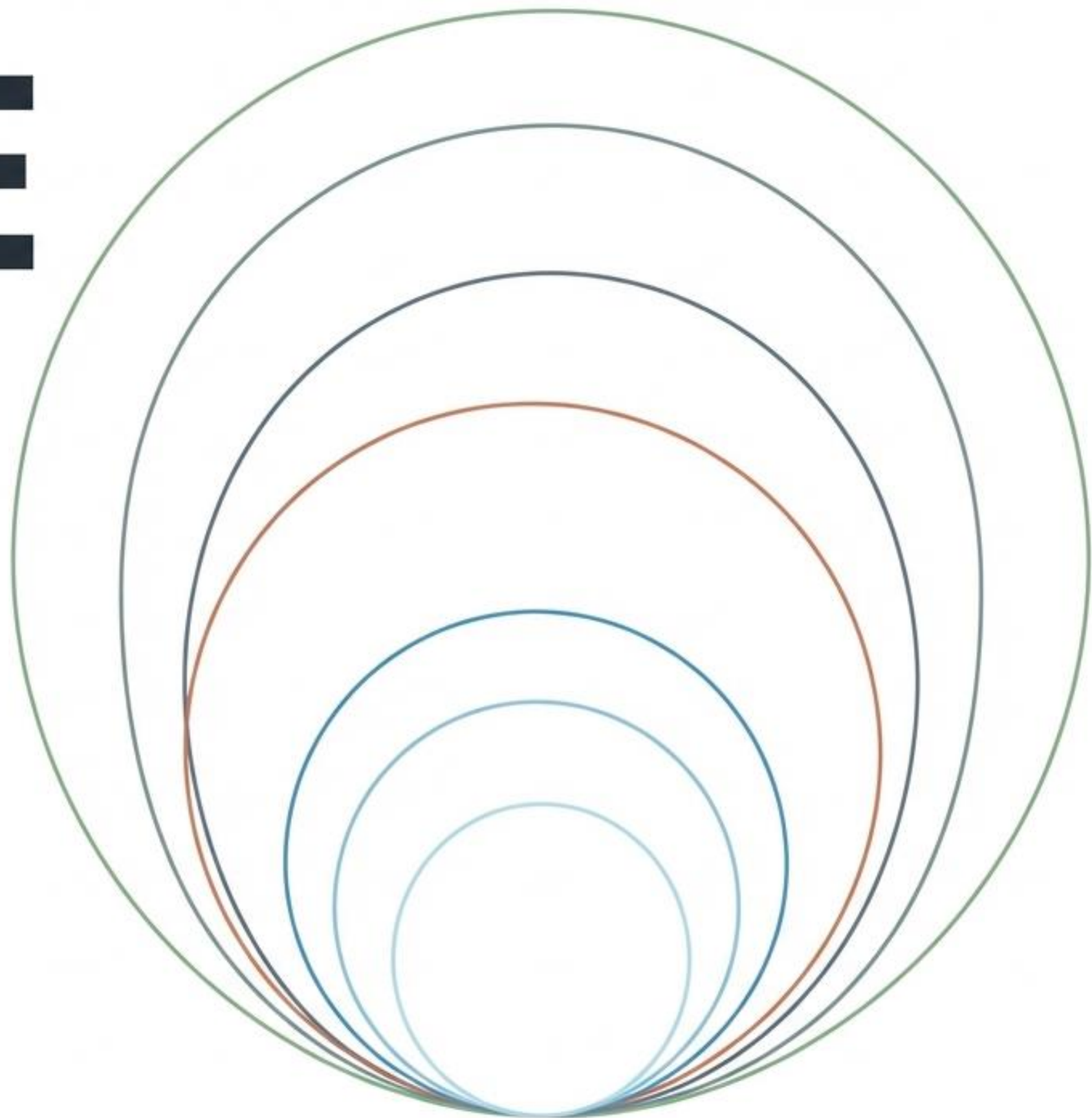


# EarthSE

A Benchmark Evaluating  
Earth Scientific Exploration  
Capability for Large  
Language Models

Wanghan Xu, Xiangyu Zhao, Yuhao Zhou, et al.  
Shanghai Jiao Tong University &  
Shanghai Artificial Intelligence Laboratory  
ICLR 2026



# Existing benchmarks evaluate scientific recall, but ignore open-ended exploration.

Benchmark	#Ques.	Source	Earth Cover	Grad-Diff	Multi-Sub	Scientific Exploration
ScienceQA (Lu et al., 2022)	21,208	High School Courses	×	×	×	×
MMLU-Pro (Wang et al., 2024b)	12,032	Exam Questions	×	×	×	×
SciBench (Wang et al., 2023)	869	10 Textbooks	×	×	×	×
GPQA (Rein et al., 2024)	448	Expert Curated	×	✓	✓	×
OceanBench (Bi et al., 2023)	12,426	Domain Texts	×	✓	×	×
ClimaQA (Manivannan et al., 2024)	3502	18 Textbooks	×	✓	✓	×
GeoBench (Deng et al., 2024)	2439	Exam Questions	×	✓	×	×
<b>EarthSE</b>	<b>4133</b>	<b>100,000 Earth Science Papers</b>	✓	✓	✓	✓

