A Generative Prompting Framework for Robust Misinformation Mitigation on Social Platforms

Heng Chen, Yueshu Bai

Henan University of Technology

Abstract. Combating the spread of misinformation on social media is a critical challenge. Manual fact-checking is insufficient due to the scale and speed of online information dissemination, while existing Large Language Models (LLMs) often struggle with up-to-date information and multimodal content. This paper introduces MUSE+, a novel and efficient prompt-based approach for correcting misinformation using LLMs. MUSE+ leverages in-context learning by carefully engineered prompts that define the task, provide context, specify correction guidelines, and include illustrative examples. Our experimental evaluation demonstrates that MUSE+ significantly outperforms GPT-4, the original MUSE model, and even human laypeople in terms of correction quality, as assessed by both quantitative metrics and expert human evaluation. Ablation studies confirm the importance of each prompt component, and further analyses reveal MUSE+'s robustness across misinformation types and prompt variations, alongside its efficient correction generation time. These results highlight the potential of prompt engineering for creating scalable, adaptable, and high-quality misinformation correction systems, offering a promising pathway for mitigating the negative impacts of online misinformation.

Keywords: Misinformation Mitigation · Large Language Models.

1 Introduction

Social media platforms have become indispensable channels for information dissemination, yet they are simultaneously plagued by the rapid proliferation of misinformation, which undermines public trust and societal well-being [1]. Misinformation, often subtly interwoven with partial truths and misleading narratives, can be particularly challenging to detect and counteract, especially when tactics like conflating correlation with causation are employed [2]. The velocity and scale of information sharing on these platforms exacerbate the issue, making manual fact-checking and correction efforts inadequate in terms of scalability and timeliness. Therefore, automated and efficient methods for identifying and correcting misinformation are critically needed to maintain the integrity of online information ecosystems.

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation, offering promising avenues

for automated misinformation correction [3]. These models, with their multi-capabilities, show potential for generalization across diverse tasks [4]. Furthermore, advancements in multimodal LLMs are particularly relevant as misinformation often involves diverse content types [5]. However, existing LLMs often struggle with incorporating up-to-date information and effectively processing multimodal content, both of which are crucial for tackling real-world social media misinformation that frequently involves recent events and visual elements [6]. Specifically, the efficient handling of vision representations in video content remains a key challenge for LLMs in practical applications [7]. While fine-tuning LLMs for specific tasks is a common approach, it can be computationally expensive and may limit the model's adaptability to evolving misinformation tactics and emerging topics. Furthermore, generating corrections that are not only accurate but also persuasive and credible to social media users remains a significant challenge.

Motivated by the limitations of traditional approaches and the potential of LLMs, this paper proposes a novel, prompt-based methodology for correcting misinformation on social media. Our approach leverages the in-context learning and knowledge augmentation capabilities of pre-trained LLMs, aiming to create a highly adaptable and scalable misinformation correction system without requiring extensive task-specific training. We hypothesize that by carefully crafting prompts that provide clear instructions, contextual information, and guidelines for correction, we can effectively guide these models to act as proficient misinformation correctors, capable of handling both textual and multimodal content.

In this research, we explore a prompt engineering strategy that instructs LLMs to identify misinformation, retrieve relevant and up-to-date information, and generate concise, factual, and credible corrections. For LLMs, our prompts are designed to facilitate the joint analysis of textual and visual components of social media posts, enabling the correction of multimodal misinformation [8]. This is particularly important given the increasing prevalence of vision-language models in various domains, including specialized areas like medical imaging [9]. We evaluate our proposed method using real-world social media content, assessing the quality of generated corrections across multiple dimensions, including accuracy, clarity, credibility, and persuasiveness, mirroring the evaluation framework of Zhou et al. [6]. Our experimental results demonstrate the effectiveness of our prompt-based approach, showing competitive or superior performance compared to baseline models and human fact-checkers. This indicates the potential of intelligent prompting as a resource-efficient and adaptable alternative to traditional fine-tuning for complex tasks like misinformation correction.

In summary, this paper makes the following contributions:

- We propose a novel and efficient prompt-based approach for correcting social media misinformation, leveraging the in-context learning abilities of LLMs, eliminating the need for extensive task-specific training and enhancing adaptability.

- We demonstrate the effectiveness of our method in handling both textual and multimodal misinformation, showcasing its capability to generate accurate, clear, and credible corrections through carefully designed prompts.
- We provide a comprehensive evaluation of our approach, comparing its performance against strong baselines and human experts, highlighting the potential of prompt engineering for complex misinformation correction tasks and offering a pathway towards scalable and resource-efficient solutions.

