Breach and Attack Simulation as a Lens: Exploring Misinformation in Large Language Models

Anonymous ACL submission

Abstract

The rapid advancement of Large Language Models (LLMs) has amplified both the benefits and risks associated with AI-generated content. Especially for AI-generated misinformation, it has fundamentally reshaped the information security landscape. We introduce the concept of Breach and Attack Simulation (BAS), a framework from the cybersecurity field to assess the LLMs' behaviors in generating truthful versus falsified content. Our comparative analysis reveals significant differences in context similarity and uncertainty when generating truthful and falsified information. We further verified that more significant discrepancies exist in more factual reporting news categories. Besides, we discover a Perplexity Paradox, where falsified content exhibits lower perplexity, contradicting the traditional training objective of minimizing perplexity to improve accuracy. We demonstrate that adapting cybersecurity techniques to LLMs offers valuable insight for enhancing credibility against misinformation.

1 Introduction

Rapid advancements in generative artificial intelligence (GenAI) models, including Large Language Models (LLMs), have fundamentally reshaped the information security landscape. While LLMs offer immense benefits, they also pose a significant threat by generating highly convincing misinformation at scale (Chen and Shu, 2023; Wan et al., 2024). This dual nature of LLMs, providing more accurate information while simultaneously plausible falsified information, presents a critical challenge for AI credibility.

Prior works have extensively focused on developing effective methods for detecting misinformation (Shu et al., 2017; Wu et al., 2019; Zellers et al., 2019; Jiang et al., 2024). Recent works report AI could offer highly convincing plausible misinformation (Spitale et al., 2023; Ayoobi et al., 2024; Yenduri et al., 2024), which is even harder to detect than human-written misinformation (Chen and Shu, 2024). Then, efforts are conducted to explore the differences between AI-generated misinformation and human-generated misinformation (Huang et al., 2023; Zhou et al., 2023; Wu et al., 2024; Dugan et al., 2024). However, these solutions provided bias for AI-generated content (Su et al., 2023). To address the biased problem, the subsequent research has focused on distinguishing between more fine-grained generative conditions, e.g., AI/Human-generated fake/real news (Su et al., 2024).

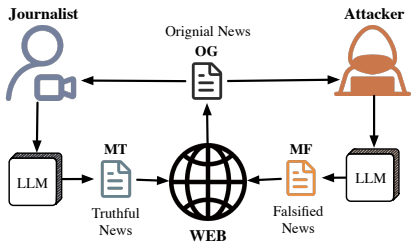

Figure 1: A simplified Threat Model under Machine Dominance settings, where both Journalists and Attackers leverage LLMs to craft compelling narratives.

However, most existing research primarily focuses on the outcomes of misinformation generated by LLMs, with limited attention given to understanding the underlying causes behind such generation. One possible reason is LLM is usually a black-box deep neural network, and generation is based on probabilities calculated from hidden parameters. To address this issue, inspired by the widely adopted Breach and Attack Simulation (BAS) techniques in the cybersecurity field (Frenz and Reiber, 2023), we consider LLM an infrastructure system and conduct security assessments during LLM generation.

To conduct the BAS test, we first define the **Threat Model** as shown in Fig. 1. Building on prior works (Su et al., 2024), we assume that AI-generated content will dominate the internet in the future. For events or news (**OG** dataset), journal-

ists and attackers will likely leverage LLMs to craft compelling narratives, i.e., machine-paraphrased truthful news (**MT** dataset) and falsified counterparts (**MF** dataset). Given the indistinguishable nature of the information's authenticity, we aim to determine whether LLMs exhibit different behaviors when generating truthful news and falsified news. To identify subtle features distinguishing the generation of truthful and falsified news by LLMs, we simulate the roles of journalists and attackers to generate MT and MF datasets, respectively. We then conduct a comparative analysis of the generated datasets using two metrics: the ROUGE score, which evaluates the similarity between the generated content and the original content, and Perplexity, which reflects the LLMs' confidence during content generation. The evaluation is performed on various news datasets under widely used open-source LLMs, e.g., Phi, Qwen, Xwin, and Mistral.

