Silver Lining in the Fake News Cloud: Can Large Language Models Help Detect Misinformation?

Raghvendra Kumar, Bhargav Goddu, Sriparna Saha, *Senior Member, IEEE*, and Adam Jatowt

***Abstract*—In the times of advanced generative artificial intelligence, distinguishing truth from fallacy and deception has become a critical societal challenge. This research attempts to analyze the capabilities of large language models (LLMs) for detecting misinformation. Our study employs a versatile approach, covering multiple LLMs with few- and zero-shot prompting. These models are rigorously evaluated across various fake news and rumor detection datasets. Introducing a novel dimension, we additionally incorporate sentiment and emotion annotations to understand the emotional influence on misinformation detection using LLMs. Moreover, to extend our inquiry, we employ ChatGPT to intentionally distort authentic news as well as human-written fake news, utilizing zero-shot and iterative prompts. This deliberate corruption allows for a detailed examination of various parameters such as abstractness, concreteness, and named entity density, providing insights into differentiating between unaltered news, human-written fake news, and its LLM-corrupted counterpart. Our findings aspire to furnish a refined framework for discerning authentic news, human-generated misinformation, and LLM-induced distortions. This multifaceted approach, utilizing various prompt techniques, contributes to a comprehensive understanding of the subtle variations shaping misinformation sources.**

***Impact Statement*—In the evolving landscape of advanced generative artificial intelligence, LLMs function as both the vigilant guardian and the deceitful manipulator when coping with misinformation. This research looks deeper into this conflicting nature of LLMs by evaluating multiple LLMs across diverse datasets. Our statistical analyses investigate the efficacy of LLMs in discerning rumor content, exploring the impact of various prompting techniques and hyperparameter sensitivity. Additionally, we assess the influence of sentiment and emotion on LLMs' ability to identify rumors. Introducing a novel perspective, we intentionally corrupt authentic news using ChatGPT, quantifying parameters such as concreteness and abstractness. Furthermore, our dual approach involves distorting human-generated fake news with LLMs, particularly GPT, enhancing the robustness of our framework. In the short run, the research provides valuable insights for improving current misinformation detection tools, enhancing their effectiveness through a better understanding of LLMs. In the long run, these findings contribute to the responsible evolution of artificial intelligence and aim to mitigate misinformation risks, fostering a more resilient and trustworthy information ecosystem as society increasingly relies on advanced generative AI technologies.**

***Index Terms*—Fake news, large language models (LLMs), misinformation detection, prompting techniques, rumor news, sentiment and emotion.**

I. Introduction

In the current digital era, which is abundant with technologies of generative artificial intelligence, there is an unprecedented surge in misinformation, which can be attributed to several key factors. First, the internet and social media platforms have democratized information sharing, allowing anyone to disseminate content without fact-checking. This has created an environment where both *fake news*, deliberately fabricated to deceive, and *rumor news*, which stems from unverified or loosely sourced information, can spread rapidly. Additionally, the algorithms used by social media platforms tend to prioritize sensational or controversial content, boosting the creation and circulation of misleading information for increased engagement and visibility. As a result, distinguishing between fact and fiction has become more challenging. Of course, one should not forget that large language models (LLMs), while powerful in generating humanlike text, still hallucinate and produce factual and other types of mistakes.

Additionally, we incorporate sentiment and emotional analysis into our methodology, recognizing the pivotal role these factors play in identifying deceptive content, intending to contribute to a more robust defence against the proliferation of misleading information in online spaces. Recent studies [1], [2], [3], [4] have shown that malicious actors can exploit these models to generate deceptive content because LLMs, crafted for humanlike text generation, pose a challenge for traditional detection methods to distinguish deceptive content due to their ability to mimic authentic communication. Furthermore, the extensive and diverse training data of LLMs may inadvertently perpetuate biased patterns, empowering malicious actors to create misleading information that appears legitimate. As LLMs become increasingly integrated into various online platforms, there is a heightened risk of amplifying the spread of

Manuscript received 7 March 2024; revised 10 June 2024 and 16 July 2024; accepted 3 August 2024. Date of publication 8 August 2024; date of current version 9 January 2025. The work of Raghvendra Kumar was supported by the Prime Minister's Research Fellowship (PMRF). The work of Sriparna Saha was supported by the "Technology Innovation Hub, Vishlesan I-Hub Foundation IIT Patna" under Project TIH/CSE/ASMO/05. This article was recommended for publication by Associate Editor Lin Liu upon evaluation of the reviewers' comments. *(Corresponding author: Raghvendra Kumar.)*

Raghvendra Kumar, Bhargav Goddu, and Sriparna Saha are with the Department of Computer Science and Engineering, Indian Institute of Technology Patna 801106, India (e-mail: raghvendra_2221cs27@iitp.ac.in; goddu_2211cs07@iitp.ac.in; sriparna@iitp.ac.in).

Adam Jatowt is with the Department of Computer Science, University of Innsbruck, 6020 Innsbruck, Austria (e-mail: adam.jatowt@uibk.ac.at).

Digital Object Identifier 10.1109/TAI.2024.3440248

misinformation,[1] underscoring the need for reliable detection mechanisms.

Research Inspiration: Motivated by the above considerations, *we reverse the problem* and investigate whether it is possible to *harness LLMs for misinformation detection*. In the second part of the article, we conduct a detailed analysis of linguistic parameters aimed at enhancing the *discriminative capability of misinformation assessment*, which is vital to accurately distinguish between authentic and false information. Specifically, we examine the concreteness and abstractness of the information—whether the content is clear and tangible or generalized and detached from specific details. We also assess the readability of the content and the proportion of named entities. By scrutinizing these metrics, we aim to identify signals that could potentially support examining the authenticity of news articles. Through systematic evaluation and comparison of these parameters across *authentic news, human-generated fake news, and LLM-generated fake news*, our aim is to identify text characteristics that improve tools for detecting misinformation, thereby enhancing public information integrity.

Contributions: In summary, our research makes the following contributions, each shedding light on critical aspects of the subject matter.

1) Evaluating the efficacy of multiple LLMs in discerning rumor content across six distinct rumor datasets.
2) Investigating the impact of various prompting techniques, such as few- and zero-shot, on the LLMs' ability to detect rumor content while also examining their sensitivity to hyperparameters such as "temperature."[2] Additionally, we explore the effects of modifying prompting instructions.
3) Assessing whether incorporating sentiment and emotions in the input data enhances or rather hampers the LLMs' capacity to identify rumor content.
4) Introducing a novel perspective on misinformation detection by deliberately corrupting authentic news using ChatGPT, employing zero-shot and iterative-style-based prompts, and quantifying parameters such as concreteness, abstractness, and the density of named entities.
5) Further enriching our analysis by intentionally distorting human-generated fake news using LLMs, particularly GPT, and conducting a parallel parameter assessment. This dual approach enhances the robustness of our framework, offering insights into the distinctive features that differentiate real news, human-generated misinformation, and LLM-induced distortions.

Based on the above contributions, we structure our article into two parts: the first part focuses on misinformation detection, evaluating the effectiveness of diverse language models (LLMs) in discerning rumor content and exploring the impact of various prompting techniques. It also assesses how incorporating sentiment and emotions in input data influences LLMs' ability to identify rumors. The second part examines the authenticity of news, intentionally corrupting authentic and human-generated fake news using LLMs. This dual-part approach provides a comprehensive understanding of information integrity, distinguishing real news from human-generated misinformation and LLM-induced distortions. The code and dataset are available through a public GitHub repo: https://github.com/Raghvendra-14/TAI-MISINFORMATION.

[1]We consider "fake news" and "rumor news" as equally harmful misinformation.

[2]The "temperature" parameter in a language model such as GPT controls the level of randomness in the generated text.

