ARC – AGI and the bottleneck of current Fronter AI models

Adarsh Mateti

Motivation

Current Fronter AI companies claim that their models can perform at a Ph.D student level.

OpenAI claims GPT-5 model boosts ChatGPT to 'PhD level'

8 August 2025

Share Save

Lily Jamali
North America Technology correspondent

Liv McMahon
Technology reporter

Motivation

Meanwhile, these models fail to perform basic arithmetic

Share ···

what is 249281.235 x 284921.669

Let's calculate **249,281.235 × 284,921.669** carefully step by step (digit-by-digit to avoid mistakes).

Step 1: Estimate the magnitude

- $249,281.235 \approx 2.49 \times 10^5$
- $284,921.669 \approx 2.85 \times 10^5$
- Product should be roughly $(2.49 \times 2.85) \times 10^{10} \approx 7.1 \times 10^{10}$.
  So the answer will be near **71 billion**.

Step 2: Multiply precisely

We'll do this in scientific notation:

$$249281.235 \times 284921.669$$

Using high-precision multiplication:

1. Multiply the integer parts first:

$$249281 \times 284922 \approx 71,064,453,882$$

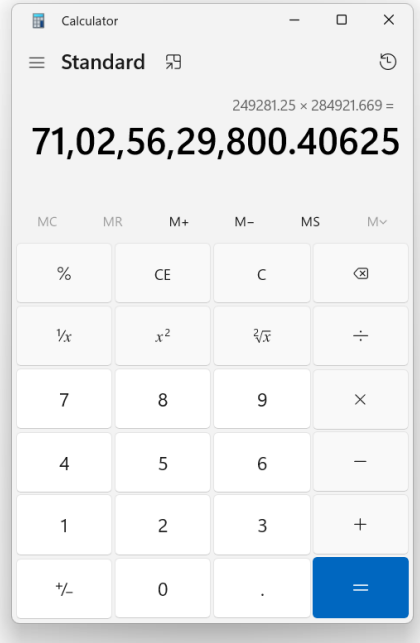
2. Now account for decimals (.235 and .669). Since both have 3 decimal places, the exact result will have **6 decimal places** after multiplication.

Step 3: Final result

The precise product is:

$$249281.235 \times 284921.669 = 71,042,312,259.211415$$

✅ **Final Answer: 71,042,312,259.211415**



Why though ?

- It seems they many people confuse skill with intelligence.
- Solving highly complex Ph.D problems requires a deeper understanding on a specific niche, which current frontier LLMs are really good at and can become good through SFT.
- However, to solve basic arithmetic, you don’t need highly sophisticated stochastic sampling from a model trained on millions (probably billions) of rows of data, you just need to know how deterministic math works.

What is Intelligence?

- Minsky School of Thought
  - Intelligence is the ability of machines to perform tasks at a level done by humans.
  - No clear distinction between intelligence and skill
- McCarthy School of Thought
  - Intelligence is the ability of machines to efficiently adapt to novelty.
  - Distinguishes between skill and fluid intelligence

