Preventing Accidental Sharing of Misinformation Using Large Language Models

Mirko Franco
mifranco@math.unipd.it
University of Padua
Padua, Italy

Valentin Grimm
valentin.grimm@th-owl.de
OWL University of Applied Sciences and Arts
Höxter, Germany

Eelco Herder
e.herder@uu.nl
Utrecht University
Utrecht, The Netherlands

Abstract

The proliferation of misinformation is one of the most pressing challenges in today's digital landscape, due to its far-reaching implications for public health, economic stability, trust in governmental institutions, and societal cohesion. Despite efforts to regulate online platforms and limit the spread of misinformation, many individuals are left behind because of their low digital literacy, level of education, and other contributing factors. In this context, we explore the use of Large Language Models (LLMs) to identify misinformation and we evaluate the capabilities of GPT-4.1-mini, as a representative example of these models. We then discuss how LLMs can help empower users to critically create and share information, thereby fostering more resilient online communities. We also present a set of possible interaction patterns for content creation and moderation.

CCS Concepts

• **Human-centered computing → Social media**; *Social networks*; • **Computing methodologies → Natural language generation**.

Keywords

misinformation, fake news, large language models, online social networks

ACM Reference Format:
Mirko Franco, Valentin Grimm, and Eelco Herder. 2025. Preventing Accidental Sharing of Misinformation Using Large Language Models. In *International Conference on Information Technology for Social Good (GoodIT '25), September 3–5, 2025, Antwerp, Belgium.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3748699.3749798

1 Introduction

The rise of online activities and the widespread adoption of Online Social Networks (OSNs) have amplified the dissemination of misinformation (i.e., news containing false or misleading information). The proliferation of misinformation constitutes one of the most pressing challenges in today's digital landscape, with effects that extend beyond the online world and impact our economies, health, trust in governments, and more. For example, increased engagement with misinformation was observed during the US presidential election campaigns of 2016 and 2020 [1]. In particular, during the 2016 campaign, the 20 top-performing false election stories on Facebook generated more shares, comments, and reactions than the 20 best-performing election stories from 19 major news outlets [35]. Similarly, during the 2018 presidential election in Brazil, the role of messaging applications in spreading misinformation became evident due to their technical affordances, such as the end-to-end encryption, minimal forwarding limits, and the lack of effective fact-checking mechanisms [3]. Even more concerning, during the COVID-19 pandemic, OSNs experienced a surge in misinformation that promoted erroneous practices and disseminated inaccurate information, thus affecting people's health and well-being [37]. Instead, in 2013, a fake news claiming that Barack Obama had been injured in an explosion wiped out $130 billion from the stock market, demonstrating the potential impact of misinformation on our economies and people's lives [33].

In order to build safer online spaces and safeguard users, online platforms implement various techniques to limit the spread of misinformation. For example, suspected misinformation violations are usually assessed by third-party reviewers, unlike traditional policy violations, which are reviewed by trained moderators. In particular, in the case of Meta, these reviewers are certified by an independent organization, namely the International Fact-Checking Network [19]. Recently, Meta began phasing out its fact-checking program and introduced a community-based (i.e., crowd-sourced) initiative in the US, similar to X's Community Notes, where ordinary users of social media comment on content deemed misleading [28]. However, the effectiveness of this approach in reducing engagement with misinformation remains unclear [6].

Unfortunately, many people are left behind in navigating online spaces, including misinformation, because of their low digital literacy, level of education, and other demographics [13]. According to [32], *accidental sharing* - defined as the dissemination of news without recognizing its falsity - is more prevalent than deliberate sharing, which occurs when individuals share information they are aware is false. These findings highlight the importance of empowering users with the skills to detect fake news and raising their awareness, as both are crucial for limiting the dissemination of misinformation and fostering resilient online communities. Moreover, Geeng *et al.* [17] observed that social media users may refrain from further investigating posts for various reasons, including challenges in navigating the user interface (UI) on mobile devices, overconfidence in the ability to identify misinformation, the cognitive effort required to conduct in-depth investigation, and a lack of interest in the content. Therefore, the design of effective user interfaces, experiences, and interactions is essential to enhance users' likelihood

GoodIT '25, Antwerp, Belgium

ACM ISBN 979-8-4007-2089-5/2025/09
https://doi.org/10.1145/3748699.3749798

of critically engaging with misinformation and to promote greater awareness, thereby aiming to limit the spread of misinformation.

Since the launch of *ChatGPT* in November 2022, Large Language Models (LLMs) - deep learning models trained on large scale *corpora* with advanced language understanding capabilities - have seen a rapid adoption and have been applied across several domains, including, but not limited to, healthcare [16, 42], education [18, 40], coding and web accessibility [2, 10, 29], content moderation [15, 22, 30], and beyond. Notable examples of LLMs include the OpenAI's GPT series (e.g., GPT-4o, GPT-4o mini, etc.), and the Meta's LlaMa (*Large Language Model Meta AI*) family.

In this context, our research aims to explore the potential of LLMs for identifying misinformation, and providing justifications and guidance for users in critically engaging with it. In particular, we evaluate the performance of GPT-4.1-mini, representative of contemporary LLMs, in detecting misinformation and extracting interpretable semantic properties, including polarization, emotionality, sentiment, and readability. Furthermore, we design and discuss a set of interaction patterns for content creation and moderation. The ultimate aim of this work is to enable AI-assisted, trustworthy content creation and engagement - empowering users to critically produce and share information, thereby reducing the spread of misinformation and fostering safer and more resilient online spaces and communities.

This work is in line with some of the Sustainable Development Goals (SDGs), which are the central component of the United Nations' 2030 Agenda for Sustainable Development. In particular, our research is related to Good Health and Well-Being (#3), Reduced Inequalities (#10), and Peace, Justice, and Strong Institutions (#16).

The remainder of the paper is organized as follows. Section 2 provides an overview of the related literature. Section 3 presents and discusses the results of the evaluation of GPT-4.1-mini in identifying misinformation, along with an analysis of semantic properties and their relationships to misinformation. A set of possible interaction patterns for content creation and moderation is presented in Section 4. Section 5 discusses the limitations of our work. Finally, we draw our conclusions and present some future research directions in Section 6.

