Closing the Loop for Embodied AI: Benchmarking Decoupled Cognition, Execution, and Composition



5 Real-World Scenarios



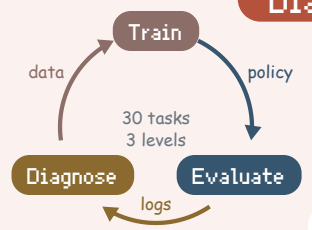
12 Capability Categories

Mind — 5 categories
Entity · Rule · Process · Spatial · Integrated

Act — 2 categories
Dynamic · Precision

Mind2Act — 5 categories
Cumulative · Repair · Feedback · Plan · Predict

Diagnosis Loop



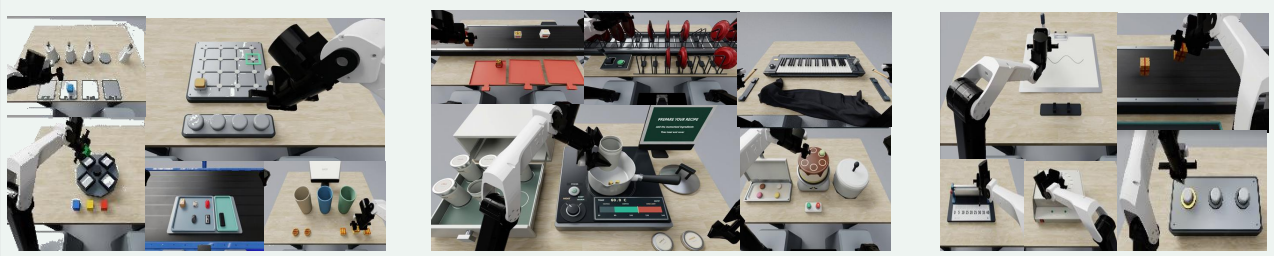
where it broke: reasoning · execution · coupling

report by: capability · pressure · domain · behaviour type

directions: capability × cause × pressure → data

GPT-6 Astra SR: Mind 52.0 → Mind2Act 1.3

30 Tasks Across 3 Views



Mind: Mem. span↑ · Logic depth↑

Mind2Act: Close loop horizon↑ · Error compound↑

Act: Dynamics↑ · Precision↑

Progressive 3 Pressure levels for each task

Benchmark Highlights

Isaac Sim · Bimanual · Easy · Medium · Hard



Process-Level Event Trace · CP01 conveyor meal assembly: did the effect propagate to the next decision?



Open-Source Ecosystem

- **Public Leaderboard** — VLA / WAM and Coding-Agent boards
- **Diverse Policies** — VLA / WAM · LLM agents · Hierarchical
- **90 Task Configs & Demo Videos** — 30 tasks × Easy / Medium / Hard
- **Open Code & Task Specs** — repository · structured data · media index
- **Mind2Act Harness** — Astra-Jev cognitive + reactive agents
- **Community** — WeChat · QQ · Feishu

Mind2Act PolicyLab

Unified Infra & Standard for Policy Development and Evaluation

20+ Policies · 3 Paradigms *Growing…*

- VLA / WAM — end-to-end
- LLM as Policy — coding agents
- Hierarchical — planner + executor

Policy (model / agent) ⇄ Standard Protocol ⇄ Env (Isaac Sim · Real*)

Leaderboard

Simulation · 15 tasks × 3 levels · Preliminary

Strong Mind ≠ strong Mind2Act

	Model	Mind	Act	Mind2Act	Overall
LLM as Policy · coding agent + harness					
1	GPT-6 Astra *Codex*	59.0	20.0	19.8	**32.9**
2	Gemini 3.8 Flash *Antigravity CLI*	20.7	19.2	3.8	**14.6**
3	DeepSeek-V4.1-Flash *Claude Code*	17.5	17.4	5.2	**13.2**
VLA / WAM · end-to-end					
1	π0.5	3.0	24.1	15.4	**14.1**
2	KinRT	3.8	16.2	7.9	**9.3**
3	Galaxea G0.5	2.8	15.6	3.8	**7.4**
4	X-VLA	0.6	7.2	7.4	**5.1**

Avg. stage score (0–100) per track over 5 tasks × 3 levels; Overall = mean over all 45 task-difficulty configs. 5 runs / config for coding agents, 50 for end-to-end policies. Preliminary internal runs (Oct 2026), not yet frozen. *Real-world protocol planned.