NIMI
NATURE-INSPIRED
MACHINE INTELLIGENCE

TIB
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

L3S Forschungszentrum Research Center

Leibniz
Universität
Hannover

What ARC-AGI Reveals About AI

From High Scores to Efficient Learning

Prof. Dr. Sahar Vahdati
TIB Technology and Science Center & Leibniz University of Hannover, Germany

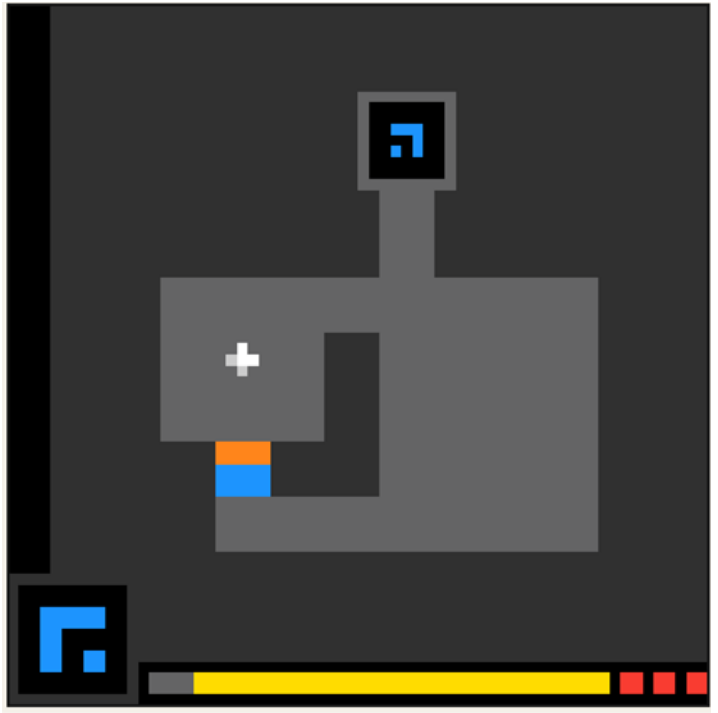
Trilateral AI Symposium, Tokyo,
28 October 2026



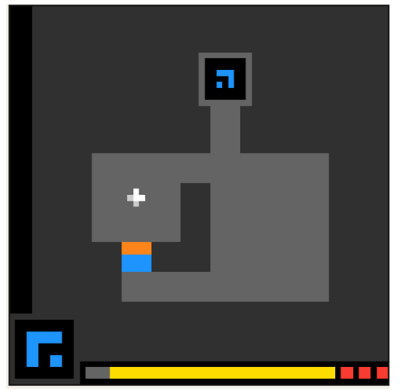
“The intelligence of a system is a measure of its skill-acquisition efficiency over a scope of tasks...”

François Chollet, On the Measure of Intelligence (2019), p. 27

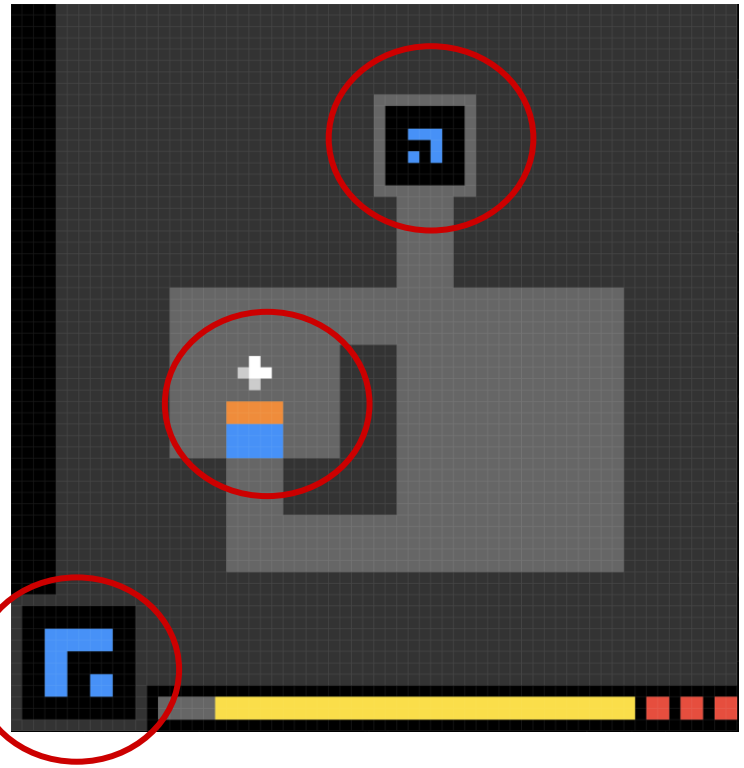
How much reliable new competence do our observations, computations and tests buy?



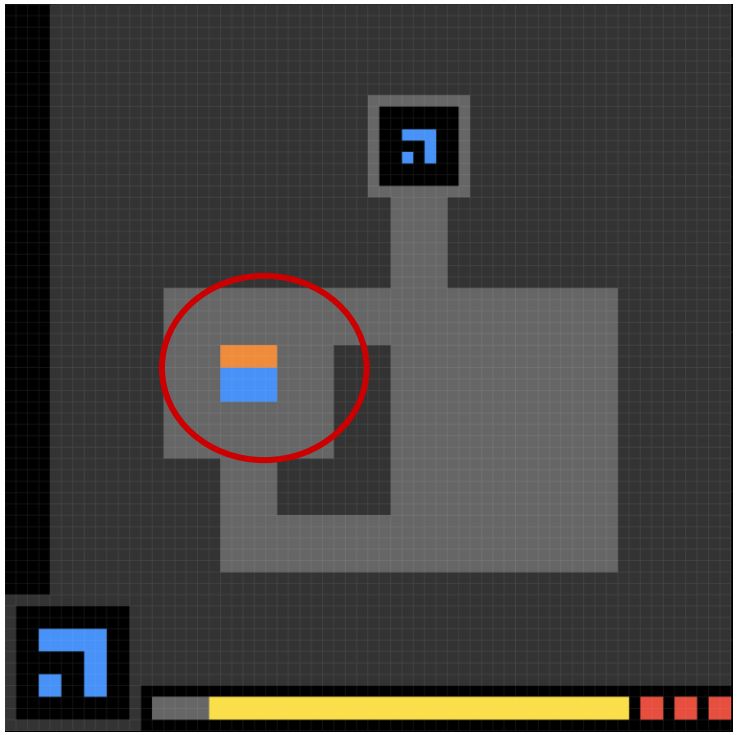
One world: LS20



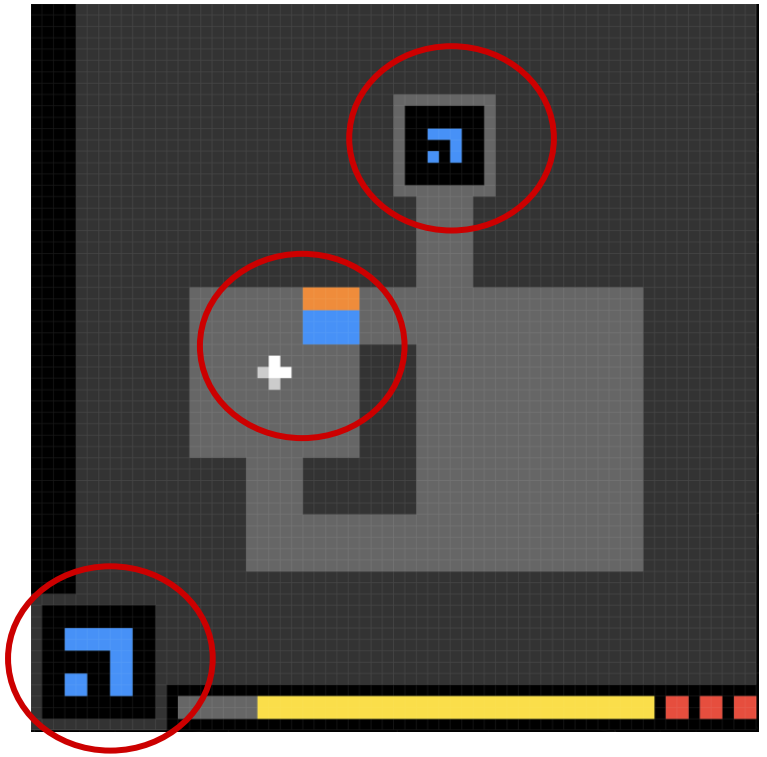
Before contact



On the white marked tile



After contact



Contact rotates the lower-left reference pattern.
The learner must discover what that change makes possible.

