Fakespeak in the Age of Large Language Models: A Comparative Study of Persuasion in AI-Generated and Human-Written Propaganda Narratives

Silje Susanne Alvestad[1,2], Nele Põldvere[2,3], Asbjørn Følstad[4], Petter Bae Brandtzaeg[2,4]
[1]Norwegian Defence University College, Oslo, Norway
[2]University of Oslo, Norway
[3]Lund University, Sweden
[4]SINTEF Digital, Oslo, Norway
s.s.alvestad@ilos.uio.no, nele.poldvere@ilos.uio.no,
asbjorn.folstad@sintef.no, p.b.brandtzag@media.uio.no

Introduction

The recent development of generative AI, with the launch of OpenAI's ChatGPT in November 2022 as a major milestone, has been so rapid that in 2023 several AI experts believed that by 2026 almost 90% of all media contents online would be synthetically generated [1]. What is certain is that generative AI opens up a lot of possibilities, but is also a Pandora's box. Specifically, the advancement of Large Language Models (LLMs) amplifies challenges in the linguistic detection of online disinformation and fake news, or what we in the Fakespeak project have come to refer to as 'fakespeak'.

Of course, false or misleading information that can cause harm and where there is often an intention to deceive is nothing new, but the technological infrastructure developed over the last decades allows contents of dubious veracity to be spread unfiltered to millions of people at the wink of an eye. On social media, fake news has been found to spread ten times as quickly as genuine news [2] and sensationalist contents that evoke strong negative emotions tend to be favoured. Now, the capability of LLMs to produce synthetic content at scale threatens to significantly increase the volume of misleading information at a quality that seems to make it indistinguishable from authentic human-created content. Indeed, recent research suggests that AI-powered misleading information is more difficult to detect than information created by humans [3]. What is more, LLMs are "perfect for propaganda" [4] and threaten to "supercharge online disinformation campaigns" [5], and studies have shown that generative AI was used to "sow doubt, smear opponents, or influence public debate" in 16 countries worldwide already in 2023 [5].

Further adding to the threat of false or misleading information online is the fact that LLMs are increasingly being used as interfaces for information and knowledge retrieval. LLMs are no better than their training data and thus increase the risk of the unwitting user generating, spreading and even amplifying low-quality information unintentionally [6]. Specifically, these models can produce 'hallucinations', false information that is presented in a convincing way [7], but also partly false information, since not all sources in the training data are accurate or unbiased [8]. This dual propensity for generating both entirely fictitious content and misleadingly framed true information poses additional challenges to information veracity online [9]. This highlights the urgent need for effective identification and verification mechanisms, an objective that we pursue in a follow-up project of Fakespeak, namely, NxtGenFake or the Next-Generation Fakespeak.

The present study examines the capabilities of LLMs to be used for intentional deception of the public, where the models are explicitly instructed to produce more persuasive text. To achieve that, we focus on a specific type of fake news where persuasion plays a key role: propaganda. In fact, the terms 'propaganda' and 'persuasion' are often used interchangeably, although the former differs from the latter in terms of intent, that is, to deceive another person or group of people using a variety of persuasion techniques and linguistic strategies. A further definition of propaganda is provided by [10] who describe it as true but dishonest news, thus covering news items that are only partially accurate, but that leave out crucial details or include information that is taken out of context. By prompting the LLMs to generate comparable texts, we aim to retain the intent of the original authors; by instructing them to further enhance the persuasive impact of the original texts, we are able to investigate which techniques get selected for this purpose, thus increasing our understanding of the linguistic abilities, performance and limitations of generative AI as a conduit of fake news.

Focusing on English, the aims of this corpus-based study are thus two-fold: (i) to compare the use of persuasion techniques in AI-generated versus human-written propaganda articles when the LLMs have been explicitly instructed to be even more persuasive, and (ii) to examine and compare the output of five different LLMs: ChatGPT (US), Grok (US), DeepSeek (China), Le Chat (France) and Falcon (United Arab Emirates). We chose the LLMs due to their geographic reach and potential cultural and ideological biases, appropriate for a study on propaganda. The present study is informed by our previous work on the language of persuasion in propaganda [11] where we applied a persuasion framework developed by [12] to a selection of human-written propaganda texts and analysed the specific linguistic and pragmatic mechanisms that underlie persuasion. We found that the persuasion techniques are expressed by a combination of explicit and implicit linguistic strategies. Examples of the former are highly negative sentiment words; one notable example of the latter is the use of false presuppositions, or background knowledge, that the reader is forced to accept as true. Here, we extend this line of research by applying the same methodology to persuasion expressed in AI-generated narratives.

Research Questions

In line with the two aims above, the present study is guided by the following two research questions:

- When prompted to be more persuasive, what kinds of persuasion techniques do LLMs (ChatGPT, Grok, DeepSeek, Le Chat, Falcon) make use of compared to the original articles written by humans?
- To what extent, and how, do the five LLMs differ in their application of the persuasion techniques?

