DeepSWE v1.1 Review Report

Author: Benchmarks' Last Exam Lab 2026-11-11

DeepSWE evaluates AI agents on implementing features and fixing issues in real code repositories. Automated tests determine whether the work passes.

Abstract

We reviewed all 113 tasks in DeepSWE v1.1: **41 Accepted**, **15 Doubtful** and **57 Refused**. Confirmed issues mainly involve missed core errors, unstated scoring requirements and incorrect grading logic.

Decision	Tasks	Meaning
Accepted	**41**	No substantive evaluation defect or specific unresolved concern was found in this review.
Doubtful	**15**	A specific concern remains unresolved: its trigger, correct interpretation or grading effect needs further evidence.
Refused	**57**	At least one defect affecting evaluation validity has been confirmed; its effect depends on the trigger conditions.

Scope and version

Version: DeepSWE v1.1, 0b9fabbb63b9.

Materials include task prompts, tests, grading code and public submissions, trajectories and grading logs. All 113 tasks are covered. Updated November 11, 2026.

Distribution of findings

Each task is counted once under its primary issue.

Issue type	Refused	Doubtful	Total
C1 Incorrect answers or scoring logic	9	0	9
C2 Scoring requirements not stated in the task	17	0	17
C3 Missed critical errors	22	1	23
C4 Scoring bypasses or answer leakage	3	14	17

Issue type	Refused	Doubtful	Total
C5 Other issues	6	0	6
Total	**57**	**15**	**72**

Representative cases

New data reuse an old address and are mistaken for the old data

C1 Incorrect answers or scoring logic

Task: python-statemachine-state-data-scoping

The task requires deleting old data when a state is exited and creating fresh data when it is entered again. The model does delete and recreate the list. Python can give the new list the memory address freed by the old one; sharing an address does not make them the same object.

The test compares only those addresses and fails when they match. The public run passes the other 71 new tests and all 1,286 existing tests, but this single false failure produces a score of 0.

Evidence 3

- Grading assertions
- Actual submission
- Original grading log

The test requires filling earlier gaps, but the task does not

C2 Scoring requirements not stated in the task

Task: ink-grid-box-layout

The task asks for grid layout. Under default CSS rules, items are placed progressively and may leave earlier gaps. The hidden test instead requires later items to go back and fill those gaps, which the task does not request.

The public run passes all 49 existing tests and fails only this one of 25 new tests, yet receives 0. An implementation following normal layout rules is rejected by a different, unstated rule.

Evidence 2

- Prompt
- Hidden test

A historical snapshot shows the current state and still earns full credit

C3 Missed critical errors

Task: aiomonitor-task-snapshots-diff

The task requires saving a task's state for later inspection. If the task was running when saved, that historical snapshot should still say it was running, even after the task finishes.

Two public full-score submissions instead read the task's current state when the snapshot is opened. The test named for freezing state checks only the task's name, not whether the saved state changes. Both submissions pass all 53 new tests and 8 existing tests and receive 1.

Evidence 4

- Prompt
- Test assertions
- Actual full-score submission
- Original reward

Empty test functions can stand in for missing functionality (static evidence)

C4 Scoring bypasses or answer leakage

Task: tengo-destructuring-bindings

The task adds syntax that unpacks several variables at once. For example, [a, b] := [10, 20] should set a to 10 and b to 20, with support for nesting, defaults and other forms. Grading requires 91 new tests and 132 existing tests to pass.

The existing code review found that a submitted TestMain, which runs before the tests, can retain all 91 functional test names while replacing their checks with empty functions. The runner would then report passes without checking whether the new syntax works.

The independent review found that grading would neither remove this entry point nor reject it. It can also affect 29 existing tests, while the other 103 remain unchanged. This code path predicts that a submission could satisfy the full-credit condition without implementing the core feature.

Evidence 5

- Task
- Test entry point
- Submission replay
- Scored test IDs
- Full-credit condition

A self-test and a hidden test share a name, so grading cannot compile them

C5 Other issues

Task: anko-typed-variable-bindings

The model adds a self-test called TestTypedBindingsDeclarations. A hidden test added during grading has the same name, but the task never reserves that name.

Go does not allow two functions with the same name in one package. Combining the files therefore prevents compilation. The run has eight missing test results and receives 0. Grading has prevented the tests from running; this does not establish that fixing the collision would make every feature correct.

