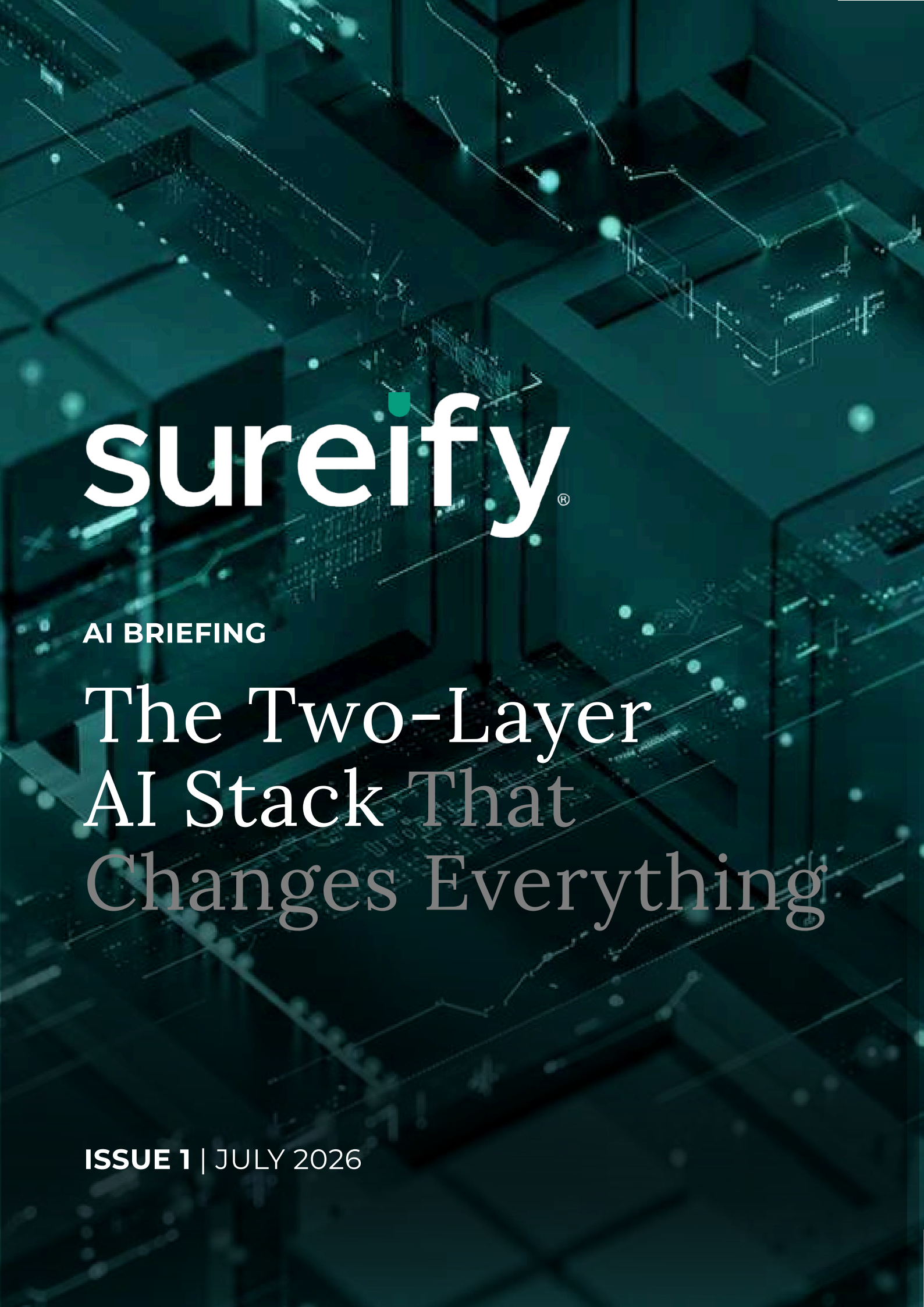


sureify®

AI BRIEFING

The Two-Layer AI Stack That Changes Everything

ISSUE 1 | JULY 2026

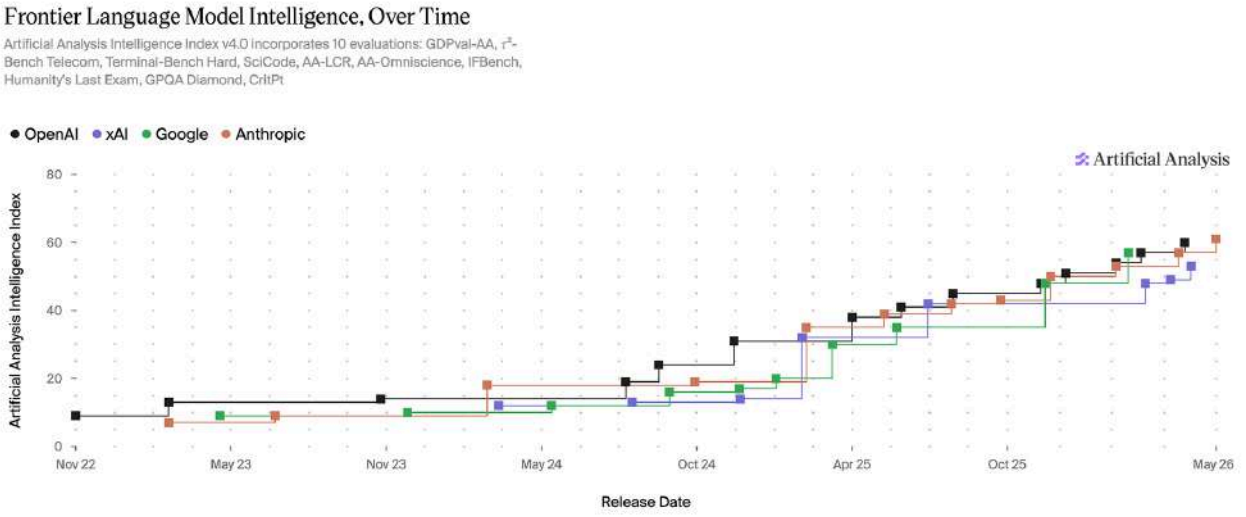
Index

The Late 2025 Revolution

On November 17, 2025, Grok shipped version 4.1. This was the first step in a progression of model breakthroughs over the next 25 days with the launch of Gemini 3, Claude Opus 4.5, and OpenAI's GPT-5.2. Shelly Palmer called December 2025 the month that "AI became ambient." Claude Opus 4.5 broke 80% on the SWE-bench, the first model to do that. GPT-5.2 beat or tied professionals on 70.9% of knowledge work tasks across 44 occupations. Gemini 3 shipped with a million-token context window.

Then the holiday break hit. Engineers, executives, and people who'd never written code all had time to play with this stuff. Claude Code went viral. People discovered they could build working software without knowing how to program. Then over the next few months, we saw the launch of Opus 4.6, GPT 5.4, and Gemini 3.1 Pro. Now the interesting thing is that all of the major models deliver frontier-quality reasoning, coding, and analysis. The differences will shift to other factors, such as pricing, ecosystem product

> "For the first time, people discovered they could build working software without knowing how to program."



Source: *artificialanalysis.ai "Frontier Language Model Intelligence, Over Time"*

