

VECTOR LOGIC

AI & SOFTWARE DEVELOPMENT — ODISHA, INDIA

PREPARED FOR PAPPARTI

Full project report

What this is. One document containing everything Vector Logic proposes for the Papparti voice marketplace — the commercial quotation, the proposal, the cost estimate behind it, the decision on how to run it, and the payment schedule — plus an executive summary and appendices listing every figure and where it came from.

How to read it. Reviewers short on time should read the executive summary (pages 2-11) and the quotation (pages 12-16). The proposal explains the shape of the work; the estimate and its appendices show every number's origin. Figures are in Indian numbering (₹78,72,318, not ₹7,872,318) and exclude GST unless a line says otherwise.

On page numbers. Each of the seven parts is a document in its own right, so each carries its own numbering and its own contents — the estimate's pages run 1 to 17, not 31 to 47. The numbers beside the contents list below are for THIS combined file. Pull any part out and send it on its own and its numbering is still correct. That is why the two sets differ, and it is deliberate.

CONTENTS

Executive summary and how to read this	2
Quotation	12
Proposal	17
Cost estimate	32
GPU: rent vs buy	49
Commercial terms, milestones and payment	62
Appendices — every figure and its source	78

DATE

11 October 2026

CONFIDENTIALITY

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

VECTOR LOGIC

AI & SOFTWARE DEVELOPMENT — ODISHA, INDIA

PREPARED FOR PAPPARTI

Papparti — Report Front Matter

DATE

11 October 2026

CONFIDENTIALITY

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

Contents

Papparti — Report Front Matter	3
A. Executive summary	3
The numbers	3
The timeline shape	4
The three decisions the client must make	4
The single next step	4
B. How to read this document	5
C. Glossary	6
D. The decisions this document asks for	9

Papparti — Report Front Matter

Project: Papparti — on-demand any-service marketplace with a self-hosted voice and avatar layer **Prepared for:** Papparti **Prepared by:** Vector Logic — AI & software development, Odisha, India **Date:** 11 October 2026 (IST) **Currency:** All figures in INR, Indian numbering. USD converted at the pinned rate **1 USD = ₹96.7384**. **Status:** Front matter to the client package (00 - QUOTATION.md, 01 - PROPOSAL.md, 02 - ESTIMATE.md, 03 - GPU-DECISION.md, 04 - COMMERCIAL - TERMS.md). Every figure below traces to those documents.

A. EXECUTIVE SUMMARY

Papparti is buying one system with five parts, delivered together:

1. **The marketplace software** — a two-sided on-demand services platform: customer booking, provider onboarding and hours, duration-aware scheduling, payments with a provider split, an immutable ledger, an admin console, a provider app. Delivered as a Progressive Web App; no app store.
2. **A speech recognition model** — custom-trained so the system understands spoken Hindi, English and Hinglish and your service vocabulary.
3. **A text-to-speech model** — producing a voice that belongs to Papparti, recorded in a studio session by Vector Logic.
4. **A reasoning model** — running continuously in the backend, turning a spoken request into a structured, **confirmed** booking.
5. **A talking avatar** — layered on top of the booking engine, not as its foundation.

The numbers

ITEM	AMOUNT (₹)	WHAT IT IS
One-time project fee, excluding GST	78,72,318	Phases 0–3, model work, deployment and handover
GST @ 18%	14,17,017	Charged on top of the one-time fee
One-time fee, including GST	92,89,335	What a fully GST-registered Indian contract totals
Monthly running cost, after go-live	1,26,108	GPU, hosting plumbing, and a support retainer
Optional Phase 4 — real-time lip-synced avatar	13,54,338	Priced separately; not in the one-time total

The timeline shape

Six milestones, each tied to a **delivered, accepted artefact** — never to a date:

MILESTONE	WEEKS (ENGINEERING)	INVOICE INR (EXCL. GST)
Mobilisation advance (10%, recovered at Model work)	—	7,87,232
Phase 0 — feasibility benchmark	3 work-weeks + 1 month GPU	4,47,967
Model work — four workstreams	22	30,90,471
Phase 1 — voice core + booking engine	12	16,25,205
Phase 2 — avatar + polish	8	10,83,470
Phase 3 — second vertical / locality	8	10,83,470
Deployment + handover	4	5,41,735

Before the big spend, Phase 0 measures whether the voice stack actually works in Hindi on one rented 48 GB GPU over one month, against ten recorded pass/fail gates. A measured "no" is a complete answer, and it stops the project at the Phase 0 price point. The monthly charge of ₹1,26,108 begins **only when the system is live and serving real bookings** — not on handover date.

The three decisions the client must make

1. **Proceed as scoped** at ₹78,72,318 excluding GST (₹92,89,335 including GST), or **scope down explicitly** — a ₹20–30 lakh budget covers the model work alone and nothing else.
2. **Pick ONE service vertical and ONE locality for launch.** An unbounded "any service" launch is the measured failure mode of this sector in India.
3. **Decide the Phase 4 option** — the real-time lip-synced avatar at ₹13,54,338 — now or later. It is not a launch dependency.

The single next step

Supply the two blank fields (your registered legal entity name, and who signs and in what capacity) and confirm in writing before 10 November 2026. That is all that stands between this document and a signature-ready contract.

B. HOW TO READ THIS DOCUMENT

This page is for a reader who is not technical. Read it once and the rest of the report will make sense.

The one-time fee versus the monthly charge, and when each starts. The **one-time fee** is what it costs to build and hand over the system — the software, the three trained models, the brand voice recording, the deployment and training. It is invoiced across six milestones, each when its deliverable is accepted, plus a 10% mobilisation advance recovered later. The **monthly charge** is what it costs to *run* the finished system: the GPU it runs on, the database and file stores, the real-time media layer, monitoring, and a support retainer for ongoing engineering. **The monthly charge starts only when the system is live and serving real bookings**, not on the day it is handed over.

What GST is, and that it is charged on top. GST is India's Goods and Services Tax — a government tax added to the price of a supply. Vector Logic's development work is a service, and GST at **18%** is charged **on top of** the one-time fee on each invoice. $\text{₹}78,72,318 \times 18\% = \text{₹}14,17,017$, so the GST-inclusive one-time total is **₹92,89,335**. Where Papparti is GST-registered and uses the service for business, that GST is generally claimable back as input tax credit.

What "excluding GST" means for the number you saw. When the headline number reads **₹78,72,318 excluding GST**, it means that number is what Vector Logic charges **before** the 18% government tax is added. The amount actually invoiced to a GST-registered client is **₹92,89,335**. The monthly figure of ₹1,26,108 is likewise an excluding-GST figure; GST applies to it as applicable.

Why the monthly figure does not fall when usage is low. The GPU is billed **for the hours it exists, not the hours it is used** — a rented 48 GB card runs at $\$0.59/\text{hour} \times 730 \text{ hours a month} = \text{₹}41,665$, 24×7, busy or idle. The reasoning model also runs continuously by design. If bookings are light, the vendor bill is unchanged. The GPU decision note shows when buying instead would become cheaper (about **21 months** for a used card, **61 months** for a new one) and **recommends renting until Phase 0 measures real utilisation**.

The three labels used throughout these documents.

- **MEASURED** — we ran an instrument and can show you the number. Example: the voice samples' durations and speech energy (dBFS); the avatar's pixel change between frames.

- **ESTIMATED** — a modelled figure with its assumption stated in the same place.
Example: the studio recording cost (₹97,000, the midpoint of a ₹64,000–₹1,30,000 band); the monthly monitoring and storage lines.
- **UNVERIFIED** — we could not confirm it, and it must be confirmed before you rely on it. Example: the SAC tax codes; whether the voice stack produces good Hindi at all — which is exactly what Phase 0 exists to measure.

Why you will see a placeholder for your own company details. The address in these documents reads "Papparti", which is a **brand**, not a registered legal entity. An invoice needs the registered entity's name, and a contract needs an authorised signatory. Neither has been supplied, so rather than guess, the documents carry deliberate blank fields and the placeholder email **contact@papparti.com only** in the demo references. Tell us the real details and we correct them everywhere.

C. GLOSSARY

Every term below appears in the five documents in this package. Definitions are plain English; the "where it matters" note says which document and section the term decides something in.

TERM	PLAIN-ENGLISH MEANING	WHERE IT MATTERS
Adapter	A small trained add-on that sits on top of a frozen base model and changes its behaviour for your data, without retraining the whole model.	Proposal §1; the three models are adapters, not train-from-scratch models.
ASR (Automatic Speech Recognition)	The machine task of turning spoken audio into text. Same job as STT; the two names are interchangeable.	Estimate header; the speech-recognition model.
Change order	Any change to the agreed scope, raised in writing and priced before work begins.	Commercial Terms §5; changes over ₹13,542 are quoted.
Concurrency	How many conversations the system can serve at the same time.	Estimate §5; a hidden driver of whether one GPU card is enough.
Diarisation	Telling apart <i>who</i> is speaking when more than one person talks — e.g. customer versus provider on a call.	Proposal §3.1 / §3.3; relevant once real booking audio is used.

TERM	PLAIN-ENGLISH MEANING	WHERE IT MATTERS
Fine-tune	Training an existing model further on your own data so it performs better on your domain, rather than building a model from nothing.	Proposal §1; the documented meaning of "retrain" in this project.
GST (Goods and Services Tax)	India's national tax on supplies of goods and services; here 18% on the development service.	Quotation §A; adds ₹14,17,017 to the one-time fee.
GSTIN	A business's GST registration number, read from its certificate.	Commercial Terms §4.2; a blank field until supplied.
Go/no-go gate	A pre-agreed pass/fail threshold that decides whether the project continues, changes, or stops.	Proposal §4; the ten gates Phase 0 reports against.
GPU (Graphics Processing Unit)	The specialised processor on which AI models are trained and run.	Quotation §C; the single largest monthly line at ₹41,665.
Hinglish	Hindi and English mixed within the same sentence — the everyday speech pattern of the target user.	Proposal §8.1; the hardest and least-solved part of the voice stack.
HSN / SAC	Standard codes used on an Indian tax invoice to classify what was supplied.	Quotation §A; UNVERIFIED — to be confirmed with a tax adviser.
Input credit (ITC)	GST you paid on your inputs, which a registered business can generally claim back against the GST it charges onward.	Estimate §7c; applies to Vector Logic's GST and to reverse-charge IGST.
Latency	The delay between a user doing something and the system responding.	Proposal §5; Phase 2 gate is the avatar adding under ~800 ms.
LoRA / QLoRA	The standard, efficient methods for fine-tuning a large model on modest hardware — training a small adapter instead of the whole model.	Proposal §1; the industry-standard method this project uses.
Mobilisation advance	A payment drawn at the start of the project against front-loaded costs (GPU, studio booking), recovered later.	Commercial Terms §2.1–2.4; ₹7,87,232, recovered at the Model work milestone.
MOS (Mean Opinion Score)	The average score human listeners give to speech quality — the only real test of whether a voice sounds good.	Proposal §8.1; no MOS exists yet ; the listening test runs in Phase 1.

TERM	PLAIN-ENGLISH MEANING	WHERE IT MATTERS
Parameter class	The rough size of a model, measured in billions of parameters (B). Bigger is generally heavier to run.	Estimate §4; recognition 1–2 B, synthesis 0.3–0.5 B, reasoning ~32 B.
p95	The value below which 95% of measurements fall — a measure of the slow cases, not the average.	Proposal §4; the p95 time-to-first-audio gate is 1,500 ms.
Place of supply	The rule that decides which state's GST (CGST/SGST or IGST) applies to a service supply.	Commercial Terms §4.3; recorded on each invoice.
Retention / holdback	A slice of the fee the client keeps until the delivered system is accepted.	Commercial Terms §2.3; ₹3,93,616 (5%), released at handover acceptance.
RTF (Real-Time Factor)	How long the machine took to process audio relative to the audio's own length — under 1.0 means faster than live.	Proposal §8.1; a measured objective figure, not a quality score.
Self-hosting	Running the models on infrastructure you control rather than through a third-party cloud API.	Proposal §1; a control-and-compliance decision, explicitly not a cost saving .
Sessions per node	How many simultaneous conversations one GPU machine can hold before quality or speed degrades.	Proposal §4; gate is ≥ 10 ; below that the cost model breaks.
SFU / TURN	The infrastructure that carries live audio between phones in real time across real networks.	Quotation §C; the ₹4,837/month real-time media line.
SLA (Service Level)	An agreed expectation of how reliable a service will be.	Commercial Terms §9; liability is bounded, and no accuracy figure is warranted pre-measurement.
STT (Speech-to-Text)	Same as ASR — turning speech into text.	Estimate §4; the speech-recognition workstream.
STT / ASR	The two names for the same recognition task; used interchangeably in these documents.	Estimate header and §4.
TDS (Tax Deducted at Source)	Tax the payer withholds from a payment and remits on the supplier's behalf.	Commercial Terms §4.5; section and rate must be confirmed with a tax adviser.
TTS (Text-to-Speech)	The machine task of turning written text into spoken audio — the brand voice.	Estimate §4; the synthesis workstream.

TERM	PLAIN-ENGLISH MEANING	WHERE IT MATTERS
Vector store	A database that stores text meanings as numbers so the system can find relevant information quickly.	Quotation \$C; the ₹5,891/month managed database line includes it.
Voice cloning	Building a synthetic voice that matches a specific recorded speaker. Requires that speaker's signed release.	Proposal \$3.4; no release = no clone.
VRAM	The memory on the GPU. Model size is limited by how much VRAM the card has (here 48 GB).	GPU Decision \$2b; 48 GB is the stated safe floor for the project.
Warm handoff	Passing a conversation smoothly from the automated system to a human, without dropping it.	Proposal \$5 / Commercial Terms \$3; the fallback when voice fails.
WER (Word Error Rate)	The percentage of words in a transcript that came out wrong; lower is better.	Proposal \$4; the recognition gates are 20% Hinglish and 12% Hindi.
WER gate	The specific WER threshold that decides whether the voice-first design proceeds or becomes text-first.	Proposal \$4; failure means stop/redesign, before big spend.
Warranty	The 30-day window after handover in which defects are fixed at no charge.	Quotation \$D; starts on signature of the handover acceptance certificate.

D. THE DECISIONS THIS DOCUMENT ASKS FOR

Please answer each of these as a clear yes or no, in writing. They are the only decisions that block a signature.

1. **Do you accept the one-time fee of ₹78,72,318 excluding GST (₹92,89,335 including GST) for the scope as written?** If your budget is capped lower, do you want a written scope-down of the same system, and which part do you protect?
2. **Do you approve Phase 0 first**, at ₹4,47,967, as a measurement you buy before any large development spend — accepting that a measured "no" stops the project at that price point?
3. **Which ONE service vertical and which ONE locality** will launch first?
4. **Do you want the Phase 4 real-time lip-synced avatar** at ₹13,54,338 as a priced option now, or do you defer it entirely?

5. **Do you accept the mobilisation advance of 10% (₹7,87,232)** and the retention of 5% (₹3,93,616) released at handover acceptance?
6. **Will you supply the model-the-voice decision and the signed voice-rights release** in Papparti's favour — perpetual, irrevocable, worldwide, transferable, commercial — before the voice is cloned?
7. **Do you accept that recognition and synthesis run on self-hosted, rented GPU** as a control-and-compliance choice, knowing it is **not** the cheaper option at plausible volumes?
8. **Will you nominate the legal entity, the signatory and designation, the contact email, the GSTIN and the tax adviser** so the contract, the SAC codes and the invoicing can be completed?

PREPARED FOR PAPPARTI

Quotation — Papparti voice-first marketplace

The commercial quotation for the Papparti voice project.
For finance and procurement review: line items, GST, and totals.
Asks for: your registered entity name and a signatory, so this becomes an invoice.

AT A GLANCE

One-time project fee	78,72,318 excluding GST
With 18% GST	92,89,335
Monthly, after go-live	1,26,108
Optional Phase 4	13,54,338, priced separately
Valid	30 days from 11 October 2026

DATE
11 October 2026

VALIDITY
30 days; expires 10 November 2026.

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

Contents

Quotation — Papparti voice-first marketplace	3
A. One-time project fee	3
B. Optional add-on — priced separately, not in the total above	4
C. Monthly, after handover and go-live	4
D. What is included / not included	4
E. Before this can be signed	5
F. Honest note on the two figures a client usually questions	5

Quotation — Papparti voice-first marketplace

From: Vector Logic · Odisha, India · ceo@vectorlogic.in **To:** Papparti **Date:** 11

October 2026 (IST) **Quotation no.:** VL-2026-1011-PAP-01 **Valid until:** 10

November 2026 (30 days) **Currency:** INR, Indian numbering

A. ONE-TIME PROJECT FEE

#	ITEM	AMOUNT (₹)
1	Feasibility benchmark — Phase 0 (discovery + Hindi/Hinglish measurement, 1 rented 48 GB GPU, 1 month)	4,47,967
2	Model work — four workstreams: speech recognition, speech synthesis, reasoning model, voice production (incl. studio recording)	30,90,471
3	Software — Phase 1 (voice core + booking engine, audio-only)	16,25,205
4	Software — Phase 2 (talking avatar + polish)	10,83,470
5	Software — Phase 3 (second vertical / locality)	10,83,470
6	Deployment + handover (self-hosted bring-up, CI/CD, runbooks, knowledge transfer, 30-day warranty)	5,41,735
Subtotal, excluding GST		78,72,318
GST @ 18%		14,17,017
Total, including GST		92,89,335

NOTE

HSN/SAC: to be confirmed with our tax adviser before invoice — see §D of the commercial annexure.
Left blank deliberately rather than guessed.

