

Recursive-Play: AI-Generated Interactive Challenges for Improving AI Agents

EnvLoop Research
research@envloop.ai

Abstract

AI-assisted task construction can supply executable challenges that evaluated AI agents do not yet complete, creating a concrete resource for subsequent improvement. Recursive-Play implements this idea with 150 interactive puzzle environments, ten mechanic families and six sequential levels per task. Author-only constructive witnesses and deterministic state verifiers establish reachability, while players receive pixels and neutral controls without the rules or solutions. Across eleven configurations and 1,650 independently eligible outcomes, 15 tasks remain unfinished by every configuration under the fixed 512-action/720-second protocol; two yield no completed level for any configuration. Independent replay verifies all 900 canonical seed-level witnesses within the action budget. Partial progress and matched traces identify different kinds of learning targets beyond binary failure. We specify a subsequent improvement loop: select verified failures, acquire corrected trajectories, train a candidate, and test it on newly frozen held-out tasks with no-regression checks. The study establishes the challenge supply and evaluation substrate; training gains remain to be tested.

1 Introduction

A useful improvement process needs a continuing supply of tasks that are valid, informative and still challenging for its learner. AI coding agents can help construct that supply: they can implement unfamiliar mechanics, generate instances and provide executable checks of success. A model then encounters the resulting world through restricted observations and must discover its rules through interaction. Creating a world with access to its construction is a different role from solving it blind.

Recursive-Play studies this practical route from AI-assisted task generation to further agent improvement. Its central finding is a verified challenge set: 15 of 150 canonical tasks remain incomplete for every one of eleven evaluated configurations, although withheld reference actions establish that each curriculum is reachable. Two tasks produce zero completed levels across all configurations. These outcomes show that AI-assisted construction can expose unresolved performance gaps under a fixed evaluation protocol; they do not require claiming that the tasks are impossible or that every AI system would fail.

The contribution has three parts: an executable, resettable task bank with independent verifiers; a matched evaluation that separates complete success, partial progress and normal budget exhaustion; and an explicit pathway from verified failures to corrective data and prospective training. The ranking table and recorded examples describe where that pathway has useful inputs. They are evidence about the generated tasks and tested configurations, rather than the endpoint of the project.

The next experiment is to turn those inputs into persistent gains: obtain verified trajectories, train a candidate and assess transfer on a distinct frozen task set. The RSI framing of [Duan et al. \[2026\]](#) makes the distinction between solving one task and retaining an improvement across tasks precise. The present work supplies and validates challenges for that later experiment.

2 Related work and scope

ARC-AGI-3 studies instruction-free interaction with compositional mechanics and human-calibrated efficiency [[ARC Prize Foundation, 2026](#)]. POET and PAIRED connect generated environments to learning or regret-driven challenge

selection [Wang et al., 2019, Dennis et al., 2020]; TextWorld and EscapeBench supply other interactive evaluation settings [Côté et al., 2019, Qian et al., 2025]. RSI Bench, RSIBench-Data, AI4AI-Bench, and PAST-Bench evaluate different improvement targets and persistence claims [Mahmoud et al., 2026, Meng et al., 2026, Chi et al., 2026, Xue et al., 2026]. STOP and the Darwin and Red Queen Gödel Machines study program improvement with distinct evaluation controls [Zelikman et al., 2024, Zhang et al., 2026, Iacob et al., 2026]. Recursive-Play contributes AI-assisted task construction, resettable interactive challenges and recorded solver outcomes as a substrate for prospective agent improvement.

3 AI-assisted construction and evaluation protocol

Constructing a verifiable challenge supply. The recorded generation plan delegates ten mechanic-family assignments to AI coding-agent sessions; it reports eight generator sessions, with two reused and at most three running concurrently. The delivered TypeScript modules implement initialization, transitions, observations, terminal predicates and author-only reference witnesses under a shared construction contract. Contracts, source review and executable audits constrain this AI-assisted workflow. Generator checkpoint identities are not recorded, so this is not a controlled same-model creation-versus-solving experiment.

A separate canonical-instance audit replays six witnesses for each of the 150 scored environment–seed pairs: 900 level instances, 20,092 reference actions and zero failed witnesses. The largest six-level reference sum is 347 charged actions, below the 512-action allowance. References establish reachability within the action allowance; they are privileged author artifacts, not a demonstration that a blind agent could discover them or an optimality proof. This separation lets AI-assisted authoring produce valid learning targets without leaking their construction to the solver.

Each environment defines initialization $s_0 = G_e(l, z)$, transition $s_{t+1} = T_e(s_t, a_t)$, observation $o_t = O_e(s_t)$, and terminal predicate $V_e(s_t)$ for level $l \in \{1, \dots, 6\}$ and seed z . A withheld constructive witness establishes reachability for each audited instance, not optimality or human discoverability. The ten author families contain 15 environments each. Players receive a palette-index grid of at most 64×64 pixels, an equivalent PNG, neutral buttons 1–6, optional click coordinates, and administrative status. Mechanic names, seeds, hidden state, and witness actions are withheld. Browser and native Godot clients render the same authoritative TypeScript game rules; scored episodes use the game/tool interface rather than a computer-use trial in either client.

Every fresh episode has at most 512 charged actions and 720 seconds from first observation across its six levels. Reset costs one action; advancing after a server-confirmed win is free. A normal `time_budget` or `action_budget` ending is a valid outcome, with attained levels retained. For model m , full success is $100W_m/150$ percent and level completion is $100L_m/900$ percent. The primary case selector takes the first independently eligible attempt for each model–environment pair regardless of score. A confirmed provider, network, authentication, or host interruption is excluded only after evidence review; its raw trace and resource use remain in an operational ledger. This rule cannot be used to turn a low score into a fresh chance.

What a 720-second ending means. An ordinary `time_budget` ending means that the six-level curriculum was unfinished when the authoritative 720-second game clock expired. It is a scored task outcome, not an engineering-failure category: any completed levels still count. The clock includes model-response waiting, tool execution, and puzzle interaction. Thus the recorded endpoint establishes incomplete completion within the allowed time; it does not determine whether reasoning, exploration, action efficiency, or ordinary serving latency was the limiting factor. These ordinary endings are selected only after independent eligibility review; the terminal label alone does not establish a clean attempt. An in-flight response may be canceled when the game deadline is reached, which by itself is not a pre-deadline engineering fault. Independently confirmed provider, transport, authentication, or host faults are recorded separately and excluded under the evidence-review rule above.

The 1,650 primary rows represent eleven configurations times the same 150 environment IDs and seeds. Prior 930 rows are reused unchanged. Three models had an earlier prespecified ten-case pilot; those ten are included in their final 150, not counted as additional experiments. The bank was exposed during development, including its ten-case panel, and is not a sealed test set. Source hashes, case-level CSV, independently audited new-cohort manifest, and operational evidence accompany the article.

4 Generated challenges that remain unfinished

The primary finding is that 15 tasks have no six-level completion among any of the eleven tested configurations. They form a verified unresolved subset of the generated bank, with two tasks producing zero cleared levels for every configuration. Table 1 supplies the wider context by comparing complete-task success with partial level progress on the same 150 environments. A task is complete only after all six sequential levels have been cleared. Comparative model tables retain this model order; the sensitivity table is not ranked by its conditional win rates. Other tables state their own ordering.

Table 1: Matched-bank results, ranked by complete-task success. Equal complete-task counts retain tied ranks; cleared levels order tied rows for display only.

Rank	Model	Complete tasks		Cleared levels		720-second endings
		Count / 150	Rate (%)	Count / 900	Rate (%)	Count
1	GPT-6 Astra	135/150	90.00	848/900	94.22	6
2	GPT-5.6 Sol	29/150	19.33	357/900	39.67	73
3	GPT-6 Sol	22/150	14.67	369/900	41.00	122
4	GPT-5.6 Terra	20/150	13.33	282/900	31.33	69
5	GPT-5.6 Luna	4/150	2.67	160/900	17.78	118
6	Qwen 3.8 Flash	2/150	1.33	114/900	12.67	125
6	GPT-6 Luna	2/150	1.33	89/900	9.89	143
8	Qwen 3.8 Max 0902	1/150	0.67	88/900	9.78	149
9	DeepSeek V4.1 Flash	0/150	0.00	63/900	7.00	150
9	Kimi K3	0/150	0.00	38/900	4.22	150
9	GLM 5.3 Flash	0/150	0.00	28/900	3.11	150

How to read the table. Each model faces 150 tasks with six levels each (900 offered levels). Complete-task rate = completed six-level tasks / 150; cleared-level rate = completed levels / 900. The final column counts ordinary 720-second endings with the curriculum unfinished; completed levels still count. These eligible endings remain in both denominators.

Worked row: GPT-6 Sol. It completes 22/150 tasks (14.67%), clears 369/900 levels (41.00%), and reaches the ordinary game deadline in 122 tasks. Thus 122 unfinished tasks does not imply 122 tasks with zero progress.

GPT-6 Astra has the largest observed full-curriculum count (135/150) and clears 848/900 offered levels. Across all eleven rows, 538 episodes end after clearing one to five levels; the win column alone would hide this partial progress. These are descriptive observations on a development-exposed bank, not evidence of a model-only effect.

Table 2: Terminal outcomes for the same primary rows. Counts sum to 150 within each model; ordinary time and action exhaustion are not classified as engineering faults.

Model	All six levels complete	720-second limit	512-action limit	Abandoned
GPT-6 Astra	135	6	9	0
GPT-5.6 Sol	29	73	47	1
GPT-6 Sol	22	122	6	0
GPT-5.6 Terra	20	69	59	2
GPT-5.6 Luna	4	118	23	5
Qwen 3.8 Flash	2	125	0	23
GPT-6 Luna	2	143	4	1
Qwen 3.8 Max 0902	1	149	0	0
DeepSeek V4.1 Flash	0	150	0	0
Kimi K3	0	150	0	0
GLM 5.3 Flash	0	150	0	0

Of 1255 ordinary 720-second endings, 484 retain at least one completed level. Removing them from the score

population would discard observed progress. The non-timeout denominator ranges from 0 to 144 across models; 3 row(s) have an undefined conditional win rate. Appendix B gives all depth distributions and family totals, and Appendix E lists each of the 1,650 primary outcomes. Offered but unvisited later levels are not separately attempted failures. Model-authored code errors do not automatically qualify as engineering exclusions.

4.1 A verified subset beyond observed completion

Table 3: Witness-valid curricula unfinished by all eleven evaluated configurations. Rows are ordered by maximum attained depth (ascending), then reference action count, then environment ID.

