When Tools Help and Hurt Scientific Simulator Agents:

A PHREEQC-MCQ-200 case study

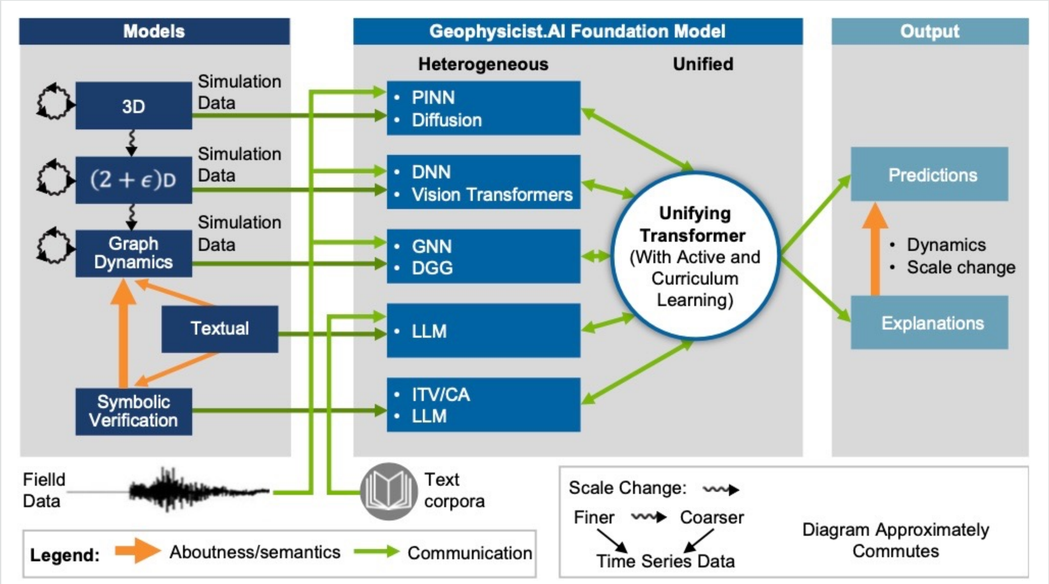
Ke Zhang, Sahchit Chundur, Jerry Peprah Owusu, MJ Qomi, Maziar Raissi
University of California, Riverside/Irvine

Geophysicist.AI

UNIVERSITY OF CALIFORNIA Office of the President

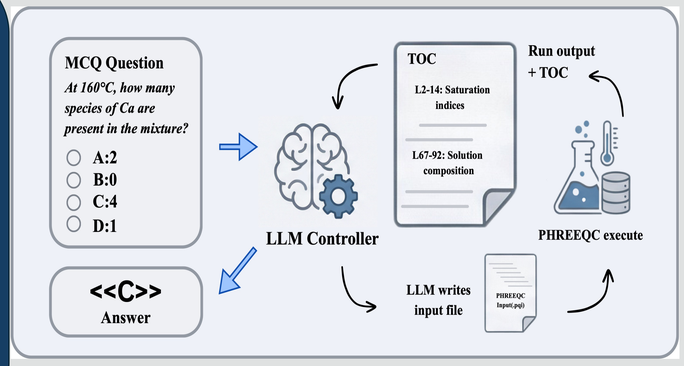
Benchmark and Agent Harness

Do tools make scientific LLM agents more reliable — or just more complex? PHREEQC is a widely used, open-source geochemistry simulator — it models aqueous speciation, chemical reactions, and reactive transport. We equip state-of-the-art LLMs with PHREEQC as an external simulator tool and test whether tool access improves scientific question answering.



Agent Workflow and results

Question: Simulate the equilibrium oxidation of Pyrite in 1 kg of pure water at 25°C. The system starts at a neutral pH (7.0) and is exposed to a constant atmospheric oxygen source (log PO2=−0.7). Allow up to 0.1 moles of Pyrite to react and the potential precipitation of Goethite and Gypsum. What is the molality of S (6) in the solution?
A) 2.008e-01, B) 2.006e-01, C) 2.003e-0, D) 2.004e-01



