

The Enterprise AI Deployment Whitepaper

AI in the workflow. Results on the ledger.

20 published industry results · why 95% of projects die · a method you can sign off on

The whole book on one page

Four parts, one thread: AI's value is not in the demo. It's on the ledger.

Part 01 · Evidence

The people and money others have already saved

20 published, verifiable deployments: the work of 700 agents, half a QC headcount, a mine earning ¥20M more a year. From global giants to 200-person companies, from China to the Middle East.

Part 02 · Why

Why 95% of projects die between pilot and production

The model is never the cause of death — never embedded, never accounted for, data locked in, no acceptance system. Four causes, each with a cure.

Part 03 · Method

How to get AI onto the ledger

Define 'good' before building; no launch below threshold; every action audited, every dollar attributed; data stays home, models come in.

Part 04 · Action

How to start

Step one is a 2–4 week bootcamp: embedded engineers stand up your first real use case on your data, charged at cost and fully credited to your first milestone. After 90 days, your desk holds a running system and a ledger that adds up.

How to read this: five minutes — this page and the number wall on page 3. Fifteen minutes — just the takeaway line on each page. Ready to act — copy the deployment checklist at the back for your team.

TAKEAWAY AI's value was never at the launch event. It's on the ledger.

01

PART 01 · EVIDENCE

The people and money others have already saved

20 numbers, all from published, real deployments.



20 numbers that already happened

Each tile is one published, real deployment.

= 700 agents AI customer service Klarna	360,000 hrs → 0 Contract review JPMorgan	Screening –60% AML monitoring HSBC	1,000 engineers +20% AI pair programming Emirates NBD · Dubai
\$500M a year AI across the value chain ADNOC · Abu Dhabi	+¥20M a year Coal-washing optimization Shandong Energy	Power –20% Water-plant tuning Jacobs	Replaced 140 inspectors AI quality control TCL CSOT
Defects –67% Line-side vision AI Pegatron	QC cost –70% AI fabric inspection 50 textile mills, China	Output +33% Final assembly Airbus	Procurement –3% · 11 days AI negotiating with 2,000 suppliers Walmart
Picking +30–180% Warehouse robots + AI DHL	40 sorters, 3× volume Automated sorting ZTO Express	1,200 staff redeployed AI security patrol Onewo (Vanke)	¥7.5M saved a year Hotel operations AI Atour Group
Forecasts 10–15 pts better 16,000 stores Yum China	Operating cost –35% Government services AI Government of Dubai	7× customers, zero hires 400-person company Synthesia	\$4.4M saved Credit-union call line Neighborhood CU

Finance
Energy & industry
Manufacturing & QC
Supply chain
Services & chains
Mid-size & SMB

Sources in the appendix. Your numbers get verified by your own audit ledger.

TAKEAWAY

These numbers already happened. The only question left is when your business gets its turn.

Tighter risk control, fewer people

SUPPORT · KLARNA × OPENAI

1 AI = 700 agents

Took over two-thirds of support chats in its first month; resolution 11 min → 2 min; \$40M more profit in a year

Klarna, 2024 ①

CONTRACT REVIEW · JPMORGAN

360,000 hrs/yr → seconds

Commercial-loan contract review fully automated — the annual workload of ~180 people

Bloomberg, 2017 ②

AML · HSBC × GOOGLE CLOUD

Screening –60%, catches x2–4

Machine learning replaced the rules engine as the primary monitoring system

Google Cloud, 2023 ③

BRANCH OPS · ICBC × PANGU

5 steps → 1

LLM generates the transaction path; 5+ minutes saved per transaction

Securities Times ④

ENGINEERING · EMIRATES NBD × GITHUB

1,000 engineers +20%

≈ 200 engineers' output added — at one of Dubai's largest banks

GitHub customer story ⑤

Run your own numbers: a 200-seat contact center → ≈120 seats of workload; a 10-person review team → 3 people saved. Your exact numbers get computed on your own data in week one of a bootcamp.

TAKEAWAY

If the most regulated industry can do it, “our industry is different” is no longer a reason.

