Level 3

Novel
Scientific
Contribution

SOTA Achievement

Level 1
Baseline Improvement

Level 0 Reproduction

Capability Levels for AI Research Agents

MLGym (Nathani et al.)

Level 3
Full Paper
Reproduction

Core-Bench* (Siegel et al.)

Level 2
Experiment Reproduction

PaperBench (Starace et al.)

Level 1
Function Reproduction
(this work)

Benchmarks by Reproduction Level

^{*} Core-Bench provides original code for reproduction.

All other benchmarks are based solely on papers.