

The AI Stack: How Vendors Are Composing AI Strategy

An Eight-Layer Reference Model and Six Emerging Archetypes for Vendors and Decision-Makers With a Stable Framework to Interpret Vendor Commitments in a Market Still in Motion

Analyst: Mitch Ashley

Publication Date: July 17, 2026

Document #: AIOMA202606

Key Points

- The AI-native software engineering market has stratified into a durable eight-layer stack, and six emerging archetypes describe how vendors are composing it.
- Buyers are committing now. A total of 54% of organizations say AI touches more than half of their software lifecycle work, and 58% expect AI to build 80% or more of their software within three years¹.
- Archetype fit is the right first-pass unit of analysis for 18- to 24-month vendor bets, while layer completeness determines whether those bets hold up in execution.

Recommendations

Vendors and enterprise technology leaders should adopt a shared reference model for the AI stack as they make their next round of strategic commitments.

- Use the eight-layer stack as the analytical unit. Read every vendor announcement against the layer it touches and the layers it leaves untouched.
- Treat the six archetypes as a current snapshot. Use them as the working shape for the next two to three quarters.
- Track the vendor's primary control point. The control point reveals where stickiness, roadmap gravity, and competitive collisions form, and where lock-in risk emerges over time.

¹ Futurum Group 2H 2026 SLE Decision-Maker Survey, N=839

What You Need to Know

Across Q1 and Q2 2026, vendors at every layer of the AI stack further revealed their strategies and made commitments that define its current shape. Microsoft adjusted its relationship with OpenAI while preserving a layered, decoupled platform posture. Google repositioned Vertex AI as part of a broader Agent Platform strategy. AWS extended Bedrock through Managed Agents and AgentCore. IBM advanced Granite plus Red Hat OpenShift. Salesforce introduced the Agent API as a SaaS-resident control plane. Anthropic expanded multi-cloud parity. This report introduces 'The AI Stack' as the reference model that Futurum uses to read those positions and those still to come.

Analysis

The AI Stack

The AI Stack establishes a durable structure for analyzing how vendors compose and evolve their AI strategies. The model treats AI strategy as a stack-layer composition problem rather than a product announcement or portfolio problem. Reading by product reduces every release to a feature comparison. Reading by layer exposes where a vendor is staking their position, where it is partnering, and where it leaves gaps competitors will fill.

Eight layers run from the consumption surfaces developers and users touch directly down to the silicon executing tensor operations. Each layer is a unit of competitive analysis and a potential point of stickiness risk. The [Agent Control Plane Framework](#), Futurum's agent governance reference model, sits at the Governance and Runtime layers and shows how those layers operate as a unified control plane in practice. The layers are durable across vendor cycles. Vendor positions within them are not.

The eight layers are as follows:

Apps: The consumption surfaces are where AI reaches end users, other agents, and integrating systems. UI applications, agents acting as applications, and programmatic interfaces through open standards such as MCP and A2A all coexist as candidate work surfaces. This is where AI value is monetized, and trust is tested.