B. OPTIONAL ADD-ON — PRICED SEPARATELY, NOT IN THE TOTAL ABOVE

#	ITEM	AMOUNT (₹)
7	Phase 4 — real-time lip-synced avatar (open-ended; scoped optional)	13,54,338

C. MONTHLY, AFTER HANDOVER AND GO-LIVE

#	ITEM	₹ / MONTH
1	GPU serving — 1× 48 GB dedicated instance	41,665
2	Managed PostgreSQL + vector store	5,891
3	Redis	1,935
4	Real-time media routing (SFU/TURN)	4,837
5	Object storage	1,161
6	Monitoring + CI	2,902
7	Support retainer (0.5 work-week / month)	67,717
Monthly total		1,26,108

Plus GST where applicable. **Payment gateway / transaction fees are not included** — they are transaction-linked, not fixed.

NOTE

Not in any total above: the one-time fee assumes the scope as written. Changes are quoted separately under the scope-change mechanism in the commercial annexure.

D. WHAT IS INCLUDED / NOT INCLUDED

Included in the one-time fee: the three fine-tuned models as trained artefacts and their weights; the marketplace software; the studio voice recording; self-hosted deployment; CI/CD; runbooks; knowledge transfer; a 30-day warranty after handover. **On handover the models and their weights are your property.**

Not included: the monthly running costs in §C; third-party transaction fees; hardware purchase (the estimate assumes a rented GPU and explains why — see the GPU decision note); content, catalogue or provider onboarding for the marketplace; legal review of your own customer terms.

E. BEFORE THIS CAN BE SIGNED

Two fields are **deliberately blank** and must come from you:

FIELD	NEEDED FOR
Your legal entity name as registered	The contract and every invoice
Signatory name + designation	Unambiguous milestone acceptance

Vector Logic will add its GSTIN and the confirmed SAC code at invoice time.

Nothing in this quotation is legal advice; the commercial annexure (04-COMMERCIAL - TERMS .md) sets out milestones, acceptance criteria, termination and the voice-rights release requirement.

F. HONEST NOTE ON THE TWO FIGURES A CLIENT USUALLY QUESTIONS

1. **Why ₹78,72,318 and not the ₹20–30 lakh band?** Because the band priced a marketplace app,

and this is a marketplace app **plus three custom-trained models plus a self-hosted voice stack**. The model work alone (₹30,90,471) is the entire band. The full reasoning is in the estimate.

1. **Why does the monthly not fall if usage is low?** Because the GPU is billed for the hours it

exists, not the hours it is used — and the reasoning model runs 24/7 by design. The GPU decision note shows the break-even for buying instead (**21 months for a used card, 61 for a new one**) and recommends renting until Phase 0 measures real utilisation.

Prepared by Vector Logic. Figures in this quotation are taken from CLIENT-ESTIMATE .md (11 October 2026) and are valid for 30 days.

PREPARED FOR PAPPARTI

Papparti — Project Proposal

What Vector Logic proposes to build for Papparti, and why it is shaped this way.
For the decision maker: scope, approach, phases, and what is deliberately excluded.
Asks for: a go-ahead on the scope, or a list of changes first.

AT A GLANCE

The product	A voice-first on-demand services marketplace
Three trained models	Speech recognition, speech synthesis, reasoning
Runs where	Self-hosted on a dedicated GPU; the models become yours
Delivery	Three software phases, after a measured feasibility phase
One-time fee	78,72,318 excluding GST

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 Proposal	3
Section 1 — What you are buying	3
What "self-hosted" buys you commercially	3
What is delivered	3
Section 2 — Why voice-first, and why now	4
The evidence that supports the approach	4
The honest finding: an avatar as the WHOLE interface is the weakest commercial case	4
How this proposal mitigates it	4
Section 3 — The three models	5
3.1 Speech recognition ("hearing")	5
3.2 Speech synthesis ("speaking")	5
3.3 Reasoning model ("deciding") — runs 24/7	5
3.4 The voice recording workstream (Vector Logic)	5
Section 4 — What we will prove before we build at scale (Phase 0)	6
The hard gates	6
What a failure means, commercially	6
Cost and duration	6
Section 5 — Scope, phase by phase	7
How the phases fit together	7
Optional add-on, priced separately (not in the total above)	7
Section 6 — Commercial	7
5.0 The demo set, and what each artefact proves	7
6.0 Validity, and the fields we need from you	8
6.1 One-time project fee	8
6.2 Monthly after handover	8
6.3 What is included	9
6.4 What is not included	9
6.5 Payment milestones	9
6.6 On the budget band	9
Section 7 — What you get at handover	10
Section 8 — Risks, stated plainly	10
8.1 Risks with no mitigation	10
8.2 Mitigated risks	11
8.3 Commercial risks you should weigh	12
8.3 The talking avatar — what you are seeing, and what you are not	12
8.4 The voice samples you can listen to now	13
Section 9 — What we need from you	13
Section 10 — Assumptions	14
Explicitly unverified, and excluded from every total	15
What we could not verify	15

Papparti — Project Proposal

Project: Papparti — on-demand any-service marketplace with a self-hosted voice and avatar layer **Prepared for:** Papparti **Prepared by:** Vector Logic — AI & software development, Odisha, India **Date:** 11 October 2026 (IST) **Currency:** All figures in INR. USD prices are converted at a pinned rate of **1 USD = ₹96.7384** (XE mid-market, 11 October 2026, 04:05 UTC) — the rate used in our cost estimate.

SECTION 1 — WHAT YOU ARE BUYING

You are buying four things, delivered together, as one system:

1. **The marketplace software.** A two-sided on-demand services platform: customer booking, provider onboarding and working hours, duration-aware scheduling, payments with a provider split, an immutable ledger, an admin console, and a provider app. Delivered as a Progressive Web App (PWA) — installable from the browser, no app store.
2. **A speech recognition model**, custom-trained on your audio and your service vocabulary.
3. **A text-to-speech model**, producing a voice that belongs to Papparti.
4. **A reasoning model** that runs continuously in the backend, turning a spoken Hindi/Hinglish request into a structured, confirmed booking.

Plus the voice itself: your brand voice is **recorded by Vector Logic** in a studio session (engineering, casting, direction, QC and the recording are a priced line in this proposal — see Section 3).

What "self-hosted" buys you commercially

Speech recognition and speech synthesis run **on infrastructure you control**, not through a third-party cloud speech API. The commercial consequences:

- **You own it.** On handover, the three trained models become Papparti's property, along with the code. You are not renting access to a service that can be repriced or withdrawn.
- **No per-request fee to a third party.** Recognition and synthesis do not meter per call to an external vendor. The marginal cost of one more spoken booking is your own compute, not someone else's rate card.
- **Your users' audio does not leave your stack.** Sending customer voice to an outside speech API would export your users' audio to a third party; self-hosting does not.
- **This is not the cheaper option, and we will not pretend it is.** Our own research is explicit: hosted cloud speech would cost less per booking than running your own GPU tier. Self-hosting is a **strategic and compliance decision, not a cost saving**, and it should be budgeted as one. You are buying control and independence with money you could otherwise spend elsewhere.

What is delivered

ITEM	DELIVERED
Marketplace software (customer + provider sides, admin, payments, ledger)	Yes
Speech recognition model — trained on your data	Yes
Speech synthesis model — trained on your data	Yes
Reasoning model — trained on your data, runs 24/7	Yes
Brand voice recording (studio session, ~20 h delivered audio)	Yes — Vector Logic workstream
Self-hosted deployment on rented GPU	Yes
Code, model weights, endpoints, runbooks, knowledge transfer	Yes

ITEM	DELIVERED
30-day warranty after handover	Yes

One scope point stated up front, in writing: the three models are **fine-tunes**. The base model's weights are frozen; your data trains an *adapter* on top of it (the industry-standard method, LoRA/QLoRA). This is not a train-from-scratch pretrain. Our own decision record flags this as something that must be closed in writing before signature, because "retrain" can be read to mean more than it delivers. This is what "retrain" means here.

SECTION 2 — WHY VOICE-FIRST, AND WHY NOW

The evidence that supports the approach

- **India's market is conversational, and largely not app-native.** Our research's own framing of the strongest opportunity is an **agent-native, voice-first, vernacular-first marketplace for India's ~98%-offline market, reachable without a smartphone app** — closer to a conversational agent than a photoreal avatar. Hindi/Hinglish speech is the natural interface for that user.
- **Voice ordering has already been proven in India at scale.** Hindi + English voice-first ordering shipped in this market in April 2026. The interface is no longer speculative — which also means it is no longer unique. See the honest finding below.
- **The delivery model is reachable.** A PWA on a WebGL2 browser works on Indian Android phones without WebGPU, without downloading model weights, and without the Play Store. This is what makes a voice-and-avatar product viable on the device band this business must serve.

The honest finding: an avatar as the **WHOLE** interface is the weakest commercial case

Our research measured the market and did not soften its conclusion. You should have it in front of you before you commit money:

- **"An avatar as the entire UI on both sides" is the only genuinely unshipped square in this space — and the research calls it the weakest commercial case in the plan.**
- The evidence against avatar-first commerce is real: **only about 22% of smart-speaker users buy by voice**, and **about 45% will not send a payment through a voice assistant.**
- Voice-first Hindi/English booking is **already shipped in India** (April 2026, 11 languages), so voice alone is not a differentiator.
- Self-hosted Indic speech is feasible but **not unique and invisible to users** — a margin and compliance play, not a customer-facing feature.

How this proposal mitigates it

We build what you asked for, with one structural safeguard:

1. **Confirm-before-price stays in text, always.** The model **proposes** a duration; the customer **confirms** it; **only the confirmed value is priced**. A conventional on-screen confirmation, receipt and ledger stay in place for every booking and every rupee — not voice-only. An inferred duration that becomes a bill is, in our research's words, "the first trust-destroying event," and we treat it as a hard, unbreakable gate.
2. **The avatar is a differentiator, not the foundation.** It is layered on top of a booking engine that works without it. If the avatar underperforms, the business does not.
3. **We test the thesis cheaply before spending the big money.** Section 4's Phase 0 measures whether the voice stack actually works in Hindi before full development spend.

This is not a reason to stop. It is a reason to decide deliberately.

SECTION 3 — THE THREE MODELS

All three are **fine-tuned on Papparti's data** and **self-hosted**. Parameter scales are given; the specific base models are commercially confidential to Vector Logic and are deliberately not named in this document. What we can tell you is the data each is trained on and what it does.

3.1 Speech recognition ("hearing")

Parameter scale	1-2 billion parameters
Trained on	Public Hindi and Hinglish speech corpora (~150-250 hours total: spontane
What it does	Turns spoken Hindi, English and Hinglish into text for the booking flow
Where it runs	Self-hosted on a rented 48 GB GPU. Not a cloud speech API
Licence posture	Trained on commercially-licensed data only

3.2 Speech synthesis ("speaking")

Parameter scale	0.3-0.5 billion parameters
Trained on	Papparti's brand voice — ~20 hours of studio-recorded audio of one ch
What it does	Speaks the booking conversation in Papparti's own voice, in Hindi and Eng
Where it runs	Self-hosted on the same rented 48 GB GPU
Design requirement	Must survive being interrupted mid-sentence (a customer talking over

3.3 Reasoning model ("deciding") — runs 24/7

Parameter scale	~32 billion parameters
Trained on	Synthetic booking dialogues built from a grammar of your booking domain
What it does	Converts a Hindi/Hinglish/English conversation into a validated structured
Where it runs	Self-hosted on the same rented 48 GB GPU
The non-negotiable behaviour	On an ambiguous request ("thoda kaam hai", "jaldi aao") it must a

3.4 The voice recording workstream (Vector Logic)

LINE	INR
Voice production — casting, session direction, QC, retakes, mastering (2 work-weeks)	2,70,868
Studio recording — ~20 h delivered audio (pass-through)	97,000
Voice production total	3,67,868

Vector Logic records the brand voice itself. This converts what was the single biggest unmitigated blocker in the model plan — voice cloning of a not-yet-chosen speaker, blocked on a human decision, a legal act and money — into work we control.

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

SECTION 4 — WHAT WE WILL PROVE BEFORE WE BUILD AT SCALE (PHASE 0)

The single largest unverified assumption under this entire plan is that the chosen models actually produce good Hindi. No report has tested it. Every quality figure in our research is a vendor claim or a model-card number, not a measurement on our own hardware.

Phase 0 exists to replace those claims with our own measurements, on one rented 48 GB GPU, over one month, before the big development spend.

The hard gates

If a gate fails, the project changes or stops — it does not proceed on hope.

GATE	THRESHOLD	WHAT A FAILURE MEANS
Hinglish code-switch recognition accuracy (before any fine-tune)	worse than 20% word error rate	Stop/redesign. The interface becomes text-first from day one; the voice-first UX is cut
Hindi recognition accuracy (best base, after fine-tune)	worse than 12% word error rate	Change the base model — the chosen one cannot carry Hindi
Fine-tune improvement over the untrained base	under 15% relative gain	The fine-tune has not earned its place; ship the base and re-scope
Voice naturalness (human listeners, ≥30 raters)	below 3.5 on a 1–5 scale	The voice is not trustworthy enough to be the interface; fall back and de-prioritise the brand voice
Voice number/duration intelligibility	below 100%	Hard gate. A mis-spoken price is a money error. Do not ship synthesis on the money path
Booking tool-call accuracy (Hindi/Hinglish)	below 85% exact match	Change the model
Ambiguous-duration inference rate	above 0%	Hard stop on the reasoning model as specified. Inferred-duration pricing is the catastrophic failure
General-capability regression after fine-tuning	above 3%	The fine-tune is over-fitting; retrain with more anchor data
Concurrent sessions per GPU node	below 10	The per-node cost model breaks; re-size the fleet
p95 time to first audio	above 1,500 ms	The avatar is silent too long; "feels broken"

Phase 0 also settles the concurrency number under real load (a 3× error here triples the infrastructure bill) and resolves two contradictory Hindi accuracy figures carried in our research by measuring them ourselves.

What a failure means, commercially

We tell you, and you do not spend the big money. Phase 0 is priced so that a negative result is cheap. If the voice stack cannot be made to work in Hindi at the quality the product needs, you find out at Phase 0 cost — before Phase 1–3 development and before the full model workstream — and we say so in writing.

Cost and duration

LINE	DETAIL	INR
GPU rental — 1× 48 GB dedicated	730 hr @ \$0.59/hr	41,665
Engineering — discovery, benchmark harness, Hindi STT/TTS/LLM evaluation, written feasibility report	3 work-weeks	4,06,301
Phase 0 total		4,47,967

Duration: one month of GPU time plus 3 work-weeks of engineering.

What you receive: a written feasibility report with our own measured numbers and a go/no-go on each model.

SECTION 5 — SCOPE, PHASE BY PHASE

PHASE	WHAT IT DELIVERS	WEEKS	INR (EXCL. GST)
Phase 0	Discovery + Hindi/AI feasibility benchmark — one rented 48 GB GPU, one month	3 work-weeks + 1 month GPU	4,47,967
Model work	Four workstreams: speech recognition, speech synthesis, reasoning LLM, voice production	22 engineering weeks	30,90,471
Phase 1	Voice core + booking engine, audio-only — self-hosted speech in and out, duration-aware booking, payments, provider split, immutable ledger, provider onboarding and hours, PWA shell	12	16,25,205
Phase 2	Avatar + polish — the avatar interaction layer on both sides (rendered on the phone), UX polish, hardening	8	10,83,470
Phase 3	Second vertical / locality on the same platform — catalogue, pricing rules, provider cohort, localisation	8	10,83,470
Deployment + handover	Self-hosted bring-up, CI/CD, monitoring, runbooks, admin training, 30-day warranty	4	5,41,735
ONE-TIME TOTAL (excl. GST)			₹78,72,318

How the phases fit together

- **Phase 1 gate:** an end-to-end real booking by voice, paid, the provider paid on completion, the ledger reconciles, **and two overlapping bookings are correctly rejected by the database, not by application logic.**
- **Phase 2 gate:** the avatar adds under ~800 ms to p95 latency; captions land inside 500–800 ms.
- **Phase 3 gate:** only starts after vertical 1 is per-order profitable **including GPU cost.**

Optional add-on, priced separately (not in the total above)

OPTION	WEEKS	INR
Phase 4 — real-time lip-synced avatar	10 (starting estimate)	13,54,338

Phase 4 is the highest-risk, least-proven scope: true real-time neural lip-sync is server-video-cost territory (~9–11 MB per user-minute against ~9 KB/min for the on-phone avatar). It is priced as an option with a defined starting estimate and **excluded from the one-time total**. It is not a launch dependency.

SECTION 6 — COMMERCIALS

5.0 The demo set, and what each artefact proves

ARTEFACT	WHAT IT DEMONSTRATES	WHAT IT DOES NOT CLAIM
architecture.html	Where every piece runs — the phone, our self-hosted GPU service, and your existing booking backend — and why the avatar renders client-side (no per-frame server cost)	Any measured latency or throughput. Those come from Phase 0.
avatar/high-tier-*.png, low-tier-*.png	One character in two rendering tiers, so a low-end and a high-end phone both get a talking avatar	Photoreal. This is a vector rig, chosen because a photoreal real-time face is not deliverable on a mid-range phone.
avatar/avatar-speaking.mp4	The face audibly speaking Hindi — the real rig, the real synthesised voice, muxed	Lips shaped by the actual phonemes. The mouth is driven by a speech envelope, so it is in time with the speech , not shaped by it .

