A NORTH STAR FOR AGI

ARC PRIZE FOUNDATION

A NORTH STAR FOR AGI

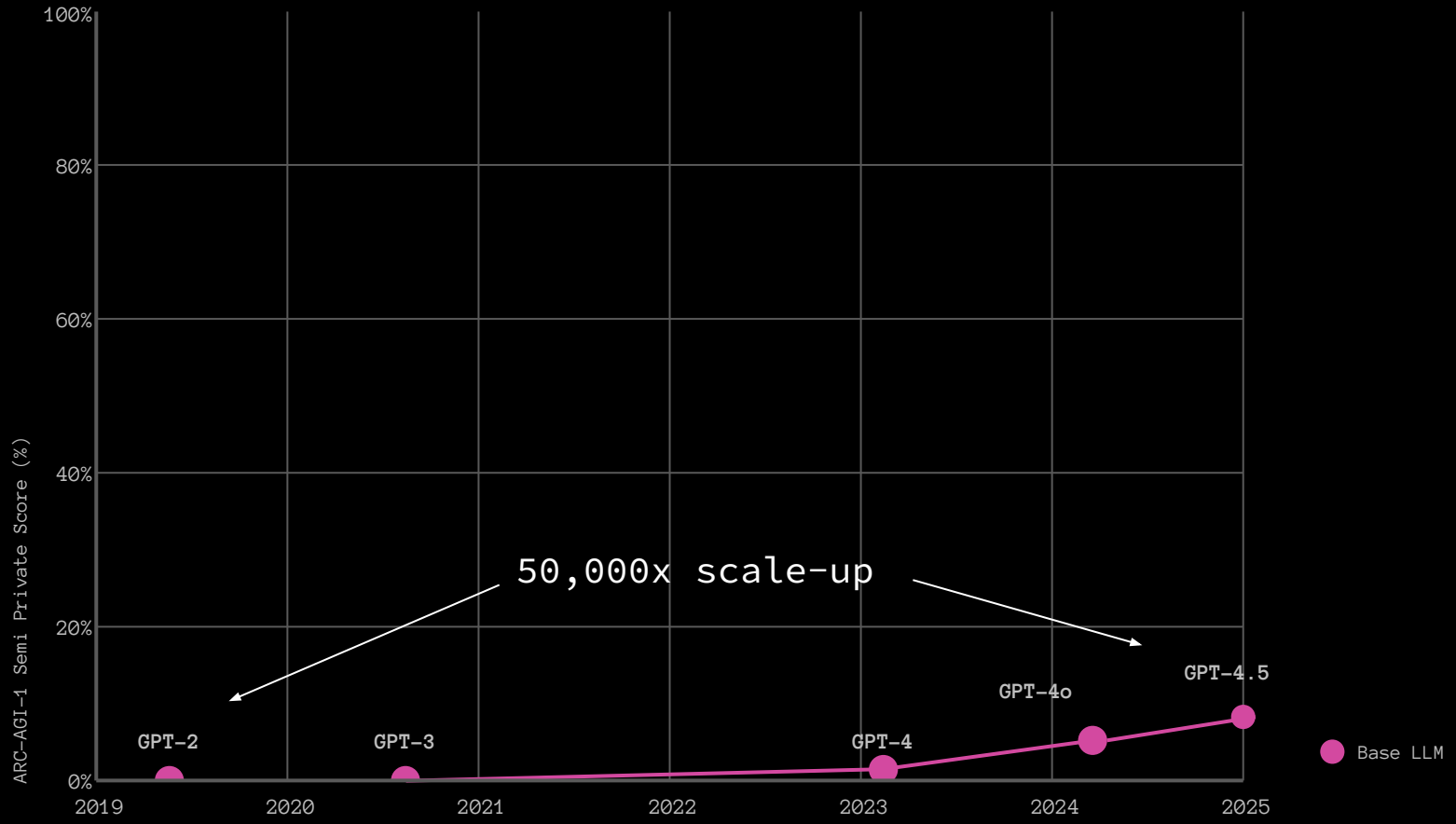
ARC PRIZE FOUNDATION

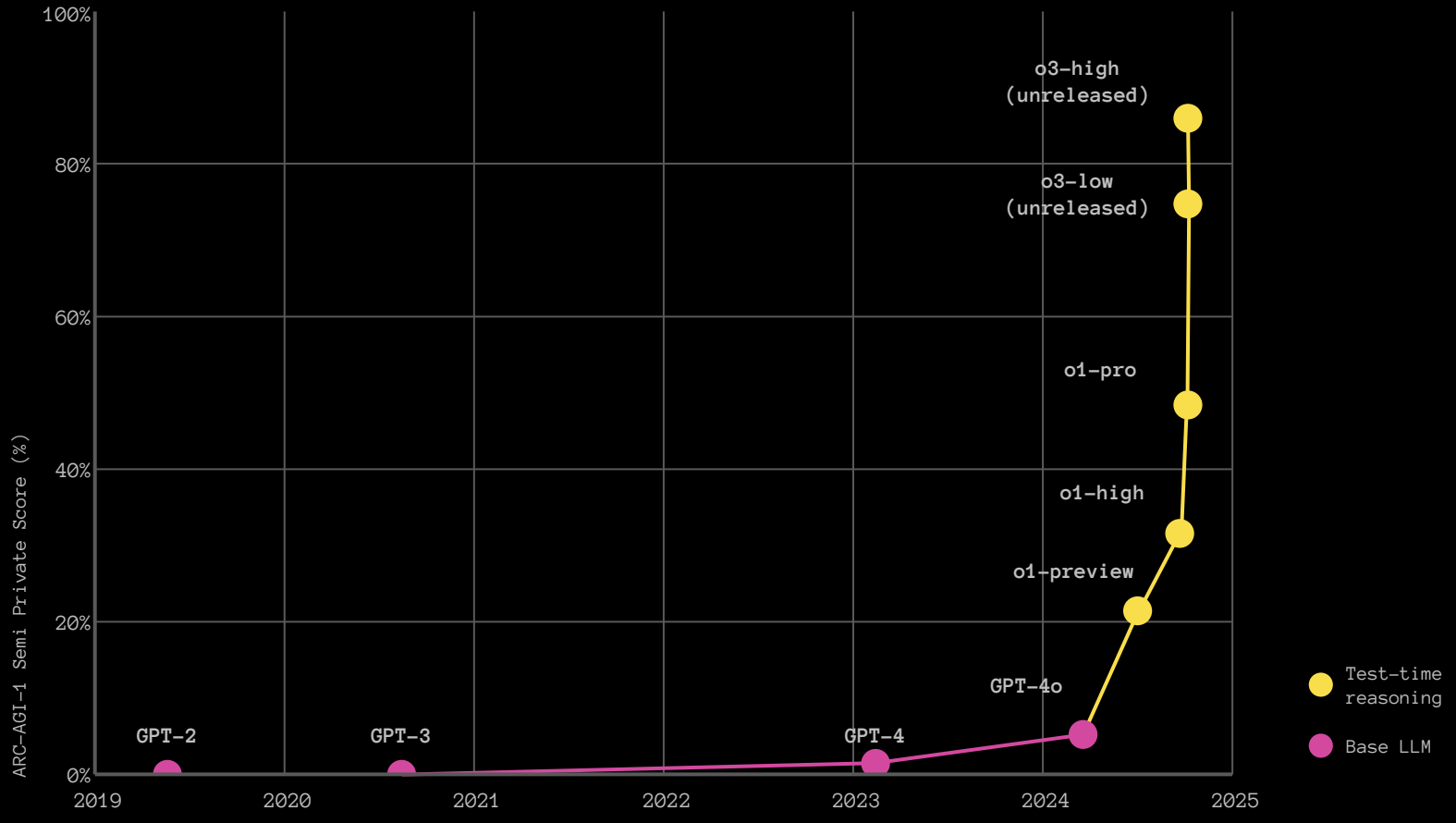
Francois Chollet
Co-Founder

Mike Knoop
Co-Founder

Greg Kamradt
President


ARC-AGI-2


100%
80%
60%
40%
20%
0%
ARC-AGI-1 Semi Private Score (%)
2019
2020
2021
2022
2023
2024
2025
50,000x scale-up
GPT-2
GPT-3
GPT-4
GPT-4o
GPT-4.5
Base LLM