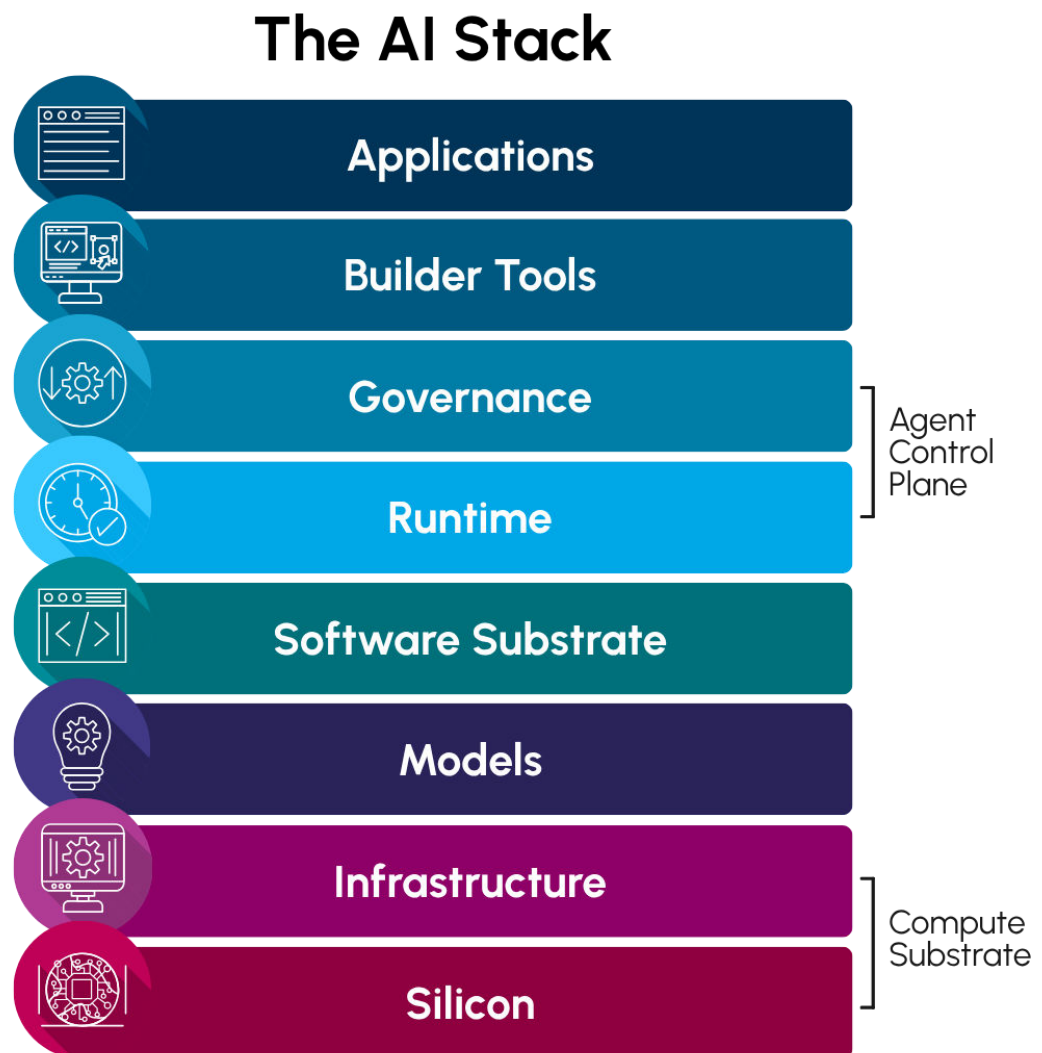
Builder Tools: The developer-facing surfaces where engineers author agent logic, write code, and define workflows: AI-native IDEs, agent development kits, evaluation frameworks, and code generation assistants. Builder tools determine developer productivity and the stickiness mechanics that follow adoption.

Governance: Identity, registry, policy, and audit for agents acting in production: what agents may do, who they act for, what evidence they generate, and how their actions are reviewed. A

total of 75% of organizations had a production incident in the last 12 months with AI as a contributing factor, yet only 43% mandate human review of AI-generated code, and agent governance ranks as the least-mature SLE practice¹. Without this layer, agent deployment is structurally ungovernable.

Runtime: Execution of agent logic and management of in-flight state – isolation, working context, resource and cost controls, lifecycle, and inter-agent communication. Runtime is where governance policy becomes execution reality, and runtime portability determines whether workloads can move across substrates.

Figure 1: The AI Stack



© 2026 Futurum. All rights reserved.

Source: Futurum Research, July 2026

Software Substrate: The operating, orchestration, data, and agent memory environments where agent execution lands, including the persistence tier purpose-built for agents. That tier is increasingly where consequential customer IP accumulates, raising the stakes of substrate choice. Open versus proprietary substrate is a defining commitment for any archetype.