**Zero existing benchmarks assess open-ended, iterative Scientific Exploration.**

# A comprehensive, three-tiered evaluation architecture.

## Tier 3: Earth-Gold (The Frontier)

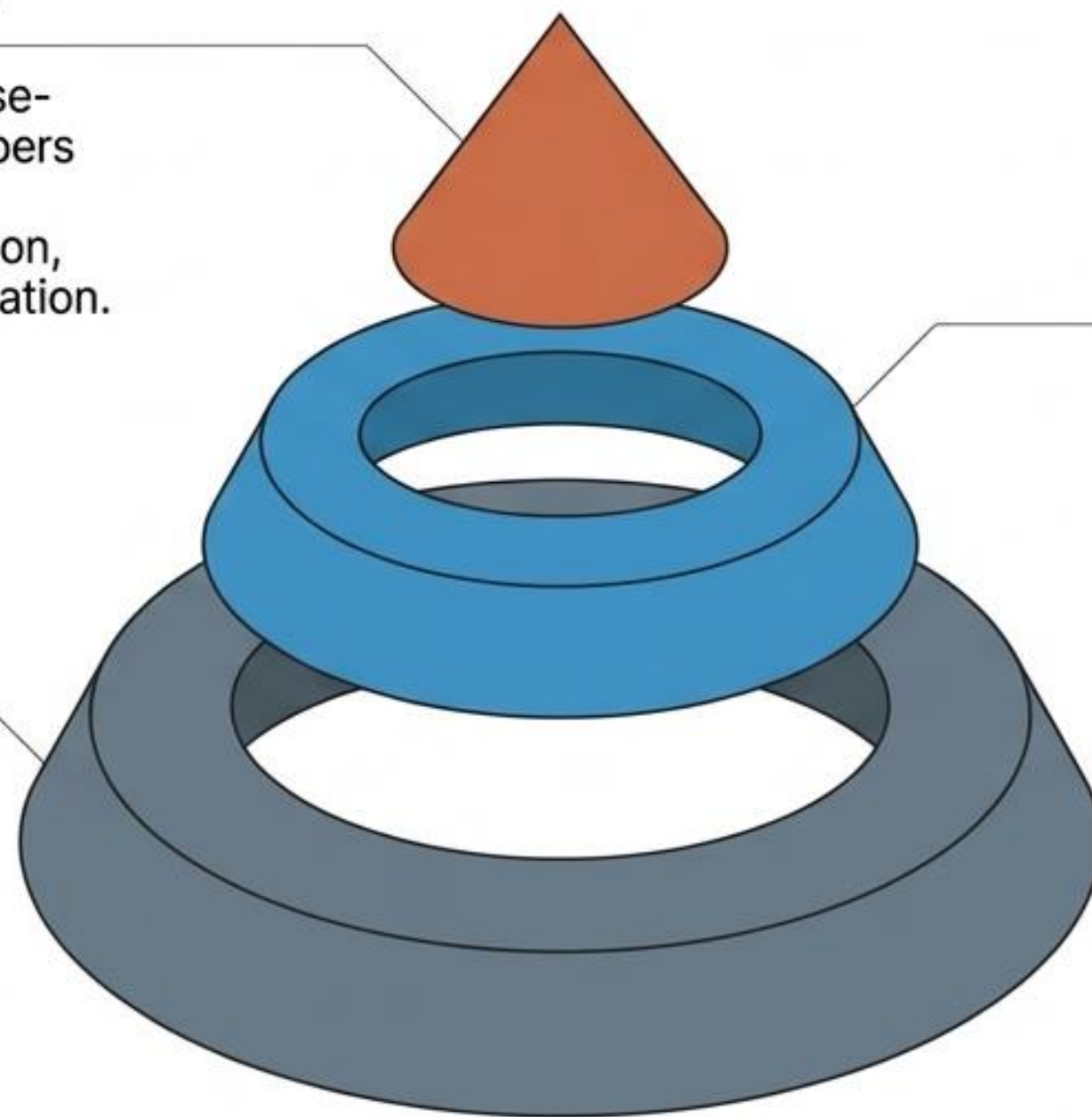
- **Scale:** Open-ended dialogues reverse-engineered from 1K Highly-Cited papers ( $P_{hc}$ ).
- **Focus:** Advanced scientific exploration, reflection, and methodological innovation.

## Tier 1: Earth-Iron (The Foundation)









- **Scale:** 4,133 QA pairs derived from 100K base papers ( $P_{base}$ ).
- **Focus:** Broad foundational knowledge across all 114 sub-disciplines.

## Tier 2: Earth-Silver (The Challenge)

- **Scale:** High-difficulty QA extracted from 10K High-Impact Journal papers ( $P_{hj}$ ).
- **Focus:** Specialized, professional-level analytical depth.