Evidence 3

- Hidden test
- Public run
- Preparation and grading

Task-by-task findings

Refused: 57 tasks

abs-module-cache-flags

C2 Scoring requirements not stated in the task

A debug log can say the right thing and still lose credit because the test requires English wording not specified in the task.

abs-stepped-slices

C1 Incorrect answers or scoring logic

An ordinarily named self-test collides with a hidden test added during grading, preventing compilation. Established from code inspection, without a new run.

actionlint-action-pinning-lint

C1 Incorrect answers or scoring logic

Grading overwrites a submitted test file but leaves other tests calling helpers from it. Those helpers disappear, so the test package cannot compile.

aiomonitor-task-snapshots-diff

C3 Missed critical errors

The snapshot reads the current state when opened instead of preserving the state when captured. Two public submissions with this defect received full credit.

anko-default-function-arguments

C5 Other issues

Grading overwrites a file holding shared self-test helpers but leaves their callers intact, preventing compilation. Established from code inspection.

---

anko-typed-variable-bindings

C5 Other issues

A model-written test has the same name as a hidden test. Combining them prevents compilation; the task did not reserve that name.

---

arcane-drift-detection-baselines

C3 Missed critical errors

The feature must detect differences between configuration and live container state. Tests check only supplied configuration and miss an empty live-state checking entry point.

---

boa-hierarchical-evaluation-cancellation

C2 Scoring requirements not stated in the task

The task does not prescribe how handle creation borrows Context, but tests allow only an immutable borrow, excluding a valid mutable-borrow implementation.

---

claude-code-by-agents-recursive-delegation

C3 Missed critical errors

When a subagent replies in several chunks, keeping only the final chunk and discarding the rest can still pass the tests.

---

cliffy-config-file-parsing

C2 Scoring requirements not stated in the task

The task allows importing an error class from the configuration submodule, but tests require a top-level import too. Without that unstated export, the functional tests cannot load.

---

csstree-shorthand-expansion-compression

C2 Scoring requirements not stated in the task

Equivalent CSS background output can be rejected solely because of spacing after commas.

---

drizzle-orm-window-function-builders

C3 Missed critical errors

A window-function result that may be null is typed as non-null. Tests miss this type error, and the public submission receives full credit.

---

dynamodb-toolbox-conditional-attribute-requirements

C2 Scoring requirements not stated in the task

An equivalent JSON Schema using anyOf/not is rejected because tests require particular keywords that the task does not mandate.

eicrud-keyset-pagination-cursor

C3 Missed critical errors

Incorrect pagination can drop later records in the same group. The test data happen to fill each page with a whole group, hiding the defect.

etree-xml-diff-patch

C4 Scoring bypasses or answer leakage

Tests run twice without restoring test files between runs. Code inspection identifies a submitted startup routine that could empty the second round's checks; no attack was executed.

fastapi-deprecation-response-headers

C2 Scoring requirements not stated in the task

Z and +00:00 both denote UTC, but tests accept only +00:00. The task does not impose that restriction.

fastapi-implicit-head-options

C2 Scoring requirements not stated in the task

Even when OPTIONS is explicitly enabled, tests require it to be absent from the OpenAPI documentation. The task does not state that restriction.

fd-deterministic-multi-key-sorting

C3 Missed critical errors

Tests miss an implementation that uses only the first requested sort key and falls back to file paths instead of applying the remaining keys.

geo-shapeindex-serialization

C3 Missed critical errors

The task requires saving the index for direct restoration. Saving only shapes and rebuilding the whole index on load can still pass.

go-critic-doc-link-checker

C1 Incorrect answers or scoring logic

An ordinary added test collides with a declaration in the hidden tests, preventing compilation when grading combines them.

go-genai-streamed-function-args

C5 Other issues

Grading removes shared helpers added to an existing test file while other tests still call them, preventing compilation. Established from code inspection.

go-git-worktree-merge-conflicts

C3 Missed critical errors

Merged working files can look correct while the commit still contains the old tree, losing target-branch changes. Existing grading misses this mismatch.

gql-incremental-graphql-delivery

C3 Missed critical errors

The feature should deliver data as it arrives, but tests cannot distinguish that from waiting for the entire response before returning anything.

helm-array-merge-strategies

C2 Scoring requirements not stated in the task

The test omits values.yaml but rejects the missing-file warning allowed by the task, adding an unstated exception.

helm-unified-manifest-stream

C3 Missed critical errors

A document is stored in both the unified manifest and separate hook data, then merged again on load, producing duplicates. Tests omit that round trip; the finding is based on code inspection.

