DOWNLOADABLE BENCHMARK BRIEF

Browser Operations 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.

51	Frontier reasoning model
AVERAGE SCORE	CURRENT LEADER
4	6 public / 10 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	71	69	49	5764ms
Fast mid-tier model	60	55	82	4874ms
Open-weight local model	43	39	76	6000ms
Small routing model	31	25	95	4732ms

REPRESENTATIVE TRACE PACKETS

Inspectable tasks behind the score

TASK	DOMAIN	SPLIT	DIFFICULTY	TOP RUN	SCORE
Pricing page extraction	Extraction	public	Medium	Frontier reasoning model	73
Multi-step demo form	Form completion	holdout	Hard	Frontier reasoning model	69
Invoice portal download	Navigation	public	Medium	Frontier reasoning model	72
Competitor feature map	Extraction	holdout	Hard	Frontier reasoning model	70

RUBRIC
Task success

RUBRIC
State verification

RUBRIC
Recovery quality

RUBRIC
Human handoff rate

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