Task	Seed	Family	Best depth out of 6	Astra depth out of 6	Reference actions
b010	677335	Circuits	0	0	240
c012	804039	Objects	0	0	278
a009	558550	Spatial	1	1	86
h008	1334612	Coordination	2	2	95
i006	1421721	Topology	2	2	110
h015	1382126	Coordination	2	2	124
j007	1548425	Composition	2	2	199
b015	716930	Circuits	2	1	231
c010	788201	Objects	3	3	76
i011	1461316	Topology	3	3	121
j001	1500911	Composition	3	3	347
c001	724849	Objects	4	4	128
a004	518955	Spatial	5	5	132
g002	1176232	Memory	5	5	148
j012	574388	Composition	5	5	163

Interpretation. All 15 tasks have zero complete six-level wins among the 11 selected episodes per task; b010 and c012 have zero completed levels for every configuration. Reference counts sum six independently replayed constructive paths at the same canonical seed, each curriculum requiring at most 512 actions. These privileged paths were withheld from evaluated agents: their existence proves reachability, not instruction-free discovery within 720 seconds or human-calibrated difficulty. Partial progress remains credited. The evidence inventory retains per-level reference counts, full model depth vectors, source hashes and quality flags.

4.2 Ranking, partial progress, and observed terminations

Figure 1 separates complete curricula from attained levels. GPT-6 Astra leads both measures (135/150 full curricula; 848/900 levels). GPT-5.6 Sol ranks second by full curricula (29/150), ahead of GPT-6 Sol (22/150), whereas GPT-6 Sol clears more total levels (369 versus 357). The distinction is substantive: GPT-6 Sol has 96 episodes at depths 1–5 and 32 at depth zero, compared with 71 and 50 for GPT-5.6 Sol. Their non-full episodes contribute 237 and 183 cleared levels, respectively. Neither measure is a replacement for the other.

The two configurations with two full curricula remain tied at primary rank 6: Qwen 3.8 Flash clears 114/900 levels and GPT-6 Luna clears 89/900. The zero-win configurations remain tied at primary rank 9 even though DeepSeek V4.1 Flash, Kimi K3, and GLM 5.3 Flash clear 63, 38, and 28 levels. Ordering tied rows by level totals is a display convention, not an additional unique full-win rank. These contrasts show why zero full wins should not be represented as zero observed progress.

Figure 3 localizes the descriptive differences. GPT-5.6 Sol clears 70/90 levels in Relations and 6/90 in Objects, with 9/15 versus 0/15 full curricula. GPT-6 Sol clears 60/90 levels in Spatial, 53/90 in Circuits, and 59/90 in Resources. GPT-6 Astra clears all 90 offered levels in Temporal, Relations, and Resources. The labels are the paper’s author categories; these development-exposed families do not isolate independent skills or identify a cognitive mechanism.

Figure 4 keeps the terminal categories visible. GPT-6 Sol has 122 ordinary 720-second endings and 6 action-budget endings; GPT-5.6 Sol has 73 and 47. GPT-6 Luna has 143 time-budget endings, and DeepSeek V4.1 Flash, Kimi K3, and GLM 5.3 Flash each have 150. These are eligible outcomes, not engineering exclusions. Across all configurations, 484 of the 1,255 time-budget endings still clear at least one level. These episodes did not finish all six levels within the allowed time; a time-budget ending does not mean no levels were completed, and cannot by itself identify the limiting cause.

Observed episode time and resource use need a separate interpretation. Astra’s median recorded episode duration is 280.600 seconds, compared with 716.471 for GPT-5.6 Sol and 720.002 for GPT-6 Sol. The clock includes provider latency, tool use, and gameplay, and many medians sit at the budget boundary; this is not decoder throughput or a model-only speed comparison. Known token totals are subtotals when usage is missing, and cached input/reasoning output are subsets rather than extra tokens. GPT monetary costs remain unpriced; API known-cost estimates coexist with retained unknown-use reservations, whose amounts are not invoices. Serving setups and tools differ, so the figures rank the observed configurations on this bank without assigning their differences solely to model capability or price efficiency.

Complete tasks and cleared levels

Same model order, two denominators; primary ties are retained.

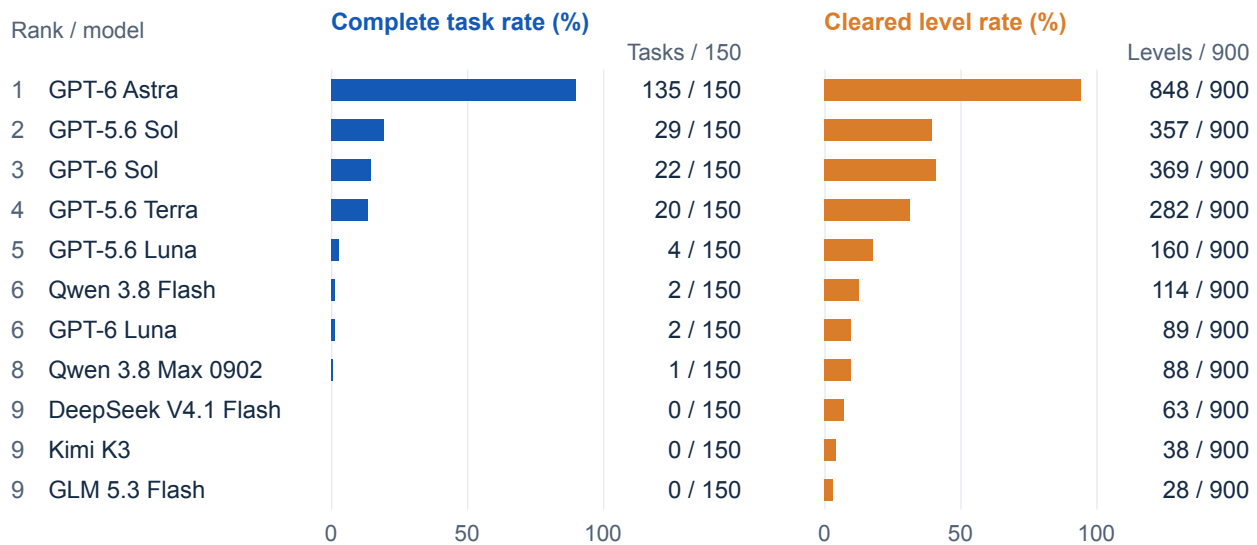


Figure 1: Aligned panels compare full-curriculum success and level completion on the same percentage scale, using 150 episodes and 900 offered levels. Exact counts appear in fixed columns rather than at the ends of small bars. Rows follow the primary full-win ranking; equal win counts retain tied ranks.

Completion depth, episode by episode

All seven depths are retained; every model has 150 episodes.

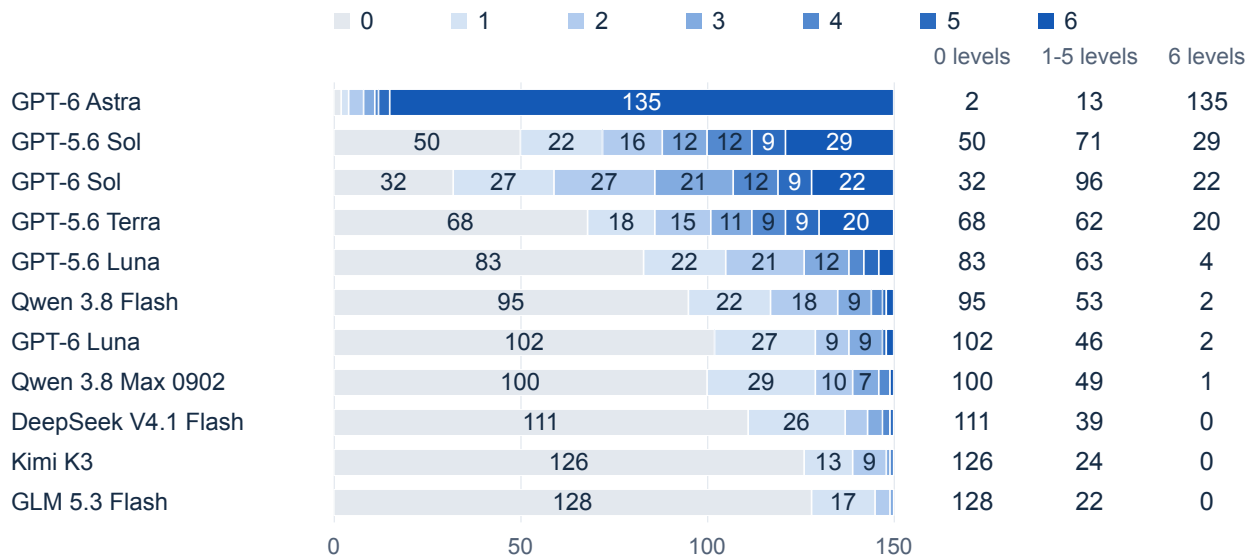


Figure 2: Completion depth retains all seven categories, with each row totaling 150 episodes. Labels within sufficiently wide segments are episode counts; aligned columns summarize zero, partial (1–5) and complete (6) depth. Later unvisited levels are not separate attempted failures.

Where levels were cleared

Cell values are cleared levels out of 90; color encodes the percentage.

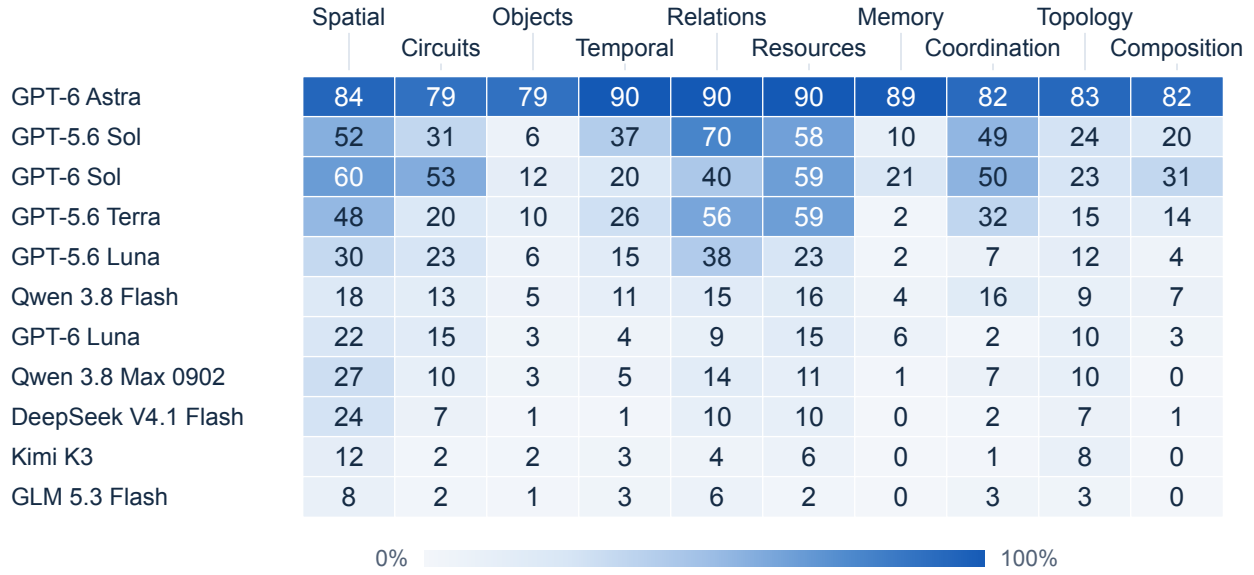


Figure 3: Level completion by authored family. Each cell offers 15 environments times six levels; printed cell values are levels cleared out of 90, and color shows the corresponding percentage. Family names follow the paper’s author mapping rather than a claim about isolated mechanisms.

