Can LLMs identify the Disinformation they create? Exploring the Role of Large Language Models in Disinformation detection

Pavan Sanjay Nichani[1,†], Ayaan Ahmad Siddiqui[1,†], Sakshi Tiwarii[1,†], Ark Ikhu[1,†] and Marina Ernst[1,*,†]

[1]*Universität Koblenz, Universitätsstr. 1, 56070 Koblenz, Germany.*

Abstract

The rapid advancements in Artificial Intelligence, especially in the development of Large Language Models, have introduced a domain abundant with opportunities while presenting considerable challenges. Although LLMs are demonstrating impressive abilities in understanding and generating human-like text, they have also contributed to a rise in disinformation. An outstanding example of this influence is COVID-19 pandemic, when fake news and conspiracy theories impacted individuals and society. This paper aims to determine whether LLMs can effectively rephrase human-like text and, utilize their knowledge base in detecting rephrased disinformation, a subtle and increasingly prevalent form of manipulated content. Furthermore, this research investigates the criteria and reasoning behind how LLMs classify text as real or fake by LLMs, aiming to assess whether LLMs can be used to counter disinformation while examining the challenges associated with their use for this purpose.

Keywords

Disinformation, Disinformation detection, LLM as Detector, Gemini Llama, LIWC

1. Introduction

LLMs are evolving rapidly, bringing along transformation in a variety of fields. They have inspired a range of applications that not only help in understanding human behavior but also provide insights into LLMs themselves by drawing upon human societal contexts [1]. A particularly intriguing aspect of LLMs is its paradoxical role in both generating and detecting fake news. On the one hand, LLMs can produce convincingly realistic yet false content, making it more challenging to identify disinformation [2]. On the other hand, these models can be utilized to build powerful tools for detecting and combating the very misinformation they may help create [3]. The significance of detecting fake news became crucial during the COVID-19 pandemic, where misinformations posed significant risks to individuals and society.

This paper aims to explore two primary research questions:

- **RQ1:** To what extent can LLMs rephrase existing disinformation that closely resembles human writing?
- **RQ2:** Can LLMs identify disinformation, and if there is a difference between human-written and LLM-generated text when detecting?

The structure of this paper is as follows: Section 2 reviews related work on generative and detection capabilities of LLMs. Section 3 outlines the methodology used to address RQ1 and RQ2. Section 4

Disinformation, Misinformation and Learning in the Age of Generative AI: Joint Proceedings of the 1st International Workshop on Disinformation and Misinformation in the Age of Generative AI (DISMISS-FAKE'25) and the 4th International Workshop on Investigating Learning during Web Search (IWILDS'25) co-located with 18th International ACM WSDM Conference on Web Search and Data Mining (WSDM 2025)

*Corresponding author.

†These authors contributed equally.

pavannichani@uni-koblenz.de (P. S. Nichani); asiddiqui@uni-koblenz.de (A. A. Siddiqui); sakshi@uni-koblenz.de (S. Tiwarii); aikhu@uni-koblenz.de (A. Ikhu); marinaernst@uni-koblenz.de (M. Ernst)

https://www.uni-koblenz.de/de/informatik/west/team/doctoral-candidates/marina-ernst (M. Ernst)

0009-0001-7041-9419 (M. Ernst)

focuses on the results related to RQ1, providing insights into the ability of LLMs to generate human-like text. Section 5 discusses the findings for RQ2, highlighting the effectiveness of LLMs in detecting fake text and common types of errors observed. Finally, Section 6 offers a conclusion and suggests directions for future research.

