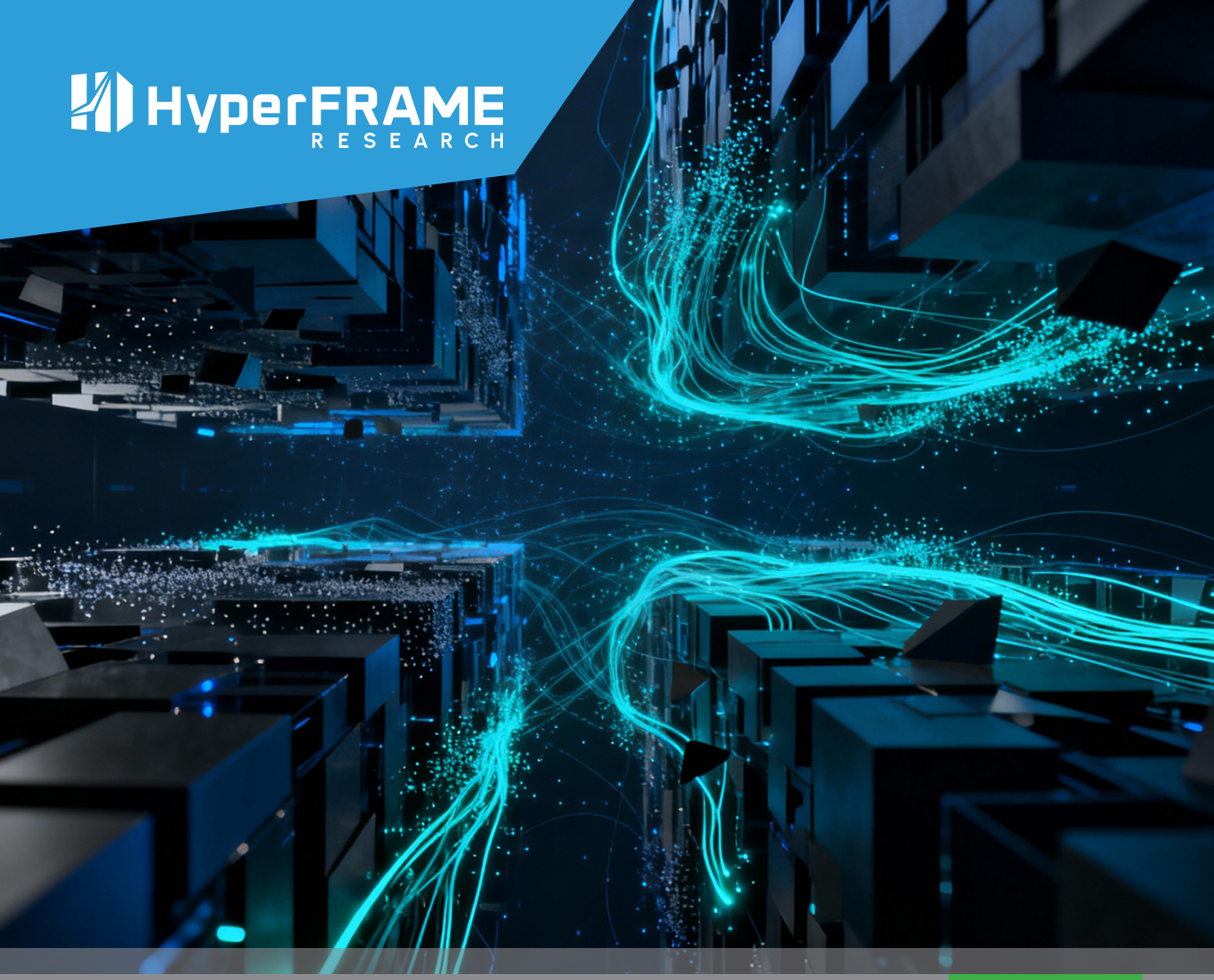
HyperFRAME RESEARCH



RESEARCH BRIEF

The AI Stack Grows Up

A 2025 Retrospective on Enterprise AI, Infrastructure, and the Rise of Governed Autonomy