The heavier the assets, the bigger the savings

GROUP-WIDE · ADNOC

\$500M in one year

30+ AI tools from exploration to sales; value measured and publicly reported

ADNOC, 2024 ⑥

MINING · SHANDONG ENERGY × PANGU

One mine, +¥20M a year

Clean-coal yield +0.2%; new use cases now ship in 1 month instead of 6

Securities Times ⑦

UTILITIES · JACOBS × PALANTIR

Plant power –20%

A further 20% saved at a water plant that had already been conventionally optimized

Palantir Impact ⑧

MAINTENANCE · CROSS-INDUSTRY RANGE

Downtime –30–50%

Predictive maintenance: cost –10–40%, asset life +20–40% (conservative ranges)

McKinsey research ⑨

Run your own numbers: one line's cost per hour of downtime × annual downtime hours × 30% — that is the opening entry for predictive maintenance alone.

TAKEAWAY In heavy industry, every percentage point is worth millions.

The QC line is where AI replaces the most people

PANEL QC · TCL CSOT

Replaced 140 inspectors

QC headcount –50%; per-image judgment 2s → 0.5s;
¥10M+ saved across three Shenzhen fabs

Leiphone, 2025 ^⑩

LINE VISION · PEGATRON × NVIDIA

Labor –7% · defects –67%

Vision AI on the assembly line cuts headcount and
scrap at once

NVIDIA case study ^⑪

FABRIC INSPECTION · 50 MILLS

Inspection cost –70%

30 inspectors → 15, at 3–5× the speed; missed defects
20–30% → under 3%

China Textile Weekly ^⑫

FINAL ASSEMBLY · AIRBUS × PALANTIR

A350 output +33%

Capacity is revenue — production bottlenecks cleared
with data

Palantir / Airbus ^⑬

Run your own numbers: a 10-person QC crew at the published –50% ≈ 5 people saved; across three shifts, 15.

TAKEAWAY The most labor-dense station is the easiest math.

One AI, negotiating with 2,000 suppliers at once

PROCUREMENT · WALMART × PACTUM

Cost **–3%**

AI negotiates with 2,000 tail suppliers in parallel; cycles of weeks → 11 days; 68% close rate

Harvard Business Review ¹⁴

PICKING · DHL × LOCUS

+30–180% per person-hour

Collaborative robots + AI dispatch, proven across 40+ warehouses and a billion picks

DHL press ¹⁵

PARCEL SORTING · ZTO EXPRESS

40 sorters, 3× the volume

Automated sorting line at 750,000 parcels a day — same crew, triple the throughput

The Paper ¹⁶

FORECASTING · CROSS-INDUSTRY RANGE

Forecast error **–20–50%**

Inventory –20–30%, logistics cost –5–20%

McKinsey ¹⁷

Run your own numbers: ¥500M annual procurement, tail categories at the published –3% ≈ ¥4M a year, scaled by your tail share.

TAKEAWAY

Supply-chain savings come a cent at a time — AI saves across 2,000 suppliers at once.

The bigger the payroll, the easier the math

PROPERTY · ONEWO (VANKE)

1,200 staff redeployed

AI remote patrol replaced on-site rounds; output per person $\times 2.3$ across 479 estates

Dute News / 36Kr ^⑱

HOTELS · ATOUR GROUP

¥7.5M saved a year

300,000+ work-hours of operations and service automated

The Paper ^⑲

RESTAURANTS · YUM CHINA

Forecasts beat managers by 10–15 pts

Demand forecasting and ordering across 16,000 stores; 95% of managers use it daily

Company disclosure ^⑳

GOVERNMENT · DUBAI

Operating cost –35%

60%+ of routine enquiries resolved by AI end-to-end; service 25% faster (15 use cases)

Digital Dubai ^㉑

Run your own numbers: a 1,000-person property or chain business at Onewo's redeployment ratio $\approx 100+$ roles reallocated.

TAKEAWAY The bigger the payroll, the easier the AI math.

