

Your service agents should scale exponentially. Your AI costs should not.

Why AI-as-a-Service returns more and costs less to run than building it or bolting on what your platform vendor ships. This guide shows how.

TRUSTED BY:



Most agentic AI business cases only do half the math

Service leaders looking to capitalize on agentic AI typically focus on what the agent can deliver: cost savings, revenue retention, and recovered margin. But that's only half the equation. To understand the real return, you also need to know what it costs to build, deploy and run the agent. The **agent value equation** makes that comparison:

NUMERATOR

DENOMINATOR

VALUE THE AGENT RETURNS
Cost savings, revenue retention, recovered margin

COST TO BUILD, DEPLOY AND RUN IT
IT build, integration, compute, data prep, upkeep

=

AGENT VALUE MULTIPLE

Value delivered per dollar spent

The top half shows the value of better service and is what drives the business case and budget request: fewer truck rolls, lower support costs, and protected revenue. The bottom half is where things can spiral quickly out of hand.

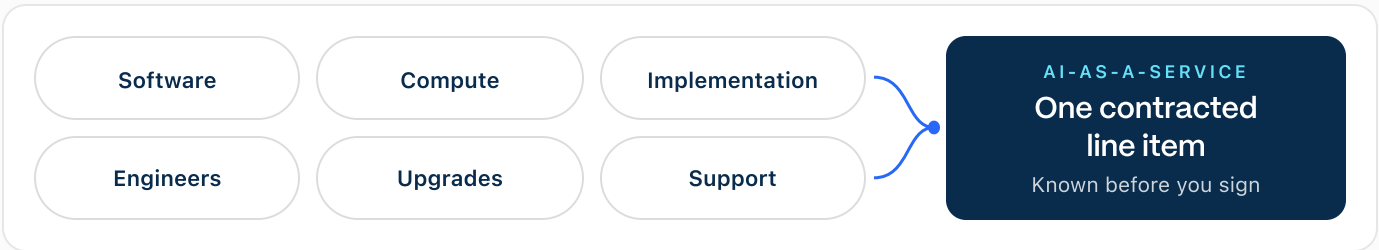
The IT build cost hides in salaries already on the books and the deployment cost hides in an integration backlog. The TCO is spread across a cloud bill, a GPU line, and the data engineering sprint just to make one source of service data (typically PDFs) machine-readable. By the time someone asks what the agent actually cost, the answer lives in four budgets across two fiscal years.

Better service outcomes are only half the business case. The full cost determines the return.

AI-as-a-Service makes both sides of the **agent value equation** easier to evaluate. The denominator becomes one number you can measure before you sign. The numerator is proven by customers already running it in production.

WHAT IS AI-AS-A-SERVICE?

AI-as-a-Service bundles everything required to deploy and operate AI for service into one contracted service. The software, compute, implementation, engineers who build and tune it, upgrades, and ongoing support are all included. Instead of managing separate vendors, budgets, and resources, you get one predictable cost.



One half you contract. One half compounds.

With AI-as-a-Service your cost is predictable because it is one contracted number rather than a program you staff. Your return is not predictable in the same way, because it keeps rising: every service decision feeds the **Service Decision Graph** and the system gets more accurate the longer it runs.

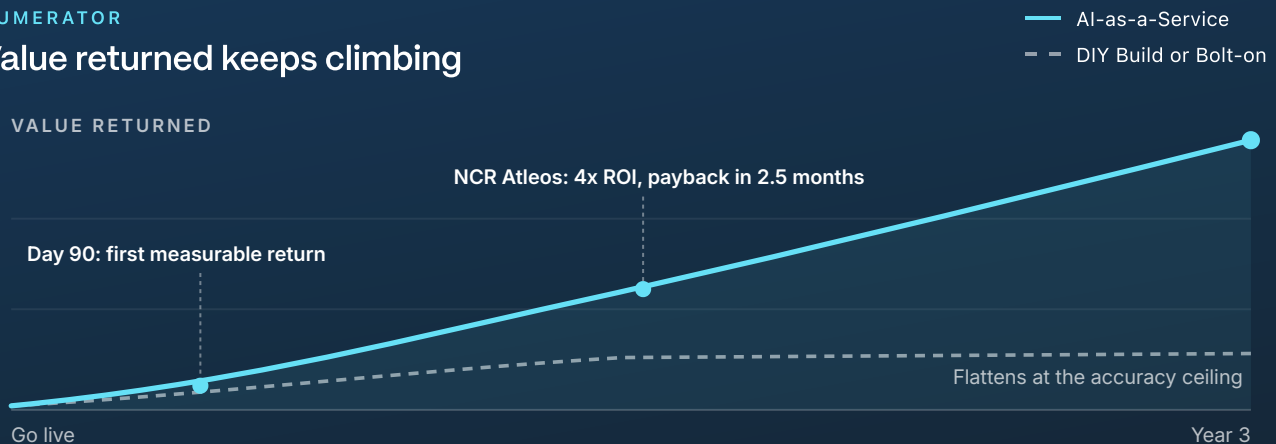
WHAT IS THE SERVICE DECISION GRAPH?

