



Earendil was created to craft
software and open
protocols that strengthen
human agency



We are a band of humans dedicated to crafting better ways to wield AI



...by building tools in the open designed to be shaped by the humans who use them



THE PROBLEM

The frontier AI stack [pre training capacity; models; RL data; harnesses; applications] is rapidly converging toward a few walled platforms.



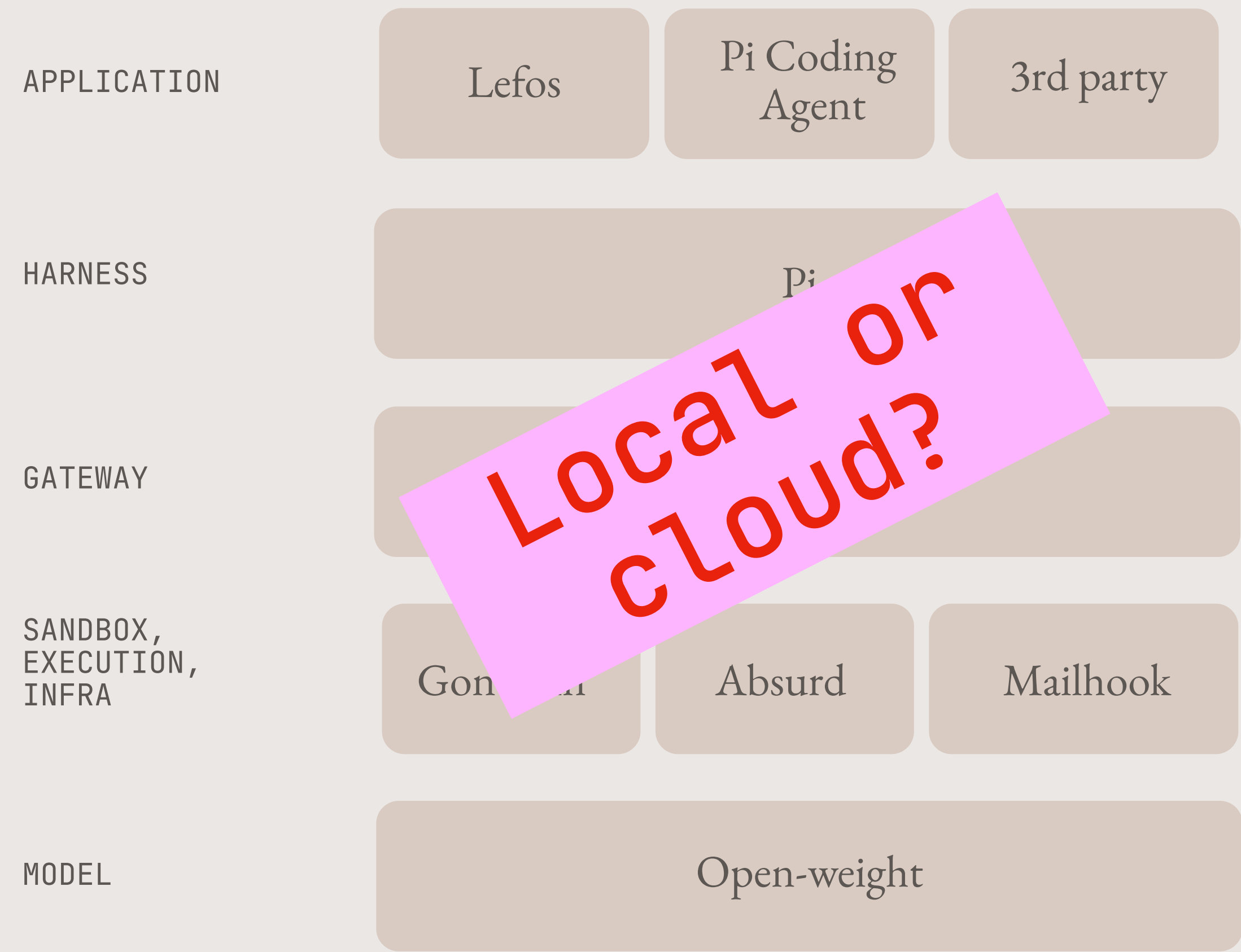
THE OTHER PROBLEM

Those same platforms are actively encouraging humans and builders to forego their agency and let the agents "do the full job".