How eligible episodes ended

Normal budget exhaustion means unfinished; engineering exclusions are separate.

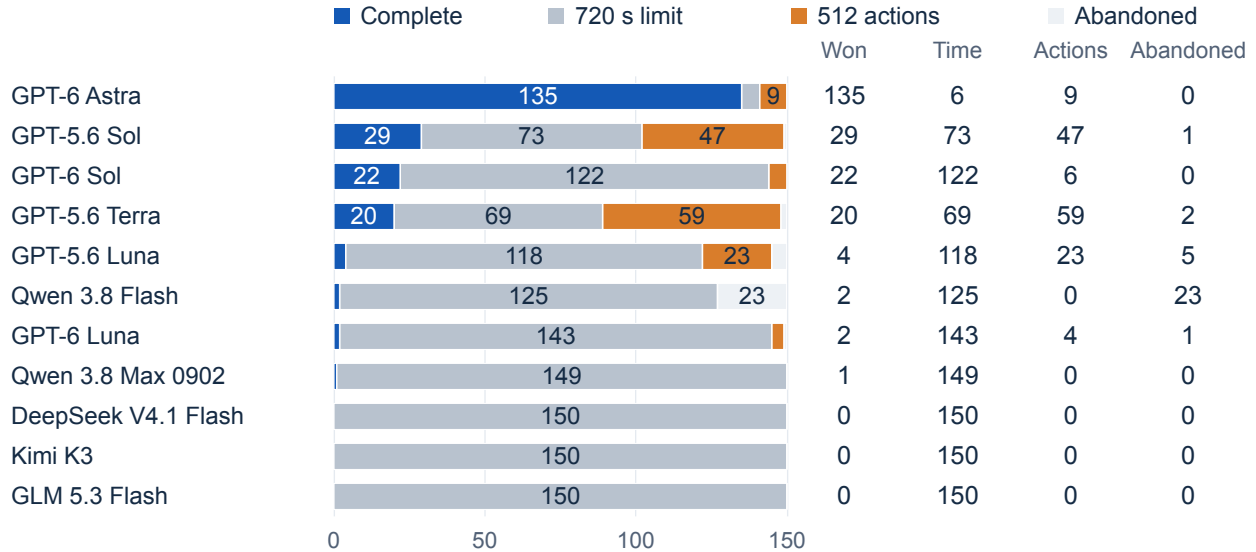


Figure 4: Mutually exclusive terminal outcomes for 150 eligible episodes per model. Fixed columns retain exact counts even for narrow segments. The time-limit category denotes an unfinished curriculum at the ordinary 720-second deadline, with partial progress retained; confirmed engineering interruptions remain separately excluded.

5 Recorded examples of improvement targets

5.1 Worked example: same puzzle, different completion depths

Figure 5 compares three selected a001 episodes (seed 503117) with an identical initial image and six-level curriculum. Astra clears six levels in 33 charged actions and 174.151 seconds; DeepSeek V4.1 Flash clears five in 38 actions before its 719.989-second game-budget ending; GPT-6 Luna clears three in 46 actions before its 720.000-second ending. The latter two are valid partial outcomes, not engineering exclusions.

In the authored spatial puzzle *Pivot cargo*, rigid colored pieces must match the target board; fixed obstacles and other pieces can block translation or rotation. The agent sees neutral numbered controls: in the source implementation, 1–4 translate the selected piece, 5 rotates from level 2 onward, and a coordinate click through 6 selects a piece. Level 4 adds a second colored body. This describes the environment source, not rules explained by a model. The shared level-4 start and differing final images expose attained levels and remaining mismatches.

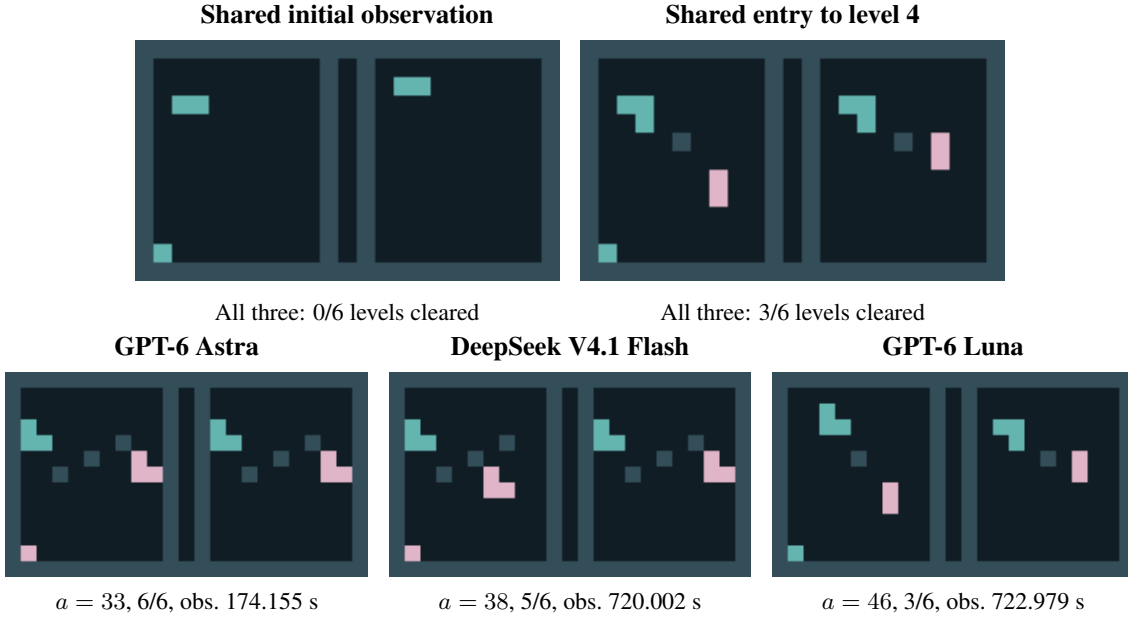


Figure 5: Actual recorded observations on a001. The shared initial and level-4-entry PNGs are each byte-identical across the three traces, so each is shown once. Initial offsets are 0.004 s (Astra and Luna) and 0.002 s (DeepSeek). Entry occurs at $a = 11, 87.398 \text{ s}$ for Astra; $a = 13, 171.253 \text{ s}$ for DeepSeek; and $a = 13, 207.874 \text{ s}$ for Luna. The three final PNGs expose retained completion depths. In each image, the left board is current state and the right board is the target; a counts charged actions. Observation offsets may follow the authoritative terminal timestamp: Luna’s last image does not extend its 720-second budget. Astra and DeepSeek show level 6, whereas Luna remains at level 4. All original nine frames and their hashes remain in the prior source and supplemental provenance; no simulation or decoded illustrative video replaces them.

Logged controls and confirmed transitions locate each win. All three clear level 1 after button 1 at action 1. Level 2 clears after button 4 at action 6 for Astra and action 8 for DeepSeek and Luna; level 3 clears after button 2 at actions 11 and 13, respectively. Thus the identical level-4-entry image is reached with different action counts and wall times. Astra then clears levels 4–6 at actions 15, 24 and 33. DeepSeek clears levels 4 and 5 at actions 16 and 32, with no sixth-level win; Luna clears no further level after entering level 4. These are recorded status transitions, not inferred internal representations or explanations of why a model became stuck.

This purposive example links binary success, depth, actions and time; it estimates neither typical behavior nor the full 150-case results. Astra and DeepSeek belong to the immutable historical cohort, Luna to the separately audited follow-up; dates, settings and serving routes differ. Faster completion alone establishes neither better cognition nor model weights as the cause. No attempt was rerun or substituted.

5.2 Task-specific reversals: Local frame rover

These three additional tasks were selected to illustrate contrasts that the aggregate table cannot show: a task-specific ranking reversal, zero completion under two different budgets, and graded progress on a constraint problem. They are purposive examples, not a representative sample. Each panel uses the same environment and seed within its task; serving routes, settings and dates still differ across configurations.



Shared initial state

Task a005, seed 526874. *Local frame rover* requires matching the left board to the right target. The rover moves forward or backward relative to its own heading; turning and special floor cells change that frame. Later levels add reflection and cargo exchange from the rover’s rear. These mechanics come from the environment source; the solver receives pixels and neutral controls.



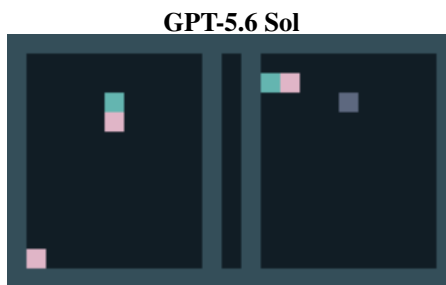
6/6 levels · 30 actions · 158.739 s
Full completion; final image at level 6



6/6 levels · 39 actions · 717.969 s
Full completion; final image at level 6



5/6 levels · 154 actions · 720.002 s
Time budget; final image at level 6

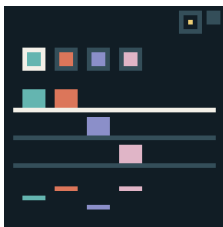


1/6 levels · 512 actions · 483.341 s
Action budget; final image at level 2

Figure 6: Actual terminal observations from a005. The shared initial PNG is byte-identical across all four episodes; Astra and Qwen also have byte-identical terminal PNGs. All displayed times are authoritative game elapsed times, not screenshot capture times. Panels can show different curriculum levels, so they are not a same-level visual-distance measure.

What the trace adds. GPT-6 Sol completes level 5 at 248.237 seconds but records no sixth-level completion before the deadline. Qwen completes level 5 much later, at 637.106 seconds, then completes level 6 at 717.969 seconds. Thus early progress alone does not predict the final result. GPT-5.6 Sol ends at the 512-action cap, well before 720 seconds, with one level completed and 123 recorded resets. This task reverses the aggregate Sol ordering: GPT-6 Sol reaches five levels here, while GPT-5.6 Sol reaches one. The aggregate table still places GPT-5.6 Sol ahead on full completions across all 150 tasks.

