



Your AI prototype works in the demo. It falls apart in production. **That is the part we build.**

Most AI pilots stall somewhere between the demo and the first audit. We take RAG systems, agents and LLM products the rest of the way for companies in the USA, UK, Germany and the EU. The success metric is agreed before any code is written, and you own every line of it at handover.

[Book a call today](#)

[Message us on WhatsApp](#)

22+

AI systems running in production

27+

enterprise clients served

\$870K+

documented client cost savings

47%

average operational efficiency gain

99.9%

production uptime

Claude Certified Architect (Anthropic)

AWS, Azure and GCP professional certificates

Top 100 global Kaggle Master

Upwork Top Rated Plus, 100% Job Success

IS THIS ON YOUR DESK?

4 problems we are usually called in to fix

- **Your pilot impressed leadership, then stalled on cost, latency or a compliance review.**

Add the evaluation set, guardrails, tracing and monitoring a reviewer needs to sign off, then put real users on it.

- **Your LLM bill is growing faster than your usage.**

Model routing, semantic caching and private open-weight models where the volume justifies them. 1 legal client now runs at 60% lower inference cost.

- **Nobody can tell you whether the system got worse last week.**

Evaluation gates in CI and drift detection on live traffic, so a quality drop blocks the release before a customer sees it.

- **Your data is not allowed to leave your network.**

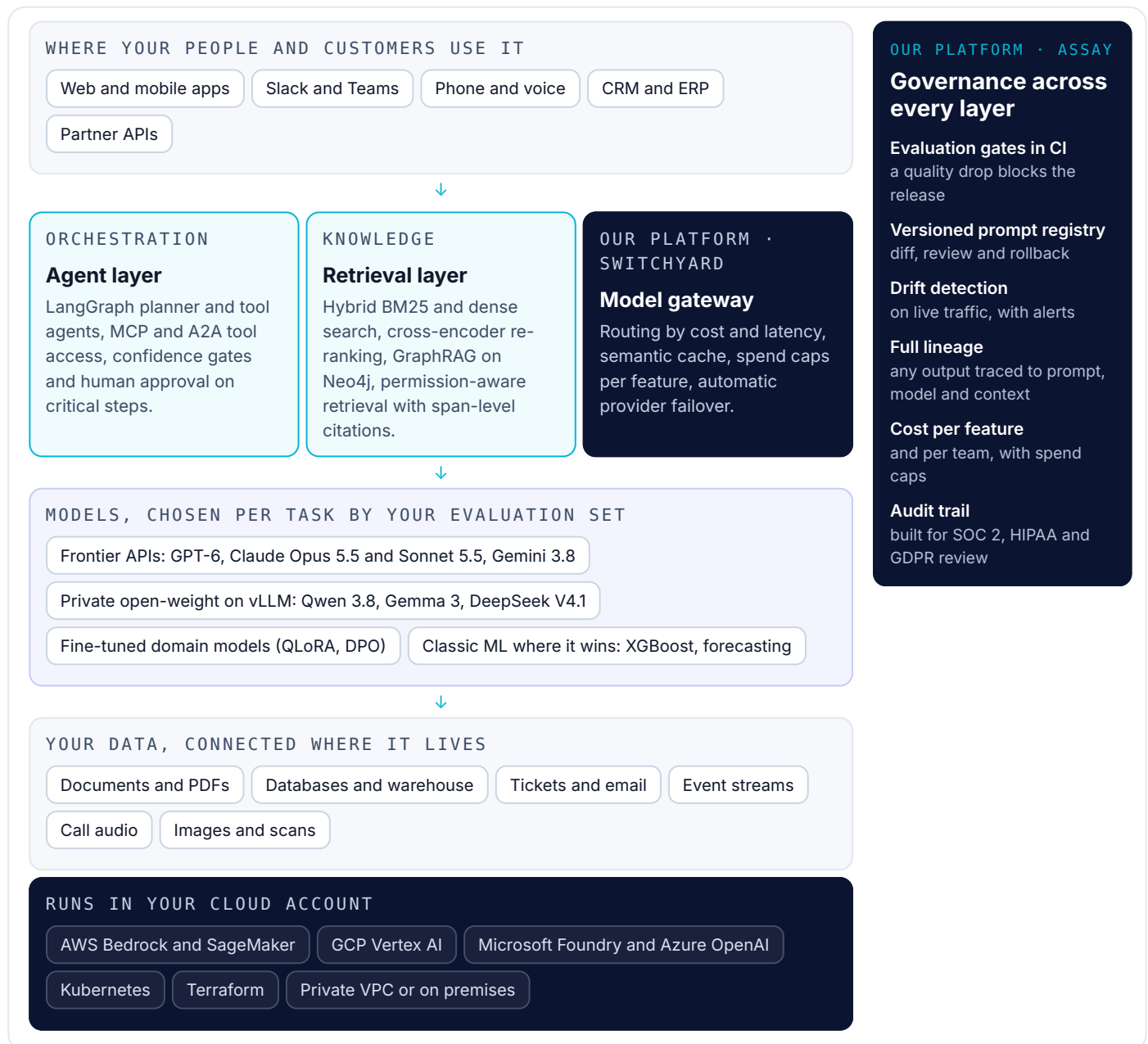
Open-weight models fine-tuned and served inside your VPC or on premises, with a full audit trail and role-based access.

