合成样例 Synthetic sample：由故障注入的模拟策略生成，只演示诊断卡的版式与逻辑，不是任何真实模型的结果。

sim-fragmented

coding agent · Mind2Act World diagnosis card

MXPS

The Strategist feedforward planner

"thinks several moves ahead before acting"

reasoning-led · plan-ahead · precision-first

Benchmark	Mind2Act World · 30 tasks × 3 levels
Valid episodes	540 (0 invalid), 333 failed
Evidence	scorer, simulator events, tool-call log
Attribution	scorer events → deterministic rules (no LLM judge in this sample)
Reliability	stage agreement 85%, κ = 0.80 (vs. synthetic ground truth, n = 332)
Human review	46 episodes, 3 conclusions queued
Reference	expectation fitted on a cohort of 4 models

Summary sim-fragmented (coding agent) succeeds on 7.2% of the closed-loop Mind2Act episodes, 12.0 pp below expectation; tested alone, reasoning (Mind) reaches 60.6% (9.5 pp above expectation) and execution (Act) 47.2% (in line with expectation); n = 180 each. 52% of the Mind2Act failures sit in the verify-and-recover stage: the outcome of an action is not checked, feedback does not update the plan, or no recovery is attempted, so reasoning and execution hold up alone but do not join in one loop; deterministic rules settle 46% of all Mind2Act failures on their own. First recommendation: read the physical outcome after every step before choosing the next, closing the perceive–reason–act–verify loop; from the attribution it is expected to raise Mind2Act success by about 6.4 pp.

1 Task performance Mind2Act first

Mind2Act headline

closed-loop reasoning and execution

7.2%

expected 19.2% · 12.0 pp below expectation ▼

n = 180

Mind component

memory, rule and spatial reasoning

60.6%

expected 51.0% · 9.5 pp above expectation ▲

n = 180

Act component

contact-rich physical execution

47.2%

expected 44.8% · in line with expectation

n = 180

Mind2Act tasks need reasoning and execution in one closed loop and are the benchmark's headline; Mind and Act test each alone and locate the weakness. Expectations come from the model's overall level and the task difficulty.

2 Behavioral profile tendencies, not ability

The Strategist (feedforward planner): leans on reasoning and plans before it acts, then executes in one go.

Shines at: multi-step sequences that need look-ahead; next to build: add checks and corrections for when outcomes drift from the plan; refine precision and timing.

Mind vs. Act — Mind vs. Act tasks (10 tasks each)

Memory — entity vs. spatial memory (2 tasks each)

Control — planning & bookkeeping vs. feedback & repair (5 tasks each)

Execution — precise execution vs. dynamic response (4 tasks each)

M reasoning-led — at home with memory, rules and space	56% reasoning-led	A execution-led — at home with contact-rich handling	M stability 100%
O entity-anchored — remembers what each object is	51% entity-anchored	L location-anchored — remembers where things are	X stability 60%
P plan-ahead — plans first, keeps track of progress	85% plan-ahead	F feedback-driven — corrects step by step from feedback	P stability 100%
S precision-first — one-shot fine actions	56% precision-first	Q agility-first — keeps up with moving targets	S stability 92%

The profile describes tendencies, not ability. Each axis compares the model's relative results on two equal-sized task sets (every task weighted equally); the percentage is the strength of the lean, 50% being none. Letters come from a bootstrap: below 80% stability the letter is X. Stability says whether the lean is consistent, not how large it is.

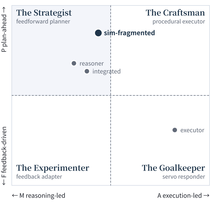


Figure 1 x: lean on the reasoning–execution axis; y: lean on the control axis, as in the bars above; the centre is no lean. Grey dots: the other 3 models of the cohort. A model whose lean on either axis is unstable is an all-rounder (X).

The Strategist feedforward planner

"thinks several moves ahead before acting"

Shines at: multi-step sequences that need look-ahead

Next to build: add checks and corrections for when outcomes drift from the plan; refine precision and timing

The Experimenter feedback adapter

"acts, observes the outcome, revises"

Shines at: changing scenes that reward adjusting on the way

Next to build: keep long sequences complete and sharpen one-shot precision

The Craftsman procedural executor

"carries every step of a procedure out cleanly"

Shines at: fixed, repetitive procedures

Next to build: build memory for changing information and induction of new rules

The Goalkeeper servo responder

"tracks the target and adjusts on the move"

Shines at: dynamic interception and feedback-based fine adjustment

Next to build: build multi-step memory and planning

3 Capabilities and robustness 11 categories; 5 difficulty pressures

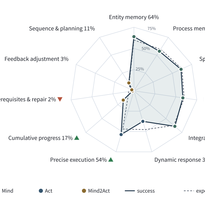


Figure 2a Success per category (solid) and its expectation (dashed), from the model's level on the view and the category's difficulty; ▲ / ▼ above / below expectation.

