Introducing cline-bench: A Real-World, Open Source Benchmark for Agentic Coding

A call for contribution to establish reproducible, practical reinforcement learning environments sourced from real open source development work — with a $1M commitment to support open source maintainers.

Nik Pash • @pashmerepat

November 20, 2025

AI models have advanced significantly, yet the field still lacks a rigorous, open source benchmark that represents real engineering work rather than synthetic, puzzle-oriented, or already-saturated tasks.

OpenAI underscores this gap clearly: “researchers use rigorous frontier evals to measure how well the models perform in different domains,” and “evals make fuzzy goals specific and explicit.”

Unfortunately, most coding benchmarks today resemble LeetCode-style puzzles: self-contained, small programs that don't capture the complexity of real development. We've all seen far too many benchmarks that ask an agent to "write me a server that generates fibonacci sequences from scratch" and winced at how irrelevant they are for day-to-day engineering work.

To support the next stage of AI research and development, we are introducing **cline-bench**, a new initiative focused on creating high fidelity benchmarks and reinforcement learning environments derived from real open source development scenarios.

Introducing cline-bench

Cline-bench is designed to create research-grade environments that capture actual engineering constraints. These include repository starting snapshots, authentic problem definitions, and automated verification criteria. Each selected task will be packaged as a reproducible environment following modern open source specifications such as those used in the Harbor (Terminal-Bench 2.0) framework and Prime Intellect's Environments Hub.

Our intention is to ensure that research on agentic coding can take place under realistic, transparent conditions that reflect the nature of real software development.

To build these environments, we look at real open source work. When you use the Cline Provider on an open source project while opted in to this initiative, we examine tasks where the model requires manual intervention or is unable to complete the work. These challenging, real-world failures become candidates for inclusion as cline-bench environments.

Cline-bench is a collaborative effort. Tasks can enter the benchmark in two ways: through opt-in usage of the Cline Provider on open source projects, and through manual contributions from engineers working in open source

It's important to note that **only open source repositories are eligible for inclusion** because the benchmark is meant to be inspected, reproduced, and studied openly. Private repositories are ineligible.

The goal of cline-bench is not to create superficial rankings but to provide a foundational research primitive that benefits the entire open source AI ecosystem. Real world tasks contain ambiguity, incomplete context, dependency friction, multi step reasoning, and the need for iterative problem solving. These conditions cannot be recreated reliably through synthetic data. Our belief is that open science requires access to resources that reflect true engineering work so that progress is measurable, replicable, and communal.

The importance of practical and open evaluation has been recognized by others in the open source AI research community. The following statements were shared with permission.

> *"Cline-bench is a great example of how open, real-world benchmarks can move the whole ecosystem forward. High-quality, verified coding tasks grounded in actual developer workflows are exactly what we need to meaningfully measure frontier models, uncover failure modes, and* ***push the state of the art****.*
>
> *We're excited about the open call for contributions and the use of shared standards like Harbor, which make it easier for the community to compare, improve, and ultimately build more capable coding agents together."*

– Shyamal Anadkat, Head of Applied Evals @ OpenAI

> *"Nous Research is focused on training and proliferating models that excel at* ***real world tasks****. cline-bench will be an integral tool in our efforts to maximize the performance and understand the capabilities of our models."*

> *benchmark available to all will help us continue to **push the frontier** coding capabilities of our LLMs."*

– *Baptiste Rozière*, Research Scientist @ *Mistral AI*

> *"We're huge fans of everything Cline has been doing to empower the open source AI ecosystem, and are incredibly excited to support the cline-bench release. High-quality open environments for agentic coding are exceedingly rare. This release will go a long way both as an evaluation of capabilities and as a post-training testbed for challenging real-world tasks, **advancing our collective understanding** and capabilities around autonomous software development."*

– *Will Brown*, Research Lead @ *Prime Intellect*

What is cline-bench for?

Put simply, cline-bench is a way to test and compare LLMs on real engineering problems instead of artificial examples. By grounding evaluation in real-world cline tasks, it becomes possible to measure capability in a way that actually reflects day-to-day software development.