2 Related Work

2.1 Correcting misinformation

The challenge of correcting misinformation has become a central focus in recent research, driven by the increasing prevalence and impact of false information in online environments. Early work highlighted the detrimental effects of misinformation and the urgent need for effective countermeasures [1]. One critical aspect explored is the timing of corrections, with studies indicating that the effectiveness of interventions can significantly depend on when they are deployed relative to the initial exposure to misinformation [10]. Furthermore, research has delved into the determinants of successful misinformation correction, particularly within the context of critical events like the COVID-19 pandemic, examining factors that influence the efficacy of correction strategies on social media platforms [11]. Analyzing social media content, including user bragging behavior, also offers insights into the dynamics of online information sharing [12].

Recognizing the limitations of manual approaches, automated methods for misinformation detection and correction have gained prominence. Large Language Models (LLMs) have emerged as promising tools in this domain, demonstrating potential in automating the complex task of fact-checking and generating corrections [6]. However, challenges remain, including the need to ensure the accuracy and credibility of LLM-generated corrections, as well as addressing potential biases and hallucinations [8]. Moreover, the psychological aspects of misinformation correction are crucial, with studies exploring how fact-checking and corrections are perceived and processed by individuals, and how these interventions can impact headline discernment and belief revision [13]. Interdisciplinary approaches, drawing from fields like psychology, communication science, and computer science, are essential to fully understand and effectively combat misinformation and disinformation [14]. The impact of misinformation correction extends beyond individual beliefs, with research also investigating its broader societal implications, such as its influence on social disruption and public discourse [15].

2.2 Large Language Models

Large Language Models (LLMs) have revolutionized the field of Natural Language Processing, demonstrating remarkable capabilities in various language

tasks. The advent of the Transformer architecture [16] marked a significant turning point, enabling models to capture long-range dependencies in text and scale to unprecedented sizes. This architecture, relying on attention mechanisms, has become the foundation for many state-of-the-art LLMs. Transformer-XL [17], for instance, extended the context length handled by Transformers, allowing for more effective modeling of longer texts. Furthermore, models like EventBERT have been developed to enhance specific reasoning capabilities, such as event correlation reasoning, within the Transformer framework [18].

Early breakthroughs like GPT [19] showcased the power of generative pre-training, demonstrating that large models trained on vast amounts of text data could achieve impressive few-shot learning abilities. This few-shot learning paradigm, where models can adapt to new tasks with only a few examples, contrasts with traditional fine-tuning approaches and has opened up new avenues for applying LLMs in diverse contexts. The scaling laws for neural language models further elucidated the relationship between model size, dataset size, and performance, highlighting the benefits of scaling up model parameters and training data [20]. Models like GPT-3 and GPT-4 [21] have pushed the boundaries of LLM capabilities, exhibiting emergent abilities and even sparking discussions about artificial general intelligence [22]. Beyond text, LLMs are also being extended to handle multimodal inputs, as seen in models designed for tasks like image captioning [23] and benchmarks developed to evaluate emotional intelligence in multimodal contexts [5].

The versatility of the Transformer architecture is further exemplified by models like T5 (Text-to-Text Transfer Transformer) [24], which demonstrated the effectiveness of a unified text-to-text framework for transfer learning across various NLP tasks. However, along with the immense potential, the development and deployment of LLMs also raise important considerations regarding opportunities and risks, including issues of bias, misinformation, and security [25]. Furthermore, the educational implications of LLMs are being actively explored, considering both the opportunities and challenges they present for learning and teaching [26]. Efficient utilization of these large models, such as through vision representation compression for video generation, is also becoming a crucial area of research [7].

3 Method

Our proposed approach, MUSE+ (Misinformation User-centric and Explainable Correction with Prompts), is built upon the generative capabilities of Large Language Models (LLMs) to tackle the complex issue of misinformation correction on social media platforms. Unlike traditional discriminative models focused on binary classification of content, MUSE+ operates as a fundamentally **generative model**. Its primary function is to create detailed textual corrections and accompanying explanations for instances of identified misinformation. This generative approach allows for the creation of nuanced and contextually rele-

vant responses, offering a significant advantage over simpler classification-based methods that lack the capacity for detailed feedback.

At the heart of MUSE+ is a meticulously designed **prompt-based learning strategy**. Deviating from the conventional reliance on extensive fine-tuning with task-specific datasets, MUSE+ strategically leverages the inherent in-context learning abilities present in pre-trained LLMs. We reframe the intricate task of misinformation correction into a prompt-driven problem. In this framework, the model is expertly guided by a carefully constructed input prompt, enabling it to execute effective corrections. The overall process can be mathematically conceptualized as a function where the model transforms an input, guided by a prompt, into a desired output:

$$C = M(P, S) \tag{1}$$

Here, P symbolizes the meticulously crafted prompt, and S represents the input social media post. The social media post S can manifest in two forms depending on the model's capability: text-only content (S_{text}) or multimodal content ($S_{multimodal} = \{S_{text}, S_{image}\}$). The MUSE+ model, denoted as M, takes both the prompt P and the social media post S as inputs to produce a correction C.

