

GPT-6 Astra jumps to 62.7% on ARC-AGI-3, five months after models scored 0.5%

Source copy kept by wheresthe.ai for the story "GPT-6 Astra jumps to 62.7% on ARC-AGI-3, five months after models scored 0.5%" (<https://wheresthe.ai/d/arc-agi-3-gpt-6-astra-2026-09/>).

Source: <https://arcprize.org/blog/astra>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 0ebde82097e109bbc269efab02bc8b86282ff762f23843d1591f12914642ceab

By Greg Kamradt

Published 03 Sep 2026

OpenAI's GPT-6 Astra on ARC-AGI-3

Summary

GPT-6 Astra scores 62.7% for \$26K on ARC-AGI-3 Semi-Private with our Standard harness Standard harness enables a model to carry forward notes it chooses to keep with it throughout the environment. , and 99.9% for \$19K with a Provider Adapter harness The Provider Adapter harness preserves opaque reasoning state between requests and uses compaction for longer conversations, allowing the model to reuse prior work. .

GPT-6 Astra surpasses the human baseline in action efficiency on ARC-AGI-3. It used fewer actions than the median tested human on 96% of levels.

A key behavior observed in GPT-6 Astra was its ability to turn unfamiliar environments into compact symbolic world models. It represented game mechanics as logical rules and developed its own domain-specific language shorthand to track state and plan actions.

ARC-AGI-3

ARC-AGI-3 is a benchmark for studying agentic intelligence through novel, abstract, turn-based environments. Agents must explore, infer goals, and build internal models of environments to effectively plan actions without explicit instructions. You can play ARC-AGI-3 yourself .

Your browser does not support embedded video.

These environments only contain core knowledge priors and are difficulty-calibrated through controlled testing with human participants. Humans can solve 100% of the environments .

The goal of the ARC-AGI series is to measure the "residual gap" between current artificial intelligence and AGI. We define AGI as a system's ability to acquire any skill a human can, as efficiently as a human can.

ARC-AGI-3 is the third generation of the ARC-AGI benchmark series . It tests agentic capabilities beyond ARC-AGI-1 and ARC-AGI-2 . Each generation expands on the one before it - as frontier AI capabilities advance, our benchmarks must advance with them.

ARC-AGI-3 tests four components of agentic intelligence:

Exploration: In real-world environments, information is rarely provided passively. Agents must actively obtain it by interacting with their surroundings.

Modeling: Agents must turn raw observations into a generalizable model that can predict future states and outcomes.

Goal-setting: Agents must identify target future states with only sparse rewards.

Planning and execution: Agents must map a path from their current state to a goal, course correcting as new information appears.

Astra Results

GPT-6 Astra achieves state-of-the-art scores on ARC-AGI-3 with both the Standard and Provider Adapter

harnesses. Higher reasoning levels generally cost less because Astra solves games in fewer actions, reducing the total number of model calls and tokens. View the full results . With our Standard harness Standard harness enables a model to carry forward notes it chooses to keep with it throughout the environment. , OpenAI's Astra (max) scores 62.7% on ARC-AGI-3 Semi-Private for \$26K. With the Provider Adapter harness The Provider Adapter harness preserves opaque reasoning state between requests and uses compaction for longer conversations, allowing the model to reuse prior work. , Astra (high) scores 99.9% for \$19K. Both are state-of-the-art scores. See the full leaderboard .

At max reasoning effort, Astra solves games more efficiently, requiring fewer actions and therefore lowering total cost relative to the other reasoning-effort levels. Reasoning effort Standard harness Standard harness enables a model to carry forward notes it chooses to keep with it throughout the environment. Provider Adapter harness The Provider Adapter harness preserves opaque reasoning state between requests and uses compaction for longer conversations, allowing the model to reuse prior work. max 62.7%, \$26,098 98.6%, \$17,332 xhigh 59.3%, \$37,317 98.4%, \$18,147 high 54.8%, \$40,705 99.9%, \$18,817 medium 38.6%, \$48,090 98.4%, \$19,285 low 17.5%, \$38,166 98.0%, \$21,298 none 35.2%, \$49,791 96.7%, \$23,457

For a cost comparison, during our controlled testing, human participants were paid \$115 per 90-minute session, plus \$5 per game completed. Participants attempted approximately nine games per session, roughly \$12.78 per attempted game before bonuses.