Not a giants' game

400-PERSON COMPANY • SYNTHESIA

7× customers, zero support hires

87% self-serve resolution; handling time 5 days → 4.6 hours

Intercom customer story ²²

REGIONAL CREDIT UNION • NEIGHBORHOOD CU

90% of calls handled within 48 hours

Voice AI live in two days; \$4.4M saved in support operations

interface.ai case study ²³

MID-SIZE CONSULTANCY • ELIASSEN

6,000 hrs/yr = 3 full-timers

Month-end close automated across 80+ branch entities

Eliassen ²⁴

JIANGSU SME • JINSHI ROBOTICS

Service tickets: 2 weeks → 1–3 days

AI ticket assistant connected to 1,000 service partners

QbitAI, 2025 ²⁵

Forrester's independent study of a composite **200-person, \$35M-revenue** business: operating cost **–20%**, net revenue **+6%**.

TAKEAWAY

A 200-person company runs the same models as a 200,000-person one. The only difference is deployment.

7× the customers. Zero new support hires.

SYNTHESIA · AI VIDEO SOFTWARE

~400 PEOPLE · MID-SIZE TECH

SCENE: CUSTOMER SUPPORT

THE SQUEEZE

Customers grew from 40,000 to 316,000 in two years; support volume rose nearly sevenfold. The old playbook said: grow the support team sevenfold too.

WHAT THE AI DID

An AI agent holds the front line: routine questions resolved outright, complex ones escalated to humans; missed answers flow back into the knowledge base, so it keeps getting sharper.

87%

of enquiries self-served

5 days → 4.6 hrs

resolution time

1,300+ hours

of labor saved in six months

+0

support headcount added

Source: Intercom customer story (vendor-published, customer named)

TAKEAWAY

Business grew 7×. Headcount grew by zero.

One hotline. \$4.4M saved in a year.

NEIGHBORHOOD CU · DALLAS, USA

~\$1B ASSETS · REGIONAL CREDIT UNION

SCENE: MEMBER HOTLINE

THE SQUEEZE

The hotline could only field about half of incoming calls: peak-hour queues, member complaints — and the payroll cost of staffing up was unaffordable.

WHAT THE AI DID

Voice AI answers every call first: balances, card blocks, transfers and other routine business completed outright; complex matters handed to human agents.

48 hours to go from 50% → 90% of calls handled	100% of calls answered by AI first	\$4.4M saved in support operations	43,000+ hours of member waiting eliminated
--	--	--	--

Source: interface.ai customer story (vendor-published, customer named)

TAKEAWAY **Live in two days. \$4.4M saved in year one.**

20 cases, three things in common

1 Each lands on a specific workflow

Not one case is “roll AI out company-wide”. Every one is a single workflow: answering calls, reviewing contracts, inspecting fabric, sorting parcels, ordering stock.

2 Each adds up on a ledger

People replaced, money saved, time cut — every case fits in one line of numbers. Projects that don't add up aren't in these 20.

3 None is a one-off project

Go-live is where the metering starts: numbers keep accruing, keep getting accepted, keep extending to the next workflow.

TAKEAWAY AI's value was never at the launch event. It's on the ledger.

02

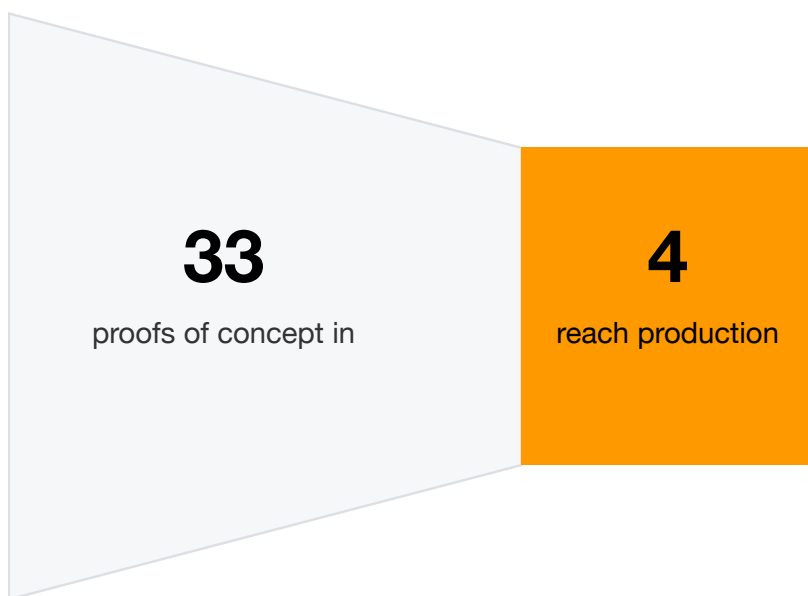
PART 02 • WHY

Why 95% of projects die between pilot and production

Part 01 showed the survivors. This part is about the majority: dazzling demos, smooth kickoffs — then a quiet death. And the causes they share.



