

# Correcting misinformation on social media with a large language model

## Authors

- Jakub Kubiak, 156049
- Maciej Janicki, 156073
- Julia Lorenz, 156066

## Introduction

Misinformation is an ongoing issue on social media platforms posing significant challenges to public discourse and individual decision-making. For a long time now, misinformation detection and correction have been tasks that required informed human intervention, which is both time-consuming and not scalable. Recent advancements in large language models (LLMs) have opened new avenues for addressing this problem, enabling automated systems to generate corrective responses to misinformation. This project aims to reproduce the method introduced in the paper *Correcting misinformation on social media with a large language model*(Zhou et al. 2024). The paper proposes a novel approach to misinformation correction called MUSE, which combines search-engine querying with LLM-based summarization. The goal of this project is to implement, test, and deploy the model on a different domain.

## Related Work

As a significant issue in the digital space, misinformation has been the focus of various studies. Prior research suggests that the main cause as well as the main problem of misinformation spread is the sheer volume and speed at which information is shared on social media (Adams et al. 2023). As a result, traditional methods of misinformation correction that often rely on manual fact-checking, however accurate, are not scalable, thus, inherently limited given the vast amount of content generated daily. Early work in computer science and natural language processing primarily focused on automated misinformation and fake news detection (Shu, Sliva, et al. 2017; Khan et al. 2021). These approaches typically frame the problem as a supervised classification task. While such methods have demonstrated promising results, they are mainly designed to identify misinformation rather than to correct it. More recently, the emergence of LLMs has enabled a shift from detection oriented systems toward generative approaches (Zhou et al. 2024; Guo et al. 2022) capable of producing natural language explanations and corrective responses.

## Datasets Description

### Dataset 1: Tweets Correction

The dataset is derived from the MUSE model's GitHub repository and is used to evaluate the generated responses. It includes the text of the original tweet, the generated corrective response, and the source of the response, as well as other metadata. The response source can be one of four categories: produced by highly helpful humans, produced by humans with average helpfulness, generated by the GPT-4 model, or generated by the MUSE model.

### Dataset 2: Fake News Detection

The second dataset is derived from the FakeNewsNet repository, which compiles news articles from multiple sources, including both genuine and fake content. For this project, we focus on the subset of articles identified as fake news from the BuzzFeed website. FakeNewsNet was introduced as part of a series of studies on fake news detection on social media (Shu, Sliva, et al. 2017; Shu, Wang, et al. 2017; Shu et al. 2018). Each record includes the title of the article and the corresponding main text.

### Dataset 3: Fact-checked Online Content

Final dataset also derived from Kaggle website contains fact-checked posts from various online platforms, covering a wide range of topics between 2008 and 2022. Each entry includes the post's title, content of the post, a truthfulness rating status, and a link to the original content.

## Method Description

The method consists of four main steps, firstly, the input content, which is a social media post is preprocessed and input into a LLM with an appropriate prompt to generate three adequate search queries. Next, these queries are used to retrieve relevant websites and sources of information from the web. Then, relevant content is extracted from the retrieved sources. Finally, the post's content along with the extracted information is fed into the LLM to generate a corrective response.

During our project we introduced an additional step before the query generation, where we prompt the LLM to rewrite the input content to make it more suitable for query generation. This step is intended to investigate whether improving the clarity and structure of the input content can lead to more effective search queries and ultimately better corrective responses.

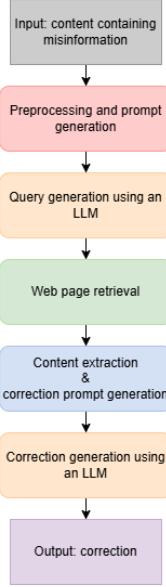


Figure 1: Method Pipeline

## Evaluation Startegy

We implemented the method step by step, starting with preprocessing the content and creating the initial prompts for query generation. The preprocessing step involved cleaning the text, removing unnecessary characters, changing temporal expressions to applicable dates from content's metadata, and formatting it appropriately for input into the LLM.

Original	Processed
<p>Before the election is stolen tomorrow, I'd like to mention that I've been to Georgia. A lot. It's not blue. Neither is Arizona.</p>	<p>before the election is stolen september 05 2024 id like to mention that ive been to georgia a lot its not blue neither is arizona</p> <p>Given a text, you are required to generate three different queries from the text for the Google Search Engine to get the most relevant web content to fact-check the text. Your answer should follow the following format:</p> <ol style="list-style-type: none"> <li>1. QUERY_1</li> <li>2. QUERY_2</li> <li>3. QUERY_3</li> </ol> <p>Where QUERY_1, QUERY_2, and QUERY_3 represent query text without quotation marks.</p> <p>If the given text is not informative enough to generate a query, you should answer "NONE".</p>

Figure 2: Preprocessing Step

Next, we focused on the query generation step. We designed prompts to instruct the LLM to generate three relevant search queries based on the preprocessed content. We evaluated the quality of the generated queries based on their relevance to the original content and their potential to retrieve useful information manually.