1 of these on your desk right now? Get on a call with me today. In 30 minutes you will know the approach I would take and roughly what it costs, whether or not you hire us.

[Book today >](#)

The architecture every engagement starts from

Each client gets only the layers their problem needs. Several systems shipped without the private model tier, and 1 shipped without a vector store because plain keyword search beat it on cost and accuracy.



OUR PLATFORM · ASSAY

Governance across every layer

Evaluation gates in CI
a quality drop blocks the release

Versioned prompt registry
diff, review and rollback

Drift detection
on live traffic, with alerts

Full lineage
any output traced to prompt, model and context

Cost per feature
and per team, with spend caps

Audit trail
built for SOC 2, HIPAA and GDPR review

HOW WE START

From first call to a measured result in 14 days

- 1**

DAY 0

Scoping call

30 minutes with the engineer who will build it. If AI is the wrong tool for the problem, we say so here.
- 2**

DAYS 1 TO 2

Written scope

Success metric, architecture, what is out of scope, fixed price and dates, in writing.
- 3**

DAYS 3 TO 10

Baseline and first slice

Your evaluation set comes first, then a working slice on your real data that you can query.
- 4**

DAYS 11 TO 14

Scored result

The slice is scored against the agreed set. If the number is there, we scale it. If not, you stop at a pilot fee.
- 5**

AFTER LAUNCH

Run and improve

Monitoring and improvement cycles on retainer, or a clean handover to your team.

Not sure which layers you need? Walk me through your stack on a call today and I will tell you which of these you can skip.

[Book today >](#)

You commit budget 1 stage at a time

STAGE 1 · 2 WEEKS

Pilot

1 workflow, 1 success metric, fixed price. You see a scored result on your own data before committing a larger budget.

STAGE 2 · 2 TO 16 WEEKS

Production build

The same architecture scaled out with integrations, guardrails, monitoring and security review, delivered in milestones you approve.

STAGE 3 · 6 MONTHS AND LONGER

Retainer or team

Monthly improvement cycles or model upgrades, or AI engineers placed inside your team, working in your time zone.

WHAT YOU OWN AT HANDOVER

The system and everything needed to run it without us

Full source with commit history

Fine-tuned weights where applicable

The evaluation set your team can re-run

Architecture diagram and decision notes

Runbook for the failure modes we expect

Dashboards and alerts already wired

Recorded handover plus 30 days of questions at no charge

IN-HOUSE PLATFORMS

The gateway and governance layers are our own code, open to your engineers

Switchyard

AI GATEWAY AND COST CONTROL

OpenAI and Anthropic compatible proxy with per-feature cost attribution, spend caps, semantic caching, routing and provider failover.

Assay

EVALUATION AND GOVERNANCE

Versioned prompt registry, CI evaluation gates, drift detection and lineage from any output back to its prompt and model version.