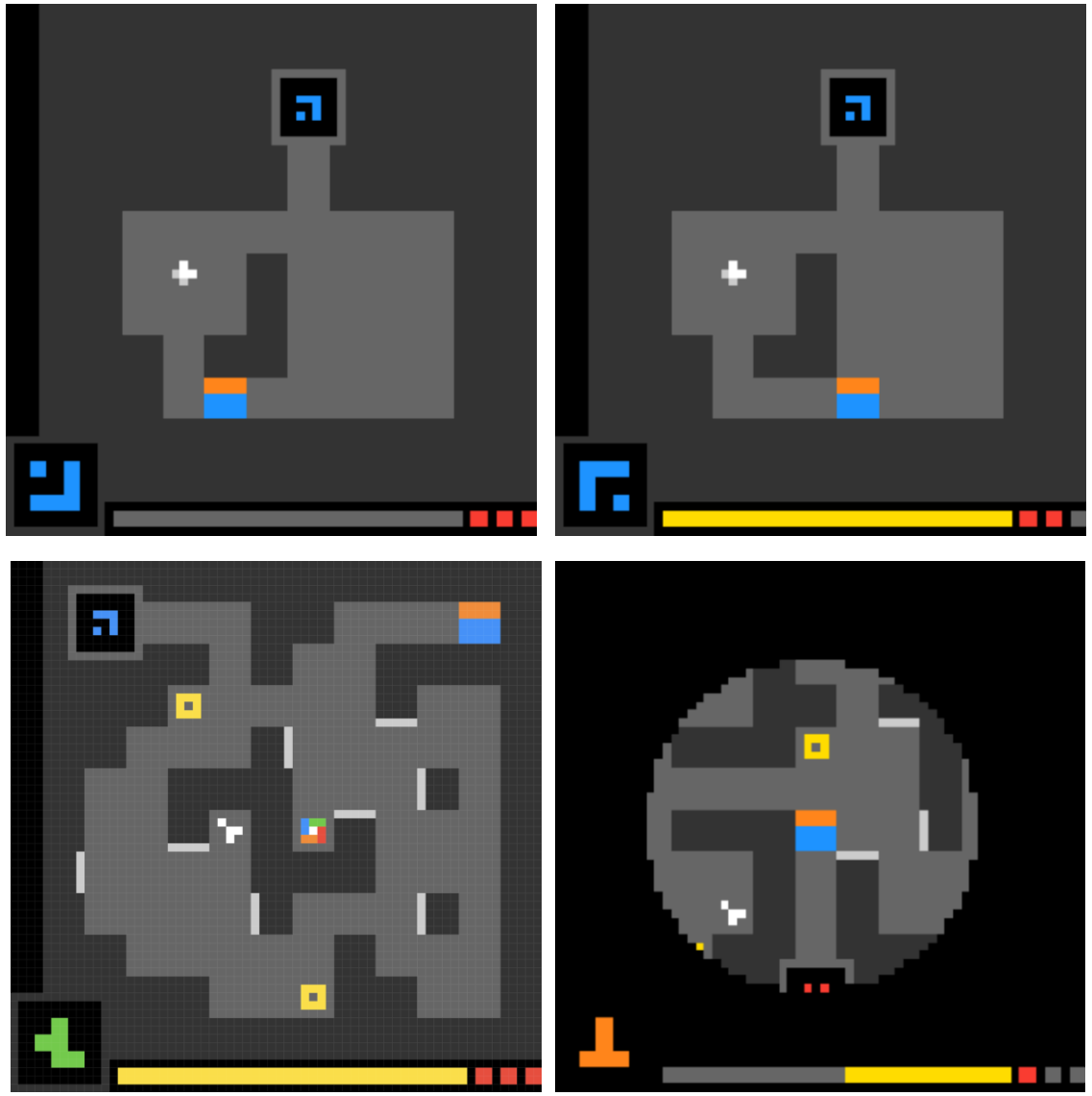
ARC-AGI-3 supplies frames and actions, not the rules

GIVEN

- 64×64 frames in 16 colors
- Available actions
- frame sequence

WITHHELD

- instructions & demonstrations
- an interpreted object vocabulary
- the mechanics
- the explicit goal
- a finite action budget



ARC Prize Foundation (2026) · public demonstration game

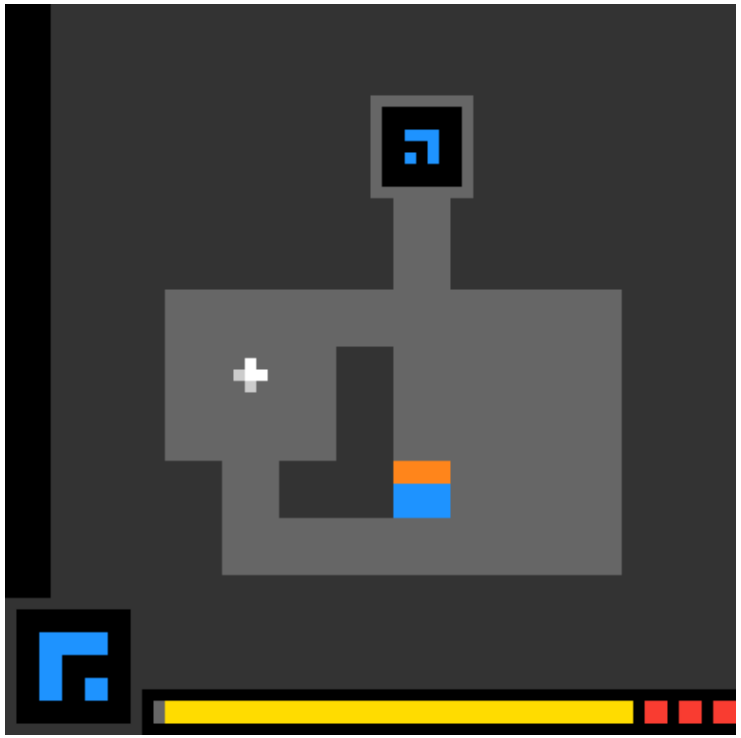
Why ARC-AGI-3?

