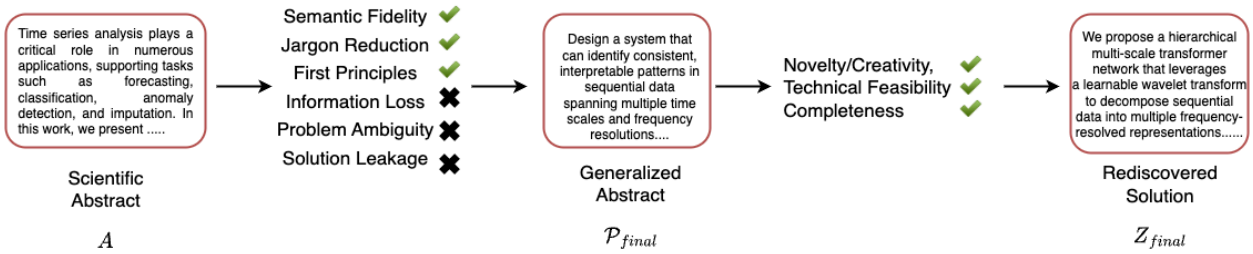
AInstein: Can AI Rediscover Scientific Concepts from First Principles?



Large language models (LLMs) have shown remarkable aptitude for scientific tasks, but it remains unclear whether this stems from genuine reasoning or sophisticated memorization. We introduce AInstein, a novel framework designed to test if LLMs can rediscover established scientific concepts from first principles. By abstracting research work into core conceptual problems, stripped of domain-specific jargon, we challenge models to solve the problem. Our framework, AInstein, operates by taking a scientific discovery, distilling it to its fundamental problem, and tasking an LLM with solving it from scratch.

The AInstein Framework

We formalize scientific rediscovery as a two-phase, multi-agent process.

1. **Generalization Phase**: A Generalizer agent ($\mathscr{G}$) takes a scientific abstract ($\mathscr{A}$) and, through iterative refinement, produces a generalized problem statement ($\mathscr{P}$) free of technical jargon and solution hints.
2. **Solution Phase**: A Solver agent ($\mathscr{S}$) receives the problem ($\mathscr{P}$) and attempts to derive a technical solution ($\mathscr{Z}$) from first principles.

Both phases are driven by an iterative refinement loop involving two models: a generative internal model ($\mathscr{M}_{\text{i}}$) and a more capable external critique model ($\mathscr{M}_{\text{e}}$). This nested critique process continues until a high-quality, converged solution is produced, simulating a rigorous scientific research process.

Experiments and Key Findings

We evaluated our framework using various LLMs as internal and external models across a dataset of scientific abstracts [1]. Performance was measured using an LLM-as-a-Judge on a 5-point scale, with a score ≥ 3 considered a success. Our primary metric, the **Rediscovery Rate (R)**, assesses whether the final solution $\mathscr{Z}$ successfully recovers the core insight of the original abstract $\mathscr{A}$. We also report the **SR-Solver** score, which measures problem-solution alignment.

Model Configuration		R [%]	SR–Solver [%]
External Model	**Internal Model**		
GPT-OSS 120B	Gemma-3-27B	49.92	62.23
	Phi-4 Reasoning+	82.56	90.00
	Qwen3-235B	53.92	85.22
	GPT-OSS 120B	82.37	**95.47**
Phi-4 Reasoning+	Phi-4 Reasoning+	82.48	90.91
	Mistral-24B	61.74	59.83
	Gemma-3-27B	68.68	88.60
	GPT-OSS 120B	**88.35**	91.90
Mistral-24B	Phi-4 Reasoning+	62.56	37.27
	Mistral-24B	76.69	87.69
Qwen3-235B	Qwen3-235B	62.31	91.57
	GPT-OSS 120B	74.85	90.04
Gemma-3-27B	Gemma-3-27B	35.12	45.70
	GPT-OSS 120B	83.97	84.21
	Phi-4 Reasoning+	81.40	87.19

Our analysis revealed two distinct modes of successful rediscovery: (1) Conceptual Convergence, where the agent independently arrived at a solution functionally equivalent to the original work, and (2) Valid but Novel Solutions, where the agent proposed creative, conceptually sound alternatives that differed from the original paper's method. This showcases reasoning that goes beyond simple retrieval.

The AInstein framework provides strong evidence that LLMs possess scientific reasoning capabilities that extend beyond memorization. Our results demonstrate that models can perform genuine conceptual rediscovery.

[1] González-Márquez & Kobak, Learning representations of learning representations, DMLR workshop at ICLR 2024 (arXiv 2404.08403)