Models: Commercial frontier, specialized, open-source, and open-weight models across cloud, on-premises, and hybrid environments. Buyers have already resolved the single-versus-multi-model axis: among enterprise builders, 56% run multiple foundation model providers in production, and just 15% rely on a single provider². Multi-model has won.

Infrastructure: Data centers across on-premises, edge, cloud, and neo-cloud environments deliver inference and compute capacity. Scarcity at this layer is reshaping vendor strategy across every layer above it.

Silicon: The GPUs, TPUs, and accelerators that execute tensor operations. Supply, generation, and architecture propagate up every layer as cost, latency, and capability constraints, and a concentrated vendor position here creates structural dependencies for everyone above it.

The Six Archetypes: How Vendors Are Currently Composing the AI Stack

The archetypes describe how vendors compose these eight layers through strategies and commercial commitments. They are emergent, not settled. Vendor moves will continue into 2H 2026, solidifying throughout 2027.

The archetypes expose where each vendor seeks to exert control, what customers implicitly commit to, and where competitive collisions will occur. They do not prove completeness or execution maturity; those require layer-by-layer analysis. Vendor placements below are descriptive, not prescriptive.

Layered Decoupled Platform: The vendor owns the integration plane while keeping the substrate, runtime, builder surface, and models separable for customers to compose. Stickiness forms through orchestration, identity, and ecosystem gravity, and customer-accumulated IP compounds that friction over time. Lock-in risk emerges as the integration plane becomes load-bearing. Microsoft illustrates it today: Microsoft IQ for context delegation, Agent 365 for cross-substrate identity and governance, Foundry as the hosted agent platform, and GitHub Copilot as the developer control surface.

Vertical Integration with Unified Control Plane: The vendor engineers the stack as a unified system from silicon through the agent platform, each layer optimized for the presence of the others. The bet is that integration depth and unified governance outweigh demand for

² ETR AI Product Series, March 2026, N=514

layer-by-layer substitutability. Google illustrates it today through TPU silicon, Gemini models, ADK, and the unified Agent Platform.

Harness-Coupled Brand Multiplication: The vendor couples multiple branded product surfaces to one shared agent harness, keeping the durable control point in the harness while brands multiply above it. The proliferation is deliberate, and the tradeoff is customer ambiguity about which brand becomes the strategic center. AWS illustrates it today through Bedrock, AgentCore, Managed Agents, Q, and Strands as branded expressions of one harness.

Hybrid Open Stack: Portability is the product – an open-source substrate, models, automation, and operator tooling so agent workloads run across hyperscaler clouds, on-premises, and sovereign environments. Differentiation comes from operational consistency and enterprise control. IBM illustrates it today through Granite, Red Hat OpenShift, Ansible, and HashiCorp.

SaaS-Anchored Agent Platform: The enterprise application becomes the agent control plane, with agents grounded in SaaS-resident data, workflows, permissions, and business objects. The advantage is proximity to business context; the constraint is dependence on the platform's data model. Salesforce illustrates it today through the Agent API, and ServiceNow through AI Agent Studio.