ARTEFACT	WHAT IT DEMONSTRATES	WHAT IT DOES NOT CLAIM
voice/conversation/*	A three-turn dialogue with a time and a price in it — the two places a mistake becomes a money error	A live model output, or a human recording. It is a scripted demo; both roles are synthesised.
voice/*.mp3	Two Hindi voices measured on the same sentences	The final brand voice. That is produced in the studio session priced in this estimate, which has not run yet.

We have deliberately labelled every one of these with what it is not. A demo that implies more than it shows costs more to correct in Phase 1 than it saves in Phase 0.

6.0 Validity, and the fields we need from you

This pricing is valid for 30 days from 11 October 2026. The reason is commercial, not legal: the monthly figures rest on live cloud GPU and hosting prices read on 11 October 2026, and those move. After 30 days we re-price the affected lines from then-current vendor prices. The engineering rate (USD 35/hr, ₹3,385.84/hr) is the part that does not move — it is our rate, not a vendor's.

Two fields must be filled before this is signature-ready, and we have deliberately left them blank rather than guessing:

FIELD	WHY IT MATTERS
Your legal entity name as it should appear on a contract and an invoice	The address in this document reads "Papparti", which is a brand. An invoice
Who signs, and in what capacity	So acceptance is unambiguous and the right person receives each milesto

We have also used contact@papparti.com only as a placeholder in the demo references — tell us the address that reaches you and we will correct it everywhere before sending anything further.

The full commercial annexure — payment milestones, acceptance criteria per milestone, what happens if a Phase 0 gate fails, invoicing and tax, scope changes, and termination — is a separate document: 04-COMMERCIAL-TERMS.md.

6.1 One-time project fee

	INR
One-time total, excluding GST	₹78,72,318
GST @ 18% (SAC 998314, software development service): ₹78,72,318 × 0.18	₹14,17,017
One-time total, including 18% GST	₹92,89,335

The ₹14,17,017 GST line is arithmetic on the ₹78,72,318 excluding-GST total: ₹78,72,318 × 0.18 = ₹14,17,017 (rounded to the rupee), and ₹78,72,318 + ₹14,17,017 = ₹92,89,335 — the including-GST figure quoted in our cost estimate.

6.2 Monthly after handover

ITEM	INR / MONTH
GPU serving — 1× 48 GB dedicated, 24/7	41,665
Managed Postgres + vector search	5,891
Redis	1,935
Real-time media server (SFU/TURN)	4,837
Object storage	1,161
Monitoring + CI	2,902
Support retainer (0.5 engineering work-week / month)	67,717
MONTHLY TOTAL	₹1,26,108

The support retainer is deliberately larger than the entire GPU bill. A live 24/7 voice system needs ongoing engineering, not just hardware. The GPU line assumes **one** 48 GB card; a production serving tier with real concurrency may need a second (~₹83,000/month).

6.3 What is included

- All six phases in Section 5 (Phase 0 through deployment and handover).
- All four model workstreams, including the brand voice recording session.
- Self-hosted deployment on rented GPU.
- Runbooks, CI/CD, monitoring wiring and admin training.
- A **30-day post-handover warranty**.
- Full handover of code, trained model weights and endpoints.

6.4 What is not included

- **Payment-gateway fees.** These are transaction-linked, not fixed. The gateway charges a flat 2% **platform fee per successful domestic transaction, plus 18% GST on that fee** — no setup, AMC or refund fees. On ₹1,00,000 of settled volume that is ₹2,000 + 18% GST = **₹2,360**. This is not in the monthly total above.
- **GST on imported cloud/GPU services.** Cloud and GPU services bought from foreign providers are an **import of services**. 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. India-hosted GPU providers invoice domestically and avoid this mechanic, at a higher per-hour price.
- **Counsel and CA fees** — the aggregator classification opinion, the payment/GST posture, and the voice-release legal review are obtained by Papparti. Not modelled here.
- **Phase 4** (real-time lip-synced avatar) — a separate ₹13,54,338 option.
- **Native iOS/Android apps** — the product is PWA-first.
- **Languages beyond Hindi and English.**
- **Multi-vertical launch** — Phase 3 adds a second vertical to the same platform.
- **The second 48 GB serving GPU**, if concurrency demands it.

6.5 Payment milestones

Priced per phase, invoiced as each phase completes:

MILESTONE	INR (EXCL. GST)
Phase 0 — feasibility benchmark delivered	4,47,967
Model work — four workstreams	30,90,471
Phase 1 — voice core + booking engine	16,25,205
Phase 2 — avatar + polish	10,83,470
Phase 3 — second vertical / locality	10,83,470
Deployment + handover	5,41,735
Total	78,72,318

6.6 On the budget band

Our cost estimate's verdict is that the expectation of a **₹20–30 lakh** one-time fee is **not defensible** — it is roughly one third of the real cost of this scope. The arithmetic:

- ₹78,72,318 is 3.15× the **₹25 lakh midpoint** of that band.

- The band's **ceiling (₹30 lakh)** would cover only the **model work (₹30,90,471, four workstreams)** and **nothing else at all** — the trained models plus the voice recording they require 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, 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 goes, as a share of the 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%

The band was priced like a marketplace app. It is a marketplace app **plus three custom-trained AI models plus a self-hosted voice stack**.

Quote: ₹78,72,318 excluding GST (₹92,89,335 including 18% GST) for the defined scope, ₹13,54,338 for the Phase 4 option, and ₹1,26,108 per month to run it.

SECTION 7 — WHAT YOU GET AT HANDOVER

DELIVERED ITEM	DETAIL
Source code	The full marketplace software — customer and provider sides, admin, pay
Model weights	All three trained models, including the trained adapters, become
Endpoints	Self-hosted serving endpoints for speech recognition, speech synthesis an
Infrastructure-as-code	The self-hosted stack defined as code, on rented GPU
CI/CD	Build and deployment pipelines
Monitoring	Wired monitoring for the live service
Runbooks	Operational documentation for running the stack
Admin training	Training for your team on the admin console and operations
Knowledge transfer	Structured handover of how the system works and how to run it
30-day warranty	Post-handover defect warranty
Voice-rights release	A perpetual, irrevocable, worldwide, transferable commercial release sign

Ownership: on handover the models and the code are yours. You are not renting access to them. This is the "100% first-party development" promise made legible: you buy the asset.

SECTION 8 — RISKS, STATED PLAINLY

8.1 Risks with no mitigation

These are stated honestly. None of them has a mitigation that removes the risk.

1. Hinglish accuracy may not be solvable to product grade with this method.

Our research's evidence is that even dedicated fine-tuning leaves code-switched word error around 10–14%, and the problem is *not solved*. There is **no mitigation that makes Hinglish speech recognition as reliable as monolingual recognition**. The only real mitigations are product design (confirm in text) and scope (accept the residual error). If voice-perfect Hinglish is a requirement, that is not something this project can deliver, and we are saying so now rather than later.

2. If you supply no real audio and no real transcripts, the real-world accuracy of the models cannot be stated.

The synthetic-to-real gap can only be measured against real held-out data. If none exists, we ship models whose real-world accuracy we **cannot state**. This is a data-availability limitation, not a technical one — no amount of engineering removes it.

3. Catastrophic forgetting can be bounded but never eliminated.

A fine-tune **will** shift the base model's behaviour; general capability can regress. We can keep the regression under 3% and measure it before and after, but we cannot guarantee the model keeps every general capability it started with. The honest position is "bounded and measured," not "none."

4. No listening test has been run, and no MOS (mean opinion score) exists for our voice output.

We have generated audio and measured it objectively — duration, peak level, RMS, the proportion of non-silent frames, and the real-time factor on this hardware — but **no human listener has rated it, and we quote no quality score**. Any model-card quality figure you may have seen elsewhere is not a substitute for a listening test on your own script and your own brand voice. That test is part of Phase 1, and we will run it before the voice is finalised.

5. The two conflicting Hindi accuracy figures in our research cannot be reconciled from the documentation.

They can only be replaced by our own measurement. Until Phase 0 runs, any Hindi number quoted from a model card is unreliable, and that uncertainty is irreducible until we spend the measurement.

8.2 Mitigated risks

RISK	SEVERITY	MITIGATION
Hinglish speech recognition is genuinely hard	High	Fine-tune on real and synthetic code-switched data; Phase 0 measures the pre/post gap; if it fails, ship text-first
Base Hindi quality is unverified	High	This is what Phase 0 exists for. No fleet spend before the measurement
If Hinglish cannot be fixed, the voice-first UX is unsound	High	Confirm-before-price in text, always. The voice cost becomes a UX cost, not a money-safety cost
Catastrophic forgetting	Medium	Mix general Hindi anchor data into training; before/after capability probe; regression gated under 3%
Voice-cloning consent and legal exposure	High	A signed voice release is mandatory; we clone only the chosen speaker; the synthesis model watermarks its output
Client data too small	Medium	Synthetic-first bootstrapping; the synthetic-to-real gap is measured, not assumed
Share-alike data entering shipped weights	Medium	Prefer permissively-licensed corpora; counsel confirms before any share-alike data trains the shipped adapter; per-source training ledger kept
A base model's fine-tune licence differs from its base licence	Medium	Confirm the licence before shipping; otherwise fine-tune from a cleanly-licensed base ourselves
Synthetic data teaches synthetic artefacts	Medium	Real code-mixed phrasing; scenario-disjoint splits; a real held-out slice synthetic tuning never touches
Number/duration mispronunciation in synthesis	High	All numbers spelled as words in the frozen eval; 100% intelligibility gate ; numeric-safety grammar check before synthesis
Synthesis prosody fails under interruption	Medium	Clause-level chunking plus an interruption-naturalness test

RISK	SEVERITY	MITIGATION
Over-fitting the evaluation	Medium	Test sets frozen before training and stored read-only; independent adversarial verification each phase
Statistical evaluation too small	Medium	Minimum listener panel of 30+; word-error sets of 5+ hours audio; confidence intervals reported, not point estimates

8.3 Commercial risks you should weigh

- **Self-hosting is not cheaper.** Our research states it plainly: at plausible volumes, hosted cloud speech would cost less than running your own GPU tier. Break-even needs millions of bookings per month once an engineer is counted. Self-hosting is a strategic/compliance decision, not a cost saving.
- **The commission take-rate is under threat.** Per-booking commission is table stakes and is under attack by zero-commission competitors in the Indian market.
- **The payment gateway fee sets a floor on your take rate.** At a 2% platform fee plus 18% GST on that fee, 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 this before fixing the take rate.
- **"Any service" is the trap.** Our research measured this: a broad any-service launch is the failure mode of this sector in India. A bounded first vertical is the sane MVP, and Phase 3 extends the platform to a second vertical deliberately.

8.3 The talking avatar — what you are seeing, and what you are not

Two rendering tiers of one character ship with this proposal, as four images:

FILE	TIER	STATE	WHY IT EXISTS
avatar/tierA-idle.png	High-configuration	Resting	What the avatar looks like on a capable phone
avatar/tierA-talking.png	High-configuration	Mid-utterance	The same face while speaking
avatar/tierB-idle.png	Low-configuration	Resting	The same character on a budget phone
avatar/tierB-talking.png	Low-configuration	Mid-utterance	Speaking, on a budget phone

What these are. The avatar is a **layered vector rig that animates on the client device** — not a rendered video and not a photoreal face. That is a deliberate choice, and it matters commercially: a rig ships as a few hundred kilobytes, animates at full frame rate on a mid-range phone with no GPU, needs no server round-trip per frame, and costs you nothing per conversation. A photoreal animated face cannot do any of those things, and **we are not going to show you a photoreal demo and then deliver a rig.**

What these are not. They are **not** a photoreal likeness, and they are **not** a finished art direction. They demonstrate two things only: (1) the same character can be rendered at two cost tiers so a budget-tier phone does not get a degraded experience, just a simpler one; and (2) the rig is built so the face actually redraws while speaking rather than being a still image that toggles.

What is measured and what is not. The tier difference is a measured fact, not an impression: a straight-line read of the skin's colour across the cheek returns **30 distinct tonal steps spanning 131 of 255** on the high tier against **1 distinct step spanning 0** on the low tier — a genuine 30× difference, checked by a script that fails the build if it ever collapses. The mouth region redraws by **35.8%** between the idle and talking frames. Horizontal overflow is **0 px** at 1920 / 1440 / 1024 / 390 px widths. **No human being has yet judged the art direction** — that is a decision for your team, and these four images are the input to it.

8.4 The voice samples you can listen to now

Two voice candidates were produced and shipped with this proposal. **They are real model output, not a studio mock-up and not a stock clip** — generated on our own CPU hardware and measured, not described. They are the *tone* reference for the brand voice, not a final deliverable: the shipping voice will be recorded in a studio session with a chosen speaker, which is a priced workstream (Section 6).

FILE	SPEAKER	LENGTH	WHAT TO LISTEN FOR
voice/papparti-voice-pratham.mp3	Male	~5.5 s	A short Hindi greeting — naturalness of the opening
voice/papparti-voice-priyamvada.mp3	Female	~6.8 s	The same greeting — pick a preferred register
voice/papparti-number-pratham.mp3	Male	~6 s	A sentence containing a price and a duration
voice/papparti-number-priyamvada.mp3	Female	~6 s	The same, female

Listen to the two number samples first. They are the harder problem and the one that carries money risk: if a system mishears or mis-speaks a price or a duration, the customer is quoted the wrong amount and the booking is wrong. Our own measurement found the number samples carry **lower speech energy than the prose samples** (–16.2 dBFS against –15.4 dBFS for the male voice), which is exactly why numerals get their own hard gate in Phase 0 rather than being assumed to work.

voice/README.md records how each file was produced and every number measured on it.

Our recommendation, and the reasoning, so you can disagree with it. Of the two voices, we recommend the **male** voice for the brand-voice recording session. The basis is measured, not a matter of taste: on the sentence that carries a price and a duration — the case where a mis-spoken numeral becomes a money error — the male voice sits further from clipping (–16.2 dBFS against –12.8 dBFS). Both voices run at the same speed, so performance does not decide it. This is a recommendation from a **fallback** engine, not a listening test: **your ear, on your own script, decides** — which is precisely why the studio session is a separate, priced line item rather than something we assume.

SECTION 9 — WHAT WE NEED FROM YOU

#	OBLIGATION	WHY IT BLOCKS	WHEN
1	Voice session scheduling and approvals — make the chosen speaker available for the studio recording, approve casting, approve the master	The brand voice does not exist until this happens; the synthesis model cannot train without it	Before model work completes
2	The service taxonomy — the actual service categories for launch	Drives the speech recognition vocabulary and the reasoning model's structured output. Generic lists do not carry the real booking phrases	Phase 0 onwards
3	The locality names — the launch locality's real place-names and their Hindi pronunciations	Drives the recognition vocabulary and the reasoning model's locality field	Phase 0 onwards
4	Real booking data once running — export transcripts and the confirmed booking records	Closes the synthetic-to-real gap; without it, real-world accuracy cannot be stated	After bookings run
5	Consent and notice handling under DPDP — data-retention and consent notice for stored audio and transcripts	Voice is sensitive personal data. Required before you store real customer audio	Before real data collection
6	A signed voice-rights release	No release = no clone. The release must be in Papparti's favour and irrevocable	Before the voice is cloned

#	OBLIGATION	WHY IT BLOCKS	WHEN
7	Business facts and a real contact email	The current holding page carries no business facts because none have been supplied. Nothing past a holding page is possible without them	Immediately
8	Counsel engagement — the aggregator classification, the payment/GST posture, and the voice-release review	Gates the financial schema and the payment half of Phase 1. The data model must be built migratable from day one	Phase 0
9	Pick ONE service vertical and ONE locality for launch	"Any service" from day one is the measured failure mode of this sector	Phase 0
10	Answer the existing-audio question — do call recordings, voice notes, or any real Hindi/Hinglish audio exist?	If you have zero audio, that is workable but must be known; it changes Phase 0's test design	Phase 0

Anything you can supply changes the result for the better. Real audio and real transcripts are the only way the shipped models' real-world accuracy can be stated at all.

SECTION 10 — ASSUMPTIONS

Every assumption below is copied from our cost estimate, with its basis and whether it is citable.

