Leveraging LLMs for Identifying types of Misinformation on Reddit

Bhavana Ramesh[1,†], Durwankur Gursale[1,†], Abram Jopaul[1,†] and Marina Ernst[1,*,†]

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

Abstract

The recent development of LLMs has demonstrated their ability to generate coherent, contextually relevant responses to a variety of tasks. However, they also pose significant risks, including disinformation, factual inaccuracies and the propagation of bias. This research aims to access the effect of prompting techniques on the detection of different types of misinformation using data collected from Reddit. Our findings suggest that the Few Shot prompting method performs well across different LLMs. However, the effect of positional bias found in our experiments indicates that prompt engineering needs to be further investigated for such a sensitive task.

Keywords

Misinformation, Fake news, Misinfrmation detection, LLMs, generative AI, Reddit

1. Introduction

The growing accessibility of LLMs and their enhanced capacity to produce credibly-sounding text also raise concerns regarding their potential misuse for generating misinformation [1]. These models are typically pre-trained on large datasets containing a combination of opinions, accurate information, outdated facts, sarcasm, and fake news.

Although LLMs are built to mimic human interactions, they lack the ability of understand the verification mechanism [2]. As a result, the models end up generating text, that contains misinformation, leading to false narratives. Ensuring factual accuracy is critical in sensitive areas such as healthcare and law, where mistakes can have serious consequences. Therefore, there is a need for post-generation verification mechanisms to mitigate the impact of large language models.

The use of LLMs to detect false data has been explored in a number of recent studies. [3, 4]. However, fewer attempts have been made to compare the predictions of SLMs with LLMs by fine-tuning an SLM model, such as Roberta, for different dataset sizes and highlighting the intrinsic characteristics of LLMs, such as position, attention, and label bias.

In this study, we aim to answer the following research questions:

- **RQ1:** How efficiently can LLMs and SLMs categorize misinformation, depending on the prompt techniques and patterns used?
- **RQ2:** Which types and categories of misinformation are more likely to be misclassified by LLMs?

The paper is structured as follows: Section 2 is dedicated to related work and the current landscape of misinformation detection, Section 3 introduces the methodology used and the data collection process. Section 4 presents the results and answers the research questions, and Section 5 concludes the paper and discusses the directions for future work.

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.

bhavanaramesh23@uni-koblenz.de (B. Ramesh); dcgursale@uni-koblenz.de (D. Gursale); bramjopaul@uni-koblenz.de (A. Jopaul); 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)

2. Related Works

When analyzing most available datasets, a common pattern emerges: they often use broad classifications such as "true" and "false". Although these datasets often contain millions of rows and are well-developed, they struggle to capture the nuanced nature of misinformation. However, some datasets, such as NELA-GT-2018 [5], LIAR [6], and FakeNewsCorpus [7], offer more detailed classifications. While some datasets, such as NELA-GT-2018 [5], consider for heterogeneous sources of information, others focus on specific areas, such as politics and gossip. Thus, these datasets provide limited context due to their restricted sources and classification types.

The ability of LLMs to generate both reliable information and disinformation has been studied extensively in recent years. Studies such as Jiang et al. [8] and Leite et al. [9] have investigated how the LLMs perform under different prompting techniques to classify disinformation. These studies are important for understanding the potential of LLMs to detect disinformation. However, these studies focused on the broad classification of disinformation as true or false. Furthermore, as mentioned above, these analyses used datasets that focused on a specific domain.

Positional bias in LLMs refers to the tendency to rely on the positions of the tokens in the input, which can adversely affect them, especially in the classification tasks [10], was a path breaking study that introduced the transformer architecture. Transformer based LLMs like GPT, Llama etc. use positional encoding which leads to the positional bias. Yu et al. [11] and Hsieh et al. [2] explained the tendency of LLMs to give priority to the tokens at the beginning and end of the input.

Even though LLMs are thought to be powerful, Hu et al. [12] reveals that they fail to outperform fine-tuned small language models(SLM), when appropriate prompting techniques are used. Rather than relying entirely on LLMs to detect disinformation, the authors propose a model in which LLMs act as advisors to the fine-tuned SLMs. Like the studies mentioned above, this analysis also classifies data into binary and focuses on GossipCop [13].