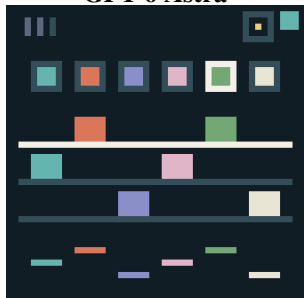
5.3 Zero completed levels under different budgets: Escrow Cycle



Shared initial state

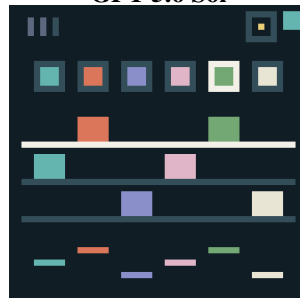
Task d009, seed 899067. *Escrow Cycle* moves colored assets between accounts. A transfer is reserved first and settles only after clock-advance actions. Later levels add unequal delays, capacity limits, atomic netting and delayed compensation. The image shows token identities at the top, account holdings in the middle, and target ownership as colored marks near the bottom. A transfer request alone does not establish that the target has been reached.

GPT-6 Astra



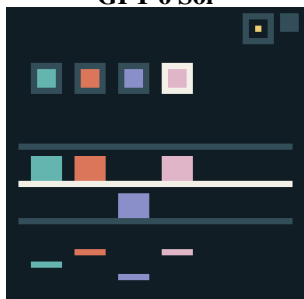
6/6 levels · 96 actions · 135.827 s
Full completion; final image at level 6

GPT-5.6 Sol



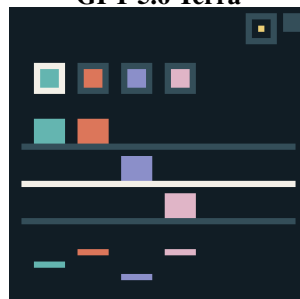
6/6 levels · 115 actions · 661.809 s
Full completion; final image at level 6

GPT-6 Sol



0/6 levels · 45 actions · 720.002 s
Time budget; final image at level 1

GPT-5.6 Terra



0/6 levels · 512 actions · 709.718 s
Action budget; final image at level 1

Figure 7: Actual terminal observations from d009, with an identical initial PNG across all four episodes. Zero completed levels means that no level win was verified, even if visible state changed. The two zero rows have different normal termination causes; neither is an engineering exclusion.

What the trace adds. GPT-5.6 Sol completes the first level at 171.963 seconds and subsequently completes all six; its final three recorded controls are two clock advances (button 3), then a batch-mode toggle (button 5). GPT-6 Sol records 45 actions without a first-level win before the 720-second deadline. Terra records 512 actions and seven resets, also without a first-level win, and reaches the action cap at 709.718 seconds. The shared zero score hides a large difference in action volume and stopping condition. It does not reveal why either configuration failed, or imply that issuing more actions is sufficient.

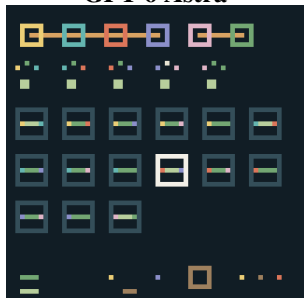
5.4 Full, partial and zero completion: Transitive islands



Shared initial state

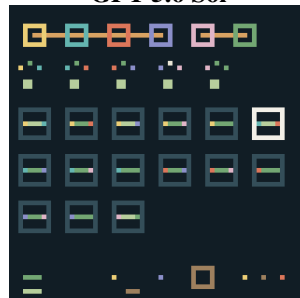
Task e005, seed 978257. *Transitive islands* requires choosing graph edges to satisfy connected and separated node-pair constraints. Later levels add an edge-count budget, path-length and degree limits, then an exact two-edge distance requirement. Nodes and active links appear at the top; pair requirements and their status indicators below; selectable edge controls occupy the middle. Buttons 1–2 change selection, 3–4 toggle an edge, and a coordinate click selects an edge.

GPT-6 Astra



6/6 levels · 41 actions · 135.514 s
Full completion; final image at level 6

GPT-5.6 Sol



6/6 levels · 101 actions · 676.009 s
Full completion; final image at level 6

GPT-6 Sol



2/6 levels · 67 actions · 720.002 s
Time budget; final image at level 3

GPT-6 Luna



0/6 levels · 9 actions · 720.003 s
Time budget; final image at level 1

Figure 8: Actual terminal observations from e005, sharing the same initial PNG. Different winning edge configurations can satisfy the same constraints: Astra and GPT-5.6 Sol both achieve 6/6 despite different final images. Their panels show level 6; GPT-6 Sol and Luna show levels 3 and 1.

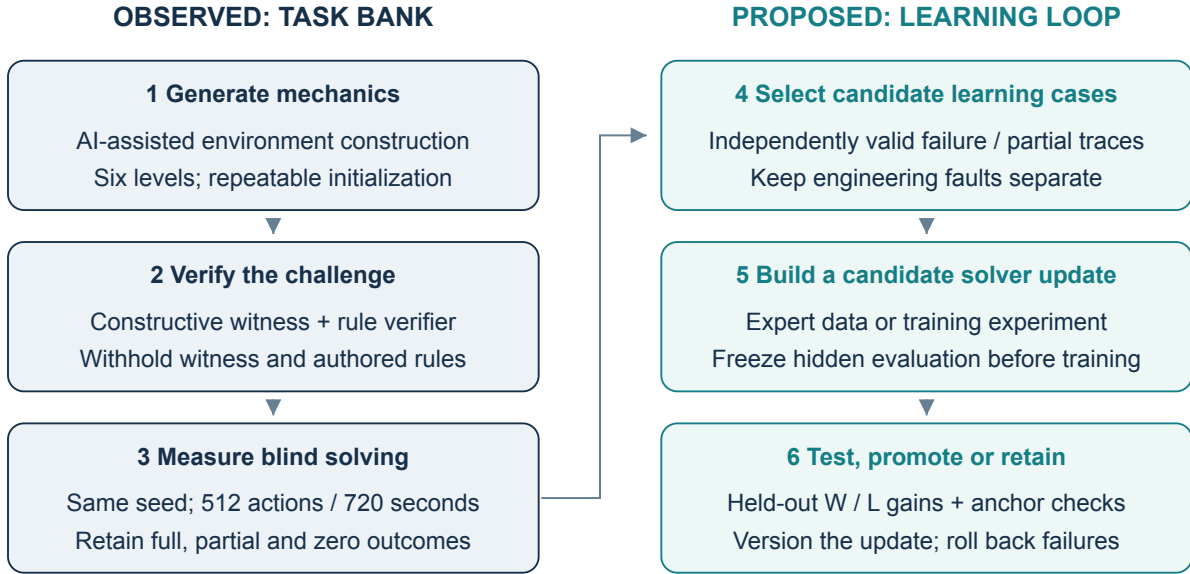
What the trace adds. GPT-6 Sol verifies its first two level wins at 213.880 and 575.905 seconds, then records no third-level win. Luna records nine actions without a first-level win. Both are normal time-budget outcomes, but only one carries partial completion credit. Astra and GPT-5.6 Sol each complete all six, with different action counts and elapsed times. The comparison makes the meaning of completion depth concrete; it does not identify model weights, reasoning quality or provider latency as the isolated cause.

All screenshots are unmodified logged PNGs. Some terminal observations were saved after the authoritative terminal timestamp; this does not grant extra game time or actions. The supplemental provenance records both clock types, all 11 outcomes for each selected task, and the selected result/event/PNG hashes. No episode was rerun, reconstructed or substituted for these panels.

6 From generated challenges to a testable improvement loop

AI-assisted generation supplies solvable, verifiable challenges that the tested players do not fully solve. Fifteen of the 150 curricula remain unfinished by all eleven configurations under the 512-action/720-second budget; in two of these tasks, no configuration completes even one level. The strongest configuration completes the other 135 curricula. These are shared challenges under the evaluated protocol, not universally impossible tasks or a ceiling on all AI systems. Serving routes, settings and dates still limit model-only attribution.

Figure 9 turns that observation into a concrete next experiment. The observed stages produce resettable mechanics, a withheld constructive witness and an executable state verifier, then record blind solving. An independently eligible failure or partial trace supplies a candidate learning case. Its observations, actions and verified level transitions can be paired with an independently checked expert solution; a witness proves reachability but does not automatically constitute good teaching data. Confirmed engineering interruptions and known/unknown resource use stay separate from the learning-case selection.



A proposed improvement loop, not a measured training result

Figure 9: The left column describes the construction, verification and blind solving measured here. The right column proposes trace selection, a solver update and an independent improvement test. No training, held-out gain or promotion outcome is reported by this study. W counts fully completed tasks and L counts verified completed levels.

Before training, freeze a separate held-out task set and seeds, withhold its rules, witnesses and evaluation traces, and retain an unchanged anchor suite. Repeating exposed public-bank cases cannot establish generalization. Compare the old and candidate solvers on identical hidden instances, fresh contexts, exact resets and the same 512-action/720-second limits.

Prespecify acceptance in terms of paired full-task wins W , completed levels L and uncertainty across the held-out tasks. Require a positive held-out gain without regression on the frozen anchor suite, while checking that gains are not explained by changed budgets or fault handling. Version accepted updates and preserve a rollback to the prior solver; rejected updates leave it in place. This makes future improvement measurable while keeping the present contribution centered on generated challenges and verified solving outcomes.

7 Time-budget sensitivity, token coverage, and resources

The time-budget-censored quantity $W_m/(150 - T_m)$ removes the T_m ordinary 720-second endings for model m . It is a *non-primary sensitivity analysis*: completion time is an outcome affected by model behavior and serving speed, and its denominator may differ substantially by model or even be zero. It cannot replace the matched 150-case comparison.

Table 4: Time-budget-censored sensitivity only. Here 720 s denotes an unfinished curriculum at the normal game limit, not an engineering fault. Undefined means every eligible episode reached that limit; missing values are not silently scored as zero.

Model	720-second endings	Other endings	Wins / other endings
GPT-6 Astra	6	144	135/144 (93.8%)
GPT-5.6 Sol	73	77	29/77 (37.7%)
GPT-6 Sol	122	28	22/28 (78.6%)
GPT-5.6 Terra	69	81	20/81 (24.7%)
GPT-5.6 Luna	118	32	4/32 (12.5%)
Qwen 3.8 Flash	125	25	2/25 (8.0%)
GPT-6 Luna	143	7	2/7 (28.6%)
Qwen 3.8 Max 0902	149	1	1/1 (100.0%)
DeepSeek V4.1 Flash	150	0	undefined
Kimi K3	150	0	undefined
GLM 5.3 Flash	150	0	undefined

For resource comparisons, known input and output token counts are reported subtotals when usage is incomplete. Cached input and reasoning output are subsets of their corresponding totals, never extra tokens. Request-level and episode-level usage counts are distinct grains and must not be added as if they were the same missingness unit. We define a descriptive generated-token index $I_{\text{tok}} = 10^6 L / T_{\text{out}}$ levels per million reported output tokens. If output usage is missing, $10^6 L / T_{\text{out,known}}$ is an upper bound on the index using total output, provided the known denominator is positive. Its lower bound remains unknown. Request-grain 150,000-token reservations and GPT 20-million-token technical holds are not observed output, proven maxima, provider quota, charges, or invoices, and cannot supply a finite positive lower bound. With a zero known denominator the displayed index is undefined. This applies separately to request-grain and episode-grain missingness. Tokenizers, tools, and hidden reasoning compute differ, so this index is not a model-only efficiency ranking. Appendix C includes every configuration’s interaction time, token coverage and monetary estimates.

