

Your **AI transformation** **partner.**

We set & execute your enterprise AI strategy at startup speed.

AI isn't scary.
Ignoring it is.

We don't sell headcount. We sell **outcomes.**

Forward-deployed engineers embedded in your team, priced against the result rather than the hours. You buy a system that works, not a number of people.

01

Priced on the result

Resolution time, deflection rate, cost per transaction — we contract against the metric that moves your P&L, and we carry the delivery risk.

02

Senior people, embedded

Inside your repos, your data and your standups from week one. No offshore pyramid, no ticket queue, no account manager in between.

03

Weeks, not quarters

Two weeks to something running in your environment. The deck is never the deliverable — the working system is.

Five layers. One team ships all of them.

Most vendors sell you one layer and subcontract the rest. We own the whole column, from the metal to the agent.

05	Agents	Self-improving agents for any function, built on your context
04	Business Context Engine	Every system your company runs on, made retrievable and permission-aware
03	Model Layer	Fine-tuning, RL and distributed training on your proprietary data
02	Serving	On-prem bare-metal GPU, or AWS / GCP at scale
01	Infrastructure	Distributed, high-throughput systems built to hold under load

ACROSS EVERY
LAYER

Evals.
Guardrails.
Observability.

If we can't measure it,
we don't ship it.

Every answer your business needs, grounded in **your own data.**

Every system your company already runs, turned into one place your people — and your agents — can actually ask. No data migration project.

01

One place to ask

Documents, tickets, databases, wikis and chat stop being twelve separate searches. One question, one answer, drawn from all of them.

YOU GET

Connector set + continuously-synced index

02

Answers you can defend

Every answer cites where it came from. And the engine knows a customer, an invoice and a ticket are the same story — so it joins the dots your staff currently join by hand.

YOU GET

Retrieval service + entity graph

03

Nobody sees what they shouldn't

Access rules come straight from the systems you already run. The engine cannot surface what a person was never allowed to open, and every query is logged.

YOU GET

Permission model + audit log

THE ASSET

1 index

across every system your company runs on. Your context is the moat — not the model, which everyone can rent.

Agents for any function — that **get better every week.**

Because they sit on the context engine, our agents know your business from day one. Because every action they take flows back as signal, they are sharper in month six than in month one.

01

Any function

Support, operations, sales, finance, compliance, back office, engineering. One agent substrate, many jobs — we build the one your P&L feels first.

YOU GET

Agent runtime + tool integrations

02

Grounded from day one

Multi-step agents that read your systems of record, decide and write back — with your policies, your history and your permissions already in context.

YOU GET

Context bindings + confidence-gated handoff

03

Self-improving loop

Every action, correction and outcome is captured as labelled signal. Evals re-run, policies and models update. The loop tightens as your data flows in.

YOU GET

Feedback loop + retraining pipeline + eval dashboard

THE COMPOUNDING EDGE

We don't ship a feature that launches. We ship a system that learns.

One deployed loop took resolution time from months to under an hour — and kept improving after go-live.

Your data never leaves your building.

Open-weight models deployed on your own bare-metal GPUs, inside your perimeter — for the workloads your regulator, your board or your contracts will not let you send to an API.

01

Sovereign

Models run inside your VPC or your data centre. Nothing crosses the boundary — built for DPDP, sectoral regulators and data-residency clauses.

YOU GET

Cluster build + deployment runbook

02

Economics

You own the cost curve. Capex on your own silicon instead of a per-token bill that scales with your success.

YOU GET

Capacity model + cost-per-token baseline

03

Performance

A tuned serving stack — continuous batching, quantization, KV-cache and paged attention — so the hardware you bought is the hardware you use.

YOU GET

Tuned inference stack + load-test report

THE GUARANTEE

0 bytes

of your proprietary data leave your perimeter. Not for inference, not for training, not for telemetry.

A model that knows your business, not the internet's.

Supervised fine-tuning, reward modelling and distributed reinforcement learning on the proprietary data only you have.

01

Adapt

SFT and LoRA on your corpus, your domain vocabulary and your house tone — so output needs editing, not rewriting.

YOU GET

Fine-tuned checkpoint + eval report

02

Align

Reward models built from your own experts' preferences, then RL against them. Your best people's judgement, encoded.

YOU GET

Reward model + preference dataset

03

Scale

Multi-node distributed training with checkpointing, reproducible runs and a versioned model registry. Repeatable, not heroic.

YOU GET

Training pipeline + model registry

WHY IT COMPOUNDS

Every point of accuracy you gain here is a point your competitor cannot buy.

Anyone can rent the same frontier model. Nobody else has your data.

Scale that holds on the **worst day of the year.**

Distributed, high-throughput systems on AWS and GCP — designed by engineers who ran payments infrastructure where downtime was measured in money.

01

Throughput

Event-driven pipelines, autoscaling inference fleets and queue-backed workloads that degrade gracefully instead of falling over.

YOU GET

Reference architecture + infra-as-code

02

Resilience

Multi-AZ, explicit SLOs, load-shedding and failover paths rehearsed before launch — not discovered during the festival peak.

YOU GET

SLO definitions + failover runbook

03

Cost

Cost per request tracked as a first-class metric: routing, caching and right-sized models so unit economics improve as volume grows.

YOU GET

Unit-economics dashboard

PEDIGREE

Built by engineers who ran payments and data infrastructure at Gojek.

Scale is not a slide for us. It is where we come from.

If we can't measure it, we **don't ship it.**

The reason AI pilots stall is that nobody can prove they work. Every system we hand over arrives with the evidence attached.

01

Eval harness

Golden datasets written with your domain experts, scored automatically, wired as regression gates in CI. A quality drop blocks the deploy.

YOU GET

Golden set + CI regression gate

02

Guardrails

PII redaction, refusal policy, citation enforcement and prompt-injection defence — tested adversarially, not assumed.

YOU GET

Policy config + red-team report

03

Observability

Traces on every request, cost and latency per call, drift and quality alerting. You see what the system did and why.

YOU GET

Tracing stack + drift alerting

HANDOVER STANDARD

Every system we hand over ships with its own test suite.

Your team can verify it, change it and run it without us.

HOW WE ENGAGE

Two weeks to a working prototype. Then we ship.

No six-month discovery. No 200-slide deck. The audit ends with something running.

WEEKS 0-2 · AUDIT

Prototype & ROI map

A working prototype plus a ranked list of use cases with real ROI math. Fixed fee. You keep the output either way.

WEEK 3+ · EMBED

Forward-deployed squad

A senior squad inside your repos and standups. Subscription-priced per squad, not per hour. Scale up or stop monthly.

ONGOING · OWN

Outcome contract

Priced against the metric that matters — resolution time, deflection rate, cost per transaction. We carry the delivery risk.

Change management is part of the work, not an upsell: workflow integration, staff training and a measured baseline so ROI is a number, not a claim.

THE BENCH

Builders from IIT & IIT. **Shipped at scale before.**

We don't staff from a résumé database. Every engineer comes from a network we've worked alongside — and we only send people we'd put on our own product.

150+

senior engineers in our deployable network

6

forward-deployed engineers live in client teams today

Gojek & Swiggy

payments, logistics & large-scale data infrastructure in our founding DNA

From AI-absent to **AI-native.**

MARK **AI**

saurabh@trymarkai.com