II. Related Studies

Exploring Misinformation: This review spans diverse aspects of fake news and rumor detection.

Machine Learning, Deep Learning, and Graph-Based Techniques: The study by Zhou and Zafarani [5] underscored the impact of fake news on democracy, focusing on automated detection methods. Likewise, Shu et al. [6] and Zhang and Ghorbani [7] conducted a comprehensive survey encompassing characteristics, algorithms, metrics, and datasets in the realm of fake news. Research by Zubiaga et al. [8] concentrated on rumor classification, encompassing detection, tracking, stance, and veracity, evaluating current methodologies. Bondielli and Marcelloni [9] analyzed automatic detection methods, addressing definitions, data challenges, and techniques. Shu et al. [10] in their work introduced the TriFN framework for precise fake news classification. The study by Zhou et al. [11] focused on early rumor detection through reinforcement learning. Sicilia et al. [12] introduced an innovative system for detecting social media rumors on Twitter, with a specific focus on healthcare using graph-based techniques. In contrast, Zubiaga et al. [13] employed context extraction from tweets, rather than direct processing, to enhance classification.

LLM-Based Techniques: Based on comprehensive evaluations of LLMs in tasks such as reasoning and hallucination, Shah et al. [14] illustrated their inherent link to misinformation. In a similar fashion, Bang et al. [15] proposed a framework to quantitatively assess ChatGPT across 23 datasets covering eight NLP tasks, revealing ChatGPT's superior performance over other LLMs in most tasks and its capability in generating multimodal content from textual prompts. However, ChatGPT shows an average accuracy of 63.41% in reasoning tasks, underscoring its limitations in logical, nontextual, and commonsense reasoning, alongside issues of hallucination shared with other LLMs. Additionally, Zhang et al. [16] investigated the widespread misuse of misinformation on social media. Their study highlighted LLMs such as ChatGPT as facilitators of deceptive content creation and interactive social bots, posing challenges to platform detection systems. The authors introduced advanced machine-learning strategies for detecting social manipulators, modeling misinformation causality, and identifying LLM-generated misinformation, suggesting significant implications for future research directions.

Examining LLMs for Misinformation Detection and Propagation: Research by Hu et al. [17] highlighted the effectiveness of LLMs such as GPT-3.5 while acknowledging BERT's superior performance for fake news detection. Similarly, the study in [18] focused on fine-tuning the Llama-2 LLM for tasks ranging from disinformation analysis to fake news detection. Abburi

et al. [19] addressed the detection of AI-generated text and its attribution to specific language models. Additionally, Sun et al. [20] introduced Med-MMHL, a comprehensive dataset for medical misinformation detection covering various diseases and including both human-generated and LLM-generated content. Pan et al. [3] explored the potential misuse of LLMs in generating misinformation, revealing an 87% degradation in open-domain question answering (ODQA) system performance. They proposed three defence strategies to mitigate these risks, which are the detection of false information, careful prompting, and collective reader scrutiny.

Lucas et al. [21] introduced a "fighting fire with fire" (F3) strategy to combat potential LLM misuse. Their approach utilized GPT-3.5-turbo to generate both genuine and deceptive content via paraphrasing and perturbation-based prompts. Zhang and Gao [22] explored leveraging LLMs with in-context learning to verify news claims. Similarly, Chen and Shu [23] investigated the capability of LLMs such as ChatGPT to generate misinformation, highlighting concerns about online safety and public trust. Jiang et al. [24] assessed current detection techniques' effectiveness and explored LLMs' potential role in defence strategies. Pelrine et al. [25] utilized GPT-4 to enhance the evaluation of information veracity amid the challenge of misinformation. Their research demonstrated GPT-4's superior performance compared to previous methods across multiple languages, introducing the "LIAR-New" dataset for context-based veracity assessment. Similarly, Pan et al. [26] introduced program-guided fact-checking (ProgramFC), a model leveraging LLMs to break down complex claims into manageable subtasks, employing specialized functions for verification. In their survey, Chen and Shu [27] systematically reviewed pre-LLM strategies for combating misinformation, discussing current initiatives and outlining future directions to effectively utilize LLMs and address LLM-generated misinformation, emphasizing interdisciplinary collaboration. Choi and Ferrara [28] presented FACT-GPT, an automated system for claim matching using LLMs that generate labeled datasets from simulated social media posts, fine-tune specialized models, and achieve competitive performance in fact-checking tasks complementing human expertise.

Sentiment and Emotional Analysis for Misinformation Detection: Research by Alonso et al. [29] showcased a comprehensive overview of sentiment analysis in fake news detection, emphasizing critical factors and future needs. Iwendi et al. [30] applied information fusion techniques to address COVID-19-related fake news. The study by Kula et al. [31] utilized deep learning for fake news detection, while Ajao et al. [32] explored sentiments' role in social network-based fake news detection. Bakir and McStay [33] scrutinized the 2016 U.S. presidential election campaign, proposing solutions and introducing the concept of "empathic media." Additionally, Mackey et al. [34] introduced a BERT-based model that enhances fake news classification using emotional cues. Furthermore, Samuel et al. [35], in their study, analyzed nearly 70 000 AI-related news headlines using NLP, ML, and LLMs to uncover dominant themes and sentiments. Their research exposes pervasive negativity and fear-evoking language in AI news, illustrating its influence on public perception and policy.

LLMs and Linguistic Semantics: Rawte et al. [36] investigated LLM hallucination and its underlying causes, aiming to uncover the relationship between linguistic factors and hallucination occurrences. Liu et al. [37] explored code-generating language models, specifically addressing the challenge of abstraction matching in data analysis contexts. Wang et al. [38] addressed the performance disparities between LLMs and supervised baselines in named entity recognition (NER) tasks. Grosse et al. [39] studied LLM generalization patterns, including sparsity, abstraction at scale, linguistic and programming capabilities, cross-lingual generalization, and role-playing behavior. Furthermore, Liu et al. [40] introduced a fake news propagation simulation (FPS) framework based on LLMs to study the dynamics of fake news spread. Their research identified patterns influenced by topic relevance and individual traits, providing insights into effective intervention strategies.

Novelty of Our Work: The cited research offers key insights into misinformation detection, covering fake news, rumors, and LLMs in conjunction with sentiments, emotions, and linguistics. We present an approach that integrates all these critical factors. In our initial research on misinformation detection, we introduce the task of rumor classification using LLMs in both zero and few-shot settings—a previously unexplored endeavor. We then deepen this investigation by incorporating sentiments and emotions to recognize their potential impact. In the second part of our analysis, we shift focus to the landscape of LLMs and textual characteristics. The existing work has primarily centered on hallucinations, neglecting linguistic aspects such as text concreteness, readability, and named entity density. Our research addresses this gap by exploring such crucial aspects.