100%
80%
60%
40%
20%
0%
ARC-AGI-1 Semi Private Score (%)
2019
2020
2021
2022
2023
2024
2025
o3-high
(unreleased)
o3-low
(unreleased)
o1-pro
o1-high
o1-preview
GPT-4o
GPT-2
GPT-3
GPT-4
Test-time
reasoning
Base LLM

2019: ARC-AGI-1 challenges deep learning

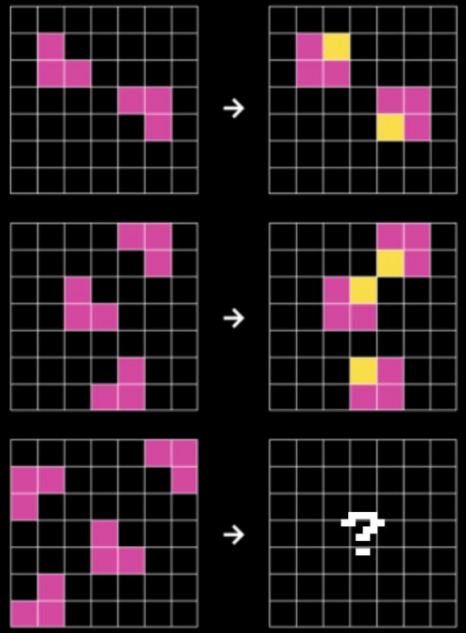
2025: ARC-AGI-2 challenges test-time reasoning

ARC-AGI-2: FULLY CALIBRATED FOR HUMAN-FACING DIFFICULTY

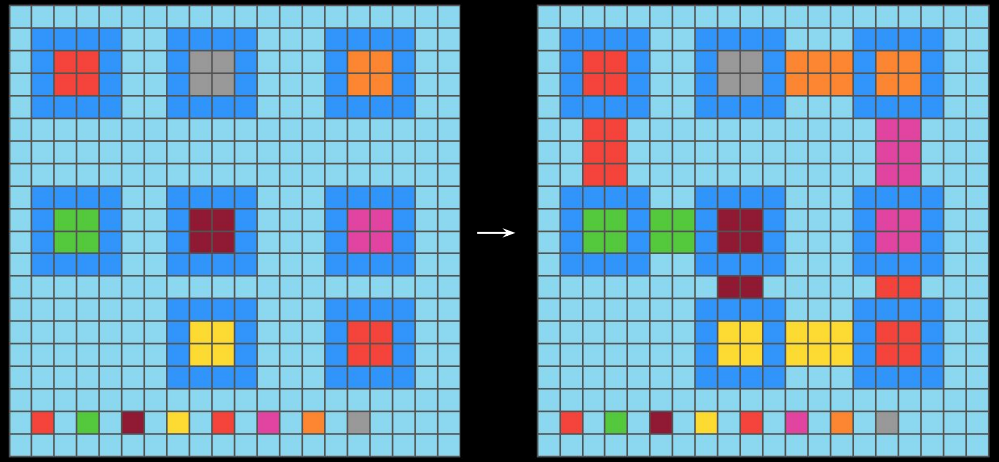
- All tasks solved by at least 2 people (out of 4-9)
- Full calibration of human performance on each eval set
  - Average single test-taker performance: 60%
  - Panel of 10 test-takers: 100%
- Public Tasks
  - 1000 Training tasks (easier) - Demonstrate format + Core Knowledge
  - 120 Evaluation tasks - Evaluate systems locally during development
- Semi-Private Tasks
  - 120 tasks - Evaluate commercial frontier systems
- Private Tasks
  - 120 tasks - Determine the winner of the competition on Kaggle

