appendix D.The detection performance of models trained on datasets generated using these two new prompts is shown in Table 5.

Although Prompt-writer allows for more humanized writing techniques, this change introduces stylistic discrepancies with the original news category; Prompt-logic maintains a closer alignment with the writing style of the original news, it increases the complexity of the task for the large language model, making it more challenging to generate high-quality content. Both changes result in a decline in the overall generation quality.

Therefore, we recommend that when constructing prompt for generation domains,ROLO should align with the style of the original source. Addition-

DATASETS	POLITIFACT			SNOPES		
	Acc	AUC	F1	Acc	AUC	F1
Prompt-writer	58.69	54.05	73.55	55.07	50.78	66.60
Prompt-logic	58.69	63.28	73.59	54.93	50.31	67.08
Prompt1	**67.38**	**81.13**	**77.72**	**59.86**	**60.79**	**74.25**

Table 5: Effectiveness study on modified prompts

ally, it is advisable to use a simpler logical structure to minimize complexity and maintain high-quality generation.

6 Conclusion and future work

In this paper, we propose a new method for automatically generating fake news. This method leverages LLM for Dual-Style Fake News Generation and Optimal Selection to obtain high-quality fake news, which is then used to replace the main sentence, resulting in a high-quality dataset that closely resembles human writing style. We evaluate our method against three existing fake news generation methods, and the results show significant improvements over strong baselines. We also provide detailed ablation experiments to validate the effectiveness of our approach.This generation framework eliminates the need for manual labeling or training new models, thus significantly addressing the issues of sparse and difficult-to-construct fake news datasets. Additionally, it offers new insights into how LLM can generate text that closely mimics human writing styles.In future work, we plan to expand our method to other languages, taking advantage of the multilingual capabilities of LLM to enhance fake news datasets across multiple languages. We also aim to explore the imitation of human writing styles by LLM in other text generation tasks, further improving their application in fake news detection and prevention.

7 Limitations

Compared to human-written fake news, although our method imitates its style, the falsehood of the content still needs to be enhanced, and the exploration of the writing intent behind fake news remains insufficient. While our approach improves the detection of human-written fake news, the ability to detect model-generated fake news still requires further investigation.

References

Alim Al Ayub Ahmed, Ayman Aljarbouh, and Myung Choi. 2021. Detecting fake news using machine learning: A systematic literature review. *Psychology (Savannah, Ga.)*, 58:1932–1939.

Mingxi Cheng, Shahin Nazarian, and Paul Bogdan. 2020. Vroc: Variational autoencoder-aided multi-task rumor classifier based on text. In *Proceedings of The Web Conference 2020*, WWW '20, page 2892–2898, New York, NY, USA. Association for Computing Machinery.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Daniel Dzienisiewicz, Filip Graliński, Piotr Jabłoński, Marek Kubis, Paweł Skórzewski, and Piotr Wierzchon. 2024. POLygraph: Polish fake news dataset. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 250–263, Bangkok, Thailand. Association for Computational Linguistics.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avi Sil. 2021. InfoSurgeon: Cross-media fine-grained information consistency checking for fake news detection. 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 1683–1698, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300.

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, Online. Association for Computational Linguistics.

Shrutika Jadhav and Sudeep Thepade. 2019. Fake news identification and classification using dssm and improved recurrent neural network classifier. *Applied Artificial Intelligence*, 33:1–11.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203, Singapore. Association for Computational Linguistics.

Sawinder Kaur, Parteek Kumar, and Ponnurangam Kumaraguru. 2020. Automating fake news detection system using multi-level voting model. *Soft Comput.*, 24(12):9049–9069.

Sheng How Kong, Li Mei Tan, Keng Hoon Gan, and Nur Hana Samsudin. 2020. Fake news detection using deep learning. In *2020 IEEE 10th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, pages 102–107.

Chandra Mouli Madhav Kotteti, XIshuang Dong, Na Li, and Lijun Qian. 2018. Fake news detection enhancement with data imputation. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 187–192.

Huang Kung-Hsiang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14571–14589, Toronto, Canada. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

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. *ArXiv*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yi-Ju Lu and Cheng-Te Li. 2020. GCAN: Graph-aware co-attention networks for explainable fake news detection on social media. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 505–514, Online. Association for Computational Linguistics.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Martschat and Katja Markert. 2018. A temporally sensitive submodularity framework for timeline summarization. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 230–240, Brussels, Belgium. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc.