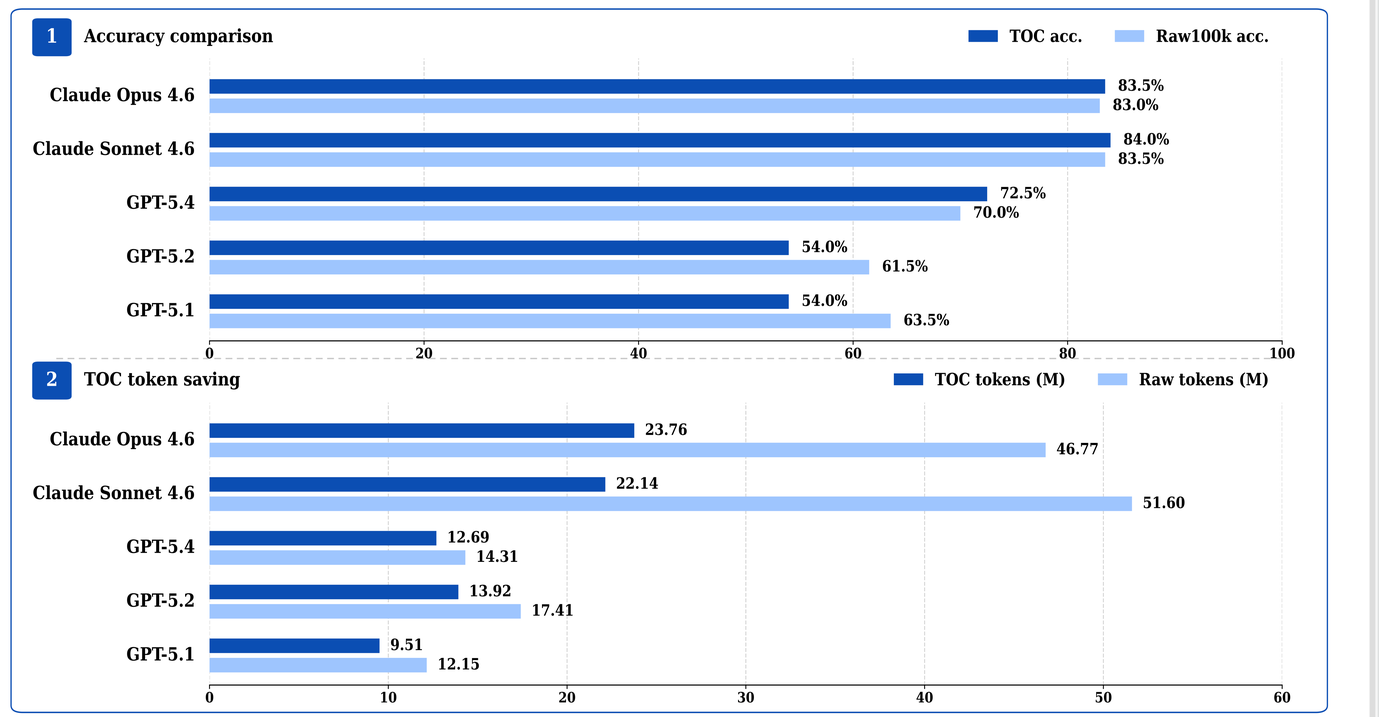
Tool access helps; CoT is not enough.

Tier	Model	Direct	CoT	Agent-TOC	Agent-Raw100k	Tool gain
Top	Claude Opus 4.6	42.0%	38.0%	83.5%	83.0%	+41.5
Top	Claude Sonnet 4.6	37.0%	43.0%	84.0%	83.5%	+41.0
Top	GPT-5.4	27.0%	35.0%	72.5%	70.0%	+37.5
Mid	GPT-5.2	27.5%	34.5%	54.0%	61.5%	+19.5
Mid	GPT-5.1	39.0%	35.0%	54.0%	63.5%	+15.0
Below-mid	Gemini 3 Flash	38.5%	40.5%	23.5%	47.0%	-17.0

Tool use is not purely additive.

Model	Kept	Gained	Lost	Retention
Claude Opus 4.6	70	97	14	83.3%
Claude Sonnet 4.6	64	104	10	86.5%
GPT-5.4	35	110	19	64.8%
GPT-5.2	31	77	24	56.4%
GPT-5.1	46	62	32	59.0%

PHREEQC raw output is ”Huge”, we use Table of contents (TOC) to save token



UC Irvine · UC Riverside · UC Berkeley · UC Santa Cruz · UC San Diego · Berkeley Lab · Lawrence Livermore National Laboratory · Los Alamos National Laboratory