

runEvaluation

September 4, 2026

runEvaluation

Kind: Function

Source: `atloria-monorepo/libs/agent-core/src/benchmark/swebench-evaluator.ts`

Run the SWE-bench Docker evaluation harness.

This executes the official harness which:

1. Pulls pre-built Docker images per task (correct Python, conda, deps)
2. Applies the agent's patch to a clean repo checkout
3. Runs the gold test patch and tests
4. Reports pass/fail per task

Signature

```
function runEvaluation(options: EvalCommandOptions, log: (msg: string) => void): EvalResult[]
```

Parameters

Name	Type
options	EvalCommandOptions
log	(msg: string) => void

Returns: `EvalResult[]`