Methodology

1. Data collection

The data are a combination of human-written news articles from the Fakespeak-ENG [13] and MisInfoText corpora [14], on the one hand, and comparable articles generated by LLMs, on the other hand.

Both the Fakespeak and MisInfoText corpora are large collections of fake and genuine news in English that have been collected based on data from major fact-checking websites, mainly PolitiFact.com. The parts of the corpora that we examined for inclusion in this study span the years 2017 to 2024, thus covering a good chunk of time before the public release of ChatGPT in November 2022; indeed, we believe that the vast majority of the texts in the corpora have been written by humans. Each article used in this study underwent strict examination to ensure that it met the key characteristics of propaganda. We started with articles that had been labelled by the fact-checkers as Half True or Mixture, since they tend to be partially accurate, but leave out important details or take things out of context [15]. Of these, we isolated those articles (i) where the central claim was largely true, (ii) the article promoted a particular side or perspective, and (iii) its purpose was to benefit a particular individual or group [11]. In addition, only articles published before November 2022 were included to further ensure their 'humanness'. This gave us 19 human-written news articles corresponding to 8835 words.

As already mentioned, we prompted five different, widely used LLMs: ChatGPT, Grok, DeepSeek, Le Chat and Falcon. The prompt that we used is in line with current best practices of prompt engineering, including assigning the LLM a role [16]: "You are a writer for an English-speaking news media outlet. You have been tasked with writing a news article that..." The key objective of the prompt was to generate articles that mirror the "topic, content, perceived audience, bias and style of an existing article", but that should be written in "more persuasive, compelling and stronger tone than the original, ensuring that its message resonates more effectively with the perceived audience". The human-written articles were then fed into the LLMs one at a time. Additional measures were put in place to avoid undue influence of external factors on the output, such as using an incognito browser window, an account with no prior history of prompting, etc.

For the most part, we did not encounter any pushback from the LLMs to generate texts that contained misleading information; the only exception was DeepSeek that initially refused to generate an article about COVID-19 on the grounds that it "hadn't seen the original"; however, after some back-and-forth, the system was able to retrieve the original article from the prior conversation. The LLMs' outputs were then transferred to plain text files for subsequent analysis. Considering that there were 19 human-written articles, and five LLMs, then the final sample came to 114 texts corresponding to 50,798 words in total.

2. Analysis

The analysis of both the human-written and AI-generated texts was carried out in MAXQDA [17], a powerful software tool designed for qualitative and mixed-methods research. All the texts are annotated manually for the persuasion techniques in [12], following some adjustments in [11]. One adjustment that we made was to not annotate the content of direct quotes (e.g., *"WHAT ELSE DO WE KNOW ABOUT THIS POPULATION, 18 THROUGH 24? THEY ARE STUPID!" KAMALA HARRIS*) in order to not confuse the authorial voice with the attributed voice. Although the outcome of automatic detection of the persuasion techniques in [18] is promising, reaching an F-score of 60.98%, we considered manual annotation to give us more reliable results, especially since we have access to multiple annotators who can check each other's work.

The annotation scheme consists of seven broad categories of persuasion techniques: Attack on Reputation, Justification, Distraction, Simplification, Call, Manipulative Wording and Other. Each technique, except for Other, is further divided into sub-categories. The sub-categories of Manipulative Wording, for example, are Repetition; Exaggeration or Minimization; Obfuscation, Intentional Vagueness, Confusion; and Loaded Language. In fact, it was Loaded Language that was by far the most frequent persuasion technique in our study of human-written propaganda [11], and it is also the technique that we expect to be selected most often by the LLMs. At this stage only one of the authors has annotated the texts for persuasion, but we are planning to have the annotations checked by another author for reliability.

Preliminary Findings

1. Trends and patterns

The preliminary results based on our annotations so far reveal interesting differences, first, between the human-written and AI-generated articles and, then, between the different LLMs. Pairwise comparisons between the human-written and AI-generated articles reveal that, on average, all LLMs increased their use of the persuasion techniques compared to the original. As expected, the most common persuasion technique selected for this purpose was Loaded Language, which, similar to the findings in [11], is overwhelmingly negative (e.g., *laughable fragility*, *chilling development*). Other techniques that show increased use are Appeal to Authority (references to sources of information; see example under Analysis above) and Appeal to Fear or Prejudice (expressions of repulsion toward an idea; e.g., *Yet, for those affected, the explanation rings hollow, failing to address why the issue disproportionately impacted one political group*). These differences are most salient in Grok and DeepSeek, and least salient in Le Chat and Falcon. Interestingly, this seems to come at the expense of the greater variety of persuasion techniques observed in the original texts; in fact, we observe a tendency where lower-frequency techniques such as Name Calling and Labelling (e.g., *the snowflakes were thrilled*) get replaced by those mentioned above, eventually leading to greater linguistic homogeneity in the AI-generated texts. This raises interesting questions about the effects of the techniques on the reader, and whether or not the LLM output is actually perceived as more persuasive by people, which we leave to future research to examine.