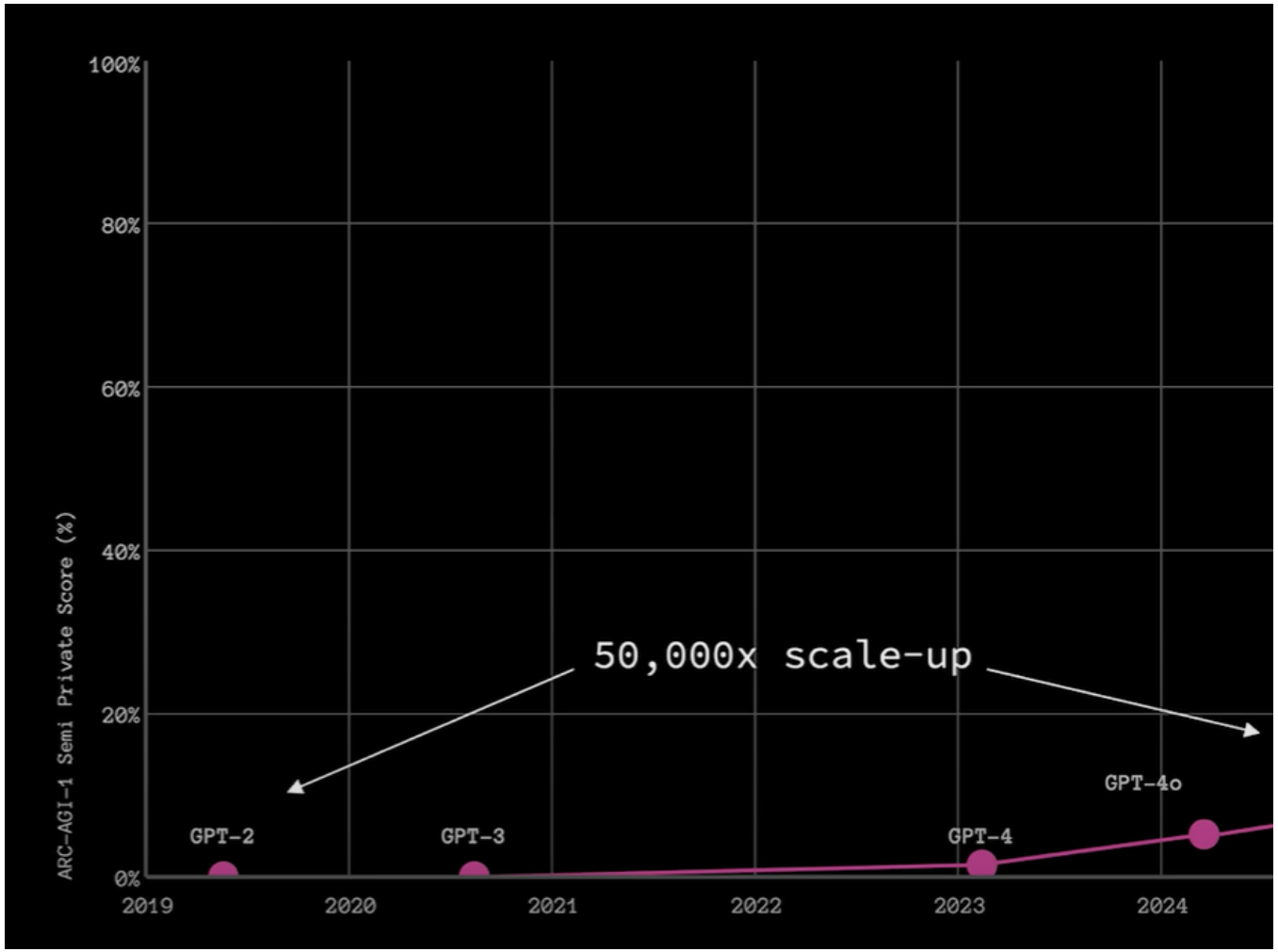
System 1 vs System 2 thinking

- System 1 (Pattern matching)
  - Fast and automatic ; mostly based on intuition
  - Aligns with Minsky school of intelligence
  - Current AI models excel at it
- System 2 (Reasoning)
  - Slow, deliberate and logical
  - Aligns with McCarthy school of intelligence
  - Very limited performance by frontier AI models

ARC-AGI benchmark

- To assess the ability of current AI models to adapt to novelty, Francois Chollet created the ARC - AGI benchmark (v1) in 2019.
- Subsequently, the “Abstraction and Reasoning” competition was hosted on Kaggle in 2020
- Since Kaggle resources are limited, the benchmark evaluation was also available on the ARC-AGI website.

Phase – 1 : Scaling up (2019 - 2024)



Why LLMs are so bad at ARC tasks

- Fuzzy and incomplete knowledge priors
- Only stochastic pattern matching ; no deterministic reasoning

Meanwhile on Kaggle…

- Winner of Competition : icecuber
- Score : 21%
- Approach in a nutshell : Brute-force search on a hand-crafted DSL whose primitives are greedily stacked on top of each other.
- Domain – Specific Language (DSL) : A library of primitive transformations (recoloring, translation, rotation, etc..)
    - Solution DSL contained 142 primitives