The eligible resource ledger retains 616 request-grain and 19 episode-grain unknown-use units. These separate grains are not added into a single missingness rate. Known monetary figures are estimates, not invoices; unknown-use reservations are holds rather than measured charges. Confirmed engineering interruptions remain excluded, with every raw trace and its known or unknown use preserved. Completed raw wins without a complete required evidence boundary are disclosed alongside the first independently eligible replacements. Appendix H retains the full phase chronology, selected-versus-complete ledger subtotals, and nonadditive resource accounting; Appendix D lists all 58 excluded raw attempts.

The observed median episode duration ranges from 280.6 seconds (GPT-6 Astra) to 720.0 seconds (GPT-6 Sol). This is a recorded end-to-end episode time contrast, not an isolated model-serving speed measurement. Appendix C reports episode wall time and operational levels per wall-hour separately from token coverage and puzzle outcomes. This pace combines provider response time, local computation, observation, and game actions. A complete, uniformly timed request-level throughput series is unavailable for the 1,650-case matrix, so we do not impute a cross-model decoding-TPS ranking.

8 What the challenge supply establishes

The largest within-model spread in full curricula appears in GPT-5.6 Sol: Relations has 9/15 and Objects has 0/15. This is a descriptive contrast across authored families, not a causal mechanism or an independent-skill diagnosis.

The within-model family patterns are descriptive because authored families share structures and were used during development. A model’s zero-depth case is not proof that it made no meaningful observation, and a normal timeout does not establish a specific cognitive cause. Conversely, a score cannot be rescued by assuming a faster service would have made an unobserved successful action possible. A new prospective experiment would need matched tool access, inference-compute accounting, and service controls.

The same case list controls puzzle exposure but does not equalize tools, reasoning settings, provider infrastructure, concurrency, or model snapshots. Model identifiers may be routed aliases. Earlier GPT phases used their recorded settings; added GPT-6 Sol and Luna used scoped local serving with an operating-system isolation boundary. Ordinary available accounts may vary, while exact requested/returned model identity and recorded numeric usage remain required. Separately retained recovery phases and confirmed interruptions are described in Appendix H; no model-only causal ranking or clean completion of an interrupted cohort is inferred. Replay and source hashes establish recorded consistency, not semantic or operating-system attestation.

The result supports a practical challenge-generation capability: AI-assisted construction produces reachable tasks with unresolved outcomes for the evaluated configurations. A verified failure can become a target for better exploration, longer-horizon control or corrective data, while retained partial progress helps choose the next learning stage. Which intervention improves a learner is an experimental question. The present bank is exposed and therefore cannot certify held-out transfer after training on its contents. A future experiment must freeze distinct acceptance tasks before training and measure paired gains under controlled tools, budgets and serving conditions. The pathway in Figure 9 is proposed; no trained candidate or persistent improvement is evaluated in this study.

9 Conclusion

Recursive-Play demonstrates an actionable starting point for AI-driven improvement: use AI-assisted task construction to produce executable, verifiably solvable challenges, then evaluate where current agents remain incomplete. Fifteen canonical curricula remain unfinished by all eleven tested configurations within the fixed budget, including two with no cleared level, while constructive witnesses establish that the task outcomes are reachable. The complete matrix and recorded traces make those gaps concrete rather than relying on an unsupported claim of difficulty.

The next step is to transform verified gaps into learning inputs and test whether they yield persistent gains on fresh acceptance tasks. Resettable environments, independent verifiers, corrected trajectories and no-regression evaluation provide a concrete route for that experiment. The contribution is thus a validated source of challenges and a measurable improvement pathway, with the effect of subsequent training left to that next test.

References

- ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. Technical report, ARC Prize Foundation, April 2026. URL https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf. Technical report dated April 22, 2026; accessed September 21, 2026.
- Yizhe Chi, Wenyi Li, Deyao Hong, Xiaoqi Wang, Mingju Gao, Kaisen Yang, Bingxiang He, Youjie Zheng, Calvin Xiao, and Qinhuai Na. AI4AI-Bench: Benchmarking LLM agents in algorithmic design for recursive self-improvement. *arXiv preprint arXiv:2608.20318*, 2026. doi: 10.48550/arXiv.2608.20318. URL <https://arxiv.org/abs/2608.20318v1>. Version 1, August 20, 2026.
- Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A learning environment for text-based games. In Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant, editors, *Computer Games: CGW 2018, Revised Selected Papers*, volume 1017 of *Communications in Computer and Information Science*, pages 41–75. Springer, 2019. doi: 10.1007/978-3-030-24337-1_3. URL https://link.springer.com/chapter/10.1007/978-3-030-24337-1_3.
- Michael Dennis, Natasha Jaques, Eugene Vinitzky, Alexandre Bayen, Stuart J. Russell, Andrew Critch, and Sergey Levine. Emergent complexity and zero-shot transfer via unsupervised environment design. In *Advances in Neural*

- Information Processing Systems*, volume 33, 2020. URL <https://papers.nips.cc/paper/2020/hash/985e9a46e10005356bbaf194249f6856-Abstract.html>.
- Yi Duan, Ying Liu, Zirui Tang, Haodong Chen, Jun Zhou, Yumou Liu, Bangrui Xu, Yukai Wu, Sidi Chen, Yuhan Zhou, Haoyu Wang, Xiaoyou Yu, Shaokun Han, Xuzhou Zhu, Le Zhou, Bolin Lu, Wei Zhou, Jiachen Liu, Nuozhou Fang, Jiaxin Tian, Ruoyu Chen, Yuxuan Li, Kai Zuo, Kaiyan Zhang, Qianyu Yang, Zijie Wang, Jiantao Qiu, Conghui He, Guoliang Li, Bowen Zhou, Zhiyuan Liu, Zhoufutu Wen, Jihua Kang, Xuanhe Zhou, and Fan Wu. The last AI built by humans: Toward genuine recursive self-improvement. *arXiv preprint arXiv:2609.11873*, 2026. doi: 10.48550/arXiv.2609.11873. URL <https://arxiv.org/abs/2609.11873v2>. Version 2, September 15, 2026.
- Alex Jacob, Andrej Jovanović, William F. Shen, Daniel Burkhardt, Meghdad Kurmanji, Nurbek Tastan, Lorenzo Sani, Niccolò Alberto Elia Venanzi, Ambroise Odonnat, Zeyu Cao, Bill Marino, Xinchu Qiu, and Nicholas D. Lane. The red queen Gödel machine: Co-evolving agents and their evaluators. *arXiv preprint arXiv:2606.26294*, 2026. doi: 10.48550/arXiv.2606.26294. URL <https://arxiv.org/abs/2606.26294v2>. Version 2, June 29, 2026.
- Anas Mahmoud, Mohammad Hossein Rezaei, Mingchen Zhuge, Chenguang Wang, Yuandong Tian, and Yunzhong He. Announcing RSI bench. Scale AI, official benchmark announcement, August 2026. URL <https://labs.scale.com/rsi-benchmark/blog/announcing-rsi-bench>. Published August 7, 2026; accessed September 21, 2026.
- Fanqing Meng, Lingxiao Du, Qiguang Chen, Ziqi Zhao, Haocheng Lu, Mengkang Hu, and Michael Qizhe Shieh. RSIBench-Data: Benchmarking data-centric research for recursive self-improvement. *arXiv preprint arXiv:2607.25886*, 2026. doi: 10.48550/arXiv.2607.25886. URL <https://arxiv.org/abs/2607.25886v1>. Version 1, July 28, 2026.
- Cheng Qian, Peixuan Han, Qinyu Luo, Bingxiang He, Xiuxi Chen, Yuji Zhang, Hongyi Du, Jiarui Yao, Xiaocheng Yang, Denghui Zhang, Yunzhu Li, and Heng Ji. EscapeBench: Towards advancing creative intelligence of language model agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 798–820. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-long.39. URL <https://aclanthology.org/2025.acl-long.39/>.
- Rui Wang, Joel Lehman, Jeff Clune, and Kenneth O. Stanley. POET: Open-ended coevolution of environments and their optimized solutions. In *Proceedings of the Genetic and Evolutionary Computation Conference*. Association for Computing Machinery, 2019. doi: 10.1145/3321707.3321799. URL <https://doi.org/10.1145/3321707.3321799>.
- Shuhan Xue, Zixin Ding, Yichen Shen, Yinjie Wang, Zhenfei Yin, Yingcheng Wu, Yuxin Chen, Mengdi Wang, and Ling Yang. PAST-Bench: Benchmarking the foundations of recursive self-improvement in personal agents. *arXiv preprint arXiv:2608.04003*, 2026. doi: 10.48550/arXiv.2608.04003. URL <https://arxiv.org/abs/2608.04003v1>. Version 1, August 4, 2026.
- Eric Zelikman, Eliana Lorch, Lester Mackey, and Adam Kalai. Self-taught optimizer (STOP): Recursively self-improving code generation. In *First Conference on Language Modeling*, 2024. URL <https://openreview.net/forum?id=46Zgqo4QIU>.
- Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin Gödel machine: Open-ended evolution of self-improving agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=pUpzQZTvGY>.

A Execution settings and measurement boundaries

Main tables identify models by name. Provider, tool, alias, reasoning, concurrency, source-hash, and isolation settings are recorded in the versioned phase manifests. Prior GPT configurations used their recorded tool, date, alias, reasoning and serving setup; ordinary available accounts may vary, without a fixed identity gate; exact requested/returned model evidence is still required; added GPT-6 Sol and Luna used a scoped local proxy forwarding to the normal local Sub2API route. Qwen Flash and GLM Flash used the original API harness, Kimi K3 and Qwen Max 0902 used Bailian, and DeepSeek V4.1 Flash used Ark. These different configurations prevent a model-only causal interpretation of score or duration gaps.

B Depth and author-family breakdown

Unless stated otherwise, model rows in the comparative tables and model columns in the complete outcome matrix follow Table 1: complete-task count first, with cleared levels ordering tied rows for display. This consistent order supports comparison across measures; it does not rank models by conditional win rate, token use, or cost.

Table 5: Primary completion-depth counts; each row sums to 150.

Model	0	1	2	3	4	5	6
GPT-6 Astra	2	2	4	3	1	3	135
GPT-5.6 Sol	50	22	16	12	12	9	29
GPT-6 Sol	32	27	27	21	12	9	22
GPT-5.6 Terra	68	18	15	11	9	9	20
GPT-5.6 Luna	83	22	21	12	4	4	4
Qwen 3.8 Flash	95	22	18	9	3	1	2
GPT-6 Luna	102	27	9	9	0	1	2
Qwen 3.8 Max 0902	100	29	10	7	3	0	1
DeepSeek V4.1 Flash	111	26	6	4	2	1	0
Kimi K3	126	13	9	1	1	0	0
GLM 5.3 Flash	128	17	4	1	0	0	0

Table 6: Full curricula by authored family; each family has 15 environments.

