NeurIPS 2026 · Evaluations & Datasets

ClawBenchPro: Benchmarking How Well Agent Harnesses Work

Yufei Liu[1]* · Yirong Zeng[2]*† · Yuxiang He[3]* · Yutai Hou[4] · Yuxian Wang[4] · Qunyao Du[2] · Wang Xu[4]
Xiao Ding[2] · Wu Ning[4] · Hao Cong[5] · Dandan Tu[4] · Qixun Zhang[1] · Ting Liu[2]

[1] Peking University [2] Harbin Institute of Technology, SCIR Lab [3] Sichuan University
[4] Huawei Technologies Co., Ltd. [5] Tsinghua University * Equal contribution. † Internship at Huawei.

HIT SCIR · Peking University · Huawei · Sichuan University · Tsinghua University

OPEN BENCHMARK

Code · Data · Harness adapters

Reproducible task environments and workspace-based grading

github.com/Herrieson/nanoclaw

CODE + DATA

TL;DR — Agent performance depends on **both the model and its harness**. ClawBenchPro tests 1,000 tasks across 43 domains with skills, noise and multi-session dependencies. Four-harness evaluation reveals **ranking reversals**, a **33.3–38.6 pp reliability gap** and an **≈11× token difference**. State recovery is the shared bottleneck: partial progress rarely becomes complete execution.

Contributions: scalable synthesis, reusable tasks, joint evaluation.

1 Background & Motivation

Agent harnesses turn language models into autonomous workers by managing tools, memory, skills and retries. Their integration choices affect whether a model can finish a professional workflow.

Existing benchmarks cover few domains: PinchBench has 53 tasks / 8 domains, and Claw-Eval 300 / 9. Narrow coverage and single-harness testing can distort model comparisons.

Key question: which model–harness pair remains reliable when tasks require tool use, noisy retrieval and continuity across sessions?

2 Benchmark Construction

2,000 seeds → 430 curated → 3,600 built → 1,080 QA → **1,000 final**



13 sectors / 43 domains, ≥20 tasks per domain. Community seeds exclude credentials (620), duplicates (510) and trivial goals (440).

Gemini 3.0 builds task assets. **QA keeps the top 30%** (mean 7.2/10); Qwen3.5-Plus / OpenClaw avg@16 retains **10–90% success**.

Persona-based queries vary user narratives.

Fig. 2 · Domain-balanced tasks from community seeds.

3 Complexity Augmentation



Fig. 3 · Turns vs. tokens.

Base: clean workspaces. **Skill:** 388 skills; 2.9 skills/task in that subset.
Environment: 45–340 files, 4–68% decoys; nesting 1.4–3.2 levels.
Multi-session: 2–4 sessions require state written to files and memos.

These settings separate **tool orchestration**, noise robustness and durable memory from success on clean tasks.

4 Evaluation Protocol & Validation

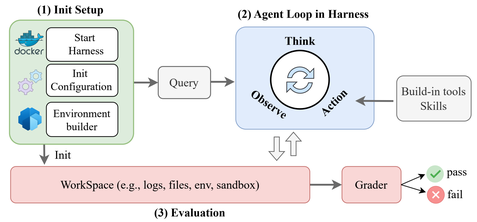


Fig. 4 · Initialize; act / observe; grade final artifacts.

Executable tasks: runtime YAML, narrative query and Python workspace builder run in isolated Docker. Models share tools and configuration within each harness.

73% deterministic checks | **27% semantic**

Avg. score measures the fraction of rubric checks passed; **perfect pass** requires every check. Graders test schemas, numerical tolerances, tone and coherence.

Why artifacts? Judge tangible outcomes, not plausible-looking traces.

5 Experimental Setup

Models / harnesses: 14 LLMs from GPT, Claude, Gemini, Qwen, DeepSeek, MiniMax and GLM run on OpenClaw, NanoClaw and Hermes-Agent; five also run on Codex.

Settings: defaults and native extended thinking. Qwen3.5-Flash provides semantic judging and API mocks.

92% **Judge–human agreement** (κ = 0.83; 200 outputs, 3 annotators). Alternate judges preserve Top-3 (τ = 0.87 / 0.81).

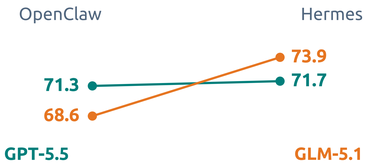
6 Results — Model–Harness Interaction

+12.0 pp

57.4 NanoClaw → 69.4 Hermes

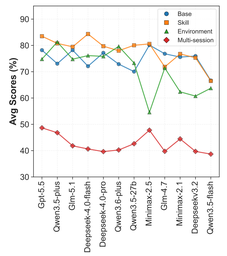
DeepSeek-V3.2

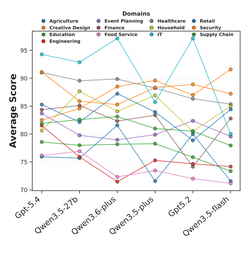
Leader flips



Harness choice changes scores and observed leaders: **model capability and orchestration interact**. Qwen3.5-Plus / Codex reaches **74.44%**, the highest observed score. Selecting a model alone misses these deployment trade-offs.

7 Robustness & Failure Diagnosis





39–49% on multi-session vs. **66–80%** on Base: every model struggles with continuity.

Figs. 7–8 · **IT / creative design lead**; engineering and supply chain lag. State recovery causes **33–42%** of failures, ahead of noisy retrieval (21–25%) and orchestration (17–24%): 100 traces, κ = 0.81.

8 Reliability, Efficiency & Repeatability

33.3–38.6 pp reliability gap

≈11× fewer tokens

OpenClaw average minus perfect-pass gap. GPT-5.5: **71.29% avg., ≈34.7% perfect** (rounded-gap derivation). Same five models: Codex **70.26% / 9.2k tokens** vs. OpenClaw 66.97% / 101.6k; tokens proxy resource use.

Repeatability: 8 NanoClaw models × 1,000 tasks × 3 trials: SD 3.6 pp, ρ = 0.96. Human reference: 80.1% vs. model mean 74.8% / best 78.6% (3 people, 30 Base tasks). **Small score gaps need caution.**

+ Limitations & Future Directions

Scope: synthetic sandboxes and API mocks, only four harnesses, and substantial evaluation cost limit deployment claims. **Static tasks risk adaptation and contamination.**

Next: refresh seeds through community maintenance, add dynamic real-world scenarios and broaden harness coverage; strengthen cross-session state management.

9 Conclusions & Takeaways

Co-evaluate models and harnesses. ClawBenchPro couples broad professional coverage with executable tasks and artifact-based grading. **Prioritize full completion and persistent state;** average scores alone hide the largest reliability failures.

Contact: xding@ir.hit.edu.cn