2 Related Work

In this section, we provide a brief overview of the relevant literature about content moderation and misinformation detection, as well as emerging approaches for interacting with large language models beyond conventional chat-based interfaces.

2.1 Content Moderation and Misinformation

Content moderation, including the detection of misinformation, is one the pillars of online social networks for keeping users safe from malicious activities [19]. The importance of the topic is further demonstrated by the growing efforts of the research community. For example, considering the ongoing decentralization of social services, Franco *et al.* [14] designed two solutions to contrast the unauthorized spread of intimate content, even incorporating blockchain and Non-Fungible Tokens (NFTs), thereby enhancing the content moderation capabilities of decentralized social platforms. Instead, Jhaver *et al.* [21] introduced the concept of personal content moderation, defined as "*a form of content moderation in which users can configure or customize some aspects of their moderation experience based on the content of posts submitted by other users*", acknowledging that a one-size-fits-all solution would not be able to accommodate the needs of the (millions of) users of OSNs. Xu *et al.* [43] proposed *Safe Guard*, an LLM-based agent for the real-time detection of hate speech in social virtual reality (VR).

The detection and mitigation of misinformation has attracted significant interest due to its potential impact on various aspects of our lives, including our economies, health, trust, and more. For example, Biselli *et al.* [4] explored the use of personalized nudges to mitigate misinformation. Xu *et al.* [41] proposed a novel fuzzy logic-based neural network for detecting fake news and conducted a comprehensive evaluation using the LIAR2 dataset. Lee *et al.* [25] investigated users' perceptions of and interactions with an AI-based chatbot for fact-checking and debunking misinformation in private messaging platforms, acknowledging the differences in addressing misinformation across various social platforms (i.e., traditional social networks vs. private messaging platforms). A comprehensive review of the current challenges and open research directions related to fake news can be found in [46].

As large language models continue to gain traction across various domains, their potential applications in content moderation and misinformation detection (and mitigation) have become a growing focus of research. For example, Franco *et al.* [15] explored the use of LLMs into content moderation pipelines and assessed the capabilities of both open-source and commercial models in content moderation, arguing that these tools can support personalized content moderation and improve communication between users and platforms. Liu *et al.* [26] proposed FMDLlama, a LLM for the detection of financial disinformation. Ernst [12] identified the use of LLMs for detecting textual misinformation as a promising research area, highlighting a wide range of potential future directions, including implications for trust in artificial intelligence, and the ability to manage multilingual content.

Recently, social media platforms have increasingly adopted crowdsourced fact-checking approaches, wherein ordinary users of social media comment on content identified as misleading. Although the effectiveness of this approach remains unclear [6, 34], it has attracted substantial research attention. For example, Costabile *et al.* [9], recognizing that traditional fact-checking mechanisms are not sufficient given the vast volume of information circulating in digital spaces, evaluated the effectiveness of using an ensemble of generative agents to perform fact-checking.

However, prior work has not focused on supporting users in identifying and critically engaging with misinformation. Our research seeks to address this gap by exploring the potential of LLMs to detect misinformation and assist users in navigating misleading content on social media. We present examples and prototypes of possible user interactions, with the goal of empowering users to critically engage with and create content.

2.2 LLM Interactions

Beyond the interaction with LLMs through a traditional chat-based interface, various emerging interactions have been explored in numerous domains, including generation of social media content, coding, and more.

While there are several works concerned with the generation of social media content (e.g., [45], [38]), work on deeper interactions with the LLM are sparse. The *corporate communication companion* [27] supports creators in writing work-related social media posts. By outlining their desired goal and context, an LLM generates a post suggestion. Additionally, users can apply different "tones", such as tailored vs. generic, to steer the LLM. *Academify-Lite* [11] follows a similar path, by generating X posts on scientific topics. Content creators provide context such as paper abstracts and key findings and can adapt the tone (e.g. humorous), to generate post suggestions.

Lee *et al.* [23] provide an extensive discussion on designing intelligent and *interactive writing* assistants. They split the system controls into two paradigms: The *implicit* paradigm, where the assistant provides responses based on the users composition of the artifact and the *explicit* paradigm, where users request specific assistance. For example, Lee *et al.* [24] presents a tool for story writing that enables automatic suggestions for text completion. *Wordcraft* [44] incorporates concepts from both paradigms and integrates several interaction techniques, including "infilling" (subphrase based alternative suggestions), "continuation" (adding text to sub-phrase), "elaboration" (adding details to sub-phrase) and "free-form style transfer" (change style of sub-phrase). An explicit interaction pattern is discussed by Clark and Smith [8] where users are provided with two alternative options and can choose one of them, while personalizing the assistant. Grimm and Rubart [18] present a tool for the creation of interactive comics and provide interaction mechanisms by adapting the story and character context and recalculating the recommendation by leveraging the stochasticity of LLMs.

For the field of programming and, more specifically, code completion driven by LLMs, Husein *et al.* [20] provide an extensive overview and discussion. The community discusses this implicit interaction mostly on the token- and line-level (e.g., [36]) but there are also endeavors to generate larger blocks of code at once (e.g., [7]). Instead, CodeA11y [29] offers code suggestions based on specific guidelines, highlighting errors in the current context, and reminds the developers of previously identified accessibility issues that have not yet been resolved.

3 LLMs for Misinformation

In this section, we report the results of our analysis on the use of LLMs for misinformation detection and mitigation, as well as the relationship between the tone and the nature of the content.

3.1 Misinformation Detection and Mitigation

To assess the effectiveness of LLMs in identifying misinformation, we evaluated GPT-4.1-mini - part of the OpenAI's GPT series - on 120 posts, comprising 60 from the LIAR2 dataset [41] and 60 from PolitiFact[1], the well-know fact-checking website. LIAR2 is an enhanced version of the LIAR dataset, originally introduced by Wang [39] in 2017. Both datasets contain posts labeled by professional fact-checkers for misinformation detection tasks collected from the PolitiFact website. In addition to the 60 posts randomly sampled from the LIAR2 dataset, we collected 60 posts directly from PolitiFact to incorporate more recent content. In each case, we collected 10 posts for each of the six categories in which the date are divided. For the sake of clarity, Table 1 summarizes the possible ratings (i.e., labels), their associated numerical values, and concise descriptions, based on PolitiFact's Truth-O-Meter scale. For the scope of this work, we used a subset of the features included in the LIAR2 dataset; these features, along with their descriptions and examples, are reported Table 2.

