POSITION · ARC-AGI · NEXT-GENERATION BENCHMARKS

ARC PRIZE RESEARCH SUMMIT 2026 · OCT 23, 2026

Score the δ, not the prior.

Freeze one model. Let it play against itself. Measure how much it grows, and whether that growth transfers.

Sundong Kim GIST AI, Korea sundong@gist.ac.kr

THE PROBLEM

A newer Qwen lifts the whole leaderboard. We are scoring the prior, not the learner.

TODAY'S WINNING RECIPE

Biggest model that fits → post-train on masses of augmented games → submit.

Memory and GPU are fixed, prior knowledge is unlimited. Under an unlimited budget, SoTA will keep coming from larger-capacity models.

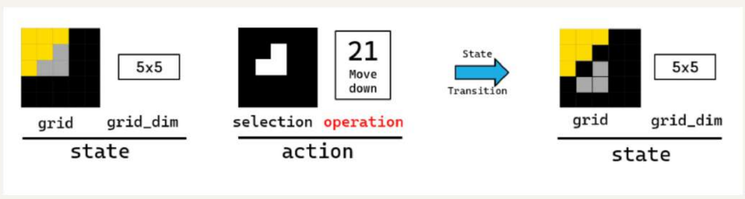
01 — 2022–2024

Make ARC executable

O2ARC + ARCLE

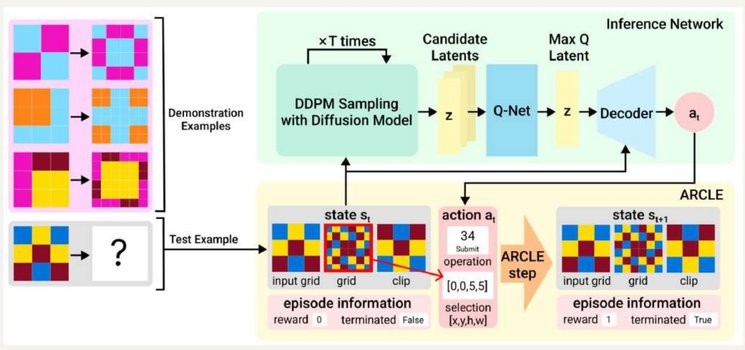
IJCAI'24 · CoLLAs'24

A web platform for solving and creating ARC tasks, and a Gymnasium environment that turns ARC into an MDP, with reward only for a correct submission.



Autoregressive policies

CoRLW'25

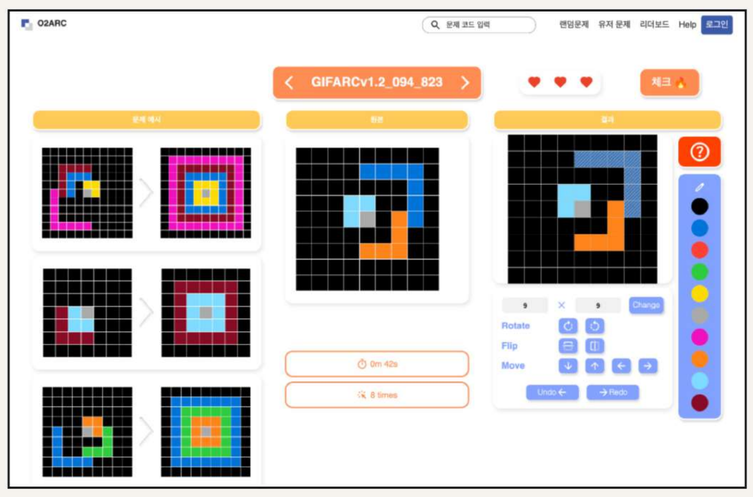


IN HINDSIGHT

Executable world models at test time are now the trend.

02 — 2024–2026

Learn from humans



ARCTraj, SOLAR

KDD'26 D&B

~15K human trajectories on 400 ARC-AGI-1 tasks. Programmatic trajectories compatible with ARCLE.

GIFARC

arXiv'25

ARC-like tasks from GIFs, with its everyday analogy. Analogy guidance shifts LLMs toward human-like solving.

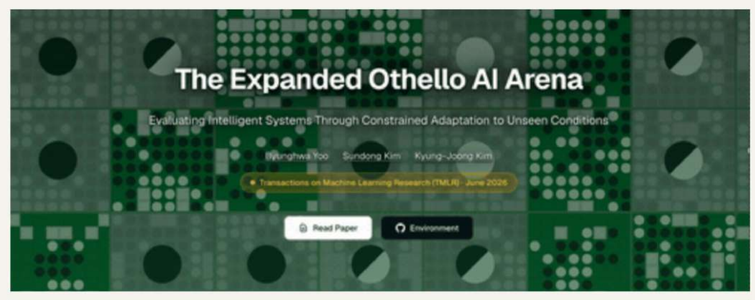


WHAT HUMANS DO

Shrink the search with priors and analogies.

03 — 2025–

Adapt under a budget

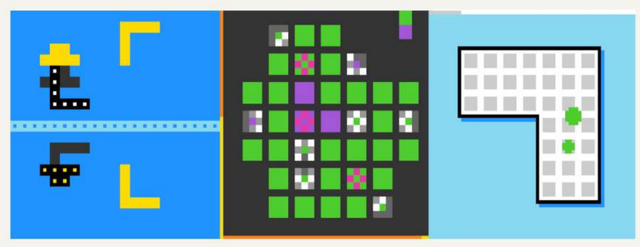


Othello AI Arena (TMLR'26): real interactions are the budget.

ARC-AGI-3 · RHAE

ARC Prize

(human ÷ AI actions)² per level.



AND NEXT

The next budget to control is the prior itself.

MEASURE SELF-EVOLUTION, NOT CAPACITY

Proposal: one frozen model, three stages

STAGE 1 · PRIOR

Pretrain freely

Any pretraining, any amount of augmented games. Submit one parametrized model plus its harness; its weights become the frozen base **θ**.

Today's Kaggle-style leaderboards are almost entirely this stage, so much that all pretraining is tolerated.

→

STAGE 2 · SELF-EVOLVE

Play your clone

θ ⇄ **θ′ = θ + ε** × ~25 games

Self-play against a clone that is the same model but slightly randomized, so it never mirrors. Build a library, fine-tune a LoRA, invent an augmentation strategy: anything that produces **δ**.

Same model, same memory and GPU budget, during the games.

→

STAGE 3 · TRANSFER

Beat your past self

New, unseen games. The evolved model **θ + δ** plays the frozen base **θ** it started from.

SELF-EVOLUTION SCORE

win rate(θ + δ vs. θ)

A bigger prior strengthens both players equally, so it cancels out.



SMALLEST TESTBED

A Puyo Puyo battle

Two players, short games, combinatorial chains. Planned chains often break and must be re-planned, so there is room to improve within a few games.

HAPPY TO SEE THE COMMUNITY REVIVING FORGOTTEN AXES

1. Accuracy
2. Accuracy at equal cost per task
3. RHAE: + real environment interactions
4. **δ: self-evolution, prior held fixed** ← this poster
5. **ARC-4/5:** define, in advance, the metrics we will need to coexist with superintelligence

LET'S TALK

How many self-play games make a fair Stage 2?
How much randomness should the clone get?
What else is a minimal testbed?

Sundong Kim
sundong.kim/publications
LinkedIn →