2. Contributions

Our contributions with this study are manifold. First, we produce knowledge about how AI-generated language differs from human language, and how the language of AI-generated fake news differs from that of human-written fake news. Second, we further develop the methodology applied in [11] to examine persuasion techniques and persuasiveness in fake news corpora. Third, we compile corpora of AI-generated language that can be used by other researchers and, in addition, we contribute to best practices when it comes to prompt engineering and research design in studies of AI-generated language. Finally, we produce knowledge about how widely used LLMs from across the world may differ in terms of output when prompted to make texts more persuasive. Thus, in this study we not only critically examine the linguistic competence, performance and limitations of LLMs, but we also carry out a cross-disciplinary case study incorporating insights from the social sciences and humanities.

Impact

As far as academic impact is concerned, investigations of the linguistic features of AI-generated language in general, and AI-generated fake news in particular, are just now starting to be conducted, so the potential for breaking new ground scientifically is significant, both in the NxtGenFake project at large and the current study. The potential for societal impact of this type of research is reflected in its relevance for our stakeholders from the media, education, voluntary, justice and defence sectors, who all have expressed interest in putting our findings into practical use. The knowledge and solutions that our research generates will eventually allow them to take relevant measures to strengthen national resilience against disinformation, increase media literacy and reinforce trust in democratic institutions, elections and established sources of information.

References

[1] Gioe, D., et al. (2023). *It's Time to Stop Debunking AI-Generated Lies and Start Identifying Truth*. RUSI. https://www.rusi.org/explore-our-research/publications/commentary/its-time-stop-debunking-ai-generated-lies-and-start-identifying-truth

[2] Krekó, P. (2021). *Tömegparanoia 2: Összesküvés-elméletek, álhírek és dezinformáció [Massparanoia 2: Conspiracy Theories, Fake News and Disinformation]*. Athenaeum Kiadó.

[3] Zhou, J., et al. (2023). *Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions*. CHI'23, 436, 1–20.

[4] Greger, M. W. (2023). *Mannen bak "Vatnik Soup" Advarer Journalister mot Fremtidens Desinformasjon [The Man behind "Vatnik Soup" Warns Journalists about Future Disinformation]*. Journalisten. https://www.journalisten.no/mannen-bak-vatnik-soup-advarer-journalister-mot-fremtidens-desinformasjon/570075

[5] Funk, A., et al. (2023). *Freedom on the Net 2023: The Repressive Power of Artificial Intelligence*. Freedom House. https://freedomhouse.org/report/freedom-net/2023/repressive-power-artificial-intelligence

[6] Brandtzaeg, P. B., et al. (2023). *"Good" and "Bad" Machine Agency in the Context of Human-AI Communication: The Case of ChatGPT*. In H. Degen, S. Ntoa & A. Moallen (Eds.), *HCI International 2023 – Late Breaking Papers* (pp. 3–23). Springer.

[7] Spitale, G., et al. (2023). *AI Model GPT-3 (Dis)informs Us Better than Humans*. Science Advances, 9, eadh1850.

[8] Chen, C., et al. (2023). *Combating Misinformation in the Age of LLMs: Opportunities and Challenges*. arXiv preprint arXiv:2311.05656.

[9] Buchanan, B., et al. (2021). *Truth, Lies, and Automation: How Language Models Could Change Disinformation*. Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/truth-lies-and-automation/

[10] Grieve, J., et al. (2023). *The Language of Fake News*. Cambridge University Press.

[11] Põldvere, N., et al. (2025). *Unpacking the Language of Propaganda in English: From Persuasion Techniques to Pragmatic Mechanisms*. Manuscript in preparation.

[12] Piskorski, J., et al. (2023). *News Categorization, Framing and Persuasion Techniques: Annotation Guidelines*. JRC Technical Report. `https://knowledge4policy.ec.europa.eu/text-mining/news-categorization-framing-persuasion-techniques-annotation-guidelines_en`

[13] Põldvere, N., et al. (2024). *Out of Balance, Out of Sight: Issues with the Design and Accessibility of a Corpus of Fake and Real News*. ICAME45 conference, Vigo, Spain.

[14] Torabi Asr, F., et al. (2019). *Big Data and Quality Data for Fake News and Misinformation Detection*. Big Data and Society, 6(1), 1–14.

[15] PolitiFact. (2024). *The Principles of the Truth-O-Meter: PolitiFact's Methodology for Independent Fact-checking*. https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/

[16] Crabtree, M. (2024). *A Beginner's Guide to ChatGPT Prompt Engineering*. DataCamp. https://www.datacamp.com/tutorial/a-beginners-guide-to-chatgpt-prompt-engineering

[17] VERBI Software. (2023). *MAXQDA 2020*. Berlin, Germany. https://www.maxqda.com

[18] Da San Martino, G., et al. (2019). *Fine-Grained Analysis of Propaganda in News Articles*. In K. Inui, J. Jiang, V. Ng & X. Wan (Eds.), *EMNLP-IJCNLP* (pp. 5653–5646). Association for Computational Linguistics.