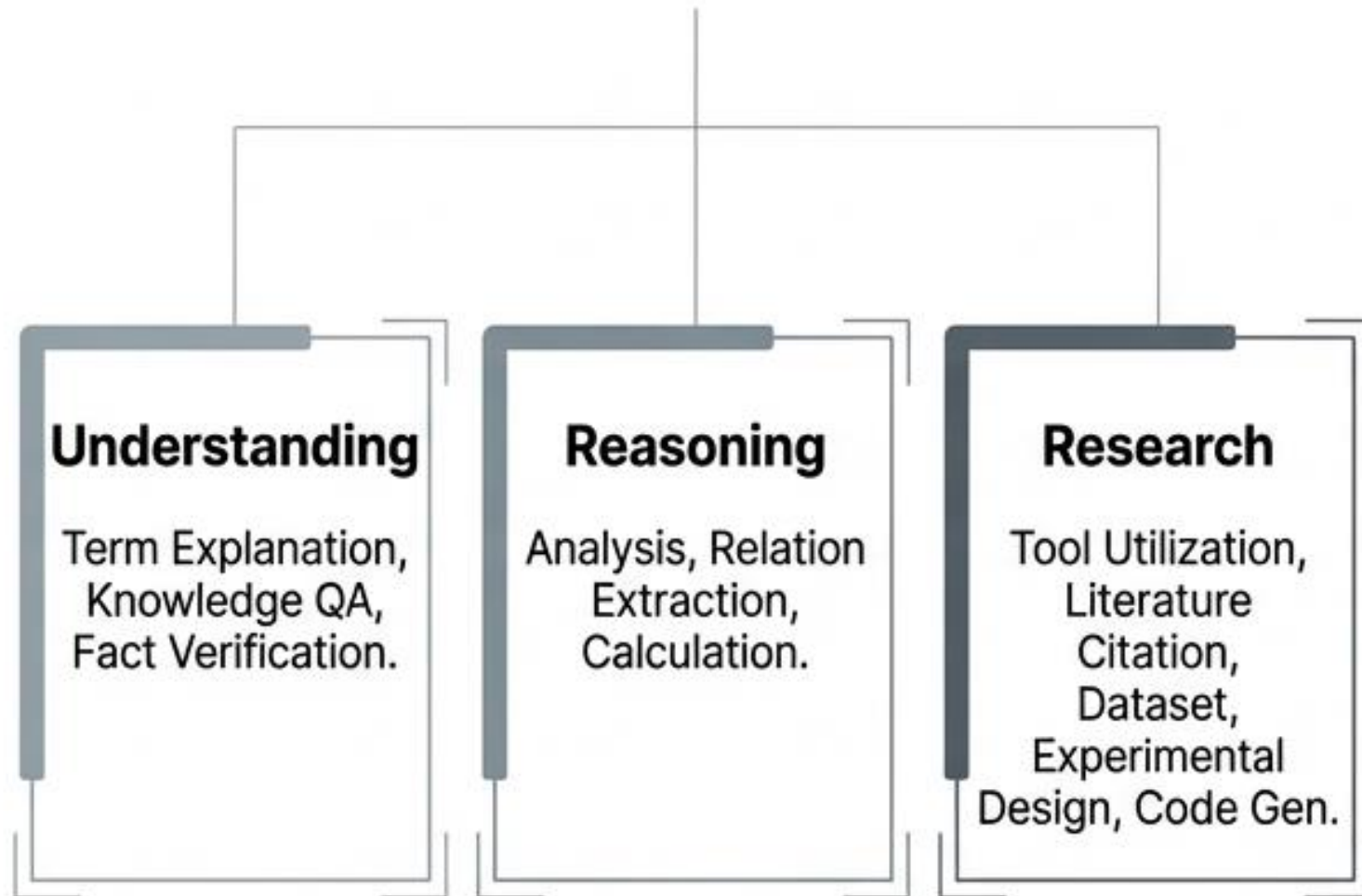


# 5 Spheres. 114 Sub-disciplines. 100,000 Academic Papers.

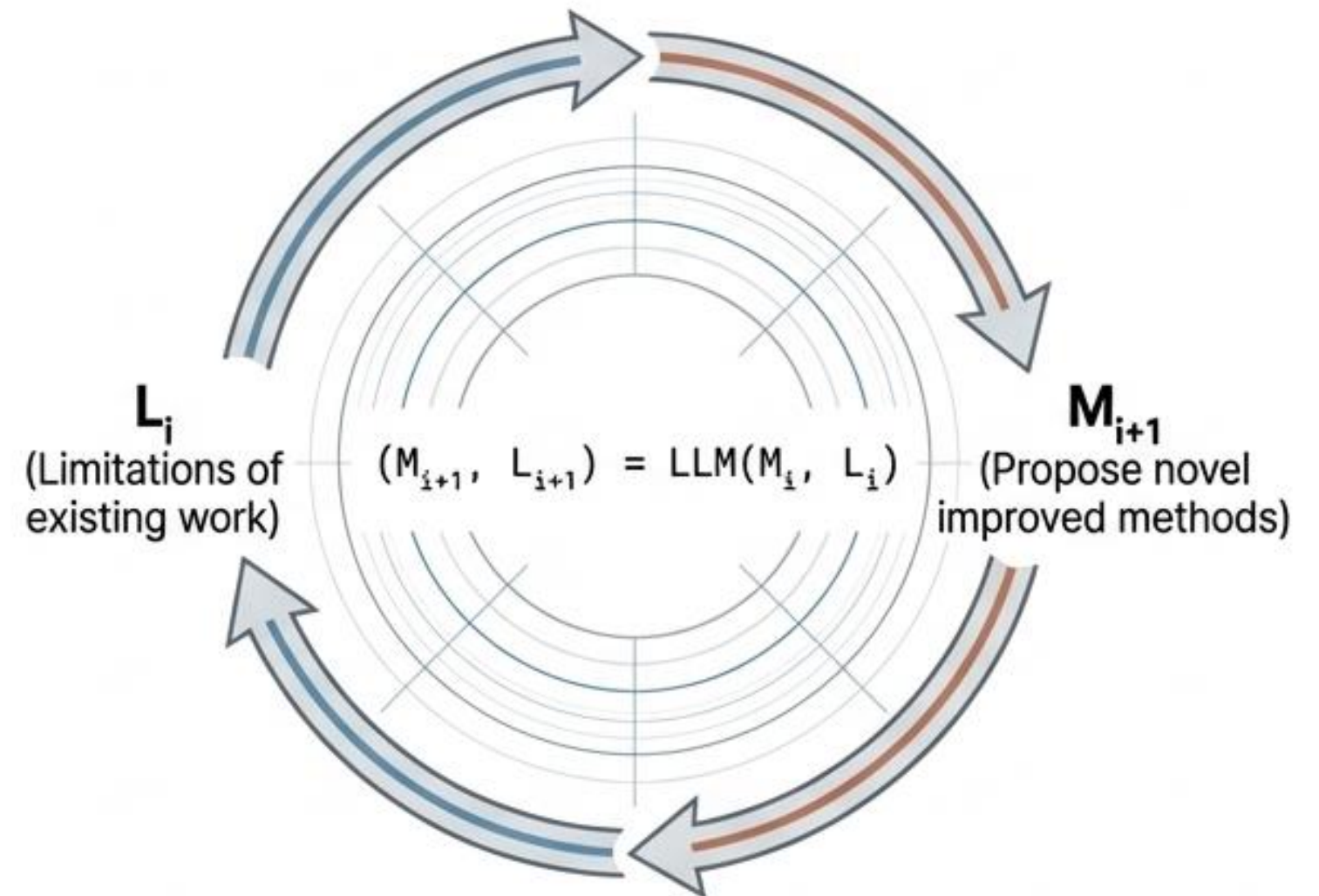
<p><b>Ecology and Biosciences</b></p> <p>Restoration Ecology Biogeochemistry Landscape Ecology Population Ecology Ecosystem Ecology Biogeography Regional Ecology Soil Ecology Environmental Urban Ecology ...</p> 	<p><b>Hydrology</b> </p> <p>Ecohydrology Groundwater Hydrology Hydrological Measurement Regional Hydrology Hydrological Meteorology Hydrological Physics Hydrogeology ...</p>	<p><b>Polar Science</b> </p> <p>Glaciology Polar Oceanography Polar Climate Science Permafrost Science Polar Ecology ...</p>	<p><b>Atmospheric Sciences</b> </p> <p>Meteorology Solar Energy Atmospheric Remote Atmospheric Physics Atmospheric Chemistry Hydrometeorology Air Quality and Pollution Weather Forecast ...</p>
<p><b>Climatology</b> </p> <p>Paleoclimatology ...</p>	<p><b>Solid Earth Geophysics</b> </p> <p>Geomagnetism Seismology Volcanology Remote Sensing Geology ...</p>	<p><b>Oceanography</b> </p> <p>Ocean Physics Ocean Chemistry Ocean Geology Ocean Biology Ocean Remote Ocean Environment ...</p>	<p><b>Geology</b> </p> <p>Sedimentology Quaternary Geology Mineralogy and Petrology ...</p>

