

PREPARED FOR PAPPARTI

Papparti — Project Cost Estimate

The cost breakdown: how the figure in the quotation was arrived at.
For finance and technical review: every line, its basis, and its source.
Asks for: nothing on its own. It supports the quotation and the proposal.

AT A GLANCE

Phases priced	Feasibility, models, software 1-3, deployment
Model work	30,90,471 across four workstreams
Monthly running cost	1,26,108 at the prices read on 11 October 2026
Engineer rate used	3,385.84 per hour, from 35 USD at 96.7384
Prices	Live vendor prices, each with a source and a read date

DATE
11 October 2026

VALIDITY
30 days; prices re-quoted
after 10 November 2026.

CONFIDENTIALITY
Confidential. Prepared by
Vector Logic for Papparti.
Not for distribution beyond
Papparti and its advisers.

Contents

Papparti — Project Cost Estimate	3
1. One-page summary	4
1a. One-time build cost (excludes GST)	4
1b. Monthly running cost after handover	4
2. GPU rental — live prices (the foundation of every compute number)	5
3. Per-phase breakdown	6
Phase 0 — Discovery + Hindi/AI feasibility benchmark	6
Phase 1 — Voice core + booking engine, audio-only	6
Phase 2 — Avatar + polish	7
Phase 3 — Second vertical / locality	7
Phase 4 — Real-time lip-synced avatar (OPTION, open-ended)	7
Deployment + handover	7
4. Model workstream breakdown (three retrained models)	8
5. Monthly running cost after handover	10
6. Engineering rate — assumption, stated explicitly	11
7. Plumbing and third-party fees	12
7a. Infrastructure plumbing (see §5 for prices)	12
7b. Payment gateway — Razorpay	12
7c. GST treatment of imported services	12
8. Verdict on the ₹20–30 lakh expectation	13
9. Assumptions	14
10. Explicitly UNVERIFIED (excluded from every total)	15
11. What could not be verified	16
Sources	16

Papparti — Project Cost Estimate

Prepared by: Vector Logic (AI & software development, Odisha, India) **Prepared for:** Papparti (client) **Date:** 11 October 2026 **Scope:** On-demand any-service marketplace. An AI avatar is the primary interaction layer on both the user side and the provider side. Speech-to-text (STT) and text-to-speech (TTS) are **100% self-hosted** (Hindi and English, no cloud speech APIs). A reasoning language model runs continuously in the backend. Duration-aware booking (30 min / 1 hr / 5 hr) with payments, provider split, and an immutable ledger. A PWA (no native app), provider onboarding, working hours, and per-booking commission.

Vector Logic delivers **100% first-party development** and hands the system over. Three AI models are part of the deliverable, each **retrained (LoRA/QLoRA fine-tune) on the client's own data**:

1. **STT model** — Hindi + English, fine-tuned on the client's audio and domain vocabulary.
2. **TTS model** — a client-brand voice, fine-tuned from a licensed base.
3. **Reasoning model** — fine-tuned on the client's service catalogue, pricing rules and Hinglish booking

language; runs 24/7 in the backend.

Plus the complete software (marketplace app, booking engine, payments, ledger, admin, provider app) and self-hosted deployment on **rented GPU** (the client's own hardware is GPU-less).

NOTE

Model naming is confidential. This document states parameter counts, dataset descriptions and capabilities only. The specific base models chosen by Vector Logic are commercially confidential and are deliberately not named here.

Currency note — pinned FX rate: all USD figures are converted to INR at **1 USD = ₹96.7384** [5], a real mid-market rate captured from XE on 11 October 2026 (04:05 UTC). Every INR figure in this document is that rate applied to a USD list price fetched live on the same day.