Authors:

Stephanie Walter
Practice Leader, AI Stack

DECEMBER 2025

Executive Summary

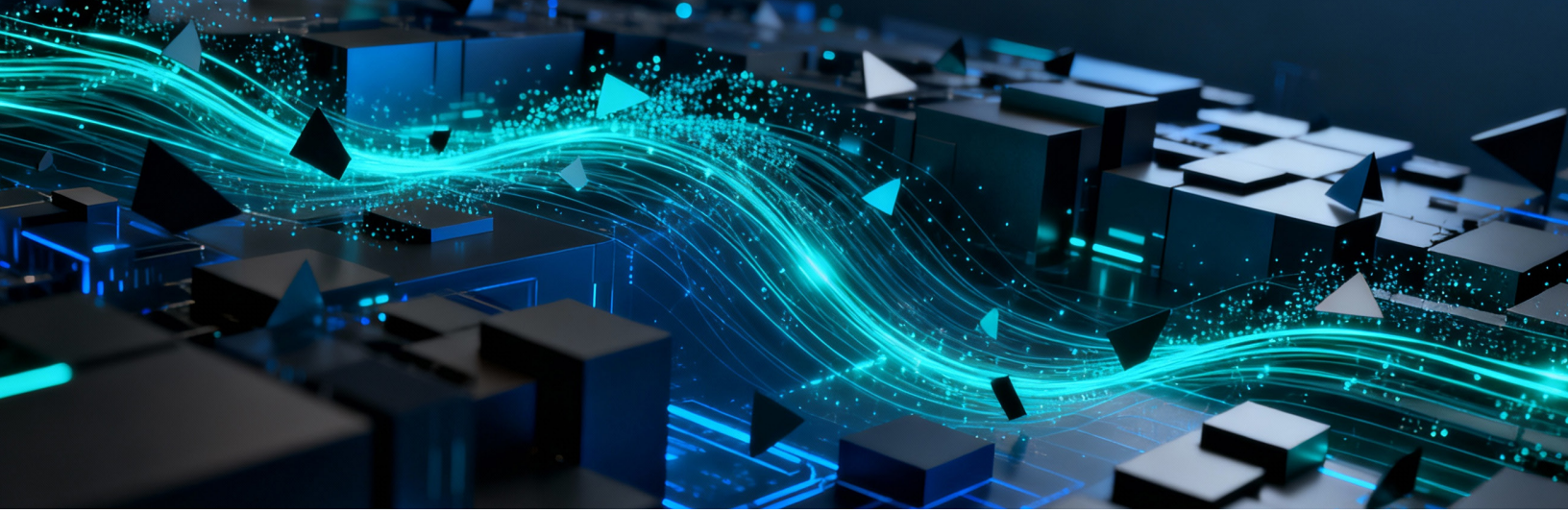
This paper synthesizes our 2025 HyperFRAME Research to examine how the AI Stack, or the software that enables enterprises to use and create AI applications, advanced from model experimentation to the systems required to operate it at scale. We began 2025 with an industry still obsessed by the raw generative capabilities of foundational models. However, as the months progressed, our analysis tracked a fundamental pivot in priorities. The conversation moved from the cognitive potential of the model to the operational integrity of the system. We have observed that 2025 was the year enterprise AI moved from a state of promise to a state of extreme pressure. Across many large enterprises, executive conversations shifted from 'what's possible' to 'what's operational, governed, and measurable,' with more scrutiny on timelines and ROI.

The core observation across this body of work is that AI progress has shifted from models to systems, platforms, and governance. The AI software stack has emerged as the primary battleground for vendor differentiation. We are no longer in an era where a slightly better benchmark score on a reasoning test can sustain a competitive advantage. Instead, the market is rewarding providers who offer deep integration and architectural stability. The transition of agents has gone from experimental playthings to operational workhorses. This shift was not driven by breakthroughs in logic alone but by the development of the scaffolding required to manage them.