ARC-AGI-1 WAS EASILY BRUTE FORCIBLE – ARC-AGI-2 IS NOT

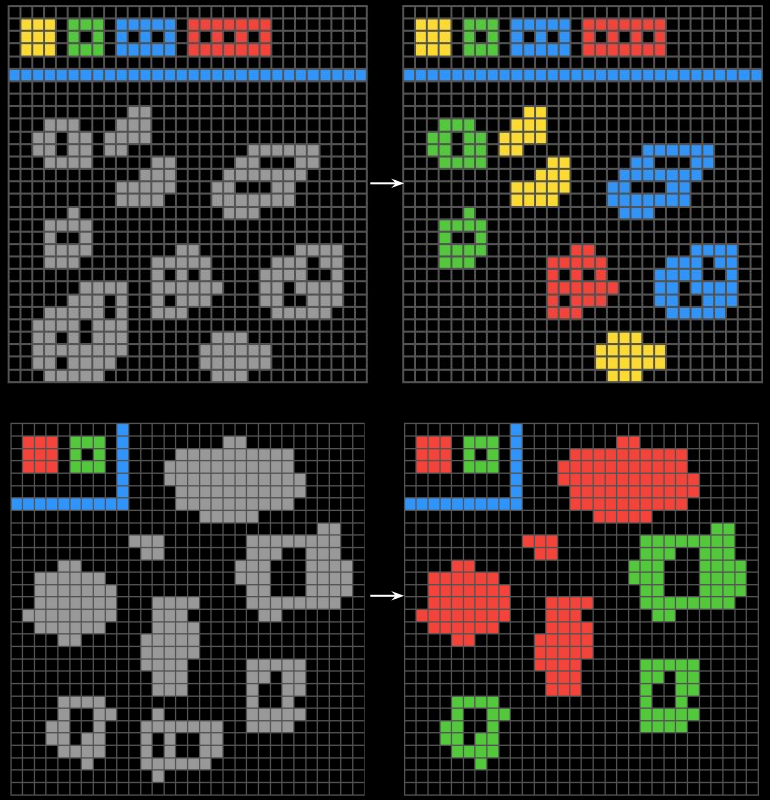
ARC-AGI-1 Task

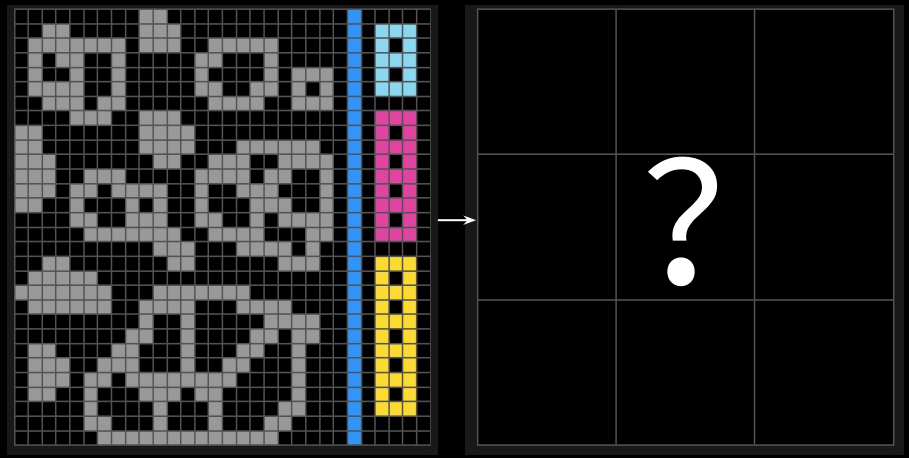


ARC-AGI-2 Task



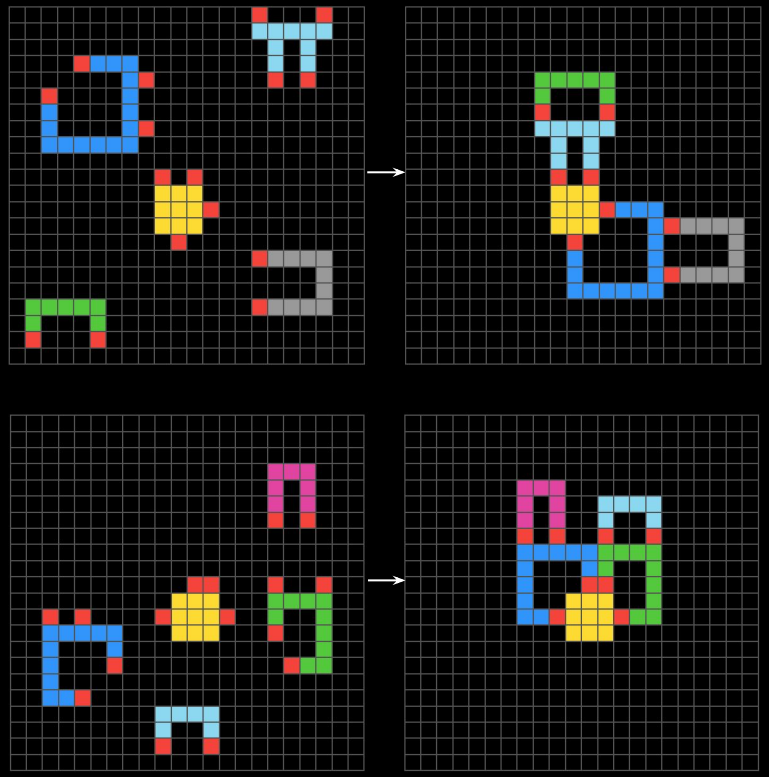
WHAT MAKES ARC-AGI-2 DIFFERENT? – SYMBOLIC INTERPRETATION

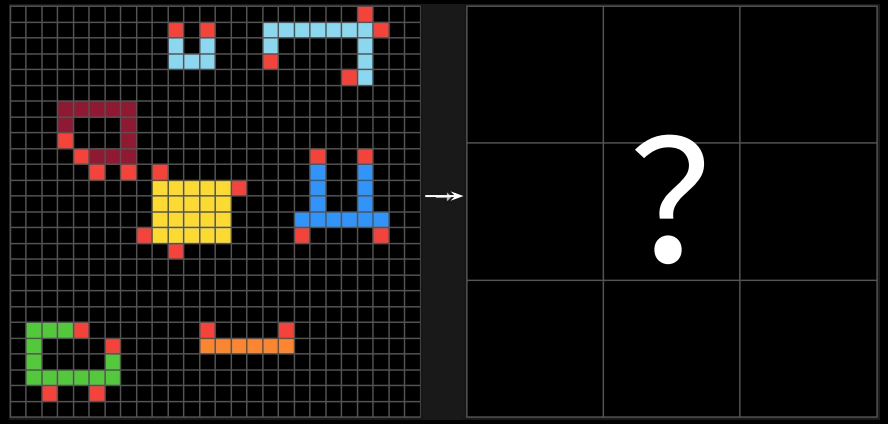




ARC-AGI-2 Public Eval Task: #e3721c99

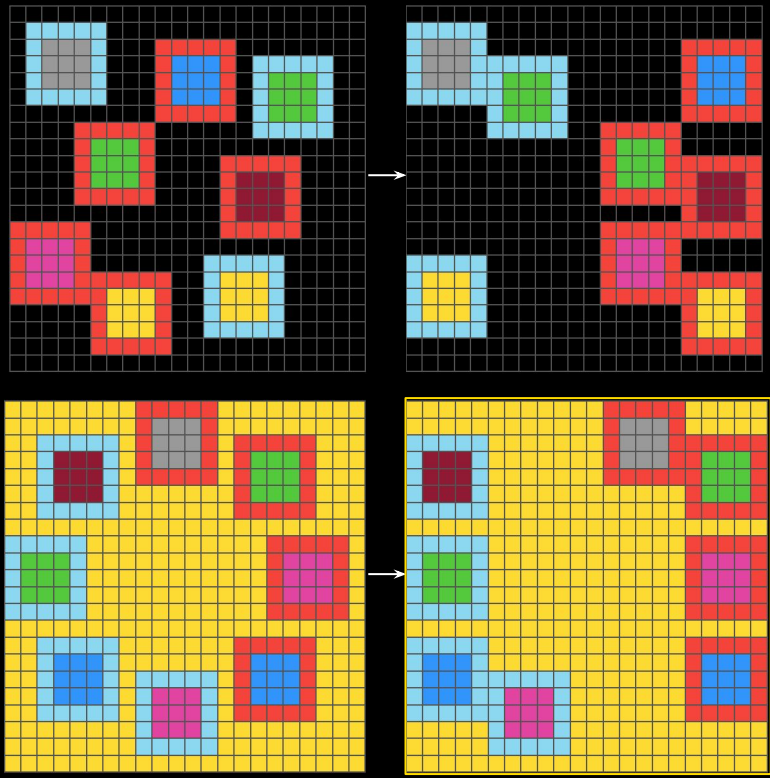
WHAT MAKES ARC-AGI-2 DIFFERENT? – MULTI-STEP COMPOSITIONAL RULES

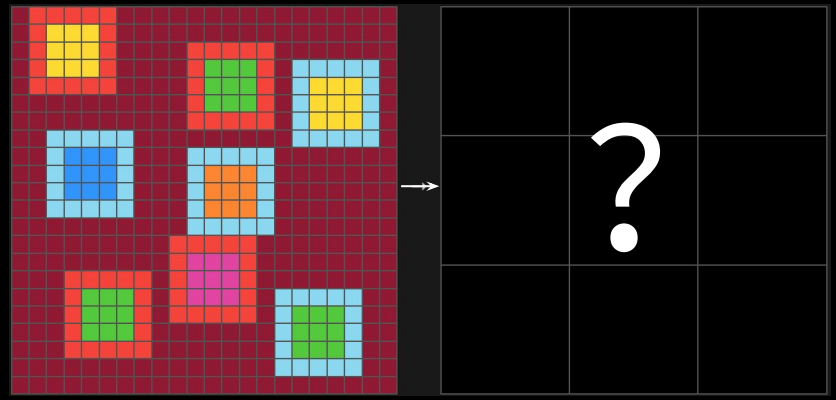




ARC-AGI-2 Public Eval Task: #cbebaa4b

