

Container viability of open-source AI agents

Generated from run records. Every figure below is a count of rows.

Abstract

We attempt to build 27 open-source AI agent repositories into containers at pinned commits and to start each resulting image in a network-denied sandbox. 22 build successfully and 22 pass a container environment check, a rate of 81.5 percent. We report the distribution of failure causes and, for each, whether the cause is a property of the agent or a limit of the harness. We do not measure whether any agent runs.

Method

Each repository is resolved to a commit SHA, cloned inside a container build at that pin, and installed using a single generic recipe. The resulting image is started with egress denied and runs an environment probe: the interpreter starts, prints one line, and exits.

The recipe starts from the `#{BASE}` image, clones at the pinned commit, and installs the first package definition it finds at the repository root or one or two directories below it, refusing to choose when several qualify. A build is abandoned after 30 minutes and the image identity is read within 60 seconds of the build completing. The environment probe runs with the network disabled.

A build that exits successfully is checked for having installed a distribution: a workspace manifest that `setuptools` installs under the name `UNKNOWN` exits zero and installs nothing, and was recorded as a success until the check existed.

The probe does not invoke the agent. The container command overrides whatever entrypoint the agent defines, so a passing check establishes that the environment works and nothing more. The repositories studied share no common entrypoint, and with no network and no credentials an invocation would measure the harness rather than the agent.

Records span 2026-09-26 to 2026-09-28. Build recipes used: 5. A recipe digest is pinned on every build row, since the recipe determines which agents are buildable and rows produced under different recipes are different measurements.

Results



Stage	Count
Repositories harvested	27
Seed URLs unresolved	0
Built	22
Build failed	9
Environment check passed	22

Stage	Count
Environment check failed	0

Failure counts are tallies of recorded rows and are not a partition of the repositories harvested. A repository that fails to build contributes a build row and no check row, so the tables below neither sum to the harvested count nor are bounded by it.

Build stage

Cause	Count	Attributable to
Written in a language this recipe does not build	2	this harness
The manifest found is a workspace root, not an installable package	2	this harness
Dependencies could not be resolved	2	the repository
Build failed	2	the repository
Requires a Python version the base image does not provide	1	this harness

Environment check stage

No environment check failures recorded.

Probe durations range from 208 to 723 milliseconds.

Threats to validity

The headline rate is not a property of the agents alone. The attribution column above separates causes that are facts about a repository from causes that are limits of this harness. A reader comparing this rate against another study must check that the harness limits are comparable, because extending the recipe changes the rate without any agent changing.

Failure classification is substring based. Unmatched failures are recorded as unclassified rather than assigned to a nearest class. The unclassified bucket also holds failures of the harness itself, including a build exceeding its timeout and the container runtime failing to start, because filing either under a build cause would attribute it to the agent.

Selection. The repositories studied were chosen by hand as agents people run together, not sampled from a ranking. The rate therefore describes that population and does not generalise to all repositories matching an agent keyword.

Limitations

Generated from run metadata rather than composed, so these cannot fall out of date with the results above.

- Sample size is 27 repositories.
- Models held constant across checks: 1.
- Build recipes in the record: 5. A rate computed across more than one recipe is not a single measurement.
- No agent was invoked, so nothing here is evidence of agent behaviour.

- No pair of agents was run, so nothing here concerns interaction.

Classification

Failure classes are assigned by matching the build log against a fixed, ordered list of needles, most specific first, so that a generic wrapper error cannot claim a failure that a more specific message explains. Two needles present in an early version were removed because they matched text that BuildKit prints in every failed build. Each class carries an attribution: whether it describes the repository or a limit of this harness.