Between the demo and the ledger lies a 95% gap



IDC: of every 33 AI proofs of concept, only 4 reach production

95%

of organizations report no measurable P&L return from their GenAI pilots (MIT, 2025)

Capability ≠ consistency

A demo needs one or two shining examples; production needs ten out of ten on every request. Chain multi-step reasoning, tool calls and external systems together, and every link's randomness compounds. (AWS, Guide to Production-Ready Enterprise AI Agents, 2026)

TAKEAWAY

AI that never reaches production is no AI at all.

Anatomy of the 95%: four causes of death — none is the model

CAUSE 01

Never embedded

The model hangs outside the business. However good the answers, if they never reach the order, risk or ticketing systems, it stays a chat window.

CAUSE 02

Nobody can do the math

Nobody can answer the CFO's one question: "how much did we save?" Budgets that can't be accounted for get cut the next year.

CAUSE 03

Data can't leave

Regulated environments can't call public-cloud APIs. The pilot runs on sanitized toy data forever — and stays a pilot forever.

CAUSE 04

No acceptance system

Nobody ever defined what "good" means. Only after go-live do you discover: win or lose, there's no way to score it.

TAKEAWAY The model is never the cause of death.

Build, buy — or co-deliver

	Build in-house	Co-deliver with TokenPower	Off-the-shelf SaaS
First production value	12–18 months+, half of it hiring	4–8 weeks to a live workflow	Fast, but stops at generic scenes
3-year total cost	Roughly 2–3× the co-delivery route	Milestone-priced, written down up front	Cheap subscription; integration extra
AI talent	Hire and retain your own	Embedded engineers + your domain experts	No one owns deployment
Model churn	4–8 major releases a year, absorbed by you	Absorbed by the multi-model gateway — no lock-in	Locked to one vendor
Compliance & audit	An audit stack from scratch	End-to-end audit & attribution built in	Usually absent
Data & IP	Yours	Yours — data stays in-domain	Depends on vendor terms

MIT, 2025: purchased tools and partnerships succeed about **67%** of the time; internal builds about a third as often. Menlo Ventures: internal builds fell from 47% to **24%** of enterprise AI use cases in one year.

TAKEAWAY The market has already voted.

Four causes, four cures

Never embedded



Embedded co-building Engineers write code inside your systems — not advice across a requirements document

Nobody can do the math



End-to-end audit Every action, every dollar, every gain: queryable and attributable

Data can't leave



Sovereign deployment The data doesn't move; the models move in — local DC / dedicated cloud / air-gapped

No acceptance system



Define “good” first A co-signed acceptance baseline; thresholds gate the launch

TAKEAWAY Every cause of death has a cure — that's the next part.

03

PART 03 · METHOD

How to get AI onto the ledger

Four cures — one section per cause.

Four steps of delivery on four layers of platform

THE DELIVERY LOOP · A CYCLE, NOT A PIPELINE



FOUR LAYERS · ONE AUDIT TRAIL

DEPLOY	Critical environments · private / sovereign / air-gapped
AGENTS	Trusted execution · decisions become actions, inside permissions and audit
GRAPH	Business-object graph · your data becomes AI-callable objects
GATEWAY	Model access & governance · routing / cost / audit / policy — live in production

• IN PRODUCTION

Delivery is a loop, not a line; the platform is a foundation, not a toolbox. Each of Part 02's four causes has its cure on this page.

TAKEAWAY

Delivery follows a process. Trust runs on a system. Neither runs on gut feel.

The first thing we build isn't code — it's the definition of “good”

Week one of a bootcamp produces this table, signed by both sides.

