COLLABORATIVE DISCUSSION 2:
LEGAL AND ETHICAL VIEWS ON ARTIFICIAL NEURAL NETWORK (ANN) APPLICATIONS

Maria Ingold
12693772

Unit 9
Machine Learning
University of Essex Online
23 December 2024

CONTENTS

MY PEER RESPONSES TO OTHERS

Peer Response: To Martyna Antas (2024)

Martyna, your initial post on Hutson's (2021) insights into the impact of Generative Pretrained Transformer (GPT) on writing is insightful, especially with its practical application across healthcare.

Your use cases of chatbots, routine communications, marketing, and localisation are applicable across a wide range of industries and benefit from your academic references. Staying with the healthcare theme for the risks provides a continuity to your story with Hutson's (2021) illustration of GPT suggesting suicide.

As a small piece of feedback, please define your acronyms at initial use. You use both Artificial Intelligence (AI) and GPT without first defining them. While I would like to see the bias issue in AI writing tied back to healthcare as well, raising it is essential, but I would find a way to transition your writing to note that this applies more generally than just healthcare.

Your point on reputation is valid and widely applicable. You may wish to raise the impact of hallucination on validity and the need for human review (Liu et al., 2024). Finally, I think the insight into skill erosion is one worth exploring further. How does it erode skills and what could be the impact, for instance.

Overall, I am impressed with the insights of your initial draft and look forward to seeing the summary post.

References

Hutson, M. (2021) ‘Robo-writers: the rise and risks of language-generating AI’, *Nature*, 591(7848), pp. 22–25. DOI: https://doi.org/10.1038/D41586-021-00530-0.

Liu, F., Liu, Y., Shi, L., Huang, H., Wang, R., Yang, Z., Zhang, L., Li, Z. & Ma, Y. (2024) ‘Exploring and Evaluating Hallucinations in LLM-Powered Code Generation’. Available from: https://arxiv.org/abs/2404.00971v2 (Accessed: 22 December 2024).

Peer Response: To Ben Zapka (2024)

Ben, thank you again for being so prompt with your initial post on Hutson’s (2021) insights into using large language models (LLMs) for writing across a range of use cases.

While you mention that Hutson referenced prompt engineering, in 2021 it was not the field that it is in 2024, and it would be interesting to see references on how the prompt influences the writing style. Your examples are good, however, highlighting idea generation, initial drafts, content structuring, and editing. You could add more on creative writing, such as Wafa et al.’s (2024) research into science fiction use of Generative Pre-Trained Transformer (GPT) by Algerian Masters’ students.

Your points on hallucinations, bias, and lack of guardrails are all valid, especially keeping a human in the loop to review created content. Lack of critical thinking and authenticity seems to tally with my research showing that the AP English Language and AP English Lit scores are still performing at 58% and 68% respectively, despite advances in chain-of-thought reasoning in the OpenAI o1 language model (OpenAI, 2024; Zhong et al., 2024).

What I see in industry concurs with your insight that brainstorming or ideation are a key benefit, but the lack of critical thinking, context, hallucinations, and bias are problematic in fields requiring accuracy like news, or as you say, administrative tasks (Sun et al., 2024). Using human review is also something I see in practice in industry. However, I would like to see more supporting research on your creative writing insights in your summary, as that will lend greater credibility to your prediction that GPT would be more effective in that instance.

References

Hutson, M. (2021) ‘Robo-writers: the rise and risks of language-generating AI’, *Nature*, 591(7848), pp. 22–25. DOI: https://doi.org/10.1038/D41586-021-00530-0.

OpenAI (2024) *Learning to Reason with LLMs*. Available from: https://openai.com/index/learning-to-reason-with-llms/ (Accessed: 22 December 2024).

Sun, Y., Tech, V., Cui, L., Lei, S. & Lu, C.-T. (2024) ‘Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges’. Available from: https://arxiv.org/abs/2403.18249v2 (Accessed: 22 December 2024).

Wafa, N., Samia, M., El Amine, H.M. & Fatiha, K.B. (2024) ‘View of Innovating Narratives Or Stifling Creativity? Assessing The Impact Of Generative Pre-Trained Transformer (GPT) On Science Fiction Writing Skills Among Master Students In Algeria’, *Educational Administration: Theory and Practice*, 6(30), pp. 1494–1507. Available from: https://www.kuey.net/index.php/kuey/article/view/5526/3879 (Accessed: 22 December 2024).

