DOWNLOADABLE BENCHMARK BRIEF

Support Agent Policy 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.

68
AVERAGE SCORE

Frontier reasoning model
CURRENT LEADER

4
TRACE PACKETS

8 public / 12 private holdout tasks
SPLIT

EXECUTIVE READOUT

Buyer decision memo

Strong candidate; inspect cost and latency before production use.

MODEL CLASS	SCORE	RECOVERY	COST	P95
Frontier reasoning model	86	82	51	5512ms
Fast mid-tier model	79	70	83	4790ms
Open-weight local model	58	47	74	6042ms
Small routing model	49	35	93	4816ms

REPRESENTATIVE TRACE PACKETS

Inspectable tasks behind the score

TASK	DOMAIN	SPLIT	DIFFICULTY	TOP RUN	SCORE
Refund policy boundary case	Refund decisions	public	Medium	Frontier reasoning model	88
Regional-language human handoff	Language handoff	holdout	Hard	Frontier reasoning model	83
Subscription downgrade save	Refund decisions	public	Medium	Frontier reasoning model	87
PII redaction escalation	Policy lookup	holdout	Hard	Frontier reasoning model	85

RUBRIC
Resolution rate

RUBRIC
Policy compliance

RUBRIC
Tone control

RUBRIC
Escalation precision

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