1. ONE-PAGE SUMMARY

1a. One-time build cost (excludes GST)

#	COMPONENT	INR (EXCL. GST)	BASIS
1	Feasibility benchmark — Phase 0 (discovery + Hindi/AI benchmark, 1 rented 48 GB GPU, 1 month)	4,47,967	3 work-weeks + 1 month GPU [1]
2	Model work — four workstreams (STT, TTS, reasoning LLM, voice production)	30,90,471	GPU-hours [9] + engineering
3	Software — Phase 1 (voice core + booking engine, audio-only)	16,25,205	12 work-weeks
4	Software — Phase 2 (avatar + polish)	10,83,470	8 work-weeks
5	Software — Phase 3 (second vertical / locality)	10,83,470	8 work-weeks
6	Deployment + handover (self-hosted bring-up, CI/CD, runbooks, knowledge transfer, 30-day warranty)	5,41,735	4 work-weeks
	ONE-TIME TOTAL (excl. GST)	₹78,72,318	sum of the rows above
	ONE-TIME TOTAL (incl. 18% GST)	₹92,89,335	+18% [15]

Optional add-on (not in the total): | Phase 4 — Real-time lip-synced avatar (open-ended) | **₹13,54,338** | 10 work-weeks, priced as an option |

1b. Monthly running cost after handover

ITEM	INR / MONTH	BASIS
GPU serving — 1× 48 GB dedicated	41,665	RunPod RTX A6000 @ \$0.59/hr × 730 hr [1]
Managed Postgres + pgvector	5,891	DigitalOcean 4 GiB / 2 vCPU @ \$60.90/mo [12]
Redis	1,935	Upstash Fixed 1 GB @ \$20/mo [11]

ITEM	INR / MONTH	BASIS
SFU / TURN (LiveKit)	4,837	LiveKit Cloud Ship tier @ \$50/mo [13]
Object storage (Cloudflare R2)	1,161	~\$12/mo ESTIMATED (500 GB + ops) [14]
Monitoring + CI	2,902	~\$30/mo ESTIMATED — see §7
Support retainer	67,717	0.5 work-week / month @ agency rate
MONTHLY TOTAL	₹1,26,108	sum of the rows above

The headline answer for the CEO: the expectation of a **₹20–30 lakh one-time fee** is **NOT defensible** — **it is roughly one-third of the real cost**. The arithmetic gives **₹78.72 lakh excl. GST** (₹92.89 lakh incl. GST) for the scope as defined, plus ₹13.54 lakh if the Phase 4 real-time avatar option is added. See §8 for the full verdict.

2. GPU RENTAL — LIVE PRICES (THE FOUNDATION OF EVERY COMPUTE NUMBER)

All figures below were read from the vendors' own live pricing pages on **11 October 2026**. The 48 GB-tier GPU is the project's stated minimum for a pilot with a client-side avatar; the research's own floor is one 48 GB card.

PROVIDER	48 GB GPU	PER HOUR (USD)	PER MONTH (USD, ×730 H)	PER MONTH (INR)	SOURCE
RunPod (Community Cloud)	RTX A6000 48 GB	\$0.59	\$430.70	₹41,665	[1]
Vast.ai (on-demand, from)	RTX A6000 48 GB	\$0.28	\$204.40	₹19,773	[2]
Vast.ai (median rate)	RTX A6000 48 GB	\$0.40	\$292.00	₹28,248	[2]
Lambda Labs	RTX A6000 48 GB	\$0.80	\$584.00	₹56,495	[3]
E2E Networks (India)	L40S 48 GB	\$1.20	\$876.00	₹84,743	[4]
E2E Networks (India)	A40 48 GB	\$1.44	\$1,051.20	₹1,01,691	[4]

Correction to prior research. The project's earlier research cited RunPod RTX A6000 48 GB dedicated at ~USD 387/month (INR 34,056 at ₹88/USD). That figure is now **stale**. RunPod's live page today lists the RTX A6000 at **\$0.59/hr**, i.e. **\$430.70/month** — and at today's FX rate that is **₹41,665/month**, about **22% higher** than the research stated, and further from it because the research also used an outdated ₹88/USD rate. The planning number in this estimate is therefore the live one, ₹41,665/month [1].