2. Related Work

LLMs, being trained on vast and diverse data, possess advanced linguistic capabilities, making it difficult to distinguish between human-generated and LLM-generated content. A comprehensive survey on the challenges of differentiating between these two types of texts was presented in a 2020 paper, following the introduction of GPT-2 [4]. As LLMs evolve, the effectiveness of prompt engineering has been demonstrated in guiding models to produce more accurate responses [5]. However, studies suggest that the format and structure of input prompts significantly influence the performance and output quality of LLMs [6]. In order to analyze LLM-generated text from a linguistic perspective, techniques like Linguistic Inquiry and Word Count (LIWC) has been used to assess the presence of specific linguistic features within generated texts [7]. Cosine similarity can be applied to compare these linguistic features between human-written and LLM-generated content, helping to assess how closely the LLM-generated text mirrors human writing. Additionally, methods like Levenshtein distance have been utilized to filter out highly similar LLM texts before being passed to advanced attribution models such as BERT and Random Forest[8].

In the domain of disinformation detection, fine-tuned models based on BERT have shown success in identifying LLM-generated text [9]. Other approaches, such as hybrid CNN-RNN deep learning frameworks, have been explored, though the effectiveness of these methods is highly dependent on the dataset and the training process [10]. The ability of LLMs to generate and detect disinformation has reached a point where, with the right prompts, models can now be tasked with determining whether a piece of content was generated by AI or written by a human [11]. In particular, the rise of automated fake news generation has driven interest in developing models [12] or with help of prompts using LLMs [13] to detect such content.

In light of the research questions posed in this study, our work aims to build on these findings by: (1) examining the extent to which LLMs can generate human-like text and its impact on detection, and (2) evaluating the effectiveness of these models in detecting disinformation, particularly in the context of a COVID-19-related dataset.

3. Methodology

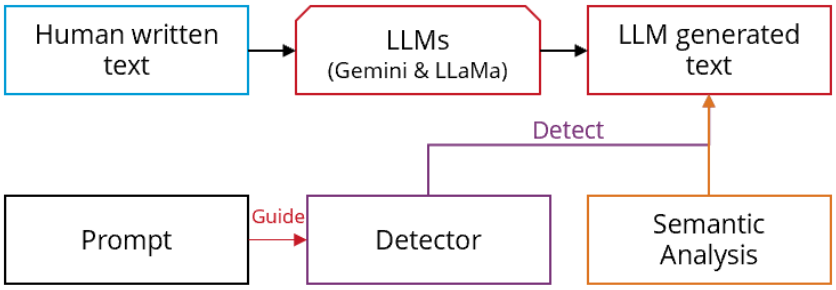


Figure 1: Overview of our methodology.

Figure 1 illustrates our approach, which begins with a human-written dataset. Using prompting techniques, we generated a rephrased LLM-generated database. Rephrasing was selected as the method for generating LLM text to reflect common strategies used by adversarial actors, who often tries to

modify and make existing misinformation to appear more factual to evade detection. This method ensures that the original disinformation persists within the newly generated content.

During the linguistic and syntactic analysis phase, we assess any potential similarities between human-written and LLM-generated data (RQ1). Finally, we employ LLMs as a detection method, using prompts to evaluate their ability to identify disinformation (RQ2).

We used two different LLMs, Gemini-1.5-flash and LLaMA-3 8b-8192, to leverage their unique strengths and ensure an unbiased evaluation of disinformation detection capabilities. This approach also allows us cross-validation of findings across different architectures, while including both commercial and opens-source solutions. Gemini-1.5-flash uses self-attention mechanism which enables it to process nuanced language and identify subtle indicators of disinformation by considering the full context of a passage. Its architecture is particularly suited for tasks requiring detailed contextual analysis [14]. LLaMA-3 8b-8192 generates text by conditioning on previous tokens, allowing it to maintain logical coherence across extended passages. This makes it highly effective at both producing human-like text and identifying inconsistencies indicative of disinformation [14].

3.1. Human-written Dataset

Our dataset comprises manually annotated dataset of 10,700 social media posts and articles related to real (5,600) and fake (5,100) COVID-19 news [15]. The dataset was pre-processed by removing hyperlinks, and posts with 12 words or fewer were excluded to avoid short tweets. This filtering resulted in a final dataset of 8,880 posts, comprising 5,275 real and 3,605 fake entries.