# Moving beyond factual recall to simulate the human scientific process.

## Foundational Competencies

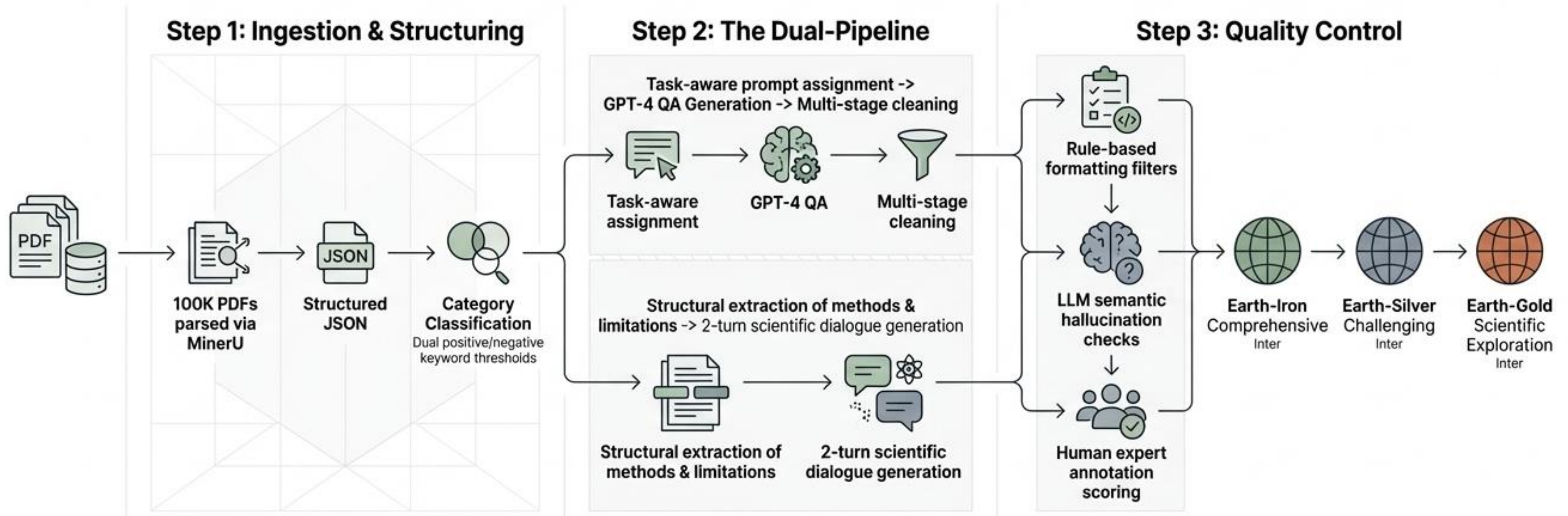


## The Exploration Loop

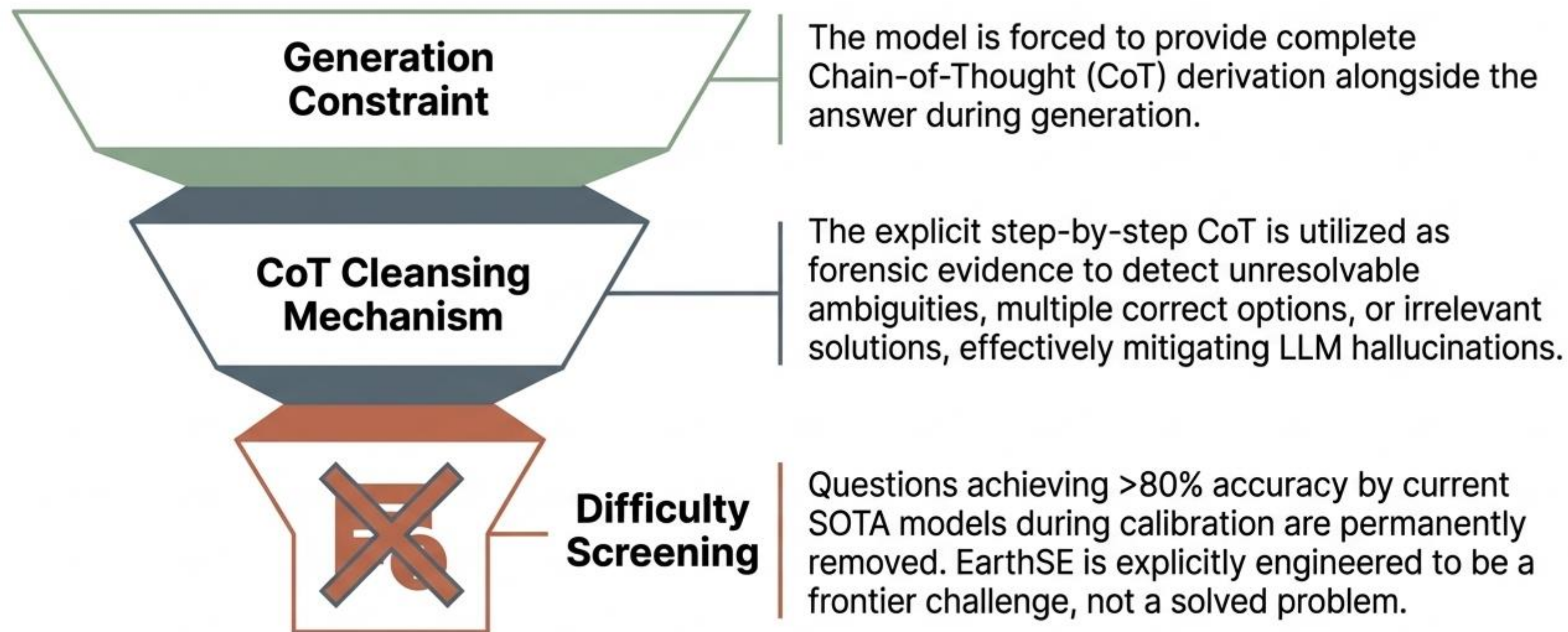


True scientific discovery requires critically analyzing the limitations of existing work to propose novel, improved methods. It is modeled as an iterative, self-critique process.