Benchmark feature	Why it matters for efficiency
Unfamiliar rules and unstated goal	Competence must develop during interaction
Actions reveal new evidence	Learning consumes the same resource as acting
Replayable mechanics and traces	We can inspect what the learner could have tested

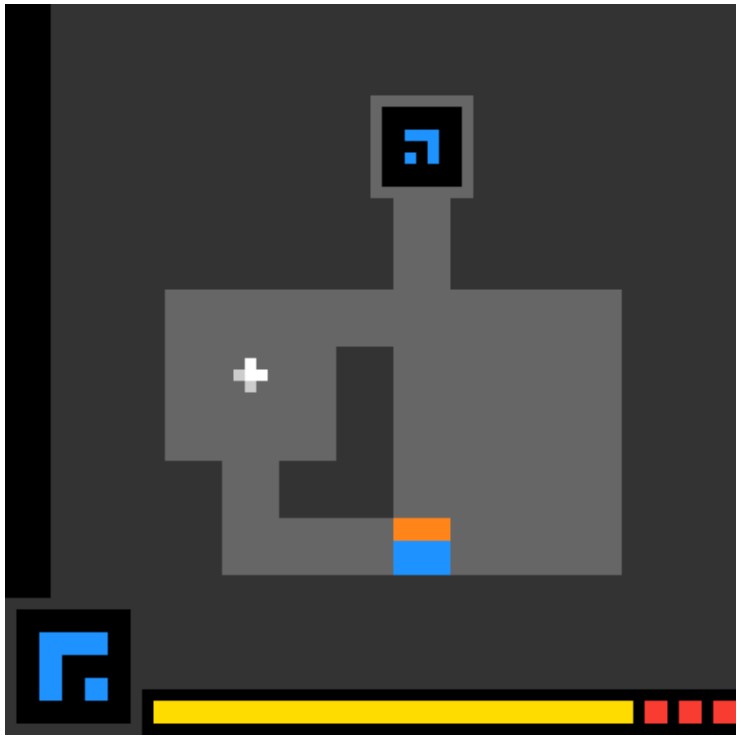
RHAE: level completion plus action efficiency relative to human baselines

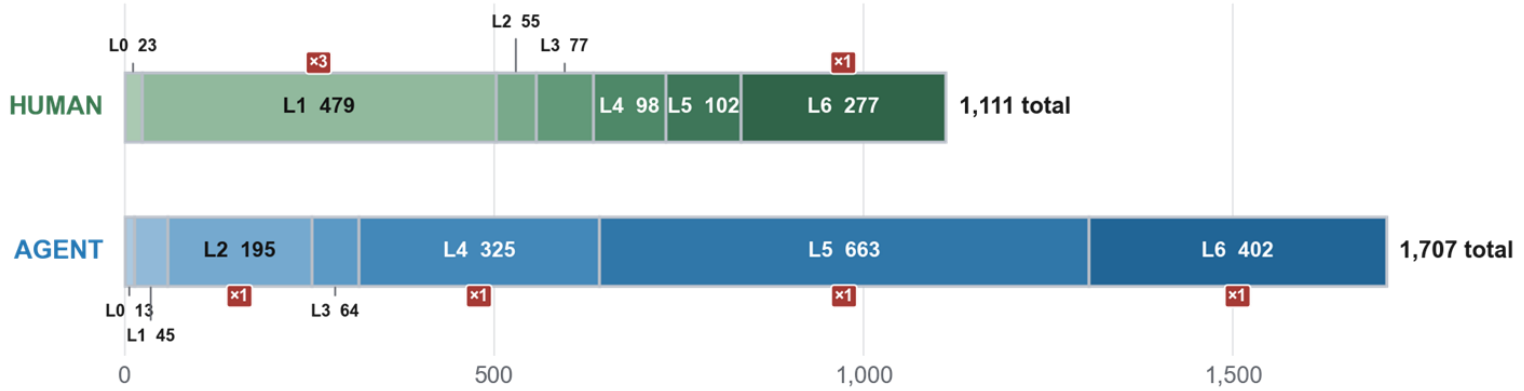
Internal inference has its own cost. An action score does not include it.

HUMAN RECORDING



ARTIFICIAL AGENT

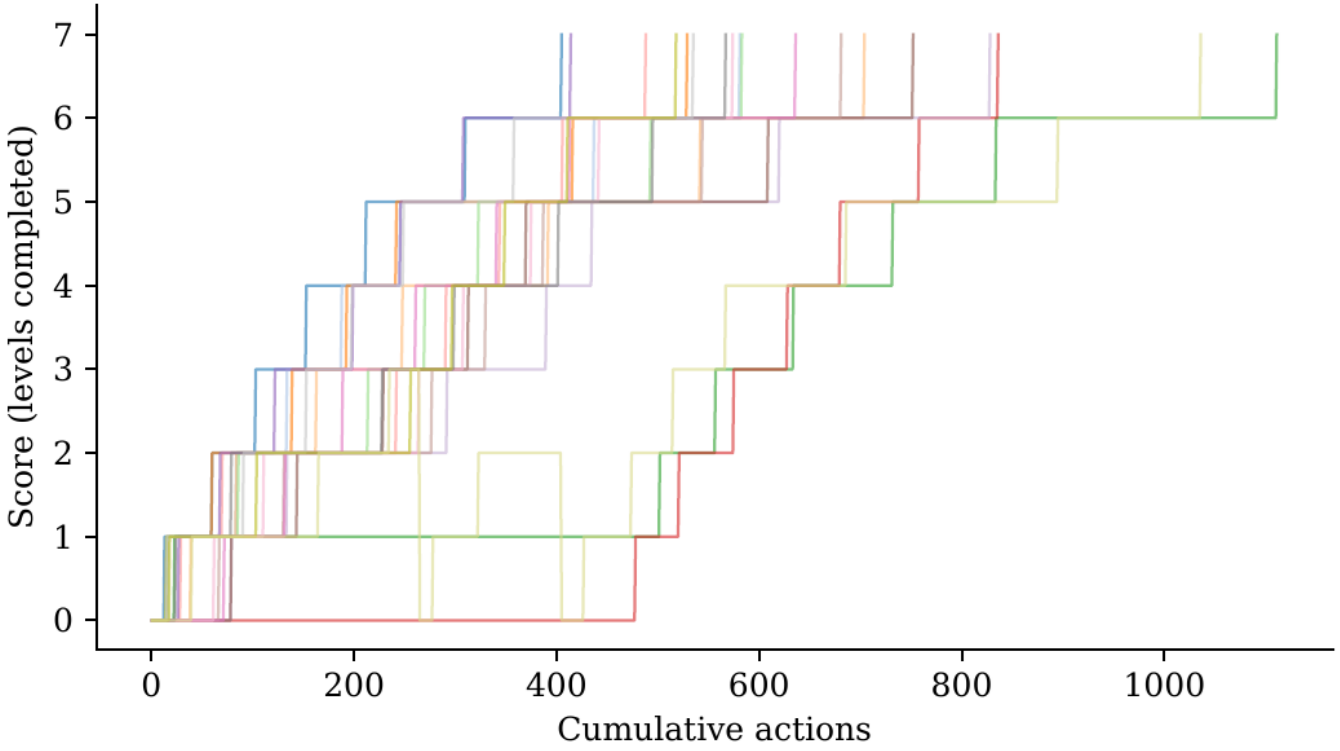




Visit website

Games as Experimental Philosophy: Human and Artificial Learning in ARC-AGI-3 Worlds, Andrei Aioanei, Syeda Khushbakht Batool, Alexander Aston, and Sahar Vahdati, ALIFE 2026