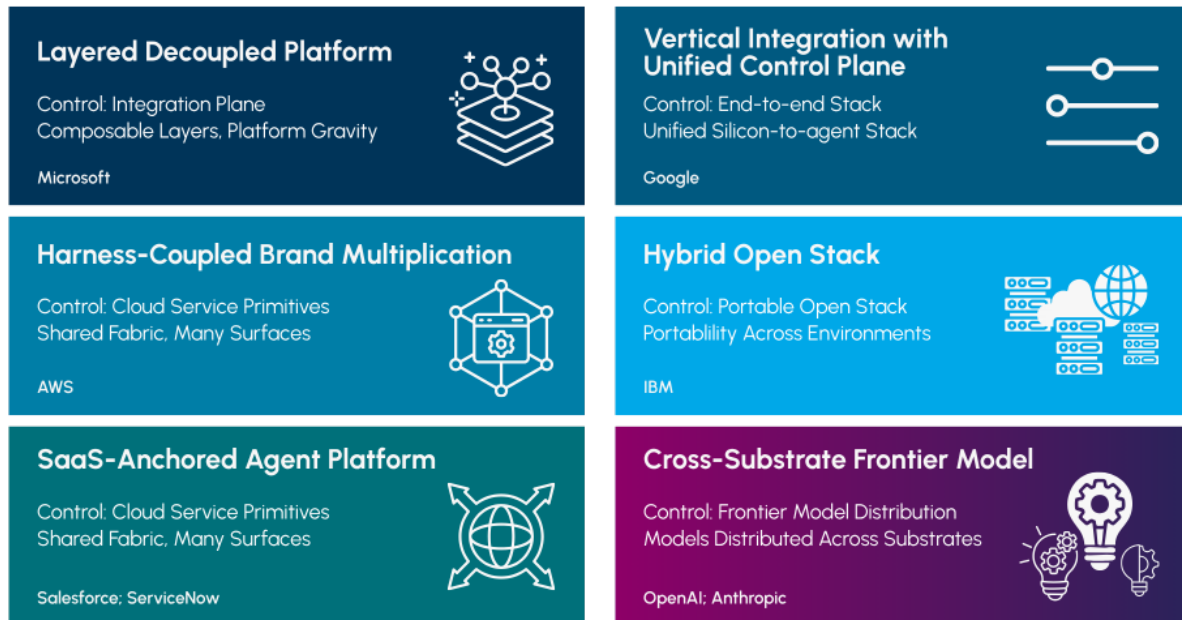
Cross-Substrate Frontier Model: The model is the primary product, and cloud substrates function as distribution channels rather than owned control points. OpenAI and Anthropic illustrate two variants: OpenAI through deep Microsoft alignment, Anthropic through multi-cloud parity across Bedrock, Vertex, and Foundry. OpenAI holds the broadest model footprint: 57% of enterprise builders use its GPT and o-series models². Anthropic leads the coding workflow: Claude Code is rolled out at 54% of organizations, nearly triple OpenAI Codex's 20%, with a 63% value-over-cost rating against Codex's 41%³. One is winning the model surface, the other the coding workflow.

³ ETR AI Product Series, January 2026

Figure 2: Six AI Stack Archetypes – How Vendors Reveal and Compose Their Strategies

Six AI Stack Archetypes

How Vendors Reveal & Compose Their Strategies



*NVIDIA is the dominant silicon dependency between the archetype set, not a peer archetype.

© 2026 Futurum. All rights reserved.

Futurum

Source: Futurum Research, July 2026

NVIDIA operates as the dominant shared silicon dependency beneath the archetype set rather than as an archetype of its own. Most AI stack strategies still depend on NVIDIA's accelerator, networking, and software ecosystem, even when vendors position their own silicon alternatives.

Presence Is Not Completeness

A vendor appearing in a layer says nothing about how fully it plays there. Archetype placement maps position, not strength, depth, or maturity.

Independent buyer data turns that distinction into a measurement. Microsoft 365 Copilot, the most widely deployed AI product ETR tracks, sits in 71% of organizations, yet only 36% of users say its value clears its cost⁴. GitHub Copilot reaches 59% of organizations, while only about half of its users say it returns more than it costs³. Presence and completeness diverge inside the same product.