Model	a	b	c	d	e	f	g	h	i	j
GPT-6 Astra	13/15	13/15	12/15	15/15	15/15	15/15	14/15	13/15	13/15	12/15
GPT-5.6 Sol	3/15	4/15	0/15	1/15	9/15	7/15	0/15	5/15	0/15	0/15
GPT-6 Sol	6/15	6/15	0/15	0/15	1/15	5/15	0/15	3/15	1/15	0/15
GPT-5.6 Terra	4/15	2/15	0/15	1/15	3/15	6/15	0/15	4/15	0/15	0/15
GPT-5.6 Luna	0/15	3/15	0/15	0/15	1/15	0/15	0/15	0/15	0/15	0/15
Qwen 3.8 Flash	1/15	1/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15
GPT-6 Luna	0/15	2/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15
Qwen 3.8 Max 0902	1/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15
DeepSeek V4.1 Flash	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15
Kimi K3	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15
GLM 5.3 Flash	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15	0/15

Table 7: Completed levels by authored family; each family offers 90 levels.

Model	a	b	c	d	e	f	g	h	i	j
GPT-6 Astra	84/90	79/90	79/90	90/90	90/90	90/90	89/90	82/90	83/90	82/90
GPT-5.6 Sol	52/90	31/90	6/90	37/90	70/90	58/90	10/90	49/90	24/90	20/90
GPT-6 Sol	60/90	53/90	12/90	20/90	40/90	59/90	21/90	50/90	23/90	31/90
GPT-5.6 Terra	48/90	20/90	10/90	26/90	56/90	59/90	2/90	32/90	15/90	14/90
GPT-5.6 Luna	30/90	23/90	6/90	15/90	38/90	23/90	2/90	7/90	12/90	4/90
Qwen 3.8 Flash	18/90	13/90	5/90	11/90	15/90	16/90	4/90	16/90	9/90	7/90
GPT-6 Luna	22/90	15/90	3/90	4/90	9/90	15/90	6/90	2/90	10/90	3/90
Qwen 3.8 Max 0902	27/90	10/90	3/90	5/90	14/90	11/90	1/90	7/90	10/90	0/90
DeepSeek V4.1 Flash	24/90	7/90	1/90	1/90	10/90	10/90	0/90	2/90	7/90	1/90
Kimi K3	12/90	2/90	2/90	3/90	4/90	6/90	0/90	1/90	8/90	0/90
GLM 5.3 Flash	8/90	2/90	1/90	3/90	6/90	2/90	0/90	3/90	3/90	0/90

C Interaction, token, and operational resources

Summed episode wall time includes service latency and gameplay. It is not batch wall time or decoder throughput. Known tokens are subtotals if usage is missing. Cached input and reasoning output are included within their respective token totals. Request and native episode usage have different grains.

Table 8: Eligible interaction totals. Levels per summed wall-hour is descriptive pace.

Model	Actions	Median s	Summed min	Levels/h	Unknown
GPT-6 Astra	23,288	280.6	828.7	61.40	0 episode
GPT-5.6 Sol	50,017	716.5	1,581.0	13.55	7 episode
GPT-6 Sol	24,517	720.0	1,722.9	12.85	0 episode
GPT-5.6 Terra	51,221	710.4	1,609.9	10.51	6 episode
GPT-5.6 Luna	38,480	720.0	1,742.9	5.51	6 episode
Qwen 3.8 Flash	5,847	720.0	1,777.4	3.85	125 request
GPT-6 Luna	8,963	720.0	1,783.0	3.00	0 episode
Qwen 3.8 Max 0902	2,722	720.0	1,796.5	2.94	149 request
DeepSeek V4.1 Flash	2,927	720.0	1,800.0	2.10	49 request
Kimi K3	1,370	720.0	1,800.0	1.27	148 request
GLM 5.3 Flash	1,257	720.0	1,800.0	0.93	145 request

Table 9: Known token subtotals in millions, with explicit usage coverage.

Model	Known/total	Input M	Output M	Cache M	Reason M
GPT-6 Astra	150/150 episode	225.453	1.255	215.536	0.510
GPT-5.6 Sol	143/150 episode	319.409	2.693	309.650	1.539
GPT-6 Sol	150/150 episode	156.087	1.887	147.199	0.890
GPT-5.6 Terra	144/150 episode	491.846	4.102	476.310	2.522
GPT-5.6 Luna	144/150 episode	506.390	4.545	485.935	2.843
Qwen 3.8 Flash	3309/3434 request	392.122	5.717	353.362	4.788
GPT-6 Luna	150/150 episode	172.207	2.784	161.486	1.979
Qwen 3.8 Max 0902	1530/1679 request	87.427	3.739	70.331	3.491
DeepSeek V4.1 Flash	2034/2083 request	133.566	8.716	112.817	8.178
Kimi K3	1014/1162 request	35.802	3.245	26.923	2.591
GLM 5.3 Flash	659/804 request	27.730	3.606	19.274	3.442

Table 10: Selected-case estimates and non-primary levels per million reported output tokens. Unpriced routes are shown as dashes, not zero.

Model	Selected CNY	Unknown hold CNY	Token index
GPT-6 Astra	—	—	675.70
GPT-5.6 Sol	—	—	≤ 132.55; lower unknown
GPT-6 Sol	—	—	195.54
GPT-5.6 Terra	—	—	≤ 68.75; lower unknown
GPT-5.6 Luna	—	—	≤ 35.20; lower unknown
Qwen 3.8 Flash	81.7797	86.3213	≤ 19.94; lower unknown
GPT-6 Luna	—	—	31.97
Qwen 3.8 Max 0902	445.2400	1,137.2558	≤ 23.54; lower unknown
DeepSeek V4.1 Flash	336.8595	85.7758	≤ 7.23; lower unknown
Kimi K3	555.9738	2,591.6665	≤ 11.71; lower unknown
GLM 5.3 Flash	21.2955	75.5633	≤ 7.76; lower unknown

D Independently excluded engineering attempts

All 58 independently reviewed raw engineering attempts remain outside primary scores and inside the operational evidence. An excluded GPT-6 Sol b002 game mechanically reached 6/6, yet predeadline response-stream identity and usage faults made it ineligible; its first valid replacement is selected even if that score is lower. This example guards against best-of-two reporting. The nine cluster-exit partial amounts are subsets of their API parent checkpoints. For the six-admission GPT partial cohort, the checkpoint known subtotal is selected-only: 3,450,954 technical tokens. The complete provider ledger is 6,343,672, including 2,892,718 from excluded attempts. The later twelve-admission checkpoint contains 10 selected cases and 10,488,646 known technical tokens; its complete provider ledger contains 11,666,058, including 1,177,412 from the two TLS-interrupted exclusions. The complete operational phase snapshot is counted once; its contained subsets are not additional charges. Unknown holds are not provider invoices. A distinct Sol b013 attempt (original55, attempt2) reached all six levels with84 actions in648443ms before its later request/native-integrity fault; that post-terminal raw win remains excluded with native aggregate usage UNKNOWN. This differs from the historical Sol b002 predeadline example above. The selected GPT provenance combines94 older cases,179 cases from the completed cohort, three selected partial6 cases, ten selected partial12 cases, and14 distinct recovery cases; ordinals2/3 are preserved.

Table 11: All excluded raw attempts by independently reviewed class, ordered by raw-attempt count (descending).

Exclusion kind	Raw attempts
excluded-scored-original-attempt	31
excluded-attempt	22
partial6-proxy-native-integrity-exclusion	3
partial12-upstream-tls-integrity-exclusion	2

Table 12: Cumulative API phase checkpoint estimates, including excluded raw attempts once. Probe exposure is separately itemized.

Model	Known CNY	Unknown hold CNY	Pending hold CNY	Checkpoints
Qwen 3.8 Max 0902	–	–	–	4
DeepSeek V4.1 Flash	–	–	–	6
Kimi K3	–	–	–	11

Table 13: Non-scored probes and any excluded exposure kept outside phase checkpoints. These are added only once in the separate operational ledger.

Model	Probe known CNY	Probe unknown hold CNY	Separate excluded CNY
Qwen 3.8 Max 0902	–	–	–
DeepSeek V4.1 Flash	–	–	–
Kimi K3	–	–	–

Table 14: All 58 excluded raw attempts, grouped by the model order in Table 1, then by environment ID and full exclusion class (stable ties). Each exact raw row remains traceable in the separately hashed operational CSV; ordering does not select by outcome or score.

Model	Case	Class	Raw result	Unknown responses
GPT-6 Sol	b002	excluded-scored-original-attempt	6/6 won	1
GPT-6 Sol	b006	excluded-scored-original-attempt	0/6 time_budget	2
GPT-6 Sol	b013	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Sol	b013	partial6-proxy-native-integrity-exclusion	6/6 won	1
GPT-6 Sol	b014	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Sol	b015	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Sol	c001	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Sol	c001	partial12-upstream-tls-integrity-exclusion	0/6 time_budget	1
GPT-6 Sol	c002	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Sol	c003	excluded-scored-original-attempt	0/6 playing	0
GPT-6 Sol	c004	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Sol	c005	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Sol	c006	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Sol	c007	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Sol	d015	excluded-scored-original-attempt	0/6 time_budget	1
GPT-6 Luna	a011	excluded-scored-original-attempt	0/6 time_budget	1
GPT-6 Luna	b003	excluded-scored-original-attempt	0/6 time_budget	2
GPT-6 Luna	b003	partial6-proxy-native-integrity-exclusion	0/6 time_budget	0
GPT-6 Luna	b004	excluded-scored-original-attempt	0/6 time_budget	UNKNOWN
GPT-6 Luna	b007	excluded-scored-original-attempt	0/6 playing	0
GPT-6 Luna	b008	excluded-scored-original-attempt	0/6 playing	0
GPT-6 Luna	b010	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Luna	b011	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Luna	b011	partial6-proxy-native-integrity-exclusion	0/6 time_budget	1
GPT-6 Luna	b012	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Luna	b013	excluded-scored-original-attempt	0/6 playing	1
GPT-6 Luna	b014	excluded-scored-original-attempt	0/6 time_budget	1
GPT-6 Luna	b015	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Luna	b015	partial12-upstream-tls-integrity-exclusion	0/6 time_budget	1
GPT-6 Luna	c001	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Luna	c002	excluded-scored-original-attempt	0/6 time_budget	0
GPT-6 Luna	c003	excluded-scored-original-attempt	0/6 time_budget	1
GPT-6 Luna	c004	excluded-scored-original-attempt	0/6 time_budget	1
GPT-6 Luna	c005	excluded-scored-original-attempt	0/6 playing	UNKNOWN
GPT-6 Luna	c006	excluded-scored-original-attempt	0/6 playing	0
GPT-6 Luna	d013	excluded-scored-original-attempt	1/6 time_budget	1
Qwen 3.8 Max 0902	b003	excluded-attempt	0/6 abandoned	UNKNOWN
Qwen 3.8 Max 0902	b004	excluded-attempt	0/6 abandoned	UNKNOWN
Qwen 3.8 Max 0902	b005	excluded-attempt	3/6 abandoned	UNKNOWN
Qwen 3.8 Max 0902	d013	excluded-attempt	—	UNKNOWN
Qwen 3.8 Max 0902	d015	excluded-attempt	—	UNKNOWN
Qwen 3.8 Max 0902	e001	excluded-attempt	—	UNKNOWN
DeepSeek V4.1 Flash	b003	excluded-attempt	0/6 abandoned	UNKNOWN
DeepSeek V4.1 Flash	b005	excluded-attempt	1/6 abandoned	UNKNOWN
DeepSeek V4.1 Flash	d013	excluded-attempt	—	UNKNOWN
DeepSeek V4.1 Flash	d015	excluded-attempt	—	UNKNOWN
DeepSeek V4.1 Flash	e001	excluded-attempt	—	UNKNOWN
DeepSeek V4.1 Flash	h015	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	b003	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	b004	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	b005	excluded-attempt	1/6 abandoned	UNKNOWN
Kimi K3	d013	excluded-attempt	—	UNKNOWN
Kimi K3	d015	excluded-attempt	—	UNKNOWN
Kimi K3	e001	excluded-attempt	—	UNKNOWN
Kimi K3	e005	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	j005	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	j006	excluded-attempt	0/6 abandoned	UNKNOWN
Kimi K3	j011	excluded-attempt	0/6 abandoned	UNKNOWN