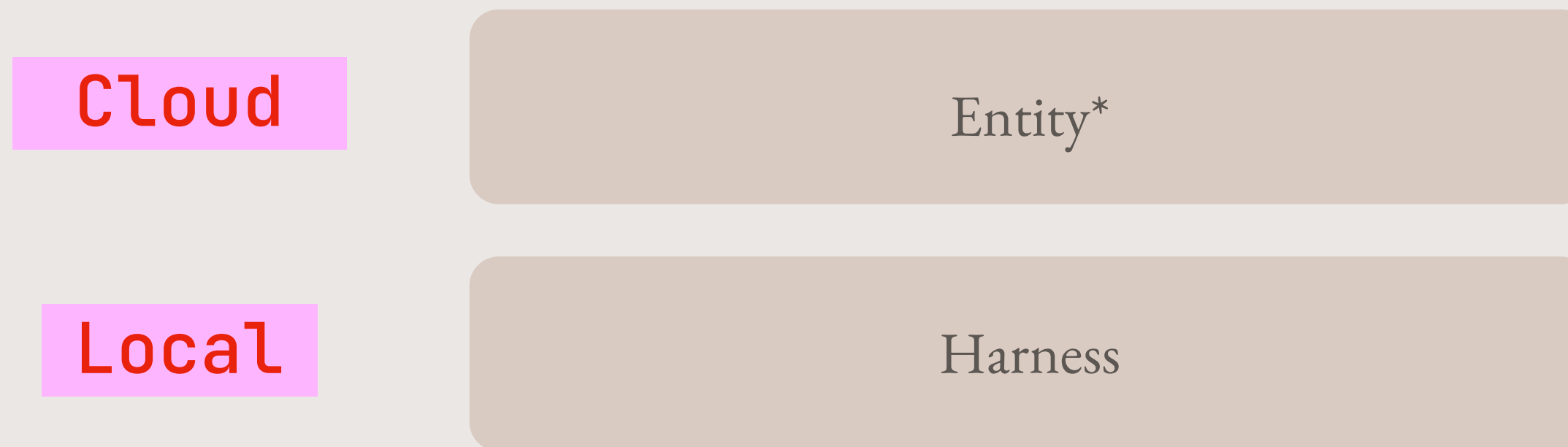
THE SOLUTION

An open source frontier AI stack that is neutral, freely composable, and maximally controlled by the end user





These architectures are already hybrid.
The harness often runs locally, while the entity leverages global state





People want to interact with entities that bring the benefits of local-first, such as context and session ownership, customizability, and privacy



...but they want that entity to feel coherent and capable across interfaces,
and to be able to manage access to relevant context and data stores globally.





Navigating these trade offs in a principled way, without compromising trustworthiness, is non trivial