Moreover, each accepted task becomes a reproducible reinforcement learning environment that can be executed, scored, and compared across different models and agentic strategies.

Engineers using models for day-to-day coding tasks, researchers, and applied AI engineers will be able to evaluate how different models perform on the same real engineering problems and measure progress over time. You can also then directly train your own models on these RL environments.

Cline-bench is primarily intended to serve three purposes:

engineering tasks rather than puzzles or synthetic benchmarks, allowing researchers and developers to assess real-world capability.

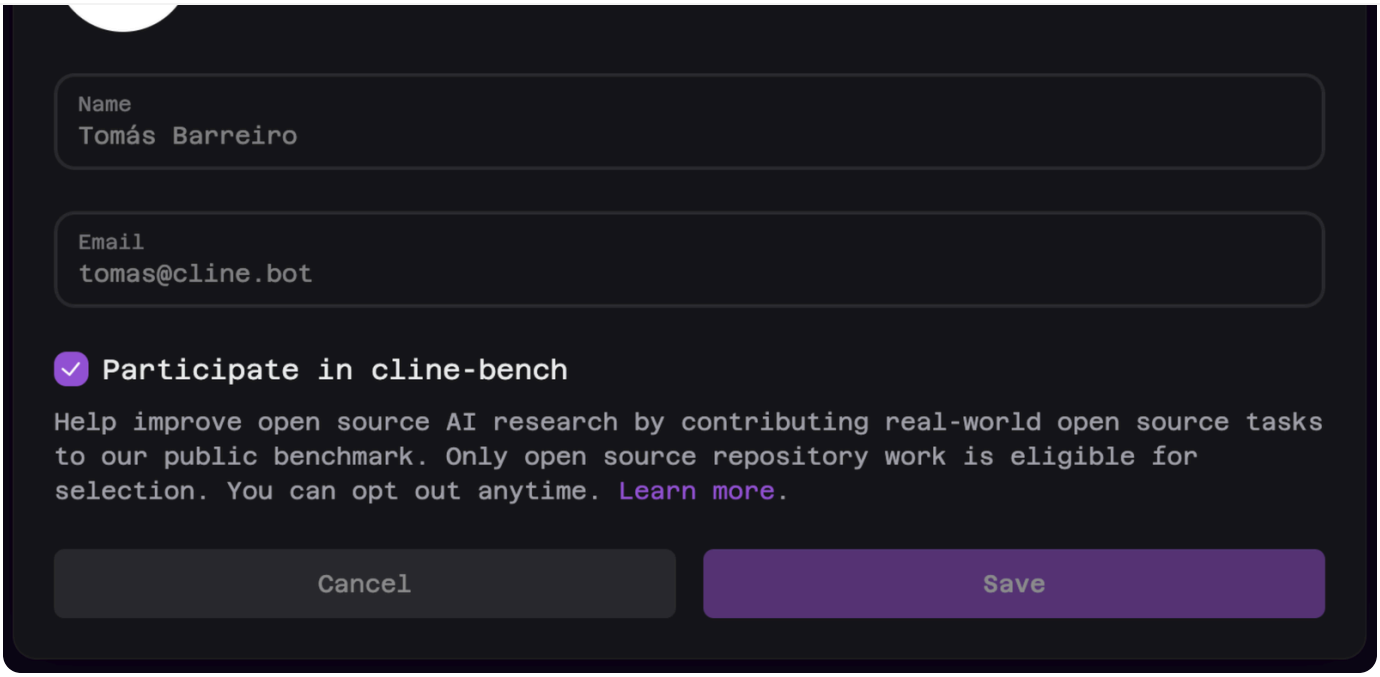
2. ***Open scientific progress***. By standardizing and publishing these environments, the broader research community can study failure modes, identify capability gaps, and share techniques to improve agentic coding performance.
3. ***Training data for downstream fine-tuning and RL research***. Because each task includes a clear initial state, a starting prompt, and a verifiable end state, it can serve as a catalyst for supervised fine-tuning, reinforcement learning, or hybrid approaches.