3.2. LLM-Generated Dataset

For LLM dataset generation, we employed Gemini-1.5-flash and LLaMA-3 8b-8192. To create diverse outputs, we used two rephrasing prompting strategies: paraphrase generation and rewriting generation [16]. Tables 1 and 2 show examples of the generated output, with **bold** words indicating the specific terms or phrases added by the LLMs during the rephrasing process.

Prompt 1: Paraphrase Generation	You are a journalist. Given a "passage", please paraphrase it. Do not generate any other word. The "passage" is <text>.
Original Text	The main mode of transmission of #Novel #Coronavirus is through droplets and it is possible that infected smokers may blow droplets carrying the virus when they exhale. Regardless of #COVID19 you should steer clear of second-hand smoke as it may cause various health problems.
gemini-1.5-flash	The novel coronavirus **spreads** primarily through droplets, **which** infected smokers may exhale. Secondhand smoke is **harmful** and **should be avoided.**
llama3-8b-8192	The **primary way** the **new** coronavirus spreads is through **tiny** droplets, and it's **conceivable** that smokers who are infected may release **virus-laden** droplets when exhaling. It's **not limited** to COVID-19; second-hand smoke can lead to various health issues, regardless of the virus.

Table 1
LLM-Generated data using prompt 1

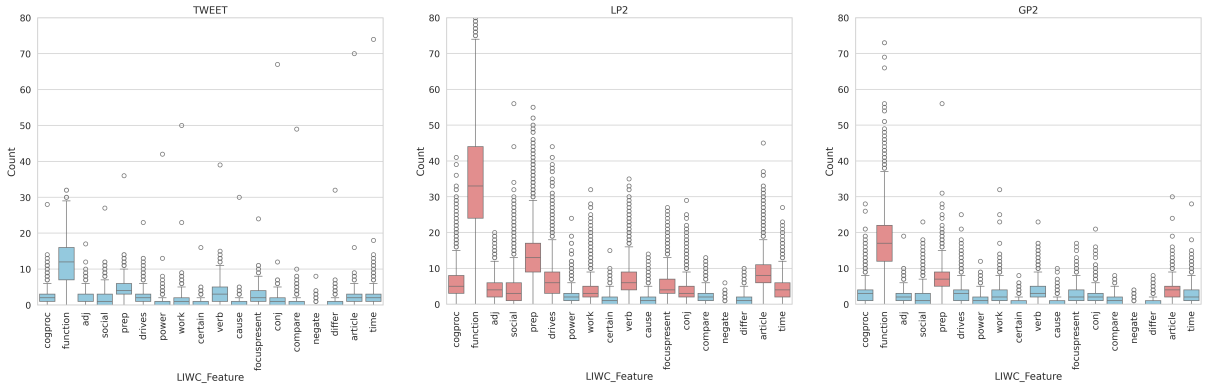
4. Generating disinformation with LLMs

To assess the similarity between human-written text and the text generated by LLMs, we conducted a linguistic analysis. This analysis aimed to determine whether our prompting methods can modify human-written content in a meaningful way and whether they may inadvertently introduce disinformation. We employed the Linguistic Inquiry and Word Count (LIWC) [17] for this purpose, leveraging its robust LIWC2015 Dictionary. As the focus of research is on rephrased disinformation, identifying subtle changes in linguistic features such as word choice, tone, syntax is important to evaluate the similarities