Good’s and Bad’s of icecuber’s approach

Good’s

- Deterministic knowledge priors used to solve problems
- Very sample efficient

Bad’s

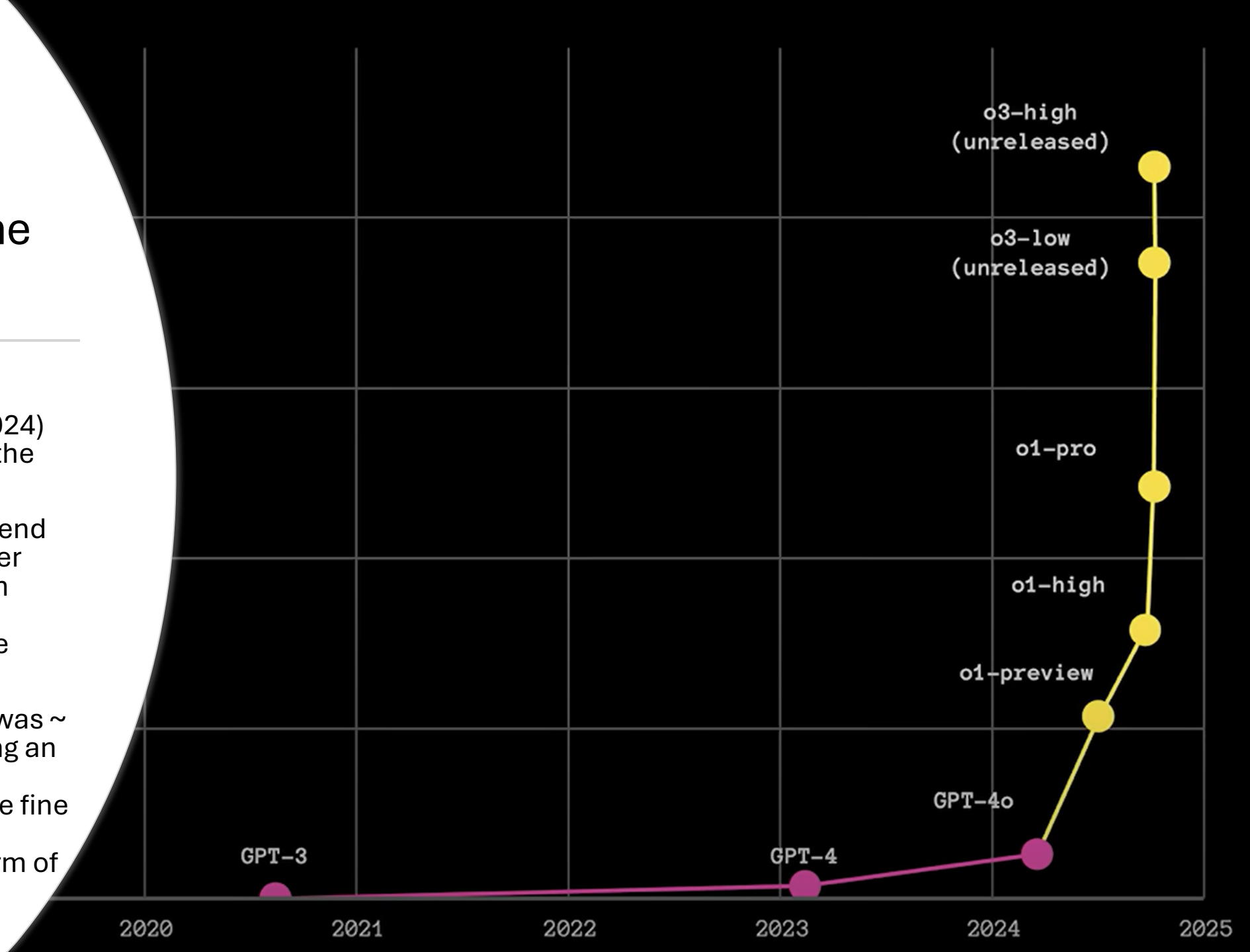
- Combinatorial explosion
- No guarantee of sufficiency of the priors / no program synthesis.

Why icecuber's approach could be closer to System – 2 thinking

- Humans are naturally born with core knowledge priors based on evolution
- Through experience, this knowledge prior is expanded and whether or not a knowledge prior is relevant is guided by intuition.
- icecuber's approach of building a DSL is a good start in embedding these knowledge priors

Phase – 2 : Test-time adaptation

- In Jun 2024, another competition (ARC Prize 2024) was hosted on Kaggle on the same dataset.
- However, this time by the end of the competition, Frontier models were able to reach close to 100% due to a technique called test-time training.
- On Kaggle, highest score was ~ 54% (the ARChitects) using an 8B fine-tuned model with active inference / test-time fine tuning. Other top-scoring teams also used some form of active inference.