III. LLMs for Detecting Misinformation

A. Datasets

We have utilized six datasets for misinformation detection. The *PHEME* dataset [41], released in 2016, captures Twitter conversations during five distinct breaking news events labeled to distinguish rumors from nonrumors. The FakeNewsNet dataset [42], a combination of *GossipCop*[3] and *PolitiFact*[4] datasets, encompasses news articles along with details of relevant social contexts. The *Snopes* dataset [43] was created in 2020 from the Snopes website, a widely recognized fact-checking platform that includes diverse textual claims, each paired with veracity labels (true, false, part true, part false, and mixed) and contextual information such as origin date and source. The Indian fake news dataset (IFND) [44], introduced in 2021, is a comprehensive resource primarily focused on events spanning from 2013 to 2021. Additionally, the researchers used an augmentation algorithm to enhance the fake news section of the IFND, creating authentic-seeming fake news statements to boost its reliability. The *ESOC COVID-19 misinformation* dataset [45] contains samples of misinformation across social media and news platforms, providing detailed information, including sources, keywords, and direct links. Table I presents

[3]https://www.gossipcop.com/
[4]https://www.politifact.com/

TABLE I
ATTRIBUTES OF THE EXPERIMENTALLY EMPLOYED DATASETS

Dataset	Domain	Application	# Labels	Modality	Media Platform
PHEME	Crisis Events	Rumor Detection	2	Text	Twitter
POLITIFACT	Political News	Rumor Detection	3	Text	Twitter
GOSSIPCOP	Entertainment Media	Rumor Detection	3	Text	Twitter
SNOPES	Mainstream News	Fact Checking	5	Text, Images	Snopes, Twitter
IFND	Mainstream News	Fake News Detection	2	Text, Images	Multiple News Websites
ESOC COVID-19	COVID	Veracity Classification	3	Text	Social Media, News Outlets

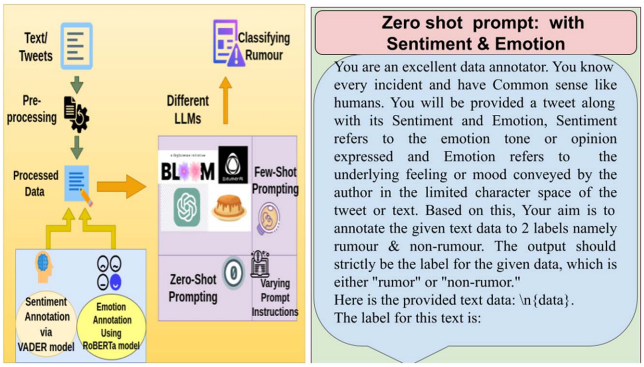

Fig. 1. Our rumor detection framework with sentiments and emotions, and examples of zero-shot prompting with and the consideration of sentiment and emotions.

an overview of the datasets utilized in our experimental study, highlighting key characteristics such as domain, application, number of labels, modality, and the primary media platform.

B. Experimental Setup

Here, we outline the LLM models used in our analysis. First, *GPT-3.5*, introduced by OpenAI in June 2020, is a paid service model with a decoder-only transformer architecture trained on over 500 billion tokens. *BLOOM*: a part of the BigScience project, is a freely accessible LLM published in 2021. It features a decoder-only transformer model and is trained on around 366 billion tokens. Then, *FLAN-T5*, an open-source instructional model, FLAN-T5 is an adapted variant of the T5 model. It has an encoder–decoder architecture and has been trained on a corpus of 1.5 trillion tokens. Last, *GPT-Neo*, released by EleutherAI, is a large-scale language model meant to replicate the performance of models such as GPT-3 but without the associated cost. It was introduced in 2021 and is available for free. GPT-Neo utilizes a similar architecture to GPT-3, employing a decoder-only transformer. Considering resource limitations and the objective to evaluate a spectrum of LLMs ranging from high to low parameter counts, we opted for the following specific models: GPT-3.5 (gpt-3.5-turbo), GPT-Neo (EleutherAI/gpt-neo-1.3B), BLOOM (bigscience/bloomz-560m), and FLAN-T5 (google/flan-t5-small).

1) Our Framework for Sentiment and Emotion Analysis: Our systematic approach, illustrated in Fig. 1, begins with preprocessing 500 randomly selected rumor texts/tweets from each dataset, with the exception of the PolitiFact dataset, which contained 432 false samples. This preprocessing includes the removal of URLs, emoticons, and hashtags. The next steps involve sentiment and emotion annotation using VADER [46] for sentiment analysis and DistilRoBERTa-base [47] for emotion detection. To ensure accuracy, we manually verified the annotations of 50 texts/tweets from each dataset, confirming satisfactory results.

2) Rationale for Choosing VADER: VADER is specifically designed for microbloglike content, making it well suited for our datasets' unique linguistic characteristics.

3) Justification for Sentiments Used: Our study employs three sentiment classes—positive, negative, and neutral—to analyze misinformation detection using LLMs. This selection captures a range of emotional influences that may accompany misinformation. Positive sentiment may indicate biased or overly positive framing, negative sentiment might highlight critical or alarming content often associated with misinformation, and neutral sentiment helps to identify factual or unbiased information. Integrating sentiment analysis enhances the accuracy and robustness of detection methods. This approach aligns with established research practices in sentiment analysis for microblogs [48], customer responses [49], and financial stock sentiment analysis [50], contributing to the comprehensive framework we aim to develop.

4) Rationale for Choosing DistilRoBERTa-Base: DistilRoBERTa-base is extensively trained on diverse English text sources, making it compatible with our datasets. Moreover, recent works [51], [52], using DistilRoBERTa for emotional analysis, further validate our choice. Last, given our computing resource constraints, VADER and DistilRoBERTa-base were practical choices.

5) Justification for Emotions Used: Our study uses a range of emotions—anger, disgust, fear, joy, neutral, sadness, and surprise—to analyze the emotional impact in misinformation detection using LLMs. Each emotion provides valuable insights: anger and disgust may signal provocative content, fear and surprise can highlight alarming elements, joy may indicate biased framing, sadness can reflect concerning narratives, and neutrality helps to identify factual information. This multidimensional approach enhances our understanding of misinformation sources. Additionally, it is important to note that except for neutrality, all the emotions we have selected are part of Paul Ekman's basic emotions [53], which are widely recognized and utilized in the field of natural language processing (NLP). Ekman's basic emotions provide a robust framework for analyzing and categorizing emotional responses, making them highly relevant and effective for our study. By incorporating these well-established emotional categories, our analysis gains depth and aligns with established research practices in NLP, as demonstrated in [54], [55], and [56].

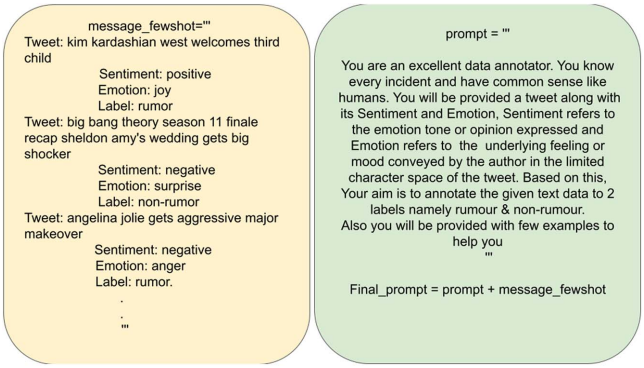

Fig. 2. Few-shot prompt example along with sentiment and emotions.

In Fig. 1, a crucial aspect comes to light: in the end part of our prompt, we direct the LLM to generate a precise output, categorizing it as either "rumor" or "nonrumor." The label extraction process is carefully crafted to trim unnecessary details. This involves parsing the output response by splitting it at the colon symbol, capturing the initial line with the predicted label, and removing any trailing spaces or words. Subsequently, this label is directly stored in a Python data frame. To enhance transparency, the extracted labels are further cross verified for the presence of "rumor" or "nonrumor" words using regex.

6) Experiments Without Sentiments and Emotions: In our first experiment setting, the preprocessed texts/tweets, without sentiment and emotion annotations, are used as input to several LLMs—GPT-3.5 [57], BLOOM [58], FLAN-T5 [59], and GPT-Neo [60]—for classification into rumor or nonrumor. Different prompt methods, including zero- and few-shot, are explored to assess the LLMs' performance under varied conditions.

7) Experiments With Sentiments and Emotions: In our next experiment setting, the same texts/tweets undergo classification using the same LLMs but with the inclusion of sentiments and emotions. This comparative analysis helps to evaluate the impact of sentiment and emotion annotations on the rumor detection capabilities of the LLMs. The emotions incorporated into our experiments spanned a range, including anger, disgust, fear, joy, neutrality, sadness, and surprise.