Human learning: equal success, unequal cost



18 / 18

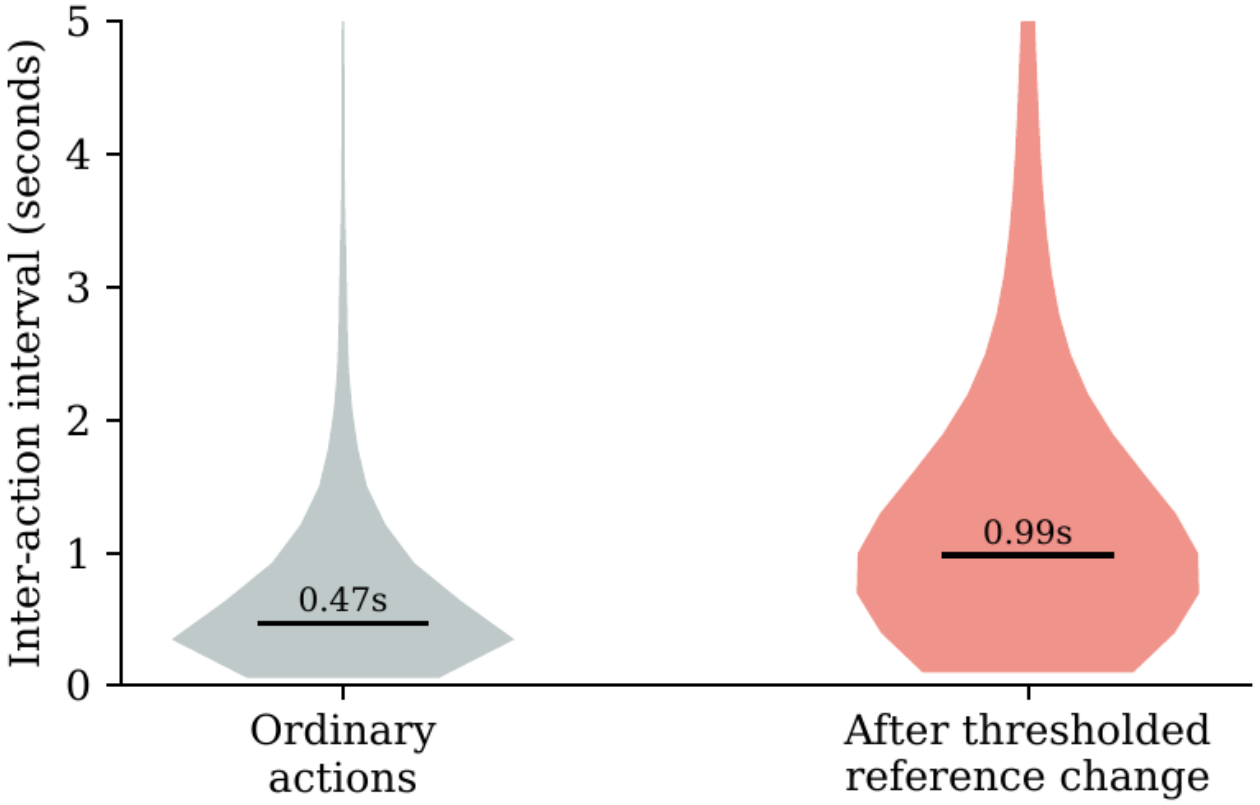
trajectories completed
all seven levels

405–1111

actions per trajectory

The learning path carries the efficiency signal!

Human players pause after reference changes



2.1×

longer median interval

Consistent with consequence checking

Association, not mechanism

ARC-AGI-3 progress within months

Date	System or milestone	Reported result	Evaluation set
25 Mar	ARC-AGI-3 launch	Frontier scores <1%	Semi-private
July	Duck release and Tycho	Tycho: 100 RHAE	25 public games
21 Aug	NVIDIA AVO	100 RHAE	25 public games
3 Sep	Astra + Provider Adapter	99.9%	55 semi-private
13 Oct	Gemini 4 Argon + Adapter	97.5%	Semi-private

Separate evaluations with different systems, test sets and budgets

Primary sources: ARC Prize, Tycho, Tufa Labs and NVIDIA. Checked 23 October 2026.

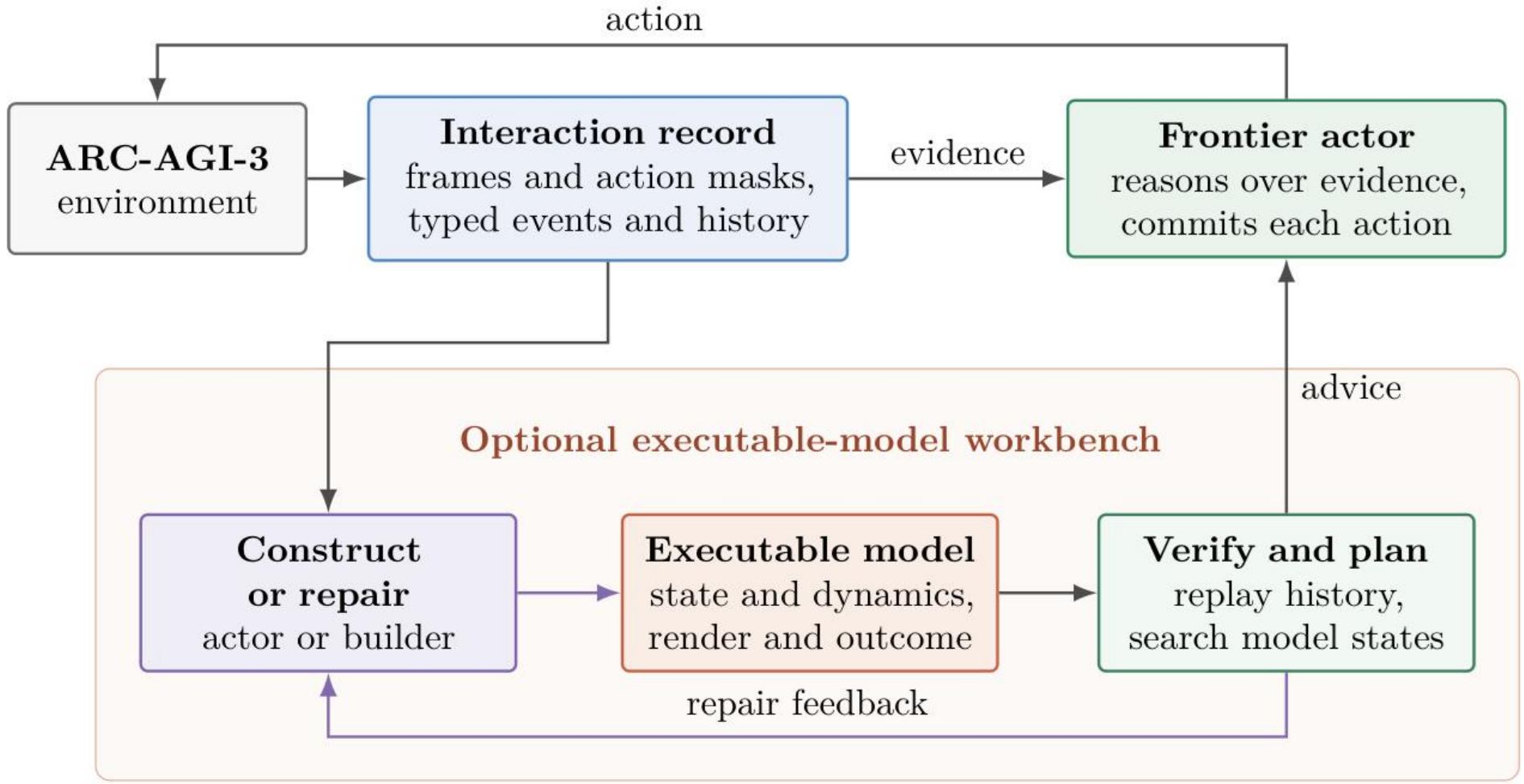
One model, two harness configurations

GPT-6 Astra, high reasoning, 55 semi-private environments

Harness	What it carries forward	RHAE	Reported cost
Standard	Visible model-selected notes	54.8%	$40,705
Provider Adapter	Opaque reasoning state and compaction	99.9%	$18,817

Evidence retained across turns can change the performance–cost profile of the complete system

Tycho: hypotheses as executable models



Tycho: fidelity to the past and useful action

Policy	Accepted transition match	Task RHAE	Cost / game
Actor delegates	16.2%	88.49	$231
Automatic repair	88.1%	83.07	$321

A future action needs the right objective and a model that is useful for that decision

One public-set run per policy. Observed-transition agreement is conditional on executable predictions.

