Peer response 1 – Martyna's post

In her initial post, Martyna outlined some possible applications of AI models capable of generating text, including examples within healthcare, communications, marketing, and sales. She also rightly points out some of the potential challenges and risks associated with widespread use of AI writers, such as producing potentially harmful content, or biased content arising from inaccurate historical training data, the potential replacing and eventually impacting vital skills such as writing and editing, and the reputation damage of AI-generated errors for companies using these systems.

In order to prevent these risks and help ensure a net positive impact from adoption of AI writers, Martyna highlights the need for responsible development and implementation, underpinned by human oversight and appropriate testing. To this, I would add the need for transparent and ethical development, and the importance of creating frameworks to ensure that companies developing AI models retain some accountability for potential harmful content produced by those models (Hutson, 2021). For example, Anthropic has explicitly made its goal to "to ensure transformative AI helps people and society flourish" (Perrigo, 2024). This ethos is embodied in its Claude model, developed based on a strong moral and ethical imperative outlined in its Constitution, which follows the principles of the Universal Declaration of Human Rights (Anthropic, 2023). This however does not mean the model is open-source, allowing the company to retain full ownership over its product and benefiting financially from it. The focus on ethical and responsible development may indeed prove an advantage, as it helps build user trust, facilitates accountability, and leads to wider adoption. At the same time, it may not only not hinder but also improve the final model performance, even in purely financial modelling use-cases (Starks, 2024).

References

Anthropic (2023) *Claude's Constitution*. Available from: https://www.anthropic.com/news/claudes-constitution.

Hutson, M. (2021) 'Robo-writers: the rise and risks of language-generating AI', *Nature*, 591(7848), pp. 22–25. Available from: https://doi.org/10.1038/d41586-021-00530-0.

Perrigo, B. (2024) ‘Inside Anthropic, the AI Company Betting That Safety Can Be a Winning Strategy’, *Time Magazine*, 30 May. Available from: https://time.com/6980000/anthropic/ (Accessed: 21 December 2024).

Starks, A. (2024) *Anthropic Dominates OpenAI: A Side-by-Side Comparison of Claude 3.5 Sonnet and GPT-4o*. Available from: https://plainenglish.io/blog/anthropic-dominates-openai-a-side-by-side-comparison-of-claude-3-5-sonnet-and-gpt-4o (Accessed: 21 December 2024).