WHAT MAKES ARC-AGI-2 DIFFERENT? – CONTEXTUAL RULE APPLICATION





ARC-AGI-2 Public Eval Task: #b5ca7ac4

ARC-AGI-2 SYSTEM PERFORMANCE

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%

* Estimate, will fully test once available

Represents

A compass pointing towards useful research direction

A playground to test few-shot reasoning architectures

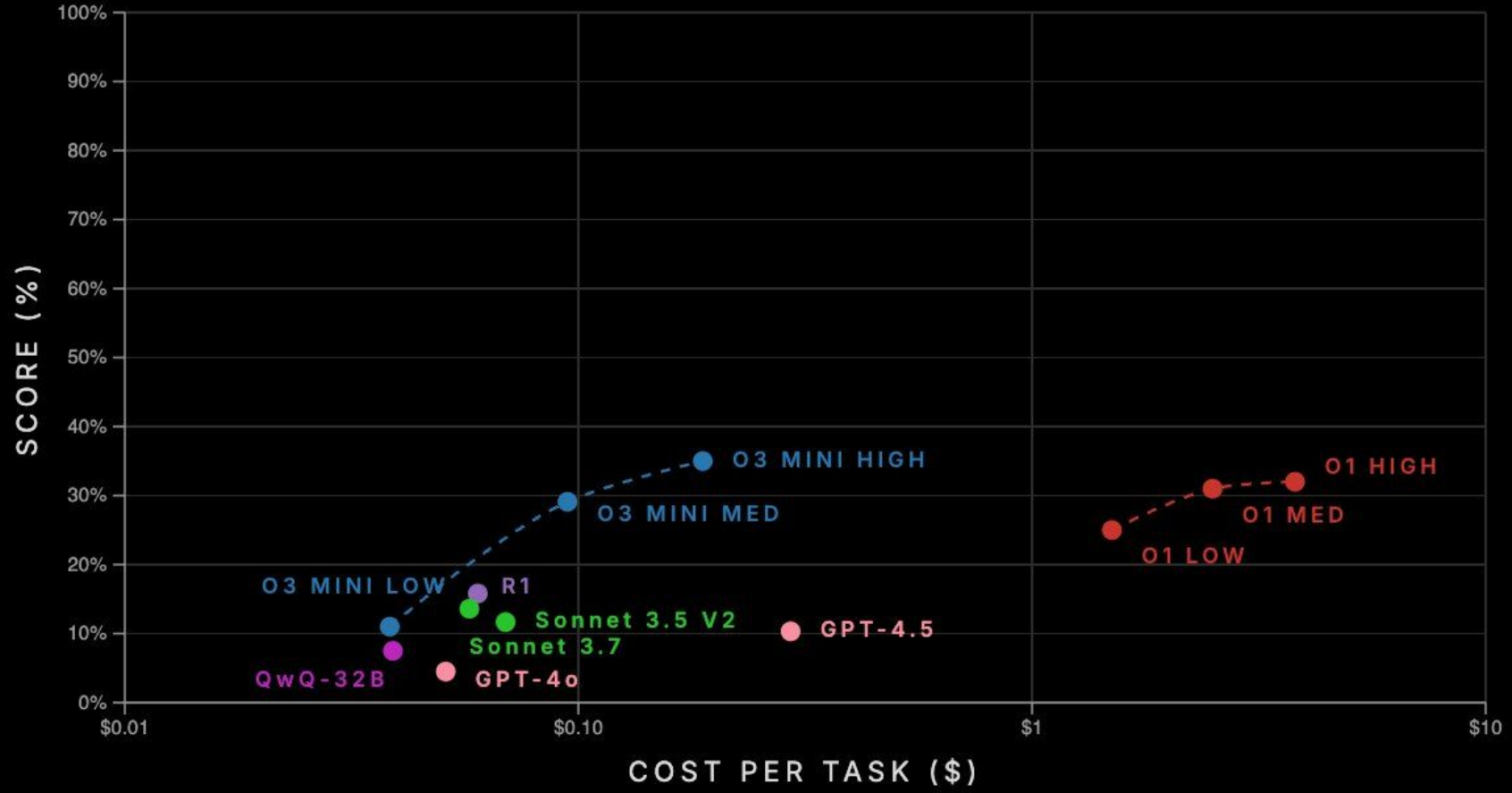
A tool to accelerate progress towards AGI

Does Not Represent

An indicator of whether we have AGI or not