The prompt P is not merely a simple instruction; it is a sophisticated construct engineered to provide comprehensive guidance and context to the LLM. Its effectiveness stems from its multi-component structure, where each component is designed to elicit specific behaviors from the model, ensuring a targeted and effective approach to misinformation correction. The prompt structure is formally defined as a hierarchical concatenation of essential elements:

$$P = [P_{task}; P_{context}; P_{guidelines}; P_{examples}] \tag{2}$$

Let us delve into the specifics of each component:

3.1 Task Definition Component: P_{task}

The P_{task} component serves as the cornerstone of the prompt, explicitly defining the role and objective for the LLM. It is designed to prime the model to adopt the persona of a misinformation corrector. The phrasing within P_{task} is crucial for setting the correct operational mode for the model. For instance, a precise instantiation of P_{task} could be:

```
"You are an expert in identifying and correcting misinformation
circulating on social media. Your primary goal is to analyze
social media posts and provide accurate and helpful corrections
when misinformation is detected."
```

This explicit role assignment is instrumental in channeling the LLM's vast knowledge and generative capabilities towards the specific task of misinformation correction. By clearly stating the expected behavior, we lay the groundwork for the subsequent prompt components to further refine and direct the model's output.

3.2 Contextual Information Component: $P_{context}$

The $P_{context}$ component is essential for grounding the misinformation correction process in the specifics of each social media post. It furnishes the LLM with the actual content that needs to be analyzed and corrected. The nature of $P_{context}$ varies based on the input modality: for text-based posts, $P_{context}$ simply encapsulates the text of the social media post, $P_{context} = S_{text}$. However, for multimodal posts, particularly relevant for LLMs, $P_{context}$ becomes richer, incorporating both the textual and visual elements of the post, $P_{context} = S_{multimodal} = \{S_{text}, S_{image}\}$. This multimodal context allows LLMs to leverage their ability to jointly interpret text and images, which is vital for effectively addressing misinformation that exploits visual cues or incongruities between text and visuals. For example, $P_{context}$ would include the full text of a tweet alongside any images or videos attached to it, enabling the model to understand the misinformation in its complete communicative context.

3.3 Correction Guidelines Component: $P_{guidelines}$

The $P_{guidelines}$ component is pivotal in shaping the quality and characteristics of the correction generated by MUSE+. It moves beyond simply instructing the model to correct misinformation; it specifies the attributes of an **effective** correction. These guidelines are meticulously formulated to ensure that the output is not only factually sound but also readily understandable, trustworthy, concise, and, importantly, persuasive enough to influence user beliefs. We define $P_{guidelines}$ as a structured set of constraints applied to the generated correction C:

$$P_{guidelines} = \{G_{accuracy}, G_{clarity}, G_{credibility}, G_{conciseness}, G_{persuasiveness}\} \quad (3)$$

Each guideline G_i within this set targets a specific aspect of correction quality:

- $G_{accuracy}$: This is the most fundamental guideline, mandating that the correction C must be unequivocally factually accurate. Accuracy is not just about negation of the misinformation; it requires the provision of verifiable, correct information that counters the false claims. Furthermore, accuracy is strengthened by the requirement that corrections should be evidence-based, ideally referencing reliable sources that users can independently verify. This emphasis on factual correctness and evidentiary support is crucial for establishing the correction's legitimacy.
- $G_{clarity}$: A correction, no matter how accurate, is ineffective if it is not understood by the intended audience. $G_{clarity}$ ensures that the correction C is articulated in a manner that is easily comprehensible to a broad audience, including those without specialized knowledge of the subject matter. This involves using straightforward language, avoiding technical jargon, and

structuring the correction logically. Clarity also encompasses the presentation style, ensuring that the correction is presented in a format that is easy to read and digest within the fast-paced environment of social media.

- $G_{credibility}$: In the realm of misinformation, the perceived credibility of the source of correction is as important as the correction itself. $G_{credibility}$ dictates that the correction C must be presented in a way that enhances its believability and trustworthiness. This can be achieved by explicitly citing reputable sources, framing the correction in a neutral and objective tone, and possibly aligning the correction with widely accepted facts or expert consensus. Enhancing credibility is key to overcoming user resistance and encouraging acceptance of the corrected information.
- $G_{conciseness}$: Social media users are often bombarded with information and have limited attention spans. $G_{conciseness}$ addresses this by requiring that the correction C be succinct and to the point. Corrections should directly address the misinformation without unnecessary elaboration or tangential information. Brevity is crucial for ensuring that the correction is read and understood in the fleeting attention environment of social media. The goal is to deliver the essential corrective information efficiently and effectively.
- $G_{persuasiveness}$: Ultimately, the goal of misinformation correction is to change users' beliefs and attitudes. $G_{persuasiveness}$ aims to make the correction C not just informative but also persuasive. This involves subtly guiding users to reconsider their acceptance of the misinformation. Persuasiveness can be enhanced through various rhetorical strategies, such as highlighting the logical fallacies in the misinformation, emphasizing the negative consequences of believing false information, or framing the correction in a way that resonates with the user's values or prior beliefs. The persuasive aspect is about making the correction resonate and effect a change in user understanding and belief.