Visit website

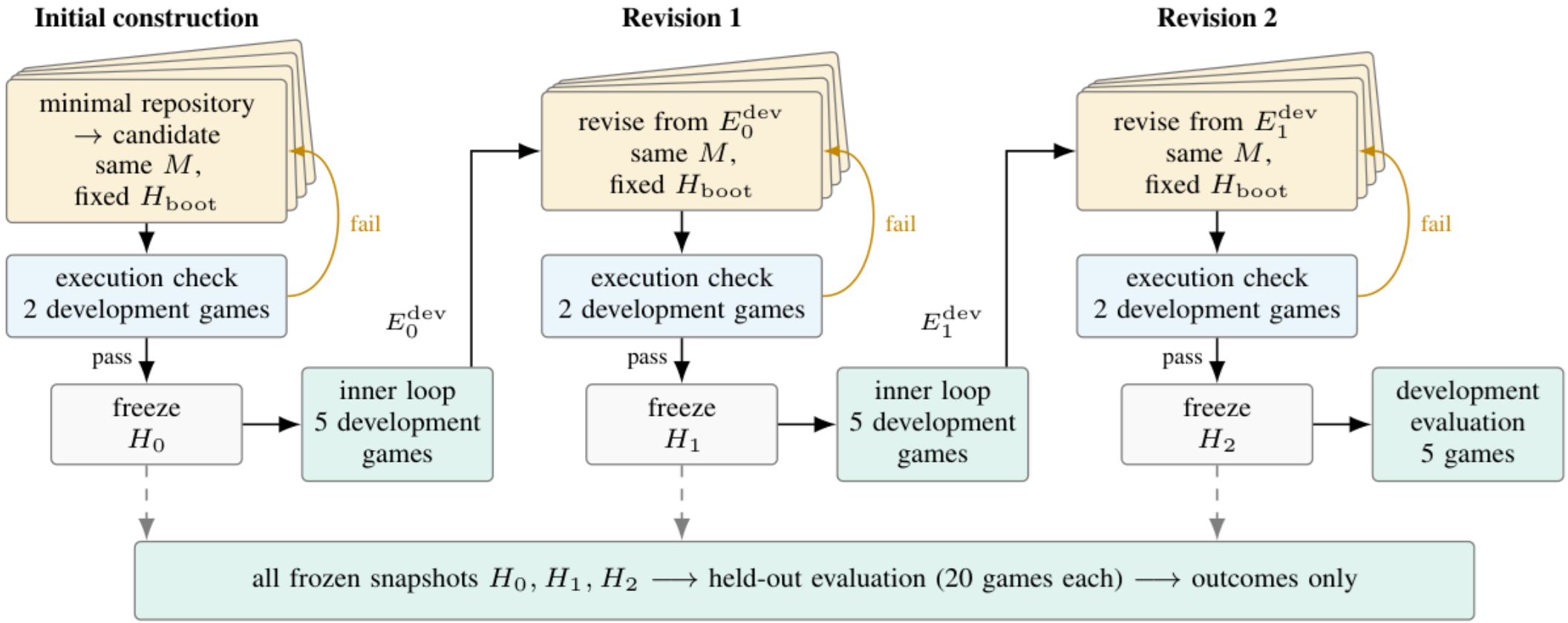
Tycho: Active abstraction with programmatic world models for ARC-AGI-3 Jens Lehmann, Andrei Aioanei, Sahar Vahdati, ARR Findings 2026

What has to generalize?

Scope	Transfer question
New level	Can the same learned rule support a changed situation?
Feedback-held-out game	Does a harness revision improve tasks outside adaptation feedback?
New domain	Does the learning process work beyond ARC’s format?

27 lineages with fixed weights and two revisions
5 development games, 20 feedback-held-out games

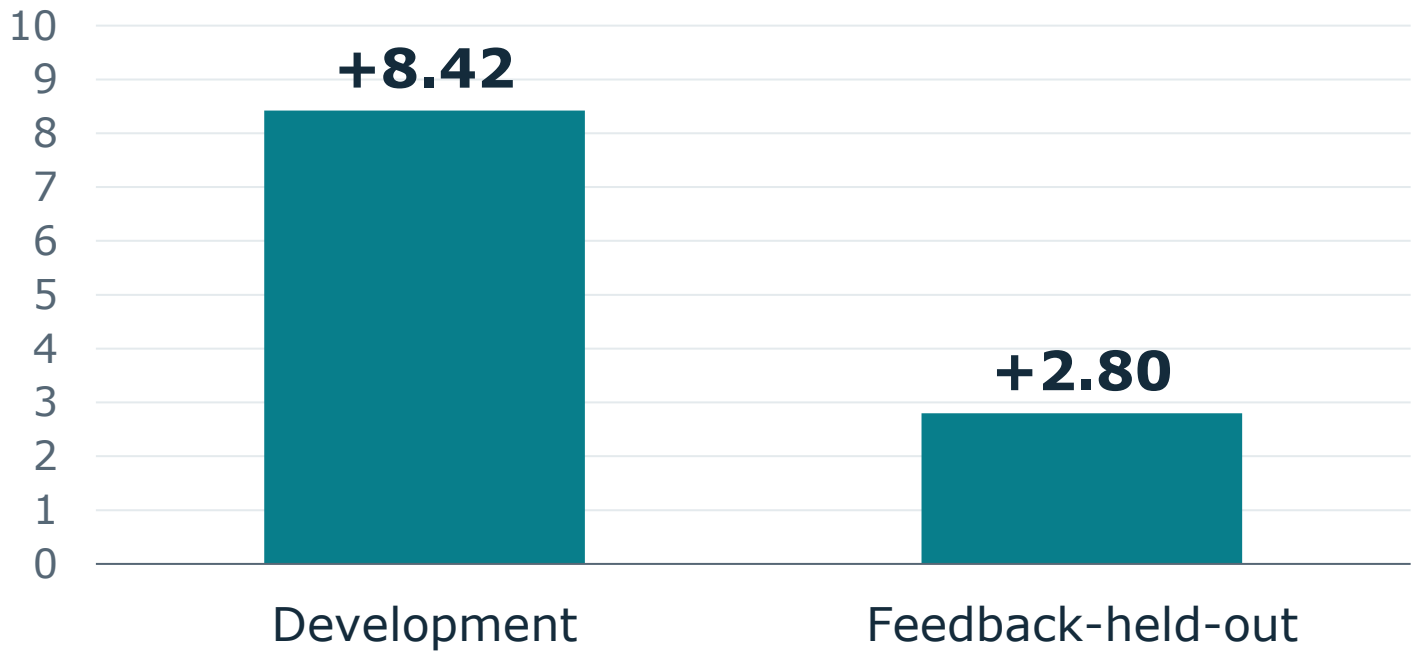
Autonomous harness adaptation protocol



Visit website

Self-written revisions: the transfer gap

Mean RHAE-point change after two revisions



Held-out 95% interval: −1.44 to +7.35

9 / 27

became cheaper and lost held-out performance

Improvement must transfer

Guessing: a rule still needs a reliable test

Qwen3.8-Flash-Next, Duck-derived harness
Public-game trace analysis under a one-GPU competition budget

50 selected stopped plays	Count
Stated a rule	50 / 50
Tested it before relying on it	11 / 50
Continued despite contrary evidence	32 / 50

Hypothesis. Predicted consequence.
Discriminating action and comparison.