Tanushree Mitra and Eric Gilbert. 2021. Credbank: A large-scale social media corpus with associated credibility annotations. *Proceedings of the International AAAI Conference on Web and Social Media*, 9:258–267.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2022. Fang: leveraging social context for fake news detection using graph representation. *Commun. ACM*, 65(4):124–132.

Bo Ni, Zhichun Guo, Jianing Li, and Meng Jiang. 2020. Improving generalizability of fake news detection methods using propensity score matching.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John

Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. 2017. Study of hoax news detection using naïve bayes classifier in indonesian language. In *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, pages 73–78.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Kai Shu, Yichuan Li, Kaize Ding, and Huan Liu. 2020. Fact-enhanced synthetic news generation. *ArXiv*, abs/2012.04778.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8 3:171–188.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36.

Vivek Singh, Rupanjal Dasgupta, Darshan Sonagra, Karthik Raman, and Isha Ghosh. 2017. Automated fake news detection using linguistic analy-sis and machine learning.

Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some like it hoax: Automated fake news detection in social networks. *ArXiv*, abs/1704.07506.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Giang Binh Tran, Tuan Tran, Nam Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013. Leveraging learning to rank in an optimization framework for timeline summarization.

Lifu Tu, Semih Yavuz, Jin Qu, Jiacheng Xu, Rui Meng, Caiming Xiong, and Yingbo Zhou. 2024. Unlocking anticipatory text generation: A constrained approach for large language models decoding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15532–15548, Miami, Florida, USA. Association for Computational Linguistics.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection.

In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*, abs/2302.11382.

Jake Williams and Giovanni Santia. 2018. Buzzface: A news veracity dataset withfacebook user commentary and egos.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

Ni Xuanfan and Li Piji. 2023. A systematic evaluation of large language models for natural language generation tasks. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 40–56, Harbin, China. Chinese Information Processing Society of China.

Ai Ming Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Hai Zhao, Hang Xu, Hao-Lun Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kuncheng Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, Mingan Lin, Nuolan Nie, Pei Guo, Ruiyang Sun, Zhang Tao, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yan-Bin Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *ArXiv*, abs/2309.10305.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. *Defending against neural fake news*. Curran Associates Inc., Red Hook, NY, USA.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

Arkaitz Zubiaga, Geraldine Wong Sak Hoi, Maria Liakata, Rob Procter, and Peter Tolmie. 2015. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PLoS ONE*, 11.

A Data Source

We chose the TL17 dataset(Tran et al., 2013)as our data source, and the statistics of the dataset are shown in the table 6(Martschat and Markert, 2018). The news articles in this dataset were collected from major news outlets such as CNN, BBC, NBC News, and other reputable news organizations, ensuring that the unaltered articles are real news with high credibility. The dataset contains multiple timelines, with each timeline corresponding to a specific news event, covering hot topics across various domains such as war, economic crises, natural disasters, and public health. The original input for all the baseline generation methods used in our experiments is the news articles from this dataset.

Datasets	Topic	TLs	Docs	Sentences
TL17	9	19	4622	273,432

Table 6: Dataset Information for TIMELINE17

B Evaluation Data

POLITIFACT: This dataset comes from the fact-checking website PolitiFact, which specializes in verifying the truthfulness of political statements. The dataset collects and evaluates statements made by political figures, political organizations, and other public figures in public settings, providing a truthfulness rating for each statement.
SNOPES: This dataset is from the fact-checking website Snopes, which is dedicated to verifying and debunking false information. Snopes collects and verifies various pieces of information and stories circulating on the internet, offering assessments of their truthfulness. The dataset includes text descriptions of the fact-checked stories, background information, and related links or citations.

C Experimental Details

The training parameter settings for RoBERTa-Large are as follows:
- LLMFAKE,GROVER,FAKEEVENT: Trained using AdamW(Loshchilov and Hutter, 2017). The learning rates for the two layers are set to 1e-05 and 5e-05, with weight decays of 1e-6 and 1e-4, respectively. Learning rate warm-up is applied, with a total of 20 epochs, 10 warmup epochs, and a batch size of 16.
- PROPANEWS: We directly used the pre-trained model provided by the dataset authors for this method.

For security reasons, the authors of FAKEEVENT only released the dataset without making the generation model publicly available. Therefore, the specific parameters used to train the FAKEEVENT-1ST model were referenced directly from their published work.

D Changed Prompt Details

To better align with the existing changes, we modified other minor parts of the prompt.This was done to more precisely guide the large language model in generating fake news content that aligns with the target objectives.For specific details, please refer to Table 7.