Which provider to use: for training bursts, Vast.ai's on-demand floor (\$0.28/hr, ₹19,773/mo) is the cheapest live option [2]; its *median* rate (\$0.40/hr) is the more realistic planning figure because the floor machines are preemptible and constrained. For **24/7 production serving**, RunPod's dedicated \$0.59/hr [1] is the reference used throughout this estimate because it is a guaranteed-uptime dedicated instance. A **fully India-hosted** option exists (E2E Networks, INR-billed) but is **2–2.4× the price** per GPU-hour [4], which buys latency-to-India and INR invoicing but should be reserved for the serving layer if that latency turns out to matter.

3. PER-PHASE BREAKDOWN

Engineering is priced in **work-weeks** at an Indian AI-agency rate. The rate basis is set out in §6.

Phase 0 — Discovery + Hindi/AI feasibility benchmark

1 rented 48 GB GPU, 1 month.

LINE	QTY	INR
GPU rental — 1× 48 GB dedicated, 1 month	730 hr @ \$0.59	41,665
Engineering — discovery, benchmark harness, Hindi STT/TTS/LLM evaluation, written feasibility report	3 work-weeks	4,06,302
Phase 0 total		4,47,967

Phase 0 answers the one question the whole plan rests on: **does the chosen stack actually produce good Hindi, on rented 48 GB hardware, under load?** It produces a measured go/no-go report before full development spend.

Phase 1 — Voice core + booking engine, audio-only

10–14 weeks of engineering; priced at 12 weeks (midpoint).

The self-hosted STT + TTS pipeline (Hindi + English), the backend reasoning model endpoint, the duration-aware booking engine (30 min / 1 hr / 5 hr), payments, provider split, the immutable ledger, provider onboarding and working hours, and the PWA shell.

PHASE 1	WEEKS	INR
Voice core + booking engine	12	16,25,205

Phase 2 — Avatar + polish

6–10 weeks; priced at 8 weeks (midpoint). The avatar interaction layer on both sides (client-side viseme rendering, per the project's own decision), UX polish, and hardening.

PHASE 2	WEEKS	INR
Avatar + polish	8	10,83,470

Phase 3 — Second vertical / locality

6–10 weeks; priced at 8 weeks (midpoint). Onboarding a second service vertical or locality onto the same platform — catalogue, pricing rules, provider cohort, localisation.

PHASE 3	WEEKS	INR
Second vertical / locality	8	10,83,470

Phase 4 — Real-time lip-synced avatar (OPTION, open-ended)

Priced as a standalone option and **excluded from the one-time total**. This is the highest-risk, least-proven scope: true real-time neural lip-sync is server-video-cost territory (~9–11 MB/user-minute vs the viseme stream's ~9 KB/min, per the project's own research), so it is quote-on-request with a defined starting estimate.

PHASE 4 (OPTION)	WEEKS	INR
Real-time lip-synced avatar	10 (starting estimate)	13,54,338

Deployment + handover

4 work-weeks. Infrastructure-as-code for the self-hosted stack on rented GPU, CI/CD, monitoring wiring, runbooks, admin training, and a 30-day post-handover warranty.

DEPLOYMENT + HANDOVER	WEEKS	INR
Bring-up, CI/CD, runbooks, KT, warranty	4	5,41,735

4. MODEL WORKSTREAM BREAKDOWN (THREE RETRAINED MODELS)

Each workstream is a distinct sale: **data preparation** → **LoRA/QLoRA fine-tune** → **evaluation** → **packaging** → **endpoint serving setup**. GPU-hour figures are pinned to published QLoRA fine-tuning benchmarks [9] and to the QLoRA paper's demonstrated feasibility of fine-tuning very large models on a single 48 GB card. The base models are not named (confidential — see header).