Prompt 2: Re-writing Generation	You are a journalist. Given a “passage”, please rewrite it to make it more convincing. The content should be the same. The style should be serious, calm and informative. The “passage” is: <text> .
Original Text	The main mode of transmission of #Novel #Coronavirus is through droplets and it is possible that infected smokers may blow droplets carrying the virus when they exhale. Regardless of #COVID19 you should steer clear of second-hand smoke as it may cause various health problems.
gemini-1.5-flash	The **primary route** of transmission for the novel coronavirus is through **respiratory** droplets. **Individuals infected** with COVID-19 may release these droplets, **potentially containing the virus**, when they exhale, **including when smoking. While the potential for transmission through secondhand smoke remains under investigation**, it is important to note that secondhand smoke is known to cause a variety of health issues. **As a precautionary measure, it is strongly advised to avoid exposure** to secondhand smoke **in all situations**.
llama3-8b-8192	The **primary means** of transmission for the novel coronavirus is via airborne droplets. **Furthermore, research suggests that** individuals infected with the virus may inadvertently spread it through the exhal**ation** of droplets when they breathe. **It is essential to note that this phenomenon is not exclusive to the current COVID-19 pandemic. Exposure** to second-hand smoke, **a common phenomenon in various settings, has been linked** to a range of serious health issues. **As a result, it is crucial to maintain a safe distance** from second-hand smoke, **regardless of the prevailing** health situation, **to minimize the risk of developing respiratory** problems and other illnesses.

Table 2
LLM-Generated data using prompt 2

between human-written and LLM generated text. The LIWC2015 Dictionary contains approximately 6,400 words, word stems, and select emoticons, categorized under linguistic or psychological sub-dictionaries. Words may belong to multiple categories. For example, the word ’cried’ falls into five categories: sadness, negative emotion, overall affect, verbs, and past focus. If this word appears in a text, it increases the count for each of these categories accordingly [18].

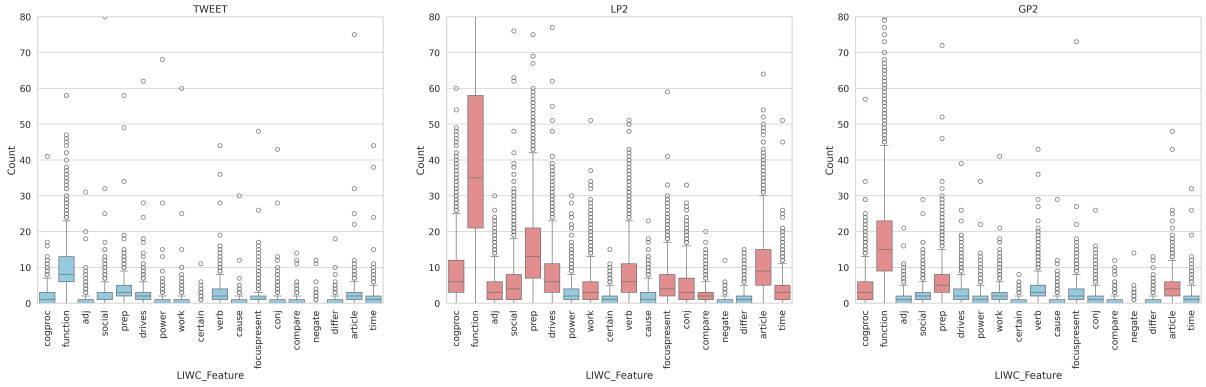


(a) Human-Written Text (b) Text Rephrased by Llama3 (c) Text Rephrased by Gemini

Figure 2: Comparison of LIWC Feature Distributions in Human-Written and LLM-Rephrased Texts for ’Real’ Label with prompt 2.

We analyzed the LIWC distributions across human-written and LLM-generated text. While LIWC features for the text generated by both LLMs with Prompt 1 are distributed very similar to the human-written ones, for Prompt 2 distributions defer significantly. Figures 2 and 3 illustrate these difference for both "Real" and "Fake" tweets respectively. The box plots depict the distribution of LIWC features in the original human-written texts and their corresponding rephrased versions generated by the LLMs.