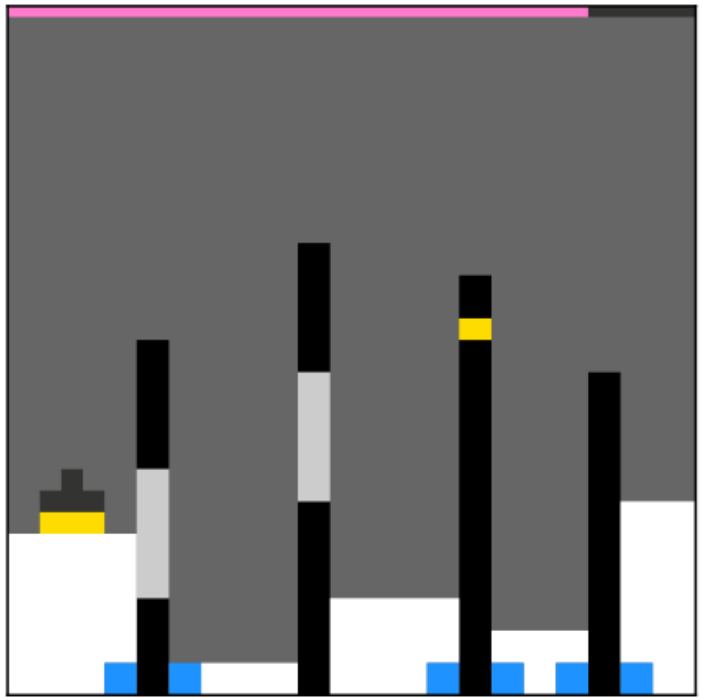
What should the harness contribute?

Harness help	What it supplies	What the experiments establish
Telling	Relations every turn	Context burden, fewer levels in GPU runs
Correcting	A fact at a selected mistake	Local action changes, whole-game gain unestablished
Pointing	A missed observation’s location	A benefit at one tested continuation context

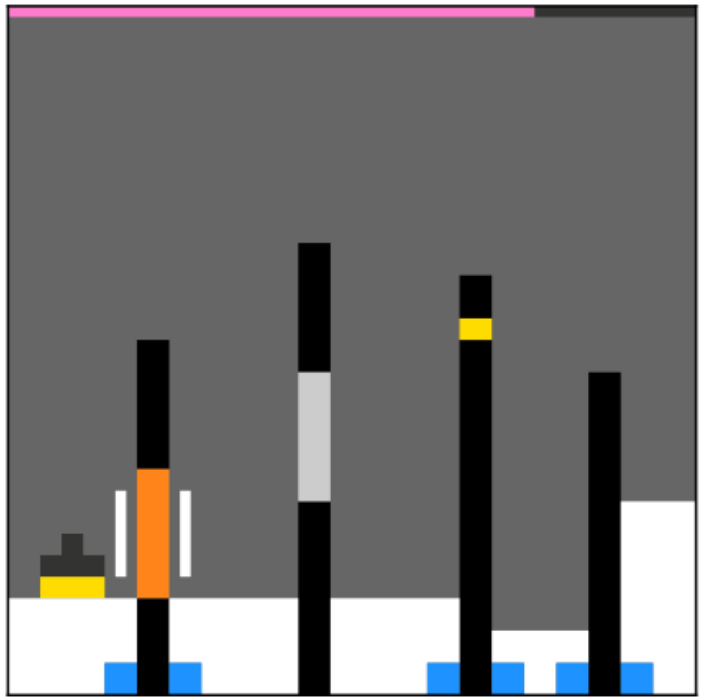
More assistance has to earn its context, inference and interaction cost

A decisive observation can disappear between turns

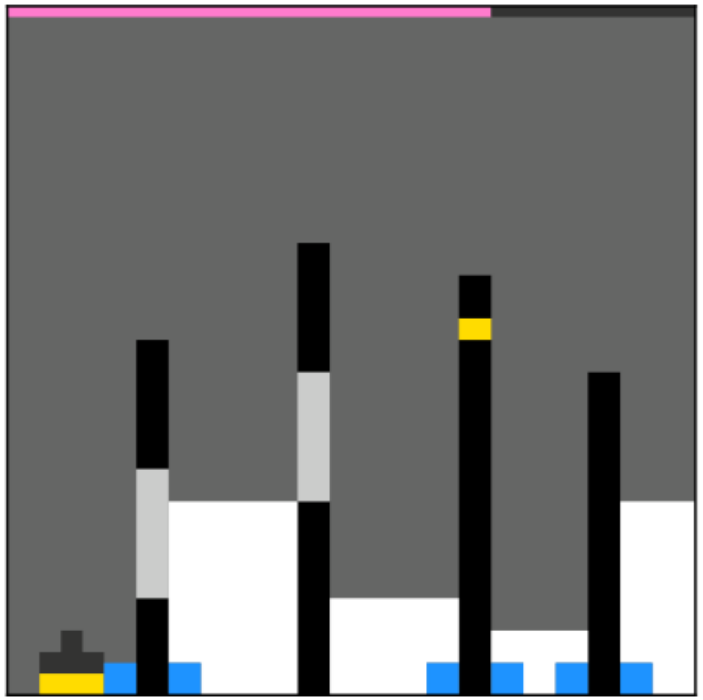
Before batch: shown



After click 4: not shown



After click 7: shown



Level completion: 9 / 20 plain 17 / 20 with a pointer

VC33 level 4. Other three turns: 14 / 30 in both arms. Hidden-set gain unestablished.

Guessing Is Not the Problem, Fig. 3 and Table 5, pp. 12–15

Learning to check also changes the cost

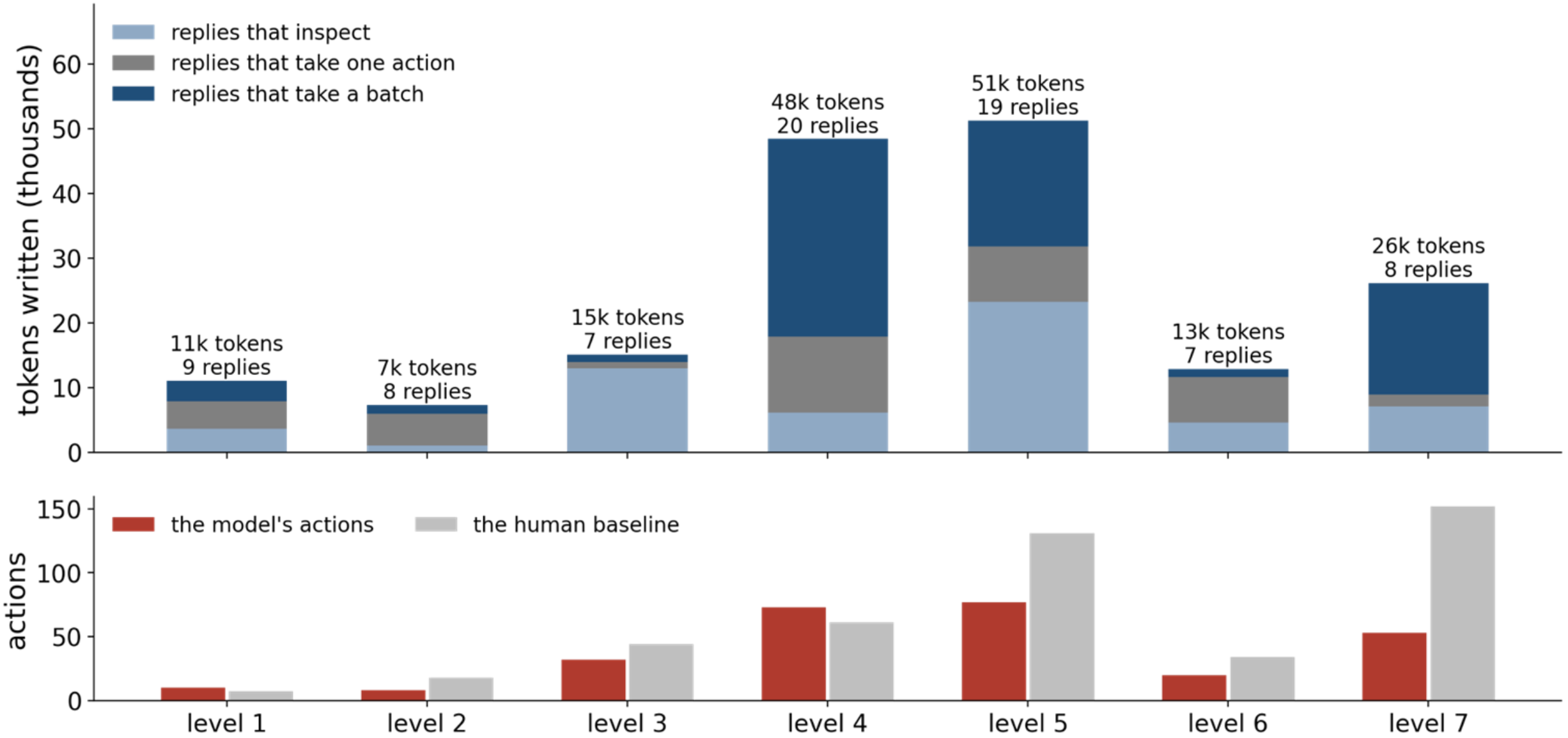
Measure	Base model	Checking adapter
Law-defining replies checked	18%	82%
Median reply tokens	378	1,370
Completed levels	141	55