WORKSTREAM	PARAMS	GPU-HOURS	GPU COST (INR)	ENGINEERING (WEEKS)	ENGINEERING (INR)	TOTAL (INR)
STT — Hindi + English, domain audio/vocab	1-2 B	40	2,283	6	8,12,603	8,14,886
TTS — client-brand voice	0.3-0.5 B	24	1,370	6	8,12,603	8,13,973
Reasoning LLM — catalogue, pricing rules, Hinglish booking	~32 B	180	10,274	8	10,83,470	10,93,744
Voice production — casting, direction, QC, master (eng) + studio recording (pass-through)	—	—	—	2	2,70,868 + 97,000	3,67,868
<i>of which: studio recording, ~20 h delivered audio</i>	—	—	—	—	97,000 (pass-through)	
Model work total		244	13,927	22	29,79,544	₹30,90,471

Voice production is included because Vector Logic records the brand voice itself (CEO decision, 11 Oct 2026). The studio recording line is the research's own modelled cost for ~20 h of delivered audio at India studio rates — **₹64,000 (low) to ₹1,30,000 (high), ₹97,000 taken as the midpoint** [see the assumptions section]. The engineering line covers casting, session direction, QC, retakes and mastering. **This converts what was the single biggest unmitigated blocker in the model plan — "voice cloning of a not-yet-chosen speaker is blocked on a human decision, a legal act and money, none of it engineering" — into work we control**, and it is why the TTS workstream also grew by one week: we now own our own data pipeline end to end rather than consuming a client recording.

One obligation this creates, and it is not optional: the voice artist must sign a **perpetual, irrevocable, worldwide, transferable commercial release** in the client's favour. The model is handed to the client at handover; if the release were revocable, the client's product would break the day the artist withdrew. Legal review of that release is a Phase 0 item.

GPU-hours are charged at the live RunPod 48 GB rate of **\$0.59/hr** [1]. The reasoning-model workstream dominates because a ~32 B model fine-tuned over multiple epochs on a large Hinglish/catalogue dataset is an order of magnitude heavier than the 7 B-class benchmark runs published in [9]; the STT workstream is heavy on *data preparation and evaluation* (transcription, WER harness, Hindi vocab) rather than raw GPU time. **The GPU compute is a small fraction (~0.5%) of each workstream — the cost is the engineering, data work and evaluation, not the rental.** This is the single most important thing to understand about the model line items.

5. MONTHLY RUNNING COST AFTER HANDOVER

ITEM	LIVE / ESTIMATED	INR / MONTH	BASIS
GPU serving — 1× 48 GB dedicated, 24/7	LIVE	41,665	RunPod @ \$0.59/hr × 730 [1]
Managed Postgres + pgvector	LIVE	5,891	DigitalOcean 4 GiB/2 vCPU @ \$60.90/mo [12]; Neon Launch usage-based alternative ~\$15/mo typical [10]
Redis	LIVE	1,935	Upstash Fixed 1 GB @ \$20/mo [11]

ITEM	LIVE / ESTIMATED	INR / MONTH	BASIS
SFU / TURN (LiveKit)	LIVE	4,837	LiveKit Cloud Ship tier @ \$50/mo [13]; self-hosted SFU on a VPS is a cheaper alternative
Object storage (Cloudflare R2)	ESTIMATED	1,161	~500 GB @ \$0.015/GB-month [14] + ops; egress is free [14]. Figure is an assumption, not a quote
Monitoring + CI	ESTIMATED	2,902	~\$30/mo placeholder — see §7
Support retainer	ASSUMPTION	67,717	0.5 work-week/month @ agency rate (\$6)
MONTHLY TOTAL		₹1,26,108	

Not included above: payment-gateway fees, which are transaction-linked, not fixed — see §7. The monthly GPU figure assumes **one** 48 GB card; the project's research notes that a production serving tier with real concurrency may need **two** 48 GB cards (~₹83,000/month at RunPod's live rate [1]).