METRIC	TARGET THRESHOLD	ACCEPTED BY
Tool / system call selection	≥ 95%	Both teams
Key-parameter extraction	≥ 98%	Both teams
Refusal of out-of-scope requests	100%	Client compliance
Latency	P50 < 2s · P95 < 5s	Client IT
Cost per transaction	< agreed ceiling	Client finance
Business metrics (examples)	Handling time –X% · self-serve ≥ Y% · false positives –Z%	Client business owner

Technical metrics

answer “is the AI doing it right”

Business metrics

answer “is the AI worth it”

Both rows, always

missing either one isn't acceptance

TAKEAWAY Define “good” before you build — blueprints before the building.

Three-step delivery: every step accepted, every step a gate

① Bootcamp

2–4 WKS · AT COST, CREDITED
IN FULL

Embedded engineers stand up your first real use case on your data; the acceptance baseline is co-signed.

You get: a running system
your board can see

A gate between every step: the
threshold is the ticket

② Gated launch

4–8 WKS · MILESTONE FEES

The first workflow wired into core systems; audit and metering on. Below threshold: no launch, no fee.

You get: savings that start
accruing

Acceptance is the baseline table's
call, not a feeling

③ Scale

QUARTERLY · PLATFORM
SUBSCRIPTION

Solution assets reused;
adjacent workflows added;
unit cost falls each quarter.

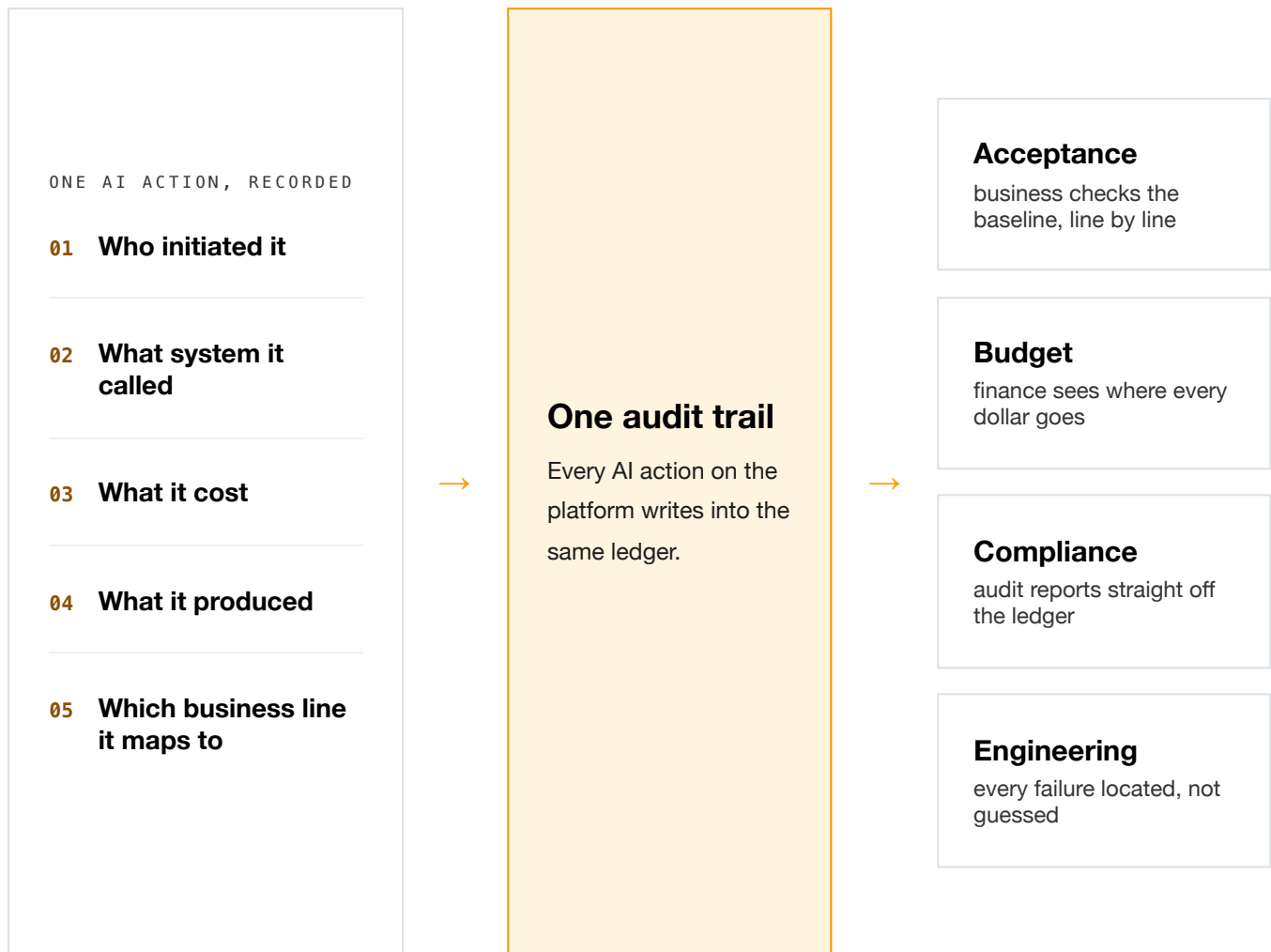
You get: expansion accepted
quarter by quarter