We report here the structure of the prompt provided to the LLM for classifying the content into one of the six possible ratings.

The prompt contains explanations of the six labels (table 1), statement, data, speaker, speaker description and their credibility record (table 2). Based on this, the LLM is prompted to only provide the number for the classification.

Using this prompt, we evaluated the capabilities of GPT-4.1-mini to identify misinformation and to provide justifications. To account for the varying severity of possible misclassifications (i.e., predicting True instead of Mostly True is less severe than predicting True instead of Pants on Fire), we computed the absolute difference between the predicted rating and the actual rating. More formally, we used the following formula:

$$Distance = abs(PredictedRating - ActualRating) \tag{1}$$

Then, we calculated the mean of the distances for each possible original rating and data source. As reported in Table 3, the mean does not exceed 1.7 in any case (although, theoretically, it can range from 0 to 5). The overall mean across all 120 posts is equal to 1.08. These results indicate that the model can generally identify misinformation, although it is not consistently reliable, and therefore should not be used as standalone tool.

Using LLMs for misinformation detection and mitigation also enables the generation of human-readable explanations for the reasoning behind a decision. For example, in response to the claim "*In President Donald Trump's first 100 days, fentanyl seizures saved "119 million" to "258 million" lives*", which is correctly labeled as "Pants on Fire", the model provides the following justifactions for its decision:

- the claim is mathematically implausible and wildly exaggerated;
- it reflects misunderstanding or misrepresentation of drug seizure impact;
- there is no supporting evidence;
- although the speaker has a limited record of false statements, the statement is particularly misleading to the public.

These reasons align with PolitiFact's professional fact-checking analysis, demonstrating that such explanations can be both plausible and valuable in helping users understand and critically evaluate misinformation.

Finally, the use of LLMs in this context enables the generation of suggestions for creating or revising content to enhance its reliability. For example, in response to the claim "*Egg prices came down 50%*" made by Donald Trump on April 2, 2025, the LLM recommended adding context, such as the timeframe and baselines for comparison, avoiding exaggerated or absolute figures without

[1] https://www.politifact.com

Table 1: The PolitiFact's Truth-O-Meter Scale

Numerical Rating	Rating	Description
0	Pants on Fire	The statement is not accurate and makes a ridiculous claim
1	False	The statement is not accurate
2	Mostly False	The statement contains an element of truth but ignores critical facts that would give a different impression
3	Half True	The statement is partially accurate but leaves out important details or takes things out of context
4	Mostly True	The statement is accurate but needs clarification or additional information
5	True	The statement is accurate and there is nothing significant missing

Table 2: Details and Example of the Data

Feature	Description	Example
statement	Statement or news text	In President Donald Trump's first 100 days, fentanyl seizures saved "119 million" to "258 million" lives.
date	Date of the news or statement	April 29, 2025
speaker	Name of the speaker or platform where the content was posted	Pam Bondi
speaker_description	Short biography of the speaker or description of the platform where the content was posted	Pam Bondi, a former Hillsborough County prosecutor, is the U.S. attorney general, sworn in Feb. 5, 2025. [...]
<rating>_counts (for each of the 6 labels)	Counts of statements previously labeled as <rating>	3
label	Rating of the content	0 (Pants on Fire)

Table 3: Mean Distances between Predicted and Actual Ratings

	0	1	2	3	4	5
LIAR2	1.4	0.4	1.1	1.2	1.0	1.7
PolitiFact	0.9	0.5	1.0	1.2	1.5	1.1
All	1.15	0.45	1.05	1.2	1.25	1.4

sourcing, and using precise language language grounded on official data. It then proposed the following revised version: "*Egg prices dropped significantly — by nearly 50% from their peak in early 2023 — after the avian flu crisis eased and supply chains recovered, according to USDA and BLS data. However, prices remain higher than pre-pandemic levels.*".

Although preliminary, these results suggest that the use of LLMs can assist users in recognizing and misinformation and empowering them with the ability to critically create and/or share content, thereby reducing the number of accidental sharing of misinformation and improving the overall content quality on social media.

3.2 Tone Analysis

For the purpose of providing guidance to users in creating social media content, simply using the six labels from the LIAR dataset is rather abstract and lacks explanatory value.

According to the EU parliament, there are six main tactics to spread disinformation [31]:

1. Playing with emotions
2. Polarizing
3. Flooding the information space
4. Taking advantage of the confirmation bias
5. Manipulating context
6. Attacking and silencing critical voices

These factors can provide valuable insights to explain why certain content is aligning with a certain label. As we do not have the necessary context from the dataset to analyze content for 3 - 6, the first two, emotional language and polarization, might be detectable through a semantical analysis.

Additionally, text complexity and reading difficulty have been discussed and analyzed in the context of fake news detection (e.g., [5]). Disinformation tends to be transported in a more superficial way and rather tries to take advantage of the confirmation bias (point 4) or might be hard to understand on purpose (point 5).

Lastly, sarcasm is a common rhetorical device to transport information in a tangible way. A sarcastic statement could be factually false, yet have the intention to emphasize on the truth.

Therefore, we conducted an analysis on the extracted 120 samples by letting the LLM generate a polarization score (0 to 1), emotionality (-1 to 1), a sentiment ("negative", "neutral", "positive"), reading difficulty (0 to 1) and sarcasm (0 to 1). We used the same prompt as for the fake news classification but asked it for the aforementioned scores instead of the fake news classification.

For the analysis of the results, first we conducted an ANOVA analysis, comparing the original labels with the generated scores. The results are shown in Table 4.