Data gravity and infrastructure have reasserted their control over the technology roadmap. The physical and logical location of data now dictates the feasibility of AI applications. Furthermore, governance has emerged not as a bureaucratic constraint but as a prerequisite for deployment. Without the ability to govern autonomous actions, enterprises simply refuse to move past the pilot phase. As we look toward 2026, the research contained in this synthesis signals a move toward execution discipline. The following analysis explores how the stack grew up and what that means for the next phase of the digital enterprise.

Key Highlights: What 2025 Taught Us

- AI maturity exposed organizational, not technical, bottlenecks.
- Infrastructure, data, and governance converged into a single stack reality.
- Agentic AI advanced only when control mechanisms followed capability.
- Vendor success correlated with stack depth and integration, not model performance alone.
- Enterprises rewarded realism over ambition and demanded proof, not promises.



The Unifying Thread Across 2025: The AI Stack as the System of Record

The most consistent theme across the 2025 research is the elevation of the AI Stack toward a system-of-record-like role for AI operations, governance, and integration. In previous years, AI was treated as an experimental add-on or a peripheral analytics tool. Today, it serves as the central coordination point for enterprise logic. We started the year documenting the chaos of the data center, where infrastructure identity was fragmented and workloads were poorly understood. By the end of the year, our research reflected a much more structured reality.

The AI Stack addresses the three most significant failures of the early generative era. First, it solves a coordination problem. Enterprises realized that having fifty different point solutions for AI was creating more friction than value. A unified stack allows these disparate tools to share context and data. Second, it addresses a governance problem. When AI is integrated into a central platform, policy can be applied uniformly rather than managed on a per tool basis. Third, it solves an economic problem. Fragmented AI usage is expensive and difficult to optimize. A central stack provides the visibility needed to manage costs effectively.

We have found that the limits of point solutions were exposed much faster than many anticipated. Organizations that tried to build their AI strategy on a collection of best-of-breed individual tools found themselves trapped in an integration nightmare. The AI Stack is now the engine room of the company. It is where decisions are made, data is processed, and autonomy is governed. Success in this environment is less about the intelligence of the model and far more about the coherence of the integration.

Infrastructure Reasserts Itself: AI Is a Systems Problem Again

For a brief period, there was a belief that AI would live entirely in a world of high-level abstractions where the underlying hardware did not matter. Our coverage this year has proven the opposite. Infrastructure has come roaring back as a primary strategic concern. The complexity of the modern data center and the chaos of AI workloads have made systems engineering fashionable again.

One of the most significant trends we observed was that Kubernetes increasingly served as a common orchestration layer for AI platforms and pipelines, especially where enterprises needed portability and governance. We also saw a noticeable resurgence in on-prem and hybrid AI strategies, driven by cost, data gravity, and control requirements. Companies like Dell have found a second wind by offering hybrid strategies that give enterprises the control they crave. Many organizations realized that moving all their proprietary data to the public cloud for the sake of AI was either too expensive or too risky.

The tension between cloud elasticity and enterprise control was a major narrative thread in 2025. Hyperscalers like AWS and Google continue to offer unmatched scale, but platforms like Oracle and its Exadata systems have gained ground by promising a tighter loop between the database and the compute. Where AI runs has become a strategic decision once again. It is no longer just about picking the cheapest cloud provider. It's about identity, workload orchestration, and the silent enablers of performance like high-speed networking and memory bandwidth.

Enterprises are finding that they cannot simply throw more GPUs at a problem if their networking fabric is outdated. As a team, we have written about how the return of hardware consciousness is shaping vendor selections. AI infrastructure is no longer invisible plumbing. It is a competitive differentiator. The firms that own their infrastructure stack or have a clear hybrid roadmap can move significantly faster than those waiting for someone else to solve their capacity issues.