- **clone_failed**: Repository could not be cloned. Attributable to the repository.
- **no_installable_manifest**: No installable manifest at the repository root. Attributable to the repository.
- **ambiguous_manifest**: Several installable subprojects, none of them obviously the agent. Attributable to this harness.
- **python_version_unsupported**: Requires a Python version the base image does not provide. Attributable to this harness.
- **workspace_root_not_installable**: The manifest found is a workspace root, not an installable package. Attributable to this harness.
- **unsupported_language**: Written in a language this recipe does not build. Attributable to this harness.
- **no_build_recipe**: No usable build recipe. Attributable to this harness.
- **dependency_conflict**: Dependencies could not be resolved. Attributable to the repository.
- **host_out_of_disk**: The machine running the build ran out of disk. Attributable to this harness.
- **host_out_of_memory**: The machine running the build ran out of memory. Attributable to this harness.
- **build_failed**: Build failed. Attributable to the repository.
- **env_check_timeout**: Container environment probe did not finish in time. Attributable to the repository.
- **env_check_failed**: Container environment probe exited non-zero. Attributable to the repository.
- **needs_credentials**: Requires credentials we do not supply. Attributable to the repository.
- **unknown**: Unclassified failure. Attributable to this harness.

Appendix A: every agent

Agent	Repository	Commit	Build	Cause	Environment check	Probe (ms)
aider-ai-aider	Aider-AI/aider	5dc9490bb35f	built		passed	214
gptme-gptme	gptme/gptme	9e3ab7cf5418	built		passed	246
continuedev-continue	continuedev/continue	5522c6f44ca0	failed	unsupported_language	not reached	
anomalyco-opencode	anomalyco/opencode	a42f393c850b	failed	unsupported_language	not reached	
openai-swarm	openai/swarm	6af0b4caf37d	built		passed	235
microsoft-autogen	microsoft/autogen	027ecf0a379b	failed	workspace_root_not_installable	not reached	

Agent	Repository	Commit	Built	Cause	Environment check	Probe (ms)
crewaiinc-crewai	crewAIInc/crewAI	4ed2abc7bbf5	failed	workspace_root_not_installable	not reached	
langchain-ai-langgraph	langchain-ai/langgraph	7daa3ab49d67	built		passed	208
yoheinakajima-babyagi	yoheinakajima/babyagi	fa8930ebe72a	built		passed	255
browser-use-browser-use	browser-use/browser-use	4cbe921673b4	built		passed	267
ther1d-shell-gpt	TheR1D/shell_gpt	a082bd5327ce	built		passed	209
huggingface-smolagents	huggingface/smolagents	227ef5e49dd	built		passed	212
foundationagents-metagpt	FoundationAgents/MetaGPT	11cdf466d042	failed	dependency_conflict	not reached	
swe-agent-swe-agent	SWE-agent/SWE-agent	3ea751c087f3	built		passed	221
smol-ai-developer	smol-ai/developer	a6747d1a6ccd	failed	build_failed	not reached	
openbmb-chatdev	OpenBMB/ChatDev	4fb2db0ea903	failed	build_failed	not reached	
pythagora-io-gpt-pilot	Pythagora-io/gpt-pilot	9b763fdaf002	built		passed	213
openinterpreter-open-interpreter	OpenInterpreter/open-interpreter	13061d2ce0ac	built		passed	253
antonosika-gpt-engineer	AntonOsika/gpt-engineer	a90fcd543eed	built		passed	244
simonw-llm	simonw/llm	764dc386c58b	built		passed	222
swe-agent-mini-swe-agent	SWE-agent/mini-swe-agent	04d809ceab9d	built		passed	222
foundationagents-openmanus	FoundationAgents/OpenManus	3309bf4e416f	failed	python_version_unsupported	not reached	
openhands-openhands	OpenHands/OpenHands	7fbb48c40679	failed	dependency_conflict	not reached	
significant-gravitas-autogpt	Significant-Gravitas/AutoGPT	5e84f064d377	built		passed	290
significant-gravitas-autogpt	Significant-Gravitas/AutoGPT	c035ab1f91b8	built		passed	290
qwenlm-qwen-code	QwenLM/qwen-code	3f5ae3ffeb72	built		passed	309
qwenlm-qwen-code	QwenLM/qwen-code	0a136f89d7f9	built		passed	309

Reproduction

Every figure resolves to a row in the published dataset, and every recorded failure to its full build log, addressed by content hash. The dataset, its schema, and the recipe digest pinned on every build row are published beside this document. Running `crucible run` against the same seed list at the same commits under the same recipe reproduces the rows; a different recipe is a different measurement and is recorded as such.