3. Methodology

3.1. Dataset

Most available datasets are built for standard binary classification rather than fine-grained classification, and our bespoke dataset focuses on distinguishing each new instance into 4 categories. We briefly describe each category:

- True: Indicates that the news item is true.
- Satire/Parody: Indicates if the content is twisted or misinterpreted in a satirical or humorous way.
- Misleading Content: Corresponds to the news where the information is intentionally manipulated to fool the audience.
- Imposter Content: Represents content generated by bots.

While Wardle et al. [14] outline additional disinformation types (e.g., Fabricated Content, False Connection), we excluded these as they rely on multimodal data. Our dataset was sourced from Reddit, a platform known for its threaded discussions and popularity, ranking among the top 20 websites globally. Using subreddits identified in Fakeddit [15], we obtained posts from July 2012 to December 2024. The collected data is published at Zenodo[1].

Data collection utilized the pullpush.io API, extracting submission titles, text, metadata (e.g., score, upvote ratio), and using tools such as Newspaper3k and BeautifulSoup4 for structured content scraping. To ensure quality, subreddit moderators enforced relevance to themes, performing an initial filter. We further refined the dataset by excluding posts with a score below 1, assuming that off-topic or inappropriate posts would either be deleted or downvoted. This multi-level processing ensures the credibility and relevance of our dataset. The final dataset contains 2343 posts with title, content, time,

[1]https://zenodo.org/records/14900167 DOI: 10.5281/zenodo.14900167

misinformation category, subreddit, score and upvote ratio. The data is fairly balanced in terms of misinformation category (Figure 1)

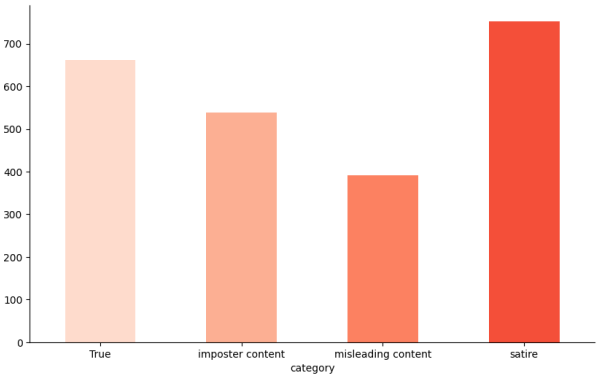

Figure 1: Comparison on 500 v/s 1000 dataset rows

3.2. Prompting

Zero-shot prompting is used to classify the given content into predefined categories without any prior training on specific examples. The model is provided with instructions to perform the classification task based solely on its existing knowledge. An example of prompt is shown in Figure 2

```
Content categories are as follows: True, Satire, Imposter Content, Misleading
Content
Analyze the type of content and return the corresponding label.
```

Figure 2: Zero-shot prompt

The chain-of-thought prompting was used to not only to classify the content, but also to explain the reasoning process, for a possible manual evaluation. The model is guided to systematically evaluate the given content by considering multiple factors before reaching a conclusion. An example of a CoT prompt is shown in Figure 3.

```
Analyze the type of the content enclosed in square brackets, and determine if it is
 true content, satire, misleading content, or imposter content. Explain your
reasoning step by step and then return the answer as the corresponding content
label "true content" or "satire" or "misleading content" or "imposter content".
   Step-by-step reasoning:
   1. Identify the primary purpose of the content (inform, entertain, deceive, etc
   .).
   2. Check for factual accuracy and sources.
   3. Determine if there is any exaggeration, humor, or irony.
   4. Check for any signs of manipulation or alteration.
   5. Determine if the content has been entirely fabricated.
```

Figure 3: CoT prompt

To address the challenge of positional bias, a binary classification was conducted, and each category was labelled separately. The example of the prompt for True information case is shown on Figure 4. Each class was isolated and classified individually to verify whether focusing on individual categories could yield better accuracy than the combined classification.This approach helps to determine whether the model favours certain types of content when labels are presented in different orders or contexts.

```
Given below the content category and definition:
 True: True information.
 Analyze the content within square brackets and determine if it is true or not.
 Return "True" if the content is true, otherwise return "None".
```

Figure 4: Binary classification prompt

To investigate positional bias in the Large Language Model (LLM), we explored all possible positional arrangements of the four content classes: True, Satire, Misleading Content, and Imposter Content. This involved changing the order of the classes in all 4! (24) possible ways.