Plumblin

CONTRACT REVIEW

7 agents extract 24 clause types with verified source spans. Low-confidence findings go to a human review queue. Evaluated on CUAD.

How you contract with us. You sign with Jillani SofTech, a registered IT services company with an engineering team. MSA and NDA on request, USD invoicing by international bank transfer, milestone payments agreed in writing. The code, weights and evaluation set are yours on final payment.

SHIPPED WITH EVERY SYSTEM

What ships with every system

An evaluation set with a number

Built from your data before the system, re-run on every change.

Guardrails

PII redaction, policy checks and confidence gates on every output.

Tracing and dashboards

Langfuse or LangSmith on every call, with quality, latency and cost in 1 view.

Alerts that fire first

Thresholds set from real traffic, so you hear about drift before a customer does.

A runbook

What to do when retrieval quality drops, a provider throttles you or a check fails.

A rollback path

Versioned prompts, models and indexes, so any release can be reversed in minutes.

Want your engineers to judge the code first? Ask for a 20-minute walkthrough of Switchyard, Assay or Plumblin. Each one runs end to end with no API key.

[Request a walkthrough >](#)

Search and automation across 5 million files that cannot leave their regions

Multinational enterprise · multi-region · AWS · client name withheld under NDA

1. PROBLEM

5 million+ files across regional data centers, S3, document systems, databases and application servers: contracts, financial statements, engineering reports, spreadsheets and email. The usual fix is 1 central index, and regional data rules ruled it out. So every cross-region question, and every workflow touching more than 1 system, depended on someone stitching answers together by hand.

