DOWNLOADABLE BENCHMARK BRIEF

Coding Agent Maintenance Suite Buyer Brief

Generated from Edxperimental Labs benchmark data: model rows, task traces, task mix, leaderboard controls, and the next evidence to collect before production.

56	Frontier reasoning model
AVERAGE SCORE	CURRENT LEADER
4	8 public / 12 private holdout tasks
TRACE PACKETS	SPLIT

EXECUTIVE READOUT

Buyer decision memo

Usable for constrained workflows with fallback routing.

MODEL CLASS	SCORE	RECOVERY	COST	P95
Frontier reasoning model	76	74	47	5848ms
Fast mid-tier model	62	58	79	4916ms
Open-weight local model	51	45	71	6210ms
Small routing model	34	28	94	4774ms

REPRESENTATIVE TRACE PACKETS

Inspectable tasks behind the score

TASK	DOMAIN	SPLIT	DIFFICULTY	TOP RUN	SCORE
Fix Command-K search regression	Frontend	public	Medium	Frontier reasoning model	78
Repair failing static build	Build	holdout	Hard	Frontier reasoning model	74
Add Playwright smoke test	Frontend QA	public	Medium	Frontier reasoning model	77
Refactor API error handling	Backend	holdout	Hard	Frontier reasoning model	75

RUBRIC
Patch correctness

RUBRIC
Regression rate

RUBRIC
Tool discipline

RUBRIC
Review readiness

LEADERBOARD CONTROLS

Controls attached to this run

Freshness	**Public sample refreshed monthly while private holdout stays sealed until replacement tasks exist.**
Leakage policy	**Do not use tasks sourced from public examples, vendor demos, or training-contaminated snippets without replacement variants.**
Repeat-run rule	**Repeat any result within five points of a leaderboard boundary across at least three seeds.**
Retirement rule	**Retire a task when frontier and mid-tier models cluster near the ceiling or when source material becomes widely circulated.**
Required provenance	**traceId, createdAt, split, source, modelVersion, runSeed, reviewerNote, retirementStatus**

Generated 2026-05-16T00:00:00+05:30 from dataset 0.1.0

sanjay@edxperimentallabs.com / saujas@edxperimentallabs.com