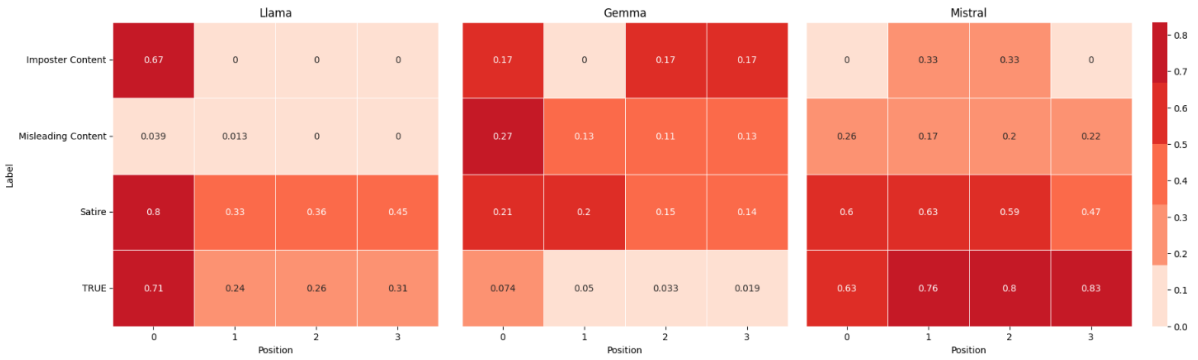

Figure 5: Heat Map - 24 Permutations of Llama, Gemma and Mistral

By classifying the content using these 24 different permutations, we aimed to determine if the LLM's performance varied based on the positional order of the categories, thereby revealing any potential biases in their responses.

3.3. SLM vs LLM

This approach evaluates the performance of SLMs and LLMs over similar dataset. RoBERTa was chosen as the SLM, and LlaMA (Llama-3-8B-Instruct), Gemma (Gemma-2-9b-it), and Mistral (Mistral-7B-Instruct-v0.3) were employed as the LLMs.

There are many reasons why RoBERTa was chosen over other SLMs. To name a few, RoBERTa has fewer restrictions on the data to be trained compared to the original BERT. RoBERTa is more focused on the goal of modelling disguised speech. Since RoBERTa uses a dynamic masking strategy, it is renowned for its robust and efficient generalization. Not to forget that RoBERTa is exceptionally flexible and can be fine-tuned to specific tasks.

Finally, after considering the pros and cons of using different models, we tested the dataset on both SLM and LLM models without any fine-tuning to witness their out-of-the-box capabilities. Later, fine-tuned RoBERTa was examined to see if larger datasets could improve its performance, and a comparison with LLMs was performed to confirm if it could outperform.

4. Results

In this section, we present the results obtained for each model. We report the recall, precision, and F1 scores obtained by all the models for different prompting techniques. Comparing the accuracy of various prompting techniques helps us find the best prompting approach. In addition, we are also interested in the variety of biases exhibited by LLMs.

Among several prompting strategies, Zero shot, Few Shot and Chain-of-Thoughts are selected to test their performance. For each prompt accuracy, recall, precision, and F1 scores were recorded across models. The results are presented in Table 1.

Table 1
Performance across models

Prompts	Llama				Gemma				Mistral			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Zero Shot	0.27	0.21	0.27	0.19	0.46	0.53	0.46	0.42	**0.4**	0.38	**0.4**	**0.37**
CoT	0.4	0.39	0.39	0.32	**0.71**	**0.71**	**0.71**	**0.71**	0.23	0.50	0.22	0.28
Few Shots	**0.53**	**0.68**	**0.54**	**0.54**	X	X	X	X	0.38	**0.51**	0.38	0.32

Llama achieves an accuracy of 53% with Few Shot, significantly outperforming the other two prompting techniques. Not only did Few Shot give the highest accuracy, but it also took all positions into account when labelling. Overall, this made it the most efficient prompting technique.

From an SLM perspective, the fine-tuned RoBERTa is run over different dataset sizes to record the efficiency of how well it can classify the labels compared to LLMs. The graph on Figure 6 shows the performance of RoBERTa with 500 and 1000 rows of data. It managed to achieve an accuracy of 23% with the 500 row dataset and 30% with the 1000 row dataset. Surprisingly, there is a spike in accuracy as the dataset size increases.