# An automated, expert-verified pipeline built on structured scientific literature.

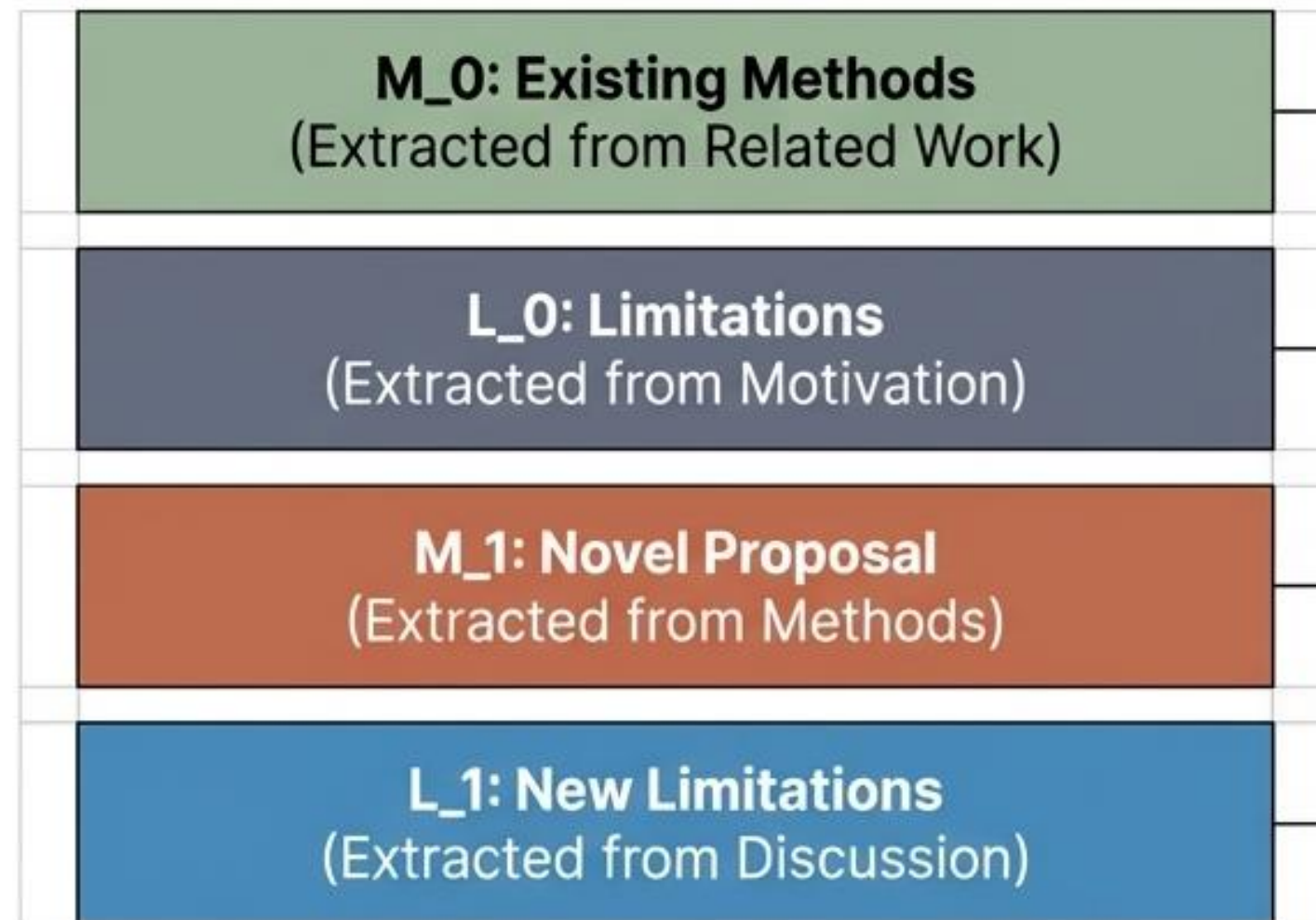


# Forcing explicit Chain-of-Thought reasoning to guarantee benchmark rigor

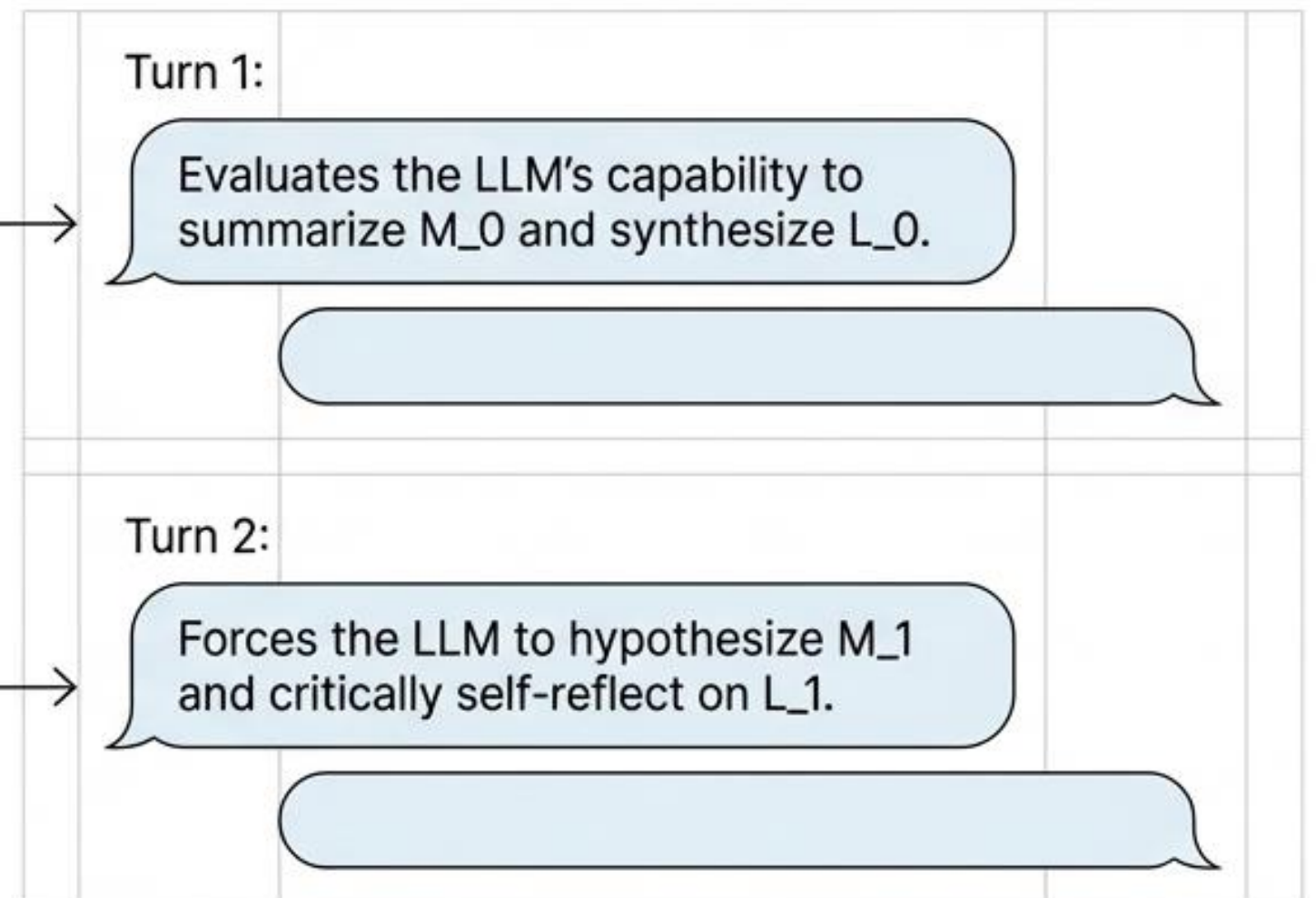


# Reverse-engineering highly-cited papers to simulate collaborative AI research

The Extraction Engine (from P<sub>hc</sub> corpus)



The Assessment Dialogue



# Standard accuracy fails open-ended discovery. Introducing the Scientific Exploration Score (SES).

$$SES = r \times d$$

$$SES = r / (10 \times (\bar{s} - 0.9))$$

## Retention (r)



Does the model preserve high-quality scientific reasoning?

Calculated via GPT-4 rank-based evaluation comparing M generated answers against the reference paper's actual methodology.

## Diversity (d)



Does the model exhibit divergent, lateral thinking?

Measured via the inverse cosine similarity of sentence embeddings across multiple generation attempts.

**Key Insight:** High SES demands models break out of generic repetitions; they must deliver both profound analytical quality and conceptual breadth.



# Performance plummets when confronted with graduate-level, specialized depth.

## The Empirical Drop



## Case Study: Stokes Number Calculation

Category: Atmosphere 🌤️

Format: Free-response ❓

Task: Calculation 🧮

If the density of LN fly ash particles is  $1800 \text{ kg/m}^3$ , the surface area of the particle is  $1.6 \times 10^{-5} \pi \text{ m}^2$ , the incident velocity is  $1.5 \text{ m/s}$ , and the viscosity of the liquid is  $0.001 \text{ Pa} \cdot \text{s}$  under dry conditions, calculate the Stokes number using the formula  $St = \frac{\rho r v_i}{\eta}$ , where  $\rho$  is the particle density,  $r$  is the particle radius,  $v_i$  is the incident velocity, and  $\eta$  is the viscosity of the liquid.