If you stop, the system, data and
golden set stay with you

TAKEAWAY

Every milestone is accepted — or it doesn't ship.

Audit isn't logging. It's the product.



The report on your desk at day 90 — people saved, money saved, money earned — **every line opens up to the actions that produced it.**

TAKEAWAY Here, trust is a queryable database.

Governance doesn't rely on the model's goodwill. It relies on the gateway.

✗ Write it into the prompt

`"Please don't access data you're not entitled to"`

Relies on the model behaving. Probabilistic; injectable around; unreproducible when it fails.

✓ Put it in the gateway

`Authorization completes before the request reaches any tool`

Deterministic. Permissions, rate limits, policy and audit all happen before the model sees the request — the model never gets to “decide” whether to comply.

“Externalize security controls: put them in a dedicated **Gateway** and Policy layer, and complete authorization before any tool is invoked — the LLM never gets the chance to ‘decide’ whether to follow the rules.”

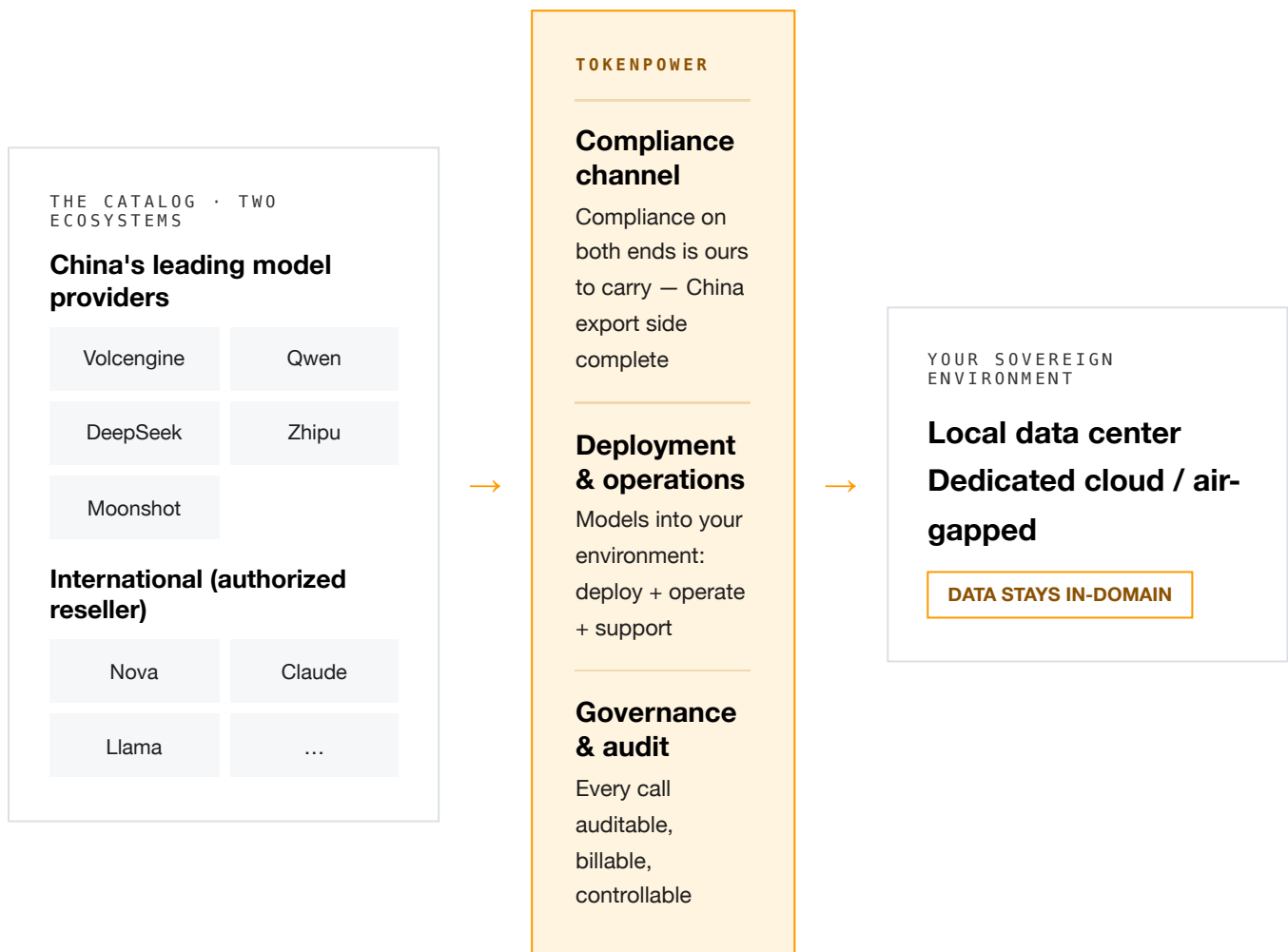
— AWS, GUIDE TO PRODUCTION-READY ENTERPRISE AI AGENTS · 2026

Our gateway layer is already running in production, handling live traffic every day.

TAKEAWAY

Governance doesn't rely on the model's goodwill. It relies on the gateway.

Data stays home. Models come in.



The in-domain loop: cloud-hosted evaluation and monitoring tools can't enter an air-gapped network — so we bring the whole toolchain inside: models run in-domain, reports come out in-domain, acceptance happens in-domain.

TAKEAWAY Data stays home. Models come in.

Models age. Golden sets appreciate.



AWS calls this class of asset “evaluation IP”

“Evaluation platforms can be bought; the authority to define ‘what counts as good’ cannot — outsource your golden set and rubrics, and you have handed that authority to someone else.” Evaluation content must stay under your control.