Table 4: ANOVA Analysis and Correlations of Tone Features with Respect to the Fake News Label

Feature	F-Statistic	Correlation	p-Value
polarization	18.63	-0.698	2.86e-14
sentiment	14.40	0.642	1.93e-11
emotionality	9.43	0.550	1.11e-07
reading_difficulty	5.97	-0.416	9.52e-05
sarcasm	1.90	-0.248	4.36e-01

The F-Statistic shows the ratio of between-group variance to within-group variance. A higher value means a higher explanatory value. The bonferroni corrected p-Value (to adjust for multiple variables), shows the significance of the results. The results show that all features but sarcasm are significantly associated with the fake news labels. Especially polarization and the sentiment are strongly significant, while emotionality and reading difficulty are less significant. Figure 1 shows the relationship between the scores and the labels. Sarcasm is not shown as it was mostly 0 with some scores between 0.1 and 0.2 for Pants on Fire or False labels.

The statistical and visual results show that there is a relevant relationship between most scores generated by the LLM (besides sarcasm) and the fake news label, supporting our approach to explain and assist users in writing appropriate content based on these features.

4 Use Cases

In the previous section, we have shown how LLMs can quite reliably identify misinformation and provide human-language justifications. Elaborating on these results, we present a number of use cases that showcase how these insights can be exploited to guide users who are about to inadvertently create or share posts that contain misinformation.

4.1 Content Sharing

Interaction mechanisms for LLM-assisted content sharing are shown in Figure 2. When attempting to share a potentially manipulative post (left), the user is informed about the problematic practices of it (top right) and can dive into detailed analysis of sub-phrases of the post (bottom right).

4.2 Content Creation with LIAR Classification

Interaction mechanisms for LLM-assisted content creation based on the LIAR2 classification pattern are shown in Figure 3. While writing a post, the LLM provides (near) real-time information about the classification on the LIAR2 scale (top left). The user can ask about the reasons of the classification (bottom left) or ask for a refactoring of the post (right).

4.3 Content Creation with Semantic Analysis

Interaction mechanisms for LLM-assisted content creation based on the semantic properties are shown in Figure 4. While writing a post, the LLM provides (near) real-time information about the semantic properties relevant in the context of fake news. The user can ask for a refactoring of the post with an explanation.

4.4 Time-Efficient LLM Interactions

Due to limited capabilities of the LLM, misunderstandings or an incomplete context, it can be valuable to add interaction mechanisms that enable collaboration with the assistant to correct and steer its recommendations. While chat-interactions are a very common pattern to interact with LLMs, we argue that this can lead to an "inconvenient" complexity, increase cognitive demands and slow down the process. Therefore, we propose three (non-exhaustive) interaction patterns with the LLM to foster user interaction while acknowledging context-specific user preferences as depicted in Figure 5. These are:

- Recalculation: Asking the LLM to rephrase the recommendation for fast (yet dirty) sampling.
- Adding context: Users can add contextual information about the content that is not available to the LLM or clarify misunderstandings of the LLM.
- Templates: The LLM automatically detects contextual information in its post-recommendation and masks them. Users can add the content of the templates field to trigger an update from the assistant.

4.5 Inline Suggestions

To create an even faster and more immersive experience, inline-suggestions can be of value (Figure 6). These are a common pattern in LLM-assisted coding applications (e.g., [36]) and story-writing (e.g., [24]). In our proposal, the assistant provides a recommendation when the user stops writing. This recommendation can replace the initial text based on a user command inline. The concepts from the previous section can be applied, by e.g., offering a template-based recommendation or showing the explanation for the recommendation. Although this approach can speed-up the writing process, it arguably reduces the explanatory/educational effect and makes it less trivial to effectively integrate interaction patterns such as "adding context" and "recalculation".

4.6 Summary of Design Patterns

To support users in critically engaging with misinformation and crafting trustworthy content, we derived a set of interaction design patterns tailored to our use case. These include LLM-powered fake-news classification with accompanying explanations as well as semantic property assessments paired with contextual justifications. Based on these semantic properties, highlighting problematic rhetorical practices (e.g., emotionally charged or polarizing) and sub-phrase-level analyzes, to explore why specific parts of posts might be toxic, is facilitated.

In addition, we discussed improved content suggestions that are transparently labeled and accompanied by explanations to promote user understanding and trust. To streamline user interaction, we

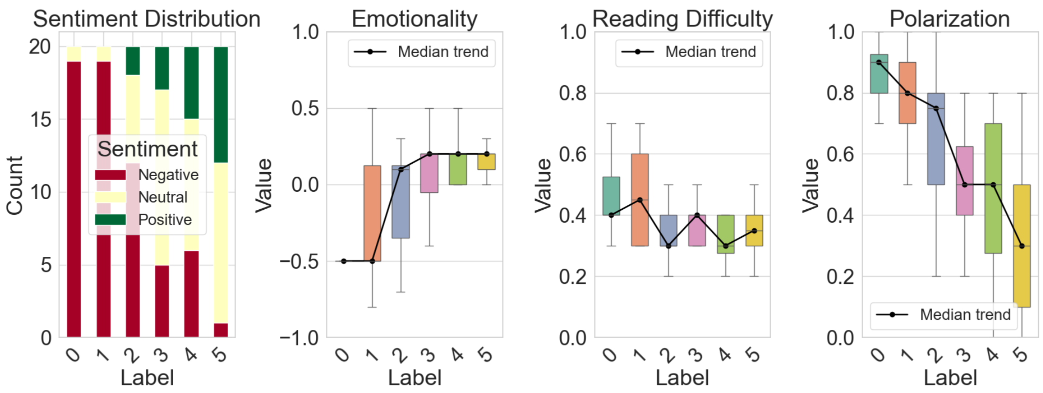
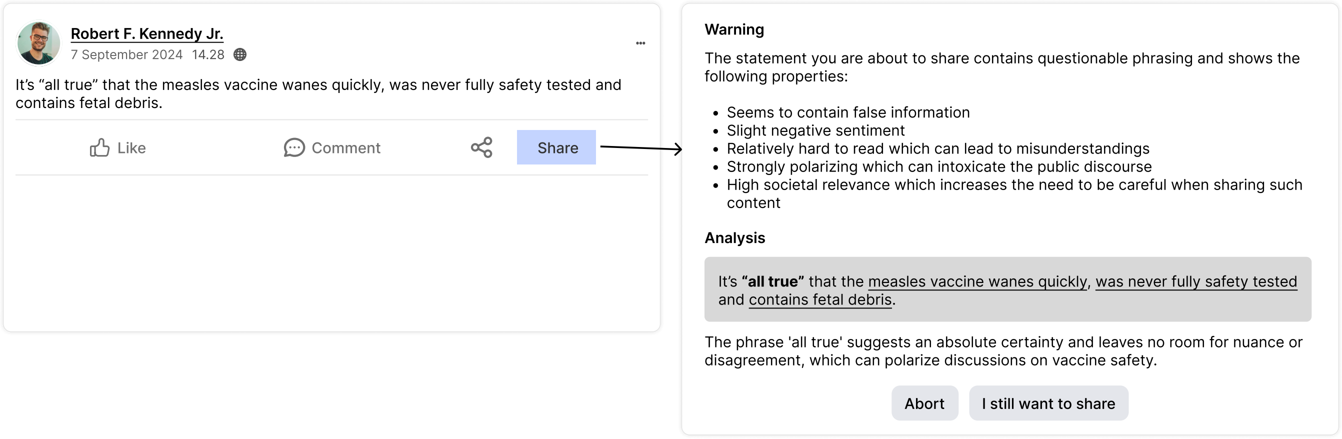