3. IMPACT

5M+

files searchable across regions, with permissions intact

97.86%

task-success benchmark target on the agreed evaluation set

7

specialised agents coordinating over A2A

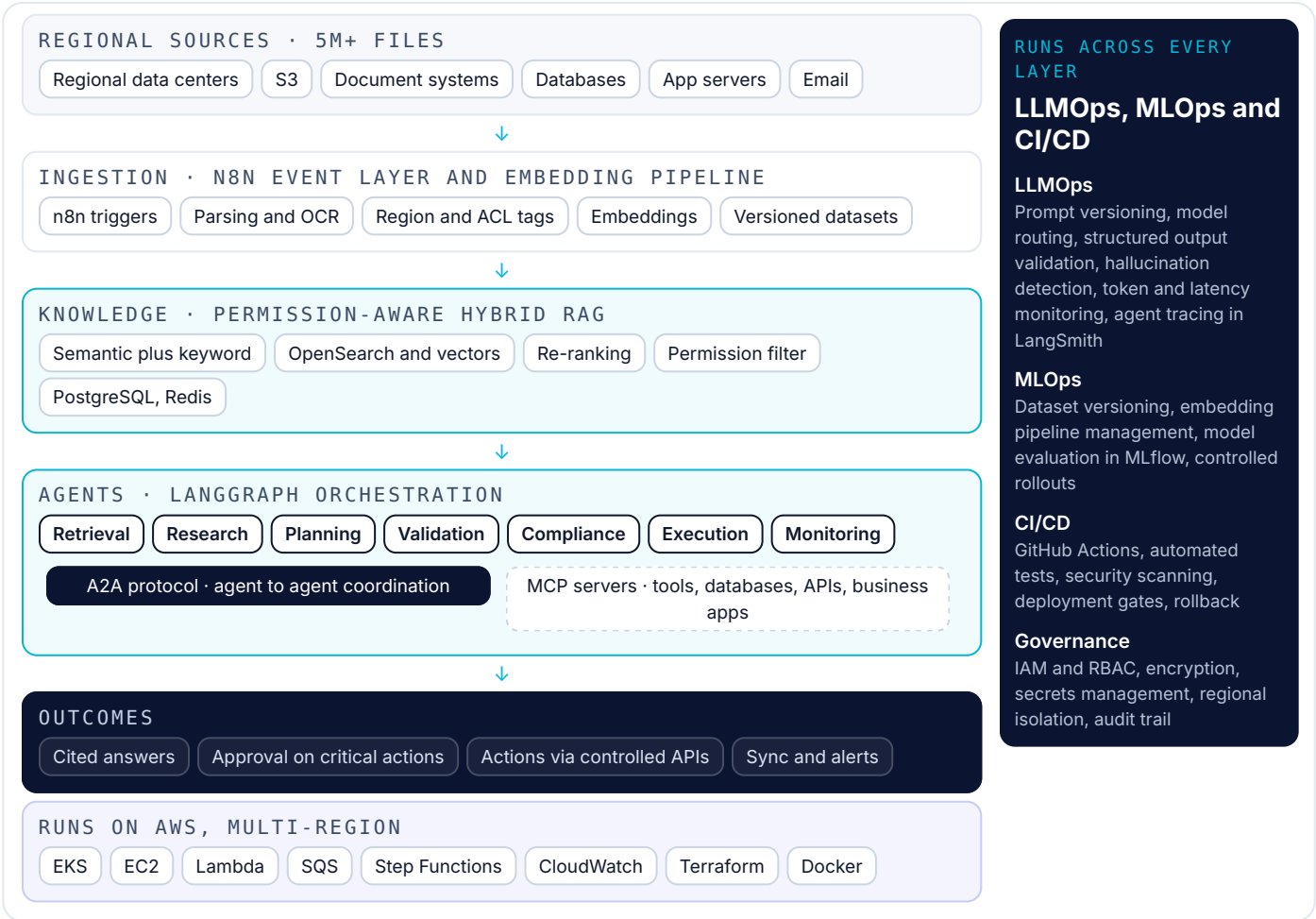
24/7

event-driven ingestion, approvals and notifications

12 months

6 delivered, 6 more contracted

HOW IT WORKS



3 DECISIONS THAT SHAPED IT

Local retrieval over a central index.

Slower to build, and it keeps every query inside the residency rules from the first day.

7 narrow agents over 1 general one.

Each agent is tested on its own, so a failure points at 1 component instead of the whole chain.

6 metrics over 1 accuracy number.

97.86% is the task-success target, with recall, faithfulness, latency, errors and permission violations reported separately.

Problem, solution and impact, measured against the client's own baseline

☐ System or data step
 ☐ AI component
 ☐ Human checkpoint
 ☒ Business outcome

LEGALTECH · USA AND UK

CASE 01

Legal LLM that associates trust without re-checking every citation

1. PROBLEM

Associates hand-checked every citation the old tool produced, so the hours it was bought to save never showed up. 1 bad citation in a filing is a malpractice risk.

2. SOLUTION

Gemma 3 27B fine-tuned with QLoRA on 400,000 of the firm's own documents, then aligned with DPO so a correct citation always scores above a confident wrong one. It runs on vLLM inside the firm's VPC, so no document leaves the building.

3. IMPACT

91% citation accuracy on jurisdiction queries

0 hallucinated citations after alignment

60% lower inference cost than the hosted model

HOW IT WORKS



Across every step: release gated on the held-out legal benchmark, LangSmith tracing, access logged per matter

STACK Gemma 3 27B QLoRA DPO Hugging Face TRL vLLM GCP Vertex AI LangSmith

HEALTHCARE · US HOSPITAL NETWORK

CASE 02

HIPAA-compliant clinical decision support at the point of care

1. PROBLEM

Records, labs, ICD mappings and literature sat in 6 formats across 4 systems. None of it could be searched in the 90 seconds a clinician has between patients.

2. SOLUTION

Hybrid agentic RAG on GCP Vertex AI across all 4 systems. Access comes first: retrieval checks the clinician's role before it reads a single record, and every answer leaves an audit entry.