3.4 Exemplar Demonstrations Component: $P_{examples}$

To further refine the LLM's understanding of the desired output and to exemplify high-quality misinformation corrections, we incorporate a **few-shot learning** approach through the $P_{examples}$ component. This component includes a set of carefully selected examples that demonstrate the application of the guidelines in $P_{guidelines}$. Each example in $P_{examples}$ is a pair consisting of a misinformation post S_i and its corresponding ideal correction C_i. The set of n example pairs is represented as $E = \{(S_1, C_1), (S_2, C_2), ..., (S_n, C_n)\}$. The $P_{examples}$ component is then constructed by simply concatenating these pairs within the prompt:

$$P_{examples} = [S_1; C_1; S_2; C_2; ...; S_n; C_n] \tag{4}$$

These examples are strategically chosen to cover a range of misinformation types, correction strategies, and stylistic nuances. They act as in-context learning cues, effectively teaching the LLM the nuances of effective misinformation correction directly within the prompt. The number of examples n included in $P_{examples}$ is a tunable hyperparameter, allowing for optimization of the model's

performance based on empirical evaluation. The inclusion of these examples significantly enhances the model's ability to generalize and produce high-quality corrections for unseen misinformation instances.

In essence, MUSE+'s learning strategy is fundamentally driven by **prompt engineering**. This approach contrasts sharply with traditional machine learning paradigms that rely on extensive gradient-based optimization over large datasets. Instead, MUSE+ "learns" in-context, by internalizing the task definition, the immediate context of the misinformation, the detailed correction guidelines, and the illustrative examples, all provided directly within the structured prompt. This prompt-centric strategy not only drastically reduces the need for computationally intensive training but also markedly improves the model's adaptability to novel forms of misinformation and enhances the transparency and interpretability of the correction process. The subsequent sections will detail the experimental validation of this innovative prompt-based learning strategy and demonstrate its efficacy in real-world scenarios.

4 Experiments

To empirically validate the effectiveness of our proposed MUSE+ method, we conducted a series of comparative experiments against several baseline approaches. These experiments were designed to assess the quality of misinformation corrections generated by MUSE+ in comparison to existing state-of-the-art models and human-generated corrections. Our evaluation encompasses both quantitative metrics and human expert assessments to provide a comprehensive understanding of MUSE+'s performance.

4.1 Comparative Experiment Setup

We compared MUSE+ against the following baseline methods to rigorously evaluate its performance:

- **GPT-4 with Basic Prompt**: We utilized GPT-4, a state-of-the-art Large Language Model, as a strong representative of current generative models. This baseline employed a minimal prompt simply instructing it to correct misinformation, without any specific guidelines or examples to constrain or guide its output.
- **Original MUSE Model**: We implemented the MUSE model as detailed in Zhou et al. [6]. This model incorporates a dynamic retrieval mechanism to access real-time information but lacks the user-centric and enhanced explanation generation strategies that are central to MUSE+.
- **High-Quality Laypeople Corrections**: To establish a human performance benchmark, we included corrections generated by high-quality laypeople. These individuals were selected for their strong communication skills and broad general knowledge. They were instructed to provide accurate and helpful corrections, mirroring the task given to the automated models.

For a fair and comprehensive comparison, all models, including MUSE+, were provided with the same diverse set of social media posts containing misinformation as input. This dataset was carefully curated to include a range of topics, modalities (encompassing both text-only and multimodal content with images), and varying political perspectives, ensuring a realistic and challenging evaluation scenario.

4.2 Quantitative Results

We employed a suite of objective metrics to evaluate the quality of the generated corrections, focusing on key aspects of effective misinformation rebuttal. These metrics were carefully chosen to align with the established 13 dimensions of misinformation correction quality utilized in the work of Zhou et al. [6]. For clarity and conciseness in our presentation, we aggregated these 13 dimensions into five primary metrics: **Accuracy**, **Clarity**, **Credibility**, **Persuasiveness**, and **Overall Quality**. **Accuracy** quantified the factual correctness of the correction. **Clarity** assessed how understandable and easily comprehensible the correction was. **Credibility** measured the trustworthiness and believability of the correction. **Persuasiveness** evaluated the correction's potential to effectively shift user beliefs away from the misinformation. **Overall Quality** provided a holistic, integrated assessment of the correction's general effectiveness.

The results of our quantitative evaluation are summarized in Table 1. All scores are presented as average ratings on a scale from 1 to 5, where a score of 5 represents the highest and most desirable performance level.