depth, and style and personality. The agentic side is moving just as fast. In February 2026, OpenAI hired Peter Steinberger, creator of OpenClaw, the open-source agent framework that went viral over the holidays and generated 1.5 million agents in three months. Sam Altman called the work "core to our product offerings" and said, "The future is going to be extremely multi-agent." Anthropic shipped Claude Cowork with enterprise plugin marketplaces. Google pushed Gemini into Wear OS, Google TV, and Android Auto. AI isn't a tool you invoke anymore. It's embedded in every surface where work happens. Sureify's Brian Mulconrey put it this way in VentureBeat back in January: "That's evolving into continuous co-production: People and agents are shaping decisions, logic and actions together, in real time."

AI agent use at scale, by industry & function; % of respondents.

Source: mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai "Exhibit 3"

Timeline of AI Model Releases

The AI Acceleration Era

From late 2025 to early 2026, a wave of model releases rapidly advanced AI capabilities across reasoning, coding, and real-world performance.

2025		
Nov 17	**Grok 4.1** xAI	Achieved a ~65% reduction in factual errors.
Nov 18	**Gemini 3** Google	Million-token context window.
Nov 24	**Opus 4.5** Claude	Broke 80% on the SWE-bench.
Dec 11	**GPT-5.2** OpenAI	Beat or tied professionals on 70.9% of knowledge work tasks across 44 occupations.
2026		
Feb 5	**Opus 4.6** Claude	Million-token context window.
Feb 19	**Gemini 3.1** Google	Scored 77.1% on the ARC-AGI-2 benchmark.
Mar 5	**GPT-5.4** OpenAI	A 33% reduction in factual inaccuracies.

