DOWNLOADABLE BENCHMARK BRIEF

Indian Enterprise Workflow 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.

69	Frontier reasoning model
AVERAGE SCORE	CURRENT LEADER
4	10 public / 14 private holdout tasks
TRACE PACKETS	SPLIT

EXECUTIVE READOUT

Buyer decision memo

Strong candidate; inspect cost and latency before production use.

MODEL CLASS	SCORE	RECOVERY	COST	P95
Frontier reasoning model	88	83	52	5638ms
Fast mid-tier model	76	66	81	4832ms
Open-weight local model	61	49	73	6126ms
Small routing model	52	36	92	4858ms

REPRESENTATIVE TRACE PACKETS

Inspectable tasks behind the score

TASK	DOMAIN	SPLIT	DIFFICULTY	TOP RUN	SCORE
GST invoice discrepancy explanation	Finance	public	Medium	Frontier reasoning model	91
Hindi-English refund escalation	Support	holdout	Hard	Frontier reasoning model	86
Vendor contract renewal risk	Legal	public	Medium	Frontier reasoning model	87
GST credit note reconciliation	Finance	holdout	Hard	Frontier reasoning model	89

RUBRIC
Outcome correctness

RUBRIC
Evidence citation

RUBRIC
Escalation judgement

RUBRIC
Cost per accepted output

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