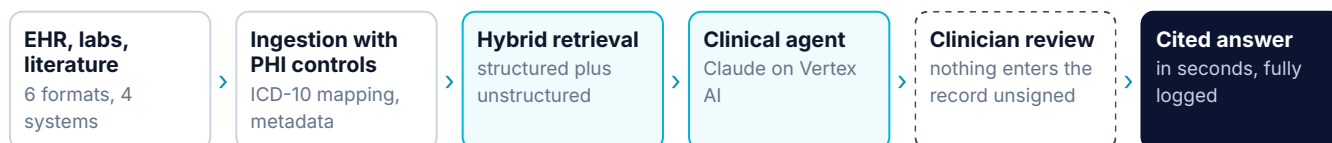
3. IMPACT

91% factual accuracy on the clinical test set

27% less patient intake time

48% less manual documentation

HOW IT WORKS



Across every step: HIPAA controls, role-based access, audit log per response, zero-downtime deployment

STACK Agentic RAG Claude on Vertex AI FHIR Qdrant PostgreSQL GCP HIPAA

Close to 1 of these? Book a call with me today and I will walk you through how the number behind it was measured.

[Book today >](#)

SAAS · USA AND UK

CASE 03

Revenue agents that work every lead across 14 disconnected sales tools

1. PROBLEM

Leads sat unqualified for days because no system owned the handoff between 14 tools, and nobody could show where the funnel was leaking.

2. SOLUTION

An n8n event layer catches every lead from all 14 tools the moment it arrives. LangGraph agents qualify it, enrich the CRM record and draft the outreach, and a rep approves every message before it goes out.

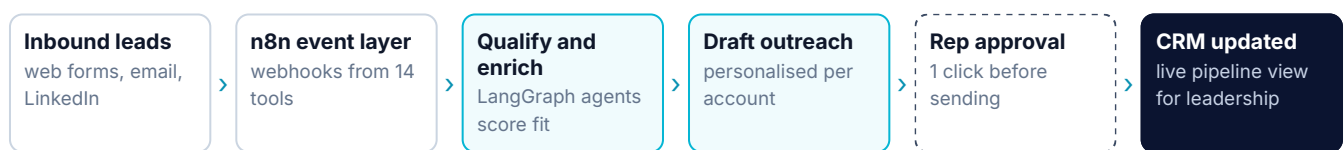
3. IMPACT

\$340K pipeline in the first 6 months

1.8x more qualified leads per month

900+ automated triggers a day

HOW IT WORKS



Across every step: 48% less manual sales ops, every agent action logged against the CRM record

STACK n8n LangGraph GPT-5.6 HubSpot API LinkedIn API PostgreSQL Langfuse

CUSTOMER EXPERIENCE · USA

CASE 04

Real-time voice agent that answers every call at peak

1. PROBLEM

Calls queued at peak and callers hung up before anyone answered. Those were leads the business had already paid to generate.

2. SOLUTION

A voice agent that picks up in under a second, answers from the client's knowledge base and books or updates the CRM during the call. Anything that needs judgement goes to a person with the full context already on screen.

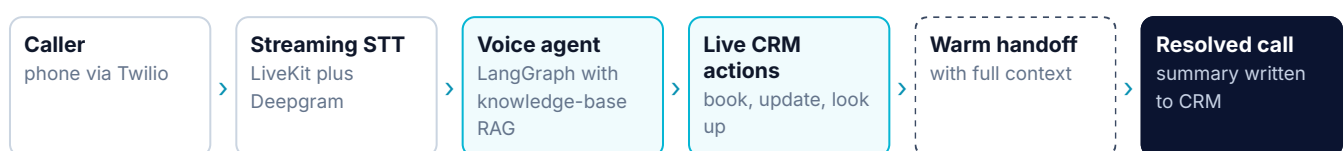
3. IMPACT

80%+ of tier-1 calls handled end to end

45% lower average handle time

<1s response latency, 24/7

HOW IT WORKS



Across every step: barge-in handling, call recording with consent, latency and resolution tracked per call

STACK Twilio LiveKit Deepgram OpenAI Realtime LangGraph RAG GCP

Prefer to start with your own case? Reply with 3 lines: the workflow, where the data lives and your deadline. I reply with a first read.

Send 3 lines >

REGTECH · GERMANY AND EU

CASE 05

Regulatory gap detection across 8 jurisdictions in 3 languages

1. PROBLEM

Every reporting cycle meant a manual document review, and outside counsel invoices went up each time the rules changed.

2. SOLUTION

Agentic RAG watches the regulatory feeds and maps each internal disclosure to the CSRD clause it has to satisfy. Gaps come back with evidence and a remediation note, and the board report writes itself in English, German and French.

