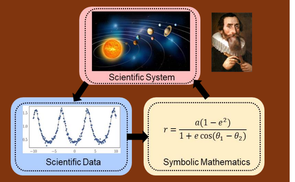
LLM-SRBench: A New Benchmark for Scientific Equation Discovery with LLMs

Parshin Shojaee*[1], Ngoc-Hieu Nguyen*[2], Kazem Meidani[3,4], Amir Barati Farimani[3], Khoa D Doan[2], Chandan K Reddy[1]

[1] Virginia Tech [2] VinUniversity [3] Carnegie Mellon University [4] Capital One *Equal contribution

🎯 Takeaway:

Existing benchmarks for scientific discovery with LLMs often **suffer from memorization bias**, misleading evaluation of the discovery capabilities.

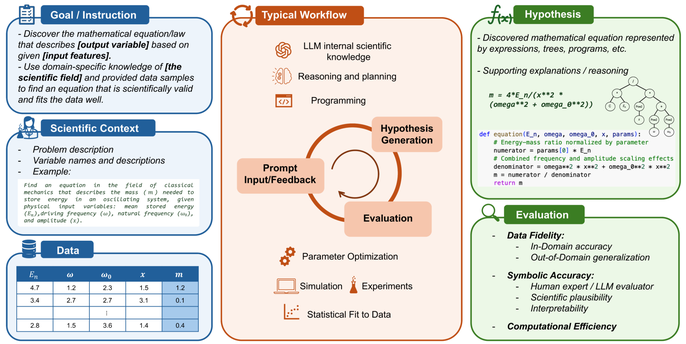


LLM-SRBench is a holistic benchmark for **scientific equation discovery with LLMs** that provides **rigorous evaluation towards driving innovation beyond memorizing known solutions**.

Scientific Equation Discovery & LLMs

Math as Language of Science: *Finding mathematical models and equations underlying scientific observations has been one of the fundamental tasks in the history of scientific discovery.*

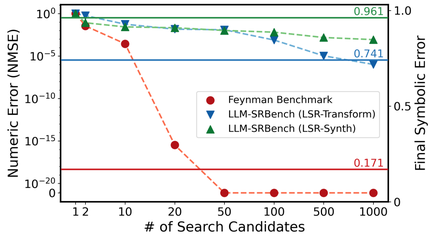
LLMs, with their vast scientific knowledge, show great promise for this task



Memorization Bias

However, current benchmarks often rely on well-known textbook scientific laws for evaluation, which are mostly memorized by LLMs.

Sharp drops in error curves and **low symbolic error** on Feynman problems suggest memorization rather than gradual exploration towards discovery.

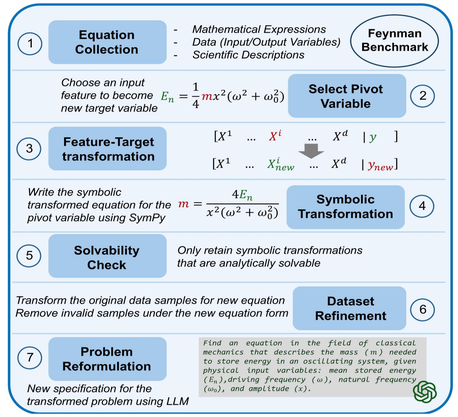


LLM-SRBench (Datasets)

Our benchmark includes two main categories of datasets:

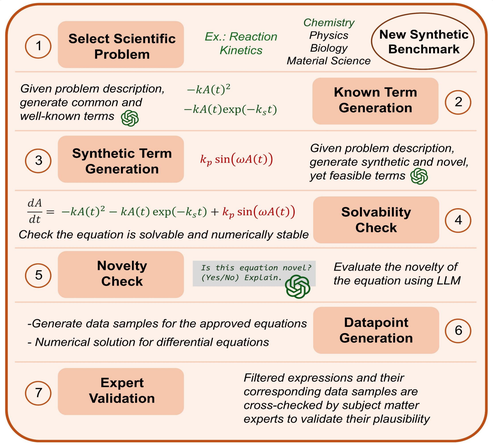
LSR-Transform: *transforms common physical models into less familiar mathematical representations*

111 problems from 100 Feynman problems



LSR-Synth: *introduces novel synthetic components into scientific models, requiring data-driven reasoning*

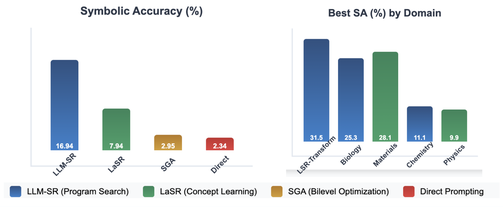
128 problems across 4 domains (Chemistry, Biology, Physics, Materials)



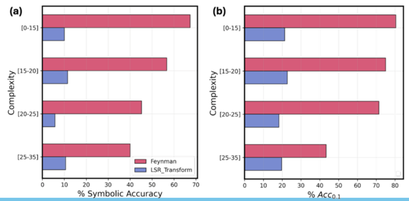
Results

- Best performing method achieves symbolic accuracy of at most **31.5%** on LSR-Transform and **28.1%** on LSR-Synth, highlighting challenging nature of scientific equation discovery!

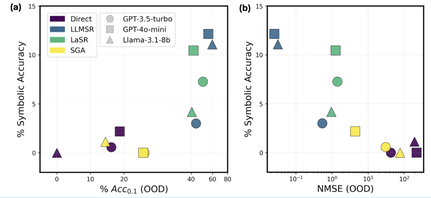
Symbolic Accuracy (methods & domains)



Different Performance at Same Complexity



Symbolic Accuracy ↔ OOD Performance



Code on GitHub: