The Promise and Perils of Autonomous AI in Science

Junyi Gao, Ewen M Harrison

Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, EH16 4UX, UK

Ewen.harrison@ed.ac.uk

In Isaac Asimov's 1956 short story, "The Last Question", a self-correcting computer called Multivac is fed data and slowly learns to answer fundamental questions of the universe. To be fair, it takes until the end of said universe to answer the most fundamental of all, but it successfully tackles everything else along the way.

In this issue of *NEJM AI*, Ifargan and colleagues present *data-to-paper*, an autonomous platform that seeks to mimic human scientific practice by guiding large language model (LLM) agents through a complete, stepwise research process, much as Asimov might have imagined. That LLMs are being used to write scientific papers should come as no surprise. Indeed, it has been estimated that at least 10% of research may already be co-authored by LLMs.[1] The current study is notable however in its attempts to comprehensively automate the end-to-end process.

How does do this? The platform runs either in autopilot, performing the entire research cycle with no human intervention, or in copilot mode, where human users interact to provide guidance and corrections. The extensive documentation that accompanies the paper shows that it can make hypotheses, design research plans, write and debug analysis code, generate and interpret results, and compose final manuscripts section by section. Two broad applications are demonstrated. In *open-goal research*, simple publicly available datasets are used and research papers autonomously produced, using self-defined hypotheses and analysis plans. In *reproducing peer-reviewed papers*, more complex datasets are provided, together with the research goals of the original studies.

As an endeavour, discovery science is challenging, requiring abstract non-linear thought, rigorous iterative methodologies, scepticism and an ability to handle uncertainty, together with cooperative approaches to feedback and criticism. Since LLMs (likely[2]) lack a meaningful world model – especially with regard to biology, social constructs, and human health – it would be surprising if they were genuinely capable of 'doing' health science.

So, are results useful? The papers certainly look like actual research – which though a low bar – is impressive in itself. When provided with simple datasets and running in open goal/autopilot mode, the authors report 8 out of 10 papers with no major errors.

With the introduction of a complex dataset, significant errors occurred in the absence of human intervention, making human co-piloting essential for ensuring accuracy.

In our own tests, we provided the model with a publicly available COVID-19 dataset.[3] Although consisting of only a single table, it is reasonably complex and comparable to a real-world example. The model ran smoothly, although got stuck while correcting the data analysis code. It explored the relationship between mortality and variables such as age, gender, and prothrombin time. It built a logistic regression model and reported associations, eventually generating a 7-page paper with two tables, one figure, and 18-page appendix of all intermediate results. Analytical errors occurred, but when pointed in the human review phase, correct analysis code and results were successfully produced.

The authors should be congratulated for their innovative work. The use of interacting agents is particularly interesting – output is passed back and forth between "performer" and "reviewer" LLM agents, facilitating iterative corrections and improvements – an approach which holds broad potential in health and beyond.[4] Keeping a "human-in-the-loop" is clearly important and desirable, ensuring accuracy and quality, particularly for more complex research goals or datasets.

The data chaining solution is equally impressive. In the generated PDF, each numerical result in the text links to its corresponding table, which in turn, links to the statistical output that produced it, and ultimately traces back to the Python code that generated it. In addition, the entire end-to-end process is saved in an annotated file which can be examined or rerun (noting the random characteristics of LLMs mean no two papers are identical). Rather than being an opaque black box, here is something which reflects the transparency goals of open-science more effectively than many human-generated papers.

There are still many areas for improvement. For instance, can we enhance the inference capabilities of LLMs to generate more valuable research questions? Could LLMs effectively incorporate widely used, human-verified clinical analysis tools to perform more sophisticated analyses? How can we improve the precision of literature retrieval to support more accurate evaluations, thus enhancing the reliability and rigor of discussions? These are ambitious goals in advancing the role of artificial intelligence as an assistant to researchers.

There will rightly be focus on the negatives, and these studies should serve as catalysts for important conversations around the implications of such technologies. The risk of error is high, particularly with when dealing with complex datasets or research goals. The heavy reliance on human oversight will limit promised short-term efficiency gains – checking for coherence and mistakes likely represents more work for the human than simply completing the work themselves.

The ease of generating papers will surely lead to an influx of low-quality or generic manuscripts, overwhelming a publication and peer review system which is already broken. P-hacking is a real risk, particularly given the widespread use of reward mechanisms in AI training. Fabrication or manipulation of results may also occur (is deep-fake science a thing yet?), potentially with plausible deniability on the part of a human overseeing a complex system only partly understood. Ethical concerns arise in many areas, but particularly around the conduct of research without genuine human intellectual contribution and guarantees will always be needed around the integrity, accountability, and responsibility for the scientific output. Using AI-generated content watermarked with an identifiable signature could help track its output, as is suggested by the authors.

As would be expected given the architecture of LLMs, much of the writing is “vanilla” and novelty was absent in most of the output, although *de novo* insights were reported to have been made. Yet, in an era where reproducibility of scientific research is under scrutiny and arguably as great a priority as novel discovery, could AI be the impartial auditor science needs? Rather than crafting new theories, AI's true potential may lie in its ability to reproduce and verify scientific findings from raw data, bridging the reproducibility gap that challenges modern research.

We see three important conclusions arising from this work. First, there is a clear need for better guidelines and standards on the ethical use of AI to produce research, including around issues of authorship, accountability, and the prevention of misconduct. Second, it highlights the importance of promoting human-AI collaboration where platforms augment human researchers, assisting with the grind of lower-level tasks while preserving the essential roles of human judgment and creativity. Third, there is an opportunity to adopt innovative approaches such as data chaining into the wider scientific endeavour, enhancing transparency, traceability, and reproducibility.

Perhaps, as Asimov foresaw, the answers may come in time, but only if we remain careful custodians of our technological creations.

1. Kobak D, González-Márquez R, Horvát E-Á, Lause J. Delving into ChatGPT usage in academic writing through excess vocabulary [Internet]. 2024 [cited 2024 Oct 29];Available from: http://arxiv.org/abs/2406.07016

2. Li K, Hopkins AK, Bau D, Viégas F, Pfister H, Wattenberg M. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task [Internet]. 2024 [cited 2024 Oct 29];Available from: http://arxiv.org/abs/2210.13382

3. Lab (HAIL) HAI. HAIRLAB/Pre_Surv_COVID_19: 2.0 [Internet]. 2020 [cited 2024 Oct 31];Available from: https://zenodo.org/records/3766350

4. Freyer O, Wiest IC, Kather JN, Gilbert S. A future role for health applications of large language models depends on regulators enforcing safety standards. The Lancet Digital Health 2024;6(9):e662–72.