Figure 1: Plots that highlight the relationship between the different semantic dimensions and fake news labels



It's "all true" that the measles vaccine wanes quickly, was never fully safety tested and contains fetal debris.

Warning

The statement you are about to share contains questionable phrasing and shows the following properties:

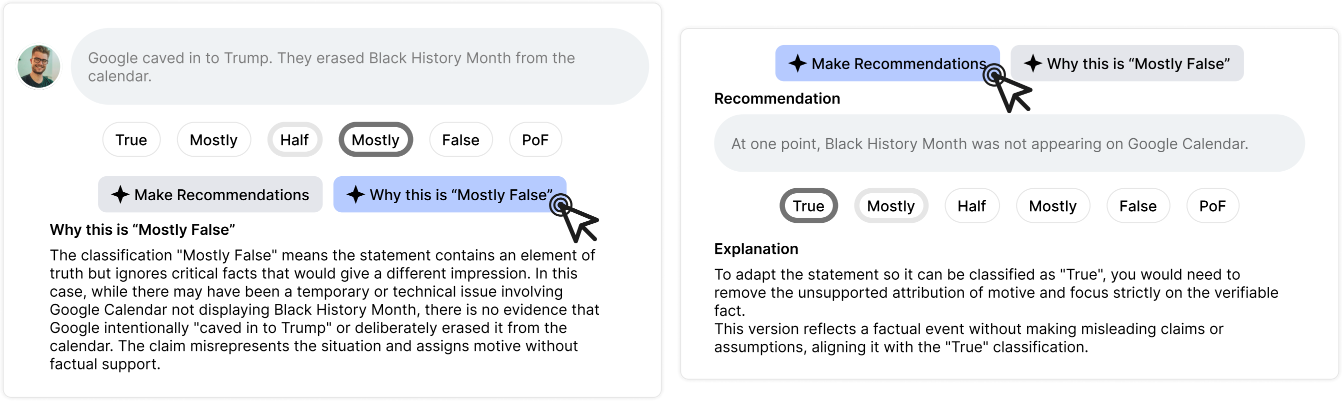
- Seems to contain false information
- Slight negative sentiment
- Relatively hard to read which can lead to misunderstandings
- Strongly polarizing which can intoxicate the public discourse
- High societal relevance which increases the need to be careful when sharing such content

Analysis

It's **"all true"** that the measles vaccine wanes quickly, was never fully safety tested and contains fetal debris.

The phrase 'all true' suggests an absolute certainty and leaves no room for nuance or disagreement, which can polarize discussions on vaccine safety.

Figure 2: Example UI "Sharing Content". Left: Social Media post, Right: Intervention Pop-up



Google caved in to Trump. They erased Black History Month from the calendar.

Why this is "Mostly False"

The classification "Mostly False" means the statement contains an element of truth but ignores critical facts that would give a different impression. In this case, while there may have been a temporary or technical issue involving Google Calendar not displaying Black History Month, there is no evidence that Google intentionally "caved in to Trump" or deliberately erased it from the calendar. The claim misrepresents the situation and assigns motive without factual support.

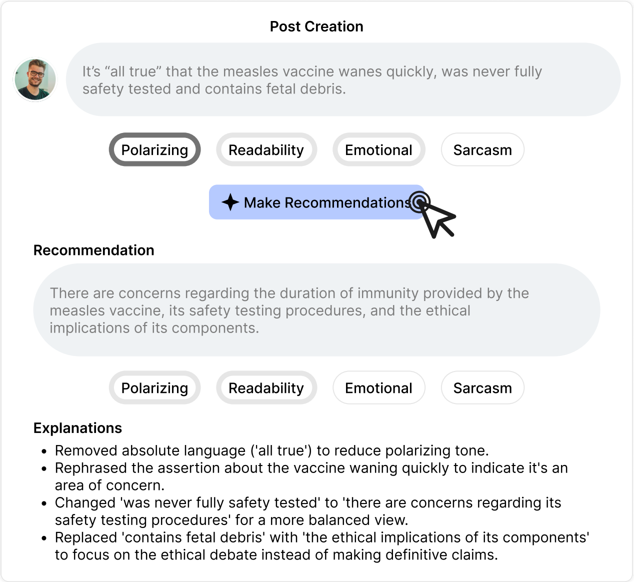
Recommendation

At one point, Black History Month was not appearing on Google Calendar.

Explanation

To adapt the statement so it can be classified as "True", you would need to remove the unsupported attribution of motive and focus strictly on the verifiable fact.
This version reflects a factual event without making misleading claims or assumptions, aligning it with the "True" classification.