Prompt 2, which rewrites the text to make it more persuasive with a more serious and informative



(a) Human-Written Text (b) Text Rephrased by Llama3 (c) Text Rephrased by Gemini

Figure 3: Comparison of LIWC Feature Distributions in Human-Written and LLM-Rephrased Texts for 'Fake' Label with prompt 2

tone, a number of LIWC categories demonstrate significant changes with notable shifts highlighted in red. Llama-generated texts generally differ more from human-generated ones. While occurrences of categories such as function words, prepositions and articles are easy to understand, others such as 'cognitive processes' (cogproc) and 'social processes' (social) require deeper analysis. It is worth noting that although Gemini stick more closely to human-written text, when it comes to fake tweets, the 'cognitive processes' category becomes more frequent.

To further quantify the similarity between human-written and LLM-generated text, the cosine similarity of their LIWC feature values was calculated. Among 8,880 tweets analyzed, 5,854 tweets (4,168 labeled "Real" and 1,686 "Fake"), had cosine similarity scores exceeding 0.8 across both LLMs and prompts. A high cosine similarity score suggests that the LLM-generated texts maintained similar feature distributions to the human-written texts, despite potential differences in the absolute frequencies of individual features. In essence, while the LLMs may have altered the intensity or frequency of certain linguistic elements, the overall pattern and structure of the feature distributions remained consistent with the human-written content. Despite high cosine similarity, manual verification of some LLM-generated text revealed factual inaccuracies. This indicates that while LLMs can produce text with linguistic patterns that closely mimic human writing, their rephrased content may still introduce factual errors, raising concerns about the risk of disinformation.

5. Detecting LLM-generated disinformation

In this section, we assess how effectively the models distinguish between real and fake information and analyze the misclassifications to identify potential patterns. Initially, the LLM-generated outputs for both prompts were manually reviewed, and a balanced subset of 800 tweets (400 real and 400 fake) was selected. This manual review ensured that no factual inaccuracies were introduced during paraphrasing process in sampled data.

To further utilize LLMs for disinformation detection, we conducted experiments with the same models using identical parameters to those used during the text generation phase. The following prompt was used for detection tasks:

Prompt: *Given a passage, determine whether it is a piece of disinformation. Only output 'YES' or 'NO'. The passage is: <text>*

Detection performance of bot models is shown in table 3 Gemini demonstrated better performance than Llama3 in identifying LLM-generated disinformation across both prompt types. As shown in Appendix A, Gemini initially exhibited balanced performance, achieving a True Positive Rate (TPR) of 65.5%and a True Negative Rate (TNR) of 90.6% for human-written texts. After rephrasing, Gemini improved in detecting real tweets, particularly with prompt GP1, where it reached a TPR of 81.8%, though its ability in detecting fake tweets slightly declined.

Table 3
Fake Tweets Detection Performance

Prompt	Gemini		Llama	
	Acc	F1	Acc	F1
Human-written	0.72	0.65	0.6	0.44
Gemini 1	0.66	0.49	0.67	0.32
Gemini 2	0.61	0.37	0.56	0.23
Llama 1	0.68	0.54	0.56	0.3
Llama 2	0.64	0.44	0.54	0.19

For Llama3, rephrasing improved real tweets detection, particularly with prompt GP2, achieving 88.4% TPR. However, its ability to detect fake tweets declined significantly, highlighting the model's sensitivity to rephrasing.

5.1. Error Analysis

To identify patterns in the model's misclassifications, we conducted a detailed error analysis.

Rephrasing Impact: Llama3 model exhibited a slightly higher number of changes when rephrasing was applied compared to the Gemini model (Table 14). This suggests that Llama3 may be more sensitive or responsive to rephrased text. Notably, the Llama3 was more successful at correcting errors through rephrasing than the Gemini model, as indicated by a higher number of corrected misclassifications. However, Llama3 also displayed a greater tendency to amplify errors under certain prompts compared to Gemini. This variation in model performance highlights the importance of prompt engineering in determining how rephrasing influences both error correction and exacerbation.