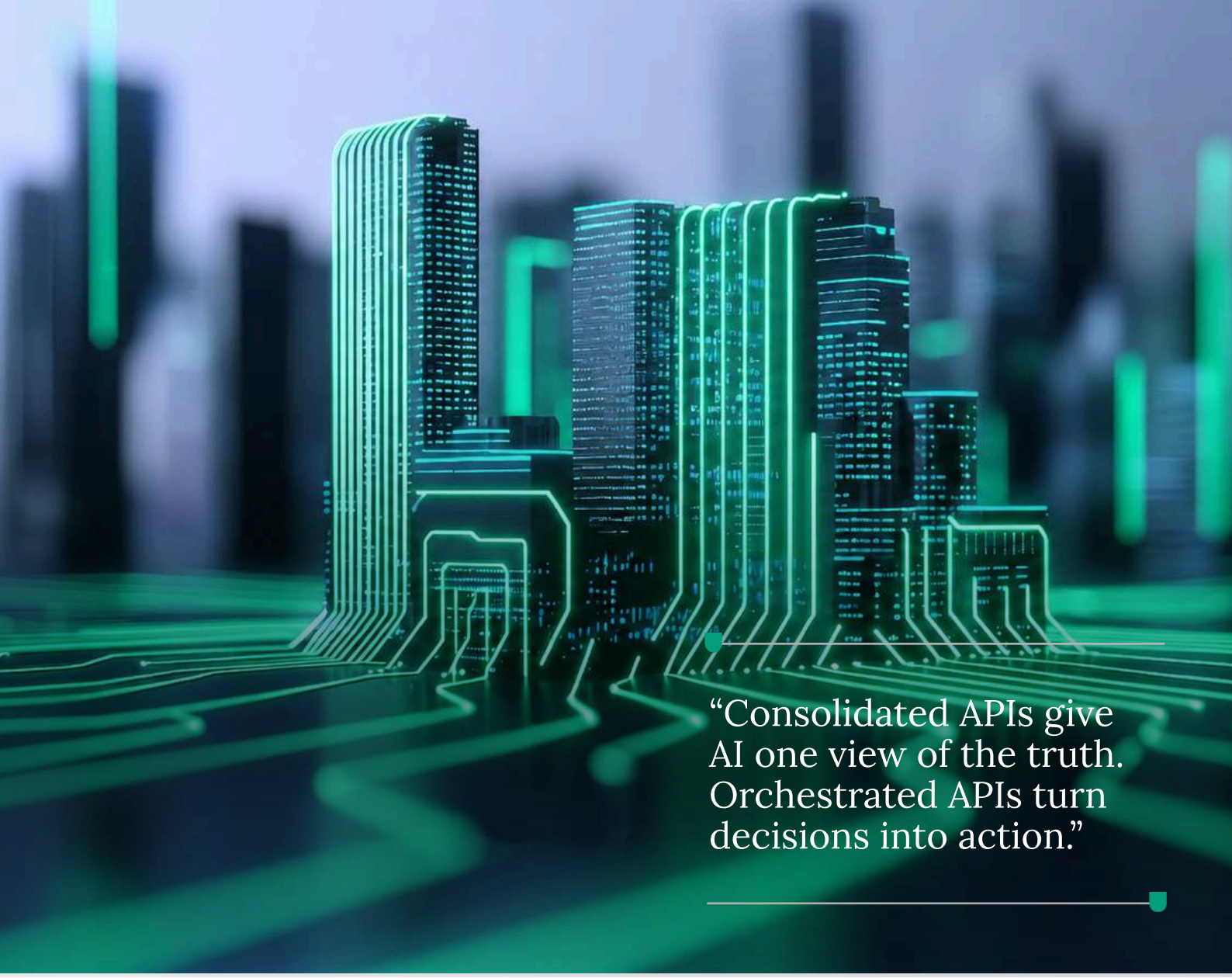
info@sureify.com | www.sureify.com

All APIs Are Not Equal

All life and annuity carriers use APIs from core system vendors, TPAs, and internally developed services. In fact, every modernization project over the past decade or so has produced waves of APIs. But not all APIs are equal. Single-system APIs expose one vendor's data in that vendor's format. That works for direct integrations between known systems. But for most carriers, your AI-powered digital experiences will need to read from and write to multiple back-end core systems. To do this, you need a different category of APIs:

Consolidated read APIs ingest from multiple back-end systems, normalize into a single canonical data model, and expose that unified view through a clean, authenticated interface. The model calls a single endpoint and returns a complete picture. It doesn't matter how many source systems contributed to that view.