A living record of how your equipment fails and how your team fixes it. It connects each product to its known problems, each problem to the fix that worked, and each fix to what happened next. Neuron7 builds it from your manuals, case history and technician notes, and the AI uses it to recommend the right fix. Every resolved case makes it smarter.

NUMERATOR

Value returned keeps climbing

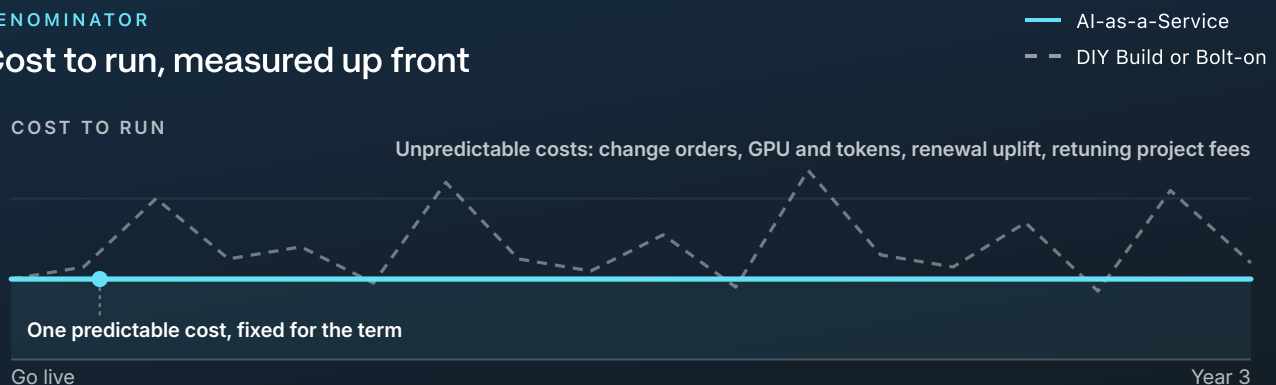
VALUE RETURNED



DENOMINATOR

Cost to run, measured up front

COST TO RUN



Curves are illustrative. Customer metrics are reported results attributed to the named customer, and individual results vary.

The next page explains why AI accuracy impacts ROI and service outcomes

60% accuracy is easy. Getting from 60% to 90% is where you actually scale.

Any competent team can demo or pilot simple retrieval on structured data. The problem is that expert decision making at scale doesn't live in there.

Why the first use-case is easy

An LLM over a vector database returns the closest matching document with reasonable relevance. That is a solved problem.

Your best-documented cases are the ones retrieval finds, and they are roughly the first 60%. The manuals cover them, the knowledge articles were written about them, and the case notes are consistent enough for retrieval to work.

A sprint-one demo pulls from that well-lit half of your data, so it looks like magic. Off-the-shelf frameworks get any competent team there in weeks.

Why real transformation is hard

Case notes were written for people, not for training. They are missing the micro-decisions that caused the resolution, and retrieval cannot surface reasoning nobody wrote down.

Manuals describe what the product is, not how it fails. No failure signatures, no branching for symptom variants, no pathway logic.

Retrieval returns the closest document. That is different from knowing which of 400 possible steps to try first, for this error code, on this product version, in this region.

The long tail is where tribal knowledge lives, and it lives in senior technicians rather than in your documents.

None of this is a model problem. It's six separate engineering programs.

THE COST OF BEING WRONG

A wrong answer in service is a second truck roll at \$800 to \$2,000, an SLA breach, or a warranty claim. The cost of being wrong is what makes 60% unusable in production, and it is why the accuracy line matters more than the launch date.

Build isn't one thing. It's six.

Each component below is its own engineering program with its own team and its own timeline. Five of them plateau. The sixth is the one almost nobody builds, and it is the one the others depend on.