In short, cline-bench provides the missing research infrastructure needed to measure meaningful progress in agentic coding and to develop models that perform better in real engineering settings.

Our goal is simple and practical: **we want coding agents to actually work**, and this is the path to making them genuinely reliable in daily development.

Privacy, security, and control

Users always retain full control of how they interact with Cline. Participation in cline-bench is optional and can be changed at any time on the Cline Provider dashboard.



https://app.cline.bot/dashboard/account

As always, you can bring your own API keys, use third party model providers, or self host your own models and use them in Cline. You have full control over your privacy posture, infrastructure, and security boundaries at all times.

Teams and Enterprise customers are already isolated by default and are not included in this initiative. Cline's zero trust architecture ensures your enterprise data stays secure and inside your network. Your usage and data are excluded from cline-bench.

Over the coming weeks we will publish contribution guidelines, environment structure documentation, and an early tranche of open source cline-bench tasks that demonstrate how cline-bench is built and validated.

Our goal is to work transparently, publicly, and alongside the open source community that has supported Cline from the beginning.

A Call for Contribution

frontier model fails on a task you're working on, that failure defines the cutting edge of what's possible today. Your real work is what matters. The problems that make you manually intervene, the tasks where even the best models struggle: these are the exact challenges that will train the next generation of AI systems.

I would like to personally invite engineers who believe in open source AI research and open scientific progress to participate. By simply using the Cline Provider while opted in, you are directly contributing and helping build a shared research resource that can benefit the entire open source community. If you regularly work on difficult real world problems in open source repositories, including commercial open source projects, your contributions are especially valuable.

In the spirit of Open Science, you will be directly attributed if your task gets selected to be part of cline-bench. You can also request to have your attribution removed at any time.

In practice, published open source cline-bench tasks will have the following data:

1. A starting snapshot (git commit hash of an open source repo in which you started working on a real world engineering task)
2. Your starting prompt (may be modified slightly to ensure evaluation fairness and removal of sensitive information)
3. Tests based on the ground truth end state - the code you actually committed at the end

Only the most challenging real-world engineering tasks will be accepted. A task that frontier LLMs struggle to complete are ideal candidates for cline-bench. If you find yourself having to intervene, or write code manually because cline was

Please consider using the Cline Provider while opted in so that your work can help shape the future of agentic coding research. If you are interested in contributing beyond using the Cline provider to work on challenging engineering tasks, please reach out in the contributor channel in our Discord.

Cline's Commitment: $1M to Support Open Source Builders

As we open cline-bench to the community, we also want to support the developers who make rigorous, real-world evaluation possible in the first place.

Open source maintainers shoulder the majority of modern software infrastructure, and the hardest, most valuable tasks in cline-bench will come from their day-to-day engineering work. We believe it's important to give back.

To accelerate open source agentic coding research, we are launching a $1M sponsorship program to support developers who contribute real-world tasks to cline-bench.

Selected contributors will receive Cline Open Source Builder Credits, designed to support your workflow while helping us build a richer, more representative benchmark for the community.

If you're an active open source contributor, you can apply for Cline Builder Credits here.

> *Cline-bench will always remain fully open source and freely accessible.*

Thank you for reading.

– Nik

Cline

Related Posts

Cline 3.38.0: Gemini 3 Pro Preview & Voice Dictations that understands your code

November 17, 2025

Cline: The Fastest Growing AI Open Source Project on GitHub in 2025, Thanks to You

November 4, 2025

Cline

Transform your engineering team with a fully collaborative AI partner. Open source, fully extensible, and built to amplify developer impact.

Stay updated on Cline's evolution

Email address

Subscribe

Product

- Docs
- Blog
- Enterprise
- MCP Marketplace
- Changelog

Community

- Discord
- Reddit
- GitHub Discussions

Support

- GitHub Issues
- Feature Requests
- Contact

Company

- Careers
- Brand
- Terms
- Privacy

© 2025 Cline Bot Inc. All rights reserved.