Orchestrated write APIs accept a single action request from a front-end, whether AI-powered or conventional, and execute it as a durable, multi-system workflow. Orchestration supports guaranteed completion, automatic rollback on failure, and a full audit trail. A beneficiary change that touches the policy admin system, the document management system, and the compliance log occurs as a single atomic transaction.



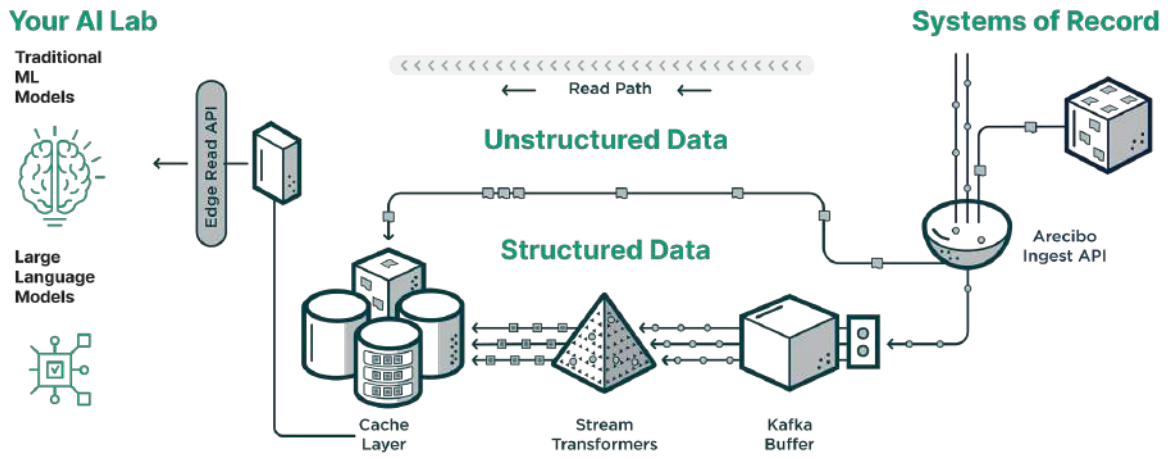
> “Consolidated APIs give AI one view of the truth. Orchestrated APIs turn decisions into action.”

The industry has invested tens of billions of dollars in recent years on modernization projects in search of these two seemingly simple capabilities. Most of those efforts have failed. Making your data ready for AI doesn’t have to involve replacing your legacy infrastructure; standing up the right API solutions can make all the difference.

CoreCONNECT™ Perspective

CoreCONNECT transforms data from your back-end core systems (e.g., policy admin systems, billing, compensation, producer management, document systems) into consolidated data models and exposes those models via APIs for use by both AI models and conventional user interfaces (Pillar 1). On the write side, CoreCONNECT Journeys, backed by Temporal, provides durable, multi-system transaction orchestration across potentially many write APIs (Pillar 2).

Pillar 1: Consolidated Read APIs



Twelve Systems, No Single View

A typical life and annuity carrier runs a dozen or more core systems. An AI model can't reliably reason about a carrier's operations if the data is scattered across 12 places and 12 formats. It needs a single API that works with a data transform architecture that's already done the integration work: Ingesting from every source, normalizing into a single model, and exposing it through a clean interface. The model calls that API the same way any other application would.

This is the data readiness problem. It's the primary reason enterprise AI projects stall. A recent EY survey of finance leaders found, "Eighty percent say insufficient AI-ready data is their most significant barrier to advancing AI within their organizations." And, last year, Sureify CEO Dustin Yoder, writing in Insurance Innovation Reporter, made a similar point: "The 'data plumbing' side of making your data AI-ready is not the most glamorous aspect of the work. Yet, it's the most significant factor in the success of your AI projects." To actually operationalize those AI pilot projects, you're going to need to make your organization ready for AI.

> "80% say insufficient AI-ready data is their most significant barrier to advancing AI within their organizations."
>
> EY

> “The future of AI isn’t just smarter models — it’s unified data exposed through clean, consumable APIs.”

Give AI Models One View of Your Data

BigDATAwire calls 2026 the year of the “context engine” or purpose-built infrastructure for managing what AI can see. MIT Sloan’s Davenport and Bean warned that AI agents are entering Gartner’s trough of disillusionment largely due to their data not being ready for AI.