Data Platforms and Databases are AI's Center of Gravity

If infrastructure is the body of the AI Stack, the data platform is its heart. Our research in mid-2025 was dominated by the evolution of the database. We witnessed a historical convergence of analytics, transactions, and artificial intelligence into a single layer. This convergence is intended to reduce latency and improve contextual access for AI workloads.

Databricks' pursuit of OLTP ambitions and its strategic acquisitions signaled a move toward being the complete data foundation for the AI era. We also tracked the significant partnership between SAP and Databricks, which aimed to bridge the gap between enterprise resource planning and advanced data science. Similarly, Oracle spent the year aggressively positioning its database as the ultimate AI execution environment. They understood that AI is most effective when it is close to the data it needs to process.

Databases are no longer just storage bins for information. They are evolving into active participants in the AI workflow. Vector search, which was once a niche feature, is now a standard requirement. But the real innovation was the ability for databases to handle the complex, multimodal data that agents require to function. MongoDB and DataStax also played crucial roles here, providing AI-native data services that allow for the rapid retrieval of context.

The platforms that dominated 2025 are those that help enterprises redefine what AI-ready data actually looks like. It is no longer just about cleaning rows and columns, but creating a dynamic environment where data can be retrieved, summarized, and acted upon in milliseconds. The center of gravity in the enterprise has shifted away from the model and toward the platform that feeds it.

Agentic AI Moves From Concept to Governed Capability

The most visible evolution of the year occurred in the field of agentic AI. We began 2025 with agents that were largely experimental and often unreliable. By the end of the year, our research showed a shift toward agents that are designed to deliver real work within governed boundaries. This was one of the most consequential shifts of the year: moving from chat interfaces toward tool-using systems designed to act.

We spent a large portion of the year tracking the rise of agent platforms like ServiceNow and IBM agentic automation. These tools are architected to act as digital labor rather than just digital assistants. They can interact with legacy systems, perform multi-step tasks, and make low-level decisions.

However, these agents only became enterprise credible when governance entered the conversation. Google Vertex AI Agent Builder and other similar tools prioritized the creation of control planes. They recognized that an agent is only as good as the guardrails that surround it.

Governance, memory, and tool control became the three pillars of agentic success. Companies like Oracle and Elastic emphasized MCP-style tool exposure and registries as governance primitives; other ecosystem players advanced similar patterns for tool control and auditability. We moved from a world of black box agents to a world of transparent, audited digital labor. This is a crucial distinction. In the early part of the year, the fear of hallucinations kept agents out of production. By late 2025, the focus was on policy violations.

Enterprises began viewing AI agents as a way to augment their workforce rather than just a tool for individual productivity. This shift in perception was earned through the hard work of building memory layers and context engineering. Agents are no longer just responding to prompts. They are executing workflows based on a deep understanding of corporate policy and historical data. They have become the interface for the governed enterprise.

Governance, Identity, and Trust: The Hidden Throughline

While agents and infrastructure grabbed the headlines, the most important work of 2025 happened in the realm of governance, identity, and trust. These are not blockers but the actual enablers of scale. Throughout the year, we tracked the emergence of trust infrastructure as a quietly decisive factor in who won and who lost in the AI race.

The convergence of identity and AI orchestration became a primary focus of HyperFRAME research. If an agent makes a mistake, the system must be able to trace that action back to a specific policy and a specific identity. We saw the rise of infrastructure identity and API governance as the silent throughline that connected every successful deployment. Security is no longer a separate department that says “no” to AI. It has become the department that provides the “how” for AI.

Organizations with mature governance frameworks can deploy AI faster than those without them. This is because they have already solved the hard questions of data residency, model bias, and tool authority. Governance is moving from a compliance tax to a scaling requirement. You cannot run ten thousand agents if you have to manually check every action they take. You need an automated system of trust.

The organizations that scaled AI in 2025 did so because they recognized that innovation without control is just a liability. They invested in tool registries and context control early. They treated governance as a core feature of their AI Stack rather than an afterthought.