30 held-out games, 48 minutes per game, one pass per arm

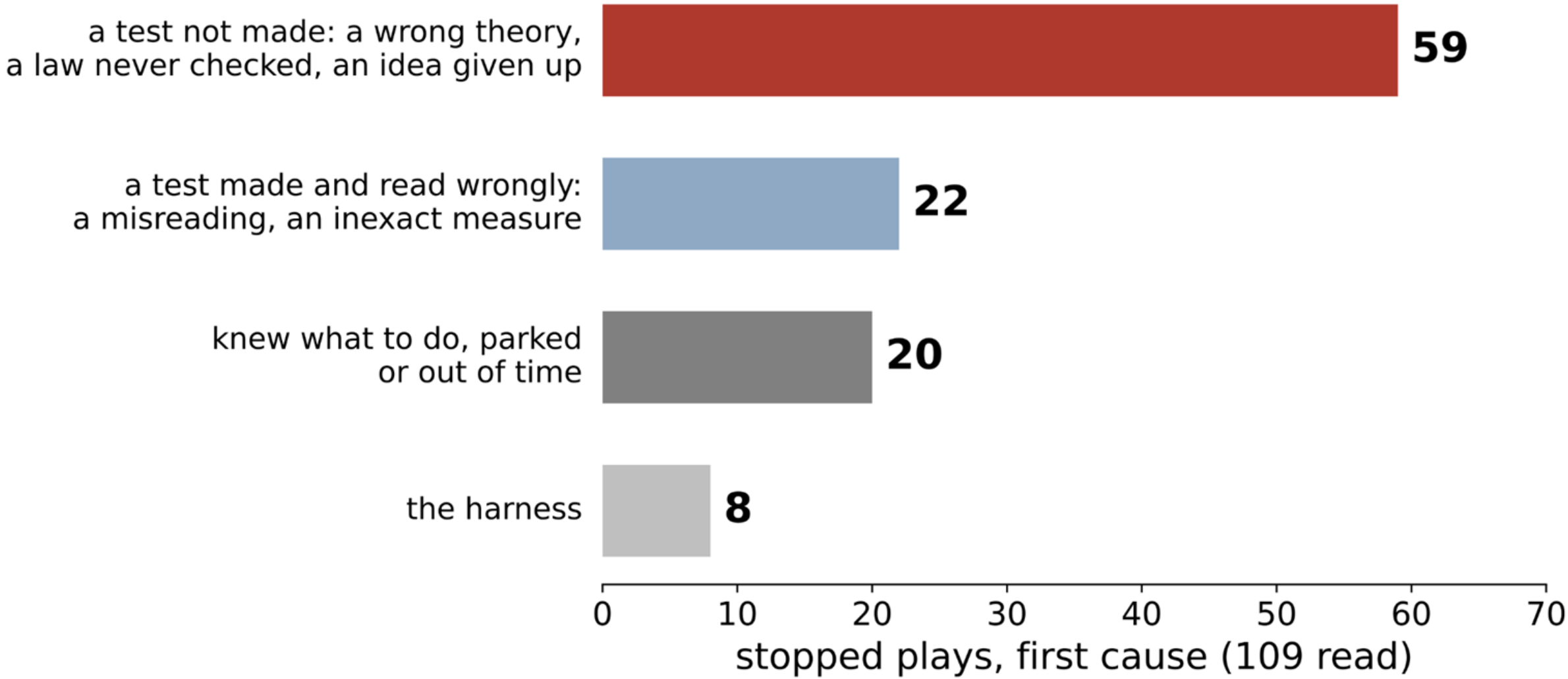
Missing: selecting useful tests and executing them accurately

Guessing Is Not the Problem, Table 7, p. 21. Preliminary training study.

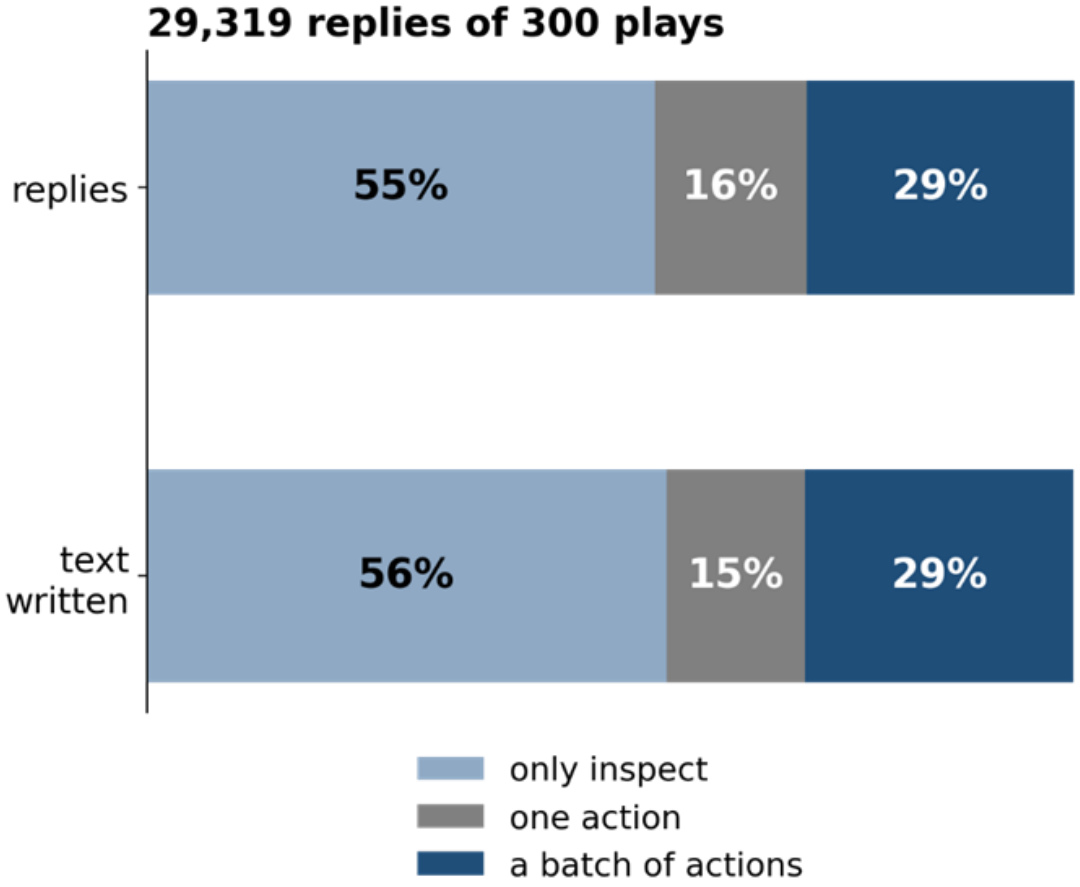
VC33, one play won in full: 172k tokens, 78 replies, 273 actions against a baseline of 447, score 100