Across industries, the enterprise platforms generating the highest returns from AI share a common architecture: Consolidate data from multiple source systems into a unified model, then expose it through clean APIs that any AI tool can consume. Salesforce recognized this with Data.Cloud, which unified customer data across CRM, commerce, and service into a single profile that Agentforce queries through APIs. ServiceNow embedded AI search across its platform by layering retrieval over a normalized data model using the same pattern: consolidate first, then let AI consume the consolidated view.

CoreCONNECT™ Perspective

CoreCONNECT follows this pattern for life and annuity carriers. Ingest data from your core systems and then normalize that data into an insurance data model. We then expose those consolidated views through OpenAPI-compliant Edge Read APIs. Those APIs can be called by AI models, conventional applications, and MCP-compatible agents.

MCP: The Protocol for AI-to-API Connections

Also in December of 2025, the Model Context Protocol was donated to the Linux Foundation. MCP is the open standard for how AI models and tools connect to external APIs. Anthropic made MCP open-source in late 2024, with support from OpenAI, Block, Google, Microsoft, AWS, Cloudflare, and Bloomberg Monthly SDK downloads now exceed 97 million.

Gartner predicts 40% of enterprise applications will feature task-specific AI agents by the end of 2026.

MCP defines how AI models discover and call APIs. When a carrier deploys any MCP-compatible tool, it looks for MCP endpoints. If a consolidated data API exists behind an MCP server, every AI tool on the market can find and use it.

> "40% of enterprise applications will feature task-specific AI agents by the end of 2026."
>
> Gartner

Beyond Structured APIs: The Knowledge Layer

Consolidated APIs handle structured data well: Policy records, rate tables, product metadata, and in-force queries. But carriers also hold massive volumes of unstructured domain knowledge, such as underwriting guidelines, regulatory requirements, servicing procedures, compensation structures, call recordings, and distribution rules.

This data doesn't fit into standard API responses, so the semantic knowledge layer becomes important for indexing content so the model can search by meaning, not exact text. In practice, this means converting carrier documents into a knowledge base that an AI model can query as naturally as querying a database via an API.

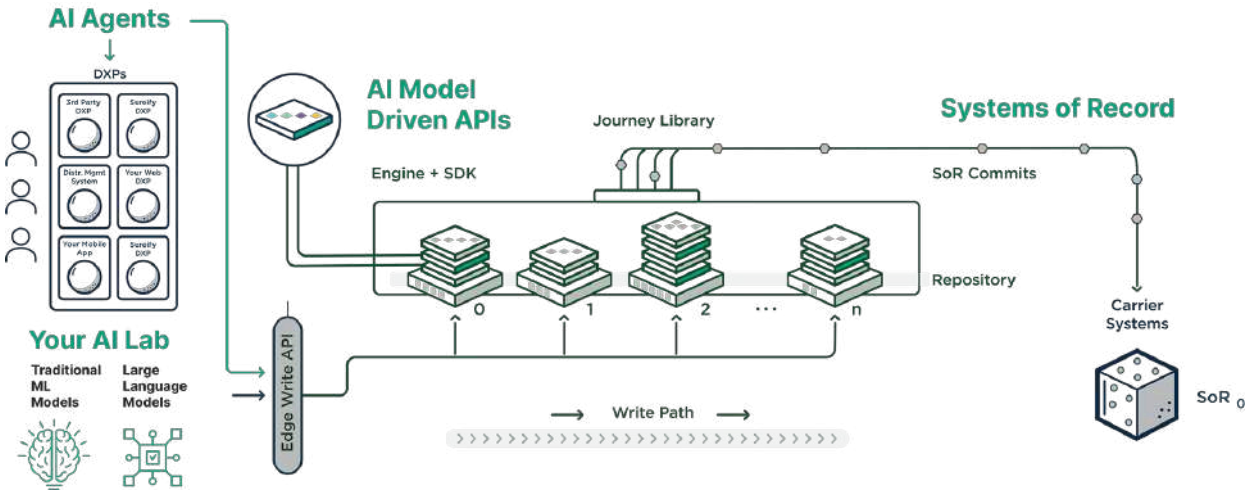
For carriers, this means product definitions, rider specifications, surrender schedules, and regulatory requirements can be made searchable by any AI model, even when the query and the source document use entirely different terminology.