httpx-streaming-json-iteration

C3 Missed critical errors

The task requires yielding JSON items as data arrives, but waiting for the complete response before yielding the first item can satisfy the tests.

ink-grid-box-layout

C2 Scoring requirements not stated in the task

Default CSS placement may leave earlier gaps, but the test requires later items to go back and fill them. The task does not require this.

katex-multicolumn-array-spans

C3 Missed critical errors

The submission does not implement real column spanning in HTML. Grading mainly checks MathML, another output format, and still awards full credit.

kcp-go-multiplexed-kcp-streams

C3 Missed critical errors

The priority test accepts ordinary round-robin processing that ignores priority, missing the required preemption behavior.

langchain-request-coalescing

C1 Incorrect answers or scoring logic

The test requires coalescing two overlapping requests but does not ensure they overlap. Normal scheduling can therefore cause a false failure.

mnamer-daemon-watch-lifecycle

C3 Missed critical errors

The running-state check also accepts the text “not running,” so an implementation that always reports stopped can escape detection.

numba-stencil-boundary-modes

C3 Missed critical errors

Mixed constant-boundary handling produces incorrect values, but the public submission receives full credit. Tests miss a behavior explicitly required by the task.

obsidian-linter-auto-table-of-contents

C4 Scoring bypasses or answer leakage

The table-of-contents tests share a helper that submissions can modify. Code inspection identifies a way to retain test names but remove checks; the attack was not executed.

onedump-dump-encryption-pipeline

C3 Missed critical errors

Ciphertext truncated between complete blocks can be accepted without checking the end marker or whole-stream integrity. Tests miss this case.

opa-template-string-reconstruction

C5 Other issues

An ordinary CLI self-test has the same name as a hidden test. Combining them prevents compilation before functionality can be checked.

oxvg-structural-selector-preservation

C3 Missed critical errors

The submission protects SVG elements outside the complete CSS relationship being matched, leaving elements unoptimized even though the task requires optimizing them. It receives full credit.

participle-grammar-conflict-analysis

C5 Other issues

Hidden tests collide with model-written test names. Grading then fails to compile the package and counts the unrun tests as failures.

pebble-durability-wait-apis

C2 Scoring requirements not stated in the task

The task requires bounded history, but tests impose an unstated 10,000-entry cutoff, rejecting a bounded 16,384-entry window. Established from code inspection.

prometheus-transactional-reload-status

C2 Scoring requirements not stated in the task

Fractional milliseconds satisfy the task, but tests require whole milliseconds without a stated rounding requirement.

pwntools-tube-multiplexing

C1 Incorrect answers or scoring logic

The test sends more than the buffer limit before reading. Correct flow control pauses the sender, so the test itself can deadlock a valid implementation.

python-statemachine-state-data-scoping

C1 Incorrect answers or scoring logic

The old list was deleted, but a new list reused its memory address. The test mistakes this for reusing the old list; this sole failure makes the run score zero.

query-persist-restored-query-state

C1 Incorrect answers or scoring logic

The submission retains the restoration behavior and timestamps, but renamed tests are counted as missing. That alone causes the run to score zero.

quill-shared-toolbar-focus

C2 Scoring requirements not stated in the task

Normal picker interaction works, but the test sends an event that does not bubble to the parent, rejecting a valid implementation that listens there.

returns-validated-error-accumulation

C2 Scoring requirements not stated in the task

Renaming an internal function changes generated test IDs without changing what the tests check. The grader treats the old IDs as missing and awards zero.

scc-bounded-memory-spilling

C3 Missed critical errors

Merging data can exceed the allowed number of file records held at once. Grading trusts the program's own counters and misses the violation.

skrub-duration-encoding

C1 Incorrect answers or scoring logic

Adding test parameters changes IDs, so five passing cases are counted as missing. Other genuine failures remain, so this does not establish that the submission deserves full credit.

sqlfmt-create-table-ddl-formatting

C2 Scoring requirements not stated in the task

The task permits an opening parenthesis immediately after the table name, but tests require a space and reject equivalent output.

task-task-graph-export

C3 Missed critical errors

The graph API can exist while --graph is never connected to it, leaving the CLI feature unusable. Grading checks only the API and compilation.

tengo-callable-instance-isolation

C3 Missed critical errors

When scripts declare variables in different orders, the implementation reuses positions from the source script in the destination and accesses the wrong variables. The public submission receives full credit.