8) Prompting Methods and Settings: We have used zero- and few-shot (with 19 examples of tweets with correct labels in the few-shot) prompting techniques. Additionally, we have employed diverse configurations to evaluate the performance of different LLMs. This includes testing with varying temperature parameters (the default value used by us is 0.7 for all LLMs) and randomly annotating sentiments and emotions to examine their impact more deeply. Our code implementation leveraged fundamental Python libraries, including Pandas, Numpy, and CSV handling. The models were executed on two types of GPUs: the NVIDIA Tesla V100-PCIE-32GB and the GeForce GTX 1080-Ti-11GB, both operating on Unix/Linux systems.

9) Few-Shot Prompt: We illustrate the specific few-shot prompt designed for the GossipCop dataset in Fig. 2, which encompasses tweets related to entertainment media gossip. Additionally, we showcase how sentiments and emotions were incorporated into the prompt. We provided a total of 19 examples in a few-shot setting, aiming to encompass a comprehensive range of permutations and combinations involving sentiments, emotions, and labels. The prompt was crafted to optimize LLMs' categorization of tweets. By specifying attributes such as comprehensive knowledge and humanlike common sense for annotators, the prompt seeks to provide LLMs with a contextually rich understanding aligned with real-world tweet complexities. Instructing LLMs to consider both sentiment and emotional information aims to guide them in capturing the emotional tone and underlying sentiment, with the aim to enhance the accuracy of classifications.

10) Zero-Shot Prompt: The zero-shot prompt is shown as follows. *You are an excellent data annotator. You know every incident and have common sense like humans. Your aim is to annotate the given text data to 2 labels, namely rumor and nonrumor. The output should strictly be the label for the given data, which is either "rumor" or "nonrumor." Here is the provided text data:* $\{data\}$. *The label for this text is as follows.*

It is strategically crafted to assist LLM in detecting rumors. The prompt provides a clear directive, establishing the LLM's role as an annotator and emphasizing its humanlike common sense. This simplicity guides the model in binary text data annotation, promoting generalization across diverse textual instances. The straightforward design enhances the LLM's adaptability and effectiveness in making informed predictions, minimizing reliance on explicit sentiment and emotion cues.

C. Results and Discussions

1) Zero Shot Versus Few Shot and Inclusion Versus Exclusion of Sentiments and Emotions: Table II highlights[5] multiple key observations from this study as follows.

a) The counter-intuitive results, with few-shot settings displaying poorer performance than zero-shot, could stem from the challenges inherent in few-shot learning, where the model makes predictions based on limited examples, making it more susceptible to noise and variations. On the other hand, zero-shot learning leverages preexisting knowledge without explicit examples, potentially benefiting from a broader understanding of the underlying patterns or may result from encountering labeled data during training. This study exemplifies that few-shot learning does not always outperform zero-shot, shedding light on the complex dynamics between prompting and learning paradigms.

b) The inclusion of sentiments and emotions (W-SE) leads to a notable accuracy drop, with the model interpreting them apparently as noise. Additionally, this incorporation does not contribute to performance enhancement compared to their exclusion (WO-SE), which is particularly evident in zero-shot scenarios. Therefore, a crucial takeaway emerges: refraining from including sentiments and emotions is advisable when using LLMs for misinformation detection.

[5]The bold values in the table indicate the best performance in the dataset-wise arrangement.

TABLE II
ACCURACY VALUES WITH DIFFERENT DATASETS AND LLMS ARE PRESENTED FOR BOTH FEW-SHOT AND ZERO-SHOT SETTINGS, WITH ''W-SE'' AND WITHOUT ''WO-SE'' INDICATING THE INCLUSION OR LACK OF SENTIMENTS AND EMOTIONS

Few-Shot								
LLM Used →	*GPT-3.5*		*BLOOM*		*FLAN-T5*		*GPT-Neo*	
DATASET ↓	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*
PHEME	0.640	**0.680**	0.014	0.008	0.592	0.002	0.556	0.016
POLITIFACT	0.326	0.286	**0.980**	0.022	0.748	0.024	0.040	0.018
GOSSIPCOP	**0.430**	0.318	0.162	0.014	0.236	0.034	0.046	0.010
SNOPES	0.274	0.176	0.076	0.004	**0.368**	0.264	0.034	0.012
IFND	**0.356**	0.344	0.098	0.010	0.068	0.022	0.024	0.010
ESOC COVID-19	**0.262**	0.086	0.034	0.022	0.176	0.060	0.026	0.014
Zero-Shot								
LLM Used →	*GPT-3.5*		*BLOOM*		*FLAN-T5*		*GPT-Neo*	
DATASET↓	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*	*WO-SE*	*W-SE*
PHEME	0.638	0.630	0.842	0.028	0.412	0.168	**0.862**	0.828
POLITIFACT	0.626	0.494	**0.974**	0.036	0.526	0.136	0.804	0.768
GOSSIPCOP	**0.506**	0.296	0.494	0.048	0.454	0.086	0.050	0.034
SNOPES	0.098	0.274	0.578	0.032	**0.612**	0.236	0.076	0.042
IFND	0.562	0.374	**0.964**	0.044	0.888	0.090	0.068	0.024
ESOC COVID-19	**0.630**	0.130	0.594	0.032	0.140	0.186	0.044	0.036

Note: The bold values indicates the best performance in the dataset-wise arrangement.

TABLE III
ACCURACY RESULTS FOR VARYING TEMPERATURE HYPERPARAMETER

LLM Used → GPT-3.5	**Temperature = 0.2**		**Temperature = 0.7**		**Temperature = 1.0**	
DATASET ↓	**Zero-Shot**	**Few-Shot**	**Zero-Shot**	**Few-Shot**	**Zero-Shot**	**Few-Shot**
PHEME (WO-SE)	0.476	0.082	0.638	**0.640**	0.430	0.126
PHEME (W-SE)	0.376	0.048	0.630	**0.680**	0.404	0.094
POLITIFACT (WO-SE)	0.194	0.422	**0.626**	0.326	0.212	0.444
POLITIFACT (W-SE)	0.170	0.264	**0.494**	0.286	0.200	0.308
GOSSIPCOP (WO-SE)	0.472	0.363	**0.506**	0.430	0.468	0.414
GOSSIPCOP (W-SE)	0.362	0.188	0.296	0.318	**0.378**	0.252

Note: The bold values indicates the highest accuracy achieved across the three datasets for different temperature hyperparameters.

c) Model performance varies across settings, with GPT-3.5 demonstrating promising results in few-shot scenarios, while GPT-Neo lags. In zero-shot instances, GPT-Neo emerges as the top performer for the PHEME dataset. Additionally, FLAN-T5 excels with the highest accuracy for SNOPES in few-shot settings, and BLOOM achieved 90+% accuracy in three combinations.

2) Varying Temperature Values: Table III showcases the observations with different temperature settings. The key takeaway from this analysis is that varying the temperature setting in the experiment often influences accuracy scores. Higher temperatures correlate with increased accuracy, as demonstrated across different datasets. However, this sensitivity to the temperature hyperparameter suggests that finding an optimal balance is crucial, as excessively high temperatures may lead to a decline in accuracy. *Additionally, a fundamental question arises: why not employ a temperature value close to 0 for a more predictable classifier? Yet, when experimenting with values such as 0.01 and 0.15, we encountered a significant issue. The output responses were blank labels, meaning that the input prompt was echoed as the output for over half of our prompts, all having blank labels. This led us to initiate our exploration with a temperature value of 0.2.*

3) Significance of Labels: Finally, bold values in Table IV highlights the fact that in both the zero- and few-shot scenarios, correctly labeled sentiments and emotions (C-SE) improve the accuracy when compared to the cases with their values being set randomly (R-SE) indicating their continued relevance, though their utility might be somewhat limited in the broader context.

