

# LLM Workshop

# Ethics Case Studies

### Jayne Fitzgerald

### Catherine Breslin

### Ryan Daniels

2024

## Can LLMs outperform humans in some tasks?

As LLMs improve in capability, researchers are looking into how and when LLMs can outperform humans at specific tasks. Here are two examples where LLMs have been shown to perform as well as the best students in their qualifying exams.

**Medicine**: Researchers showed that OpenAI’s models are capable of successfully answering first- and second-year medical student exam questions. The exam questions were case-study based, where medical information is presented and then students must write free-text answers to a series of 2-7 questions. The LLM model’s answers were blindly scored by faculty members using the original exam marking rubric. With a small amount of prompt engineering, GPT-4 scored on average 4 points higher than students answering the same exam questions (Strong et al, 2023).

**Law**: Legal language is complex, and it can take years of study for law students to correctly interpret legal documents. In the US, passing ‘the bar exam’ is a prerequisite to practicing law. There are three components of the bar exam, one is multi-choice and two are essay based. Researchers collected questions from previous exams, for all three parts, and then had LLM answers rated by professionals. When combining results of the three parts, GPT-4 scored highly enough to pass the exam, which was approaching the 90th percentile of results when compared with scores from a recent administration of the exam (Katz et al, 2024).

|  |
| --- |
| **Discussion Questions**   * When and where should LLMs be used? * What are the limitations and risks of using LLMs in a field like medicine or law? * What impact will LLMs have on education? * If an LLM passes an exam, is it as qualified as a person who also passed? * How should we evaluate LLMs? |

### References

Katz, D.M., Bommarito, M.J., Gao, S. and Arredondo, P., 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the* Royal Society A, 382(2270), p.20230254.

Strong E, DiGiammarino A, Weng Y, et al. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. JAMA Intern Med. 2023;183(9):1028–1030

## Bias and the use of LLMs for Medical Care

Researchers across disciplines are finding ways to use LLMs in their work. Successful academic research informs the commercial products and models that get built and deployed or open-sourced. Once models are publicly released, they then go on to form the foundation of the next iteration of research.

In medicine, for example, there has been tremendous research effort dedicated to LLMs for medical tasks, and success in the research leads to real-world implementations such as Google’s Med-Gemini – a family of LLMs specialised to medicine (Saab et al, 2024).

One known limitation of LLMs concerns bias. That is, they reflect the societal biases of the data they were trained on. A recent study looked at the racial biases present in four commercial LLM systems in the context of medicine. There are several misconceptions about the effect of race on medical treatment, which have been debunked but still persist in the medical field. For example, the misconception that there are different ways to calculate lung capacity based on race. In this study, researchers showed that four commercial LLMs were able to propagate inaccurate race-based content when responding to scenarios that were explicitly designed to check for race-based misconceptions in medicine. Additionally, the models were inconsistent in their responses when asked the same question multiple times, making it harder to detect (Omive et al, 2023).

|  |
| --- |
| **Discussion Questions**   * Why are the LLMs showing bias? * Could this bias have been predicted before deployment? What tests and measures should have been put in place? * Whose responsibility is it to conduct these tests when using LLMs in research? * Should you use technology in research that has a known bias? |

### References

Omiye, J.A., Lester, J.C., Spichak, S. et al. Large language models propagate race-based medicine. npj Digit. Med. 6, 195 (2023).

Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E. and Chaves, J.Z., 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint arXiv:2404.18416*.

## LLMs for Academic Writing

The rise of LLMs has led to speculation about how models like ChatGPT are being used in academic writing. Recent analysis shows that since the launch of ChatGPT there’s been a steady increase in LLM generated content in academic papers. In Computer Science for example, an estimated 17.5% of papers on Arxiv now use text from LLMs (Liang et al, 2024). As AI is integrated into writing tools like Microsoft Word and Google Docs, it becomes harder and harder to avoid. However, LLM tools can introduce errors and inaccuracies. Journals and conferences have taken different stances on the use of LLMs in academic writing:

1. The journal ***Science*** has warned researchers that submitting manuscripts that have been produced using AI tools amounts to scientific misconduct (Holden, 2023)
2. The journal ***Nature*** has declared that it will not accept any papers listing ChatGPT or any other AI software as author, but hasn’t banned these types of tool completely.
3. The publisher ***Elsevier*** permits using LLMs “*to improve the readability and language of the research article, but not to replace key tasks that should be done by the authors, such as interpreting data or drawing scientific conclusions*” (Elsevier 2023)
4. The conference ***ICML*** has prohibited the use of generated text unless part of the paper’s experimental analysis (ICML, 2023)

|  |
| --- |
| **Discussion Questions**   * What scientific research tasks (if any) are acceptable to complete with an LLM? * Why are some tasks acceptable while others are not? Does this differ across disciplines? * How much output from an LLM is appropriate to use in research writing? * Are there other parts of the research process where LLMs are permissible? * Can AI ever be an author? |

### References

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z. and Yang, D., 2024. Mapping the increasing use of llms in scientific papers. arXiv preprint arXiv:2404.01268.

ICML.Clarification on large language model policy LLM.<https://icml.cc/Conferences/2023/llm-policy>, 2023.

H. Holden Thorp.Chatgpt is fun, but not an author.Science, 379(6630):313–313, 2023.doi: [10.1126/science.adg7879](https://arxiv.org/html/10.1126/science.adg7879).URL <https://www.science.org/doi/abs/10.1126/science.adg7879>.

Nature, Nature Editorial Policies (AI), [https://www.nature.com/nature-portfolio/editorial-policies/ai](https://www.nature.iciescom/nature-portfolio/editorial-policies/ai)

Elsevier, The use of generative AI and AI-assisted technologies in writing for Elsevier, 2023, <https://www.elsevier.com/en-gb/about/policies-and-standards/the-use-of-generative-ai-and-ai-assisted-technologies-in-writing-for-elsevier>