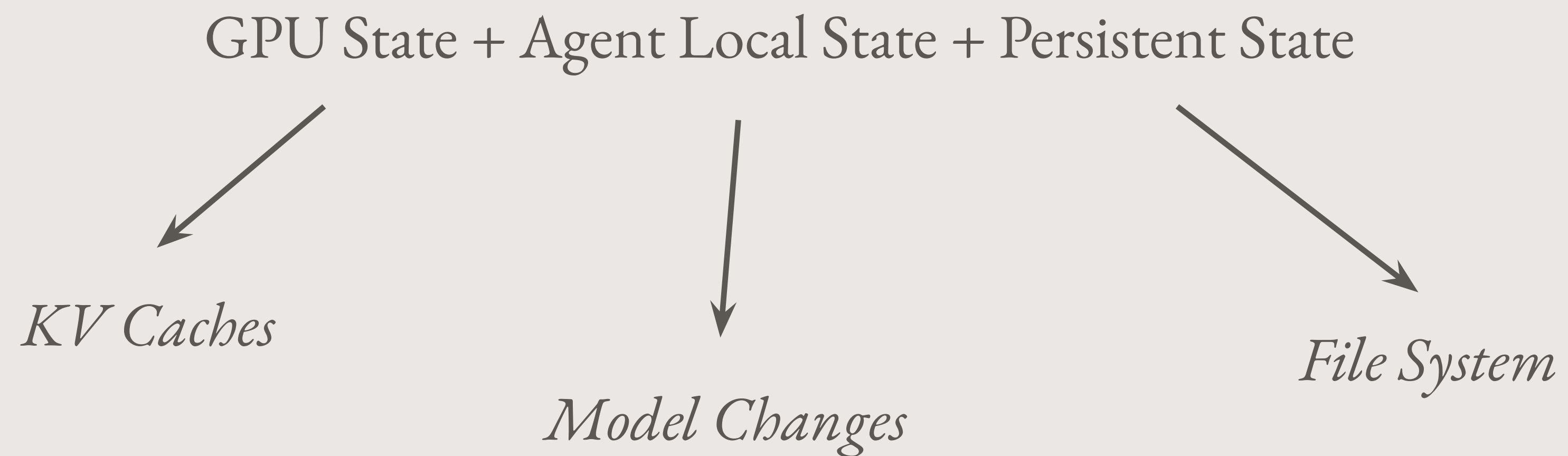


There are many agent harnesses, but this one is yours

pi.dev



Agentic Loops are a weird form of state synchronization





A session / session transcript is an append-only log

```
{
  "type": "session",
  "version": 3,
  "id": "demo-001",
  "timestamp": "2026-07-12T09:00:00.000Z",
  "cwd": "/project"
}
{
  "type": "model_change",
  "id": "state-1",
  "parentId": null,
  "timestamp": "2026-07-12T09:00:00.000Z",
  "provider": "openai",
  "modelId": "gpt-5.3-codex"
}
{
  "type": "message",
  "id": "user-1",
  "parentId": "state-1",
  "timestamp": "2026-07-12T09:00:02.000Z",
  "message": {
    "role": "user",
    "content": [
      {
        "type": "text",
        "text": "Add a dark-mode toggle to Settings."
      }
    ]
  }
}
{
  "type": "message",
  "id": "assistant-1",
  "parentId": "user-1",
  "timestamp": "2026-07-12T09:00:06.000Z",
  "message": {
    "role": "assistant",
    "content": [
      {
        "type": "text",
        "text": "I'll add the toggle and persist the preference."
      }
    ],
    "model": "gpt-5.3-codex",
    "stopReason": "stop"
  }
}
```




The append-only log becomes GPU State + Agent State



I can take a session transcript and take it to a different machine



I should be able to take a session log and bring it to a different model



Session log turned into model specific API calls



```
POST /v1/completions
```

```
  prompt: user: What is the capital of France?
```

```
    └─ hash A → cold start, full compute
```

```
response: The capital of France is Paris.
```

```
POST /v1/completions
```

```
  prompt: user: What is the capital of France?
```

```
        asst: The capital of France is Paris.
```

```
        user: What is its population?
```

```
    └─ hash A → KV cache hit (reuse)
```

```
    └─ hash B → fresh compute
```

```
response: Paris has a population of about 2.1 million.
```



conversation:

system: You are an assistant answering questions of a user

user: What is the capital of France?

asst: The capital of France is Paris.

user: What is its population?

expanded prompt:

<|system|>

You are an assistant answering questions of a user

<|user|>

What is the capital of France?

<|assistant|>

The capital of France is Paris.

<|user|>

What is its population?

<|assistant|>



They increasingly don't like the power of these APIs.

- Hidden server side tokens (*web search results are not revealed*)
- Hidden reasoning (*to fight distilling*)
- Some API encourage server side session logs (*store = true*)
- Assistant Prefill blocked (*for safety*)
- Some session logs are retained (*cybersecurity risks*)



Others increasingly fight agents.

- #vibecoding on lobsters and other luddite communities
- Agentic engineering puts pressure on Open Source
- Fight of Cloud IPs vs Residential Proxies
- Misaligned Incentives for Agent Authentication



How can we break free from the behemoths?



We can start with Open Harnesses,
continue by retaining your Data,
and ultimately support Open+Local Models



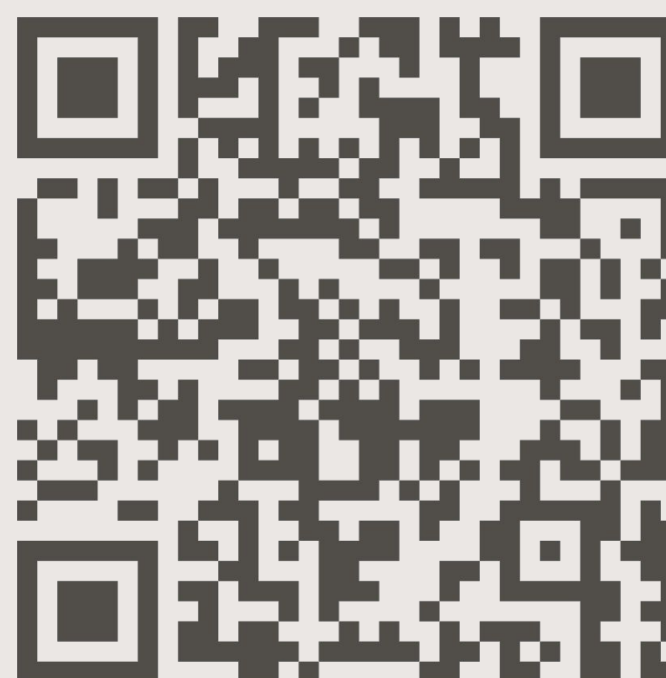
But the experience has to be user friendly.
It cannot be worse than ChatGPT, Codex and Claude Code



Run the harness where you want
Connect it to the World



<https://lucumr.pocoo.org/2025/11/22/llm-apis/>





earendil.com/purpose