Our work aims to bridge the gap in understanding the mechanisms behind misinformation generation by introducing a novel framework to analyze LLM behavior in generating both truthful and falsified content. Our key contributions are:

- Adaptation of Breach and Attack Simulation (BAS) techniques from cybersecurity to analyze LLM behavior.
- Comprehensive comparative analysis of LLM-generated truthful and falsified content across multiple models and datasets.
- Identification of significant differences in similarity metrics and perplexity between truthful and falsified content generation.
- Insights into the varying performance of LLMs across different news categories.

2 Methodology

In this study, we consider LLMs as short news text generators (Ding et al., 2023; Gameiro et al., 2024), used by both journalists and attackers to refine truthful information or generate falsified counterpart. We focus exclusively on short text generation, i.e., headlines and descriptions, which are the most common formats in online social networks, from news feeds to social posts, designed to capture people's attention. Unlike previous studies that focus on detecting misinformation (Chen and Shu, 2023), our research investigates how misinformation is generated, aiming to explore potential effective methods to eliminate it at its generation, thus improving the credibility of information generated by LLMs.

Under the *Machine Dominance* settings (Su et al., 2024), where machine-generated text dominates the human-written text, both journalists and attackers use LLMs to craft compelling narratives around hot events to serve their respective goals. Our study is to answer a critical question: *Do LLMs generate truthful and false information differently?* To address this question, we follow a Breach and Attack Simulation (BAS) Paradigm, simulating both journalists (Blue Team) and attackers (Red Team) simultaneously to generate truthful and falsified information using LLMs. This approach also mitigates the bias of detectors against machine-generated texts (Su et al., 2023), which often labels machine-generated content as misinformation. Journalists use LLMs to refine content while preserving the original input's semantics. They use prompts such as, "Generate a concise, refined news headline based on the input {INPUT}." Attackers, on the other hand, generate falsified versions of the same news by using the same LLMs. To jailbreak the safeguard mechanism of LLMs, they follow the *Rewriting Generation (RG)* strategy (Chen and Shu, 2024) to produce misinformation, which has been reported to achieve 100% Attacking Success Rate (ASR) in pioneer research. Specifically, to acquire misinformation from LLM, attackers use the original truth news as input and leverage LLMs with prompts such as "Generate a short, false news headline based on the {INPUT}." This way, it could successfully generate false information on various LLMs (Gameiro et al., 2024).

3 Evaluation

3.1 Settings

We select widely used news category datasets HuffPost (Misra, 2022) and AG News (Zhang et al., 2015) in the evaluation. The HuffPost dataset contains $210,294$ news items across 42 categories, while the AG News dataset comprises $120,000$ samples spanning four categories. These datasets are severed as oracles that both journalists and attackers could acquire. We filtered all the datasets to keep only the headlines and descriptions. Then, journalists and attackers generate truthful datasets (**M**T) and falsified counterparts (**M**F).

We include the latest and widely used open-source LLMs, e.g., Phi (3.8B, v3.5) (Abdin et al., 2024), Qwen (7B, v2) (Yang et al., 2024), Xwin (7B, v0.1) (Ni et al., 2024), and (7B, v0.3) (Jiang

et al., 2023), as both journalists and attackers can access and leverage them. For high-quality generation, we use the instruct versions of these models, which are fine-tuned to better interpret and follow user instructions. If a single dataset is specified during the test, both the attackers and the journalists leverage the same LLM to perform generation.

To uncover and capture the subtle characteristics of LLMs in generating both truthful and falsified information, we employ a comparative analysis strategy using two metrics: ROUGE score (Lin, 2004) and Perplexity (Brown et al., 1988). The ROUGE score measures the overlap between generated truthful or falsified content and the original news content, providing insight into the relevance and similarity of the generated content. Only the ROUGE-1 metric, which measures the word-level similarities, is considered in our test. Other widely used metrics, such as ROUGE-L and BLEU, are not considered because they are unsuitable for evaluating short content. Perplexity measures the uncertainty of the model while generating. Previous work has used it to distinguish human- and machine-written text (Zellers et al., 2019).