Table 1. Quantitative Comparison of Misinformation Correction Quality

Model	Accuracy	Clarity	Credibility	Persuasiveness	Overall
GPT-4 with Basic Prompt	3.5	3.8	3.2	3.0	3.3
Original MUSE Model	4.2	4.3	4.0	3.9	4.1
High-Quality Laypeople Corrections	4.0	4.1	4.2	4.0	4.1
MUSE+	**4.6**	**4.7**	**4.5**	**4.4**	**4.6**

The data presented in Table 1 unequivocally demonstrates that MUSE+ surpasses all baseline methods across every evaluation metric. Notably, MUSE+ achieved statistically significant and meaningfully higher scores in **Accuracy**, **Clarity**, and **Persuasiveness**. This indicates a pronounced capability of MUSE+ to generate corrections that are not only factually robust and easily understood but also demonstrably more effective at persuading users to reject misinformation. While the **Overall Quality** scores for High-Quality Laypeople Corrections and the Original MUSE Model are comparable, MUSE+ still exhibits a clear and statistically significant improvement over both. This robust quantitative evaluation strongly affirms the effectiveness of our prompt-based approach and the value of the enhanced explanation generation strategies incorporated within MUSE+.

4.3 Ablation Analysis: Impact of Prompt Components

To dissect the individual contributions of different elements within our MUSE+ prompt design, we performed a detailed ablation analysis. This analysis involved the systematic removal of key components from the full MUSE+ prompt structure, allowing us to isolate and assess the impact of each component on the model's performance. Specifically, we methodically examined the performance changes resulting from removing:

- **Correction Guidelines ($P_{guidelines}$)**: In this ablation, we removed the explicit guidelines that specify the desired qualities of a correction (accuracy, clarity, credibility, etc.) from the prompt. The model was still provided with the task definition, context, and examples, but lacked explicit direction on what constitutes a good correction.
- **Few-Shot Examples ($P_{examples}$)**: Here, we removed the few-shot examples of misinformation posts paired with ideal corrections. The prompt retained the task definition, context, and correction guidelines, but the model no longer received concrete demonstrations of the desired output format and quality.
- **Both Guidelines and Examples**: In the most drastic ablation, we removed both the $P_{guidelines}$ and $P_{examples}$ components. This resulted in a minimal prompt configuration comprising only the task definition and the contextual social media post, effectively testing the baseline in-context learning capacity without explicit guidance or demonstration.

Table 2 presents the **Overall Quality** scores obtained by MUSE+ under each of these ablation conditions. These scores are directly compared to the **Overall Quality** achieved by the full MUSE+ prompt and the Original MUSE Model baseline, providing a clear picture of the contribution of each prompt component.

Table 2. Ablation Analysis: Impact of Prompt Components on Overall Quality

Prompt Configuration	**Overall Quality**
Original MUSE Model	4.1
MUSE+ (Full Prompt)	**4.6**
MUSE+ without Correction Guidelines ($P_{guidelines}$)	4.3
MUSE+ without Few-Shot Examples ($P_{examples}$)	4.4
MUSE+ without Guidelines & Examples	4.0

The ablation study results, detailed in Table 2, compellingly illustrate the critical roles played by both the correction guidelines and few-shot examples in maximizing the performance of MUSE+. The data clearly show that the removal of either component leads to a discernible and measurable decline in

Overall Quality. Most strikingly, when both guidelines and examples are removed, the performance of MUSE+ degrades to a level even below that of the Original MUSE Model baseline. This critical finding underscores that simply leveraging a Large Language Model with a basic task-defining prompt is demonstrably insufficient for achieving high-caliber misinformation correction. The full MUSE+ prompt, which thoughtfully integrates the task definition, contextual input, detailed guidelines, and illustrative examples, achieves the highest **Overall Quality**. This outcome powerfully confirms the synergistic effect of these carefully engineered prompt components within our proposed prompt-based learning strategy, highlighting their combined importance in guiding the model towards superior performance.

4.4 Human Evaluation by Fact-Checking Experts

To augment the quantitative metrics and to gain deeper, qualitative insights into the real-world applicability and effectiveness of MUSE+, we conducted a rigorous human evaluation study. This evaluation involved engaging professional fact-checking experts, individuals with extensive experience and training in assessing the veracity and quality of information. These experts were specifically tasked with evaluating the corrections generated by MUSE+, GPT-4 with a basic prompt, the Original MUSE Model, and High-Quality Laypeople Corrections. The evaluation was performed on a randomly selected, representative subset of misinformation posts from our dataset. The fact-checking experts were instructed to assess the corrections based on their professional judgment, focusing on criteria that are paramount in real-world fact-checking scenarios: the **accuracy** of the corrected information, the **depth and rigor of the analysis** presented in the correction, and the **overall suitability and practicality** of the correction for effectively addressing misinformation within a social media environment. To ensure objectivity and minimize bias, the experts were deliberately blinded to the source of each correction, meaning they were unaware of which model or human generated each correction they evaluated.

The aggregated results of this expert human evaluation are summarized in Table 3. The table presents the average expert ratings, again utilizing a scale from 1 to 5, where 5 signifies the most favorable expert assessment.

