UnitSense

You price flat. You pay per token.

The margin layer for AI-native products — cost per customer, per feature, across your whole AI stack.

Bilal Baqar — CEO

bilal@unitsense.ai

Salman Sikandar — President & CTO

salman@unitsense.ai

You sell in units your customers understand. You buy in units your vendors invented.

Per seat, per agent, per clinic, per minute of product — while the costs underneath run per token, per character, per GPU-second. Someone has to reconcile the two. Today, that someone is a spreadsheet.

Flat price out

Your revenue per customer is fixed the day the contract is signed.

Variable cost in

Your cost per customer moves with every conversation, call, and session they run.

Month-end surprise

The gap between the two shows up weeks later, aggregated, on six different invoices.

THE PROBLEM ISN'T THAT AI COSTS TOO MUCH — IT'S THAT YOU'RE PRICING BLIND.

Three ways pricing blind catches up with you.

01

Whale customers

Your heaviest users can be your least profitable. Without per-customer cost, your best logo and your worst margin can be the same account — and you can't tell.

02

Pricing on nerves

Repricing conversations run on gut feel. Cap it? Tier it? Raise it? Without numbers, every option feels risky, so nothing changes.

03

The diligence question

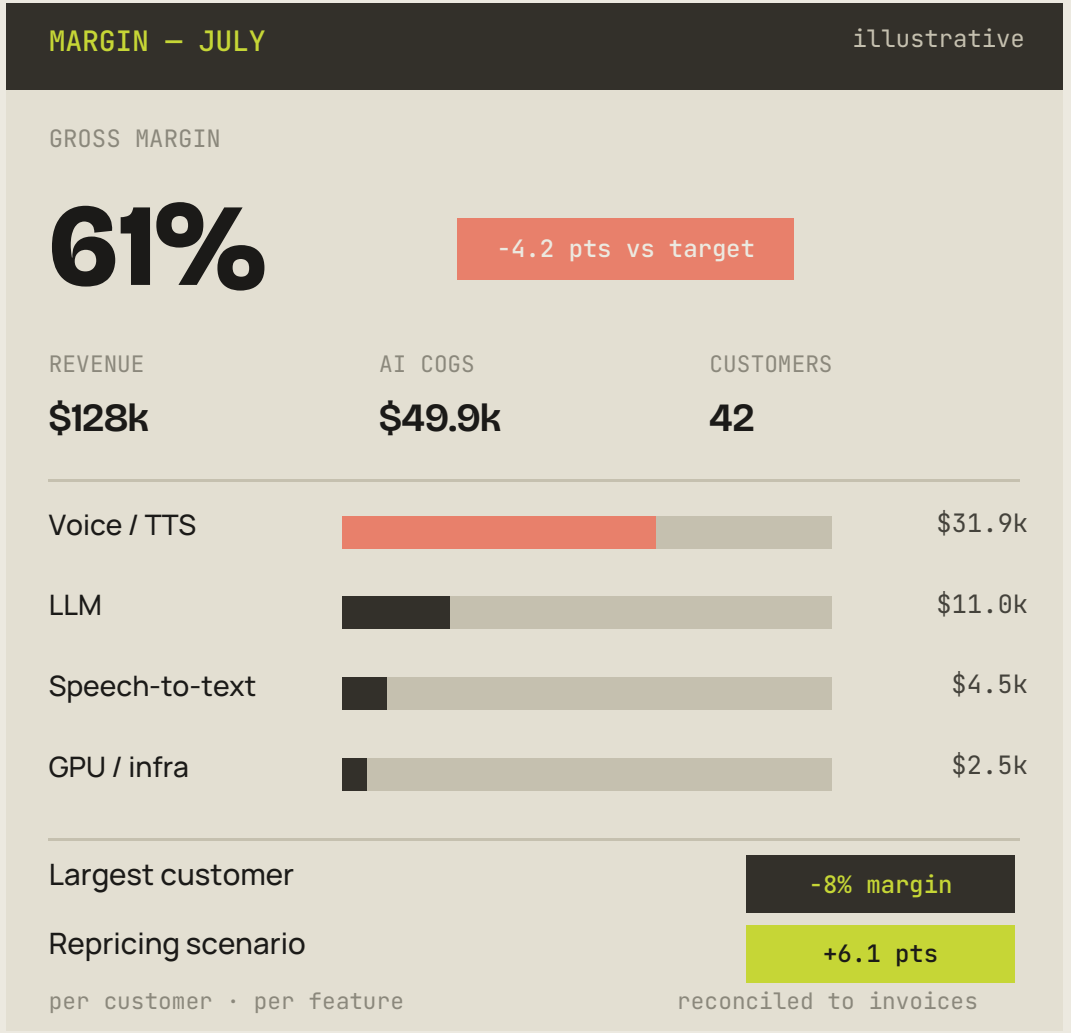
At your next round, someone will ask for unit economics per customer. “We don't track that” is not an answer you want to give with a term sheet on the table.