Ownership, written into the contract

Your golden set is yours — data, labels and acceptance standards are your assets. What we keep is the method and the engineering, which is why the second project in an industry costs about half.

TAKEAWAY Models age. Golden sets appreciate.

We charge for outcomes because outcomes are audited

THE STANDARD PATH · MILESTONE FEES

Every step priced in writing, before work starts

- 01 Bootcamp: at cost, credited in full to your first milestone
- 02 Launch: fees follow the co-signed milestones
- 03 Scale: platform subscription, renewed quarterly

OPTIONAL · OUTCOME SHARE

No result, no fee

When results land, we share the **audited** savings. We can price this way because the audit platform itself is the measuring instrument.

Futurum, 2026: outcome-based pricing already covers 21.7% of enterprise AI contracts.

TAKEAWAY We charge for outcomes because outcomes are audited.

04

PART 04 · ACTION

Start with a bootcamp

You don't have to decide everything at once. Step one is a 2–4 week bootcamp — embedded engineers stand up your first real use case on your data, charged at cost and credited in full to your first milestone.



90 days on site: what lands on your desk



WEEK 4

A running system

Your first real use case, on your own data — not a deck, not a report



WEEK 8

A workflow in production

Wired into core systems; savings start accruing



DAY 90

An audit report

People saved, money saved, money earned — every line traceable



THROUGHOUT

Acceptance by milestone

Data stays in-domain · every entry attributable · below threshold: no launch, no fee

Bootcamp pricing: at cost, credited in full to your first milestone. Week one starts with the co-signed acceptance baseline — **if the numbers don't add up, we stop there.**

TAKEAWAY

After 90 days you don't have a report. You have a system and a ledger.