Step 1: Calculate particle radius  $A = 4\pi r^2 \Rightarrow r = \sqrt{\frac{A}{4\pi}} = 2 \times 10^{-3} \text{ m}$

Step 2: Calculate Stokes number  $St = \frac{\rho r v_i}{\eta} = \frac{1800 \times 2 \times 10^{-3} \times 1.5}{0.001} = 5400$

Answer: 5400

DeepSeek-V3:

1. Find particle radius (r):

1. Surface area of sphere:  $A = 4\pi r^2$

2.  $1.6 \times 10^{-5} \pi = 4\pi r^2$

3.  $r = \sqrt{\frac{1.6 \times 10^{-5}}{4}} = \sqrt{4 \times 10^{-6}} = 2 \times 10^{-3} \text{ m}$

2. Calculate Stokes number (St):

1. Formula:  $St = \frac{\rho r v_i}{\eta}$

2.  $St = \frac{1800 \times 2 \times 10^{-3} \times 1.5}{0.001} = 5400$

Llama-3.1-405B:

Surface Area =  $4\pi r^2$ , where  $r$  is the radius of the sphere.

We can rearrange the formula to solve for  $r$ :

$r = \text{sqrt}(\text{Surface Area} / 4\pi)$

Plugging in the value for the surface area, we get:

$r = 2 \times 10^{-3} \text{ m}$

Now that we have the particle radius, we can calculate the

Stokes number using the formula:

$St = \rho r v_i / \eta$

Plugging in the values, we get:

$St = 1800 \text{ kg/m}^3 \times 2 \times 10^{-3} \text{ m} \times 1.5 \text{ m/s} = 5.4$

While models like DeepSeek-V3 successfully navigate multi-step formulas, others like Llama-3.1-405B fundamentally misuse domain-specific formulas mid-derivation, exposing brittle mathematical reasoning outside general knowledge.