Zhong, T., Liu, Z., Pan, Y., Zhang, Y., Zhou, Y., Liang, S., Wu, Z., Lyu, Y., Shu, P., Yu, X., Cao, C., Jiang, H., Chen, H., Li, Y., Chen, Junhao, Hu, H., Liu, Y., Zhao, Huaqin, Xu, S., Dai, H., Zhao, L., Zhang, R., Zhao, W., Yang, Z., Chen, Jingyuan, Wang, P., Ruan, W., Wang, H., Zhao, Huan, Zhang, J., Ren, Yiming, Qin, S., Chen, T., Li, J., Zidan, A.H., Jahin, A., Chen, M., Xia, S., Holmes, J., Zhuang, Y., Wang, J., Xu, B., Xia, W., Yu, J., Tang, K., Yang, Y., Sun, B., Yang, T., Lu, G., Wang, X., Chai, L., Li, H., Lu, J., Sun, L., Zhang, X., Ge, B., Hu, X., Zhang, Lian, Zhou, H., Zhang, Lu, Zhang, S., Liu, N., Jiang, B., Kong, L., Xiang, Z., Ren, Yudan, Liu, J., Jiang, X., Bao, Y., Zhang, W., Li, X., Li, G., Liu, W., Shen, D., Sikora, A., Zhai, X., Zhu, D., Zhang, T. & Liu, T. (2024) ‘Evaluation of OpenAI o1: Opportunities and Challenges of AGI’. Available from: https://arxiv.org/abs/2409.18486v1 (Accessed: 22 December 2024).

PEER RESPONSES TO ME

Peer Response: From Linga Murthy Kanuri (2025)

Your evaluation of the large language models (LLMs), their scope and use cases, hallucinations, and biases related to them is relevant and valid.

The unveiling of OpenAI's o1 is a huge step forward in how artificial intelligence is seen and used. The o1 Model, deployed in December 2024, can solve complex mathematical equations, conceiving scientific solutions and programming. It achieves this by using the "Chain-of-thought" technique of solving problems, which translates into solving more complicated solutions using smaller avenues, thus increasing performance on various benchmarks (OpenAI, 2024).

One of the pitfalls of LLMs, including the o1 Model, is hallucination, the ability of a model to easily generate incorrect responses to prompts while making them seem plausible. This brings to light the need for a well-trained human with sufficient critical thinking skills to supervise content produced by AI. This is paramount, especially in research and news publication (Liu et al., 2024; Sun et al., 2024).

Prompt engineering and reinforcement learning from human feedback (RLHF) are in the most experimental stages, but they may prove helpful in easing hallucinations in the LLMs. Purposeful LLM implementations that include verifying facts in real additions that go through the Model can assist in substantively ensuring that information generated by the AI is accurate (Godofprompt.ai, 2024).

Various techniques are being researched to abate these biases, including adversarial training, bias detection algorithms, and diverse and representative training datasets. Besides, introducing model training processes with transparency will help discover and rectify biased output; regular audits should be mandatory (Fang et al., 2024).

While the benefits of AI in developments like the o1 Model are great, they equally come with problems that continue to raise significant current research and ethical issues.

In this respect, moving forward with AI technologies will require balancing innovation with responsibility to ensure these tools serve to augment human capabilities, not at an ethical compromise.

References:

1) OpenAI. (2024). OpenAI o1 System Card. Retrieved from https://cdn.openai.com/o1-system-card-20241205.pdf

2) Liu, X., et al. (2024). Mitigating LLM Hallucinations: A Multifaceted Approach. Retrieved from https://amatria.in/blog/hallucinations

3) Sun, T., et al. (2024). Tackling Hallucination in Large Language Models: A Survey of Cutting-Edge Techniques. Retrieved from https://www.unite.ai/tackling-hallucination-in-large-language-models-a-survey-of-cutting-edge-techniques/

4) Godofprompt.ai. (2024). 9 Prompt Engineering Methods to Reduce Hallucinations. Retrieved from https://www.godofprompt.ai/blog/9-prompt-engineering-methods-to-reduce-hallucinations-proven-tips

5) Fang, B., et al. (2024). Bias in Large Language Models: Causes and Mitigation Strategies. Retrieved from https://arxiv.org/abs/2412.16720

Peer Response: From James Adams (2025)

Hi Maria,

Your post provides an excellent description of some of the use cases and limitations for LLMs. I was particularly interested in the point that you described regarding the use of LLMs in creative pursuits, such as writing science fiction, a favourite genre of mine.

Understanding the practical applications and limitations is essential for the successful application of this technology in the workplace, an issue that I am exploring within the small business where I work, considering the suitability of different models for different tasks. (Yang et al., 2023) provide an excellent overview of some of the practical considerations of using LLMs in a real-world environment, and they also highlight some key themes such as bias, safety and reliability.

This feeds into a wider topic regarding the effective and knowledgeable use of AI in the workplace – referred to as AI literacy, a key principle in the EU AI Act 2023. Cetindamar et al. (2022) provide a useful taxonomy for defining AI literacy, breaking it down into 4 key areas: technology-related, work-related, human-machine-related, and learning-related. As individuals and organisations become more experienced with LLMs, I am hopeful that some of the implicit issues within the current generation of LLMs can be managed through implementation, so that people can see the benefits of using this technology safely.

I would be interested to hear your thoughts regarding a key challenge of the successful deployment of LLM technology in your business and industry.

Reference List

Cetindamar, D. et al. (2022) ‘Explicating AI literacy of employees at digital workplaces’, IEEE transactions on engineering management, pp. 1–14. Available at: https://doi.org/10.1109/tem.2021.3138503.

Yang, J. et al. (2023) ‘Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond’, arXiv [cs.CL]. Available at: https://doi.org/10.48550/ARXIV.2304.13712.