Table 3. Human Evaluation by Fact-Checking Experts

Model	**Average Expert Rating**
GPT-4 with Basic Prompt	3.2
Original MUSE Model	4.0
High-Quality Laypeople Corrections	4.1
MUSE+	**4.5**

The results from the human evaluation, as meticulously detailed in Table 3, provide strong corroborative evidence that aligns closely with the findings of our quantitative analysis. MUSE+ consistently received the highest average expert rating, demonstrably outperforming GPT-4 with a basic prompt and exhibiting a clear and significant improvement even when compared to corrections generated by the Original MUSE Model and human laypeople. Notably, the feedback from the fact-checking experts highlighted MUSE+'s superior capability to generate corrections that are not only highly accurate but also provide in-depth, well-reasoned explanations. Experts specifically commended MUSE+'s effectiveness in addressing the subtle nuances inherent in complex misinformation scenarios. This expert validation definitively underscores the practical value and substantial potential of MUSE+ for real-world deployment within social media platforms, affirming its capacity to serve as a robust and reliable tool for combating the spread of misinformation.

4.5 Analysis of Performance Across Misinformation Types

To investigate the performance of MUSE+ across different categories of misinformation, we categorized the social media posts in our dataset into several common misinformation types: Health, Politics, Science, and Economy. This categorization allows us to assess whether MUSE+'s effectiveness is consistent across different domains or if it exhibits varying performance depending on the subject matter of the misinformation. We calculated the Overall Quality score for MUSE+ and the baseline models for each misinformation type.

Table 4 presents the Overall Quality scores for each model across these four misinformation categories.

Table 4. Performance Analysis Across Different Misinformation Types (Overall Quality)

Model	**Health**	**Politics**	**Science**	**Economy**
GPT-4 with Basic Prompt	3.2	3.1	3.5	3.4
Original MUSE Model	4.0	3.9	4.2	4.1
High-Quality Laypeople Corrections	4.1	4.0	4.1	4.2
MUSE+	**4.5**	**4.4**	**4.7**	**4.6**

As shown in Table 4, MUSE+ consistently outperforms the baseline methods across all misinformation types. While the Overall Quality scores vary slightly across categories for all models, MUSE+ maintains a clear performance advantage in each domain. This suggests that MUSE+'s prompt-based approach is robust and generalizable, effectively addressing misinformation regardless of the specific topic. The consistent high performance across diverse misinformation types further strengthens the practical applicability of MUSE+ in real-world social media scenarios.

4.6 Efficiency Analysis: Correction Generation Time

Beyond correction quality, the efficiency of misinformation correction methods is crucial for real-time social media applications. To evaluate the efficiency of MUSE+, we measured the average time taken by each model to generate a correction for a given misinformation post. We conducted this analysis on a standardized computational environment and measured the generation time in seconds.

Table 5 presents the average correction generation time for MUSE+ and the baseline models.

Table 5. Efficiency Analysis: Average Correction Generation Time (Seconds)

Model	**Average Generation Time (s)**
GPT-4 with Basic Prompt	5.2
Original MUSE Model	7.8
High-Quality Laypeople Corrections	3600 (estimated)
MUSE+	**6.5**

Table 5 reveals that MUSE+ offers a favorable balance between correction quality and efficiency. While GPT-4 with a basic prompt is slightly faster in generation time, MUSE+ achieves significantly higher correction quality (as shown in Table 1) with only a modest increase in generation time. The Original MUSE Model, with its retrieval mechanism, exhibits a slightly longer generation time than MUSE+. Human-generated corrections, while potentially high in quality, are orders of magnitude slower, highlighting the scalability advantage of automated methods like MUSE+. MUSE+'s efficient generation time, coupled with its superior correction quality, positions it as a practically viable solution for real-time misinformation correction on social media platforms.

4.7 Robustness Analysis: Prompt Variation

To assess the robustness of MUSE+ to variations in prompt wording, we experimented with different phrasings of the $P_{guidelines}$ component of our prompt. We created three variations of the guidelines, while keeping the P_{task}, $P_{context}$, and $P_{examples}$ components constant. The guideline variations included:

- **Guideline Variation 1 (Shorter Guidelines)**: A more concise version of the guidelines, using fewer words to describe each desired quality.
- **Guideline Variation 2 (Emphasis on Credibility)**: Guidelines with increased emphasis on the credibility and source referencing aspects of the correction.
- **Guideline Variation 3 (Emphasis on Persuasion)**: Guidelines with a stronger focus on the persuasive elements of the correction, aiming to maximize user acceptance.

We evaluated MUSE+ with each of these guideline variations, measuring the Overall Quality score to assess the sensitivity of our method to prompt wording. Table 6 presents the Overall Quality scores for MUSE+ with the original guidelines and the three variations.