TABLE IV
ACCURACY RESULTS FOR CORRECTLY AND RANDOMLY LABELED SENTIMENTS AND EMOTIONS (C-SE AND R-SE, RESPECTIVELY)

LLM Used → GPT-3.5	**Zero-Shot**		**Few-Shot**	
DATASET ↓	**C-SE**	**R-SE**	**C-SE**	**R-SE**
PHEME	0.630	0.540	**0.680**	0.306
POLITIFACT	**0.494**	0.336	0.286	0.120
GOSSIPCOP	0.296	0.282	0.318	**0.360**

Building on these established results, the second part of our analysis investigates specific selected textual and linguistic traits. In particular, we try to support the distinction of a given text as being authentic, human-written fake content, or the one processed by LLMs. By probing these dimensions, we aim to deepen our understanding of the complex interplay between linguistic features and the trustworthiness of textual content.

IV. ANALYSIS OF LLM-PRODUCED MISINFORMATION

A. Datasets

For this study, we utilized the SNOPES dataset, randomly selecting a set of 500 aauthentic news articles alongside 500 human-written false articles. The SNOPES dataset offers

various examples of misinformation, covering various topics and formats. Additionally, its longitudinal nature allows us to track the evolution of misinformation trends over time. Furthermore, as a reputable fact-checking organization, SNOPES ensures the accuracy and reliability of the information within the dataset. While datasets such as PHEME and ESOC COVID-19 comprised tweets/concise textual snippets and were consequently excluded from our analysis, Gossipcop and IFND exhibited less diversity than SNOPES. It is essential to note that this does not diminish the value of these datasets; however, for our specific task, SNOPES emerged as the most suitable and comprehensive choice.

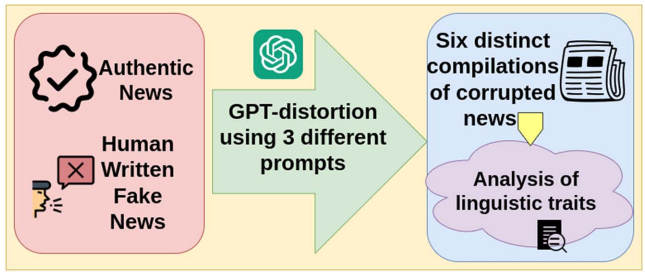

Fig. 3. Our framework for inspecting abstractness, concreteness, readability, and named entity density across authentic news, human-crafted fake news, and ChatGPT-distorted news.

B. Experimental Setup

Given the variable lengths of articles in the SNOPES dataset, our experiments were conducted in two settings. In the first setting, all linguistic experiments were performed on the overall length of the articles. In the second setting, we standardized the length of all articles to 200 words. If a sentence was truncated in the process, it was entirely removed. Therefore, in the normalized scenario, the maximum length is constrained to 200 words. Our overall framework depicted in Fig. 3 looks into the notions of abstractness, concreteness, readability, and named entity density, navigating through authentic news, human-crafted fake news, and the distorted versions of both by ChatGPT.

Abstractness and Concreteness Assessment: Abstractness refers to the degree to which a concept, idea, or piece of information is generalized, theoretical, or detached from specific, tangible details. Concreteness, on the other hand, pertains to the degree of specificity, tangibility, and focus on observable facts or events in language. We can say that the relationship between abstractness and concreteness is often considered inversely proportional. For example, if we consider the statement—"love is a powerful emotion." In this statement, the term "love" is presented in a generalized manner, lacking specific details or tangible elements. In contrast, consider the following concrete statement—"The gentle embrace and supportive actions of a friend during tough times embody the true essence of love." Here, the concept of love is illustrated with specific, tangible details, making it more concrete and vivid. For calculating the abstractness and concreteness, our methodology is inspired by the works of [61] and [62] and consists first of identifying verbs, adjectives, and adverbs—representative of abstract words by using the part-of-speech tagging module.[6] Then, the abstractness score of content is calculated as the ratio of abstract words to the total number of words in the text. Finally, the concreteness score is derived by subtracting the abstractness score from "1."

Readability Assessment: The readability of a text refers to its ease of comprehension for the intended audience. It involves the clarity, simplicity, and coherence of the writing, ensuring that readers can readily understand the content. For example, the following piece of text can be considered to have high readability: "The straightforward recipe with simple instructions allowed even novice cooks to prepare a delicious meal effortlessly," whereas "The intricacies of the culinary instructions, laden with complex terminology and convoluted steps, posed a challenge for those with limited cooking experience." can be considered as having low readability. In our assessment of article readability, we utilize two key metrics: the Flesch reading ease [64] and the Coleman-Liau index [65]. Both metrics share the common objective of offering easily interpretable readability scores that reflect the educational level required for comprehension.

Named Entity Density (NED) Estimation: NED evaluation involves assessing the proportion of named entities within a given text, providing insights into the specificity degree and the prominence of identifiable entities. Named entities can include names of people, organizations, locations, dates, and other unique identifiers. A higher NED indicates a text rich in specific details, while a lower ratio may suggest more general or abstract content. For NED inspection, we again leveraged the Spacy library to compute the NED, employing diverse entity types for analysis. The entities considered encompassed a wide spectrum, including individuals (PERSON), organizations (ORG), locations (GPE, LOC), dates (DATE), and various other categories such as "WORK_OF_ART," "MONEY," and "QUANTITY."

1) Dual Approach Analysis:

a) Distorting authentic news by ChatGPT: In adopting a dual approach for robust analysis, the intentional distortion of authentic news articles using ChatGPT serves as a controlled experiment to analyze the model's generative capabilities. By subjecting genuine news content to manipulation, we gain insights into how the language model synthesizes information and how much it deviates from the original text. This process allows us to assess the model's susceptibility to introducing misinformation and the potential impact on linguistic features, providing a valuable baseline for comparison.

b) Corrupting fake news by ChatGPT: While distorting authentic news is an intuitive step, deliberately corrupting already fake news articles introduces a deeper layer to our investigation. This approach explores the model's ability to manipulate further the misinformation that has already undergone human-generated distortion. By subjecting preexisting fake news to ChatGPT, we aim to understand the interplay between human-generated misinformation and AI-induced alterations. This unique perspective should offer insights into the model's adaptability and its potential to amplify or modify existing false narratives. Furthermore, it provides a deeper understanding of how the model interacts with and modifies

[6]We used the Spacy library [63].

Fig. 4. Two zero-shot prompts and the Iterative-style prompt used by us for distorting news.

TABLE V
COMPUTED ''ABS'' (ABSTRACTNESS SCORE) AND ''NED''' (NAMED ENTITY DENSITY) VALUES FOR THE OVERALL DISTORTED AND NORMALIZED NEWS SETS, ALONG WITH THE ORIGINAL AUTHENTIC AND FAKE NEWS

Category	Overall		Normalized	
Metrics	Abs	NED	Abs	NED
Auth	0.563	0.**101**	0.437	**0.100**
+ Dv-ZSP1	0.412	0.060	0.411	0.060
+ Dv-ZSP2	0.411	0.067	0.411	0.067
+ Dv-ItSP	0.404	0.048	0.397	0.057
Fake	**0.743**	0.098	**0.746**	0.097
+ Dv-ZSP1	0.401	0.079	0.401	0.078
+ Dv-ZSP2	0.414	0.082	0.414	0.081
+ Dv-ItSP	0.392	0.057	0.384	0.070

Note: (Concreteness score) = 1 - Abs. The bold values signify highest abstractness and NED.

content that already contains misinformation, shedding light on the potential harmonies or conflicts between human-generated and AI-induced distortions.