(*in theory, you can solve ARC-AGI without full AGI!*)

ARC-AGI-1 SEMI-PRIVATE EVAL



2024 PAPER AWARD WINNERS

1ST PLACE - $50K

"Combining Induction and Transduction for Abstract Reasoning".

Li et al.

2ND PLACE - $20K

"The Surprising Effectiveness of Test-Time Training for Abstract Reasoning".

Akyürek et al.

3RD PLACE - $5K

"Searching Latent Program Spaces".

Bonnet & Macfarlane

RUNNERS UP - $2.5K

"The LLM ARChitect: Solving ARC-AGI Is a Matter of Perspective".

Franzen et al.

"Omni-ARC".

Barbadillo

ARC PRIZE 2025

MARCH 24 - NOVEMBER 3

ARC-AGI-2

$50 COMPUTE PER SUBMISSION

NO INTERNET ACCESS

OPEN SOURCE REQUIRED FOR WINNERS

$75K PAPER PRIZE, $50K HIGH SCORE, $600K GRAND PRIZE

ARC-AGI-3

ARC PRIZE

JOIN THE MISSION

Early Testing & Model Cards

Including ARC-AGI-1/2 performance model cards helps communicate reasoning capabilities (see e.g. o3)

Help Build ARC-AGI-3

Join the ARC-AGI-3 co-design committee

Goal: The benchmark effectively reflects a model's strengths, identifies growth areas, and serves as a tool for the community.

THANK YOU.