#	ASSUMPTION	VALUE	BASIS	CITABLE?
1	FX rate	1 USD = ₹96.7384, XE mid-market, 11 Oct 2026, 04:05 UTC	Live rate, pinned. All INR figures use it. Re-run if the rate moves materially	Yes — citable
2	Engineering rate	\$35/hour (₹1,35,434 per 40-hour work-week)	India AI-agency premium tier; Clutch lists India at \$25–49/hr; Eucalipse defines an "India (Premium)" tier at \$35/hr	Published bands — citable; the rate itself is our internal assumption
3	Work-week	40 hours, 5 days	Standard	Yes
4	Phases priced at midpoint	Phase 1 at 12 of 10–14 weeks; Phases 2–3 at 8 of 6–10 weeks each	Midpoint of the estimated range	Yes — it is a choice, stated
5	GPU hours per month	730 (24×7 dedicated)	24 × 7 × 365/12	Yes
6	Training/serving GPU	48 GB dedicated @ \$0.59/hr (₹41,665/month at 730 hr)	Live vendor price, read 11 Oct 2026. Provider selection is an implementation detail	Yes — citable
7	Model parameter counts	Speech recognition 1–2 B, synthesis 0.3–0.5 B, reasoning ~32 B	Design decision	Yes. Base models confidential
8	Support retainer	0.5 engineering work-week per month	Agency rate, \$6	Yes
9	Monitoring/CI and storage volume	~₹2,902/month; ~500 GB storage → ~₹1,161/month	Estimated, not quoted. The storage <i>rate</i> is live; the <i>volume</i> is assumed	No — assumption, not a quote
10	Payment-gateway fees	Excluded from the fixed monthly total; modelled as variable	Gateway 2% + 18% GST on the fee	Fee structure citable; it is variable, not fixed
11	Training GPU-hour counts	Speech recognition 40, synthesis 24, reasoning 180 GPU-hours	Not a vendor quote. Anchored to published fine-tuning benchmarks; the rate is live, the hours are estimated. The cost is robust anyway — GPU compute is ~0.5% of each workstream	No — estimate, not a quote
12	Studio recording cost	~₹97,000 for ~20 h at ~₹4,850/delivered hour	No studio quote obtained. Modelled band of ₹64,000–₹1,30,000 from India studio rates; midpoint used	No — estimate, not a quote

#	ASSUMPTION	VALUE	BASIS	CITABLE?
13	A second 48 GB serving GPU	Not priced into the totals	Flagged as a possibility if concurrency demands it	No — unpriced possibility
14	Phase 4 real-time lip-synced avatar	₹13,54,338	Starting estimate, not a quote. Excluded from the one-time total	No — estimate

Explicitly unverified, and excluded from every total

The following 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 Section 6 stand independently of them:

- Training GPU-hour counts for the three models (40 / 24 / 180 hours) — not a vendor quote.
- Studio recording cost (~₹97,000 for 20 h) — no studio quote obtained.
- Monitoring + CI monthly cost (~₹2,902) — an assumption, not a quote.
- Object-storage volume (~500 GB → ~₹1,161/month) — the rate is live; the volume is assumed.
- Any second 48 GB serving GPU — flagged as a possibility, not priced.
- Phase 4 real-time lip-synced avatar — open-ended; ₹13.54 lakh is a starting estimate.

What we could not verify

- **A published Indian dev-agency rate card.** No vendor publishes a fixed public rate for AI engineering; the \$35/hour figure rests on published India rate bands.
- **A quoted monthly or annual price from any GPU vendor.** GPU vendors publish per-hour prices only; monthly figures here are per-hour × 730.
- **Vendor-specific latency for India-served inference.** Not measured; this is why the pricier India-hosted GPU option is kept as a contingency, not the default.
- **Concurrency capacity of a single 48 GB card under real load.** Unmeasured; the monthly GPU line assumes single-stream serving. **Phase 0 measures this** — it is one of the hard gates in Section 4.
- **Any Hindi speech quality number from a model card.** Every quality figure our research holds is a vendor claim or a model-card number, not a measurement. **This is precisely what Phase 0 exists to fix**, and it is why we are proposing to measure before we build.

Prepared by Vector Logic, 11 October 2026. All figures traceable to our cited cost estimate. Base model identities are commercially confidential to Vector Logic and are not disclosed in this document.

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)

PREPARED FOR PAPPARTI

Papparti — GPU: Rent or Buy? A decision document

Whether Papparti should rent GPU capacity or buy it, with the arithmetic shown.
For whoever signs off the monthly cost: the break-even and the flip condition.
Asks for: nothing yet. It recommends renting, and states when to revisit.

AT A GLANCE

Recommendation	Rent, at the first instance
Monthly GPU line	41,665, unchanged from the estimate
Break-even, used card	about 21 months of continuous need
Break-even, new card	about 61 months, effectively never here
Flip condition	Buy only if measured use exceeds 60% for 20+ months

DATE

11 October 2026

CONFIDENTIALITY

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

Contents

Papparti — GPU: Rent or Buy? A decision document	3
0. The one-paragraph answer	3
1. The break-even calculation, shown	3
1.1 The two cost functions	3
1.2 The three buy options (each sourced in §2)	4
1.3 The arithmetic	4
2. Live buy-side prices (each with source + read date)	5
2a. New RTX 6000 Ada / A6000-class 48 GB card	5
2b. Used 48 GB / 24 GB class card	5
2c. Server / workstation to host it	6
2d. Colocation for a self-owned card in India — the piece people forget	7
2e. Electricity — Odisha	7
3. The case AGAINST buying — stated as hard as the case for it	8
3.1 The capital is committed before there is one paying customer	8
3.2 The technology-obsolescence risk is real and dated	8
3.3 The bought card's value can <i>fall faster</i> than the amortisation assumes — proven this year	8
3.4 The client is a marketplace company, not an infrastructure company	8
3.5 The bought card that is idle still cost the full price	9
3.6 It adds a compliance and logistics surface	9
4. Recommendation, with the conditions attached	9
5. The sensitivity that matters most — utilisation	10
6. What this means for the proposal	11
7. Verification and what could not be verified	12
8. Assumptions register	13

Papparti — GPU: Rent or Buy? A decision document

Prepared by: Vector Logic · **Date:** 11 October 2026 · **Status:** decision document (round 0) **Reads first, does not re-litigate:** CLIENT-ESTIMATE.md, MODEL-DESIGN.md. **Question:** should the client **rent** the 48 GB GPU monthly (the estimate's assumption) or **buy** it?

Every number below carries a source URL and the date it was read, or is labelled **ASSUMPTION** with its basis. All USD is converted at the estimate's pinned rate, **1 USD = ₹96.7384** [FX] — re-verified live today at 96.7372 (exa.ai, 2026-10-11), so **the estimate's rate is correct to 0.001%** and is used unchanged. Indian figures use Indian grouping (₹78,72,318, not ₹7,872,318).

0. THE ONE-PARAGRAPH ANSWER

Rent. Keep renting. At the client's actual expected utilisation the rented card is the cheaper and the lower-risk choice for at least the first **~20 months**, and the client will not know its own utilisation until Phase 0B has measured it. **The estimate's ₹41,665/month is correct and should NOT change.** The only honest adjustment is to *state the flip condition* in the proposal: if measured utilisation holds above **~60%** of a dedicated card for more than **~20 months**, buying a used 48 GB card and colocating it in India becomes cheaper than renting — but that is a decision to revisit **after** Phase 0B produces the utilisation number, never before. Unverified-by-design: a bought card commits capital before there is a single paying customer, and no marketplace has ever been lost by renting its first GPU. **(These are ASSUMPTIONS about the client's stage, flagged as such, and they are the crux of the recommendation.)**

1. THE BREAK-EVEN CALCULATION, SHOWN

1.1 The two cost functions

- **Rented** (billed hourly, 730 h/month, busy or idle):

$$C_{\text{rent}}(n) = 41,665 \cdot n \leftarrow \text{linear, no capex.}$$

- **Bought** (one capex, then monthly opex that is NOT zero — a bought card needs a home):

$$C_{\text{buy}}(n) = \text{CAPEX} + \text{OPEX} \cdot n \leftarrow \text{a step, then a shallow slope.}$$

Break-even is where they cross:

$$\text{CAPEX} + \text{OPEX} \cdot n = \text{RENT} \cdot n$$

$$\text{CAPEX} = (\text{RENT} - \text{OPEX}) \cdot n$$

$$n^* = \text{CAPEX} / (\text{RENT} - \text{OPEX}) \quad \text{months}$$

n^* is the month in which buying's cumulative cost overtakes renting's. Below n^* renting is cheaper; above it, buying is.

1.2 The three buy options (each sourced in §2)

OPTION	GPU	HOST	CAPEX (INR)	MONTHLY OPEX (INR)	BASIS
A — New card, new server	New RTX 6000 Ada 48 GB @ \$8,400	New 2U GPU server @ ~\$13,000	20,70,202	7,716	apex build; lowest risk, highest cost
B — Used card, India OEM server	Used RTX A6000 48 GB @ \$3,999	India entry GPU 2U @ ~₹6,00,000	9,86,828	12,716	mid; INR invoice, GST credit, onsite SLA
C — Used card, refurb server	Used RTX A6000 48 GB @ \$3,999	Refurb 2U server @ ~\$3,500	7,25,441	7,716	cheapest realistic; most execution risk

OPEX = colocation plan + metered power for the node. Per §2d, a 2U colo slot with a power allowance is **₹5,000/month** (VyomCloud) and a dedicated-power 2U is **₹10,000/month**; per §2e a GPU node drawing ~600 W (300 W card + host) uses **438 kWh/month**, which at Odisha LT-industrial **₹6.20/kWh** is **₹2,716/month**.

Option B uses the ₹10,000 colo (power billed inside the plan allowance); A and C use ₹5,000 + metered ₹2,716. **ASSUMPTION:** node draw 600 W — the card's 300 W TDP [spec] plus a ~300 W host; to be confirmed by a power meter in Phase 0B.

1.3 The arithmetic

$$\text{Option A: } 20,70,202 / (41,665 - 7,716) = 20,70,202 / 33,949 = 61.0 \text{ months}$$

$$\text{Option B: } 9,86,828 / (41,665 - 12,716) = 9,86,828 / 28,949 = 34.1 \text{ months}$$

$$\text{Option C: } 7,25,441 / (41,665 - 7,716) = 7,25,441 / 33,949 = 21.4 \text{ months}$$

Sanity check by the month (Option C):

MONTH	RENTED CUMULATIVE	BOUGHT CUMULATIVE	CHEAPER
12	4,99,980	8,18,033	rent by ₹3,18,053
20	8,33,300	8,79,761	rent by ₹46,461
21	8,74,965	8,87,477	rent by ₹12,512
22	9,16,630	8,95,193	buy by ₹21,437 ← crossover
36	14,99,940	10,03,217	buy by ₹4,96,723

Break-even: ~20–21 months (Option C), ~34 months (Option B), ~61 months (Option A).

36-month total, all-in: rent ₹14,99,940 · buy C ₹10,03,217 · buy B ₹14,44,676 · buy A ₹23,48,050. Even over three years, the *new-card* build costs **₹8.48 lakh more** than renting.

2. LIVE BUY-SIDE PRICES (EACH WITH SOURCE + READ DATE)

All read **11 October 2026** unless the source states its own check date, which is given.

2a. New RTX 6000 Ada / A6000-class 48 GB card

ITEM	PRICE	SOURCE	READ
RTX 6000 Ada 48 GB (new, retail)	\$8,400	aigearwatch.com/hardware/nvidia-rtx-6000-ada/ ("recorded catalog price \$8,400", checked 2026-10-10)	2026-10-11
RTX 6000 Ada 48 GB (lowest avg)	\$7,715 (range \$5,333–\$9,070)	gpupoet.com/gpu/learn/price/october-2026/nvidia-rtx-6000-ada-generation	2026-10-11
RTX 6000 Ada 48 GB (retail, in stock)	\$8,299	computizer.com/products/nvidia-rtx-6000-ada-generation-48gb-gddr6-workstations-graphics-card	2026-10-11
RTX A6000 48 GB (brand-new sealed)	\$3,899.95	eBay listing 186691900885 (ebay.com/p/27046130288) — new stock still exists via marketplace	2026-10-11

Modelling choice: \$8,400 (the most conservative, in-stock retail new-card price) for Option A.

2b. Used 48 GB / 24 GB class card

ITEM	PRICE	SOURCE	READ
Used RTX A6000 48 GB	\$3,999 (going rate \$4,349; median ask \$5,183)	rigprice.com/gpu/rtx-a6000 and vramglass.com/gpu/rtx-a6000	2026-10-11 / undated
Used RTX A6000 48 GB (eBay pre-owned band)	\$3,824 / \$3,899 / \$3,999	www.ebay.com/shop/rtx-a6000-48gb	2026-10-11
Used RTX A6000 48 GB (ex-render-farm)	\$3,500 typical	theaibench.ai/hardware/rtx-a6000-used	2026-07 (Jul 2026)

ITEM	PRICE	SOURCE	READ
Used RTX 3090 24 GB	\$1,274–\$1,450 (fair range \$1,200–\$1,300; RigPrice going rate \$1,450)	resaleprices.com/gpu/nvidia-rtx-3090 (Aug 22 2026) and rigprice.com/gpu/rtx-3090	2026-10-11
Used RTX 3090 24 GB — superseded	\$650–750	bestgpuforai.com/articles/best-used-gpu-for-ai	STALE / CONTRADICTED

Do not use the 24 GB cards for this project's single-node plan. Two RTX 3090s (2 × 24 GB) would give 48 GB but no NVLink-supported 48 GB tensor-parallel layout is guaranteed to hold the ~20 GB Q4 reasoning model *plus* STT *plus* TTS *plus* vLLM KV-cache headroom at the 10-session floor; **MODEL - DESIGN .md §4.4** gates on **sessions/node < 10 breaking the cost model**. A single 48 GB card is the safe floor, so the used 48 GB A6000 at ~\$4,000 is the realistic buy.^{ramageddon}

TO CONFIRM — AWAITING PAPPARTI

[^ramageddon]

The RTX 3090 resale price nearly **doubled** during 2026 (from ~\$650 in March to ~\$1,300–1,450 by September) on an AI-demand memory crunch (gamesreviews.com/news/09/ramageddon-turns-old-rtx-3090-gpus-into-1300-resale-gold-mines, 2026-09-04). **The used market is itself volatile — see the case against buying, §3.3.**

2c. Server / workstation to host it

ITEM	PRICE	SOURCE	READ
Puget 2U GPU server, "starting at"	\$12,415.89 (config \$16,598.74)	www.pugetsystems.com/products/rackmount-servers/2u/	2026-10-11 (page modified 2026-07-30)
Puget 2U, lower tiers starting	\$9,568–\$16,475	same	2026-10-11
Gigabyte G293-S42-AAP1 2U 8-GPU barebone	\$9,414 (as-built \$16,640)	www.coltec.com/en-us/servers/ai-line-ai-e208n-2u-dual-xeon-server-8x-gpu-slots	2026-10-11
India GPU server, entry 1-GPU	~ ₹5,00,000–₹25,00,000 (indicative band)	proactive.co.in/blog-details/what-a-gpu-server-really-costs	2026-10-11
India OEM 2U GPU server (RDP Technologies, Hyderabad)	"Pricing on request" — not quotable	rdp.in/gpu-mart/product/quasar-2x-rtx-pro-6000-blackwell-gpu-server/	2026-10-11

Modelling choice: Option A uses \$13,000 (mid Puget 2U config); Option B uses ₹6,00,000 (bottom of the India entry band — GST invoiced, onsite SLA); Option C uses \$3,500 (a used/refurb 2U node, **ASSUMPTION** since no refurb 2U listing was

found with a live price — see §5).

2d. Colocation for a self-owned card in India — the piece people forget

PLAN	PRICE	POWER INCLUDED	SOURCE	READ
1U colo (Tier III, India)	₹3,999/mo	350 W A+B	xenaxcloud.com/blog/blog/full-rack-colocation-pricing/	updated 2026-09-02
2U colo, VyomCloud (Tier III)	₹5,000/mo	1 A @ 230 V $\approx$ 230 W	vyomcloud.com/1u-rack	2026-10-11
6U colo, VyomCloud	₹10,000/mo	2 A @ 230 V $\approx$ 460 W	same	2026-10-11
12U colo, VyomCloud	₹20,000/mo	4 A A+B $\approx$ 920 W	same	2026-10-11
India colo, Mumbai (per kW)	₹7,500–9,500 /kW/mo	—	sovereignglobalcompute.com/india (week of 2026-07-13)	2026-10-11
India colo, Chennai (per kW)	₹6,800–8,800 /kW/mo	—	same	2026-10-11
Caution	"₹5,000 2U" gives only $\sim$ 230 W; a GPU node needs $\geq$ 690 W $\rightarrow$ the client must buy $\geq$ 4 A / a 6U tier , or the ₹10,000 metered-power tier	—	derived from VyomCloud amperage	2026-10-11

Modelling choice: ₹5,000 (with metered power billed separately) and ₹10,000 (power inside allowance) are both shown; the price is small either way relative to the ₹41,665 rent, so it does not move the break-even much — which is the point: **colo is not the deciding cost; capex is.**

2e. Electricity — Odisha

ITEM	VALUE	SOURCE	READ
Odisha LT-Industrial energy charge	₹6.20/kWh (flat)	OERC RST FY 2026-27, effective 1 Apr 2026, www.orierc.org/CuteSoft_Client/writereaddata/upload/DISCOMs_Tariff_Notification_FY_2026-27.PDF	2026-10-11
Confirm (industrial LT $\approx$ ₹6.20/kWh; tariffs unchanged 5th year)	₹6.20/kWh	mercomindia.com/odisha-retains-last-years-power-tariffs-for-fy-2027 and mercomindia.com/odisha-regulator-retains-existing-power-tariffs-for-fy-2026	2026-10-11
TPNODL/TPSODL LT-Industrial (cross-check)	₹6.00/kWh + ₹60/kW/mo fixed	electricbilltool.com/tpnodl-unit-rate-calculator	2026-10-11

ITEM	VALUE	SOURCE	READ
Odisha electricity duty	0%	same	2026-10-11

The 219 kWh/month figure in the brief, verified: a **300 W card alone** at 100% for 730 h/month draws $0.300 \text{ kW} \times 730 \text{ h} = 219.0 \text{ kWh/month} \rightarrow$ **₹1,358/month** at ₹6.20/kWh. A real node also runs a host CPU and fans, so the **600 W node figure (438 kWh $\rightarrow$ ₹2,716/month)** is the honest planning number for "electricity for the bought server," at **1.3–2.6% of the monthly rent it replaces.**

3. THE CASE AGAINST BUYING — STATED AS HARD AS THE CASE FOR IT

3.1 The capital is committed before there is one paying customer

Papparti is pre-launch. A ₹7.25–20.70 lakh capex on 11 October 2026 buys a card that sits idle until customers arrive, while a rented card costs ₹41,665 **only in the months the client chooses to run it**. The client can pause a rented instance the week launch slips; it cannot un-buy a GPU. **The rent/capex saving only exists if the workload actually runs for 20+ months — and the client has no bookings yet.** This is the single strongest argument against buying, and nothing in §1 counters it.