2) Deliberate Distortion Techniques:

a) Zero-shot prompt: We chose the zero-shot prompt illustrated in Fig. 4. The initial prompt explicitly instructs the language model to fabricate information in a genuine news article. However, the first prompt yielded outputs that, in addition to modifying the content in the original input news article, add multiple new sentences to the news articles. To rectify this, we experimented with a second-zero-shot prompt, explicitly directing the model to refrain from adding any extra sentences.

Prompt 1: The initial prompt employs a straightforward and concise approach, informing the model about the experimental nature of the task and instructing it to fake information in a given genuine news article. However, the simplicity of this prompt led to unintended consequences, such as adding extra sentences to the original content.

Prompt 2: Recognizing the challenges faced with the first prompt, the second prompt followed a more detailed and explicit set of guidelines. By explicitly specifying the role of an "information corrupting agent" and providing strict instructions not to add any extra sentences to the original information, this prompt aims to mitigate the issue observed in the initial results. The motivation for the inclusion of clear directives was to enhance the model's understanding of the desired outcome, fostering a more controlled and targeted distortion of information.

b) Iterative-style-based prompt: Additionally, for a more in-depth examination of how ChatGPT corrupts the original news article, we devised an iterative-style prompt. Initially, we presented the article to ChatGPT, instructing it to distort the news iteratively. Following the first response, we fed it back as input to ChatGPT for further distortion. This iterative process was repeated for three cycles on each news article, allowing us to observe the gradual evolution of the distortion. We selected a cycle iteration of three, as beyond this point, we observed a diminishing quality in the distorted outputs. It became apparent that prolonged iterations led to an increase in noise and a decline in the fidelity of the generated content. This decision aimed to strike a balance between obtaining meaningful distortions and avoiding excessive noise accumulation during the iterative process.

c) Detailed description of iterative-style-based prompt: The initial prompt serves as the cornerstone for an iterative distortion experiment. It outlines the research context, indicating that the task involves distorting information in the news article iteratively. The provided news article is the starting point for the iterative transformation process. Once ChatGPT generates a response based on this prompt, the subsequent prompt builds upon the model's output from the previous iteration. This iterative cycle continues, with each new prompt being informed by the generated response in a continuous loop. This approach tracks the evolving nature of distortion over successive interactions, shaping the gradual transformation of the original news article through each iterative step.

C. Results and Discussions

1) Abstractness, Concreteness, and NED: In Table V, we examine the computed *Abstractness* score (Abs), *Concreteness* score (Conc), which is 1—Abs, and *NED* values for various news sets, encompassing both overall and normalized scenarios, alongside original authentic and fake news. Also, in Tables V and VI, "Auth" represents "authentic/true news." "Dv-ZSP1" stands for "distorted via zero-shot prompt-1," "Dv-ZSP2" means "distorted via zero-shot prompt-2," and "Dv-ItSP" stands for "distorted via iterative-style prompt." Last, "Fake" denotes "fake/false news." All of the prompting were performed using GPT-3.5.

We make the following observations from Table V.

a) The scores reveal interesting patterns in text characteristics. In authentic news, a balanced interplay between *abstractness* and *concreteness* is evident. On the contrary, fake news exhibits a distinctive pattern, with higher *abstractness* and lower *concreteness*. Consequently, a noteworthy takeaway emerges: when encountering a news article with a high *abstractness* (low *concreteness*), there is a higher likelihood that it is a product of human-generated misinformation.

b) Additionally, our observation indicates that LLM distortion tends to elevate *concreteness* while diminishing

TABLE VI
FLESCH READABILITY EASE SCORE (FR-SC) AND COLEMAN-LIAU READABILITY SCORE (CL-SC) FOR BOTH THE OVERALL DISTORTED AND NORMALIZED NEWS SETS, AS WELL AS THE ORIGINAL, AUTHENTIC, AND FAKE NEWS

Category →	Overall		Normalized	
Metrics→	FR-Sc	CL-Sc	FR-Sc	CL-Sc
Auth	34.49	13.67	34.48	13.67
+ Dv-ZSP1	53.11	13.71	53.08	13.71
+ Dv-ZSP2	46.81	13.84	46.81	13.84
+ Dv-ItSP	87.50	12.75	82.36	12.69
Fake	**89.40**	**10.56**	**83.47**	**10.54**
+ Dv-ZSP1	53.05	13.78	53.04	13.78
+ Dv-ZSP2	53.04	14.19	53.04	14.19
+ Dv-ItSP	86.50	11.43	82.69	11.42

Note: High Coleman-Liau index: complex text. Low scores: simpler content. High Flesch reading ease: easily understandable. Low scores: complex text.[7]

abstractness. This insight leads us to a key takeaway: if a news article exhibits *concreteness* close to 0.6 and *abstractness* close to 0.4, it likely underwent LLM-based distortion. The presence of a low NED score further substantiates this inference.

c) Corruption through LLM prompts unveils a unique pattern marked by a moderate level of *abstractness*, higher *concreteness*, and a low NED, evident in both authentic and fake news scenarios. This distinct pattern serves as a suggestion of an LLM-induced corruption when compared to the original article. Access to authentic articles is pivotal, as observations on the corrupted text with the original further bolster the identification of LLM influence in the distortion process. When a reference to the original article is absent, a concreteness score exceeding 0.6 and a low NED can still aid in recognizing the LLM's involvement.

2) Readability: The following key takeaways can be obtained from Table VI.

a) In the general context, authentic news exhibits a balanced readability, reflected in moderate FR-Sc and CL-Sc values. Conversely, fake news consistently demonstrates higher FR-Sc and lower CL-Sc values, indicating easier readability than real news. These high-readability scores for fake news suggest that these articles may employ language and writing styles that appeal to a wider audience, potentially contributing to their spread and impact.

b) The "Dv-ItSP" scenario, based on the iterative-style prompt, results in higher FR-Sc and lower CL-Sc values in both overall and normalized settings. This suggests that the iterative-style prompt might contribute to making the text more readable. Consequently, if a user has access to the original article and observes a substantial increase in readability in the distorted article, there is a higher likelihood that it has undergone distortion.

c) The "Dv-ZSP1" and "Dv-ZSP2" scenarios, involving zero-shot prompts, result in higher FR-Sc values when compared to authentic news, implying improved readability, although the impact on CL-Sc is relatively modest.

[7]The bold values in FR-Sc, signifying the highest values, and CL-Sc, representing the lowest values, are indicative of enhanced readability.

V. CONCLUSION AND LIMITATIONS

We repeat the main findings and observations from our study as follows. The *zero-shot settings outperform few-shot settings*. The *inclusion of sentiments and emotions negatively impacts the accuracy*, suggesting a need to refrain from their incorporation for improved misinformation detection when using LLMs. Model-specific variations highlight *GPT-3.5's strength in few-shot scenarios*, *GPT-Neo's excellence in zero-shot instances*, and *FLAN-T5's notable accuracy with the PHEME dataset in few-shot settings*. Furthermore, our subsequent text analysis reveals that *high abstractness in news articles may indicate human-generated misinformation*, while *LLM-induced distortions are characterized by increased concreteness and low named entity density*. Recognizing these patterns should increase the accuracy of misinformation detection models. However, we must recognize inherent limitations in our study, specifically regarding the analyzed datasets and scenarios, potentially constraining the generalizability of our findings to broader contexts and diverse datasets.

APPENDIX

Traditional Approaches for Rumor Detection

Traditional rumor classification methods, ranging from statistics to machine learning, shaped early misinformation detection. We compare these machine-learning approaches to have a comprehensive analysis. Also, to ensure methodological consistency across datasets, we used BERT [66] for textual embeddings; we trained with 1000 positive and 1000 negative samples per dataset (except Politifact), maintaining a balanced 75:25 train-test split. Our diverse set of classifiers enhances analysis and model robustness by capturing different data aspects. We implemented this approach using Python libraries (scikit-learn, pandas, and numpy) with default scikit-learn values for consistency.

Discussion: In our analysis (Table VII), the gradient boosting classifier proved optimal for datasets including Pheme, PolitiFact, Snopes, and Esoc Covid-19, demonstrating robust predictive capabilities across diverse information sources. The random forest model excelled with the IFND dataset, particularly in handling news focused on celebrities and politicians. The neural net model showed superior performance in classifying content within the GossipCop dataset, highlighting its effectiveness with gossip-oriented sources. Despite slightly trailing LLMs, GPT-Neo achieved exceptional accuracy in zero-shot settings on the PHEME dataset, excluding sentiment and emotional cues from its analysis. Similarly, GPT-3.5 performed well on GossipCop and Esoc Covid-19 datasets without specific training for the task, showcasing its adaptability across different domains. Overall, LLMs outperformed traditional techniques in three datasets, underscoring their potential for advancing misinformation detection.

TABLE VII
RESULTS FROM DIFFERENT MODELS ACROSS DATASETS

Approaches ↓ and Datasets →	***Pheme***	***Politifact***	***GossipCop***	***Snopes***	***IFND***	***Esoc Covid-19***
Logistic Regression	0.280	0.146	0.172	0.198	0.224	0.256
Rbf SVM	0.362	0.302	0.328	0.354	0.382	0.406
Decision Tree	0.514	0.458	0.484	0.514	0.536	0.562
Neural Net	0.446	0.414	**0.542**	0.566	0.492	0.518
Nearest Neighbors	0.232	0.158	0.284	0.212	0.236	0.162
QDA	0.444	0.314	0.348	0.366	0.392	0.418
Random Forest	0.522	0.472	0.496	0.522	**0.548**	0.574
Gradient Boosting	**0.650**	**0.526**	0.520	**0.578**	0.504	**0.618**
Naive Bayes	0.312	0.174	0.222	0.220	0.252	0.278
AdaBoost	0.388	0.432	0.356	0.382	0.408	0.434

Note: Bold values indicate top-performing results.

Readability Score Metrics

We employed the Flesch reading ease [64] and the Coleman-Liau index [65], with their mathematical formulations provided as follows.

Flesch Reading Ease:

$$\text{Flesch Reading Ease} = 206.835 - 1.015 \times \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \times \left(\frac{\text{total syllables}}{\text{total words}}\right). \quad (1)$$

Coleman-Liau Index:

$$\text{Coleman-Liau Index} = 0.0588 \times \text{L} - 0.296 \times \text{S} - 15.8. \quad (2)$$

In these formulas, "total words" refers to the number of words in the text, "total sentences" refers to the number of sentences, "total syllables" refers to the number of syllables, "L" is the average number of letters per 100 words, and "S" is the average number of sentences per 100 words.

We benefited from using the Flesch reading ease and Coleman-Liau index to assess article readability. Flesch measures overall comprehension ease, considering sentence length and syllable count, while Coleman-Liau focuses on average letters and sentences per 100 words, providing complementary insights.

REFERENCES

[1] Z. Epstein, A. A. Arechar, and D. Rand, "What label should be applied to content produced by generative AI?," Jul. 2023.

[2] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative language models and automated influence operations: Emerging threats potential mitigations," 2023, *arXiv:2301.04246*.

[3] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Wang, "On the risk of misinformation pollution with large language models," in *Findings Association for Computational. Linguistics: EMNLP*, H. Bouamor et al., Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1389–1403. Accessed: Jun. 7, 2024. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.97

[4] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2023, pp. 1–20.

[5] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Comput. Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.

[7] X. Zhang and A. A. Ghorbani, "An overview of online fake news: Characterization, detection, and discussion," *Inf. Process. Manage.*, vol. 57, no. 2, 2020, Art. no. 102025.

[8] A. Zubiaga, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media," 2016, *arXiv:1610.07363*.

[9] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Inf. Sci.*, vol. 497, pp. 38–55, Sep. 2019.

[10] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proc. 12th ACM Int. Conf. Web Search Data Mining*, 2019, pp. 312–320.

[11] K. Zhou, C. Shu, B. Li, and J. H. Lau, "Early rumour detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.*, (Vol. 1: Long Short Papers), 2019, pp. 1614–1623.

[12] R. Sicilia, S. Lo Giudice, Y. Pei, M. Pechenizkiy, and P. Soda, "Twitter rumour detection in the health domain," *Expert Syst. Appl.*, vol. 110, pp. 33–40, Nov. 2018.

[13] A. Zubiaga, M. Liakata, and R. Procter, "Exploiting context for rumour detection in social media," in *Social Inform.: 9th Int. Conf. (SocInfo)*, Oxford, U.K., 2017, pp. 109–123.

[14] S. B. Shah et al., "Navigating the web of disinformation and misinformation: Large language models as double-edged swords," *IEEE Access*, early access, May 29, 2024, doi: 10.1109/ACCESS.2024.3406644.

[15] Y. Bang et al., "A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity," in *Proc. 13th Int. Joint Conf. Natural Lang. Process. 3rd Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics* (Vol. 1: Long Papers), J. C. Park et al. Eds., Nusa Dua, Bali: Association for Computational Linguistics, Nov. 2023, pp. 675–718. Accessed: Jun. 7, 2024. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.45

[16] Y. Zhang, K. Sharma, L. Du, and Y. Liu, "Toward mitigating misinformation and social media manipulation in LLM era," in *Proc. Companion ACM Web Conf. (WWW)*, New York, NY, USA: ACM, 2024, pp. 1302–1305, doi: 10.1145/3589335.3641256.

[17] B. Hu et al., "Bad actor, good advisor: Exploring the role of large language models in fake news detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 20, Mar. 2024, pp. 22105–22113. Accessed: Jun. 7, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/30214

[18] B. M. Pavlyshenko, "Analysis of disinformation and fake news detection using fine-tuned large language model," 2023, *arXiv:2309.04704*.

[19] H. Abburi, M. Suesserman, N. Pudota, B. Veeramani, E. Bowen, and S. Bhattacharya, "Generative AI text classification using ensemble LLM approaches," 2023, *arXiv:2309.07755*.

[20] Y. Sun, J. He, S. Lei, L. Cui, and C.-T. Lu, "Med-MMHL: A multimodal dataset for detecting human- and LLM-generated misinformation in the medical domain," 2023, *arXiv:2306.08871*.

[21] J. Lucas, A. Uchendu, M. Yamashita, J. Lee, S. Rohatgi, and D. Lee, "Fighting fire with fire: The dual role of LLMs in crafting and detecting elusive disinformation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14279–14305. [Online]. Available: https://aclanthology.org/2023.emnlp-main.883

[22] X. Zhang and W. Gao, "Towards LLM-based fact verification on news claims with a hierarchical step-by-step prompting method," in *Proc. 13th Int. Joint Conf. Natural Lang. Process. 3rd Conf. Asia-Pacific Chapter Assoc. Comput. Linguistics* (Vol. 1: Long Papers), J. C. Park et al., Eds., Nusa Dua, Bali: Association for Computational Linguistics, Nov.

2023, pp. 996–1011. Accessed: Jun. 7, 2024. [Online]. Available: https://aclanthology.org/2023.ijcnlp-main.64

[23] C. Chen and K. Shu, "Can LLM-generated misinformation be detected?" in *Proc. 12th Int. Conf. Learn. Representations*, 2023.

[24] B. Jiang, Z. Tan, A. Nirmal, and H. Liu, "Disinformation detection: An evolving challenge in the age of LLMs," in *Proc. SIAM Int. Conf. Data Mining (SDM),* Philadelphia, PA, USA: SIAM, 2024, pp. 427–435.