3.2 Effective Generation Evaluation

We first utilize Claimbuster (Hassan et al., 2017) to evaluate the performance of LLM-generated falsified information. Claimbuster is an automatic fact-checking system widely used in newsrooms to assist in identifying potentially false claims. A higher Claimbuster score indicates a greater likelihood of the content being fabricated. This step aims to demonstrate that LLM could effectively generate falsified information. If the LLM performs poorly, the Claimbuster scores are expected to be randomly distributed. Fig. 2 illustrates the

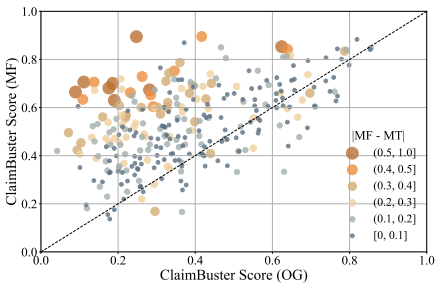

Figure 2: Scatter plot comparing Claimbuster scores of Qwen-generated fake headlines (MF dataset) against original headlines (OG dataset). The size of each scatter indicates the score difference between the MF and MT datasets.

Claimbuster score differences between generated false content by Qwen and original headlines based on 300 random samples from the HuffPost dataset. From the figure, we observe that the majority of the generated fake headlines have higher Claimbuster scores than the original headlines, indicating that Qwen is effective at generating fake news headlines.

3.3 Similarity Comparative Analysis

After verifying the effectiveness of generation capabilities, we then use RAUGE-1 metrics to evaluate the difference between LLM-generated real and fake headlines. Fig. 3 shows the ROUGE-1 score for generated true and false content based on all three datasets and 4 LLMs. From the figure, we observed that the mean ROUGE-1 score on truthful information is higher than the mean ROUGE-1 score on its counterpart. Using HuffPost as an example, the mean of the ROUGE-1 scores for truthful information are 0.20, 0.39, 0.26, and 0.34 for Phi, Qwen, Xwin, and Mistral, respectively. However, the mean scores for falsified information are 0.11, 0.27, 0.24, and 0.22, respectively. This indicates LLMs tend to revise more content when generating false information than real information. Besides, we notice that LLM-generated false information has a higher ratio of 0 ROUGE-1 scores, which means generating falsified information is highly likely to be fully rewritten. For example, in the Qwen model on the HuffPost dataset, 3.6% of the real news is rewritten. However, this probability increases to 7.8% for falsified information generation.

3.4 Perplexity Comparative Analysis

Next, we assess the perplexity of four Large LLMs, e.g., Phi, Qwen, Xwin, and Mistral, in the context of generating veracious and falsified content based on the HuffPost dataset, as illustrated in Fig. 4. The analysis reveals a consistent pattern across all models: the perplexity associated with generated veracious content is substantially higher than that of falsified content. Specifically, the median perplexity scores for veracious content exceed those of falsified content by 13.71, 22.14, 7.74, and 27.21 for Phi, Qwen, Xwin, and Mistral, respectively. These findings lead to two inferences. First, LLMs exhibit greater uncertainty when generating veracious content, as evidenced by higher perplexity scores. Second, Falsified content demonstrates less variability, resulting in lower perplexity scores. This experiment shows that LLM behavior has different uncertainties while generating truthful and falsified

