Behind the Words: Dissecting ChatGPT's Capability of Manipulative Tactics in Forming Product Reviews

Naw Eh Htoo, Sabid Bin Habib Pias (Co-Mentor), Apu Kapadia (Mentor)

OpenAI developed a large language model that analyzes, understands, and generates human language. They trained it using large data texts created by humans. As a result, ChatGPT has the potential to manipulate users through product reviews [1]. Studies have also shown that certain strategies can significantly impact customers' trust and can influence decisions and perceptions [3, 6]. Ill-intended salespersons can instruct GPT models to create manipulative text that influences users' purchase decisions. To understand how GPT models generate deceitful fake reviews, we included negative and positive reviews from Amazon and Best Buy in custom GPT models, designed to write simulated fake reviews [2] for Apple HomePod, Google Nest Audio, and Amazon Echo. These models focused on tactics like exaggeration, omission, half-truths, misleading statements, and expert claims, based on Niemi and Pullins's research on sales [4].

Custom GPT models generated fake reviews. We utilized Word Cloud to do frequency analysis and qualitative analysis to identify manipulative patterns. Frequency analysis revealed the most common words related to sound, voice recognition, and smart home integration. World Cloud's result showed that positive reviews frequently generated positive words, while negative reviews generated negative words frequently. Qualitative analysis found exaggerated reviews used strong vocabulary, omission and half-truth reviews selectively highlighted features, false reviews included non-existent features, and expert claim reviews followed a specific structure while claiming expertise.

These reviews can distort perceptions, cause misunderstandings, set unrealistic expectations, and sway decisions. Studying this is crucial to raise awareness about fake reviews and questioning AI ethics. Future studies should look into ways to recognize manipulative reviews, assess how varying user personalities are affected by specific manipulation techniques [5], and prevent possible manipulation.

References

[1] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3):674–684, 2012.

[2] Rami Mohawesh, Shuxiang Xu, Son N Tran, Robert Ollington, Matthew Springer, Yaser Jararweh, and Sumbal Maqsood. Fake reviews detection: A survey. *Ieee Access*, 9:65771–65802, 2021.

[3] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. What yelp fake review filter might be doing? In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 409–418, 2013.

[4] Jarkko Niemi and Ellen Bolman Pullins. Tell me more: how salespeople encourage customer disclosure. *Journal of Business & Industrial Marketing*, 36(5):717–728, 2020.

[5] Sabid Bin Habib Pias, Alicia Freel, Timothy Trammel, Taslima Akter, Donald Williamson, and Apu Kapadia. The drawback of insight: Detailed explanations can reduce agreement with xai. *arXiv preprint arXiv:2404.19629*, 2024.

[6] Sabid Bin Habib Pias, Ran Huang, Donald S. Williamson, Minjeong Kim, and Apu Kapadia. The impact of perceived tone, age, and gender on voice assistant persuasiveness in the context of product recommendations. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, CUI '24, New York, NY, USA, 2024. Association for Computing Machinery.