COLLABORATIVE DISCUSSION 2: LEGAL AND ETHICAL VIEWS ON ANN APPLICATIONS

(PEER RESPONSES)

Murthy Kanuri
Machine Learning
University of Essex

1 Response from Peers

1.1 Response from Georgios Papachristou

Linga in her post summarizes in a very comprehensive way Hutson's (2021) article, presenting the benefits and risks of AI writers. Highlighting the importance of responsible usage of AI, Linga concludes that such technologies can be used as complementary and not substitutes of human beings.

One area where Large Language Models (LLMs) may offer gains to human beings complementing their understanding of new concepts is critical thinking; "the ability to analyse and evaluate information". Alarcon-Lopez et al., (2024) studied the influence of ChatGPT on students' critical thinking and mainly on the dimensions of analysis, inference and explanation. Based on their results, the use of ChatGPT improved the students' sustainable energy literacy as well as their interest for using such technologies.

On the same page, van Rensburg (2024) found that ChatGPT could be used as a model for critical thinking skills, but highlighted that involvement of an educator is of critical importance at least to facilitate the process.

That being said, it is concluded that LLMs can be useful assistants in critical thinking development and in turn complement human-beings understanding of concepts that would help them take the best-informed decisions.

References:

Alarcon-Lopez, C., Krutli, P., & Gillet, D. (2024). Assessing ChatGPT's Influence on Critical Thinking in Sustainability Oriented Activities. *024 IEEE Global Engineering Education Conference (EDUCON)* (pp. 1-10). Kos Island, Greece: IEEE.

Hutson, M. (2021) Robo-writers: the rise and risks of language-generating AI. *Nature.* Available from: **https://www.nature.com/articles/d41586-021-00530-0** [Accessed 11.12.2024]

van Rensburg, J. (2024). Artifcial human thinking: ChatGPT's capacity to be a model for critical thinking when prompted with problem-based writing activities. *Discover Education, 3*(42).

2 My Response to Peers

2.1 Response to Georgios Papachristou

I agree with your thoughtful discussion on the potential and risks of AI writers like GPT-3, as highlighted by Hutson (2021). You have outlined how LLMs can be used in various fields, such as creative writing, education, and customer support. The fact that these AI tools can help with tasks like generating ideas, drafting content, and even offering constructive feedback is a valuable resource for many users. As you have pointed out, these tools can be especially useful in education to help create lectures and mock exams, which would then help improve the learning experience for students and tutors (Lund et al., 2023).

I could not agree more with your observation of the inherent risks with these AI tools, particularly their tendency to contain biases and deficiencies in common sense. Since many AI models produce content without knowing the proper contexts, as pointed out by Hutson (2021), it perpetuates harmful stereotypes and misinformation, especially on sensitive topics.
The other big concern is the risk of personal data being extracted from the training data of AI models, especially in the face of a surging increase in data breaches and privacy issues brought forth by Bender et al. (2019).

This discussion should be continued, as it reveals how we can mitigate these risks. You have rightly said that whatever the AI generates should be reviewed by a human before re-sharing or application, but what practical ways are there for responsible usage of these tools? Perhaps introducing strict ethical guidelines, including improving transparency about the data used to train these models and installing robust systems that detect and remove noxious outputs, would help.

Also, while AI frees humans from most of their workload, one must consider what this will do to human creativity and job displacement in the long run. How will we make AI a tool for enhancement and not a replacement?

Overall, your post balances the many benefits and risks associated with AI writing tools, and I appreciate the depth you brought to this discussion.

References

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S., 2019. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp.1-9.
- Hutson, M., 2021. Robo-writers: the rise and risks of language-generating AI. Nature, 591, pp.22-25.
- Lund, M. L., et al., 2023. The role of AI in enhancing educational experiences. International Journal of Education Technology, 15(3), pp.115-130.

2.2 Response to Maria Ingold

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 d iscover 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