
SAFETY, EVALUATION AND ALIGNMENT JANUARY 15, 2026

SCIPREDICT: CAN LLMs PREDICT THE OUTCOMES OF RESEARCH EXPERIMENTS IN NATURAL SCIENCES?

Udari Madhushani Sehwag, Elaine Lau, Haniyeh Ehsani Oskouie, Shayan Shabihi, Erich Liang, Andrea Toledo, Guillermo Mangialardi, Sergio Fonrouge, Ed-Yeremai Hernández Cardona, Paula Vergara, Utkarsh Tyagi, Chen Bo Calvin Zhang, Pavi Bhattar, Nicholas Johnson, Furong Huang, Ernesto Gabriel Hernández Montoya, Bing Liu

SciPredict: Can LLMs Predict the Outcomes of Research Experiments in Natural Sciences?

Accelerating scientific progress depends on developing and efficiently allocating resources towards the most promising research directions. In experimental sciences, this often means predicting which experiments will yield meaningful results before committing to costly physical validation. Although existing benchmarks evaluate AI systems on knowledge recall, simulated environments, or theoretical reasoning, assessing their ability to predict outcomes of practical experiments remains underexplored. We introduce SciPredict, a benchmark evaluating whether we can rely on current AI systems to predict experimental outcomes in three key domains: physics, biology, and chemistry. The benchmark comprises of 405 questions derived from recently published empirical studies (post-March 2025), which spans 33 subdomains, requiring models to reason about real experimental systems. Unlike most benchmarks that assess whether AI has reached human-level performance, experimental outcome prediction represents a domain where AI systems could substantially exceed human capabilities, integrating vast cross-domain knowledge, processing complex parameter interactions, and identifying non-obvious patterns that individual researchers cannot readily perceive. This raises two critical questions: can models predict experimental outcomes with sufficient accuracy? and can we identify which predictions are trustworthy? Our analysis reveals fundamental limitations on both fronts. Our evaluations on frontier models show that models accuracy ranges between 14%–26% and accuracy of human domain experts is H 20%. Although some frontier models exceed human performance model accuracy is still far below what would enable reliable experimental guidance. Second, even within this limited performance, models cannot distinguish reliable predictions from unreliable ones. Models only achieve H 20% accuracy even when they self-report very high confidence in their answer and high feasibility in question (i.e., perceiving as it is highly feasible to predict the outcome without running the practical experiment). In contrast, human experts demonstrate strong calibration: the accuracy of human experts increases as they get more confident in their answers and accuracy increases from H 5% on questions they judge infeasible to H 80% on questions they consider feasible to answer without experimentation. Our findings demonstrate that while frontier models are comparable to human experts in raw predictive accuracy, they fundamentally lack the calibration awareness required for reliable deployment in experimental planning. SciPredict establishes a rigorous evaluation framework for experimental outcome prediction and demonstrates that achieving superhuman performance in experimental science requires not just better predictions, but better awareness of prediction reliability

LINKS

<https://arxiv.org/pdf/2604.10718>