[25] K. Pelrine et al., "Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6399–6429. Accessed: Jun. 7, 2024. [Online]. Available: https://aclanthology.org/2023.emnlp-main.395

[26] L. Pan et al., "Fact-checking complex claims with program-guided reasoning," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics* (Vol. 1: Long Papers), A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds., Toronto, ON, Canada: Association for Computational Linguistics, Jul. 2023, pp. 6981–7004. Accessed: Jun. 8, 2024. [Online]. Available: https://aclanthology.org/2023.acl-long.386

[27] C. Chen and K. Shu, "Combating misinformation in the age of LLMs: Opportunities and challenges," 2023, *arXiv:2311.05656*.

[28] E. C. Choi and E. Ferrara, "Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation," in *Proc. Companion Proc. ACM Web Conf.*, New York, NY, USA: ACM, 2024, pp. 1441–1449, doi: 10.1145/3589335.3651910.

[29] M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares, "Sentiment analysis for fake news detection," *Electronics*, vol. 10, no. 11, 2021, Art. no. 1348.

[30] C. Iwendi, S. Mohan, E. Ibeke, A. Ahmadian, and T. Ciano, "COVID-19 fake news sentiment analysis," *Comput. Elect. Eng.*, vol. 101, 2022, Art. no. 107967.

[31] S. Kula, M. ChoraŚ, R. Kozik, P. Ksieniewicz, and M. Woźniak, "Sentiment analysis for fake news detection by means of neural networks," in *Proc. Comput. Sci.–ICCS: 20th Int. Conf.*, Amsterdam, The Netherlands, Jun. 2020, Springer International Publishing, 2020, pp. 653–666.

[32] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment aware fake news detection on online social networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP),* Piscataway, NJ, USA: IEEE Press, 2019, pp. 2507–2511.

[33] V. Bakir and A. McStay, "Fake news and the economy of emotions: Problems, causes, solutions," *Digit. Journalism*, vol. 6, no. 2, pp. 154–175, 2018.

[34] A. Mackey, S. Gauch, and K. Labille, "Detecting fake news through emotion analysis," in *Proc. 13th Int. Conf. Inf., Process, Knowl. Manage.*, 2021, pp. 65–71.

[35] J. Samuel, T. Khanna, and S. Sundar, "Fear of artificial intelligence? NLP, ML and LLMs based discovery of AI-phobia and fear sentiment propagation by AI news," Apr. 2024.

[36] V. Rawte, P. Priya, S. Tonmoy, S. Zaman, A. Sheth, and A. Das, "Exploring the relationship between LLM hallucinations and prompt linguistic nuances: Readability, formality, and concreteness," 2023, *arXiv:2309.11064*.

[37] M. X. Liu et al., "'What it wants me to say': Bridging the abstraction gap between end-user programmers and code-generating large language models," in *Proc. CHI Conf. Human Factors Comput. Syst.*, 2023, pp. 1–31.

[38] S. Wang et al., "GPT-NER: Named entity recognition via large language models," 2023, *arXiv:2304.10428*.

[39] R. Grosse et al., "Studying large language model generalization with influence functions," 2023, *arXiv:2308.03296*.

[40] Y. Liu, X. Chen, X. Zhang, X. Gao, J. Zhang, and R. Yan, "From skepticism to acceptance: Simulating the attitude dynamics toward fake news," 2024, *arXiv:2403.09498*.

[41] A. Zubiaga, G. Wong Sak Hoi, M. Liakata, and R. Procter, "Learning reporting dynamics during breaking news for rumour detection in social media," 2016, *arXiv:1610.07363*.

[42] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big Data*, vol. 8, no. 3, pp. 171–188, 2020.

[43] N. Vo and K. Lee, "Where are the facts? Searching for fact-checked information to alleviate the spread of fake news," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds., Association for Computational Linguistics, Nov. 2020, pp. 7717–7731. [Online]. Available: https://aclanthology.org/2020.emnlp-main.621

[44] D. K. Sharma and S. Garg, "IFND: A benchmark dataset for fake news detection," *Complex Intell. Syst.*, vol. 9, no. 3, pp. 1–21, 2021.

[45] J. N. Shapiro, J. Oledan, and S. Siwakoti, "ESOC COVID-19 misinformation dataset," 2020. Accessed: Jun. 8, 2024. [Online]. Available: https://esoc.princeton.edu/publications/esoc-covid-19-misinformation-dataset

[46] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, no. 1, 2014, pp. 216–225.

[47] J. Hartmann, "Emotion English distilroberta-base," 2022. Accessed: Jun. 8, 2024. [Online]. Available: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

[48] S. Elbagir and J. Yang, "Twitter sentiment analysis using natural language toolkit and VADER sentiment," in *Proc. Int. Multiconference Eng. Comput. Sci.*, vol. 122, no. 16. sn, 2019.

[49] A. Borg and M. Boldt, "Using VADER sentiment and SVM for predicting customer response sentiment," *Expert Syst. Appl.*, vol. 162, 2020, Art. no. 113746.

[50] S. Sohangir, N. Petty, and D. Wang, "Financial sentiment lexicon analysis," in *Proc. IEEE 12th Int. Conf. Semantic Comput. (ICSC)*, 2018, pp. 286–289.

[51] D. F. Vamossy and R. Skog, "EmTract: Extracting emotions from social media," 2023, *arXiv:2112.03868*.

[52] P. Liu and S. Lv, "Chinese RoBERTa distillation for emotion classification," *Comput. J.*, vol. 66, no. 12, pp. 3107–3118, 2023.

[53] P. Ekman et al., "Basic emotions," in *Handbook of Cognition and Emotion*, vol. 98, nos. 45–60, 1999, p. 16.

[54] E. Batbaatar, M. Li, and K. H. Ryu, "Semantic-emotion neural network for emotion recognition from text," *IEEE Access*, vol. 7, pp. 111866–111878, 2019.

[55] W. Graterol, J. Diaz-Amado, Y. Cardinale, I. Dongo, E. Lopes-Silva, and C. Santos-Libarino, "Emotion detection for social robots based on NLP transformers and an emotion ontology," *Sensors*, vol. 21, no. 4, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1322

[56] A. R. Murthy and K. A. Kumar, "A review of different approaches for detecting emotion from text," in *Proc. IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1110, no. 1. IOP Publishing, 2021, p. 012009.

[57] T. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neural Inf. Process. Syst.,* vol. 33, 2020, pp. 1877–1901.

[58] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, and D. E. A. Hesslow, "BLOOM: A 176B-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[59] H. W. Chung et al., "Scaling instruction-finetuned language models," *J. Mach. Learn. Res.*, vol. 25, no. 70, pp. 1–53, 2024.

[60] S. Black, L. Gao, P. Wang, C. Leahy, and S. R. Biderman, "GPT-Neo: Large scale autoregressive language modeling with mesh-tensorflow," Mar. 2021.

[61] F. Strik Lievers, M. Bolognesi, and B. Winter, "The linguistic dimensions of concrete and abstract concepts: Lexical category, morphological structure, countability, and etymology," *Cognit. Linguistics*, vol. 32, no. 4, pp. 641–670, 2021.

[62] R. Flesch, "Measuring the level of abstraction," *J. Appl. Psychol.*, vol. 34, no. 6, 1950, Art. no. 384.

[63] I. Montani, M. Honnibal, A. Boyd, and S. Van Landeghem, "explosion/spaCy: v3.7.2: Fixes for APIs and requirements". Zenodo, Oct. 2023.

[64] R. Flesch, "A new readability yardstick," *J. Appl. Psychol.*, vol. 32, no. 3, 1948, Art. no. 221.

[65] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring," *J. Appl. Psychol.*, vol. 60, no. 2, 1975, Art. no. 283.

[66] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Lang. Technol.,* (Vol. 1: Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. Accessed: Jun. 8, 2024. [Online]. Available: https://aclanthology.org/N19-1423