Harness adapters explained: training a LoRA with RL inside the agent harness you actually use

23 November 2026 · 7 min read · blogs.iictp.org/harness-adapters-rl-lora-agent-harness-fine-tuning-explained

Short answer: a harness adapter is a small LoRA adapter trained with reinforcement learning inside a specific agent harness, so a frozen model becomes more accurate and uses fewer tokens for your tasks. A Hugging Face post of 20 November 2026 reports that, for Qwen3.8-27B under the pi harness, Terminal-Bench 2.1 pass@1 rose from 39% to 55% while average tokens per task fell from 2.9K to 1.5K. These are the authors' own results.

TL;DR The short version

- A harness adapter is a small LoRA trained with reinforcement learning inside the agent harness you actually use.
- The authors report Qwen3.8-27B in the pi harness rising from 39% to 55% on Terminal-Bench 2.1 while using fewer tokens.
- Results are self-reported on limited tasks, but it is a cheaper route than full fine-tuning and keeps the base model reusable.

Read online, listen to it, or change the text size: https://blogs.iictp.org/harness-adapters-rl-lora-agent-harness-fine-tuning-explained

What is a harness adapter?

An agent harness (such as pi, OpenCode or Hermes Agent) wraps a model with tools, prompts and a loop. The post's claim is that general models are not tuned for any one harness, so training a LoRA with RL on tasks run inside the real harness squeezes out extra accuracy and efficiency. Weights of the base model stay frozen.

How it works

- **Train in the real harness.** Rather than re-implement the agent loop in a training framework, the harness runs as shipped; a proxy between the harness and vLLM records exactly the tokens it produced.
- **Tasks and verifiers.** Tasks use the Harbor format. Of 1,127 Hub candidate tasks, 858 passed validation, including wrong-answer controls that catch weak verifiers.
- **Difficulty filter.** Of 442 scored tasks, the base model solved 151 every time and 120 never; the 171 solved only sometimes formed the useful training set.
- **Training stack.** TRL's AsyncGRPOTrainer, OpenEnv and harbor, run on Hugging Face Jobs, with one LoRA per harness and a "super adapter" made by concatenating three rank-32 adapters into one rank-96 adapter.

Results reported by the authors

- For Qwen3.8-27B with pi on Terminal-Bench 2.1: 39% to 55% pass@1, with token use falling from about 2.9K to 1.5K per task.
- Each harness adapter beats the base model in its own harness on both benchmarks, on SWE-bench Pro (hard subset) and Terminal-Bench 2.1, for accuracy and efficiency.
- The merged super adapter, with thinking off, raised mean pass@1 from 39.5% to 49.0% over six harness–benchmark pairs (71 tasks newly solved against 34 lost, $p < 0.001$).
- Thinking mode, which was not trained, still rose from 55.5% to 63.5% mean pass@1.

Limitations and open questions

- Results are self-reported on a limited set of tasks, one base model and three harnesses.
- The authors list an open question: what the model actually learned to do, since it might exploit harness-specific habits.
- Benchmark gains do not guarantee gains on your private tasks.

How to try it, step by step

1. Pick one harness and one model you already use.
2. Collect or generate tasks with trustworthy verifiers and filter by difficulty.
3. Train a small LoRA with RL inside the harness.
4. Evaluate on held-out tasks and track token cost as well as accuracy.

What it means for teams running open-weight agents

It is a cheaper route than full fine-tuning and keeps the base model reusable. See how to evaluate an LLM application and AI agents explained.

Frequently asked questions

- **Is this fine-tuning the whole model?** No. A LoRA adapter is trained and the base weights stay frozen.
- **Can one adapter serve several harnesses?** The authors merged three into one super adapter that runs unchanged under pi, Hermes Agent and OpenCode.
- **Are the artefacts public?** The post says adapters, GGUFs, RL tasks and a plugin are in a Hugging Face collection.

Related reading

- AI agents explained
- How to evaluate an LLM application
- Running AI workloads on Kubernetes
- AI Practitioner

Sources and currency

Sources were read on 23 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Hugging Face. Introducing harness adapters: adapt your favorite model to your harness to improve performance and cost. 20 November 2026. huggingface.co/blog/harness-adapters-at-scale.

Published by IICTP, an independent technical credential body, on 23 November 2026. The current version of this post is at https://blogs.iictp.org/harness-adapters-rl-lora-agent-harness-fine-tuning-explained. More guides: blogs.iictp.org.