Prompt-writer: You're a writer, there is a story: {sentence}. Please make changes in the story to make it have the opposite meaning. The format of the output is: 'The new story is:.'
Prompt-logic: You are a veteran journalist and here is a fake news story that is contrary to the facts of the original news. The fake news is: {sentence}. Please speculate what the original news is. The format of the output is: 'The original news is:'.

Table 7: Content of the Changed Prompt

E Case Study

In Table 8, we present specific examples of articles generated under different methods, with red text representing the generated fake news. We observe that the use of GROVER results in a large amount of fake news, which deviates significantly from the original meaning and fails to establish a coherent context with the real news. PROPANEWS generates relatively simple fake information, which may be due to the fact that its trained generation model lacks the richness of world knowledge compared to LLM.

Method	Generated Article
GROVER	BP has announced that it has over US$2bn (£662m) in legal fees pending against rival pipeline company TransCanada Corp for its role in the fatal oil spill at the Deepwater Horizon rig site in the Gulf of Mexico. "BP remains focused on its ongoing business," BP's deputy CEO, Joe Gevinson, said at a news conference on Wednesday. "It is our position that we can call upon TransCanada in the best interests of the world to restore the confidence and clean energy that was lost in the Gulf disaster." We know that this is a high stakes battle and that we have been very highly valued by the department and the BP officials during the investigation. We do not see any need to raise any more fines and it will be the British people who will decide who will pay, as will the British authorities who will meet the responsibilities of being responsible for this disaster. TransCanada, the federal energy minister, said it would also be fined for providing technical support to BP on the Gulf spill, contrary to a previous ruling from the Natural Resources Defence Council. The British government has so far fined BP over the disaster in the Gulf of Mexico for illegally building four pipeline links to their Britannia oil field offshore of New Orleans. In February, the Environmental Protection Agency fined BP, the chief executive of BP and lead contractor for two BP units responsible for the Gulf of Mexico site for illegal installations in the Florida Keys. President Obama is pushing for a law that could force the oil company to pay the full cost of the clean-up operation. At least four million gallons have leaked from the well that was damaged in an explosion more than three weeks ago.
PROPANEWS	The sound of gunfire during anti-government marches in Syria . Amateur video from Friday which could not be independently verified by Reuters show demonstrators under attack in Syria , The demonstrators try to take cover. A man goes down. Friday was by far the bloodiest day in more than a month of demonstrations to demand an end to President Bashar al-Assad 's rule Activists say at least 100 people died. The protests went ahead despite Assad 's decision this week to lift emergency law , in place since his Baath Party seized power some 48 years ago. Activists say more than 100 people have been killed. Assad has ejected most foreign media from the country so independent reports of violence are difficult to verify. Demonstrators are using the internet to get out their message. Deborah Lutterbeck , Reuters Gunfire and tanks in Syria -LRB- 2:03 -RRB- Apr 24 - Amateur video purports to show Syrian protesters under fire , with several seriously injured. -LRB- Transcript -RRB- An ultra-low latency infrastructure for electronic trading and data distribution A connected approach to governance , risk and compliance Our next generation legal research platform.
Ours	British Energy Behemoth Fails to Stop Gulf Oil Leak, Continues Environmental Disaster! Despite efforts to implement a 'top hat' procedure, the leaking wellhead remains uncontained, exacerbating the massive spill in the Gulf of Mexico. Concerns mount over the ongoing environmental crisis and the company's inability to control the situation. The cap sits on the wellhead's lower marine riser package (LMRP) section, the pipe that takes oil and gas from the well to a drill ship on the surface. The cap is designed to stop the flow of oil until a relief well can be drilled. The blowout on the Deepwater Horizon rig on 20 April killed 11 workers and led to the leak of millions of gallons of oil into the Gulf of Mexico, which has now reached the Florida Keys and is threatening to eclipse the 1989 Exxon Valdez disaster in Alaska as the worst environmental disaster in US history. The oil slick is threatening to eclipse the 1989 Exxon Valdez disaster as the worst environmental disaster in US history, and the damage to the gulf's fragile coastline has been described as "potentially catastrophic". The Exxon Valdez oil spill off Alaska was the worst environmental disaster in US history, with an estimated 11m gallons (40m litres) of oil spewing into the Gulf of Mexico before it was capped in 1989. A US charity, the Marine Spill Response Corp (MSRC), which is coordinating the response to the spill, said it was working with BP and the US Coast Guard to help clean up the oil.

Table 8: A qualitative comparison between the generated articles from different approaches. The texts marked in red indicate disinformation.