Qeios

Peer Review

Review of: "PhysAgent: A Multi-Agent Approach to the Automated Discovery of Physical Laws"

Navid Ghaffarzadegan[1]

1. Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, United States

The paper aims to automate the process of scientific discovery by designing agents that represent a student and a mentor, creating interactions that lead to hypothesis generation and analysis. The authors test their framework on Kepler's and Newton's laws by generating synthetic data consistent with these laws and providing it to the program for rediscovery.

In this formulation, "automating discovery of a law" is understood as automating the interaction between student and mentor. The underlying assumption is that if such a process is modeled effectively, a scientific law can be uncovered. Thus, the challenge is framed primarily as one of simulating the "social process" of discovery. This is indeed an important part of scientific practice, connecting to broader issues in "team science." At the same time, in the real world, discovery can fail for many reasons: researchers may be working within the wrong framework, neglecting to question deep assumptions and paradigms, focusing on insufficient indicators, or simply missing a breakthrough that might emerge only through the recombination of prior ideas and new ideas. I am not necessarily criticizing the paper or asking for unrealistic expansions—the authors clearly state that their goal is only to automate the interaction process—but this limitation should be more explicitly acknowledged in the paper.

I am also not convinced that the paper presents strong empirical tests. While the use of synthetic data is acceptable, there is no particular reason to base it on established physical laws. A more rigorous evaluation would also be to make up/assume an artificial law (not corresponding to any known physical principle; something that is wrong in our universe), generate synthetic data from that made-up law, and then test whether the system can correctly identify that law within this made-up universe. Such an approach would more convincingly demonstrate that the model goes beyond leveraging preexisting scientific knowledge in the LLM. It should also show that the model is not just about uncovering

 doi.org/10.32388/4GRTNC

correlations which could have been done by regression models. So some of these laws can deal with more complicated non-linear patterns and relationships.

Additional robustness checks could include providing data that does not correspond to any law (a placebo test) to examine whether the software erroneously “discovers” a spurious law.

Finally, the evaluation should better approximate realistic empirical tests. Specifically, the authors should introduce measurement noise. A systematic analysis could then be performed by gradually increasing the noise (e.g., via the standard deviation) until the software fails to recover the law, and reporting the extent to which the method remains reliable under noisy data.

I do not intend to expand the scope of this paper too much, and I recognize that the authors are primarily offering illustrative examples. However, incorporating at least a small set of tests designed to “break” the software would strengthen the paper considerably and is itself in line with the scientific method.

Declarations

Potential competing interests: No potential competing interests to declare.