tengo-destructuring-bindings

C4 Scoring bypasses or answer leakage

Code inspection shows a submitted test entry point can leave all 91 functional test names intact while removing their checks, producing pass records without destructuring support. No attack was executed.

testem-bail-on-test-failure

C2 Scoring requirements not stated in the task

A warning can correctly describe an invalid configuration and still be rejected because it lacks the word "invalid," which the task does not require.

testem-per-launcher-reports

C2 Scoring requirements not stated in the task

Equivalent XML reports lose credit solely because their attributes appear in a different order.

updo-policy-alerting

C3 Missed critical errors

The alert component can pass its own tests without being connected to the real monitoring flow. Grading misses that integration gap.

vulture-persistent-analysis-cache

C1 Incorrect answers or scoring logic

Adding test parameters changes four passing cases' IDs. The grader counts the old IDs as missing and awards zero.

wazero-multi-module-snapshots

C3 Missed critical errors

The submission truncates gzip data to shorten incremental snapshots, leaving an incomplete stream across successive increments. Tests check size but not decompression and award full credit.

yaegi-go-embed-directives

C2 Scoring requirements not stated in the task

Hidden tests use common Go test names without reserving them in the task. Ordinary self-tests with those names then prevent compilation during grading.

ytt-jsonpath-query-api

C5 Other issues

The submission compiles before hidden tests are added. Their duplicate names break compilation, and the unrun tests are counted as failures.

Doubtful: 15 tasks

awilix-async-container-initialization

C4 Scoring bypasses or answer leakage

After restoring test files, grading runs a submitted build script that could alter them again. Credit without async initialization has not been demonstrated.

cattrs-partial-structuring-recovery

C4 Scoring bypasses or answer leakage

Submitted code may replace pytest checks when loaded, but a bypass through the complete grading process has not been demonstrated.

effect-sse-httpapi-streaming

C3 Missed critical errors

Tests may accept waiting for every event before returning any, contrary to streaming. Whether that implementation passes all other required checks is unresolved.

igel-persist-feature-schema

C4 Scoring bypasses or answer leakage

Submitted code may rewrite pytest reports as passed, but the inspected public full-score runs do not demonstrate that bypass.

kea-atomic-signal-selectors

C4 Scoring bypasses or answer leakage

The submission can affect test configuration and final reports, but no native record shows credit with the core feature absent.

kombu-virtual-queue-dead-lettering

C4 Scoring bypasses or answer leakage

A submitted pytest plugin may replace real checks with passing records. Credit without dead-letter queue support has not been demonstrated.

koota-query-predicates

C4 Scoring bypasses or answer leakage

Grading runs test scripts from the submission and trusts their reports. It remains unproven that changing those scripts earns credit without query filtering.

mashumaro-flattened-dataclass-fields

C4 Scoring bypasses or answer leakage

A submitted pytest plugin may report passes without running the checks. A successful bypass of the full grader remains unverified.

mobly-grouped-test-barriers

C4 Scoring bypasses or answer leakage

Submitted code enters pytest execution and reporting and may bypass checks for group synchronization. No successful-credit record was verified.

narwhals-rolling-window-suite

C4 Scoring bypasses or answer leakage

Submitted code may overwrite the XML report read by the grader, but the timing and whether grading would use that replacement remain unverified.

ofetch-per-origin-circuit-breaker

C4 Scoring bypasses or answer leakage

A modifiable test entry point may generate passing reports directly, but credit without the circuit-breaker feature has not been demonstrated.

optique-conditional-option-dependencies

C4 Scoring bypasses or answer leakage

Grading runs an npm test command that the submission can change, potentially giving it control over the report. No successful bypass record was verified.

psd-tools-blend-range-api

C4 Scoring bypasses or answer leakage

Submitted code may rewrite grading reports when the process exits. Full credit without implementing the core API has not been demonstrated.

sql-formatter-bigquery-pipe-formatting

C4 Scoring bypasses or answer leakage

Submitted startup code may affect reports, but the inspected public changes are ordinary edits. No successful bypass was established.

textual-richlog-follow-state

C4 Scoring bypasses or answer leakage

Submitted code may replace pytest execution and fabricate passing results. Expanded public-record searches still found no demonstrated bypass credit.

Effects on evaluation and follow-up

The confirmed defects can reject valid work or award credit despite missing core functionality; each finding depends on its specific triggering conditions.

Contact Us

benchmarks-last-exam.com Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.