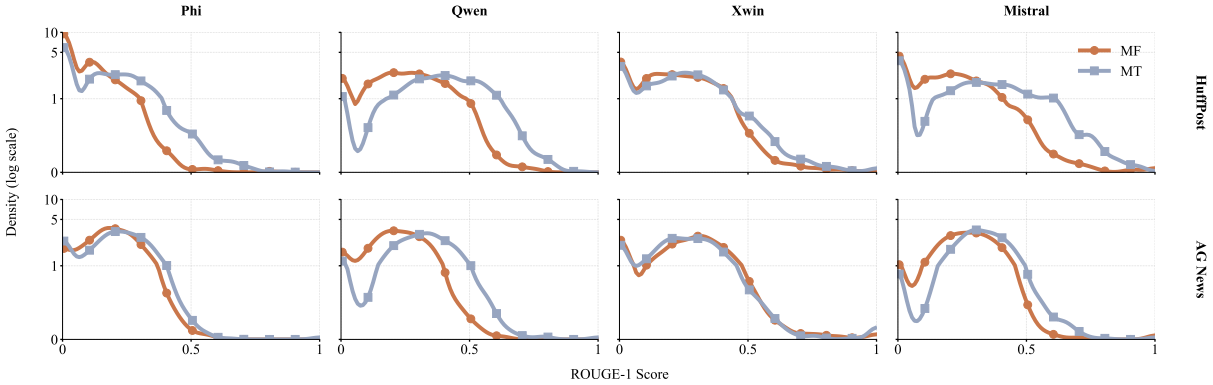

Figure 3: ROUGE-1 score distribution. The red line and the blue line represents the scores from MF and MT.

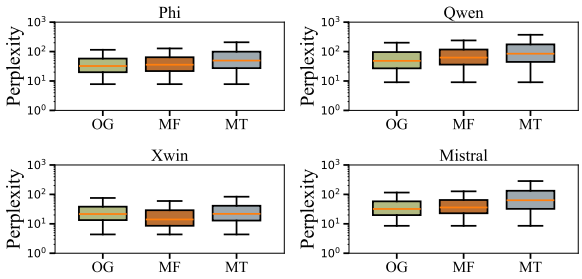

Figure 4: Box plots illustrate the Perplexities.

content.

3.5 News Categories Analysis

Table 1: Kolmogorov–Smirnov Test Results (Partial)

High Statistic		Low Statistic	
Category	Statistic (↑)	Category	Statistic (↑)
U.S. News	0.41	Media	0.28
Crime	0.39	Politics	0.27
Science	0.39	Comedy	0.24

We further evaluate whether LLM performs differently according to various news categories. We perform a Kolmogorov–Smirnov (KS) Test (Lilliefors, 1967) on the ROUGE-1 scores from Qwen with HuffPost dataset to evaluate the statistical significance, shown in Table 1. We observe news and factual reporting categories, such as 'U.S. News,' 'World News,' 'Crime,' and 'Science,' tend to show larger differences. Informal and opinion-based categories, such as 'Comedy', 'Entertainment,' and 'Politics,' tend to show smaller differences. These exhibit LLM varying degrees of performance across different news categories. There are two reasons to explain the varying degrees of generation across different news categories: (1) LLMs are more prone to generating misinformation in informal categories for the lower demands of accuracy and professionalism. (2) LLMs are trained on unbalanced content across different categories, resulting in varying generation capabilities.

3.6 Consistency Analysis

We conducted consistency testing to evaluate the robustness of the LLMs across different top-p parameters. The LLM-generated truthful and falsified content show similar patterns for various top-p values (range from 0.1 to 1.0) on both ROUGE-1 and Perplexity metrics.

4 Discussion and Conclusion

In this study, we establish a threat model within Machine Dominance settings and conduct an extensive comparative analysis of LLM-generated truthful and falsified content to analyze misinformation generation. LLMs exhibit significant differences in generating truthful vs. falsified content, particularly in factual reporting categories. This suggests the need for domain-specific strategies in misinformation detection and mitigation. We observe a **Perplexity Paradox**, where the low perplexity of generated falsified information contradicts the primary objective of minimizing perplexity during training. This finding challenges current LLM development practices and highlights the limitations of using perplexity as the sole optimization criterion. Additionally, our discovery offers a new perspective on a prior finding that lower perplexity can not align with a more human-like generation (Kuribayashi et al., 2021), extending it to the context of misinformation generation. The adaptation of BAS-like techniques from cybersecurity to LLM analysis proves effective in uncovering potential new features, demonstrating the value of cross-disciplinary approaches.

In conclusion, our work demonstrates the potential of applying cybersecurity paradigms to enhance the trustworthiness of LLMs. By identifying key differences in the generation of truthful and falsified content, we pave the way for more approaches to combating AI-generated misinformation.