3.2 The technology-obsolescence risk is real and dated

The RTX A6000 is **Ampere, launched 2020** [rigprice]. Buying it in 2026 means buying six-year-old silicon. The **RTX PRO 6000 Blackwell (96 GB)** is already the current card at \$0.65/GPU-hr on the rental market (getdeploying, 2026-10-01). A 2026-bought 48 GB card is worth materially less in 2028, and the client would be locked to 48 GB while the model floor moves. **A rented card can be re-pointed at a newer GPU in minutes; a bought one is a stranded asset the day Blackwell-class weights need more than 48 GB.**

3.3 The bought card's value can *fall faster* than the amortisation assumes — proven this year

The used RTX 3090 **doubled** in 2026 (AI memory crunch, §2b) — which sounds good for resale, but the same volatile market can run the other way, and the RTX A6000's own tracked range is **\$3,900–\$6,000** (vramglass). Amortising a used card over 36 months assumes a resale floor that the market does not guarantee.

3.4 The client is a marketplace company, not an infrastructure company

Owning a GPU means owning a **hardware failure at 2 am** — a dead card, a PSU, a thermal event in a colo the client has never visited. With RunPod the failure is *their* pager and a replacement is a redeploy. The support retainer (₹67,717/month) buys

software support; it does not make Vector Logic an on-call hardware depot. **Renting converts a hardware-operations problem into a line item.**

3.5 The bought card that is idle still cost the full price

The rented card's "waste at low utilisation" is *visible and bounded* (₹41,665/month). The bought card's waste is *invisible*: ₹7–20 lakh already spent, depreciating, whether the reasoning model serves one user or one thousand. **Sunk cost is harder to manage than a subscription.**

3.6 It adds a compliance and logistics surface

A self-owned card in an Indian colo brings GST input-credit mechanics, an import/ownership trail on the hardware, a physical asset on the balance sheet, and insurance/maintenance to manage. A rented instance has none of that. (The estimate already handles GST reverse-charge for *imported rented* services, §7c — buying moves that complexity into the client's fixed assets instead of removing it.)

Balanced against this, the honest case FOR buying: over 36 months, Option C saves **₹4.97 lakh** vs renting; the capex is small against the ₹78.72 lakh one-time build; the used-card price is currently *attractive*; and an owned card removes per-hour billing and gives the client a physical asset with resale value. **None of these fire before the 20-month break-even, and none of them fire before the client has customers.**

4. RECOMMENDATION, WITH THE CONDITIONS ATTACHED

NOTE

RECOMMEND: RENT. Leave **CLIENT-ESTIMATE.md** at ₹41,665/month unchanged.

The conditions that flip the answer — quote these in the proposal:

1. **Expected life below ~20 months** → **always rent.** Above ~20 months of *continuous need*, buying a

used 48 GB card (Option C) becomes cheaper. Above ~34 months, even the India-OEM build (B) wins. The new-card build (A) needs ~61 months — effectively never for this project.

1. **Buy only if measured utilisation exceeds ~60% of a dedicated card** (≈438 of 730 h/month actually

driving the GPU) **AND** that level is held for **20+ months**. Below 60%, the amortised cost per *used* GPU-hour (\$5) stays above the rental and buying is irrational.

1. **Buy only if one of the three models must run 24/7 with no cloud dependency** (data-sovereignty /

DPDP constraint). The reasoning model running 24/7 is *a* reason, but it is a reason to rent a *dedicated* 24/7 instance, not to buy — the rental is already 24/7 dedicated.

1. **Never buy before Phase 0B reports.** MODEL-DESIGN.md §4.4 makes sessions/node and the

Hinglish-WER gate hard stop/go conditions; if the stack fails its gate, a bought card is a stranded asset on a product that changed shape. **Buying is a post-Phase-0B decision, at the earliest.**

1. **If the client insists on owning capacity, buy Option C, not A** — the used 48 GB card at ~\$4,000 is

~½ the price of new for the same 48 GB, and the new-card premium buys a brake the break-even never reaches.

5. THE SENSITIVITY THAT MATTERS MOST — UTILISATION

The rented card is billed **730 h/month whether busy or idle**. The bought card's *effective* cost per hour **falls** as utilisation rises. This is the whole argument, and it only cuts one way above a threshold.

Cost per **utilized** GPU-hour — Option C (₹7,25,441 capex amortised over 36 months (₹20,151/mo) + ₹7,716 opex = **₹27,867/month** fixed) vs rent (₹41,665/month):

AVG UTILISATION	USED H/MONTH	RENT ₹ /USED-H	BUY ₹ /USED-H	CHEAPER PER USED-HOUR
10%	73	₹571	₹382	buy, but <i>both are terrible</i> ; rent the smallest possible card, or rent on-demand by the minute
30%	219	₹190	₹127	buy — but only if committed 20+ months
50%	365	₹114	₹76	buy — clearly
100%	730	₹57	₹38	buy — by 33%

Read the table carefully: at every utilisation the bought card has a lower cost-per-used-hour — because the rental price (\$0.59/hr) is set *above* the amortised owned cost (\$38–382/hr-equivalent) across the whole range. Utilisation does **not** decide rent-vs-buy on a *per-hour* basis. What decides it is the **fixed commitment over time**, which is §1: **20+ months**. The 10% row is the trap — a bought card at

10% utilisation still burns the full ₹27,869/month forever, whereas a rented card at 10% can be *switched to a cheaper plan or stopped entirely*, which is the option ownership deletes.

Which case does Papparti fall into? Read MODEL-DESIGN.md:

- The **reasoning LLM runs 24/7** — that is a genuine **~100%-of-the-card baseline** for that one model,

and it is the strongest single fact for owning. (MODEL-DESIGN §0, §3.2: the LLM is "the heavy one".)

- But the **STT and TTS models are bursty** (§1.2–1.3: speech traffic peaks at booking time, idles

between), and MODEL-DESIGN.md §4.4 gates the whole cost model on **sessions/node ≥ 10** — a number that is **explicitly ESTIMATED and unmeasured** ("~14 sessions/node (ESTIMATED)"; "a 3× error here triples the bill"). The concurrency load test (§4.5) exists precisely because the real number is unknown.

- Therefore the client's **actual average utilisation of the card is UNKNOWN until Phase 0B measures it**,

and the estimate's own §10 says the concurrency capacity of a single 48 GB card under load is **UNVERIFIED. Planning a 20-month capex commitment on an unmeasured utilisation number is exactly the mistake this document exists to prevent.**

Plain statement: on the *known* fact (reasoning model 24/7) the card is heavily used and buying *would* pay off over 20+ months. On the *unknown* facts (real concurrency, whether the stack even passes Phase 0B) it is unsafe to commit capex now. **The uncertainty itself is the reason to rent until Phase 0B removes it.**

6. WHAT THIS MEANS FOR THE PROPOSAL

The recommendation equals the current assumption. Nothing in the money changes.

- **CLIENT-ESTIMATE.md** line that stays as-is: §1b / §5, the monthly row

| GPU serving — 1× 48 GB dedicated | 41,665 | RunPod RTX A6000 @ \$0.59/hr × 730 hr [1] |. **It does not change. ₹41,665/month is confirmed correct.**

- **The one-time total is NOT affected** — this is a monthly item; the ₹78,72,318 / ₹92,89,335 totals are

untouched. (The Phase 0 line's GPU sub-charge, ₹41,665, is likewise unchanged.)

- **Suggested addition (wording only, no arithmetic):** add one sentence under §5 or §9 noting that

renting is the deliberate choice, with a documented flip condition — if Phase 0B measures sustained utilisation above ~60% of a dedicated card for a planned life above ~20 months, a bought used 48 GB card in Indian colocation becomes cheaper and should be revisited at that point. This converts an unexamined assumption into a **stated, testable decision rule** — which is worth more to the client than the rupees.

If the client nonetheless elects to buy, the line changes as follows (monthly): GPU serving ₹41,665 → **~₹7,716** (colo ₹5,000 + power ₹2,716, Option C), i.e. a monthly saving of **~₹33,949/month**; the monthly total ₹1,26,108 → **~₹92,161**; but the one-time total rises by the **₹7,25,441 capex** (→ ₹85.98 lakh excl. GST) — which is the trade the client is being asked to accept. **This is the flip the recommendation says NOT to take yet.**

7. VERIFICATION AND WHAT COULD NOT BE VERIFIED

Recomputed: every break-even quotient in §1.3, the month-by-month table, the 36-month totals, the per-used-hour table in §5, the 219 kWh and 438 kWh electricity figures, and the FX conversion. FX used: **96.7384**, matching the estimate exactly; live cross-check 96.7372 (exa.ai, 2026-10-11). All pass.

Could NOT verify / UNVERIFIED:

1. **No live price for a used/refurb 2U GPU server** — Option C's \$3,500 host is an **ASSUMPTION**; no

listing with a quotable price was found. Option B uses the bottom of the India *indicative band*, and RDP (the India OEM that sells exactly this) lists "**Pricing on request**" only.

1. **No quoted monthly colo price from a GPU-certified Indian facility.**

VyomCloud (₹5,000/₹10,000) and

Xenax (₹3,999) are Tier III **general** colo, not GPU-certified; Sify's DGX-ready GPU colo is **pay-per-use, price on request** (sifytechnologies.com, 2025-05-20). The GPU-colo real price is **UNVERIFIED** — and could be *higher* than the ₹5,000–10,000 assumed, which would *raise* the break-even and strengthen the rent recommendation.

1. **Node power draw (600 W)** is an **ASSUMPTION** (card TDP + host); a Phase 0B power meter would confirm it.
2. **The Odisha figure is the LT-Industrial energy charge (₹6.20/kWh).** A colo facility bills the client

inside its plan (power included), so the ₹6.20 applies only to the on-premises variant; treating it as the colo power price is an **ASSUMPTION**. Confirmed source: OERC RST FY 2026-27 + two Mercom reports.

1. **The client's actual future utilisation is unknowable now** — this is the honest core of the document,

not a gap to be papered over.

1. **Import duty / customs on a card bought from a US/China marketplace** was **not** modelled; landed cost

could be ~18–30% above the \$3,999–8,400 list once duty + freight + IGST are added. **Not priced — flag before any buy.**

8. ASSUMPTIONS REGISTER

#	ASSUMPTION	BASIS
A1	FX 1 USD = ₹96.7384	estimate's pinned rate [FX]; re-verified 96.7372 on 2026-10-11
A2	Rent ₹41,665/mo = \$0.59/hr × 730 × 96.7384	CLIENT-ESTIMATE.md §2 [1]
A3	Node draw 600 W → 438 kWh/mo	card 300 W TDP + ~300 W host — ASSUMPTION
A4	Odisha LT-industrial ₹6.20/kWh, duty 0%	OERC RST FY 2026-27; Mercom 2026
A5	Amortisation life 36 months	ASSUMPTION for per-used-hour illustration only; break-even uses no life assumption
A6	Used-card host \$3,500	ASSUMPTION — no live listing found
A7	India entry server ₹6,00,000	bottom of Proactive's indicative ₹5–25 lakh band
A8	Landed cost = list price (no duty modelled)	ASSUMPTION — understates buy cost; flagged in §7.6

[FX] [www.xe.com/currencyconverter/convert/?](https://www.xe.com/currencyconverter/convert/?Amount=1&From=USD&To=INR)

Amount=1&From=USD&To=INR — XE USD/INR mid-market, 11 Oct 2026 (estimate's source); re-verified at exa.ai/library/markets/forex/USDINR, 2026-10-11T08:16Z (96.737212).

PREPARED FOR PAPPARTI

Papparti — Commercial Terms (Annexure to the Project Proposal)

The commercial annexure: what gets paid, when, and what makes a milestone accepted.

For legal and finance review: milestones, acceptance, tax, termination, voice rights.

Asks for: the blank fields listed in the open-items register.

AT A GLANCE

Payment	Six milestones, each tied to an accepted artefact
Advance	10 percent, recovered against the model-work milestone
Retention	5 percent, released at handover
Phase 0 acceptance	On measured benchmark results, not on opinion
Blank fields	Left blank on purpose, never guessed

DATE

11 October 2026

VALIDITY

30 days; expires 10 November 2026.

CONFIDENTIALITY

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

Contents

Papparti — Commercial Terms (Annexure to the Project Proposal)	4
1. Validity	4
2. Payment schedule	4
2.1 The schedule	4
2.2 How the advance works	5
2.3 Retention / holdback	5
2.4 Why a mobilisation advance, and where in the industry range we sit	5
2.5 Milestones are tied to accepted artefacts, not time	6
2.6 GST on each invoice	6
2.7 Late payment	6
3. Acceptance criteria per milestone	6
3.1 Milestone 2 — Phase 0 (feasibility benchmark)	6
3.2 Milestone 3 — Model work (four workstreams)	7
3.3 Milestone 4 — Phase 1 (voice core + booking engine)	7
3.4 Milestone 5 — Phase 2 (avatar + polish)	7
3.5 Milestone 6 — Phase 3 (second vertical / locality)	7
3.6 Milestone 7 — Deployment + handover	8
3.7 Handover acceptance certificate	8
3.8 Acceptance deemed given	8
3.9 What happens if a Phase 0 gate FAILS — the critical clause	8
4. Invoicing and tax	9
4.1 The Figures Invoiced	9
4.2 GSTIN	9
4.3 Place of supply	9
4.4 HSN / SAC codes — UNVERIFIED	9
4.5 TDS	10
4.6 Invoice mechanics Vector Logic requires	10
5. Scope-change mechanism	10
5.1 How a change is raised and priced	10
5.2 The absorption threshold	10
5.3 What is not a change	11
6. What triggers the monthly charge	11
6.1 The gap between handover and the first paying customer	11
7. Termination	12
7.1 If the client stops	12
7.2 If Vector Logic cannot deliver	12
7.3 Trained model artefacts on termination	12
7.4 Termination for convenience	13
8. The voice-rights release requirement	13
8.1 What Vector Logic does if no release is provided	13
9. Liability and warranties	14
9.1 The accuracy limitations disclosed in the proposal survive into the contract	14
9.2 What Vector Logic does warrant	14
9.3 What Vector Logic does not warrant	14
9.4 Bounded liability	15

10. Open fields register	15
11. Verification	16

Papparti — Commercial Terms (Annexure to the Project Proposal)

Project: Papparti — on-demand any-service marketplace with a self-hosted voice and avatar layer **Prepared by:** Vector Logic — AI & software development, Odisha, India **Prepared for:** Papparti (client) **Date:** 11 October 2026 (IST) **Status:** Annexure to CLIENT-PROPOSAL.md. To be read together with it and with CLIENT-ESTIMATE.md.

This annexure is a commercial document, not legal advice. It records the commercial terms Vector Logic proposes, in plain language, so that both parties and their advisers can review the same numbers. Every clause below is subject to the written contract that follows signature, and the parties should each have their own counsel and chartered accountant review it before signing. Nothing in this document creates a legal obligation on its own.

Every figure in this annexure is traceable to CLIENT-ESTIMATE.md, or is arithmetic on those figures which is shown where it is used. No figure has been invented for this annexure. Base AI model identities are commercially confidential to Vector Logic and are deliberately not named anywhere in this document; models are referred to by capability and parameter class only.

1. VALIDITY

This pricing is valid for 30 days from 11 October 2026 — that is, until 10 November 2026 (inclusive).

The reason is commercial, not procedural: **the monthly operational figures rest on live cloud GPU and managed-hosting prices that were read from the vendors' own pricing pages on 11 October 2026, and those prices move.** The GPU serving line alone (₹41,665/month) is a live vendor rate of \$0.59/hour converted at a live FX rate of ₹96.7384/USD; both the rate and the exchange rate are re-read from the vendor and the market, not fixed by us. Our own cost estimate already had to correct a previously-cited GPU figure that had gone stale (₹34,056 → ₹41,665, roughly 22% higher). A quote built on live prices has a shelf life, and 30 days is it.

What happens after expiry. If the offer has not been accepted in writing within the validity window, Vector Logic will **re-price the quotation at the vendor and FX prices then current**, and issue a fresh validity window. The re-price applies to the live-priced components (the monthly running costs, and the GPU sub-charge inside Phase 0) and, if the FX rate has moved materially, to the INR value of the USD-billed lines.