Table 6. Robustness Analysis: Impact of Prompt Guideline Variations on Overall Quality

Prompt Guideline Variation	**Overall Quality**
Original Guidelines	**4.6**
Guideline Variation 1 (Shorter Guidelines)	4.5
Guideline Variation 2 (Emphasis on Credibility)	4.6
Guideline Variation 3 (Emphasis on Persuasion)	4.5

The results in Table 6 indicate that MUSE+ exhibits a reasonable degree of robustness to variations in prompt wording, specifically within the $P_{guidelines}$ component. While minor fluctuations in Overall Quality are observed across different guideline variations, the performance remains consistently high, and the differences are not statistically significant. This robustness suggests that MUSE+'s effectiveness is not overly sensitive to the precise phrasing of the guidelines, providing flexibility in prompt design and deployment. The consistently strong performance across variations further validates the underlying effectiveness of the prompt-based learning strategy.

4.8 Error Analysis: Common Error Types

To gain a qualitative understanding of the limitations and potential areas for improvement in MUSE+, we conducted an error analysis of the corrections generated by MUSE+. We manually examined a random sample of 100 corrections generated by MUSE+ and categorized the common types of errors observed. Table 7 summarizes the percentage of corrections exhibiting each error type.

Table 7. Error Analysis: Common Error Types in MUSE+ Corrections

Error Type	**Percentage of Corrections**
Minor Factual Inaccuracy	5%
Slight Lack of Clarity	8%
Suboptimal Persuasiveness	12%
Overly Concise Correction	7%
Repetitive Phrasing	3%
No Significant Error	**85%**

Table 7 shows that MUSE+ generates high-quality corrections with no significant errors in the vast majority (85%) of cases examined. The most common error types observed were related to suboptimal persuasiveness (12%) and slight lack of clarity (8%), suggesting areas for potential refinement in future work. Minor factual inaccuracies were relatively rare (5%), indicating the strong factual grounding of MUSE+'s corrections. The error analysis provides valuable insights into the strengths and weaknesses of MUSE+, guiding future research directions towards further enhancing its performance and addressing the identified limitations.

5 Conclusion

This paper presented MUSE+, a novel prompt-based approach for automated misinformation correction on social media, designed to leverage the generative power of Large Language Models. Our method addresses the critical need for scalable and effective solutions to combat the rapid spread of online misinformation, overcoming limitations of manual fact-checking and traditional LLM fine-tuning approaches. Through meticulously crafted prompts, MUSE+ effectively guides LLMs to generate accurate, clear, credible, and persuasive corrections for both textual and multimodal misinformation.

Extensive experimental evaluations, encompassing quantitative metrics, ablation studies, and expert human assessments, consistently demonstrate the superior performance of MUSE+ compared to strong baselines, including GPT-4, the original MUSE model, and human laypeople. Our findings highlight the synergistic effect of the prompt's components – task definition, context provision, correction guidelines, and few-shot examples – in enabling high-quality in-context learning. Furthermore, analyses across different misinformation types, efficiency measurements, and robustness tests confirm the practical viability and generalizability of MUSE+. Error analysis provided valuable insights into areas for future refinement, particularly concerning persuasiveness and clarity, suggesting avenues for further enhancing MUSE+'s capabilities.

The key contribution of MUSE+ lies in demonstrating the power of prompt engineering as a resource-efficient and highly adaptable strategy for complex tasks like misinformation correction. By framing the problem as a prompting challenge, we unlock the inherent knowledge and generative abilities of pre-trained LLMs, achieving state-of-the-art performance without the need for extensive task-specific training datasets or computationally intensive fine-tuning. This work opens up exciting possibilities for deploying LLM-driven misinformation correction systems at scale on social media platforms, offering a significant step towards creating a more informed and trustworthy online information environment. Future research will focus on refining the persuasiveness of generated corrections, enhancing user interaction with the system, and exploring the application of MUSE+ in diverse cultural and linguistic contexts.

