How well can large language models perform literature reviews?

Prof. David Schultz
Department of Earth and Environmental Science
University of Manchester
david.schultz@manchester.ac.uk

Could a large-language model (LLM) research and write a scientific paper? Judging by the many LLMs being marketed to researchers, the answer appears to be yes. But, how effective are they? LLMs can accelerate scholarly writing (idea generation, outlining, language polishing), but they also introduce risks including hallucinated citations, bias, confidentiality breaches, and uneven impacts on non-native English speakers. This proposed project aims to address this question, the answer to which will be of global interest to students, academics, and researchers.

We will test several LLMs (e.g., OpenAI's Prism, Google's Notebook LLM, Microsoft CoPilot, Elicit, Silvi), both their free and paid versions, and compare their outputs to systematic literature reviews coauthored by Schultz on the Spanish plume weather pattern and the effect of weather on pain for those living with chronic pain. We will evaluate the LLM outputs against the exemplars against the completeness against inclusion/exclusion criteria, ability to synthesize across the literature and gain new perspectives, and critique of past work. Our project aims to provide students, educators, and researchers with clear, evidence-based guidance on when, or even whether, LLMs can effectively research and write a literature review. As such, we expect a high-quality and high-visibility peer-reviewed journal article in a high-impact journal.

The scientific results of the proposed research will guide us toward future research questions, such as quantifying the benefits and harms of LLM-assisted writing across different manuscript tasks and examining equity and integrity impacts on non-native English authors.

We require access to GPUs on the Bede supercomputer to perform this task. The workload involves LLM inference, which relies heavily on GPU acceleration. Standard CPUs are not suitable for handling this type of computation efficiently, so GPU resources (e.g., on Bede) are necessary for our experiments.