Figure 3: Example UI "Content Creation LIAR". Left: Explanation for the classification, Right: Recommendation and explanation to improve the content

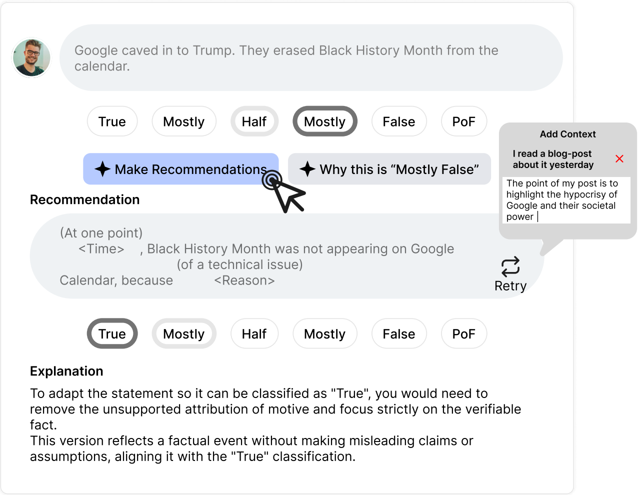


It's "all true" that the measles vaccine wanes quickly, was never fully safety tested and contains fetal debris.

There are concerns regarding the duration of immunity provided by the measles vaccine, its safety testing procedures, and the ethical implications of its components.

- Removed absolute language ('all true') to reduce polarizing tone.
- Rephrased the assertion about the vaccine waning quickly to indicate it's an area of concern.
- Changed 'was never fully safety tested' to 'there are concerns regarding its safety testing procedures' for a more balanced view.
- Replaced 'contains fetal debris' with 'the ethical implications of its components' to focus on the ethical debate instead of making definitive claims.

Figure 4: Example UI "Content Creation Semantics". Content creation area, classifications, recommendation (with updated classifications), explanations (top-to-bottom)



Google caved in to Trump. They erased Black History Month from the calendar.

I read a blog-post about it yesterday

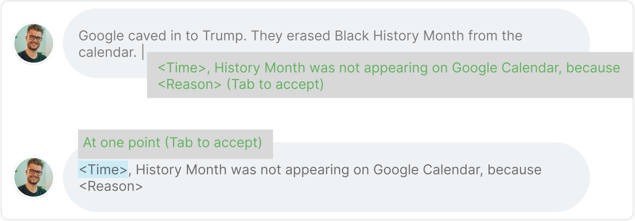
The point of my post is to highlight the hypocrisy of Google and their societal power

(At one point)
<Time> , Black History Month was not appearing on Google
(of a technical issue)
Calendar, because <Reason>

To adapt the statement so it can be classified as "True", you would need to remove the unsupported attribution of motive and focus strictly on the verifiable fact.
This version reflects a factual event without making misleading claims or assumptions, aligning it with the "True" classification.

Figure 5: Example UI "LLM Interactions". Showing template-based recommendations and quick interactions with the LLM-based assistant (Retry, "Add Context")

outline four complementary mechanisms for LLM collaboration: (1) Recalculation, enabling quick regeneration of suggestions; (2) Context enrichment, where users can clarify intent or background information; (3) Template-based editing, which dynamically highlights adjustable components in LLM outputs; and (4) Inline suggestions, offering in-situ improvements during the writing process. Together, these design patterns aim to make LLM-supported content creation both efficient and educational, empowering users to better recognize and avoid misinformation tactics.



Google caved in to Trump. They erased Black History Month from the calendar.

<Time>, History Month was not appearing on Google Calendar, because <Reason> (Tab to accept)

At one point (Tab to accept)

<Time>, History Month was not appearing on Google Calendar, because <Reason>

Figure 6: Example UI "Inline Suggestions". Top: initial inline, template-based suggestion, Bottom: Inline suggestion for template field

5 Limitations

As this is a preliminary study, there are several limitations with respect to the technical evaluation. We used a rather small sub-dataset with 120 samples for the fake news classification and tone analysis. The dataset itself contains mostly U.S.-centric political content. Content from other domains and geographical regions might yield different results. Moreover, only a single LLM architecture, namely GPT-4.1-mini, has been used and only a single, hand-crafted prompt for both, fake news classification and tone analysis. With respect to the tone analysis, we only provide an indirect validation of the results by comparing them with the original labels for the fake news classification and assess them based on the actual and expected outcome.

For the use case designs, we present several interaction designs and ways to utilize LLMs for assisting in the content creation process. Yet, we did not conduct any user evaluations or usability tests and can not make any claims about the effectiveness for end-users and how intuitive they are to use. If these concepts can actually improve users' critical thinking, reduce misinformation or increase trust in the moderation process and how they effect cognitive burden compared to chat-based interactions stay open questions for future work.

Additionally, the use of LLMs presents numerous technical and ethical challenges. Despite their advanced language understanding capabilities, they are still prone to hallucinations, i.e., generating plausible but factually inaccurate responses, and from limited knowledge recency, which hinders their ability to process the most up-to-date information. Moreover, even with safety alignment measures and protective tools in place, LLMs can still generate harmful and offensive language, raising concerns in critical domains such as content moderation. Last but not least, their deployment involves substantial financial and environmental costs, particularly related to energy consumptions and infrastructure demands. The implications for users' privacy and safety also remain insufficiently understood.

6 Conclusion

The spread of misinformation remains a critical concern in today's digital environment, undermining public health efforts, destabilizing economies, eroding trust in governments, and threatening the cohesion of our societies. While regulatory interventions and platform-based countermeasures have been introduced, they often fail to address the needs of users with limited digital literacy or lower levels of education. Therefore, news tools and strategies are needed to both detect misinformation effectively and empower individuals to engage with content more critically, aiming to create safer and more resilient online communities.

In this work, we have investigated the role of LLMs, namely GPT-4.1-mini, in identifying misinformation and extracting interpretable semantic features, such as polarization, sentiment, emotionality, and readability. Our findings suggest that LLMs can reasonably identify misleading content, provide human-readable explanations for their classification, and offer constructive suggestions for revising content to enhance credibility and accuracy. We have also proposed and discussed a set of interaction patterns that illustrate how such models might be integrated into tools to guide content creation, provide feedback, and enable more informed sharing decisions.