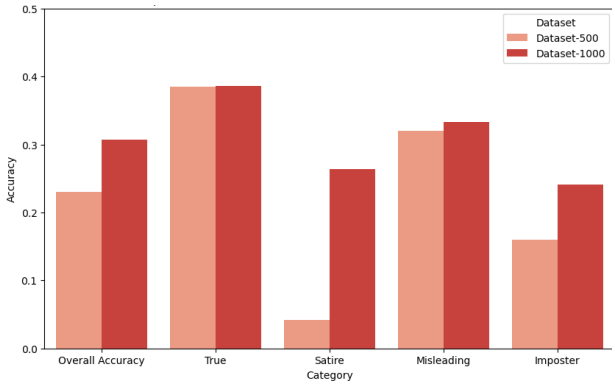

Figure 6: Comparison on 500 v/s 1000 dataset rows

4.1. Positional Bias

When experimenting with the Zero Shot prompting technique by placing labels at different positions, it is noticeable that not all positions receive a similar weight. This is due to the fact that LLMs are built using a sequence-based architecture, which assigns disproportionate weights to certain tokens based on their sequence in the input.

The heat maps shown above in Figure 5 demonstrate how the accuracy varies with respect to position when labels are called in sequence using the Zero Shot prompting technique. It is noteworthy that Mistral outperforms all other models by showing less positional bias.

To cross-validate the previous observation on positional bias, a binary classification was performed using the prompt shown in Figure 4 was performed. The results of the binary classification, presented in Table 2, show that the accuracy of all three models is generally higher when the prompt is focused on detecting only one type of content. However, in this case, Imposter content was not identified at all, which requires further investigation.

The observed pattern, known as attention bias, occurs when either the beginning, the end, or the middle of a sequence receives greater emphasis.The graph in Figure 7 demonstrates 2 shapes typical for attention bias. Llama and Gemma depicts a U-shaped curve indicating the importance given to start and end of the sequence, whereas Mistral presents an inverted U-shaped curve, indicating the attention given to the middle of the sequence. In the case of Llama and Gemma, the attention bias is called as

Table 2
Accuracy of Models on Binary Classification

Prompts	Llama	Gemma	Mistral
True	0.907	0.458	0.85
Satire	0.81	0.2	0.84
Misleading	0.957	0.28	0.72
Imposter	0	0	0

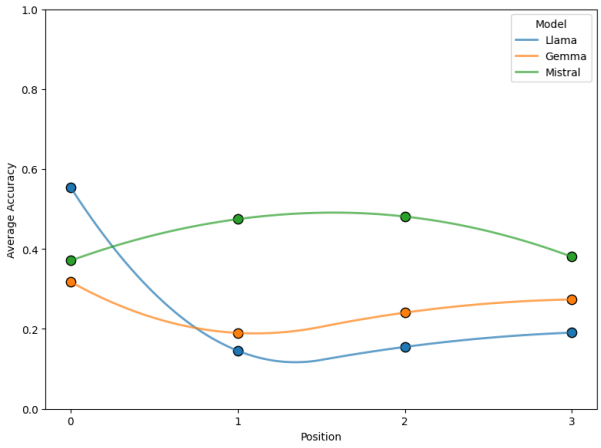

Figure 7: Attention Bias

U-shaped bias. On the other hand, the attention bias in Mistral is known as Found-in-the-middle.

4.2. Performance by category

Considering the F1-scores and accuracy of each label, Misleading content has an accuracy of 0.12 and an F1 score of 0.21 for Llama (Table 3), accuracy of 0.03 and an F1 score of 0.06 for Mistral (Table 4). Thus, by looking at the low F1 scores, we can say that the content does not cover all information needed by LLM to categorize it as Misleading, making it the most deceptive disinformation category in this setting.