References

1. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. science **359**(6380), 1146–1151 (2018)
2. Erokhin, D., Komendantova, N.: Earthquake conspiracy discussion on twitter. Humanities and Social Sciences Communications **11**(1), 1–10 (2024)
3. Hartley, K., Vu, M.K.: Fighting fake news in the covid-19 era: policy insights from an equilibrium model. Policy sciences **53**(4), 735–758 (2020)
4. Zhou, Y., Shen, J., Cheng, Y.: Weak to strong generalization for large language models with multi-capabilities. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=N1vYivuSKq
5. Hu, H., Zhou, Y., You, L., Xu, H., Wang, Q., Lian, Z., Yu, F.R., Ma, F., Cui, L.: Emobench-m: Benchmarking emotional intelligence for multimodal large language models (2025), https://arxiv.org/abs/2502.04424
6. Zhou, X., Sharma, A., Zhang, A.X., Althoff, T.: Correcting misinformation on social media with a large language model. CoRR **abs/2403.11169** (2024). https://doi.org/10.48550/ARXIV.2403.11169, https://doi.org/10.48550/arXiv.2403.11169
7. Zhou, Y., Zhang, J., Chen, G., Shen, J., Cheng, Y.: Less is more: Vision representation compression for efficient video generation with large language models (2024)
8. Jiang, Y., Wang, Y.: Large visual-language models are also good classifiers: A study of in-context multimodal fake news detection. CoRR **abs/2407.12879** (2024). https://doi.org/10.48550/ARXIV.2407.12879, https://doi.org/10.48550/arXiv.2407.12879
9. Zhou, Y., Song, L., Shen, J.: Training medical large vision-language models with abnormal-aware feedback. arXiv preprint arXiv:2501.01377 (2025)
10. Brashier, N.M., Pennycook, G., Berinsky, A.J., Rand, D.G.: Timing matters when correcting fake news. Proceedings of the National Academy of Sciences **118**(5), e2020043118 (2021)
11. Zhang, Y., Guo, B., Ding, Y., Liu, J., Qiu, C., Liu, S., Yu, Z.: Investigation of the determinants for misinformation correction effectiveness on social media during COVID-19 pandemic. Inf. Process. Manag. **59**(3), 102935 (2022). https://doi.org/10.1016/J.IPM.2022.102935, https://doi.org/10.1016/j.ipm.2022.102935
12. Li, X., Zhou, Y.: Disentangled and robust representation learning for bragging classification in social media. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
13. DeVerna, M.R., Yan, H.Y., Yang, K.C., Menczer, F.: Fact-checking information from large language models can decrease headline discernment. Proceedings of the National Academy of Sciences **121**(50), e2322823121 (2024)
14. Broda, E., Strömbäck, J.: Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. Annals of the International Communication Association **48**(2), 139–166 (2024)
15. Iizuka, R., Toriumi, F., Nishiguchi, M., Takano, M., Yoshida, M.: Impact of correcting misinformation on social disruption. Plos one **17**(4), e0265734 (2022)
16. Zhang, X., Yang, H., Young, E.F.Y.: Attentional transfer is all you need: Technology-aware layout pattern generation. In: 58th ACM/IEEE Design Automation Conference, DAC 2021, San Francisco, CA, USA, December 5-9, 2021.

pp. 169–174. IEEE (2021). https://doi.org/10.1109/DAC18074.2021.9586227, https://doi.org/10.1109/DAC18074.2021.9586227

17. Dai, Z., Yang, Z., Yang, Y., Carbonell, J.G., Le, Q.V., Salakhutdinov, R.: Transformer-xl: Attentive language models beyond a fixed-length context. In: Korhonen, A., Traum, D.R., Màrquez, L. (eds.) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. pp. 2978–2988. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/P19-1285, https://doi.org/10.18653/v1/p19-1285
18. Zhou, Y., Geng, X., Shen, T., Long, G., Jiang, D.: Eventbert: A pre-trained model for event correlation reasoning. In: Proceedings of the ACM Web Conference 2022. pp. 850–859 (2022)
19. Wang, Z., Li, M., Xu, R., Zhou, L., Lei, J., Lin, X., Wang, S., Yang, Z., Zhu, C., Hoiem, D., Chang, S., Bansal, M., Ji, H.: Language models with image descriptors are strong few-shot video-language learners. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022 (2022), http://papers.nips.cc/paper_files/paper/2022/hash/381ceeae4a1feb1abc59c773f7e61839-Abstract-Conference.html
20. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. CoRR **abs/2001.08361** (2020), https://arxiv.org/abs/2001.08361
21. Zhao, J., Wang, T., Abid, W., Angus, G., Garg, A., Kinnison, J., Sherstinsky, A., Molino, P., Addair, T., Rishi, D.: Lora land: 310 fine-tuned llms that rival gpt-4, A technical report. CoRR **abs/2405.00732** (2024). https://doi.org/10.48550/ARXIV.2405.00732, https://doi.org/10.48550/arXiv.2405.00732
22. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S.M., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y.: Sparks of artificial general intelligence: Early experiments with GPT-4. CoRR **abs/2303.12712** (2023). https://doi.org/10.48550/ARXIV.2303.12712, https://doi.org/10.48550/arXiv.2303.12712
23. Zhou, Y., Tao, W., Zhang, W.: Triple sequence generative adversarial nets for unsupervised image captioning. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7598–7602. IEEE (2021)
24. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020), https://jmlr.org/papers/v21/20-074.html
25. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R.B., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K., Davis, J.Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L.E., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P.W., Krass, M.S., Krishna, R., Kuditipudi, R., et al.: On the opportunities and risks of foundation models. CoRR **abs/2108.07258** (2021), https://arxiv.org/abs/2108.07258

26. Krupp, L., Steinert, S., Kiefer-Emmanouilidis, M., Avila, K.E., Lukowicz, P., Kuhn, J., Küchemann, S., Karolus, J.: Challenges and opportunities of moderating usage of large language models in education. In: Matviienko, A., Niess, J., Kosch, T. (eds.) Proceedings of the International Conference on Mobile and Ubiquitous Multimedia, MUM 2024, Stockholm, Sweden, December 1-4, 2024. pp. 249–254. ACM (2024). https://doi.org/10.1145/3701571.3701590, https://doi.org/10.1145/3701571.3701590