Supplementary Material

Anonymous Author(s)*

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**;
• **Information systems** → **Data mining**.

ACM Reference Format:
Anonymous Author(s). 2024. Supplementary Material. In *Proceedings of the ACM Web Conference 2024 (KDD '24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 SCINEWS DATASET ANALYSIS

We further analyzed the proposed SciNews Dataset:

- For human-written articles part: maximum number of sentences in an article is 557; minimum number of sentences is 6; The average number of sentences per article is 54.49. **The average number of words per sentence within all the news articles is 19.39.**
- For LLM-generated part: maximum number of sentences in an article is 35; minimum number of sentences is 1; The average number of sentences per article is 8.24; and **average number of words per sentence within all the news articles: 21.88.**

Then, we visualized the distribution of sentence length as well as the average number of sentences in the dataset.

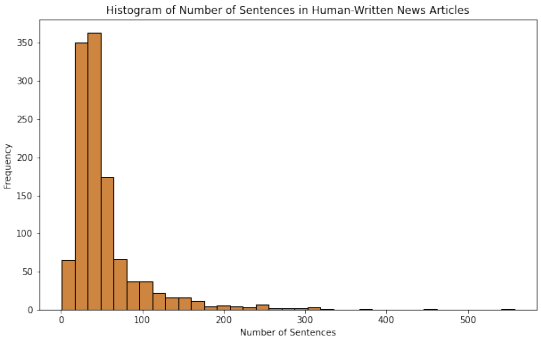

Figure 1: The number of sentences in Human-Written Articles.

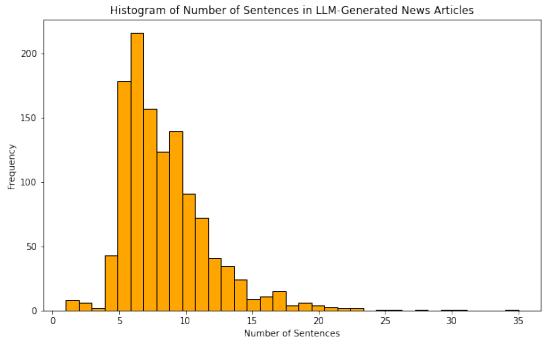

Figure 2: The number of sentences in LLM-Generated Articles.

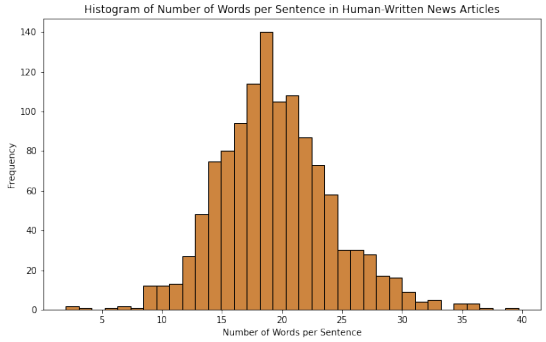

Figure 3: The number of words per sentence in Human-Written Articles.

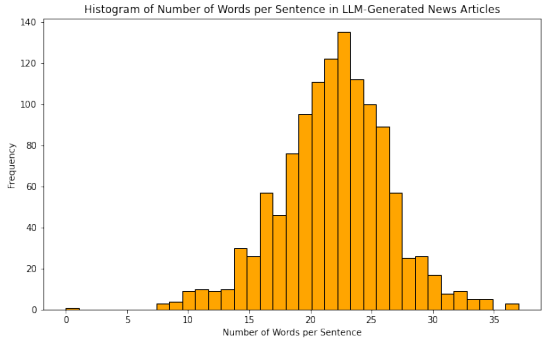

Figure 4: The number of words per sentence in LLM-Generated Articles.

KDD '24, August 25 to 29, 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

In comparing figure 1 and figure 2, it is evident that the human-written article is longer than the LLM-generated article. This discrepancy arises due to the token limit imposed on LLM outputs. Since the input prompt includes the abstract from a scientific paper, a significant portion of the token allocation is consumed, thereby limiting the length of the LLM-generated article. Despite this, the distributions figure 1 and figure 2 are remarkably similar. Further analysis of figure 3 and figure 4 reveals a high consistency in the distribution of sentence lengths, suggesting that LLMs are capable of producing articles that closely mimic human writing.