3. IMPACT

84% gap classification accuracy

52% less manual ESG auditing per cycle

2.3x faster regulatory reporting

HOW IT WORKS



Across every step: LLMOps audit trail on every AI decision, GDPR data handling, monitored for drift

STACK Agentic RAG Azure OpenAI Azure AI Search LangGraph LLMops EN / DE / FR GDPR

ENTERPRISE SAAS · USA AND UK

CASE 06

Knowledge search that replaced 3 wikis and tribal memory

1. PROBLEM

Answers lived across 3 wikis, a shared drive and a few long-tenured people. Onboarding took months and every resignation took knowledge with it.

2. SOLUTION

Hybrid search plus a knowledge graph for questions that span several documents. Permissions are checked at retrieval time, so people only see what they could already open, and every answer names its source.

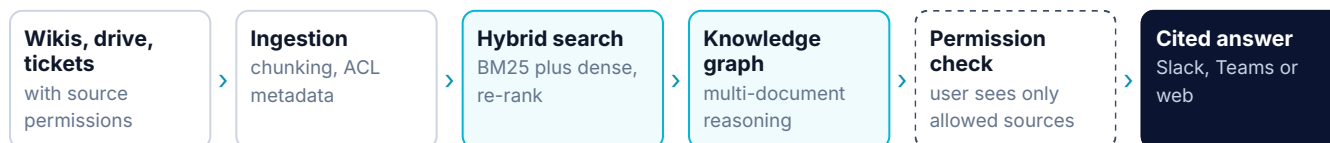
3. IMPACT

54% less internal search time

61% better retrieval precision

8,000+ queries a day

HOW IT WORKS



Across every step: 37% fewer knowledge-gap tickets, audit trail per response, quality scored weekly

STACK LangGraph RAG Fusion Neo4j Qdrant Elasticsearch Slack and Teams Langfuse

6 more engagements, detail available under NDA: LLMops governance for 5 production models at a US fintech with SOC 2 ready lineage · 7-channel support automation resolving 61% of queries without escalation · DevOps agents with 38% less debugging time · a private LLM inside an EU client's VPC with 100% data residency · retail forecasting with 62% better demand accuracy · program governance with 28% better on-time delivery.

Want numbers like these on your own data? The first scored result takes 14 days. Book a call today and we can agree the metric on the call.

[Book today >](#)

6 systems assembled from parts we have already shipped

Each reuses components from the engagements above, which keeps the first milestone to a few weeks on your own data.

FINTECH AND PAYMENTS

First milestone: **3 to 4 weeks**

Real-time fraud and transaction risk triage

1. PROBLEM

Rules engines flag too many transactions, analysts drown in false positives, and writing up each case takes longer than deciding it.

2. SOLUTION

A gradient-boosted model scores every transaction in the stream. An agent pulls history, device and KYC evidence for flagged ones and writes the case summary for the analyst.

3. VALUE

- Fewer false positives reach analysts
- Case review shrinks to approve or decline
- Every decision explained for auditors

Transaction stream
Kafka events

Feature store
velocity, device, geo

ML risk score
XGBoost, milliseconds

Evidence agent
history, KYC, linked accounts

Analyst decision
approve or block

Explained outcome
logged for regulators

INSURANCE, HEALTHCARE, LOGISTICS

First milestone: **2 to 4 weeks**

Claims and document intake with no retyping

1. PROBLEM

Claims, invoices and forms arrive as scans and email attachments, and staff retype every field into the core system.

2. SOLUTION

Vision-language extraction reads each document, validation rules check every field, and only low-confidence documents go to a person. Clean ones post straight to the core system.

3. VALUE

- Clean documents processed with no retyping
- Staff handle exceptions only
- Field-level audit trail on every document

Email, portal, scans
PDF, image, handwriting

VLM extraction
fields plus tables

Rules and schema checks
policy, limits, duplicates

Confidence gate
per field score

Exception queue
low confidence only

Core system
posted by API

B2B SAAS AND IT SERVICES

First milestone: **2 to 3 weeks**

RFP and security questionnaire agent

1. PROBLEM

Every RFP and security questionnaire pulls sales engineers off live deals for days, and answers drift from what the policies say.