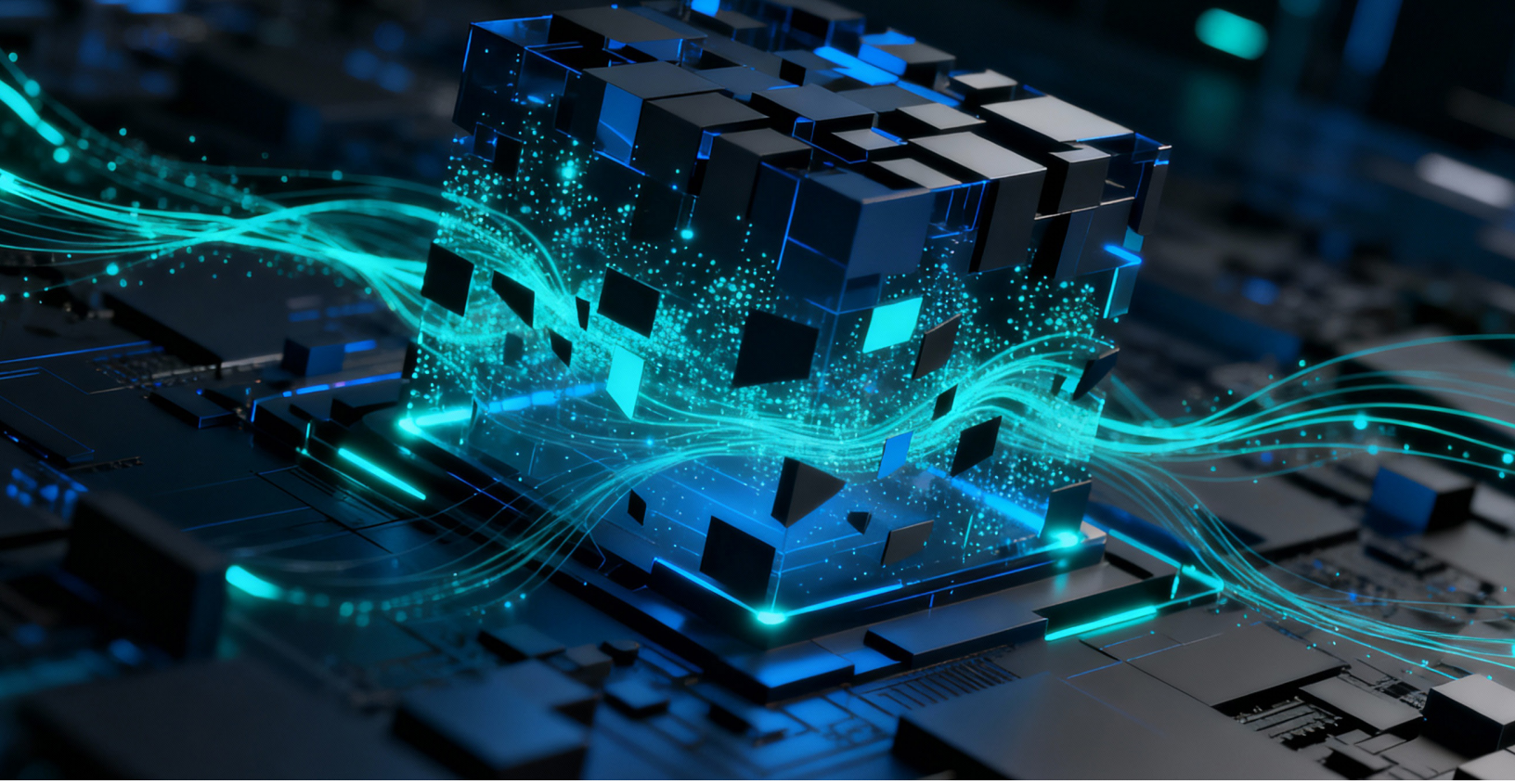
Platforms, Not Features, Won In 2025

The competitive landscape of 2025 was a lesson in the power of platforms. We saw a clear shift away from feature based competition toward platform coherence. The major players like AWS, Google, Oracle, and IBM positioned themselves not just as model providers, but as stack providers.

