SEAR: Self-Evolving Agentic Robotics

LLM agents can control robots, even text-only ones. SEAR rigorously tests how they self-evolve.

Xinyi Wang*† · Bangzheng Li*† · Hanyu Wang‡ · Lin Zhang · Andrea Yaoyun Cui · Zhenggang Tang · Shyamal Buch§ · Dejia Xu†§

*Co-first authors †Project leads ‡Real robot lead §Co-last authors

Luma AI

CoRL

sear.bot

01 Abstract

To our knowledge, SEAR is the first large-scale (4+ years of agent runtime) controlled study of self-evolving robot agents.

Frontier language-model agents can now improve robots on their own: they write control code, test it, and keep what works as skills and memories for the next task.

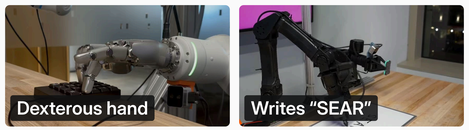
We built **SEAR, a new benchmark and harness**, to answer two questions:

(1) Is the agent really learning, or just trying more often?

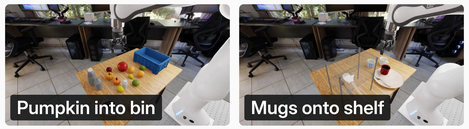
(2) Does what it learns still help on new tasks, in a new simulator, and on a real robot?

7	330	5	4+
LLMs	sim tasks	real tasks	years of agent time

On real robots

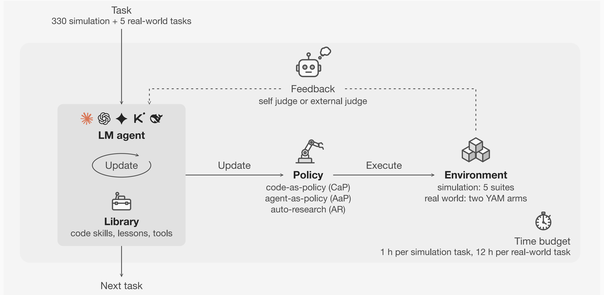


In simulation (RoboLab, Isaac Sim)



02 Benchmark and Harness

The self-evolving loop



SEAR runs the agent in all three modes

Code-as-policy
The agent writes code; no LLM runs on the robot.

Agent-as-policy
The LLM itself acts, one step at a time.

Auto-research
The agent trains a VLA on its own rollouts.

330 tasks in five simulated families, two physics engines



robosuite	LIBERO	LIBERO-PRO	LIBERO-RIG	RoboLab
7 tasks	40 tasks	160 tasks	18 tasks	105 tasks

Controls against a misleading curve

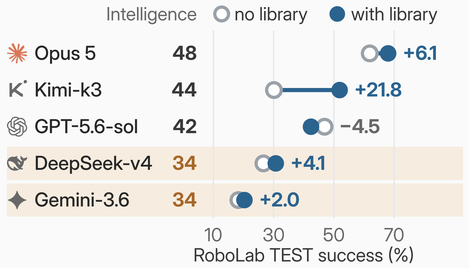
More attempts	→ **best-of-3 at the same budget**
Practice gains	→ **five generalization tiers**
Privileged state	→ **a sealed judge, unseen states**
One lucky run	→ **three replicates of every simulated LLM run**
A bigger budget	→ **matched wall clock**

03 Result Highlights

More analysis in the paper →

1 Better LLM, better robot?

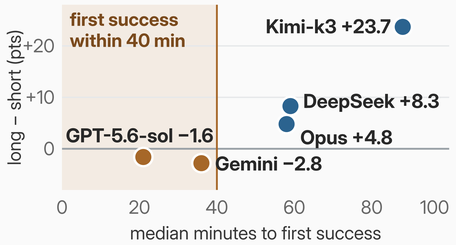
Same LLM intelligence score can yield different robots. Above a capability floor, better LLM still matters, but library can matter more.



Intelligence: Artificial Analysis Intelligence Index [10] (Kimi-k3, DeepSeek: max setting). DeepSeek is text-only.

2 Learning, or trying again?

One long session beats three short ones only when the first success comes late



RoboLab: one 2-h session vs. the best of three 40-min attempts.

3 Swap in a better library

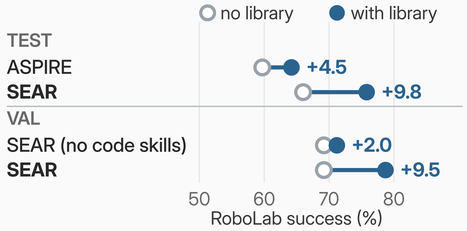
Experience saved in a library can transfer between models



GPT-5.6-sol, RoboLab TEST (%), 3 runs per library.

4 Harness matters: SEAR vs. ASPIRE [1]

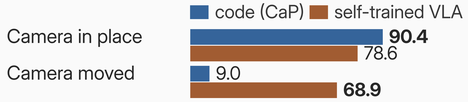
SEAR's library adds more: skills kept as code that later policies import



Opus 5 in Claude Code for both; same order, budget, toolbox, and judge. First 20 RoboLab tasks, 3 repeats.

5 Auto-research: the agent's own VLA

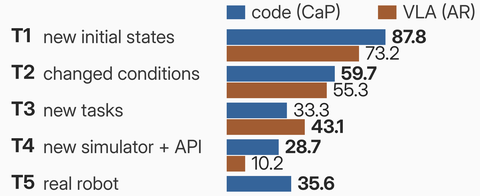
Agents ignore the off-the-shelf VLA, but a VLA they train themselves covers code's blind spot



Toolbox $\pi_{0.5}$: called by 2% of no-library RoboLab policies. Bars: Opus 5 under LIBERO-RIG's hidden faults.

6 Skills travel

Control code and grasp routines carry across tasks and simulators



Opus 5, same time budget for code and VLA; mean over simulated family pairs. T1–T2 frozen; T3–T4 written after self-exploration; T5 practiced on the robot.

7 Real-robot results

Models learn in their own style: some do best acting live, others writing code first



Banana into Bowl · Push-T · Cup Upright



Block Stacking · Spell "SEAR"

Held-out score, %	Frozen code CaP	Live agent AaP
Fable 5.1	70.2	66.4
GPT-6-astra	41.4	88.0

REFERENCES

[1] Lu et al. **ASPIRE.** arXiv 2607.00272, 2026. [2] Liang et al. **Code as Policies.** ICRA 2023. [3] Zhu et al. **robosuite.** arXiv 2009.12293, 2020. [4] Liu et al. **LIBERO.** NeurIPS 2023. [5] Zhou et al. **LIBERO-PRO.** arXiv 2510.03827, 2025. [6] Cui et al. **LIBERO-RIG.** arXiv 2610.18843, 2026. [7] Yang et al. **RoboLab.** arXiv 2604.09860, 2026. [8] Black et al. **$\pi_{0.5}$.** arXiv 2504.16054, 2025. [9] Sundermeyer et al. **Contact-GraspNet.** ICRA 2021. [10] Artificial Analysis. **Intelligence Index.** 2026.

Paper
latest version

Code
GitHub