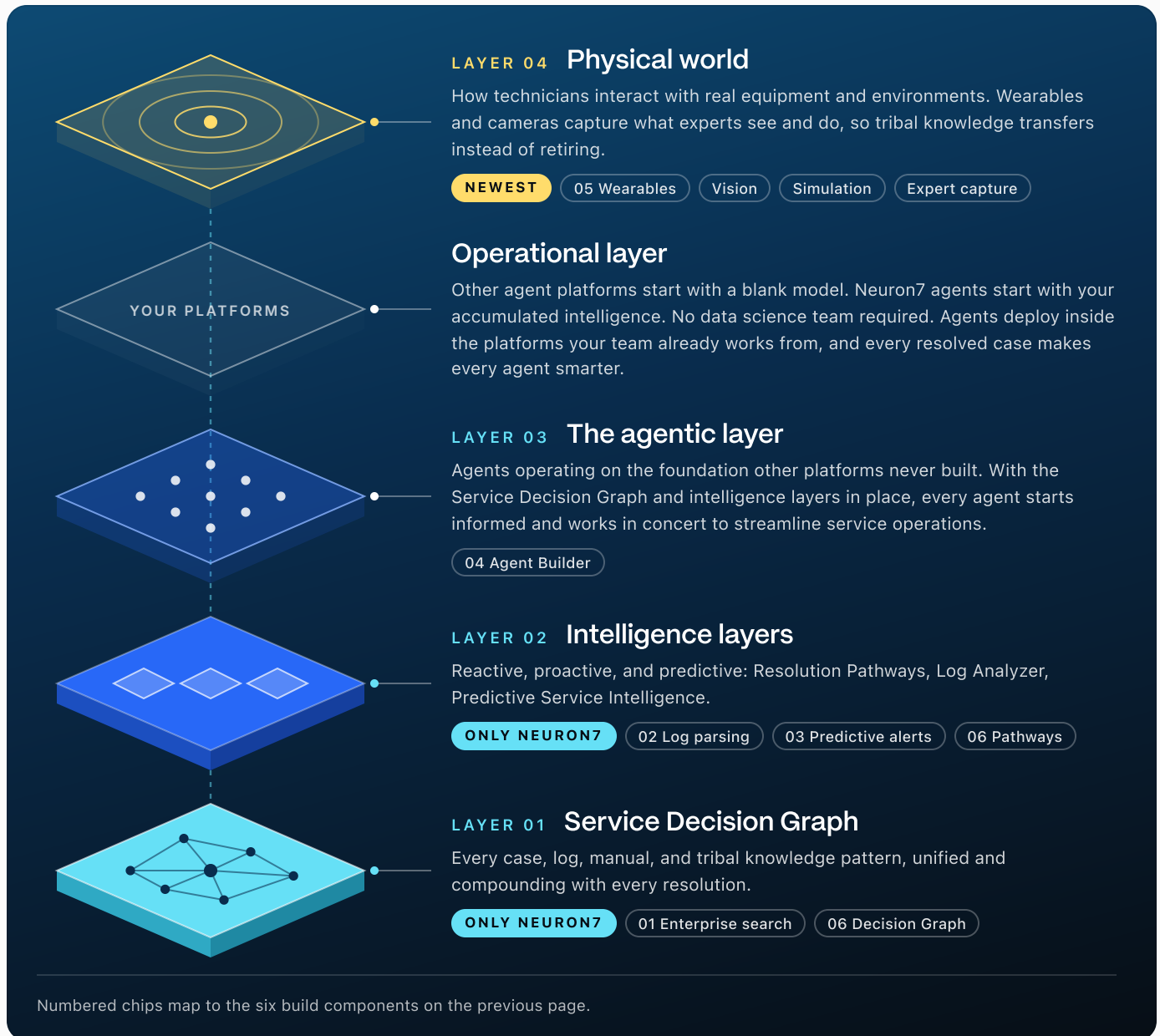
	COMPONENT	WHAT TEAMS BUILD	WHERE IT STALLS
01	Enterprise search over your knowledge base	An LLM plus a vector database over manuals, articles and case history.	Around 60% usability. Retrieval is not reasoning.
02	Log and IoT telemetry parsing	Splunk or Datadog-style pipelines, custom parsers, alerting rules.	Detection is not resolution. The technician learns that error E-127 fired, not what to do about it.
03	Predictive asset alerts	Telemetry ingestion with threshold or anomaly-detection models.	Alert noise. Doesn't work for air-gapped or disconnected assets.
04	Low-code Agent Builder	Single-step workflow agents with limited context.	Complex multi-step service processes.
05	Physical world and wearables NEW	Smart-glass apps, computer vision models and video capture of expert work.	Footage and sensor data that never connect to the resolution they show.
06	Deterministic pathways atop a Service Decision Graph	Deterministic rules engines, work instructions and SOPs	Lack of SME governance and service org-wide context

Neuron7 is all six, running together, one service decision layer.