Margin, at the unit you actually sell.

01 Cost per unit you sell — per call, per minute, per session, per seat

02 Margin per customer and per feature

03 The customers who break your pricing model



Your margin lives across the whole stack. No single vendor can see it.

Your OpenAI dashboard doesn't know ElevenLabs exists. A voice session touches five vendors before it reaches the customer — the margin answer only exists where all five meet.

= your real cost per call

often the largest line

The long-call trap

Session cost grows faster than session length — context builds as a conversation runs, so one 9-minute call costs more than three 3-minute calls. Per-minute or per-seat pricing can't see that. Per-customer margin can.

WE RECONCILE EVERY LAYER TO ITS ACTUAL INVOICE — SO THE NUMBER IS ONE FINANCE CAN DEFEND.

Repricing, rehearsed before you ship it.

Seeing the leak is half the job. UnitSense models the fix on your own usage: what to charge, what to cap, what the next tier looks like.

SCENARIO — TIER CAP + MODEL SWAP illustrative

TODAY	→	MODELED
61%	→	67.1%

- Cap included minutes at P85 usage
- Route tier-1 summaries to a smaller model
- Whale account → custom tier

How the biggest AI platforms price this

The emerging pattern for products with open-ended AI usage: sell included AI units inside the subscription; when a customer burns through them, they move up a package. Simple for the buyer, safe for your margin.

UnitSense gives you the usage curves to set those units with confidence instead of guesswork.

Built for the stack you already run.

MODELS	**OpenAI · Anthropic · Google Gemini · AWS Bedrock · Azure OpenAI · Mistral**
VOICE & SPEECH	**ElevenLabs · Cartesia · Hume · Deepgram · AssemblyAI**
VOICE AGENTS & TRANSPORT	**Vapi · Retell · LiveKit · Pipecat · Twilio**
GPU & CLOUD	**AWS · GCP · Azure · CoreWeave · Modal · RunPod**
VECTOR & DATA	**Pinecone · Qdrant · Weaviate · Chroma**
CODING TOOLS	**Claude Code · Codex · Cursor · GitHub Copilot · Windsurf**
REVENUE	**Stripe · Orb · Lago · Chargebee · QuickBooks**

How we connect: LiteLLM · Bifrost · Python SDK · JavaScript SDK · our open-source agent · provider admin APIs · CSV import · invoice ingestion

…and the rest of your stack. Where a vendor exposes usage, we pull it. Where they don't, we ingest the invoice.

A short call to your first margin picture.

01

A short call

Your stack and how you price. That's the whole agenda — no demo theater.

02

Your margin picture

We come back with your margin view built on your stack's shape and realistic numbers. Yours to keep either way.