6. ENGINEERING RATE — ASSUMPTION, STATED EXPLICITLY

Every engineering line in this document is priced at **\$35 per hour**, applied to a 40-hour work-week, converted at the pinned rate [5]:

$$\begin{aligned}
 \$35/\text{hr} \times ₹96.7384 &= ₹3,385.84 \text{ per hour} \\
 ₹3,385.84 \times 8 \text{ hr/day} &= ₹27,087 \text{ per day} \\
 ₹3,385.84 \times 40 \text{ hr/week} &= ₹1,35,434 \text{ per work-week}
 \end{aligned}$$

Why \$35/hr is defensible:

- **Clutch's Software Development Pricing Guide** lists India at **\$25–\$49/hour** [7].
\$35 sits inside

that band and above its midpoint.

- **Eucalipse (2025)** puts India at **\$18–40/hr, average \$29**, and defines an "**India (Premium)**"

tier at **\$35/hr** [8] — the AI/ML agency tier, which is what this project requires.

- This is an **AI / voice / LLM engineering** engagement, not generic web development; the rate is set at

the **premium** end of the Indian band deliberately.

This is an ASSUMPTION Vector Logic's CEO must confirm, because no vendor publishes a fixed public rate card. If the correct internal rate is different, every engineering figure scales linearly and the reader can re-run the arithmetic: **one work-week = ₹1,35,434 at \$35/hr**. To re-run at any rate R (USD/hour): **total = 36 engineering work-weeks $\times R \times 96.7384 \times 40$** . (Total billable engineering across the project is **36 work-weeks**: 3 + 12 + 8 + 8 + 4 phases, plus 19 model workstream weeks, excluding the Phase 4 option.)

7. PLUMBING AND THIRD-PARTY FEES

7a. Infrastructure plumbing (see §5 for prices)

Postgres + pgvector, Redis, an SFU/TURN layer (LiveKit or self-hosted equivalent), object storage, monitoring and CI. Live prices are cited for Postgres [12], Redis [11], the SFU [13] and object storage [14]; monitoring/CI and the R2 storage volume are **ESTIMATED** (assumptions, flagged as such, not baked into a "verified" total).

7b. Payment gateway — Razorpay

Razorpay charges a single flat **2% platform fee per successful domestic transaction**, with **18% GST on that platform fee**, and no setup, AMC or refund fees [6]. This is a **variable** cost that scales with transaction volume, not a fixed monthly line, so it is **not added to the monthly total**. Worked example: on ₹1,00,000 of settled GMV, the fee is **₹2,000 + 18% GST = ₹2,360**. This should be modelled against the per-booking commission to confirm the take-rate stays positive — see §8.

7c. GST treatment of imported services

Cloud and GPU services bought from foreign providers (RunPod, Vast.ai, Lambda, and R2) are an **import of services** under Section 2(11) of the IGST Act. The Indian recipient must **self-assess 18% IGST under the Reverse Charge Mechanism**, pay it in cash, and generally claim it back as input tax credit [15]. Two consequences for this estimate:

- **Foreign GPU/SaaS bills carry an 18% IGST reverse-charge obligation in cash**, recoverable as ITC if

Papparti is GST-registered and the service is used for business [15].

- **India-hosted E2E Networks** invoices are domestic (INR, GST on the invoice) and avoid the

reverse-charge mechanics, at the higher per-hour price noted in §2 [4].

Vector Logic's own development fee is an Indian supply of software development service (SAC 998314, **18% GST**), so the one-time total is quoted **both excluding and including 18% GST** in §1a.

8. VERDICT ON THE ₹20-30 LAKH EXPECTATION

The expectation is not defensible. It is roughly one-third of the real cost of the defined scope.