2 ANALYSIS TO THE LLM-GENERATED ARTICLES

The ROUGE score is a measure of textual overlap between a generated text and a reference widely used in measuring hallucination [2, 3]. This inspired us to utilize ROUGE to evaluate our generated results. In section 3.4, we utilized the ROUGE source to validate the content in our LLM-generated part. After manually reviewing a sample of generated articles, we calculated and applied ROUGE score intervals to filter the entire dataset effectively.

Additionally, we employed the Bert-Score (another commonly used assessment method for hallucination), which assesses contextual understanding and lexical polysemy, providing a nuanced evaluation, to evaluate the LLM-generated part again. After calculating, the average Bert-Score for 'Abstract - Generated False Article' was 0.8269, while 'Abstract - Generated True Article' scored higher at 0.9127.

By examining the distributions presented in Figure 5, we observed a significant difference between the two Bert-Scores. This discrepancy not only validates the effectiveness of our filtration process but also confirms the authenticity of the fake articles we generated

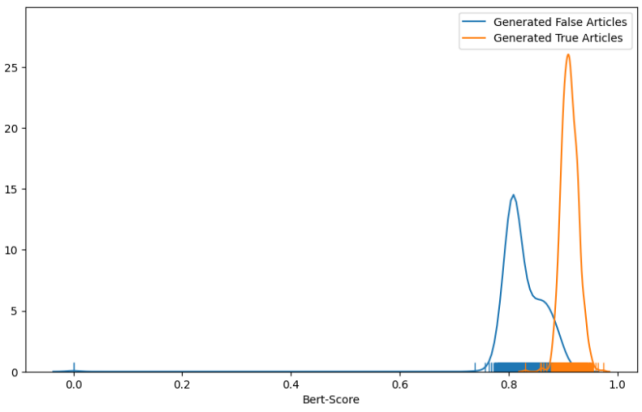

Figure 5: The orange line is the Bert-Score distribution between 'Abstract - Generated True Article' and the blue line is the Bert-Score distribution between 'Abstract - Generated False Article'.

3 SUPPLEMENTARY EXPERIMENTS

3.1 Baseline Experiment

The SciNews dataset aims to address a significant gap left by previous datasets, which involved manual claim generation steps while no original articles were provided (such as SciFact [5] and CheckCovid [6]). This limitation makes it challenging to directly apply these datasets within our framework. Despite these challenges, we have established a baseline using BERT-based models to enhance our analytical rigor. We treat it as an NLR task. Given a news or summarized news paragram and relevant selected evidence from the evidence corpus, the reasoning model acts as an evaluator to identify a pair of news/summarized news and related evidence as true or false. The model input will be $[News < SEP > EvidenceSentence]$ or $[SummarizedNews < SEP > EvidenceSentence]$. The 'Summarized News' and 'Evidence Sentences' come from our best-performing experimental step (SIF with Zero-Shot setting by using GPT-4). We choose two pre-trained models as baseline: BERT [4] and SciBERT [1]. For SciBERT, it trained using masked language modeling on a large corpus of scientific text. We would like to understand how different the models are that include domain information versus those that do not include domain information.

From table 1, preliminary comparisons using BERT and SCIBERT indicated that while SCIBERT showed improved results, it still did not surpass our best-performing setups. In the baseline results, the LLM-generated part was also lower than the Human-Written articles. This again illustrates the difficulty of detecting LLM-generated scientific misinformation. However, the superior performance of SciBERT suggests that pre-trained models enriched with scientific content are more effective in detecting misinformation in scientific news. This observation implies that augmenting large language models (LLMs) with additional scientific data could further enhance their capability. Consequently, fine-tuning the LLAMA2-7B model on a specifically tailored scientific corpus merits additional investigation.