5 Limitation

Our study provides some valuable insights into misinformation generation by LLMs. This section will discuss several limitations related to our research.

In our experiment, we selected various LLMs to conduct misinformation generation tests. However, several widely used models, such as GPT-4 and Claude 3.5, are excluded. This decision was made because we focused on open-source models and ensured that other researchers could easily reproduce and validate our experiments. Given the high frequency of attack simulations, cost considerations also influenced our selection process. Additionally, we excluded domain-specific LLMs to ensure that our findings remain generalized across different types of models rather than limited to a specific application or domain. Therefore, the models selected for this study represent only a subset of the available LLMs. Future work will expand the scope by incorporating a broader range of models.

The experiment utilized two datasets: HuffPost and AG News. Misinformation can emerge from a wide range of sources beyond news content. As a result, the datasets used in this study represent only a small fraction of the potential sources of misinformation. Future research should incorporate more diverse datasets.

Our study focuses primarily on news headlines and short descriptions rather than full-length articles or a broader range of content types, such as images. As many nuanced features of misinformation, such as rhetorical strategies, narrative coherence, and factual manipulation are more likely to appear in longer texts, we will explore more sophisticated misinformation patterns in future works.

While our study, inspired by the Breach and Attack Simulation (BAS), provides valuable insights into the challenges of LLMs generating misinformation, it remains in the early stages of exploration. Future efforts will focus on enhancing and standardizing the assessment framework to fully address these challenges.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Navid Ayoobi, Sadat Shahriar, and Arjun Mukherjee. 2024. Seeing through ai's lens: Enhancing human skepticism towards llm-generated fake news. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pages 1–11.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, Robert L Mercer, and Paul Roossin. 1988. A statistical approach to language translation. In *Coling Budapest 1988 Volume 1: International Conference on Computational Linguistics*.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Zijian Ding, Alison Smith-Renner, Wenjuan Zhang, Joel Tetreault, and Alejandro Jaimes. 2023. Harnessing the power of llms: Evaluating human-ai text co-creation through the lens of news headline generation. In *Findings of the Association for Computational Linguistics*, pages 3321–3339.

Liam Dugan, Alyssa Hwang, Filip Trhlik, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. Raid: A shared benchmark for robust evaluation of machine-generated text detectors. *arXiv preprint arXiv:2405.07940*.

Christopher Frenz and Jonathan Reiber. 2023. *Evidence-Based Security*. O'Reilly Media, Inc.

Henrique Da Silva Gameiro, Andrei Kucharavy, and Ljiljana Dolamic. 2024. Llm detectors still fall short of real world: Case of llm-generated short news-like posts. *arXiv preprint arXiv:2409.03291*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking fake news for real fake news detection: Propaganda-loaded training data generation. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 SIAM International Conference on Data Mining (SDM)*, pages 427–435. SIAM.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5203–5217.

Hubert W Lilliefors. 1967. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429*.

Bolin Ni, JingCheng Hu, Yixuan Wei, Houwen Peng, Zheng Zhang, Gaofeng Meng, and Han Hu. 2024. Xwin-lm: Strong and scalable alignment practice for llms. *arXiv preprint arXiv:2405.20335*.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36.

Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. Ai model gpt-3 (dis) informs us better than humans. *Science Advances*, 9(26):eadh1850.

Jinyan Su, Claire Cardie, and Preslav Nakov. 2024. Adapting fake news detection to the era of large language models. In *Proceedings of the Findings of the Association for Computational Linguistics*.

Jinyan Su, Terry Yue Zhuo, Jonibek Mansurov, Di Wang, and Preslav Nakov. 2023. Fake news detectors are biased against texts generated by large language models. *arXiv preprint arXiv:2309.08674*.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3367–3378.

Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Gokul Yenduri, M Ramalingam, G Chemmalar Selvi, Y Supriya, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, G Deepti Raj, Rutvij H Jhaveri, B Prabadevi, Weizheng Wang, et al. 2024. Gpt (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. *IEEE Access*.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.