LIWC Analysis: We also explored whether linguistic features contribute to misclassification (refer Figure 4). High counts in categories such as "Article," "Pronoun," and "Drives" in TP suggest these are characteristic of genuine tweets correctly classified. In contrast, categories like "Relativ," "Prep," and "Verb" appeared less frequently in FPs than in TPs and TNs, suggesting that while present, these features are less prominent in genuine tweets misclassified as fake. Similarly, "Article," "Drives," and "Cogproc" (Cognitive Processes) were significant in FNs, similar to their presence in TPs. This points to the challenge of distinguishing fake content that closely mimics the structure of real tweets.

LLM Analysis: We further refined our analysis by focusing on tweets where both Gemini and Llama3, regardless of the prompt used, changed their classification from the gold label. This filtering resulted in a subset of 29 tweets. To understand the reasoning behind these classification changes, we employed the chain-of-thought prompting technique, instructing Gemini and Llama to explain their decision-making processes. The models were prompted to consider factors such as tone, context, wording, or implied meaning that influenced their updated classifications.

Prompt: *You are analyzing tweets to classify them based on their content. Below are the details of an original tweet and its rephrased version. Please read carefully and answer the following:*

Original Tweet: <text> Original Classification by you: <text>

Now, consider the rephrased version:

Rephrased Tweet: <text>

Rephrased Classification: <text>

Question: Explain why you changed your classification for the rephrased tweet compared to the original tweet. Discuss any differences in tone, context, wording, or implied meaning that led to the new classification.

Manual review of the models' responses, revealed that certain aspects of the rephrased tweets, such as inclusion of words like 'joke', 'disturbing' or mentions of news agencies, prompted the model to alter their classification. Similarly, provocative or emotionally charged language (e.g., exaggerated claims) often led to a change in labels. In contrast, rephrased tweets with neutral, factual tone or including credible source also triggered change in classification. For detailed overview, refer to Figure 5

in Appendix B.

6. Conclusion

In conclusion, LLMs such as Llama3 and Gemini can generate human-like text with high linguistic similarity to human-written content. While these models show potential for detecting disinformation, their susceptibility to rephrased content indicate the need for further development. Our experiments confirm LLMs' ability to create human-like text, though their performance in detecting disinformation is more complex. While Gemini showed greater accuracy in identifying real content, but both models displayed weaknesses, particularly when dealing with rephrased text. Factors such as prompt selection, text length, and model type can impact their effectiveness. The results from linguistic analysis using LIWC suggests that while LLMs can closely replicate the linguistic patterns of human writing, but they remain prone to errors, especially when subtle text modifications are introduced. This raises concerns about their reliability as standalone detectors, as adversarial actors could exploit these vulnerabilities by employing techniques like rephrasing.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, DeepL Write in order to: Grammar and spelling check, Paraphrase and reword, Improve writing style. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] Y. Leng, Y. Yuan, Do llm agents exhibit social behavior?, arXiv, 2024. URL: https://arxiv.org/abs/2312.15198.

[2] E. Papageorgiou, C. Chronis, I. Varlamis, Y. Himeur, A survey on the use of large language models (llms) in fake news, Future Internet 16 (2024). URL: https://www.mdpi.com/1999-5903/16/8/298. doi:10.3390/fi16080298.

[3] D. Sallami, From deception to detection: The dual roles of large language models in fake news, arXiv, 2024. URL: https://arxiv.org/abs/2409.17416.

[4] G. Jawahar, M. Abdul-Mageed, L. Lakshmanan, V.S., Automatic detection of machine generated text: A critical survey, in: D. Scott, N. Bel, C. Zong (Eds.), Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 2296–2309. URL: https://aclanthology.org/2020.coling-main.208/. doi:10.18653/v1/2020.coling-main.208.

[5] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Curran Associates Inc., Red Hook, NY, USA, 2022.

[6] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. C. Schmidt, A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023. URL: https://arxiv.org/abs/2302.11382. arXiv:2302.11382.

[7] M. Sandler, H. Choung, A. Ross, P. David, A linguistic comparison between human and chatgpt-generated conversations, 2024. URL: https://arxiv.org/abs/2401.16587. arXiv:2401.16587.