⁴ETR AI Product Series, May 2026, N=516

The Lens of the CIO Technology Buyer

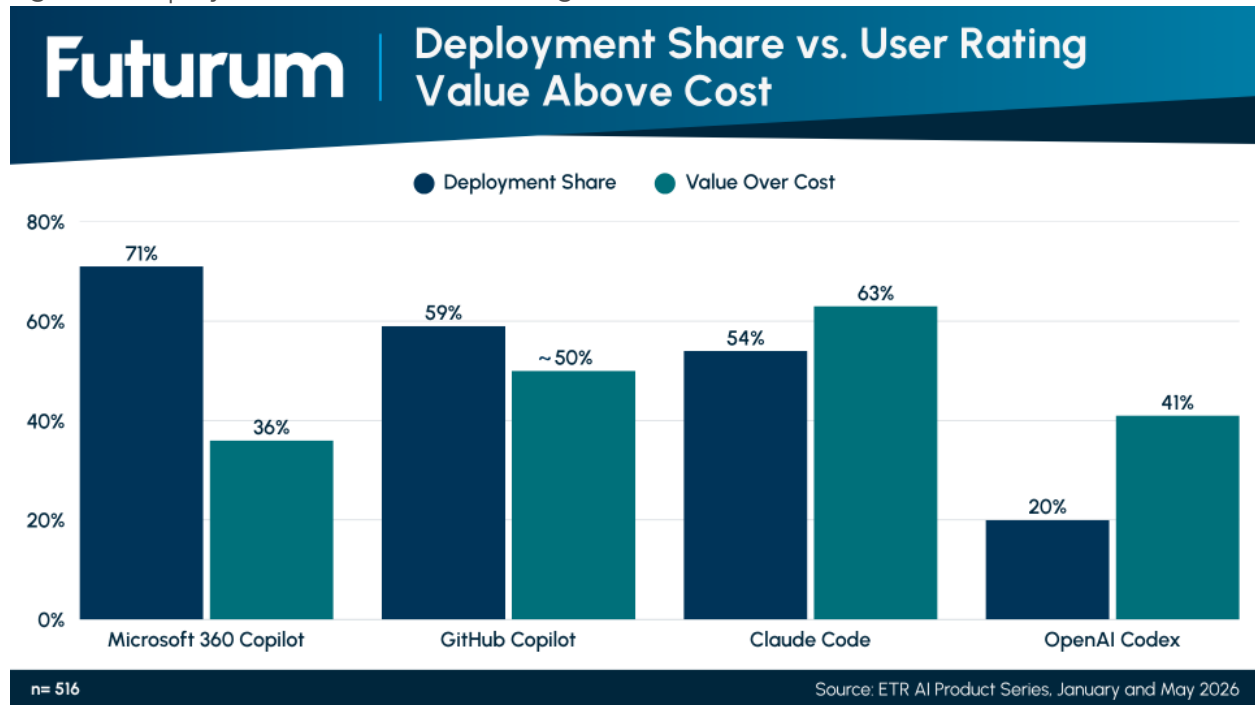
Every CIO evaluating an AI platform is choosing an archetype, whether the evaluation names it or not. The purchase looks like a product selection; the commitment underneath is structural, determining which layers the organization inherits, which the vendor owns, and which stickiness mechanics tighten over time. Buyers are already funding that commitment, ranking AI investment (31%) above engineering headcount (19%) as the top spending priority¹.

CIOs should test archetype alignment on four fronts:

- Whether the vendor's control point fits the organization's existing investments and operating model.
- Whether the layers of the organization depend more on production maturity than on positioning.
- Which substitution boundaries does the organization need to preserve as the stack evolves?
- Whether the archetype's roadmap direction reinforces or undermines the organization's own.

The CIO's accountability concentrates on the agent control plane, regardless of which archetype wins the evaluation. Incidents, audit evidence, and runaway spend all surface at Governance and Runtime, the two layers where buyer practice is weakest today. A CIO who tests vendors on control plane completeness first, and capability second, inverts the evaluation that most vendors are built to win. That inversion is where the buyer's leverage lives.

Figure 3: Deployment Share vs. User Rating Value Above Cost



Source: ETR AI Product Series, January and May 2026

Why The AI Stack Now

Buyers are not waiting for the stack to settle. Among organizations, 54% say AI now touches more than half of their software lifecycle work, and 40% report AI generated or substantially modified the majority of production code merged in the last 90 days¹.

Spending follows the adoption: total SLE spend is expected to grow 14% in 2027, with 41% of organizations planning to significantly increase Develop-stage investment, and a 58% majority expect AI to build 80% or more of their software within three years¹.