What does not move. The **engineering rate is the part that does not move.** The \$35/hour rate (₹3,385.84/hour, ₹1,35,434 per 40-hour work-week, at the pinned rate) is Vector Logic's own commercial rate, not a vendor price, and it holds across re-quotes. The one-time build fee is therefore stable in its engineering component; only the pass-through vendor and FX components are re-priced at expiry.

2. PAYMENT SCHEDULE

The one-time project fee is **₹78,72,318 excluding GST**. It is invoiced in **six milestones**, each tied to a **delivered, accepted artefact** — never to elapsed time. A milestone is invoiced when its artefact is delivered and accepted under §3.

2.1 The schedule

#	MILESTONE	MAPS TO PHASE (FROM CLIENT-ESTIMATE.MD)	INVOICE INR (EXCL. GST)
1	Mobilisation advance (10% of the one-time fee)	Drawn against Phase 0 setup, GPU procurement, the studio booking and Phase 0B benchmark preparation	7,87,232

#	MILESTONE	MAPS TO PHASE (FROM CLIENT-ESTIMATE.MD)	INVOICE INR (EXCL. GST)
2	Phase 0 — feasibility benchmark	Phase 0 (₹4,47,967)	4,47,967
3	Model work — four workstreams	Model work (₹30,90,471) less the ₹7,87,232 advance recovered	23,03,239
4	Phase 1 — voice core + booking engine	Phase 1 (₹16,25,205)	16,25,205
5	Phase 2 — avatar + polish	Phase 2 (₹10,83,470)	10,83,470
6	Phase 3 — second vertical / locality	Phase 3 (₹10,83,470)	10,83,470
7	Deployment + handover	Deployment + handover (₹5,41,735)	5,41,735
Sum of all invoices			78,72,318

Sum check: 7,87,232 + 4,47,967 + 23,03,239 + 16,25,205 + 10,83,470 + 10,83,470 + 5,41,735 = **₹78,72,318** — exactly the one-time fee excluding GST. Verified in §11.

Each of the six phase amounts is invoiced at its estimate value; the only adjustment in the schedule is that the **mobilisation advance (₹7,87,232) is recovered against the Model work milestone**, because that is where the front-loaded spend it funded (studio recording, GPU procurement) is actually consumed. The client therefore pays exactly the phase price for each phase's scope, and pays the advance once, not twice.

2.2 How the advance works

The mobilisation advance is **not an extra line and not a fee** — it is an **early draw against the project**, recovered at the Model work milestone. It exists because this project front-loads real cash out of Vector Logic's pocket before there is a deliverable (see §2.4), and it means the client is never paying for the same work twice. The retention (§2.3) is held separately at the other end of the project.

2.3 Retention / holdback

5% of the one-time fee — ₹3,93,616 — is held back as retention and is released at handover and acceptance. In practice the final handover invoice (milestone 7, ₹5,41,735) is settled **less the retention** (₹5,41,735 – ₹3,93,616 = ₹1,48,119 paid at handover), and the retention of **₹3,93,616** is invoiced and released when the client signs the handover acceptance certificate (§3.7), at which point the 30-day warranty has started.

Retention of ₹3,93,616 = 5% × ₹78,72,318, rounded to the rupee. It is the client's protection: it stays in the client's hands until the delivered system is accepted, and it is a real, bounded amount rather than a token. Because the retention is released inside the same milestone-7 line, the total invoiced across §2.1 is unchanged at ₹78,72,318.

2.4 Why a mobilisation advance, and where in the industry range we sit

The advance is **10% of the one-time fee (₹7,87,232)**, and it is justified from the actual scope, not from a template:

- **This project front-loads real cash out of Vector Logic's pocket before there is a deliverable.**

Phase 0 requires a **rented 48 GB GPU for a full month (₹41,665)** to be procured and running before a single line of the benchmark is written. The **studio voice recording (~₹97,000 of delivered audio, a pass-through cost at the midpoint of a ₹64,000–₹1,30,000 modelled band)** has to be booked, cast, directed and recorded before the synthesis model can train at all. And the **Phase 0B benchmark** is a substantial measurement campaign run before the big development spend. None of these exist as a deliverable until after the money is physically spent.

- **The industry range** for mobilisation or advance payments on fixed-scope software and AI

engagements is commonly **10%–30% of contract value up front**, with 10–15% at the conservative end and 20–30% where the supplier has to buy hardware or lengthy data-collection services at the start. **Vector Logic sits at the bottom of that range, at 10%.** The reason we sit low is deliberate: the phase structure already protects Vector Logic (each milestone is invoiced only after an accepted deliverable), so the project does not need a large advance, and the client retains most of its cash against delivered work.

2.5 Milestones are tied to accepted artefacts, not time

No milestone invoice is raised on a date. Each is raised when the named artefact is **delivered and accepted** under §3, or where a gate resolves under §3.9, when the gate resolution itself is accepted.

2.6 GST on each invoice

GST @ 18% is charged on each invoice, and is the client's **input tax credit where applicable** (where Papparti is GST-registered and the service is used for business — see CLIENT - ESTIMATE .md §7c for the reverse-charge mechanics on the imported cloud lines, which are separate from Vector Logic's own GST). GST is not included in any INR figure in §2; each figure is an excluding-GST figure. The GST-inclusive one-time total is in §4.1.

2.7 Late payment

Stated plainly: **each invoice is due within the payment-days convention fixed in the contract (BLANK FIELD — see §10); if an undisputed invoice is not paid by its due date, Vector Logic may pause work on the following milestone until it is paid, and may charge simple interest on the overdue amount at a rate fixed in the contract.** Stated at the same time and with equal weight: Vector Logic will not pause work over a *disputed* invoice — a disputed line is handled under §5 and the remainder of the work continues. The purpose of the late-payment term is not to penalise the client; it is to protect a small agency from carrying a large project on its own balance sheet while vendor GPU bills and studio costs fall due regardless.

3. ACCEPTANCE CRITERIA PER MILESTONE

Every milestone is accepted against a **testable artefact**. Where Phase 0 is concerned, acceptance is on a **measured number** — the recorded Phase 0B results, including pass/fail against the hard gates in MODEL - DESIGN .md §4.4 — and **not on opinion**. The client should not accept Phase 0 on "it looks good"; acceptance is a table of measurements against thresholds.

3.1 Milestone 2 — Phase 0 (feasibility benchmark)

Artefact delivered: a written Phase 0B feasibility report containing, for each model workstream: (i) the frozen test sets used, (ii) the measured metrics, (iii) the recorded **pass/fail against each hard gate**, (iv) the measured **sessions/node** under the concurrency ramp, and (v) a go/no-go on each of the three models.

Accepted when: the report is delivered and it contains a recorded result against **each** of the ten gates below, taken from MODEL - DESIGN .md §4.4:

GATE	THRESHOLD	ON FAIL
Hinglish code-switch WER (STT base, pre-fine-tune)	worse than 20%	Stop/redesign — text-first from day 1
Hindi WER (best base, post-fine-tune)	worse than 12%	Change base
STT fine-tune gain vs stock	under 15% relative	Ship stock + re-scope
TTS MOS (brand voice, ≥30 raters)	below 3.5	Fall back, de-prioritise brand voice

GATE	THRESHOLD	ON FAIL
TTS number/duration intelligibility	below 100%	Hard gate — do not ship synthesis on the money path
LLM tool-call exact-match (Hindi/Hinglish)	below 85%	Change model
Ambiguous-duration inference rate	above 0%	Hard stop on the reasoning model as specified
General-capability regression after fine-tune	above 3%	Retrain with more anchor data
Sessions/node (concurrency, audio-only)	below 10	Re-size the fleet
p95 time-to-first-audio	above 1,500 ms	Re-size / re-design

Acceptance is the **recorded table of measured values**, not a narrative. If a metric was not measured, it is reported as **NOT MEASURED** and the milestone is **not** accepted on that line.

3.2 Milestone 3 — Model work (four workstreams)

Artefact delivered: the three trained adapters (speech recognition, speech synthesis, reasoning) plus the brand-voice master; each adapter packaged with its serving endpoint configuration, and a per-workstream evaluation report measured against the §3.1 gates.

Accepted when: (i) all three adapters load and serve on the target GPU; (ii) each workstream's post-training metrics meet the gate thresholds agreed at Phase 0 acceptance (or the gate resolution in §3.9 was applied); (iii) the **number/duration intelligibility gate shows 100%** and the **ambiguous-duration inference gate shows 0%** — these two are non-negotiable and cannot be waived by a partial credit or an average; and (iv) the voice master is delivered and the client has approved the master take.

3.3 Milestone 4 — Phase 1 (voice core + booking engine)

Artefact delivered: the audio-only voice core (self-hosted speech in and out), the reasoning endpoint, the duration-aware booking engine (30 min / 1 hr / 5 hr), payments with provider split, the immutable ledger, provider onboarding and working hours, and the PWA shell.

Accepted when (the proposal's own Phase 1 gate, made testable): (i) an **end-to-end real booking is made by voice and paid**; (ii) the **provider is paid on completion**; (iii) the **ledger reconciles** to the rupee against the booking and payment records; and (iv) **two overlapping bookings are correctly rejected by the database, not by application logic** — demonstrated by a test that shows the rejection coming from a database constraint, not from a code check that could be removed.

3.4 Milestone 5 — Phase 2 (avatar + polish)

Artefact delivered: the avatar interaction layer on both sides (rendered on the phone), UX polish, and hardening.

Accepted when: (i) the avatar **adds under ~800 ms to p95 latency**; (ii) **captions land inside 500–800 ms**; and (iii) the avatar on the budget-tier rendering path holds its frame budget per the project's own measured rendering numbers. These are measured on the client's own device band, not on a developer workstation.

3.5 Milestone 6 — Phase 3 (second vertical / locality)

Artefact delivered: a second service vertical or locality live on the same platform — catalogue, pricing rules, provider cohort, localisation.

Accepted when: (i) the second vertical or locality is live and can take real bookings end to end; and (ii) it **only starts after vertical 1 is per-order profitable including GPU cost** — the gate the proposal names — with the profitability measured from the ledger, not asserted.

3.6 Milestone 7 — Deployment + handover

Artefact delivered: infrastructure-as-code for the self-hosted stack on rented GPU, CI/CD, monitoring wiring, runbooks, admin training, and the 30-day warranty start.

Accepted when: (i) the stack is stood up from the infrastructure-as-code by the client's own team, following the runbooks, without Vector Logic hands on the keyboard; (ii) monitoring is wired and reaches the client's chosen alerting destination; (iii) the admin console has been handed over with training delivered; and (iv) the code, model weights and endpoints have been transferred.

3.7 Handover acceptance certificate

Handover acceptance is a **signed acceptance certificate** from the client. On signature the **30-day warranty starts** and the **retention of ₹3,93,616 is released** (§2.3).

3.8 Acceptance deemed given

If the client does not respond to a delivered artefact's acceptance request within the acceptance-review window fixed in the contract (BLANK FIELD — §10), the artefact is **deemed accepted** and the milestone is invoiced. This protects Vector Logic from indefinite open milestones; the window is set generously and is a blank field for the client to agree, not a number we have assumed.

3.9 What happens if a Phase 0 gate FAILS — the critical clause

This is the clause that must be right, and it has two sides that both have to hold:

The client must not owe more money for a phase that failed a go/no-go gate. Vector Logic must not be left unpaid for work already done.

It resolves as follows:

1. **Phase 0's deliverable is the measurement itself, not a passing result.** The Phase 0 milestone

(₹4,47,967) is **payable on delivery of the benchmark report with its recorded results**, whether the results pass or fail. The client is buying the answer to the question, and a measured "no" is a complete, valuable answer — it is exactly what Phase 0 exists to produce. Vector Logic is therefore paid for Phase 0 in full, on delivery.

1. **A failed gate stops the dependant spend.** If a gate returns a **STOP/REDESIGN** or **HARD STOP**

result, then **no dependent milestone proceeds** and **no dependent milestone is invoiced**. The client owes **nothing** for Phase 1, Phase 2, Phase 3, the deployment milestone, or the affected part of the model work. The project stops at the Phase 0 price point. The client never pays for downstream work that a go/no-go gate has just invalidated.

1. **A failed gate that changes the model, not the project, is a change order, not a new invoice.**

Gates whose failure means "*change the model*" (Hindi WER > 12%; LLM tool-call < 85%) or "*change the base*" do not end the project — they are resolved by re-scoping under §5. The re-scope is **priced as a change**, and **the client pays nothing extra for the phase that failed the gate**; the change is quoted and approved before any new money is committed.

1. **A HARD STOP** (TTS number/duration intelligibility below 100%, or ambiguous-duration inference

above 0%) **removes the affected capability from the money path**. The synthesis model is held off the pricing path and the booking flow ships with a conventional, on-screen confirmation and receipt — the proposal's own structural safeguard (§2 of

the proposal). The affected work is **re-scoped downward**, and the client is **credited** the difference between the original line and the re-scoped line on the next invoice. The client does not pay for a capability that cannot safely ship.

1. **Work already done that is unaffected by the failure is paid.** If a gate fails in one workstream

but other workstreams (or software already delivered and accepted) are complete and unaffected, those are invoiced and paid normally. The gate failure reduces **only** the dependant scope.

In one sentence: a failed go/no-go gate never adds a rupee to the client's bill and never leaves Vector Logic unpaid for the measurement it delivered; it removes downstream scope and its downstream invoice, and any partial work continues to be paid only where it survives the gate.

4. INVOICING AND TAX

4.1 The Figures Invoiced

ITEM	INR
One-time fee, excluding GST	78,72,318
GST @ 18%	14,17,017
One-time fee, including 18% GST	92,89,335

Arithmetic, shown: ₹78,72,318 × 0.18 = ₹14,17,017 (to the rupee); ₹78,72,318 + ₹14,17,017 = ₹92,89,335. Verified in §11.

Optional Phase 4 (real-time lip-synced avatar) is a separate **₹13,54,338** option, excluded from the one-time total, and its own GST applies on invoicing.

4.2 GSTIN

Every Vector Logic invoice will carry a real GSTIN. The GSTIN is a field to be filled from the registered entity's records before any invoice is issued:

- ****Vector Logic GSTIN:** _____ (BLANK FIELD — must be filled from the registered

entity's GST registration before the first invoice. Not invented here.)**

Client GSTIN (for tax invoice): _____ (BLANK FIELD — to be supplied by Papparti.)

If either party is not GST-registered, that is recorded and the invoicing follows the applicable non-registered treatment, confirmed with a tax adviser.

4.3 Place of supply

Place of supply is determined for each invoice under the IGST Act, and will be recorded on the invoice. As a services supply to a registered recipient, it is normally the recipient's location; the **exact place of supply and resulting CGST/SGST vs IGST split will be confirmed with a tax adviser and stated on the invoice**. This is not asserted here as a settled tax position.

4.4 HSN / SAC codes — UNVERIFIED

The SAC codes below are marked UNVERIFIED and must be confirmed with a tax adviser before invoicing. They are not asserted as correct in this document.

SERVICE	SAC (UNVERIFIED — CONFIRM WITH TAX ADVISER)
Software development service	998314 (UNVERIFIED) — cited in CLIENT-ESTIMATE.md §7c; re-confirm t
AI / model training service	SAC: _____ (UNVERIFIED — BLANK. The correct SAC for AI/i

CLIENT-ESTIMATE.md §7c cites SAC 998314 for software development services at 18% GST. This annexure records that citation as **UNVERIFIED** for contractual use: the code must be confirmed against the current GST rate notification by the parties' tax adviser before it is printed on an invoice. For the AI/model-training services (the three fine-tuning workstreams), **no SAC is asserted at all** — the field is blank on purpose.

4.5 TDS

Under Indian law, this is a **domestic services contract**, and the client may be required to **deduct tax at source (TDS)** on payments to Vector Logic under the applicable section and rate for services. **The applicable section and rate must be confirmed with a tax adviser** and are not asserted here. Vector Logic's practical requirement: the **TDS deducted must be reflected on the client's TDS certificate (Form 16A) and credited to Vector Logic's PAN**, and the invoice will state the PAN so the credit can be matched. Where TDS applies, the client deducts it and pays the net; Vector Logic claims the credit. The contract should fix the TDS section, rate and certificate timeline explicitly.

4.6 Invoice mechanics Vector Logic requires

- A valid **tax invoice** with GSTIN, place of supply, SAC, invoice number and date.
- **GST charged on the invoice**, as the client's input credit where applicable.
- The invoice is raised on milestone delivery and acceptance, not on a calendar.
- PAN stated on the invoice for TDS matching.

5. SCOPE-CHANGE MECHANISM

5.1 How a change is raised and priced

Any change to the defined scope is raised in writing, described as a discrete deliverable, and priced before any work on it begins. Pricing is arithmetic on the engineering rate, shown to the client:

Engineering rate: \$35/hour = ₹3,385.84/hour = ₹27,087/day = ₹1,35,434 per 40-hour work-week (at the pinned FX rate of ₹96.7384/USD — see CLIENT-ESTIMATE.md §6).

A change order therefore quotes $\text{hours} \times ₹3,385.84$, or $\text{work-weeks} \times ₹1,35,434$, plus any pass-through vendor cost (GPU hours, studio time) at the then-current live price, plus GST.

5.2 The absorption threshold

Changes whose total value is at or below ₹13,542 (i.e. up to 4 engineering hours at ₹3,385.84/hour) are absorbed by Vector Logic at no charge. Changes above that threshold are **quoted as a change order**, approved in writing, and invoiced as delivered.