[8] K. Jones, J. R. Nurse, S. Li, Are you robert or roberta? deceiving online authorship attribution models using neural text generators, Proceedings of the International AAAI Conference on Web and Social Media 16 (2022) 429–440. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/19304. doi:10.1609/icwsm.v16i1.19304.

[9] M. Szczepański, M. Pawlicki, R. Kozik, M. Choraś, New explainability method for bert-based model in fake news detection, Scientific Reports 11 (2021). doi:10.1038/s41598-021-03100-6.
[10] J. A. Nasir, O. S. Khan, I. Varlamis, Fake news detection: A hybrid cnn-rnn based deep learning approach, International Journal of Information Management Data Insights 1 (2021) 100007. URL: https://www.sciencedirect.com/science/article/pii/S2667096820300070. doi:https://doi.org/10.1016/j.jjimei.2020.100007.
[11] A. Bhattacharjee, H. Liu, Fighting fire with fire: Can chatgpt detect ai-generated text?, SIGKDD Explor. Newsl. 25 (2024) 14–21. URL: https://doi.org/10.1145/3655103.3655106. doi:10.1145/3655103.3655106.
[12] I. Vykopal, M. Pikuliak, I. Srba, R. Moro, D. Macko, M. Bielikova, Disinformation capabilities of large language models, 2024, pp. 14830–14847. doi:10.18653/v1/2024.acl-long.793.
[13] B. Jiang, Z. Tan, A. Nirmal, H. Liu, Disinformation Detection: An Evolving Challenge in the Age of LLMs, 2024, pp. 427–435. doi:10.1137/1.9781611978032.50.
[14] S. H. Baskaran, A Comparison of Transformer and Autoregressive LLM Designs, International Journal of Research Publication and Reviews, 2023, pp. Vol 4, no 11, pp 19–26. URL: https://ijrpr.com/uploads/V4ISSUE11/IJRPR19003.pdf.
[15] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an Infodemic: COVID-19 Fake News Dataset, Springer International Publishing, 2021, p. 21–29. URL: http://dx.doi.org/10.1007/978-3-030-73696-5_3. doi:10.1007/978-3-030-73696-5_3.
[16] C. Chen, K. Shu, Can llm-generated misinformation be detected?, 2024. URL: https://arxiv.org/abs/2309.13788. arXiv:2309.13788.
[17] A. Koutsoumpis, J. K. Oostrom, D. Holtrop, W. van Breda, S. Ghassemi, R. E. de Vries, The Kernel of Truth in Text-Based Personality Assessment: A Meta-Analysis of the Relations Between the Big Five and the Linguistic Inquiry and Word Count (LIWC), American Psychological Association, 2022, pp. 148(11–12), 843–868. URL: https://psycnet.apa.org/fulltext/2023-55252-004.html. doi:https://doi.org/10.1037/bul0000381.
[18] J. W. Pennebaker, R. L. Boyd, K. Jordan, K. Blackburn, The development and psychometric properties of liwc2015, 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content.

A. LLM as Detector: Performance Analysis

This section presents the confusion matrix for the LLMs used as detectors. We provided the original human-written text to both Gemini and Llama, followed by the output generated from Prompt 1 by both models, and then the output from Prompt 2 by both models, effectively creating a cross-evaluation. The tables below illustrate the confusion matrix for each scenario.