E Every model–environment outcome

Each cell is completed depth 0–6 followed by W (all levels won), T (ordinary 720-second ending), A (512-action ending), or F (abandoned). No cell is inferred or untested.

Table 15: All 1,650 independently eligible outcomes on 150 matched environments. Rows follow environment ID; model columns follow the primary complete-task order in Table 1, including the displayed ordering of tied models.

Case	Seed	Astra	Sol5	Sol6	Terra5	Luna5	QFlash	Luna6	QMax	DeepSeek	Kimi	GLM
a001	503117	6W	6W	6W	6W	3T	1T	3T	6W	5T	1T	3T
a002	503117	6W	5T	6W	6W	5T	0T	2T	1T	2T	0T	0T
a003	511036	6W	4T	6W	2T	3T	3T	3T	3T	2T	1T	1T
a004	518955	5T	4T	4T	4T	2T	3T	0A	2T	3T	2T	0T
a005	526874	6W	1A	5T	6W	5T	6W	3T	0T	4T	2T	0T
a006	534793	6W	6W	6W	4T	0T	0T	1T	3T	0T	0T	0T
a007	542712	6W	5T	5T	1T	1T	1T	1T	4T	1T	2T	1T
a008	550631	6W	6W	6W	6W	5T	2T	5T	2T	3T	2T	1T
a009	558550	1A	0A	1A	0A	0A	0T	0A	0T	1T	0T	0T
a010	566469	6W	2T	6W	2T	0T	0F	0T	0T	0T	0T	0T
a011	574388	6W	5T	0T	0A	0T	0T	0T	0T	0T	0T	0T
a012	582307	6W	1A	3T	3T	1T	0T	1T	3T	1T	1T	0T
a013	590226	6W	1T	2T	3T	1T	0T	1A	1T	0T	1T	1T
a014	598145	6W	2T	2T	2T	2T	2T	2T	2T	2T	0T	1T
a015	606064	6W	4T	2T	3A	2T	0F	0T	0T	0T	0T	0T
b001	511036	6W	6W	6W	6W	6W	2T	6W	0T	1T	0T	0T
b002	613983	6W	0A	0T	0A	0T	0T	1T	1T	0T	0T	0T
b003	621902	6W	0T	4T	0A	0T	3T	0T	0T	0T	0T	0T
b004	629821	6W	6W	6W	0A	0T	0T	0A	0T	1T	0T	0T
b005	637740	6W	6W	6W	6W	6W	6W	6W	3T	3T	2T	1T
b006	645659	6W	3T	2T	0A	0T	0T	0T	0T	0T	0T	0T
b007	653578	6W	0T	6W	0T	0A	0T	0T	0T	0T	0T	0T
b008	661497	6W	1A	6W	1A	1T	0T	0T	1T	1T	0T	1T
b009	669416	6W	1T	6W	1A	1A	0T	1T	2T	0T	0T	0T
b010	677335	0A	0T	0A	0F	0T	0T	0T	0T	0T	0T	0T
b011	685254	6W	0A	3T	1T	0T	0T	0T	0T	0T	0T	0T
b012	693173	6W	0T	0A	0A	0T	0T	0T	0T	0T	0T	0T
b013	701092	6W	6W	4T	4T	1T	0T	0T	2T	0T	0T	0T
b014	709011	6W	1A	2T	0A	6W	0T	1T	0T	0T	0T	0T
b015	716930	1A	1A	2T	1T	2T	2T	0T	1T	1T	0T	0T
c001	724849	4A	0A	0T	0T	0T	0T	0T	0T	0T	0T	0T
c002	732768	6W	0A	0T	0A	0T	0T	0T	0T	0T	0T	0T
c003	740687	6W	1T	3T	0T	3T	3T	2T	3T	1T	2T	1T
c004	518955	6W	0A	0T	0A	1T	0F	0T	0T	0T	0T	0T
c005	748606	6W	0A	1T	4T	0A	0T	0T	0T	0T	0T	0T
c006	756525	6W	3T	3T	3T	0A	0F	0T	0T	0T	0T	0T
c007	764444	6W	0A	1T	0A	0A	1T	1T	0T	0T	0T	0T
c008	772363	6W	0A	0T	0A	0A	0T	0T	0T	0T	0T	0T
c009	780282	6W	1A	0T	0T	1T	0T	0T	0T	0T	0T	0T
c010	788201	3T	0T	0T	0A	0T	0T	0T	0T	0T	0T	0T
c011	796120	6W	0T	0T	0A	0A	0F	0T	0T	0T	0T	0T
c012	804039	0T	0A	0T	0T	0T	0T	0T	0T	0T	0T	0T
c013	811958	6W	0A	0T	3T	1T	1T	0T	0T	0T	0T	0T
c014	819877	6W	1T	3T	0A	0T	0T	0T	0T	0T	0T	0T
c015	827796	6W	0A	1T	0A	0A	0T	0T	0T	0T	0T	0T
d001	835715	6W	5T	1T	0T	0T	1T	0T	1T	0T	0T	0T
d002	843634	6W	0T	2T	0A	0T	0T	0T	0T	0T	0T	0T
d003	851553	6W	4T	0T	0A	0T	0T	0T	0T	0T	0T	0T
d004	859472	6W	4T	2T	4T	4T	2T	1T	0T	0T	0T	0T
d005	867391	6W	0A	5T	5T	3T	2T	0T	1T	0T	0T	0T
d006	875310	6W	0T	1T	0A	0T	0T	0T	0T	0T	0T	0T
d007	883229	6W	2T	2T	2T	2A	2T	1T	0T	0T	1T	1T
d008	891148	6W	0T	0T	0A	0T	0T	0T	0T	0T	0T	0T
d009	899067	6W	6W	0T	0A	0A	0T	0T	0T	0T	0T	0T
d010	906986	6W	5T	2T	5T	0F	0T	0T	1T	0T	0T	0T
d011	914905	6W	3T	1T	1T	3T	2T	1T	1T	0T	0T	1T
d012	922824	6W	0A	0T	0A	0A	0T	0T	0T	0T	0T	0T
d013	930743	6W	2T	2T	2T	2T	1F	1T	1T	1T	1T	1T
d014	526874	6W	1T	1T	1A	1T	1F	0T	0T	0T	1T	0T
d015	938662	6W	5T	1T	6W	0T	0F	0T	0T	0T	0T	0T
e001	946581	6W	3A	3T	5T	2T	1T	1T	1T	1T	0T	0T