The three most common concerns, answered on paper

You own

Domain knowledge · data · decision rights
IP (including your golden set)

We run

Platform · model operations
Governance & audit · delivery accountability

Data security

Data stays home; models come in

Local DC / dedicated cloud / air-gapped; export-side compliance complete

Unclear ROI

Every action audited, every dollar attributed

The ledger is open at any time; reports trace line by line

Vendor dependence

No lock-in — the contract is segmented

Milestone-segmented contracts: after each acceptance, continuing is your call; delivered systems, data and golden sets are yours

Partners

AWS (authorized reseller of cloud & model products) · Microsoft · Volcengine · Qwen · DeepSeek · Zhipu · Moonshot — two model ecosystems, one catalog, neutral by design.

TAKEAWAY Concerns don't need persuading. They need terms and a ledger.

The deployment checklist (copy it for your team)

- | | |
|--|---|
| ☐ 01 Is there an acceptance baseline table signed by both sides? | ☐ 02 Is the first workflow the one where failure costs most? |
| ☐ 03 Is the AI wired into core systems — or hanging outside as a chat window? | ☐ 04 Does every AI action leave a trace? |
| ☐ 05 Can every dollar of cost be attributed to a business line? | ☐ 06 Has data left your domain? Does compliance know? |
| ☐ 07 Models ship 4–8 major releases a year — who absorbs that? | ☐ 08 Are there explicit thresholds for false positives and misses? |
| ☐ 09 Is launch gated by thresholds — or by gut feel? | ☐ 10 Do production failures flow back into improvement? |
| ☐ 11 If you stopped tomorrow, who keeps the assets? | ☐ 12 On day 90, can your CFO see a ledger? |

If three of the twelve have no answer, a bootcamp is worth booking.

Book a bootcamp — in 2–4 weeks, your leadership sees a running system.

TOKENPOWER.AI

Where every number comes from

Grades: (A) media / official / research-institution reporting | (B) vendor-published named-customer case (not independently audited). Full links on request.

CASE	SOURCE	GRADE
Klarna	Klarna press, Feb 2024 (note: some human agents rehired May 2025 — hence this book’s stance: governed human–AI collaboration, not zero-staffing)	(A)
JPMorgan COiN	Bloomberg, Feb 2017	(A)
HSBC x Google Cloud	Google Cloud newsroom, Jun 2023	(A)
ICBC x Pangu	Securities Times / Huawei	(A)
Emirates NBD x GitHub	GitHub customer story	(B)
ADNOC	ADNOC press release, Jan 2024	(A)
Shandong Energy x Pangu	Securities Times (Jining No.2 mine)	(A)
Jacobs x Palantir	Palantir Impact	(B)
TCL CSOT	Leiphone / China.com Tech, 2025	(A)
Pegatron x NVIDIA	NVIDIA case study	(B)
AI fabric inspection (50 mills)	China Textile Weekly, 2025	(A)
Airbus x Palantir	Palantir / Airbus Skywise public materials	(B)
Walmart x Pactum	Harvard Business Review / Bloomberg, 2022–2024	(A)
DHL x Locus	DHL Group press releases, 2024–2026	(A)
ZTO Express	The Paper	(A)
Onewo (Vanke)	Dute News / 36Kr, 2025	(A)
Atour Group	The Paper (case awards), 2025	(A)
Yum China	Company disclosure, 2025	(A)
Government of Dubai	Digital Dubai (DCAI) public report	(A)
Synthesia x Intercom	Intercom (fin.ai) customer story	(B)
Neighborhood CU	interface.ai customer story	(B)
Eliassen	Eliassen case study	(B)
Jinshi Robotics	QbitAI, Jul 2025	(A)

Methodology references: AWS, Guide to Developing and Deploying Production-Ready Enterprise AI Agents (2026) · MIT NANDA, The GenAI Divide (2025) · IDC x Lenovo, AI CIO Playbook (2025) · Menlo Ventures (2025) · Forrester TEI, SMB (2024) · Futurum Group survey, n=838 (2026) · McKinsey predictive-maintenance research

TokenPower

TokenPower

See clearly, then decide.

TOKENPOWER.AI · DUBAI · 2026