Independent buyer telemetry shows the same trajectory. At the models layer, the share of enterprise builders using Anthropic's Claude more than doubled in a year, from 21% to 48%, while OpenAI's 40-point lead narrowed to nine^{2,5}.

Buyers are funding the stack now before its shape settles throughout late 2026 and into 2027. The AI Stack will be revised on a rolling basis, with this AIR as the foundational reference.

What to Watch

- The archetype set will evolve. Expect new variants, splits within existing archetypes, and possible mergers as vendors revise positioning, with splits likely within two quarters.

⁵ ETR AI Product Series, March 2025, N=472

- The work surface category is in active evolution. Agents compose tools, model interfaces, protocol layers, and skills into new surfaces such as the terminal, chat, and browser. Archetypes that own the integration plane underneath will ride out the evolution; those committed to a single composition are more exposed.
- Stickiness is observable in 2026. Lock-in is not. The real test, what happens when a buyer of meaningful size tries to leave, has not run for any archetype. The first significant migration attempts through 2027 will show which archetypes remain genuinely substitutable.
- Vendor events will be archetype validation moments: Build, re:Invent, Dreamforce, Think, and major model releases will reveal whether each vendor is deepening commitment, drifting, or shifting.
- A new archetype may emerge for IDE-as-control-plane. AnySphere (Cursor), Replit, and Windsurf are pushing the builder tool layer toward control-plane ownership and may justify a seventh archetype within six months. NVIDIA's neutrality will also be tested: expansion beyond silicon into runtime, governance, or substrate would force a revisit of its status.

Futurum Recommends

This AIR introduces 'The AI Stack' and the six emerging archetypes. The model is the analytical foundation for Futurum's coverage of AI-native software engineering through 2026, including the Agent Control Plane Framework, the Seven Principles of Observability-Native, and the Agentic AI Open Standards Report. Vendors and technology leaders can engage with the Futurum Software Lifecycle Engineering practice for vendor stack assessment, archetype fit analysis, stickiness and lock-in risk evaluation, and strategic commitment review.

Other Insights from Futurum

[Software Lifecycle Engineering Market to Reach \\$226 Billion by 2030 as Enterprise Hands AI the Keyboard Before the Guardrails](#)

[Google I/O: Did Google Just Ship the Full AI Stack?](#)

[Microsoft Build 2026 – The Platform, Integration Plane, and Developer Surface](#)

[Red Hat Brings Developers, Product, and Operations to the Center of Agentic AI](#)

[AWS Pushes the Agent Stack: Quick, Connect Verticals, OpenAI on Amazon Bedrock](#)

About Us

About the Authors

Mitch Ashley is VP & Practice Lead of CIO & Technology Buyers and Software Lifecycle Engineering for The Futurum Group. A multi-time CIO and CTO with 30+ years leading technical organizations, Mitch built and operated production systems spanning cybersecurity for the U.S. Department of Defense, PKI services for the broadband and 5G industries, SaaS platforms, large-scale telecom and banking systems, and a national broadband network. His work with AI began early, developing expert systems that diagnosed and repaired complex mainframe environments. That operator foundation grounds his analysis in operational consequence, covering the technology buyer's world of software engineering, cybersecurity, DevOps, cloud, and AI. He can be reached at mashley@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization, including partners, resellers, external consultants, etc., in any media format

Posting on a website that is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.

SEO Metadata

Excerpt: Mitch Ashley, VP and Practice Lead at Futurum, introduces 'The AI Stack', an eight-layer reference model and six vendor archetypes for reading AI strategy commitments from Microsoft, Google, AWS, IBM, Salesforce, OpenAI, and Anthropic.

Meta Description: The AI stack has stratified into eight durable layers and six vendor archetypes, which reveal how vendors compose AI strategy and where control forms.

Keyphrase Focus: AI stack

Tags: AI Stack, AI stack archetypes, agent control plane, agent governance

Vendors: Microsoft, GitHub, Google, AWS, IBM, Red Hat, OpenAI, Anthropic, NVIDIA