Continued on next page

Case	Seed	Astra	Sol5	Sol6	Terra5	Luna5	QFlash	Luna6	QMax	DeepSeek	Kimi	GLM
e002	954500	6W	6W	6W	6W	6W	0T	0T	0T	0T	0T	0T
e003	962419	6W	6W	3T	2A	3T	3T	1T	2T	2T	0T	2T
e004	970338	6W	6W	4T	6W	2T	4T	1T	4T	2T	2T	1T
e005	978257	6W	6W	2T	1A	2T	1T	0T	1T	1T	0T	0T
e006	986176	6W	6W	2T	5T	2T	0F	1T	0T	1T	0T	0T
e007	534793	6W	4T	1T	0A	2T	1T	0T	0T	0T	0T	0T
e008	994095	6W	5T	4T	5T	2T	2F	0T	0T	1T	0T	0T
e009	1002014	6W	4T	1T	4T	0T	1T	0T	1T	0T	0T	0T
e010	1009933	6W	6W	3T	5T	4T	0T	2T	3T	0T	0T	2T
e011	1017852	6W	6W	5T	5T	4T	1T	2T	1T	1T	1T	0T
e012	1025771	6W	0A	1T	2T	4T	0T	0T	0T	0T	0T	0T
e013	1033690	6W	6W	1T	4T	3A	1T	0F	1T	1T	1T	1T
e014	1041609	6W	6W	4T	6W	0T	0T	1T	0T	0T	0T	0T
e015	1049528	6W	0A	0T	0A	2T	0T	0T	0T	0T	0T	0T
f001	1057447	6W	6W	5T	6W	1F	2T	2T	1T	1T	0T	0T
f002	542712	6W	1T	3T	5T	3T	0T	0T	0T	0T	0T	0T
f003	1065366	6W	6W	6W	6W	2T	1T	3T	1T	1T	0T	1T
f004	1073285	6W	6W	3T	1T	0T	0T	0T	0T	0T	0T	0T
f005	1081204	6W	6W	6W	5T	3T	1T	3T	3T	1T	0T	0T
f006	1089123	6W	3A	3T	3T	3T	2F	1T	2T	1T	0T	0T
f007	1097042	6W	2T	2T	2T	1T	0T	0T	0T	0T	0T	0T
f008	1104961	6W	0T	3T	4T	0T	0T	0T	0T	0T	0T	0T
f009	1112880	6W	5T	4T	3T	0T	0T	0T	1T	0T	0T	0T
f010	1120799	6W	6W	6W	6W	5T	5T	3T	1T	4T	4T	1T
f011	1128718	6W	6W	6W	6W	2T	1F	0T	1T	0T	1T	0T
f012	1136637	6W	0A	2T	0A	1T	0T	0T	0T	0T	0T	0T
f013	1144556	6W	6W	6W	6W	1T	0T	2T	0T	1T	0T	0T
f014	1152475	6W	4T	2T	0T	0T	0T	0T	0T	0T	0T	0T
f015	1160394	6W	1A	2T	6W	1T	4T	1T	1T	1T	1T	0T
g001	1168313	6W	3T	3T	0A	0T	0T	3T	0T	0T	0T	0T
g002	1176232	5A	0T	0T	1T	0T	0T	0T	0T	0T	0T	0T
g003	550631	6W	2A	0T	0A	0T	0T	0T	0T	0T	0T	0T
g004	1184151	6W	3T	3T	0T	0A	0T	0T	0T	0T	0T	0T
g005	1192070	6W	0T	1T	0T	2T	0T	0T	0T	0T	0T	0T
g006	1199989	6W	1T	4T	0A	0A	3T	1T	0T	0T	0T	0T
g007	1207908	6W	0A	5T	0A	0T	0F	1T	1T	0T	0T	0T
g008	1215827	6W	0T	1T	0T	0T	1F	0T	0T	0T	0T	0T
g009	1223746	6W	0A	0T	0T	0A	0T	0T	0T	0T	0T	0T
g010	1231665	6W	0A	0T	0A	0T	0T	0T	0T	0T	0T	0T
g011	1239584	6W	0A	1T	0A	0A	0T	1T	0T	0T	0T	0T
g012	1247503	6W	1T	0T	0T	0T	0F	0T	0T	0T	0T	0T
g013	1255422	6W	0T	1T	0T	0T	0F	0T	0T	0T	0T	0T
g014	1263341	6W	0A	2A	1A	0T	0T	0T	0T	0T	0T	0T
g015	1271260	6W	0T	0A	0A	0A	0F	0T	0T	0T	0T	0T
h001	1279179	6W	5T	4T	0F	1T	1T	0T	0T	0T	0T	0T
h002	1287098	6W	4T	6W	6W	0T	2T	1T	1T	1T	0T	0T
h003	1295017	6W	2F	3T	2T	2T	2T	1T	1T	0T	1T	1T
h004	1302936	6W	2T	2T	0A	1T	2T	0T	0T	0T	0T	0T
h005	1310855	6W	6W	6W	6W	0T	0T	0T	0T	0T	0T	0T
h006	1318774	6W	6W	4T	0T	0T	0T	0T	1T	0T	0T	0T
h007	1326693	6W	0T	0T	0T	0A	0T	0T	0T	0T	0T	0T
h008	1334612	2A	2T	1T	0T	0F	0T	0T	0T	0T	0T	0T
h009	1342531	6W	6W	5T	6W	0T	1T	0T	2T	0T	0T	0T
h010	1350450	6W	6W	5T	6W	1T	4T	0T	1T	0T	0T	0T
h011	558550	6W	6W	4T	2T	2T	0T	0T	0T	0T	0T	2T
h012	1358369	6W	2T	2T	2T	0T	0T	0T	0T	0T	0T	0T
h013	1366288	6W	1A	6W	1A	0T	3T	0T	0T	0T	0T	0T
h014	1374207	6W	0A	0T	0A	0T	0T	0T	0T	0T	0T	0T
h015	1382126	2A	1A	2T	1A	0T	1T	0T	1T	1T	0T	0T
i001	566469	6W	0A	0A	0T	0T	0T	0T	0T	0T	0T	0T
i002	1390045	6W	1T	1T	1A	1T	0F	1T	1T	1T	0T	0T
i003	1397964	6W	2A	2T	2A	2T	2T	2T	2T	0T	2T	2T
i004	1405883	6W	4T	1T	3A	0T	0T	0T	0T	0T	0T	0T
i005	1413802	6W	0A	1T	1T	0T	0T	0T	0T	0T	0T	0T
i006	1421721	2A	2A	2T	2A	2F	2T	2T	2T	2T	2T	0T
i007	1429640	6W	0A	1T	0A	1T	0T	1T	0T	0T	0T	0T
i008	1437559	6W	2T	0T	0A	0A	0T	0T	0T	0T	0T	0T
i009	1445478	6W	1T	1T	0T	0T	0F	0T	0T	0T	0T	0T
i010	1453397	6W	0A	0T	0T	0T	0T	0T	0T	0T	0T	0T
i011	1461316	3T	3A	1T	2T	0T	0F	0T	0T	0T	0T	0T
i012	1469235	6W	4T	0T	0T	0T	1T	0T	0T	0T	0T	0T
i013	1477154	6W	3T	4T	3T	3T	3T	3T	4T	3T	3T	0T

Continued on next page

Case	Seed	Astra	Sol5	Sol6	Terra5	Luna5	QFlash	Luna6	QMax	DeepSeek	Kimi	GLM
i014	1485073	6W	0T	6W	0A	2T	0T	0T	0T	0T	0T	0T
i015	1492992	6W	2T	3T	1A	1T	1T	1T	1T	1T	1T	1T
j001	1500911	3A	0T	2T	0A	0T	0T	0T	0T	0T	0T	0T
j002	1508830	6W	3T	3T	1T	1T	0T	0T	0T	0T	0T	0T
j003	1516749	6W	3T	2T	0A	0T	0T	0T	0T	0T	0T	0T
j004	1524668	6W	3T	3T	0A	0T	3T	0T	0T	0T	0T	0T
j005	1532587	6W	1A	1T	1T	0T	0T	0T	0T	0T	0T	0T
j006	1540506	6W	4T	3T	4A	3T	0T	3T	0T	1T	0T	0T
j007	1548425	2T	2T	2T	0A	0T	0F	0T	0T	0T	0T	0T
j008	1556344	6W	0A	3T	3T	0T	0F	0T	0T	0T	0T	0T
j009	1564263	6W	0A	2T	0A	0F	0T	0T	0T	0T	0T	0T
j010	1572182	6W	0T	0T	3T	0T	0T	0T	0T	0T	0T	0T
j011	1580101	6W	0A	3T	0T	0T	2T	0T	0T	0T	0T	0T
j012	574388	5T	0A	1T	0A	0A	0T	0T	0T	0T	0T	0T
j013	1588020	6W	1T	1T	0T	0A	0T	0T	0T	0T	0T	0T
j014	1595939	6W	1T	0T	0A	0A	0F	0T	0T	0T	0T	0T
j015	1603858	6W	2T	5T	2T	0T	2T	0T	0T	0T	0T	0T

F Provenance and reproduction

The separate final 720-case manifest, independent final review, frozen 930-case source, case/resource CSV files, and operational raw-attempt ledger bind these tables to the evidence package. The TeX source plus generated staging files reproduce the paper without model or game calls. No training or recursive self-improvement outcome is claimed.

G Final GPT operational coverage and traceability

The final GPT operational ledger records 346,120,731 known technical tokens and 320,000,000 bounded unknown technical holds, plus 5 V6-only legacy individual unknown provider responses with unquantified token use; the separate original22 response, canary native episode and earlier unpinned quantities remain separately unknown. These are operational evidence units, not priced invoices or model quota. The final eleven-models-gpt-operational.json records native phase usage and proxy coverage at their respective grains. The case/resource CSV, excluded-raw CSV, operational-components JSON and unscored-probes JSON retain every included or excluded attempt identifier and probe. Historical holds are not new charges; cache and reasoning subsets are not added twice.

H Operational chronology and resource accounting

The following paragraphs are retained verbatim from the preceding illustrated edition. Their detailed phase accounting and recovery chronology supplement the evidence boundaries summarized in the main text.

H.1 Usage coverage and accounting

The eligible resource ledger retains 616 request-grain and 19 episode-grain unknown-use units. These separate grains are not added into a single missingness rate. Known monetary figures are estimates, not invoices. Unknown-use reservations are capacity holds, not measured charges; an unpriced route is never assigned zero cost. Independently confirmed in-game engineering interruptions, including local process exits and proxy faults, were excluded from primary scores. The frozen GPT admission policy also requires a complete native and proxy-integrity boundary: one recovery attempt reached a trusted six-level win before a later local boundary fault, but lacked a completed native turn and complete numeric proxy evidence. That raw win and its post-terminal chronology remain explicitly visible among excluded attempts; it is not labeled as a pre-terminal loss. The first independently eligible replacement is selected regardless of score. Attempts consumed before a failure remain distinct from never-admitted positions, and replacement ordinals are retained without renumbering. All excluded raw attempt identifiers, cumulative phase resources, and non-scored probes appear in Appendix D and the separate operational JSON and CSV ledgers. These include the historical GPT bounded unknown technical holds, five V6-only legacy individual unknown responses with no quantified token total, alongside the separately retained original-22 response, canary native episode, and earlier unpinned quantities, and the actual final phase ledger; earlier intermediate known totals are not substituted for final use. API exclusion subtotals are recorded inside their corresponding parent checkpoint amounts. The six-admission GPT partial checkpoint records only the three selected outcomes (3,450,954 known technical tokens); its complete provider ledger records 6,343,672, including 2,892,718 from excluded attempts. The later twelve-admission GPT checkpoint likewise contains only the ten selected outcomes (10,488,646 known technical tokens); its complete provider ledger records 11,666,058, including 1,177,412 from the two excluded attempts. Each complete phase snapshot is included once in cumulative operational use, and these contained subsets are not added again.

H.2 Execution and recovery chronology

The same case list controls exposure but does not equalize solver tools, reasoning settings, provider infrastructure, concurrency, or model snapshots. Model identifiers may be routed aliases. Client-observed request throughput includes network and first-token delay; interrupted requests are selected out of complete-request throughput summaries. The prior source-access boundary was instructional with log review, while the added GPT-6 Sol and Luna phases used a separately documented operating-system isolation and a scoped local proxy forwarding to the normal local Sub2API route. Ordinary available accounts may vary; no fixed account identity or complete per-account quota condition is imposed. Every selected GPT response still requires the exact requested and returned model identity and reported numeric usage at its recorded evidence boundary. A later recovery used birth-aware retirement checks after a historical process identifier was observed on an unrelated system process. This observation proves a contemporaneous retirement-check false positive, not the cause of an earlier boundary exception whose detailed reason was not durably recorded. The final provenance combines the completed 179-case cohort, three independently eligible cases from a six-admission partial cohort, ten independently eligible cases from a twelve-admission partial cohort, and a distinct 14-case recovery. The two excluded attempts in the twelve-admission cohort have explicit upstream stream errors and contemporaneous TLS read-error logs; their partial usage and unknown quantities remain visible. The final recovery uses a separate local HTTP/1 transport profile validated by exact-model requests; the earlier TLS error does not establish HTTP/2 as its cause. No clean 27- or 24-case completion record is fabricated. These differences are disclosed in the methods appendix, and no model-only causal ranking follows from the table. Replay and source hashes check recorded consistency but are not semantic or operating-system attestation.