Table 3
Llama - Few Shots

Prompts	Accuracy	Precision	Recall	F1-Score
True	0.389	0.67	0.39	0.49
Satire	0.769	0.69	0.77	0.73
Misleading	**0.121**	0.71	0.12	**0.21**
Imposter	0.879	0.68	0.88	0.76

Table 4
Mistral - Few Shots

Prompts	Accuracy	Precision	Recall	F1-Score
True	0.131	0.78	0.13	0.22
Satire	0.93	0.32	0.93	0.47
Misleading	**0.032**	0.25	0.03	**0.06**
Imposter	0.44	0.7	0.44	0.54

5. Conclusion

This research focuses on the classification of misinformation with LLMs and how prompting techniques can influence the model's decision making. The contribution of this study to the field starts with the novel dataset of textual misinformation. The results of the experiments suggest that the structure of the prompt has a significant impact on the performance of the models in detecting the type of misinformation. While a Few shot prompt yields the best results in the context of this study, it is evident that a correct prompt technique alone cannot ensure effective classification. Further investigation revealed the presence of the positional bias, which skews the classification and prevalence of some types over others. From the misinformation types perspective, Misleading and Imposter content appeared to be the most difficult to identify. These findings open up many possibilities for further work. The first possible direction is to run the experiment on a larger corpus of data and to include state-of-the-art solutions, both commercial and open source. Another approach is to further investigate different prompting methods.

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. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, W. Wang, On the risk of misinformation pollution with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 1389–1403. URL: https://aclanthology.org/2023.findings-emnlp.97/. doi:10.18653/v1/2023.findings-emnlp.97.

[2] C.-Y. Hsieh, Y.-S. Chuang, C.-L. Li, Z. Wang, L. Le, A. Kumar, J. Glass, A. Ratner, C.-Y. Lee, R. Krishna, T. Pfister, Found in the middle: Calibrating positional attention bias improves long context utilization, 2024, pp. 14982–14995. doi:10.18653/v1/2024.findings-acl.890.

[3] M. G. Buchholz, Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset, 2023. URL: https://arxiv.org/abs/2306.08190. arXiv:2306.08190.

[4] K. Pelrine, A. Imouza, C. Thibault, M. Reksoprodjo, C. Gupta, J. Christoph, J.-F. Godbout, R. Rabbany, Towards reliable misinformation mitigation: Generalization, uncertainty, and GPT-4, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 6399–6429. URL: https://aclanthology.org/2023.emnlp-main.395. doi:10.18653/v1/2023.emnlp-main.395.

[5] M. Gruppi, B. D. Horne, S. Adalı, Nela-gt-2020: A large multi-labelled news dataset for the study of misinformation in news articles, arXiv preprint arXiv:2102.04567 (2021).

[6] W. Y. Wang, "liar, liar pants on fire": A new benchmark dataset for fake news detection, in: R. Barzilay, M.-Y. Kan (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 422–426. URL: https://aclanthology.org/P17-2067/. doi:10.18653/v1/P17-2067.

[7] A. Pathak, R. Srihari, BREAKING! presenting fake news corpus for automated fact checking, in: F. Alva-Manchego, E. Choi, D. Khashabi (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for

Computational Linguistics, Florence, Italy, 2019, pp. 357–362. URL: https://aclanthology.org/P19-2050/. doi:10.18653/v1/P19-2050.

[8] B. Jiang, Z. Tan, A. Nirmal, H. Liu, Disinformation detection: An evolving challenge in the age of llms, in: Proceedings of the 2024 SIAM International Conference on Data Mining (SDM), SIAM, 2024, pp. 427–435.

[9] J. A. Leite, O. Razuvayevskaya, K. Bontcheva, C. Scarton, Weakly supervised veracity classification with llm-predicted credibility signals, 2024. URL: https://arxiv.org/abs/2309.07601. arXiv:2309.07601.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.

[11] Y. Yu, H. Jiang, X. Luo, Q. Wu, C.-Y. Lin, D. Li, Y. Yang, Y. Huang, L. Qiu, Mitigate position bias in large language models via scaling a single dimension (2024). doi:10.48550/arXiv.2406.02536.

[12] B. Hu, Q. Sheng, J. Cao, Y. Shi, Y. Li, D. Wang, P. Qi, Bad actor, good advisor: exploring the role of large language models in fake news detection, AAAI'24/IAAI'24/EAAI'24, AAAI Press, 2024. URL: https://doi.org/10.1609/aaai.v38i20.30214. doi:10.1609/aaai.v38i20.30214.

[13] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media, Big Data 8 (2020) 171–188. doi:10.1089/big.2020.0062.

[14] C. Wardle, H. Derakhshan, Information disorder: Toward an interdisciplinary framework for research and policymaking, volume 27, Council of Europe Strasbourg, 2017.

[15] K. Nakamura, S. Levy, W. Y. Wang, Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6149–6157. URL: https://aclanthology.org/2020.lrec-1.755/.