What stops a play



Where a reply's tokens go



only inspect: read the board's objects

```
seg = current_frame.segmentation
for n in seg['nodes']:
    print(n['id'], n['color'], n['pixels'])
```

```
→ 0 P 64 ...  1 W 972 ...  2 G 2264 ...
```

one action: a click, then check it

```
r = action([{'action': 'MOUSE', 'row': 62, 'col': 40}])
print('done', r.get('level_completed'))
```

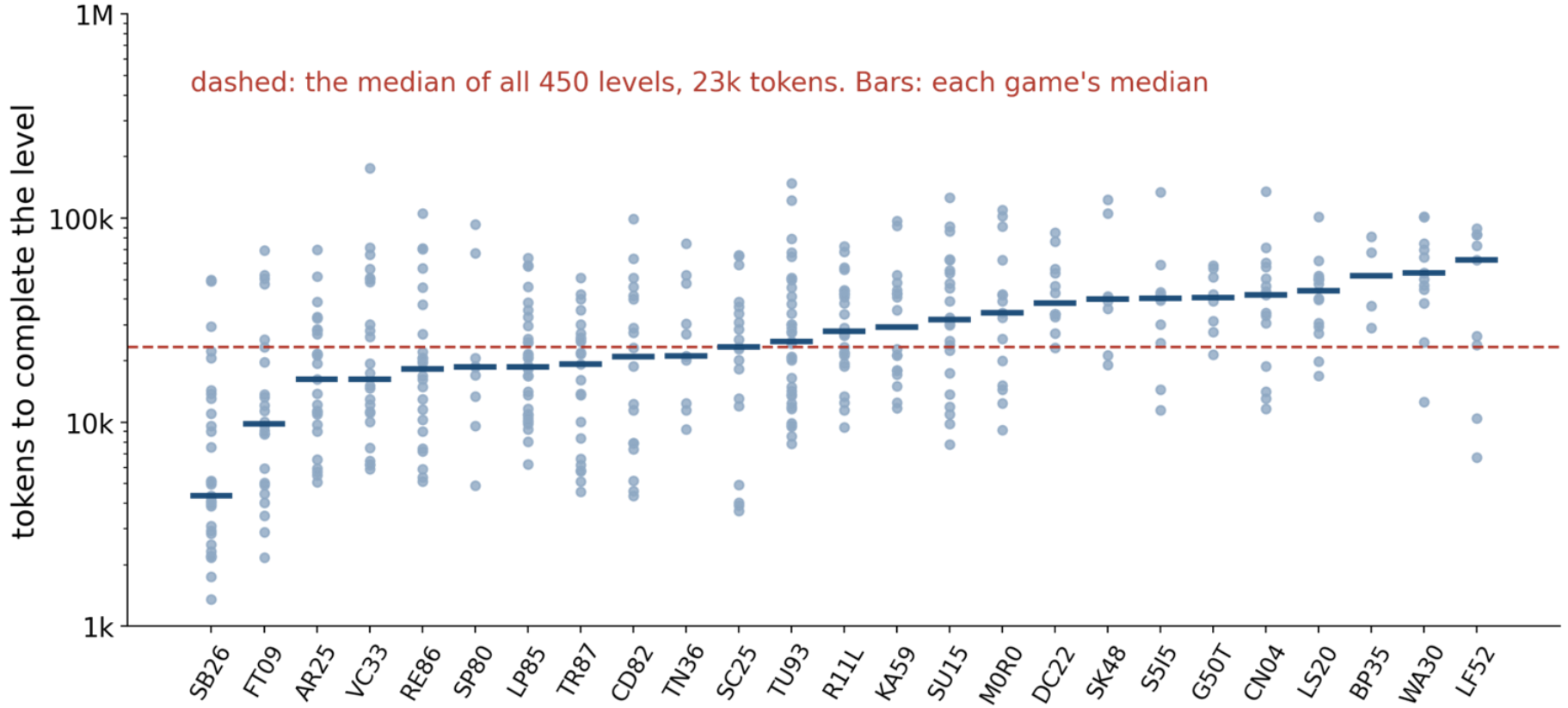
```
→ [action] MOUSE(row=62, col=40) -> LEVEL COMPLETED
```

a batch: four moves in one call

```
res = action(['UP', 'DOWN', 'UP', 'DOWN'])
print(res['state'], res['level_completed'])
```

```
→ [action] batch of 4 actions, 4 executed: UP, DOWN, UP, DOWN
```

What a level costs



Each dot is one completed level; the bar is the game's median; the dashed line the median of all 450.

Current capability and generalization gaps

Dimension	Demonstrated progress	Remaining evaluation question
Task performance	Ceiling or near-ceiling in named protocols	Transfer beyond this bounded game format
Resource efficiency	Strong action efficiency	Full inference, time and engineering cost
Autonomous adaptation	Models construct and revise harnesses	Reliable gains outside adaptation feedback

A transferable learning process needs economical, accurate tests

AI for science: the value of the next experiment

Inquiry step	Scientific requirement
Hypothesis	A candidate mechanism with testable consequences
Experiment	Conditions under which rival explanations disagree
Comparison	Accurate measurement against a stated prediction
Transfer	A finding or workflow that survives new conditions

Efficiency objective: resources per reliable new finding

Concluding notes

Efficiency	Count the full cost of acquiring a reliable capability: interaction, computation, and human engineering.
Accuracy	Choose informative tests, observe outcomes accurately, and revise beliefs from evidence.
Generalization	Test whether the learning process transfers to unfamiliar tasks and conditions with less redesign.

Aioanei, Andrei, Syeda Khushbakht Batool, Alexander Aston, Adrian Aioanei, and Sahar Vahdati.
Games as Experimental Philosophy: Human and Artificial Learning in ARC-AGI-3 Worlds.
Artificial Life Conference Proceedings, 2026.

Lehmann, Andrei Aioanei, Sahar Vahdati,,
Tycho: Active abstraction with programmatic world models for ARC-AGI-3 Jens
ARR Findings 2026

Jens Lehmann, Sahar Vahdati
Can LLMs Improve Their Own Harness?
arxiv October, 2026

NIMI

NATURE-INSPIRED
MACHINE INTELLIGENCE

AI for Science Lab





Sahar Vahdati



Adrian Aopanei



Mario Marven



Andrei Aioanei



Jason Li



Preetam Gattogi



Dmitry Muromtsev



Adrian Rolf



Zarin Nishat



Monireh Moshavash



Igor Kusnetsov



Khushbakht Batool



Kossi Amouzoivi



Hamze Yusuf



Lena Jurkschat



Mohammad Rezaei



Mahmoud Elsamadony



Linjie Zhong