The arithmetic, from the cited rows:

- **One-time total (excl. GST): ₹78,72,318** — that is **3.15× the ₹25 lakh midpoint** of the CEO's band.
- The band's **ceiling** (₹30 lakh) would cover only the **model work (₹30.90 lakh, four workstreams)**

and **nothing else at all** — the retrained AI models plus the voice production they require now **exceed the entire ₹30 lakh ceiling on their own**, before a single line of marketplace software, booking engine, payments, ledger, admin, provider app, deployment or handover is written.

- To fit ₹30 lakh, the scope would have to be cut to roughly **one** of: the model work, or phases 0–1 of

the software. It cannot cover both.

Where the money actually goes (as a share of the ₹78.72 lakh one-time total):

COMPONENT	INR	SHARE
Software development (Phases 1–3)	37,92,145	51%
Model work (4 workstreams, incl. voice production)	30,90,471	39%
Feasibility benchmark (Phase 0)	4,47,967	6%
Deployment + handover	5,41,735	7%
Total	78,72,318	100%

Why the band is low, in one line: the CEO priced this like a marketplace app. It is a marketplace app **plus three custom-trained AI models plus a self-hosted voice stack** — and the AI half of that (\$4) is both the differentiator and nearly the entire ₹20–30 lakh band on its own.

Monthly after handover: ₹1,26,108, of which the largest single fixed item is the **support retainer (₹67,717)** — deliberately larger than the entire GPU bill, because a live 24/7 voice system needs ongoing engineering, not just hardware. The GPU serving cost (₹41,665) is the second largest.

One margin check the client should run before signing: Razorpay's 2% + 18% GST on the platform fee [6] is a per-transaction cost. If Papparti's per-booking commission is below ~2.4% of booking value, the gateway fee alone exceeds the platform's take on that transaction. Model the commission against \$7b before fixing the take rate.

Recommendation: present the honest number. **₹78.72 lakh excl. GST / ₹92.89 lakh incl. GST** for the defined scope, **₹13.54 lakh** for the Phase 4 option, and **₹1.26 lakh/month** to run it. If the client's budget is genuinely capped at ₹20–30 lakh, scope it down explicitly — the three-model workstream is the line to protect, because it is the part no one else is selling them.

9. ASSUMPTIONS

1. **FX rate** — 1 USD = ₹96.7384, XE mid-market, 11 Oct 2026, 04:05 UTC [5].

Pinned; all INR figures

use it. Re-run if the rate moves materially.

1. **Engineering rate** — \$35/hour, India AI-agency premium tier [7][8]. **CEO must confirm** (§6).

Arithmetic shown so it can be re-run at any rate.

1. **Work-week** — 40 hours, 5 days. A work-week = ₹1,35,434 at the assumed rate.

2. **Phases priced at midpoint** of their given week ranges: Phase 1 at 12 of 10–14 weeks; Phases 2–3 at

8 of 6–10 weeks each.

1. **GPU hours/month** — 730 (24×7 dedicated). Training GPU-hours are per \$4.

2. **Training GPU** — RunPod RTX A6000 48 GB @ \$0.59/hr [1] for all training and serving figures.

3. **Model parameter counts** — STT 1–2 B, TTS 0.3–0.5 B, reasoning ~32 B. Base models confidential.
4. **Support retainer** — 0.5 engineering work-week per month.
5. **Monitoring/CI and R2 volume** — estimated, not quoted (§7a).
6. **Payment gateway fees** — excluded from the fixed monthly total; modelled as variable (§7b).

10. EXPLICITLY UNVERIFIED (EXCLUDED FROM EVERY TOTAL)

The following lines are **assumptions or published-benchmark estimates, not live quotes**. They are **not** added to any verified sum; where they appear in a table they are labelled, and the totals in §1 stand independently of them.