Most of this fee pays for the participant's time and willingness to take the test, rather than the energy their brain uses (a closer proxy to compare with AI). If we look at only the brain's energy, and price it as electricity, the estimate drops to about 0.6 cents per session, or 0.067 cents per game attempted. 1

Analysis

Beyond the scores, Astra's replays show how it turns unfamiliar game mechanics into useful working models. Three findings stood out: the compact algebraic notation it develops, its action efficiency compared with humans, and the custom tools it builds.

Custom Algebraic Notation

When playing ARC-AGI-3, Astra chooses which strategy notes it would like to carry forward. It tracked objects, coordinates, rules, and unfinished plans, while also using a custom domain-specific language notation it generated for the environments.

We've seen similar behavior in other models , but Astra's notes stood out for their precision and information density. It distilled the scene into a compact code-like symbolic model: where objects were, how they interacted, and exactly which actions needed to happen in what order. This is an on-the-fly algebraic shorthand rather than a fully fledged programming language. For example:

Game state: L8: hub q2 (8↓). Lengths: 14=1... records the level, a local rotation index, and mechanism lengths. s5i5, frame 219

Multi-step plans: extend8 to3; retract10 to2; shorten8 to1 records an ordered sequence of changes to the color-8 and color-10 mechanisms. s5i5, frame 219

Controls and coordinates: 9--=(39,4), rotate=(49,18), 14+=(59,11) maps operations to the coordinates of the controls that perform them. s5i5, frame 235

Time and position: Turn 5: P=(24,20), empty, facing west combines a turn counter with the player's location, carrying state, and orientation. wa30, frame 708

Astra playing s5i5 , using its on-the-fly algebraic shorthand to track state and plan actions. Action Efficiency Compared to Humans

Before launching ARC-AGI-3, we tested approximately 500 members of the general public to establish a human baseline for action efficiency, or simply, how quickly did people solve each environment. Participants were not selected for puzzle-solving experience or ability. 2

For each level, we defined the "human baseline" using the median action count among players who completed it. This gives us a reference for comparing human and AI performance. An AI that needs more actions is less action-efficient, while one that needs fewer actions is more action-efficient.

In the Provider Adapter harness The Provider Adapter harness preserves opaque reasoning state between requests and uses compaction for longer conversations, allowing the model to reuse prior work. , Astra (max) used fewer actions than the human baseline on 96.0% of levels and used 51.7% fewer actions per level on average . This is a material milestone. This means by ARC-AGI-3's measure of action efficiency, Astra matched and surpassed human parity.

As an aside, before we launched ARC-AGI-3, we hypothesized that action efficiency would remain a dividing line between humans and AI. We anticipated that even when an AI solved an environment, it might require substantially more exploration (actions) than a person. That remains true of brute-force approaches, but frontier AI shows a more binary-like pattern. Once frontier AI “understands” the mechanics, it generally executes within the range of human efficiency.

Astra's Action Efficiency Compared to Humans

Each dot represents one level that Astra (max) completed. Points below the solid line indicate fewer actions than the human baseline.

The plot above compares the number of actions Astra used to complete each level with our human baseline. This reinforces why ARC-AGI-3 measures action efficiency, not just task completion. A completion-only score would tell us that Astra completed an environment, but not how efficiently it learned to solve them.

Most benchmarks only measure cost efficiency, which measures the computational resources used, but action efficiency measures how much experience with an environment was required.