Neuron7 combines service data, product knowledge, retrieval, workflow logic, and expert resolution knowledge in one service decision layer. Six years of product innovation sits underneath every answer your service team or customer sees, with 90%+ accuracy and measurable ROI you can take to the board.

What makes Neuron7 agents different

Your agents are only as good as the intelligence they have access to. Most simply retrieve answers from documents. Neuron7 agents reason from outcomes, built on every case, log, and resolution your team has ever closed.



FOUR LAYERS,
ONE GRAPH

Only Neuron7 delivers layers 1 and 2. That foundation is what makes layer 3 and on accurate enough to trust.

Build in-house, bolt-on, or AI-as-a-Service with Neuron7

	Building in-house	Bolt-on platform	Neuron7 AI-as-a-Service
PREDICTABLE COSTS	No	Partly	Yes
WHAT IT COSTS TO RUN			
AI and ML engineering	You hire it and staff it	You rent it or use off-the-shelf models	Included
Knowledge made machine-readable	12 to 18 months of internal work	You have to connect to their data clouds	Included, done by our team
GPU and compute	Your bill, moves with the market	Inside the platform price, opaque	Locked in
Integration to CRM, FSM and ERP	Your roadmap	Your IT team	Included
Tuning, monitoring and upgrades	Ongoing headcount, 10 to 20% a year	Limited control, on their schedule	Included, with proactive tuning
Upfront investment	\$500K to \$2M and up	Platform fees plus customization	Inside the subscription
WHAT IT RETURNS			
Time to production	12 to 18 months	4 to 8 weeks, limited scope	6 to 12 weeks, full capability
Accuracy on complex service	Reached by trial and error	60 to 80%	90%+, service specific
Domain expertise	Built from scratch	Generic, built for CRM	Purpose-built for complex service
Outcomes	40% of agentic projects abandoned	More vendor lock-in, simple automations	44% faster resolutions, 4x ROI in months

Gartner puts the abandonment rate for agentic AI projects near 40%.
Only Neuron7 moves both sides of the agent value equation, ensuring project success.

What's included?

From first data load to ongoing accuracy. This is what Neuron7 means by AI deployment as a service.

01

Production-ready software

Native integration with Salesforce, ServiceNow, Microsoft and SAP. A service-specific roadmap, knowledge drawn from your data and your people, and a feedback loop showing what works at scale.

02

White glove implementation

Data ingestion from cases, manuals and knowledge articles. Configuration and testing, CRM package install, and change management for the teams who will use it.

03

Support around the sun

24x7 coverage with fast response and resolution times, plus continuous monitoring. Upgrades carry no separate line item and no project of their own.

04

A cost line that holds

GPU cost locked in against market volatility. No per-user or per-transaction licensing. Infrastructure costs sit inside the price, with no reliance on your IT team to keep it running.

05

Enterprise-grade deployment

Secure single-tenant deployment in your environment, high-availability SLAs, proactive infrastructure monitoring, SOC 2 Type II compliant.

06

Accuracy that keeps improving

90%+ accuracy held as your data shifts. Orchestration and fine-tuning matched to each use case, custom dashboards, and a named success team that stays with the account.

KEY PARTNER INTEGRATIONS

Salesforce

Agentforce and Slack

ServiceNow

Now Assist

Microsoft

Dynamics and Teams

More platforms

ServiceMax, IFS, Oracle and others

The results service teams report with Neuron7

RETURN ON SCALE

4x ROI

NCR ATLEOS

Achieved a 4x return on investment, with payback in just 2.5 months across its field service organization in 48 countries.

MARGIN EXPANSION

44% faster

CIENA

Reduced resolution time by 44%, while increasing case deflection from 10% to 38% and CSAT by 14 points.

DOWNTIME REDUCTION

3 hrs → 3 sec

TRANSLOGIC

Reduced troubleshooting time for hospital tube systems from 3 hours to 3 seconds, with 96% resolution accuracy.

COST SAVINGS

24% lower

TERUMO BCT

Reduced average parts costs by 24%, while deflecting 20% of escalations in new markets.

These are entry points. The Service Decision Graph learns from every resolution your team completes, so accuracy rises quarter over quarter while the cost line stays where you signed it.

Run the numbers on your own service data with our team.