Original	Processed	Responses
<p>"I think it is unacceptable for the Chinese today to be larger in wind power and in solar panels and photovoltaic manufacturing than all of Europe put together – forget the United States."</p>	<p>I think it is unacceptable for the chinese July 01 2010 to be larger in wind power and in solar panels and photovoltaic manufacturing than all of europe put together forget the united states</p> <p>Given a text, you are required to generate three different queries from the text for the Google Search Engine to get the most relevant web content to fact-check the text. Your answer should follow the following format:</p> <p>1. QUERY_1 2. QUERY_2 3. QUERY_3 Where QUERY_1, QUERY_2, and QUERY_3 represent query text without quotation marks.</p> <p>If the given text is not informative enough to generate a query, you should answer "NONE".</p>	<p><b>Response 1</b> China wind power capacity vs Europe July 2010 statistics</p> <p><b>Response 2</b> China solar panel and photovoltaic manufacturing output 2010 comparison Europe</p> <p><b>Response 3</b> Global wind and solar manufacturing market share China vs Europe 2010</p>

Figure 3: Query Generation Step

Following the query generation, we implemented the web scraping step to retrieve information from websites that were returned by the search engine using the generated queries. And using the extracted information, we combined the preprocessed content and the retrieved information into a new prompt to generate the corrective response.

Original Content	Prompt	LLM Verdict
<p>I think it is unacceptable for the Chinese today to be larger in wind power and in solar panels and photovoltaic manufacturing than all of Europe put together – forget the United States."</p>	<p>You are required to respond to a text given some facts as references. Your response should satisfy all the following requirements:      - You response should explain where and why the text is or is not informed or potentially misleading.      - You should prioritize the facts (1) very close to the date the text was created, (2) very recently, and (3) listed at the beginning of "Facts".      - You should show the URLs that support your explanation. You should not number the URLs.      - Your response should be informative and short.      - Your response should start with "This text is".</p> <p>Given text: [TEXT]</p> <p>Relevant context: [RELEVANT_CONTEXT]</p>	<p>This text is potentially misleading and exaggerated in at least two ways:</p> <p><b>Where it is misleading</b></p> <ul style="list-style-type: none"> <li>On solar/PV manufacturing: according to data from the International Energy Agency (IEA), in 2010 China did not yet "own" more of the global PV manufacturing industry, but its lead relative to Europe was far smaller. The 2010 data shows that China's share of global PV manufacturing capacity was significant but not dominant — not a margin that clearly exceeds "all of Europe" combined. (IEA)</li> <li>On wind: In 2010, the installed wind-power capacity in China was <b>31.1 GW</b>, according to data aggregated by global-wind databases. (Wikipedia) At that time, Europe still maintained substantially more installations, so to claim China's wind industry was "larger than all of Europe put together" is not supported by data.</li> </ul>

Figure 4: Correction Generation Step

## Outcomes and Conclusions

We successfully recreated the MUSE methodology for correcting misinformation on social media using large language models. We implemented the full pipeline, including preprocessing, search query generation, web-based information retrieval, and corrective response generation.

We applied this method to different datasets from various domains, demonstrating its versatility. The generated corrective responses were evaluated manually

based on their relevance and accuracy in addressing the misinformation present in the original content.

The method’s main advantages include its ability to gather information from multiple sources, providing a more comprehensive basis for corrections. Using multiple queries and sources allows for a broader perspective on the topic being addressed. Additionally, using datasets containing different types of content, rather than solely tweets, showcases the method’s adaptability to various contexts.

However, there are a few limitations to consider. The main challenge is the requirement for API keys to access LLMs and search engines, which come at a cost. The trustworthiness of the scraped content is another concern, as the quality of the corrective responses heavily depends on the reliability of the sources retrieved, while the authors of the original paper addressed this by implementing a source credibility assessment step, there is still a risk of incorporating biased or inaccurate information. What is more, LLMs tend to produce overly long and nuanced answers, which usually does not match the typical tone and style of online discourse, as well as may discourage users from reading the entire response. Finally, the method is limited by the publicly available knowledge on the web, which may not always be up-to-date or comprehensive.

To conclude, our project replicated the MUSE approach for misinformation correction using LLMs, demonstrating its effectiveness across different datasets. As misinformation spread is a viable concern in today’s digital age and human fact-checking is unattainable at such scale, this method demonstrates that LLMs can be an effective supporting tool for combating misinformation on social media.

## References

- MUSE GitHub repository: <https://github.com/Social-Futures-Lab/MUSE/tree/main>  
FakeNewsNet Kaggle repository: <https://www.kaggle.com/datasets/mdepak/fakenewsnet>  
Verified Posts Kaggle repository: <https://www.kaggle.com/datasets/mattimansha/verified-posts-fact-checking-online-content>
- Adams, Zoë, Magda Osman, Christos Bechlivaniidis, and Björn Meder. 2023. “(Why) Is Misinformation a Problem?” *Perspectives on Psychological Science* 18 (6): 1436–63. <https://doi.org/10.1177/17456916221141344>.
- Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos. 2022. “A Survey on Automated Fact-Checking.” *Transactions of the Association for Computational Linguistics* 10 (February): 178–206. [https://doi.org/10.1162/tacl\\_a\\_00454](https://doi.org/10.1162/tacl_a_00454).
- Khan, Junaed Younus, Md. Tawkat Islam Khondaker, Sadia Afroz, Gias Uddin, and Anindya Iqbal. 2021. “A Benchmark Study of Machine Learning Models for Online Fake News Detection.” *Machine Learning with Applications* 4: 100032. <https://doi.org/10.1016/j.mlwa.2021.100032>.

- Shu, Kai, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. “FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media.” *arXiv Preprint arXiv:1809.01286*.
- Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD Explorations Newsletter* 19 (1): 22–36.
- Shu, Kai, Suhang Wang, and Huan Liu. 2017. “Exploiting Tri-Relationship for Fake News Detection.” *arXiv Preprint arXiv:1712.07709*.
- Zhou, Xinyi, Ashish Sharma, Amy X. Zhang, and Tim Althoff. 2024. *Correcting Misinformation on Social Media with a Large Language Model*. <https://arxiv.org/abs/2403.11169>.