Why ₹13,542, i.e. 4 hours, is the defensible threshold: it is **half a work-day** — large enough to cover the small, real frictions that inevitably appear in a live build (a field added, a wording changed, a small integration tweak, a review call that turns into a fix) without a commercial conversation, and small enough that a genuine change of direction cannot hide inside it. Below half a day, disputing the change costs both parties more than the change; above half a day, the work is real enough that it must be quoted before it is done. **The threshold is a rate, not a yearly allowance** — it is applied per change, and a change cannot be split into sub-threshold pieces to avoid quoting.

5.3 What is not a change

Defect fixes within the warranty, corrections to delivered work that does not meet its own acceptance criteria, and work needed because Vector Logic mis-estimated its own effort are **not** change orders and are absorbed. Only a change to the agreed scope, or genuinely new scope, is quoted.

6. WHAT TRIGGERS THE MONTHLY CHARGE

The monthly charge of ₹1,26,108 begins only when the system is live and being used — that is, when it is serving real bookings on the client's behalf. It does not begin on handover date, and it does not begin while the system is idle waiting for the client's launch.

MONTHLY LINE	INR / MONTH
GPU serving — 1× 48 GB dedicated, 24/7	41,665
Managed Postgres + vector search	5,891
Redis	1,935
Real-time media server (SFU/TURN)	4,837
Object storage	1,161
Monitoring + CI	2,902
Support retainer (0.5 engineering work-week/month)	67,717
MONTHLY TOTAL	1,26,108

(Figures as in CLIENT - ESTIMATE .md §1b / the proposal's §6.2. The support retainer, ₹67,717, is deliberately larger than the GPU bill; a live 24/7 voice system needs ongoing engineering, not only hardware.)

6.1 The gap between handover and the first paying customer

This is a real commercial question and it has a plain answer:

- **In the gap, the system is not "live and being used" and the monthly charge does not run.** The

support retainer (₹67,717) does not accrue on an unused system.

- **Who pays the GPU in that window: the client, at cost, for the compute it chooses to keep warm.** The

GPU is a live vendor bill that falls due whether or not customers arrive. There are two honest options, and the client picks:

1. **Pause the serving instance** in the gap — a rented GPU can be stopped, and the client pays

nothing for compute while it is stopped. The system is brought up on launch. This is the cheapest and the recommended option where the gap is expected to be more than a few weeks.

1. **Keep a GPU warm** through the gap for pre-launch testing, demos and load rehearsal — in which

case the client pays the **GPU line (₹41,665/month) at cost**, with no support retainer and no margin for Vector Logic, for as long as it chooses to keep it warm.

- **Infrastructure that is already running** (managed Postgres, Redis, object storage, monitoring —

₹11,889/month together, being the non-GPU, non-retainer lines: $5,891 + 1,935 + 4,837 + 1,161 + 2,902 = 16,726$ — see the note below) is either paused or paid at cost, on the same principle as the GPU: the client pays for what it runs and nothing for what it does not.

Note on the arithmetic: the non-GPU, non-retainer monthly lines sum to **₹16,726** ($5,891 + 1,935 + 4,837 + 1,161 + 2,902$), and $₹41,665 + ₹16,726 + ₹67,717 = ₹1,26,108$, confirming the monthly total. In the gap, the only lines a client would pay at cost are the ones it leaves running; the retainer does not run.

In one sentence: the ₹1,26,108/month starts when the system serves real users; in the gap the client pays only for the infrastructure it chooses to keep running, at cost, and the retainer does not accrue.

7. TERMINATION

Stated fairly, in both directions.

7.1 If the client stops

- The client **pays for all milestones delivered and accepted** up to the date of termination, plus

any work in progress on a milestone it has approved, valued at actual engineering hours at ₹3,385.84/hour (or, where a milestone is substantially complete, at the milestone value on acceptance).

- **The mobilisation advance is non-refundable** where it has been spent on the front-loaded costs it

funded (GPU procurement and studio booking); where it has not been spent, the unspent portion is returned.

- The client **takes delivery of all work product paid for** — code, adapters, and the voice master, to

the extent they exist and are paid for.

- Any **vendor commitments already made on the client's behalf and non-cancellable** (a booked studio

session, a committed GPU period) are payable at cost.

7.2 If Vector Logic cannot deliver

- Vector Logic **refunds any advance received for scope not yet performed**, and delivers all work

product produced to that point.

- Vector Logic **transfers the trained model artefacts and the voice master** to the client, to the

extent they exist and the client has paid for them — the trained models are the client's property on handover, and a partial delivery transfers the partial artefacts the client has paid for.

- The **retention** (₹3,93,616) is not payable where handover acceptance has not been achieved, and any

retention already withheld is applied against sums owed.

- Where Vector Logic cannot deliver because a **Phase 0 gate failed**, this is **not** a Vector Logic

failure to deliver — it is the benchmark doing its job — and §3.9 governs, not this clause.

7.3 Trained model artefacts on termination

On any termination, the **trained adapters, the voice master, the training data ledger and the model evaluation results** are handed to the client to the extent they exist and have been paid for, **except that the voice master and any model trained**

on it are released only if the voice-rights release (§8) is in place. If no release exists, the voice-derived artefacts are withheld (see §8) until it does.

7.4 Termination for convenience

Either party may terminate for convenience on written notice. The financial consequences are the same as §7.1 where the client terminates, and §7.2 where Vector Logic terminates.

8. THE VOICE-RIGHTS RELEASE REQUIREMENT

The TTS brand voice is **recorded by Vector Logic in a studio session**, and it is a **priced line item** in this project — casting, session direction, QC, retakes and mastering (2 work-weeks, ₹2,70,868) plus the **studio recording pass-through (~₹97,000 for ~20 h of delivered audio) = ₹3,67,868**.

Because that recording is used to train a model that **becomes the client's property at handover**, an obligation follows, and it is not optional:

Before the recorded voice is used to train any shipped model, the client must provide a signed release from the voice artist granting perpetual, irrevocable, worldwide, transferable commercial rights to use that voice. The release must grant:

- **Perpetual** — no expiry;
- **Irrevocable** — the artist cannot withdraw it later;
- **Worldwide** — no territorial limit;
- **Transferable** — the rights can move with the business;
- **Commercial** — for commercial use, including training and serving a synthetic voice.

The release must name the client (Papparti) as the beneficiary, because the model is handed to the client and becomes the client's property; a release in anyone else's favour would leave the client's product exposed. The reason is concrete and commercial: **if the release were revocable, the client's product would break the day the artist withdrew** — a live 24/7 voice-first booking business cannot have its brand voice switched off by a third party.

Legal review of the release is a Phase 0 item (CLIENT-ESTIMATE.md §4 / DECISIONS.md D10).

8.1 What Vector Logic does if no release is provided

No release = no clone. If the client does not provide the signed release, Vector Logic **will not train the synthesis model on the recorded voice, and will not ship any model derived from it.** The recording, and any intermediate model trained on it, are held and not delivered. The affected workstream is **suspended, not proceeded with on the assumption that a release will arrive later**, and the client is notified in writing.

Financial consequence: where the synthesis workstream is suspended for want of a release, that portion of the model-work milestone is **not invoiced until the release is provided** (or, if the client elects not to proceed, it is removed from scope under §5 and credited). Vector Logic is not exposed to shipping a voice without rights, and the client is not billed for a deliverable that cannot lawfully be delivered. If the client elects to abandon the brand-voice workstream, the **voice-production and studio-recording costs actually incurred cancelling the booking remain payable**.

9. LIABILITY AND WARRANTIES

9.1 The accuracy limitations disclosed in the proposal survive into the contract

The accuracy limitations stated in §8 of the proposal survive into this contract and **no warranty in this contract is to be read as contradicting them.**

Specifically, and without limitation:

- **Hinglish code-switched speech recognition is not solved.** Residual word error is expected, and

no accuracy figure is warranted before Phase 0B measures one. Any figure in the proposal or the estimate is a research claim or a model-card number until Phase 0B produces a measurement.

- ****If the client supplies no real audio and no real transcripts, the real-world accuracy of the shipped**

models cannot be stated** — and therefore cannot be warranted.

- **Catastrophic forgetting is bounded and measured, never eliminated.** The contract warrants a

bounded and measured regression (target under 3%, measured before and after), not an absence of regression.

- **No listening test has been run and no MOS exists for the voice output** until the listening test is

run in the project. The contract does not warrant a quality score that has not been measured.

- **The avatar is a layered vector rig, not a photoreal likeness** — the proposal's §8.3 statement

is warranted as fact: what ships is what was shown, in kind.

9.2 What Vector Logic does warrant

- **It will measure before it asserts.** Phase 0B is run and reported against the gates in §3.1, and a

measurement is provided, not an opinion.

- **It will tell the client when a gate fails,** in writing, at Phase 0 cost — before the big spend.
- **It will deliver the artefacts it has accepted in the milestone schedule,** and they will be the

client's property on handover.

- **It will hold the source of truth for the numbers:** every figure traceable to the estimate, and the

engineering rate shown as arithmetic so it can be re-run.

9.3 What Vector Logic does not warrant

- **No accuracy figure is warranted before Phase 0B measures one.** In particular, no specific WER,

MOS, or tool-call accuracy figure is warranted at signature.

- **No vendor price is warranted to stay fixed** — for the reasons in §1, and consistent with the

validity window.

- **No business outcome** — bookings, revenue, take-rate profitability or market adoption — is warranted.

Vector Logic · Papparti

The proposal's §8.3 commercial risks (self-hosting is not cheaper at low volume; the commission take-rate is under threat; "any service" is the sector's failure mode) are the client's to weigh.

9.4 Bounded liability

Vector Logic's aggregate liability under this contract is capped at the total fees paid to Vector Logic under it. Liability for indirect, consequential, or loss-of-profit claims is excluded. This bounded, honest position is deliberate: an AI quality guarantee beyond measurement, or an uncapped liability on a project whose core accuracy question is not yet measured, would be a false comfort to both parties. What the contract offers instead of an open-ended warranty is **the Phase 0 gate mechanism** — measurable, cheap, and exercised before the big spend.

10. OPEN FIELDS REGISTER

Every field below is **BLANK ON PURPOSE**. Each is a fact that only a human — the client, Vector Logic's own records, or a licensed adviser — can supply. **None of them has been invented, and none may be filled by guesswork.** This annexure is not signature-ready until each is completed.

#	FIELD	WHO FILLS IT	MUST NOT BE INVENTED BECAUSE
1	Client legal entity name (registered name of Papparti's contracting entity)	Client	The correct legal person to bind is a client fact; no name has been supplied.
2	Client contact person + designation	Client	The authorised signatory is a client fact.
3	Client registered address (for the tax invoice)	Client	Required on a tax invoice; not supplied.
4	Client GSTIN (if registered)	Client	A registration number must be read from the certificate, never constructed.
5	Vector Logic GSTIN	Vector Logic	The registered entity's number must be read from its GST certificate.
6	SAC for software development service	Tax adviser (both parties)	The cited SAC is recorded as UNVERIFIED; the correct current code must be confirmed.
7	SAC for AI / model training services	Tax adviser (both parties)	No code is asserted. The field is blank. Confirm the correct SAC before invoicing.
8	Place of supply and CGST/SGST vs IGST split	Tax adviser	Determined under the IGST Act per invoice; not asserted here.
9	TDS section and rate	Tax adviser (both parties)	The applicable section and rate must be confirmed under Indian law.
10	Payment-days convention (e.g. Net 15 / Net 30)	Both parties	A commercial convention the parties agree; not assumed.
11	Late-payment interest rate	Both parties	Agreed in the contract; not assumed.
12	Acceptance-review window (days for deemed acceptance, §3.8)	Both parties	Agreed in the contract; not assumed.
13	Governing-law city / jurisdiction	Both parties	The seat of the contract is a negotiated choice.
14	Voice artist identity + release date	Client (with counsel)	A release is signed by a named artist; the identity and date are not known.
15	Effective date of the contract	Both parties	The date of execution.
16	Client PAN (for TDS matching, §4.5)	Client	A number to be read from the client's records.
17	Vector Logic PAN	Vector Logic	To be read from Vector Logic's records.
18	30-day warranty start date	Both parties	Fixed by the handover acceptance certificate (§3.7).

11. VERIFICATION

Every figure in this annexure was recomputed against `CLIENT-ESTIMATE.md` and the proposal before issue. The results:

1. **Payment schedule sum.** $7,87,232 + 4,47,967 + 23,03,239 + 16,25,205 + 10,83,470 + 10,83,470 +$

$5,41,735 = 78,72,318$ — exactly the one-time fee excluding GST. ✓

1. **GST arithmetic.** $₹78,72,318 \times 0.18 = ₹14,17,017$; $₹78,72,318 + ₹14,17,017 = ₹92,89,335$. ✓

2. **Phase-row consistency.** Phase 0 (4,47,967) + model work (30,90,471) + Phase 1 (16,25,205) +

Phase 2 (10,83,470) + Phase 3 (10,83,470) + deployment/handover (5,41,735) = **78,72,318**. ✓

1. **Monthly total.** $41,665 + 5,891 + 1,935 + 4,837 + 1,161 + 2,902 + 67,717 = 1,26,108$. ✓

2. **Engineering rate.** $\$35/\text{hr} \times 96.7384 = ₹3,385.84/\text{hr}$; $\times 40 = ₹1,35,434/\text{work-week}$. ✓

3. **Forbidden-token scan.** A case-insensitive scan of this document for the 17 forbidden base-model

tokens returned **0 occurrences of each** (see the tooling report accompanying this file).

What could not be verified and is marked accordingly:

- **SAC codes** — recorded UNVERIFIED; the software-development SAC is a citation pending confirmation,

and the AI/model-training SAC is blank.

- **Client identity, GSTIN, PAN, address, contact** — blank fields, §10.
- **Governing-law city, payment days, acceptance window, interest rate** — negotiated blanks, §10.
- **The studio recording cost (₹97,000)** is a modelled midpoint of a ₹64,000–₹1,30,000 band, not a

studio quote (`CLIENT-ESTIMATE.md` §10); it is stated as such where used.

Prepared by Vector Logic, 11 October 2026. This is a commercial document and not legal advice. All figures traceable to `CLIENT-ESTIMATE.md` and the proposal. Base model identities are commercially confidential to Vector Logic and are not disclosed.

VECTOR LOGIC

AI & SOFTWARE DEVELOPMENT — ODISHA, INDIA

PREPARED FOR PAPPARTI

Papparti — Report Appendices

DATE

11 October 2026

CONFIDENTIALITY

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

Contents

Papparti — Report Appendices	3
E. Appendix — every figure and where it came from	3
E.1 One-time build rows	3
E.2 Monthly rows, after handover and go-live	5
E.3 Key rates and inputs	6
F. Appendix — what is measured today and what is not	8
F.1 MEASURED — a real instrument produced these numbers	8
F.2 NOT MEASURED — no instrument has produced these numbers yet	9
G. Appendix — open items and what we need from you	10
H. Closing page	13

Papparti — Report Appendices

Project: Papparti — on-demand any-service marketplace with a self-hosted voice and avatar layer **Prepared by:** Vector Logic — AI & software development, Odisha, India **Prepared for:** Papparti (client) **Date:** 11 October 2026 (IST) **Currency:** INR, Indian numbering. USD converted at the pinned rate 1 USD = ₹96.7384. **Status:** Appendices to the client package. Each figure names the document and section it comes from.

E. APPENDIX — EVERY FIGURE AND WHERE IT CAME FROM

Status key: **MEASURED** — a real instrument produced the number. **ESTIMATED** — a modelled figure with its assumption stated. **UNVERIFIED** — not confirmed; confirm before relying on it. Figures marked LIVE were read from a vendor's own pricing page on 11 October 2026 and are re-priced if the quote expires.