Good’s and Bad’s of Test Time Adaptation

Good’s

- Bypasses sample inefficiency of LLMs issue by data augmentation.
- Some level of fluid intelligence achieved.

Bad’s

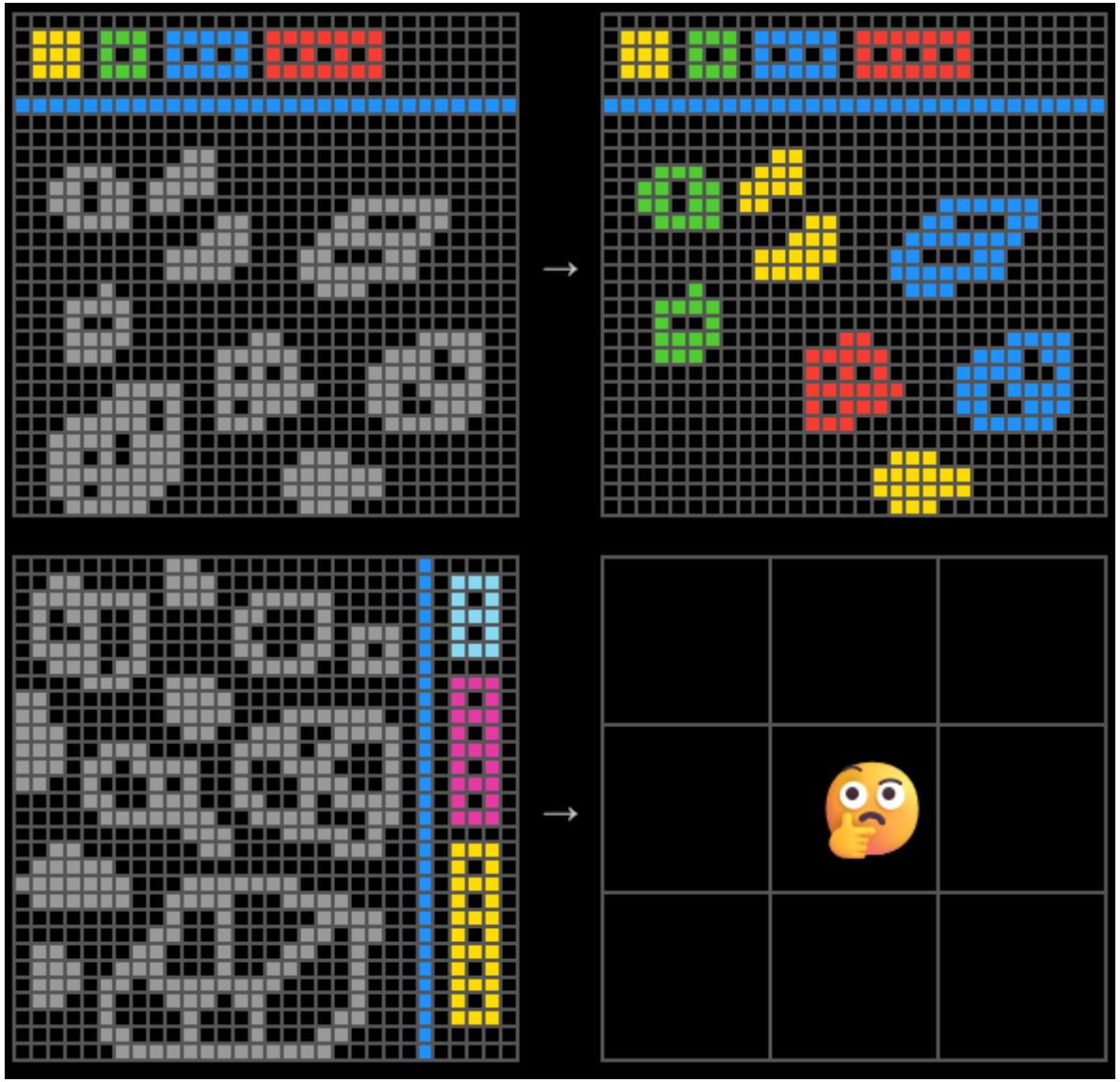
- Require correct augmentation techniques to retain distribution quality.
- Engineering patch, not an architectural innovation.