2. SOLUTION

An agent drafts each answer from approved past responses, policies and product docs, cites the source, and flags questions with no approved answer for a subject expert.

3. VALUE

- Sales engineers review drafts instead of writing them
- Answers stay consistent with current policy
- Gaps in your answer library surface automatically

RFP upload
Excel, Word, portal

Question parser
splits and classifies

Answer retrieval
approved library, policies

Cited draft
per question

Expert review
flagged items only

Submission
in the buyer's format

1 of these on your roadmap? Book a call today and leave with the first milestone and a fixed price for it.

[Book today >](#)

OPERATIONS, FINANCE, HR

First milestone: **3 to 4 weeks****Back-office agent for systems with no API****1. PROBLEM**

Critical work still lives in vendor portals and legacy screens, so staff copy data between them by hand all day.

2. SOLUTION

A computer-use agent operates the browser like a trained clerk: it reads the task, navigates the portal, fills forms and checks the result, with approval on sensitive steps.

3. VALUE

- Legacy portals automated without an integration project
- Every step recorded with screenshots
- Staff move to exception handling

Task queue
email, ticket,
schedule



Planner agent
breaks task into
steps



**Computer-use
agent**
Playwright
browser



Result checks
field and total
validation



Approval
payments and
deletions



Task closed
with full replay log

HEALTHCARE PROVIDERS

First milestone: **4 to 6 weeks****Ambient clinical documentation****1. PROBLEM**

Clinicians spend evenings finishing notes, and coding errors from rushed documentation come back as denied claims.

2. SOLUTION

With patient consent, the visit is transcribed in real time, a draft note and ICD-10 and CPT code suggestions are generated, and the clinician edits and signs before anything reaches the EHR.

3. VALUE

- Less after-hours charting for clinicians
- Coding suggestions checked against the note
- PHI stays inside your cloud

Visit audio
consented,
encrypted



Medical STT
streaming,
speaker labels



Note generation
SOAP draft



**Code
suggestions**
ICD-10 and CPT



Clinician signs
edits before
saving



EHR write-back
via FHIR API

ANY REGULATED COMPANY

First milestone: **3 to 5 weeks****Private company assistant that respects permissions****1. PROBLEM**

Staff paste company data into public chatbots because the internal tools cannot answer their questions.

2. SOLUTION

A company assistant in Teams, Slack or web that answers only from sources each user is already allowed to see, uses MCP tools for Jira, CRM and SharePoint, and runs on a private model when data must stay in your network.

3. VALUE

- Company data stays inside approved systems
- Answers respect existing access rights
- Cost and usage visible per team

**Employee via
SSO**
Teams, Slack, web



**Permission-
aware search**
mirrors source
ACLs



Assistant agent
MCP tools: Jira,
CRM, SharePoint



Model gateway
private model or
approved API



Guardrails
PII and policy
checks



Cited answer
logged per user

1 of these on your roadmap? Book a call today and leave with the first milestone and a fixed price for it.

[Book today >](#)

In their words

★★★★★

Jillani SofTech delivered a mission critical system that changed how our organisation operates. Their revenue execution platform created the pipeline intelligence we had tried to build for 2 years. The results were immediate and measurable.

Michael Stevens

CTO · TECHCORP SOLUTIONS · USA

★★★★★

The clinical intelligence platform operates at a level we expect from top tier enterprise vendors. Factual accuracy, HIPAA compliant architecture, and real time performance had a direct impact on patient care quality and throughput across our network.

Dr. Rachel Chen

CHIEF MEDICAL OFFICER · HEALTHTECH INNOVATIONS · USA

★★★★★

Most vendors talk about compliance. These engineers build for it. Their platform reduced our risk exposure significantly while improving audit readiness across multiple jurisdictions.

Lisa Thompson

CHIEF COMPLIANCE OFFICER · GLOBAL ENTERPRISES · UK

★★★★★

The systems are solid, the documentation thorough, and the team stayed accountable well past delivery.