Furthermore, it should be noted that in the comparison experiments, we need to divide the validation set and the test set, so the performance in dealing with out-of-domain data is yet to be evaluated. While our LLM experiment is directly tested using all 2.4k data, the zero-shot performance of LLM is unintentionally better, which helps to build a misinformation detection platform more easily.

3.2 Experiment Results on LLAMA2-70B

The computation cost for LLama2-70B on the task that we have is very high. We tried our best to complete a set of experiments on the LLAMA2-70B in the zero-shot setting.

From table 2, we found LLAMA2-70B performance slightly superior to LLAMA2-13B, and 13B also outperformed the 7B model. However, the overall results of LLAMA2-70B were still underwhelming. In contrast, GPT-3.5 (340B) demonstrated significant improvements, with GPT-4 delivering the best performance, indicating a strong correlation between increased model parameters and enhanced reasoning capabilities.

Table 1: Performance results of baseline models. 'N+ES' denotes 'News + Evidence Sentence', 'SN+ES' denotes 'Summarized News + Evidence Sentence'.

Models	Input Text	Human-Written				LLM-Generated				Overall			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
BERT	N+ES	53.33	53.33	100	69.56	49.44	49.44	100	66.17	51.37	51.37	100	67.87
	SN+ES	54.72	54.72	100	70.74	53.33	53.33	100	69.56	54.02	54.02	100	70.17
SciBERT	N+ES	72.66	73.33	76.10	74.69	69.44	95.35	44.79	60.99	71.05	84.34	60.45	67.84
	SN+ES	76.11	76.23	80.20	78.17	72.22	67.26	85.39	75.24	74.17	71.75	82.79	76.71

Table 2: Performance results of LLAMA2-70B on Zero-Shot Prompt.

LLAMA2-70B	Prompt Strategy	Human-Written				LLM-Generated				Overall			
		Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
	Zero-Shot	67.08	60.04	99.00	75.00	53.58	52.80	97.20	68.40	60.33	56.42	98.10	71.70

3.3 Results of separating LLAMA2-Article and GPT3.5-Article.

For the LLM-generated part, we conducted further individual analyses on the detection performance of articles generated by LLAMA2-7B and GPT-3.5, respectively. We choose the result with the best prediction for the LLM part (D2I under CoT setting by using GPT-4):

- The results for LLAMA2-7B part are as follows: Accuracy: 73.00%, Precision: 65.29%, Recall: 98.00%, and F1 Score: 78.43%.
- In comparison, the results for GPT-3.5 part, detailed in include: Accuracy: 66.00%, Precision: 59.76%, Recall: 98.00%, and F1 Score: 74.23%.

The results indicate better detection performance for articles generated by LLAMA2-7B compared to those by GPT-3.5, suggesting that it is easier to detect articles generated by LLAMA2-7B. However, these results still fall short of the detection rates achieved on human-written articles. As we said in Appendix B, even a relatively small model like the 7B can produce high-quality scientific misinformation, posing a significant risk to public safety.

4 EXPLAINABILITY ANALYSIS

We have conducted further analyses of the explanations generated by LLMs. Our observations reveal that the reasoning process frequently described by the LLM as 'found by comparison of **Accuracy** or **Generalization**...' aligns precisely with what we have defined as 'scientific validity.' This alignment suggests that our definition of scientific validity could be leveraged by CoTs to strengthen reasoning through In-Context Learning, thereby enhancing the LLM's differentiation between true and false scientific claims.

Additionally, we noted that terms such as 'Accuracy' or 'Generalization' are most frequently mentioned in the LLM's reasoning outputs. Our experiments also demonstrate that each definition we have established is consistently referenced in the reasoning process. These results were achieved by the synergistic function of all our defined terms.

We provide more results in the folder.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* (2019).

[2] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. *arXiv preprint arXiv:2402.03744* (2024).

[3] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974* (2020).

[6] Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023. Check-covid: Fact-checking COVID-19 news claims with scientific evidence. *arXiv preprint arXiv:2305.18265* (2023).