Difficulty pressure: Easy → Hard

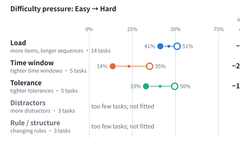


Figure 2b Each pressure changes one difficulty factor. Dots are fitted success (one intercept per task and a shared slope, logit P = a_task − β · level, which removes the difficulty differences between tasks); the right column is the Easy → Hard drop. Only load covers enough tasks to be called on its own; time window and tolerance qualify weaknesses only.

4 Failure attribution where the loop breaks

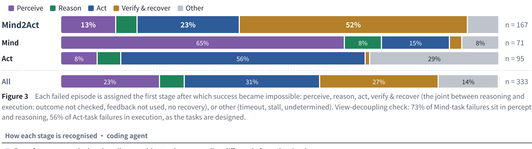


Figure 3 Each failed episode is assigned the first stage after which success became impossible: perceive, reason, act, verify & recover (the joint between reasoning and execution: outcome not checked, feedback not used, no recovery), or other (timeout, stall, undetermined). View-decoupling check: 73% of Mind-task failures sit in perception and reasoning, 56% of Act-task failures in execution, as the tasks are designed.

How each stage is recognised · coding agent

Perceive	the log describes an object, colour or reading differently from the simulator
Reason	plan, memory or rule at odds with the task: missing or misordered steps, misremembered state, a refuted plan repeated
Act	simulator events: reach short, empty grasp, slip, misplacement, collision, missed time window
Verify & recover	after a failure event the log moves on without checking, updating or recovering
Other	timeout, stall, or not enough evidence (sent to review)

Most frequent failure signatures

Task	Failure code	Episodes
E1	M-PERCEIVE perception error	8
F1	M-PERCEIVE perception error	7
CP01	C-NO-VERIFY advanced without verifying the outcome	7
P2	M-PERCEIVE perception error	5
CP02	M-PERCEIVE perception error	5

49% recovery after slips (76 slips) sim

30% failures that place or press on a wrong object sim

0.04 collisions per episode sim

5.5 tool calls per episode log

87% misses acknowledged in the tool log log

Attribution reliability

85% stage agreement

0.80 Cohen's κ

81% failure-code agreement

14% undetermined (to review)

Against the injected ground truth of the synthetic data (not human labels), 332 failed episodes; undetermined calls are left out of the agreement. Main disagreements (harness→label): Perceive→Reason 35, Verify & recover→Reason 2.

5 Recommended interventions agent modules

Mind2Act
verify & recover
priority 1

Evidence
reasoning and execution each hold up alone but fail when joined in one loop (52% of these failures)

Module: outcome verification
read the physical outcome after every step before choosing the next, closing the perceive–reason–act–verify loop

Expected gain
Mind2Act success 7.2% → about 13.6% (+6.4 pp)
if these 87 failures stop happening and the rest stays as is; ceiling +48.3 pp

Re-evaluation
re-evaluate on the Mind2Act tasks

Affected domains
Home & Food Service; Instruments & Tool Use; Assembly & Physical Reasoning; Companion & Edutainment

Mind2Act
overall
priority 2

Evidence
cannot reach, grasp, place or press reliably (46% of these failures)

Module: execution control
approach in small corrective steps and re-read the camera; check gripper opening after grasping; align to the millimetre before placing

Expected gain
Mind2Act success 7.2% → about 8.7% (+1.4 pp)
if these 32 failures stop happening and the rest stays as is; ceiling +17.8 pp; together with priority 1 about +12.6 pp

Re-evaluation
re-evaluate on the Mind2Act tasks, focusing on Prerequisites & repair, Feedback adjustment

Affected domains
Home & Food Service; Instruments & Tool Use; Assembly & Physical Reasoning; Companion & Edutainment

Act
Dynamic response
preliminary lead: few tasks, to confirm — to confirm

Evidence
cannot reach, grasp, place or press reliably (63% of these failures)

Module: execution control
approach in small corrective steps and re-read the camera; check gripper opening after grasping; align to the millimetre before placing

Expected gain
Act success 47.2% → about 52.0% (+4.8 pp)
if these 17 failures stop happening and the rest stays as is; ceiling +9.4 pp

Re-evaluation
re-evaluate on the Act tasks

Affected domains
Manufacturing & Logistics; Instruments & Tool Use

Expected gain by competing risks: each removed failure is credited with the success rate of the episodes of the same task that did not fail this way; causes are assumed independent, so this is an estimate, not a promise. The ceiling counts every removed failure as a success. Steps on one view add up.

6 Scope and limitations

- Conclusions describe the model on the Mind2Act World tasks only and do not extend to its general ability.
- Expectations come from an additive logit model (model level + task difficulty) fitted on a cohort of 4 models; strengths and weaknesses are relative and called on 95% intervals.
- Conclusions are drawn on the three task views; each of the 11 categories has 2–6 tasks, so category differences are leads only.
- For a coding agent the tool-call log is used as well; it is one piece of evidence, and the physical outcome is taken from the simulator.
- Expected gains use competing risks and assume independent causes; they rank and size the steps, they do not promise them.
- Not called for lack of evidence: Distractors and Rule / structure pressures (too few tasks to call).
- 3 uncertain conclusion(s) are queued for human review, which this card does not wait for.