Table 4
Gemini as detector for Original Tweets

	Pred. Real	**Pred. Fake**
Actual Real	378	21
Actual Fake	199	202

Table 5
Llama as detector for Original Tweets

	Pred. Real	**Pred. Fake**
Actual Real	350	49
Actual Fake	274	127

Table 6
Gemini as detector for Prompt 1 by Gemini

	Pred. Real	**Pred. Fake**
Actual Real	540	37
Actual Fake	120	103

Table 7
Gemini as detector for Prompt 2 by Gemini

	Pred. Real	**Pred. Fake**
Actual Real	563	14
Actual Fake	142	81

Table 8
Gemini as detector for Prompt 1 Tweets by Llama

	Pred. Real	Pred. Fake
Actual Real	543	34
Actual Fake	90	133

Table 9
Gemini as detector for Prompt 2 by Llama

	Pred. Real	Pred. Fake
Actual Real	551	26
Actual Fake	124	99

Table 10
Llama as detector for Prompt 1 Tweets by Gemini

	Pred. Real	Pred. Fake
Actual Real	583	41
Actual Fake	122	54

Table 11
Llama as detector for Prompt 2 by Gemini

	Pred. Real	Pred. Fake
Actual Real	602	22
Actual Fake	139	37

Table 12
Llama as detector for Prompt 1 by Llama

	Pred. Real	Pred. Fake
Actual Real	577	47
Actual Fake	126	50

Table 13
Llama as detector for Prompt 2 by Llama

	Pred. Real	Pred. Fake
Actual Real	599	25
Actual Fake	153	23

B. Error Analysis

The table below illustrates the impact of text rephrasing on classification by the LLMs.

Table 14
Rephrasing Impact

	gemini-1.5-flash GP1	gemini-1.5-flash GP2	llama3-8b-8192 LP1	llama3-8b-8192 LP2
Classification was changed after rephrasing	157	156	173	178
Misclassification was corrected after rephrasing	20	20	41	48
Misclassification worsened after rephrasing	32	12	32	20

True Positive		True Negative		False Positive		False Negative	
Category	Count	Category	Count	Category	Count	Category	Count
function	1818	function	5052	function	686	function	3615
prep	543	relativ	1807	relativ	248	prep	1087
relativ	401	prep	1453	prep	233	relativ	885
verb	277	verb	870	verb	130	verb	645
article	202	cogproc	426	cogproc	70	article	404
pronoun	138	space	409	drives	63	drives	338
drives	138	time	385	space	56	cogproc	276
cogproc	135	social	384	focuspresent	41	social	275
social	134	drives	350	article	34	space	223
space	80	focuspresent	344	social	32	pronoun	200

Figure 4: Error Analysis: Top 10 LIWC Categories for each classification

The figure 5 below outlines the reasons for classification changes, based on a manual review of each output generated using the chain-of-thoughts technique.

Original Tweet Characteristics	Rephrased Tweet Characteristics	Keywords/Reasoning for Change
Original tweet **lacks explicit indication** of falsity or context, possibly misleading.	Rephrased tweet **explicitly states the word "joke"** and provides more context, making the intention clear.	Explicit Statement and Added Context: **Clear labeling (e.g., "joke") and context remove ambiguity**, indicating no intent to misinform.
Uses **provocative or emotionally charged language (e.g., "shows"** or exaggerated claims).	**Adopts more neutral language, such as "has surfaced,"** and uses words like "disturbing" to maintain a factual tone.	Neutral and Factual Language: **Balanced presentation without sensational language** makes the tweet seem more credible.
Contains **unnecessary, misleading, or hyperbolic statements**.	Rephrased tweet **removes such statements and uses concise, direct language** while adding relevant context.	Removal of Misleading Statements: **Simplified language** focusing on factual content adds credibility.
Informal, generic, and biased tone with emotional language and agenda-driven phrases.	Professional, objective tone with **formal language, avoiding personal pronouns** or emotive words.	Professional and Objective Language: A formal tone without personal bias **enhances the tweet's trustworthiness**.
Overuse of emotional or exaggerated language without specific references.	Precise language, **inclusion of specific sources** like news agencies, and a neutral tone.	Addition of Specific References and Neutral Tone: Specific details and credible references **reduce perceived bias and add reliability**.
Explicit and direct claims without supporting evidence or sources.	Objective tone with **added context from credible news agencies**, avoids sensational language.	Objective Tone and Credible Source: Providing a credible source, such as a news agency, and **avoiding sensationalism shifts perception to "real."**

Figure 5: Error Analysis: Change in classification