Make carrier knowledge as accessible as an API.

CoreCONNECT™ Perspective

CoreCONNECT positions carriers to expose MCP servers that consolidate carrier data to any AI tool. MCP servers work naturally in front of CoreCONNECT's existing Edge Read APIs and authentication framework, making policy records, product definitions, and in-force book queries discoverable by any MCP-compatible models or agents. That can include existing models such as ChatGPT, Gemini, Claude, and Grok, or something that doesn't exist yet.

Pillar 2: Orchestrated Write APIs



Most deployed AI systems today can only answer questions. When a representative asks an AI assistant to update a beneficiary, and the model correctly determines what needs to happen, it must call a write API that triggers a reliable workflow to update the policy administration system, store documents in the document management system, and update a compliance audit log.

If one system fails, the transaction rolls back. If a regulator asks what happened six months later, the audit trail is there.

This isn't just an AI challenge. The same problem exists today when a policyholder submits a change through a portal, mobile app, or call center. Any front-end interaction that modifies data across multiple back-end systems needs durable orchestration. The late-2025 model releases made it more urgent because models can now initiate actions faster, at higher volume, with greater autonomy.

Why Execution Is the Bottleneck

MIT Sloan’s Davenport and Bean, January 2026: AI agents “make too many mistakes for businesses to rely on them for any process involving big money.” An AI might correctly determine that a policy needs a rider added.

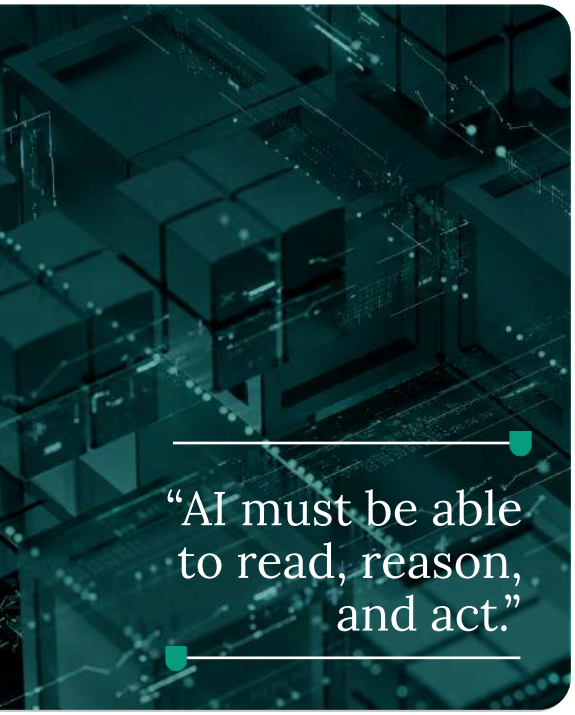
But if the API call to the admin system times out and the confirmation letter still goes out, you have a compliance problem.

Google Cloud’s 2026 AI Agent Trends report frames the enterprise challenge as building “digital assembly lines” rather than one-off prompts. Assembly lines require reliable execution at every step.

Deloitte found that only 20% of organizations are growing revenue through AI. The other 80% are stuck between a working demo and a system they’d actually trust in production.

CoreCONNECT™ Perspective

CoreCONNECT Journeys, supported by the Temporal workflow engine, provides a robust architecture for transactions that guarantee completion, automatically roll back on failure, and provide a full audit trail across back-end systems that includes both conventional transaction execution and AI model interactions. By using the same APIs to serve both AI-initiated actions and conventional transactions, carriers get unified support across portals and call centers. You don't need separate infrastructure for AI and human workflows.



MCP Works for Write Operations Too

MCP isn't just for reading data. An MCP server can expose executable tools alongside discoverable data. A single MCP endpoint lets an AI agent (GPT, Gemini, Claude, or, other open-source models) both look up a policy and initiate a servicing transaction via the same standardized protocol.

The APIs that consolidate data from your core systems are only half of the AI data readiness challenge. You also need support for writing to your back-end system with execution guarantees, rollback capabilities, and audit trails that regulators require.

CoreCONNECT™ Perspective

CoreCONNECT Journeys can be exposed as MCP tools. Any MCP-compatible AI agent can initiate a multi-system workflow. We also recognize that there may be many MCP server interactions within any given journey, such as a new representative onboarding journey that interacts with a large distributor MCP server to assign a contracting hierarchy. Our Temporal workflow engine ensures that every workflow completes correctly and in compliance.

