Initial post: Collaborative discussion 2 – Risks and benefits of AI writers

Large language models (LLMs) are Artificial Intelligence (AI) algorithms trained on exceedingly large amounts of publicly-available text available online. These models typically make use of deep learning approaches to establish relationships based on their training data to respond to textual inputs provided by human users. Such algorithms have seen immense development since the early 2020s with expansion of computing power and the amount of textual data available for training, with notable examples including OpenAI's ChatGPT, Google Gemini, and Microsoft Copilot.

Outputs from LLMs can range from short pieces of text to large documents, and have been tested in numerous applications from summarising legal documents to writing computer code, poetry, or music sheets, all with surprising accuracy and realism. The potential for such "AI writers" is immense, from streamlining legal cases, facilitating administrative tasks involving large numbers of dense documents, helping creative professionals overcome writer's block, and supporting students performing online research (Cui et al., 2023; Gosden, 2023; Hubert et al., 2024; Musumeci et al., 2024).

Nonetheless, these applications also entail several potential risks. Like any statistical model, LLMs have been known to make false predictions (usually termed "hallucinations" as they involve generating incoherent or factually wrong text). It's also been suggested that they simply reproduce text without necessarily showing an understanding of those outputs, owning them the nicknames "stochastic parrots" or "bullshiters" by different ethicists (Bender et al., 2021; Hicks et al., 2024). They can also replicate underlying biases in the training data, namely in relation to race, gender, religion, and other characteristics, potentially reinforcing those biases. Perhaps more concerning, LLMs have occasionally produced directly harmful content. Finally, the complexity of the underlying model makes it extraordinarily difficult to understand why a certain problematic output was produced, or to identify and correct the underlying training instance or model parameter that originated it (Hutson, 2021).

In conclusion, LLMs have heralded a new age in AI technology, with potentially significant benefits and risks for society and humanity as a whole. The balance between those two dimensions will depend on appropriate and responsible development and risk mitigation. This may be achieved by instilling common sense, causal reasoning, or moral judgement during model training, ensuring transparent model development, improving training data to remove biased content, or actively tracking algorithm usage and performance (Hutson, 2021).

References:

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ⍰', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency*, Virtual Event Canada: ACM, pp. 610–623. Available from: https://doi.org/10.1145/3442188.3445922.

Cui, J. et al. (2023) 'Chatlaw: A Multi-Agent Collaborative Legal Assistant with Knowledge Graph Enhanced Mixture-of-Experts Large Language Model'. arXiv. Available from: https://doi.org/10.48550/ARXIV.2306.16092.

Gosden, E. (2023) 'AI is better than people, warns Octopus Energy boss Greg Jackson', *The Times*, 8 May. Available from: https://www.thetimes.com/business-money/technology/article/ai-is-better-than-people-warns-octopus-energy-boss-greg-jackson-bzbhjc6vm (Accessed: 27 June 2024).

Hicks, M.T., Humphries, J. & Slater, J. (2024) ‘ChatGPT is bullshit’, *Ethics and Information Technology*, 26(2), p. 38. Available from: https://doi.org/10.1007/s10676-024-09775-5.

Hubert, K.F., Awa, K.N. & Zabelina, D.L. (2024) ‘The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks’, *Scientific Reports*, 14(1), p. 3440. Available from: https://doi.org/10.1038/s41598-024-53303-w.

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.

Musumeci, E., Brienza, M., Suriani, V., Nardi, D. & Bloisi, D.D. (2024) ‘LLM Based Multi-agent Generation of Semi-structured Documents from Semantic Templates in the Public Administration Domain’, in H. Degen and S. Ntoa (eds) *Artificial Intelligence in HCI*. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), pp. 98–117. Available from: https://doi.org/10.1007/978-3-031-60615-1_7.