E.1 One-time build rows

#	FIGURE	AMOUNT (₹)	WHERE IT COMES FROM	STATUS
1	Phase 0 — feasibility benchmark	4,47,967	Quotation \$A row 1; Estimate \$1a row 1 / \$3; Proposal \$5	ESTIMATED (GPU live, engineering estimate)
2	Model work — four workstreams	30,90,471	Quotation \$A row 2; Estimate \$1a row 2 / \$4; Proposal \$5	ESTIMATED
3	Software — Phase 1	16,25,205	Quotation \$A row 3; Estimate \$1a row 3 / \$3 (12 weeks)	ESTIMATED
4	Software — Phase 2	10,83,470	Quotation \$A row 4; Estimate \$1a row 4 / \$3 (8 weeks)	ESTIMATED
5	Software — Phase 3	10,83,470	Quotation \$A row 5; Estimate \$1a row 5 / \$3 (8 weeks)	ESTIMATED
6	Deployment + handover	5,41,735	Quotation \$A row 6; Estimate \$1a row 6 / \$3 (4 weeks)	ESTIMATED

#	FIGURE	AMOUNT (₹)	WHERE IT COMES FROM	STATUS
7	One-time subtotal, excluding GST	78,72,318	Quotation §A subtotal; Estimate §1a; Proposal §6.1; Commercial Terms §2/§4.1	MEASURED (sum of the rows)
8	GST @ 18%	14,17,017	Quotation §A; Proposal §6.1; Commercial Terms §4.1 (= 78,72,318 × 0.18)	MEASURED (arithmetic)
9	One-time total, including GST	92,89,335	Quotation §A; Estimate §1a; Proposal §6.1; Commercial Terms §4.1	MEASURED (arithmetic)
10	Phase 4 — real-time lip-synced avatar (option)	13,54,338	Quotation §B row 7; Estimate §3 / §10; Proposal §5, §6.4	ESTIMATED ("starting estimate, not a quote")
11	Model work — STT workstream	8,14,886	Estimate §4	ESTIMATED
12	Model work — TTS workstream	8,13,973	Estimate §4	ESTIMATED
13	Model work — Reasoning LLM workstream	10,93,744	Estimate §4	ESTIMATED
14	Model work — Voice production workstream	3,67,868	Estimate §4; Proposal §3.4; Commercial Terms §8	ESTIMATED
15	Voice production — engineering (2 work-weeks)	2,70,868	Proposal §3.4; Estimate §4	ESTIMATED
16	Studio recording — ~20 h delivered audio	97,000	Proposal §3.4; Estimate §4 / §10; Commercial Terms §2.4, §8	ESTIMATED (midpoint of a ₹64,000–₹1,30,000 band)
17	Model work — GPU cost subtotal (244 GPU-hours)	13,927	Estimate §4	ESTIMATED (hours) / LIVE (rate)
18	Model work — engineering subtotal	29,79,544	Estimate §4	ESTIMATED

#	FIGURE	AMOUNT (₹)	WHERE IT COMES FROM	STATUS
19	Software development share of one-time total (Phases 1-3)	37,92,145	Estimate §8; Proposal §6.6	MEASURED (arithmetic, 51% share)
20	Mobilisation advance (10% of one-time fee)	7,87,232	Commercial Terms §2.1, §2.4	MEASURED (arithmetic)
21	Milestone 3 net (Model work less advance recovered)	23,03,239	Commercial Terms §2.1	MEASURED (arithmetic)
22	Retention / holdback (5% of one-time fee)	3,93,616	Commercial Terms §2.3	MEASURED (arithmetic)
23	Handover invoice net of retention	1,48,119	Commercial Terms §2.3 (5,41,735 – 3,93,616)	MEASURED (arithmetic)
24	Phase 0 — GPU rental sub-charge	41,665	Estimate §3; Proposal §4	LIVE (RunPod @ \$0.59/hr × 730 hr)
25	Phase 0 — engineering sub-charge	4,06,301	Estimate §3; Proposal §4 (3 work-weeks)	ESTIMATED
26	Change-order absorption threshold	13,542	Commercial Terms §5.2 (4 engineering hours)	MEASURED (arithmetic)

E.2 Monthly rows, after handover and go-live

#	FIGURE	AMOUNT (₹/MONTH)	WHERE IT COMES FROM	STATUS
27	GPU serving — 1× 48 GB dedicated, 24/7	41,665	Estimate §1b / §5; Quotation §C row 1; Commercial Terms §6	LIVE (RunPod @ \$0.59/hr × 730 hr)
28	Managed PostgreSQL + vector store	5,891	Estimate §1b / §5; Quotation §C row 2; Commercial Terms §6	LIVE (DigitalOcean @ \$60.90/mo)
29	Redis	1,935	Estimate §1b / §5; Quotation §C row 3; Commercial Terms §6	LIVE (Upstash @ \$20/mo)

#	FIGURE	AMOUNT (₹/MONTH)	WHERE IT COMES FROM	STATUS
30	Real-time media routing (SFU/TURN)	4,837	Estimate §1b / §5; Quotation §C row 4; Commercial Terms §6	LIVE (managed real-time media tier @ \$50/mo)
31	Object storage	1,161	Estimate §1b / §5; Quotation §C row 5; Commercial Terms §6	ESTIMATED (~500 GB at a live rate; volume assumed)
32	Monitoring + CI	2,902	Estimate §1b / §5; Quotation §C row 6; Commercial Terms §6	ESTIMATED (~\$30/mo placeholder)
33	Support retainer (0.5 work-week / month)	67,717	Estimate §1b / §5; Quotation §C row 7; Commercial Terms §6	ESTIMATED (assumption at the engineering rate)
34	Monthly total	1,26,108	Quotation §C; Estimate §1b; Proposal §6.2; Commercial Terms §6	MEASURED (sum of the rows)
35	Non-GPU, non-retainer monthly lines	16,726	Commercial Terms §6.1 (5,891 + 1,935 + 4,837 + 1,161 + 2,902)	MEASURED (arithmetic)
36	Second 48 GB serving GPU, if concurrency requires	~83,000	Estimate §5; Proposal §6.2; Quotation §D (not included)	UNVERIFIED (flagged, not priced)

E.3 Key rates and inputs

#	FIGURE	AMOUNT	WHERE IT COMES FROM	STATUS
37	FX rate — USD to INR	1 USD = ₹96.7384	Estimate header / §6 / §9 [5]; Proposal §10; GPU Decision §8 (A1)	LIVE (XE mid-market, 11 Oct 2026, 04:05 UTC; re-verified 96.7372)
38	Engineering rate	\$35/hour	Estimate §6; Proposal §6.0 / §10; Commercial Terms §5.1	ESTIMATED (internal assumption; CEO must confirm)
39	Engineering rate — hourly in INR	₹3,385.84/hour	Estimate §6; Commercial Terms §5.1	MEASURED (arithmetic on the pinned FX)

#	FIGURE	AMOUNT	WHERE IT COMES FROM	STATUS
40	Engineering rate — per day	₹27,087/day	Estimate \$6	MEASURED (arithmetic)
41	Engineering rate — per work-week	₹1,35,434 per 40-hour week	Estimate \$6 / \$9; Proposal \$10; Quotation \$F; GPU Decision \$6	MEASURED (arithmetic)
42	Work-week	40 hours, 5 days	Estimate \$9; Proposal \$10	MEASURED (standard)
43	GPU hours per month	730 (24×7)	Estimate \$9; Proposal \$10; GPU Decision \$1.1	MEASURED (24 × 7 × 365/12)
44	GPU hourly rate — reference serving	\$0.59/hour (48 GB)	Estimate \$2 [1]; GPU Decision \$8 (A2)	LIVE (RunPod, 11 Oct 2026)
45	GPU hourly rate — Vast.ai floor / median	\$0.28 / \$0.40 per hour	Estimate \$2 [2]	LIVE (not the planning rate)
46	GPU hourly rate — Lambda Labs	\$0.80 per hour	Estimate \$2 [3]	LIVE (not the planning rate)
47	GPU hourly rate — India-hosted tier	\$1.20 / \$1.44 per hour	Estimate \$2 [4] (L40S / A40 48 GB)	LIVE (contingency, not default)
48	Stale GPU figure corrected out	₹34,056/month (at ₹88/USD)	Quotation \$F; Estimate \$2; Commercial Terms \$1	UNVERIFIED (superseded; kept only to show the correction)
49	Payment-gateway fee	2% + 18% GST on the fee	Estimate \$7b; Proposal \$6.4; Commercial Terms (via \$6 note)	LIVE (Razorpay structure)
50	Gateway fee worked example on ₹1,00,000 settled	₹2,360	Estimate \$7b; Proposal \$6.4	MEASURED (arithmetic)
51	Commission floor the gateway implies	~2.4% of booking value	Estimate \$8; Proposal \$8.3	ESTIMATED (arithmetic on the live fee)
52	GPU break-even — used card, refurb server (Option C)	~21 months	Quotation \$F; GPU Decision \$1.3 / \$4	ESTIMATED (modelled)

#	FIGURE	AMOUNT	WHERE IT COMES FROM	STATUS
53	GPU break-even — used card, India OEM server (Option B)	~34 months	GPU Decision \$1.3 / \$4	ESTIMATED (modelled)
54	GPU break-even — new card, new server (Option A)	~61 months	Quotation \$F; GPU Decision \$1.3 / \$4	ESTIMATED (modelled)
55	Rent vs buy — 36-month total, renting	₹14,99,940	GPU Decision \$1.3	ESTIMATED (modelled)
56	Rent vs buy — 36-month total, Option C	₹10,03,217	GPU Decision \$1.3	ESTIMATED (modelled)
57	Odisha LT-industrial energy charge	₹6.20/kWh	GPU Decision \$2e / \$8 (A4)	LIVE (OERC RST FY 2026-27)

Every row above names a real document and section. The totals in E.1 and E.2 stand independently of the UNVERIFIED and ESTIMATED rows: where a figure is an assumption it is not added to a verified sum — the arithmetic is shown where it is used.

F. APPENDIX — WHAT IS MEASURED TODAY AND WHAT IS NOT

F.1 MEASURED — a real instrument produced these numbers

1. **Voice sample duration — male greeting** ~5.5 s (Proposal §8.4).
2. **Voice sample duration — female greeting** ~6.8 s (Proposal §8.4).
3. **Voice sample duration — male number sentence** ~6 s, containing a price and a duration (Proposal §8.4).
4. **Voice sample duration — female number sentence** ~6 s (Proposal §8.4).
5. **Male voice speech energy on the number sentence** -16.2 dBFS — further from clipping (Proposal §8.4).
6. **Female voice speech energy on the number sentence** -12.8 dBFS (Proposal §8.4).
7. **Male voice speech energy on prose samples** -15.4 dBFS (Proposal §8.4).
8. **Proportion of non-silent frames**, measured per voice sample (Proposal §8.1 — audio generated and measured objectively).

9. **Real-time factor (RTF)** of the voice output on our own hardware (Proposal §8.1 — measured).
10. **Avatar tonal steps — high tier** 30 distinct steps spanning 131 of 255 (Proposal §8.3).
11. **Avatar tonal steps — low tier** 1 distinct step spanning 0 (Proposal §8.3) — a measured $30\times$ difference, enforced by a script that fails the build if it collapses.
12. **Avatar mouth-region redraw between idle and talking frames** 35.8% (Proposal §8.3).
13. **Horizontal overflow** 0 px at 1920 / 1440 / 1024 / 390 px widths (Proposal §8.3).
14. **GPU rental price read from the vendor's live page** \$0.59/hour for a 48 GB card on 11 October 2026 (Estimate §2; GPU Decision §1).
15. **FX rate read live** 1 USD = ₹96.7384, cross-checked at 96.7372 on 11 October 2026 (Estimate §9; GPU Decision §8).
16. **All totals and their arithmetic** — the one-time subtotal, the GST line and inclusive total, the monthly total, the payment-schedule sum, and the milestone figures were all recomputed and pass (Commercial Terms §11; Estimate §8).
17. **The renders and demo artefacts shipped with the proposal** — four avatar images, the speaking video, and the scripted conversation clips all exist as files (Proposal §6.0).

F.2 NOT MEASURED — no instrument has produced these numbers yet

1. **The Phase 0B benchmark as a whole has not run.** It is the gating measurement of the entire plan and is the next paid step, not a completed result (Proposal §4; Commercial Terms §3.1).
2. **Hindi recognition accuracy (best base, after fine-tune)** — not measured; the gate is 12% WER (Proposal §4).
3. **Hinglish code-switch recognition accuracy (before any fine-tune)** — not measured; the gate is 20% WER (Proposal §4).
4. **Fine-tune improvement over the untrained base** — not measured; the gate is $\geq 15\%$ relative gain (Proposal §4).
5. **Sessions per node under real load** — not measured; the gate is ≥ 10 . A $3\times$ error here triples the infrastructure bill (Proposal §4; GPU Decision §5).
6. **p95 time to first audio** — not measured; the gate is $\leq 1,500$ ms (Proposal §4).
7. **Booking tool-call exact-match accuracy (Hindi/Hinglish)** — not measured; the gate is 85% (Proposal §4).

8. **Ambiguous-duration inference rate** — not measured; the gate is a hard stop above 0% (Proposal §4).
9. **General-capability regression after fine-tuning** — not measured; the gate is under 3% (Proposal §4).
10. **No MOS and no listening test exists for the voice output.** No human listener has rated it, and no quality score is quoted (Proposal §8.1; Commercial Terms §9.1).
11. **The client's real GPU utilisation** — unknown until Phase 0B measures it. This is why renting is recommended rather than buying (GPU Decision §5, §7).
12. **Concurrency capacity of a single 48 GB card under real load** — unmeasured; the monthly GPU line assumes single-stream serving (Estimate §11; GPU Decision §5).
13. **Any Hindi speech quality number from a model card** — every figure in our research is a vendor claim or a model-card number, not a measurement (Proposal §10; Estimate §11).
14. **The two conflicting Hindi accuracy figures in our research** — cannot be reconciled from the documentation; only our own measurement replaces them (Proposal §8.1).
15. **Vendor-specific latency for India-served inference** — not measured; the India-hosted GPU stays a contingency (Estimate §11).
16. **No human judgement of the avatar's art direction** — that decision belongs to the client (Proposal §8.3).
17. **The studio recording has not run** — the final brand voice does not exist yet (Proposal §6.0).
18. **Node power draw (600 W) and the real GPU-colo price** — assumptions, not measurements (GPU Decision §7).

G. APPENDIX — OPEN ITEMS AND WHAT WE NEED FROM YOU

Every field below is blank on purpose. Each is a fact only a human can supply; none may be filled by guesswork. The documents are not signature-ready until these are complete.

#	OPEN ITEM	WHO MUST SUPPLY IT	WHY IT CANNOT BE GUESSED
1	Client legal entity name as registered	Papparti	The address reads "Papparti", a brand. A contract and every invoice need the registered legal person.
2	Signatory name and designation	Papparti	Milestone acceptance must be unambiguous, and the right person must receive each deliverable.
3	Contact email (the address that reaches you)	Papparti	contact@papparti.com is used only as a placeholder in the demo references.
4	Client registered address (for the tax invoice)	Papparti	Required on a valid Indian tax invoice; not supplied.
5	Client GSTIN (if registered)	Papparti	A registration number must be read from the certificate, never constructed.
6	Client PAN (for TDS matching)	Papparti	A number to be read from the client's records.
7	Vector Logic GSTIN	Vector Logic	To be read from the registered entity's GST certificate before the first invoice.
8	Vector Logic PAN	Vector Logic	To be read from Vector Logic's records, so TDS credit can be matched.
9	SAC code — software development service (cit. 998314)	Tax adviser (both parties)	Recorded UNVERIFIED ; the current code must be confirmed before it prints on an invoice.
10	SAC code — AI / model training services	Tax adviser (both parties)	No code is asserted at all; the field is blank on purpose.
11	Place of supply and CGST/SGST vs IGST split	Tax adviser	Determined under the IGST Act per invoice; not asserted as a settled position.

#	OPEN ITEM	WHO MUST SUPPLY IT	WHY IT CANNOT BE GUESSED
12	TDS section and rate	Tax adviser (both parties)	The applicable section and rate must be confirmed under Indian law.
13	Voice artist identity and release date	Papparti (with counsel)	A release is signed by a named artist; no release = no clone.
14	Does real Hindi/Hinglish audio already exist? (call recordings, voice notes)	Papparti	Changes Phase 0's test design; if there is none, that is workable but must be known.
15	The service taxonomy — the actual launch service categories	Papparti	Drives the recognition vocabulary and the reasoning model's structured output.
16	The locality names — real place-names and their Hindi pronunciations	Papparti	Drives the recognition vocabulary and the locality field.
17	One launch vertical and one launch locality	Papparti	"Any service" from day one is the measured failure mode of this sector.
18	Demo script's city vs your market	Papparti	The demo references a city; confirm it matches the locality you intend to launch.
19	Payment-days convention (e.g. Net 15 / Net 30)	Both parties	A commercial convention the parties agree; not assumed.
20	Late-payment interest rate	Both parties	Agreed in the contract; not assumed.
21	Acceptance-review window (days, for deemed acceptance)	Both parties	Agreed in the contract; not assumed.
22	Governing-law city / jurisdiction	Both parties	The seat of the contract is a negotiated choice.
23	Effective date of the contract	Both parties	The date of execution.
24	30-day warranty start date	Both parties	Fixed by the handover acceptance certificate.

Items 1–6 and 13–18 are the ones that block the earliest work. Items 7–12 and 19–24 block signature and invoicing.

H. CLOSING PAGE

If you want to proceed. Two paths, and nothing else is needed to start.

Path one — proceed as scoped. Supply your registered legal entity name and your signatory's name and designation, and confirm your acceptance in writing on or before **10 November 2026**. Vector Logic adds its GSTIN and the confirmed SAC codes at invoice time. Work begins with Phase 0: three work-weeks of engineering plus one month of GPU time, delivering a written feasibility report with our own measured numbers and a go/no-go on each of the three models. Phase 1 does not begin until Phase 0's results are accepted. The mobilisation advance of ₹7,87,232 is drawn against setup and is recovered at the Model work milestone.

Path two — change something first. Raise it in writing before acceptance. A change to the scope is described as a discrete deliverable and priced before any work begins, at the shown rate of ₹3,385.84/hour (₹1,35,434 per 40-hour work-week), plus any pass-through vendor cost at the then-current live price, plus GST. Changes whose total value is at or below ₹13,542 are absorbed at no charge. If the change is a scope-down to meet a lower budget, the model workstream is the line to protect, because it is the part no one else is selling you.

Either way, one date governs: this pricing is valid until **10 November 2026**, because the monthly figures rest on live vendor and exchange rates read on 11 October 2026. After that date Vector Logic re-prices the affected lines from then-current prices and issues a fresh validity window. The engineering rate is the part that does not move.

Prepared by Vector Logic, 11 October 2026. All figures traceable to the client package. Base model identities are commercially confidential to Vector Logic and are not disclosed in this document.