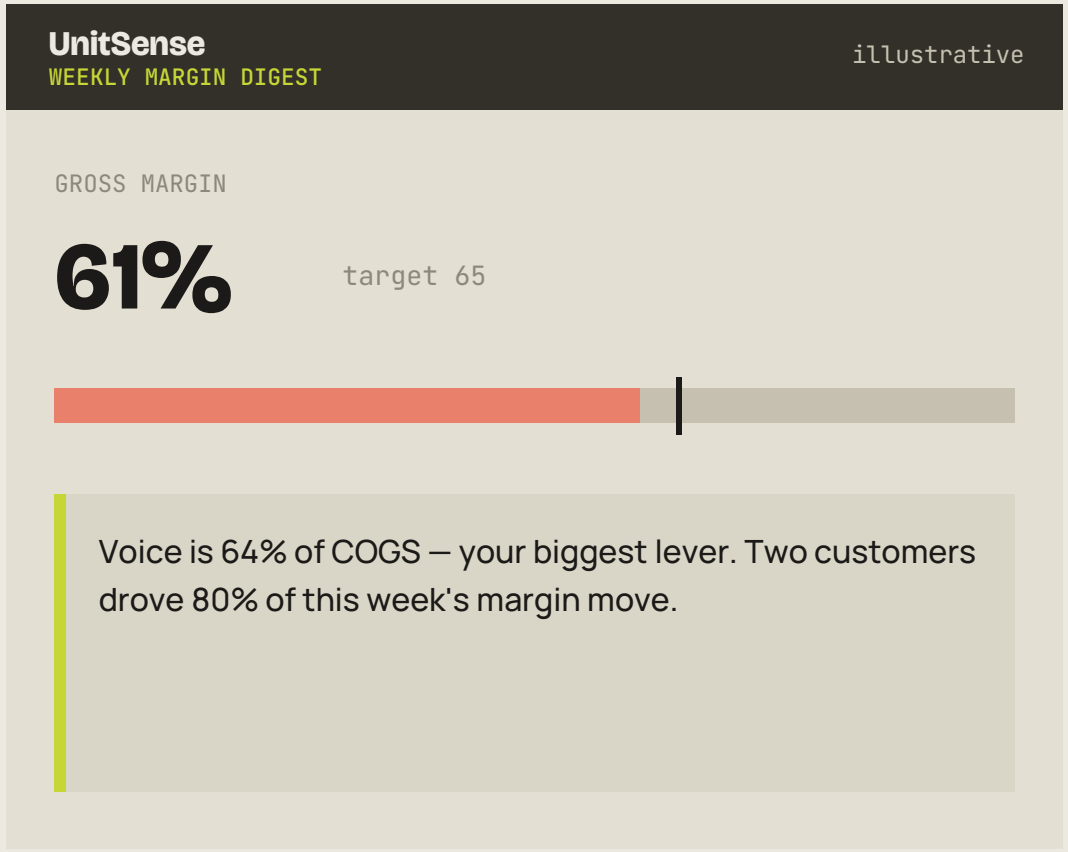
03

Built around you

If it's useful, we integrate against what you actually run. We work closely with a small number of early companies — the product gets shaped around your stack, not the other way around.

Your margin, in your inbox, every Monday.

No dashboard ritual. The digest lands with gross margin vs. target, what moved this week, and your single biggest cost lever — readable in 60 seconds, forwardable to your CFO.



We've run AI infrastructure at scale — now we're building its missing ledger.

Bilal Baqar

CO-FOUNDER & CEO

Director of AI Infrastructure at CTDS — built a petabyte-scale, zero-trust Kubernetes GPU platform. UChicago MS CS, Chicago Booth MBA. Previously PLUMgrid (enterprise SDN).

Salman Sikandar

CO-FOUNDER, PRESIDENT & CTO

Platform Engineering Lead at CTDS. Built and runs the UnitSense multi-tenant platform end to end. Previously PLUMgrid, Motorola.

WE WERE THE BUYERS OF THIS PRODUCT BEFORE WE WERE THE BUILDERS.

See your numbers.

We'll show you what your margin picture could look like — built on your stack and how you price.

Setup a demo — unitsense.ai/book

more at unitsense.ai/margin

UnitSense · bilal@unitsense.ai · salman@unitsense.ai