Evan Solomon

CEO · EFS NETWORKS · USA

TECHNOLOGY WE WORK IN

Current stack, chosen per engagement by your evaluation set

Models	GPT-6 and GPT-5.6 · Claude Fable 5.1, Opus 5.5 and Sonnet 5.5 · Gemini 3.8 · open-weight Qwen 3.8, DeepSeek V4.1, Gemma 3
Agents and automation	LangGraph · OpenAI Agents SDK · Claude Agent SDK · Google ADK · CrewAI · n8n · MCP and A2A
Retrieval	Qdrant · pgvector · Pinecone · Weaviate · Neo4j GraphRAG · hybrid search · cross-encoder re-ranking
Fine-tuning and serving	QLoRA · LoRA · DPO · Hugging Face TRL · vLLM · private VPC and on-premises deployment
LLMOps	Langfuse · LangSmith · Arize Phoenix · MLflow · CI evaluation gates · drift detection
Voice and vision	LiveKit · Twilio · Deepgram · ElevenLabs · OpenAI Realtime · vision-language models · Playwright computer-use agents
Cloud	AWS Bedrock and SageMaker · GCP Vertex AI and BigQuery · Microsoft Foundry and Azure OpenAI · Kubernetes · Terraform

CREDENTIALS YOU CAN CHECK

Claude Certified Architect, Foundations · verify on Credly

AWS, Azure and GCP professional certificates

Top 100 global Kaggle Master and KaggleX BIPOC mentor

LinkedIn Top Voice in AI, 24 times

Upwork Top Rated Plus · 100% Job Success · \$100K+ earned

NVIDIA, AMD and GitHub developer programs

Need this cleared by procurement first? Reply with your MSA or vendor form and we will return it signed or marked up within 2 working days.

Send your MSA >

Published pricing, so you can budget before the call

Architecture Review

From \$2,500

FIXED PRICE · 2 WEEKS

Written architecture, build or buy call, and the monthly running cost at your real volume. The lowest-risk way to start.

Focused Build

From \$3,500

2 TO 4 WEEKS

1 RAG system, agent, voice agent or automation, scoped to 1 success metric and shipped with evals and monitoring.

Enterprise AI Platform

\$15,000 to \$60,000+

8 TO 16 WEEKS

Multi-agent systems, fine-tuned models, LLMOps and governance, integrated with your existing stack.

Retainer and Team

\$4,800 to \$6,000

PER MONTH · 6 MONTHS MINIMUM

Dedicated engineering capacity, monitoring and improvement cycles, or AI engineers placed inside your team.

Ranges are planning guides. Your final price is confirmed in writing after the scoping call, based on data volume, integrations and compliance requirements.

BEFORE YOU ASK

The 4 questions buyers raise most

Who owns the code and the model weights?

You do, on final payment. No licence that restricts you and no dependency that forces you to keep paying us to run it.

How do we know your numbers are real?

Ask us to walk you through the evaluation behind any figure in this profile. Client references are shared under NDA.

Our data cannot leave our environment.

Then it will not. We run open-weight models on vLLM inside your VPC, and we have done this for a regulated EU client with 100% data residency.

How fast can we start?

Scoping call this week, written scope within 2 working days, build starts when you sign. A first scored result lands inside 14 days.

.JS JILLANI SOFTECH
ENTERPRISE AI ENGINEERING COMPANY

If you have 30 minutes today, let's look at your system together.

I take these calls myself. You get the approach I would take, what I would leave out and a rough cost, with no sales rep in between. If AI is the wrong answer for your problem, you will hear that on the call too.

FASTEST

Book a call for today >

Take the first open slot.
calendly.com/jillanisoftech

SAME-DAY REPLY

WhatsApp us >

+1 501 420 2439. Send 1 line about the problem and I reply the same day.

PREFER EMAIL

Reply with 3 lines >

The workflow, where the data lives, and your deadline. I send back a first read and times to talk.

Muhammad Ghulam Jillani

Founder and CEO, Principal Generative AI Architect, Jillani jillanisoftech.com · LinkedIn · Fully remote, serving the USA, UK, Germany and the EU

m.g.jillani@jillanisoftech.com