## The Exploration Bottleneck: LLMs default to generic safety over specific hypotheses.

**The Data Reality:** Most SOTA models achieve a retention rate of <50% and severely low diversity scores (<2.0), indicating rigid, repetitive thinking.



I'm interested in exploring changes and variability in Arctic sea ice thickness, volume, and multiyear ice coverage from 1958 to 2018. Could you summarize key prior research in this area and highlight their main limitations?



### Reference answer:

Certainly. Research in this field has relied on submarine missions (since 1958) for early ice thickness estimates, later supplemented by satellite observations like **ICESat** and **CryoSat-2 (CS-2)**. Multiyear ice (MYI) coverage has also been mapped using scatterometer data from **QuikSCAT** and **ASCAT**...



### Bad answer:

Research on Arctic sea ice from 1958 to 2018 indicates significant declines in thickness, volume, and multiyear ice coverage, driven by climate change. Key studies utilized satellite data and models, revealing trends like reduced multiyear ice and thinner overall ice...



Given these limitations, what alternative methodology would you recommend, and what challenges might it introduce?



### Reference answer:

... **Monthly gridded CS-2 fields** could then be paired with MYI coverage maps to analyze ... However, this method has drawbacks, the reliance on **two temporally overlapping but independently derived datasets (MYI and volume)** may introduce artifacts in trend attribution, particularly for pre-satellite eras.



### Bad answer:

I recommend using a combination of remote sensing, in situ measurements, and advanced modeling techniques like machine learning. This approach could enhance spatial and temporal resolution but may introduce challenges such as integrating diverse data sources, ensuring consistency.

# Inference-time scaling unlocks latent capabilities in highly challenging questions.

Inspired CoT Steps	FIB (%) ↑
DeekSeek-V3	
0 (Baseline)	12.80
1	21.60
2	29.60
3	<b>45.60</b>
GPT-4o	
0 (Baseline)	18.40
1	25.60
2	38.80
3	<b>55.60</b>

## The Experiment

Providing initial Chain-of-Thought (CoT) steps as hints during inference on the hardest Earth-Silver FIB questions.

## Results

**DeepSeek-V3:** Accuracy surges from **12.8%** (0 steps) to **45.6%** (3 steps)

**GPT-4o:** Accuracy jumps from **18.4%** (0 steps) to **55.6%** (3 steps)

**Takeaway:** Without modifying base model weights, leveraging CoT-augmented reasoning at inference time is a highly viable strategy to bridge the gap in complex Earth Science tasks.

# Three critical barriers at the frontier of AI scientific discovery



## 1. The Calculation Deficit

LLMs possess broad domain knowledge but lack robust, specialized mathematical reasoning for multi-step physics and chemistry equations native to Earth sciences.



## 2. The Specificity Problem

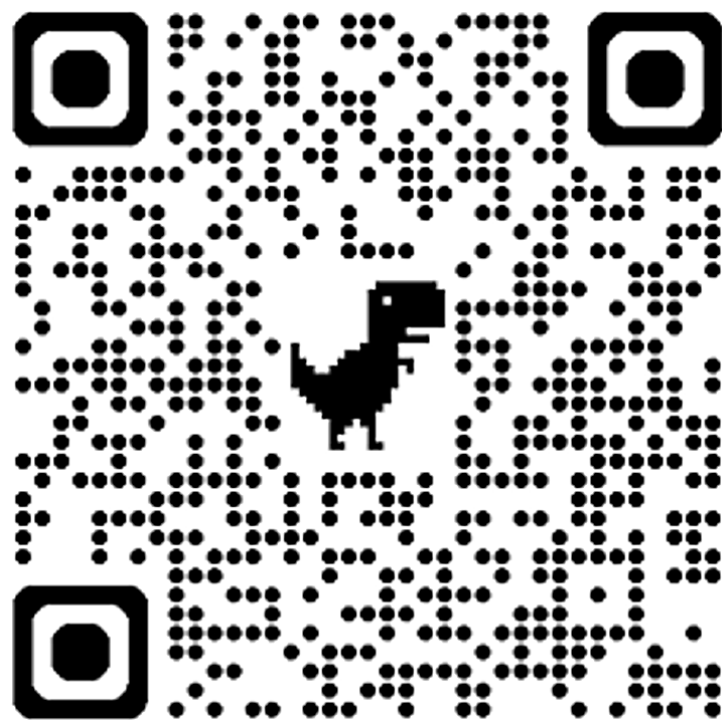
In open-ended exploration, models collapse into generic "safe" advice. They act as high-level summarizers rather than proposing concrete, actionable scientific architectures or datasets.



## 3. The Divergence Failure

Low SES diversity scores prove models struggle with lateral, divergent thinking. They repeatedly traverse identical conceptual paths, a fatal flaw for true scientific innovation.

# EarthSE establishes the new standard for evaluating AI in the Earth Sciences.



Access the 100K paper corpus, evaluation code,  
and the complete Iron, Silver, and Gold datasets.

**[huggingface.co/ai-earth](https://huggingface.co/ai-earth)**

Shanghai Artificial Intelligence Laboratory  
ICLR 2026