Astra's results show that it needed fewer interactions than the human baseline to execute a solution.

Custom Tools in Agent Harness

We also evaluated Astra in the PRO-LONG harness (paper), an early ARC-AGI-3 red-teaming partner. In this advanced setup, Astra had access to a sandbox where it could execute custom code 3 .

We observed Astra create a custom set of tools for each game: board parsers, game-state models, search algorithms, planners, and persistent notes. For more involved runs, Astra even produced small, game-specific software libraries.

For example, in tu93 , a maze-like game with guards and moving patrols, Astra started with navigation and built `maze_solver.py` . It added combat rules in `combat_solver.py` , modeled moving patrols in `patrol_solver.py` , and used `sync_state.py` to check its predictions against observations.

Examining Astra's performance in PRO-LONG is useful because we see what it can do with external tools. However, this represents different evaluation conditions from our controlled human testing. Our testing participants did not have a code interpreter, scratch pad, etc., so PRO-LONG's results should be understood as the combined performance of the model and its tools.

Astra playing tu93 in the PRO-LONG harness. Two Harnesses, Two Questions

Our Standard harness for ARC-AGI-3 asks how models compare under the same minimal, provider-neutral interface. It provides all the information required to solve each game, but leaves the model responsible for deciding what to preserve in its visible notes. We believe a future AGI should be able to solve ARC-AGI-3 under these conditions. The shared interface also gives us a consistent, apples-to-apples comparison across providers.

Alternatively, there is a separate question: how well does a model perform when it can use the context-management features its provider designed for it? For Astra, this means preserving the opaque reasoning state (which we don't see) between requests and using compaction to manage longer conversations.

With the Provider Adapter harness, Astra's best observed score on ARC-AGI-3 Semi-Private increased from 62.7% to 99.9%. Looking across Public and Semi-Private and all reasoning levels, Provider Adapter runs were approximately 3.66x faster by aggregate recorded elapsed time and used 49% fewer total tokens across the 167 game-reasoning pairs both harnesses solved.

Going forward, we will report both Standard harness and Provider Adapter harness results on the ARC-AGI leaderboard, with each evaluation condition clearly labeled. Our open-source testing repository and testing policy document both approaches.

ARC-AGI Series

ARC-AGI-3 continues to be a useful playground for researchers and agents to explore unfamiliar environments, discover rules, and learn through interaction. Astra's results are also a major milestone worth celebrating. From our perspective, Astra represents a noticeable step-function change in frontier model capabilities.

When we launched ARC-AGI-3, we made it clear that saturating the benchmark would not represent "proof of achieving AGI." Therefore, while we believe Astra represents meaningful progress towards generalization, we are not claiming that it is AGI.

The ARC-AGI benchmark series is designed to evolve in tandem with frontier AI. This creates a feedback loop between emerging research questions and advances in AI capabilities. ARC-AGI-3 was our first interactive benchmark, which asked AI to efficiently synthesize causal world models and achieve goals without specific instructions. Astra clears this bar. At the same time, ARC-AGI-3 has a tightly bounded scope and format, and its environments have deterministic, closed-ended mechanics and goals. It does not represent the complexity and open-endedness of the real world.

We are actively exploring the questions that should shape the next generation of benchmarks, including how to evaluate recursive self-improvement and open-ended innovation. Astra's progress helps clarify which AI capabilities are out of reach and which questions remain open.

Thank you to François Chollet, Mike Knoop, Matt Mazur, Ethan Bond, and Derek Smith for early review of this post.

Assuming 20 W of brain metabolic power and an electricity price of \$0.20/kWh: $0.020 \text{ kW} \times 1.5 \text{ hours} = 0.030 \text{ kWh}$, worth \$0.006 per session, or $\$0.006 \div 9 \approx \$0.00067$ per attempted game. See the ARC-AGI-3 human testing paper . No evidence of trying to break out of the sandbox was observed.