

Task Validity

Tool

- T.1. Versions of all tools (e.g., Python) are clearly specified.
- T.2. Required API tools are consistently accessible during evaluation.
- T.3. Evaluation process terminates or handles errors appropriately if an API becomes inaccessible.

Env.

- T.4. Residual data or state are fully cleared between runs.
- T.5. Agent is completely isolated from any ground truth information.
- T.6. Setup does not change over time (e.g., no live website).

Implementation

- T.7. Annotated ground truth is verified for correctness.
- T.8. Each task is verified to be solvable.
- T.9. Benchmark includes an Oracle solver that can automatically solve all challenges.
- T.10. Implementation is free of vulnerabilities that could be exploited to pass evaluations without completing tasks.