We plan to extend our research in several directions. Primarily, we aim to assess the effectiveness of LLMs on a more comprehensive dataset, even considering different contexts, languages, and cultures. Moreover, we would like to investigate the efficacy of our prototypes and revising mechanism through a user study. We also plan to generalize our approach to different types of data, including images and videos.

References

[1] Omar Alkhalili and Stefan A. Robila. 2021. Tracking the Impact of Fake News on US Election Cycles. In *Proceedings of the Conference on Information Technology for Social Good* (Roma, Italy) *(GoodIT '21)*. Association for Computing Machinery, New York, NY, USA, 192–197. doi:10.1145/3462203.3475920

[2] Barry Bassi, Giovanni Delnevo, Mirko Franco, Ombretta Gaggi, Salvatore Gatto, Silvia Mirri, and Kelvin Olaiya. 2025. Supporting Accessibility Auditing and HTML Validation using Large Language Models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing* (Catania International Airport, Catania, Italy) *(SAC '25)*. Association for Computing Machinery, New York, NY, USA, 27–31. doi:10.1145/3672608.3707912

[3] Fabrício Benevenuto and Philipe Melo. 2024. Misinformation Campaigns through WhatsApp and Telegram in Presidential Elections in Brazil. *Commun. ACM* 67, 8 (Aug. 2024), 72–77. doi:10.1145/3653325

[4] Tom Biselli, Katrin Hartwig, and Christian Reuter. 2025. Mitigating Misinformation Sharing on Social Media through Personalised Nudging. *Proc. ACM Hum.-Comput. Interact.* 9, 2, Article CSCW136 (May 2025), 44 pages. doi:10.1145/3711034

[5] Anshika Choudhary and Anuja Arora. 2021. Linguistic feature based learning model for fake news detection and classification. *Expert Systems with Applications* 169 (May 2021), 114171. doi:10.1016/j.eswa.2020.114171

[6] Yuwei Chuai, Haoye Tian, Nicolas Pröllochs, and Gabriele Lenzini. 2024. Did the Roll-Out of Community Notes Reduce Engagement With Misinformation on X/Twitter? *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 428 (Nov. 2024), 52 pages. doi:10.1145/3686967

[7] Matteo Ciniselli, Nathan Cooper, Luca Pascarella, Denys Poshyvanyk, Massimiliano Di Penta, and Gabriele Bavota. 2021. An Empirical Study on the Usage of BERT Models for Code Completion. In *2021 IEEE/ACM 18th International Conference on Mining Software Repositories (MSR)*. 108–119. doi:10.1109/MSR52588.2021.00024

[8] Elizabeth Clark and Noah A. Smith. 2021. Choose Your Own Adventure: Paired Suggestions in Collaborative Writing for Evaluating Story Generation Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 3566–3575. doi:10.18653/v1/2021.naacl-main.279

[9] Luigia Costabile, Gian Marco Orlando, Valerio La Gatta, and Vincenzo Moscato. 2025. Assessing the Potential of Generative Agents in Crowdsourced Fact-Checking. arXiv:2504.19940 [cs.CL] https://arxiv.org/abs/2504.19940

[10] Giovanni Delnevo, Manuel Andruccioli, and Silvia Mirri. 2024. On the Interaction with Large Language Models for Web Accessibility: Implications and Challenges. In *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*. 1–6. doi:10.1109/CCNC51664.2024.10454680

[11] Maha Eldamnhory. 2024. *Academify-Lite: LLM-powered social media post generator for academic content.* Master's thesis. M. Eldamnhory.

[12] Marina Ernst. 2024. Identifying textual disinformation using Large Language Models. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval* (Sheffield, United Kingdom) *(CHIIR '24)*. Association for Computing Machinery, New York, NY, USA, 453–456. doi:10.1145/3627508.3638315

[13] Mirko Franco, Salah A. Falioun, Karen E. Fisher, Ombretta Gaggi, Yacine Ghamri-Doudane, Ayat J. Nashwan, Claudio E. Palazzi, and Mohammed Shwamra. 2022. A Technology Exploration towards Trustable and Safe Use of Social Media for Vulnerable Women based on Islam and Arab Culture. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good* (Limassol, Cyprus) *(GoodIT '22)*. Association for Computing Machinery, New York, NY, USA, 138–145. doi:10.1145/3524458.3547259

[14] Mirko Franco, Ombretta Gaggi, Barbara Guidi, Andrea Michienzi, and Claudio E. Palazzi. 2023. A decentralised messaging system robust against the unauthorised forwarding of private content. *Future Generation Computer Systems* 145 (2023), 211–222. doi:10.1016/j.future.2023.03.025

[15] Mirko Franco, Ombretta Gaggi, and Claudio E. Palazzi. 2025. Integrating Content Moderation Systems with Large Language Models. *ACM Trans. Web* 19, 2, Article 18 (May 2025), 21 pages. doi:10.1145/3700789

[16] Marco Furini, Michele Mariani, Sara Montagna, and Stefano Ferretti. 2024. Conversational Skills of LLM-based Healthcare Chatbot for Personalized Communications. In *Proceedings of the 2024 International Conference on Information Technology for Social Good* (Bremen, Germany) *(GoodIT '24)*. Association for Computing Machinery, New York, NY, USA, 429–432. doi:10.1145/3677525.3678693

[17] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376784

[18] Valentin Grimm and Jessica Rubart. 2024. Authoring Educational Hypercomics assisted by Large Language Models. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 88–97. doi:10.1145/3648188.3675124

[19] Alon Halevy, Cristian Canton-Ferrer, Hao Ma, Umut Ozertem, Patrick Pantel, Marzieh Saeidi, Fabrizio Silvestri, and Ves Stoyanov. 2022. Preserving integrity in online social networks. *Commun. ACM* 65, 2 (Jan. 2022), 92–98. doi:10.1145/3462671

[20] Rasha Ahmad Husein, Hala Aburajouh, and Cagatay Catal. 2025. Large language models for code completion: A systematic literature review. *Computer Standards & Interfaces* 92 (2025), 103917. doi:10.1016/j.csi.2024.103917

[21] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. Personalizing Content Moderation on Social Media: User Perspectives on Moderation Choices, Interface Design, and Labor. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 289 (Oct. 2023), 33 pages. doi:10.1145/3610080