> “In February 2026, we saw AI adoption reach 47.6% of U.S. businesses.”
>
> ramp

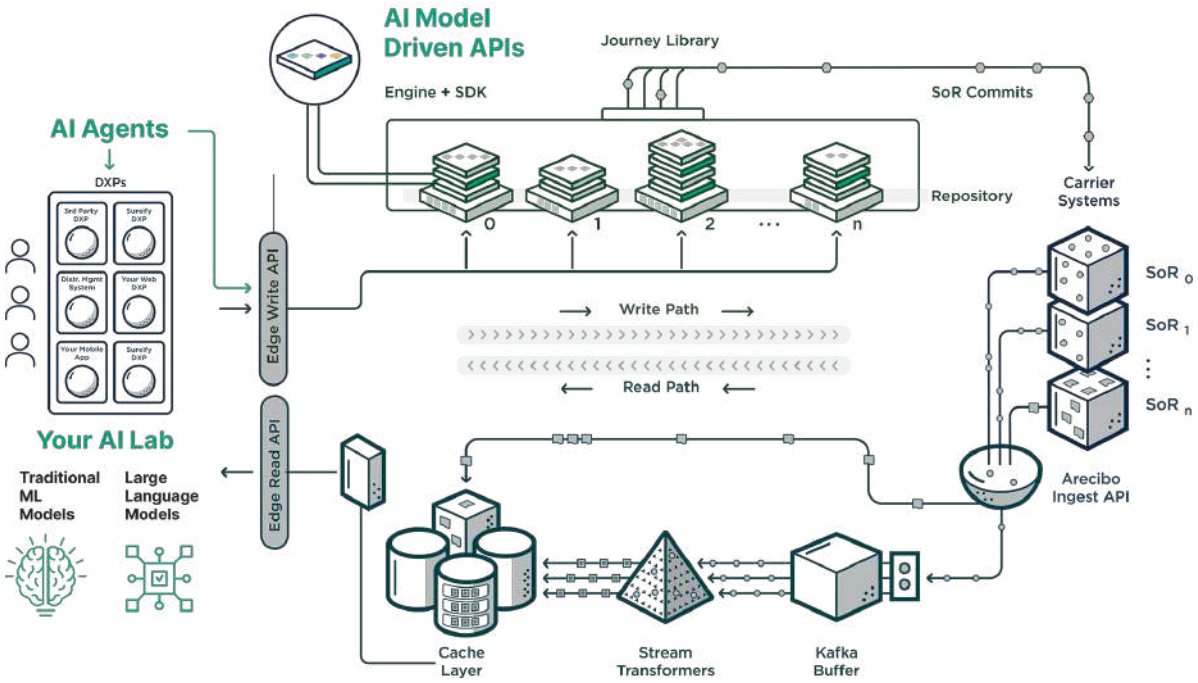
From Pilots to Procurement

Ramp’s AI Index tracks corporate spending across more than 50,000 businesses. In February 2026, we saw AI adoption reach 47.6% of U.S. businesses. In 2023, contract values for AI averaged $39,000; by 2026, that figure is $1 million.

For carriers, this means product definitions, rider specifications, surrender schedules, and regulatory requirements can be made searchable by any AI model, even when the query and the source document use entirely different terminology.

As these tools reach carriers, they’re exposing the gaps. A frontier model can reason about a policy question, but can’t update the policy administration system. An AI agent can surface a customer inquiry, but can’t execute a multi-system servicing transaction. The write API, backed by durable orchestration, turns these tools from question-answers into infrastructure that agentic AI solutions can rely on.