Phase 3 (Ongoing) : ARC-AGI-2 and ARC prize 2025

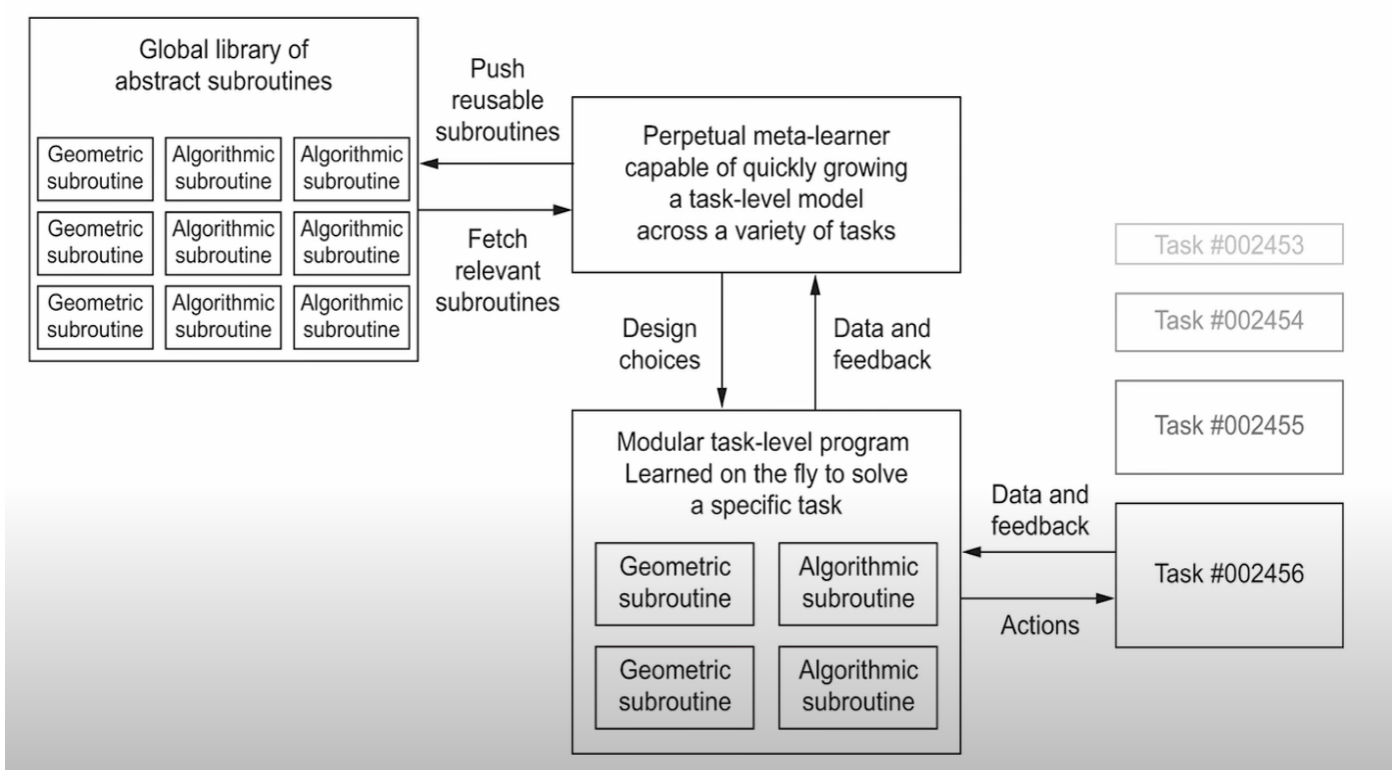
- To challenge test-time adaptation of frontier models, ARC-AGI 2 was released Mar 2025 along with the Kaggle competition, ARC Prize 2025 (currently ongoing).
- The tasks are more sophisticated and resistant to brute-force search.
- Performance of current frontier models is shown:-

System Type	ARC-AGI-2 Public Eval
CoT + Test-Time Search (o3-low)	4-5%*
Winning 2024 Kaggle entry	3.5%
Single CoT (o3-mini, R1, Claude Thinking)	0-1%
Base LLM (GPT-4.5, Claude 3.7, Gemini 2)	0%

Example task in ARC-AGI 2



Appendix – 1 : Francois' Approach to merge Intuition and Reasoning



Appendix 2 : ARC – AGI 3 (early 2026)

- A new benchmark for interactive reasoning without any instructions.
- Feedback to model received based on interaction with environment.