[22] Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. LLM-Mod: Can Large Language Models Assist Content Moderation?. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 217, 8 pages. doi:10.1145/3613905.3650828

[23] Mina Lee, Katy Ilonka Gero, John Joon Young Chung, Simon Buckingham Shum, Vipul Raheja, Hua Shen, Subhashini Venugopalan, Thiemo Wambsganss, David Zhou, Emad A. Alghamdi, Tal August, Avinash Bhat, Madiha Zahrah Choksi, Senjuti Dutta, Jin L.C. Guo, Md Naimul Hoque, Yewon Kim, Simon Knight, Seyed Parsa Neshaei, Antonette Shibani, Disha Shrivastava, Lila Shroff, Agnia Sergeyuk, Jessi Stark, Sarah Sterman, Sitong Wang, Antoine Bosselut, Daniel Buschek, Joseph Chee Chang, Sherol Chen, Max Kreminski, Joonsuk Park, Roy Pea, Eugenia Ha Rim Rho, Zejiang Shen, and Pao Siangliulue. 2024. A Design Space for Intelligent and Interactive Writing Assistants. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 1054, 35 pages. doi:10.1145/3613904.3642697

[24] Mina Lee, Percy Liang, and Qian Yang. 2022. CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 388, 19 pages. doi:10.1145/3491102.3502030

[25] Tuan-He Lee and Susan R. Fussell. 2025. Countering Misinformation in Private Messaging Groups: Insights From a Fact-checking Chatbot. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP10 (Jan. 2025), 30 pages. doi:10.1145/3701189

[26] Zhiwei Liu, Xin Zhang, Kailai Yang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2025. FMDLlama: Financial Misinformation Detection Based on Large Language Models. In *Companion Proceedings of the ACM on Web Conference 2025* (Sydney NSW, Australia) *(WWW '25)*. Association for Computing Machinery, New York, NY, USA, 1153–1157. doi:10.1145/3701716.3715599

[27] Zhuoran Lu, Sheshera Mysore, Tara Safavi, Jennifer Neville, Longqi Yang, and Mengting Wan. 2024. Corporate Communication Companion (CCC): An LLM-empowered Writing Assistant for Workplace Social Media. arXiv:2405.04656 [cs.HC] https://arxiv.org/abs/2405.04656

[28] Meta. 2025. Community Notes: A New Way to Add Context to Posts. https://transparency.meta.com/features/community-notes. Accessed 20th May, 2025.

[29] Peya Mowar, Yi-Hao Peng, Jason Wu, Aaron Steinfeld, and Jeffrey P Bigham. 2025. CodeA11y: Making AI Coding Assistants Useful for Accessible Web Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 45, 15 pages. doi:10.1145/3706598.3713335

[30] Mohammad Nadeem, Laeeba Javed, Shahab Saquib Sohail, Erik Cambria, and Amir Hussain. 2024. Are Foundation Models the Next-Generation Social Media Content Moderators? *IEEE Intelligent Systems* 39, 6 (2024), 70–80. doi:10.1109/MIS.2024.3477109

[31] European Parliament. 2025. Spotting disinformation: Six tactics used to fool us. https://www.europarl.europa.eu/topics/en/article/20250227STO27081/spotting-disinformation-six-tactics-used-to-fool-us. Accessed 28th May 2025.

[32] K. Peren Arin, Deni Mazrekaj, and Marcel Thum. 2023. Ability of detecting and willingness to share fake news. *Scientific Reports* 13, 1 (2023). doi:10.1038/s41598-023-34402-6

[33] Kenneth Rapoza. 2017. Can 'Fake News' Impact The Stock Market? https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#4b3542d52fac. Accessed 20th May, 2025.

[34] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 1736–1746. doi:10.1145/3511808.3557279

[35] Craig Silverman. 2016. This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook. Accessed 20th May, 2025.

[36] Ze Tang, Jidong Ge, Shangqing Liu, Tingwei Zhu, Tongtong Xu, Liguo Huang, and Bin Luo. 2024. Domain Adaptive Code Completion via Language Models and Decoupled Domain Databases. In *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering* (Echternach, Luxembourg) *(ASE '23)*. IEEE Press, 421–433. doi:10.1109/ASE56229.2023.00076

[37] Samia Tasnim, Mahbub Hossain, and Hoimonty Mazumder. 2020. Impact of rumors and misinformation on COVID-19 in Social Media. *Journal of Preventive Medicine and Public Health* 53, 3 (2020), 171 – 174. doi:10.3961/JPMPH.20.094

[38] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining* 15, 1 (2025), 1–30.

[39] 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)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. doi:10.18653/v1/P17-2067

[40] Xinyu Jessica Wang, Christine P. Lee, and Bilge Mutlu. 2025. LearnMate: Enhancing Online Education with LLM-Powered Personalized Learning Plans and Support. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 373, 10 pages. doi:10.1145/3706599.3719857

[41] Cheng Xu and M-Tahar Kechadi. 2024. An Enhanced Fake News Detection System With Fuzzy Deep Learning. *IEEE Access* 12 (2024), 88006–88021. doi:10.1109/ACCESS.2024.3418340

[42] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1, Article 31 (March 2024), 32 pages. doi:10.1145/3643540

[43] Yiwen Xu, Qinyang Hou, Hongyu Wan, and Mirjana Prpa. 2024. Safe Guard: an LLM-agent for Real-time Voice-based Hate Speech Detection in Social Virtual Reality. arXiv:2409.15623 [eess.AS] https://arxiv.org/abs/2409.15623

[44] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. doi:10.1145/3490099.3511105

[45] Yuying Zhao, Yu Wang, Xueqi Cheng, Anne Marie Tumlin, Yunchao Liu, Damin Xia, Meng Jiang, and Tyler Derr. 2025. Amplifying Your Social Media Presence: Personalized Influential Content Generation with LLMs. arXiv:2505.01698 [cs.SI] https://arxiv.org/abs/2505.01698

[46] Xinyi Zhou and Reza Zafarani. 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Comput. Surv.* 53, 5, Article 109 (Sept. 2020), 40 pages. doi:10.1145/3395046