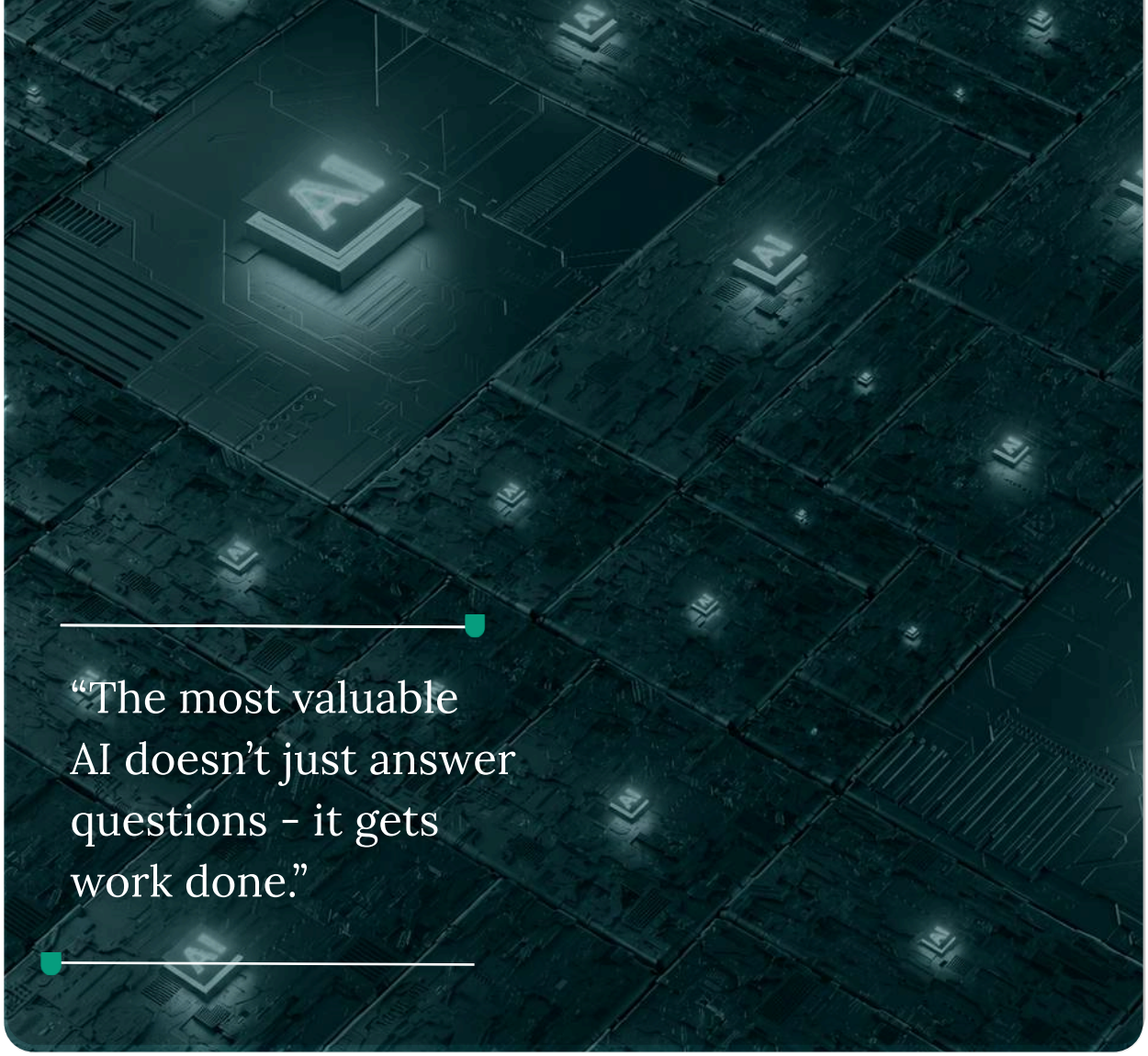
How the Two Pillars Work Together



Consider what happens when an insurance representative asks an AI assistant: "Which of my clients should I call about the new FlexGrowth Plus MYGA?"

"Context powers AI. Action creates value."

Pillar 1. First, the model calls a consolidated read API and returns the representative's book of business, each client's existing products and maturity dates, the FlexGrowth Plus specifications, state-specific suitability requirements, and replacement rules.



Pillar 2. The model then calls a write API that triggers a workflow that updates records in the CRM, initiates a suitability review in the compliance system, and logs the appropriate actions in the audit trail. If one of those systems is unavailable, it automatically retries, and if a step fails permanently, the transaction is rolled back.

With consolidated read APIs (Pillar 1), you can avoid replacing back-end infrastructure and invest in real sales and service improvement initiatives. And by empowering AI models with API tools that orchestrate complex back-end updates and agentic behaviors (Pillar 2), you are making it possible to drive complexity out of your organization.

The late 2025 model releases accelerated the timeline. Carriers are fielding requests right now from representatives, distributors, and internal teams who’ve already used GPT, Gemini, and Claude in their personal workflows. They want the same capabilities against carrier data.

sureify®

Ready to move AI beyond the pilot phase?

Learn more about how **Sureify** is providing the foundation to enable AI in life and annuity, transforming legacy constraints into a scalable, agentic future.

sureify.com/ai

info@sureify.com | www.sureify.com