The rivalry between TorchTPU and CUDA continued, but the real story was in the partnerships that redefined competition. We saw Grok running on Oracle and IBM collaborating with Oracle to deliver hybrid solutions. These moves were architected to deliver stability to an enterprise market that was tired of choosing between competing silos. We observed that the erosion of single vendor dominance did not lead to a displacement of the giants. Instead, it led to a more complex, interconnected ecosystem where everyone has to play well together.

Developer experience emerged as a massive factor in vendor selection. The cost of switching between AI platforms is now much higher because the stacks are so deep. If you have built your data pipeline, your agent governance, and your workload orchestration on one platform, moving to another is a monumental task. This has created a new kind of sticky relationship between vendors and enterprises.

AI maturity has finally exposed the fact that most bottlenecks are organizational rather than technical. We have learned that infrastructure, data, and governance have converged into a single unified stack that cannot be managed in pieces. Agentic AI only works when control mechanisms are as sophisticated as the models themselves. We have seen that vendor success is now correlated with the depth of the integration they offer rather than the raw performance of a single model. Ultimately, the enterprise has rewarded realism over ambition this year.



Looking Forward: What This Body of Work Signals for 2026

As we move into 2026, the primary theme will be the transition from experimentation to optimization. Our research suggests that the coming year will be defined by a focus on ROI accountability. The massive investments made in 2024 and 2025 will be expected to yield measurable results. We are moving from a period of platform build-out to a period of operational excellence.

We predict that 2026 will see a reduction in the number of AI announcements and an increase in the number of operational benchmarks. The industry will move from agent enablement to agent management. We will see the rise of selective abstraction, where companies choose to own the parts of the stack that provide competitive advantage and outsource the rest. Governance tooling will become a default requirement for any purchase.

There will be a much sharper focus on the economics of AI. CFOs are now asking for granular cost models for inference and training. The era of the blank check for AI is over. Our 2025 coverage indicates that the organizations that will win in 2026 are those that have built a coherent, governed, and integrated AI Stack.

Execution discipline is the new gold standard. AI is not magic. It is infrastructure. It is data. It is governance. And it is finally growing up. The strategic advantage in 2026 will belong to those who can manage the complexity of the AI Stack and turn the promise of 2025 into the performance of 2026.

ABOUT HYPERFRAME RESEARCH:

HyperFRAME Research delivers in-depth research and insights across the global technology landscape, spanning everything from hyperscale public cloud to the mainframe and everything in between. We offer strategic advisory services, custom research reports, tailored consulting engagements, digital events, go to market planning, message testing, and lead generation programs.

Our industry analysts specialize in rigorous qualitative and quantitative assessments of technology solutions, business challenges, market forces, and end user demands across industry sectors. HyperFRAME Research collaborates closely with your Analyst Relations, Product, and Marketing teams to build and amplify your thought leadership, positioning your expertise to enhance brand and product recognition. Through content that engages readers, viewers, and listeners alike, we ensure your voice resonates across channels.

CONTACT HYPERFRAME RESEARCH:

Steven Dickens
CEO & Principal Analyst | HyperFRAME Research

Email Address:

steven.dickens@hyperframeresearch.com

Telephone Number:

+1 845 505 1678

X: @StevenDickens3

LinkedIn: Steven Dickens

BlueSky: Steven Dickens

CONTRIBUTORS

Stephanie Walter
Practice Leader, AI Stack

INQUIRIES

Contact us if you would like to discuss this report and HyperFRAME Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts, but must be cited in-context, displaying author's name, author's title, and "HyperFRAME Research." Non-press and non-analysts must receive prior written permission by HyperFRAME Research for any citations.

LICENSING

DISCLOSURES

HyperFRAME Research provides research, analysis, advising, and consulting to many high-tech companies, including those mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.


HyperFRAME
RESEARCH