- **Training GPU-hour counts for the three models** (40 / 24 / 180 hours in §4) — *not a vendor quote*.
- **Studio recording cost** (~₹97,000 for 20 h at ~₹4,850/delivered hour) — the research modelled a

₹64,000–₹1,30,000 band from India studio rates; the midpoint is used. *No studio quote obtained*. They are engineering estimates anchored to published QLoRA benchmarks (7 B QLoRA = \$0.75–\$1.50 per run on a single 24 GB card, 70 B QLoRA = \$24–\$64 per run) [9] and to the QLoRA paper's 48 GB feasibility result. The *rates* used (\$0.59/hr) are live [1]; the *hours* are estimated. **The cost of these lines is robust anyway — GPU compute is ~0.5% of each workstream.**

- **Monitoring + CI monthly cost** (~₹2,902 / ~\$30) — an assumption, not a quote.
- **Object-storage volume** (~500 GB → ~₹1,161/mo) — the *rate* (\$0.015/GB-month, free egress) is live

[14]; the *volume* is assumed.

- **Any second 48 GB serving GPU** — flagged as a possibility, not priced into the totals.
- **Phase 4 real-time lip-synced avatar** — open-ended; the ₹13.54 lakh is a **starting estimate**, not a

quote, and is excluded from the one-time total.

11. WHAT COULD NOT BE VERIFIED

- **A published Indian dev-agency rate card.** No vendor publishes a fixed public rate for AI engineering;

the \$35/hr figure rests on the Clutch India band [7] and the Eucalipse India-premium tier [8], and is recorded as an assumption the CEO must set (§6).

- **A quoted monthly/annual price from any GPU vendor.** RunPod, Vast.ai and Lambda publish **per-hour**

prices only [1][2][3]; monthly figures here are per-hour $\times$ 730 and are labelled as such.

- **E2E Networks monthly plans.** E2E lists India per-hour prices [4] but its monthly/annual plans are

"Contact Sales" — not quotable.

- **Vendor-specific latency for India-served inference.** Not measured; it is the reason the (pricier)

India-hosted option in §2 is kept as a contingency, not the default.

- **Concurrency capacity of a single 48 GB card under real load.** The project's own research flags this as

unmeasured; the monthly GPU line assumes single-stream serving.

SOURCES

[1] www.runpod.io/product/cloud-gpus — Runpod GPU Cloud Pricing (RTX A6000 48GB \$0.59/hr, updated Aug 27 2026) [2] vast.ai/pricing — Vast.ai Live GPU Pricing (RTX A6000 48GB from \$0.28/hr median \$0.40) [3] lambdalabs.com/service/gpu-cloud/pricing — Lambda GPU Cloud VM Pricing (RTX A6000 48GB \$0.80/hr) [4] www.e2enetworks.com/pricing — E2E Networks India GPU Cloud Pricing (L40S 48GB \$1.20/hr, A40 48GB \$1.44/hr) [5] www.xe.com/currencyconverter/convert/?Amount=1&From=USD&To=INR — XE USD/INR mid-market rate (1 USD = 96.7384 INR, 11 Oct 2026) [6] www.razorpay.com/pricing — Razorpay India Pricing (2% platform fee + 18% GST) [7] clutch.co/developers/pricing — Clutch Software Development Pricing Guide (India \$25-49/hr) [8] eucalipse.com/articles/software-development-costs-by-country-2025 — Eucalipse Software Development Cost by Country 2025 (India \$18-40/hr avg \$29) [9]

www.heulistic.com/blog/llm-fine-tuning-cost-breakdown-model-size — Heulistic LLM Fine-Tuning Cost Breakdown by Model Size (2026) [10] neon.com/pricing — Neon Postgres + pgvector Pricing [11] upstash.com/pricing/redis — Upstash Redis Pricing [12] www.digitalocean.com/pricing/managed-databases — DigitalOcean Managed PostgreSQL Pricing [13] livekit.io/pricing — LiveKit Cloud Pricing (SFU/TURN self-host or cloud) [14] developers.cloudflare.com/r2/pricing — Cloudflare R2 Object Storage Pricing (zero egress) [15] startupadvisory.in/blog-gst-rcm-import-of-services.htm — GST on Import of Services under RCM (18% IGST)