# The Intelligent Infrastructure: Navigating the Agentic AI Era with Compute Orchestration

An Executive Guide to Optimizing AI Workloads, Controlling Costs, and Accelerating Innovation.

Clarifai ebooks | 2025

# Index

# 1. Executive Summary

The era of Artificial Intelligence is no longer a distant future; it is here, fundamentally reshaping industries and driving unprecedented opportunities for innovation and competitive advantage. Yet, the true impact of AI hinges not just on groundbreaking models and agentic AI workflows, but on the efficiency, scalability, and intelligence of its underlying infrastructure. As enterprises worldwide embark on AI initiatives and automation agents of increasing complexity and scale, they are confronting a stark reality: without a robust and optimized compute foundation, the promise of AI can quickly turn into a burden of spiraling costs and operational bottlenecks.

Consider the ambitions of tech giants like OpenAI and Microsoft, who officially announced The Stargate Project in January—a new initiative to address the growing need for compute in AI. This collaboration underscores a critical truth: advanced AI, especially the next generation of autonomous agents, demands a significant scaling of compute infrastructure, requiring unprecedented investments in both new AI hardware and software development, and the clean energy to power these data centers. **Traditional compute management approaches are simply not equipped to handle the unpredictable, high-volume demands of modern AI.**

> This is where **Compute Orchestration (CO) emerges as the strategic imperative**. This discipline is poised to unlock $2.6 trillion to $4.4 trillion in annual value for the global economy, transforming enterprises into 'agentic' powerhouses where autonomous systems drive unprecedented efficiency and innovation. Compute Orchestration is the intelligent, automated approach to managing and optimizing AI workloads across diverse and distributed compute resources. It transforms a chaotic landscape of disparate GPUs and CPUs into a unified, autoscaling, efficient, and cost-effective fabric ready for the demands of the AI era.

At Clarifai, we have spent over a decade at the forefront of AI operations, pioneering solutions that empower enterprises to create, control, and optimize their AI workloads on any compute. Our unique position stems from our comprehensive platform, which is purpose-built to be the unified AI infrastructure for any model, any compute, any tool, and crucially, any agent development kit. This guide will illuminate the challenges and present a clear path forward, demonstrating how Clarifai's Compute Orchestration can unlock the full potential of your AI investments, control costs, and accelerate innovation while ensuring mission-critical reliability.

# 2. The New AI Reality

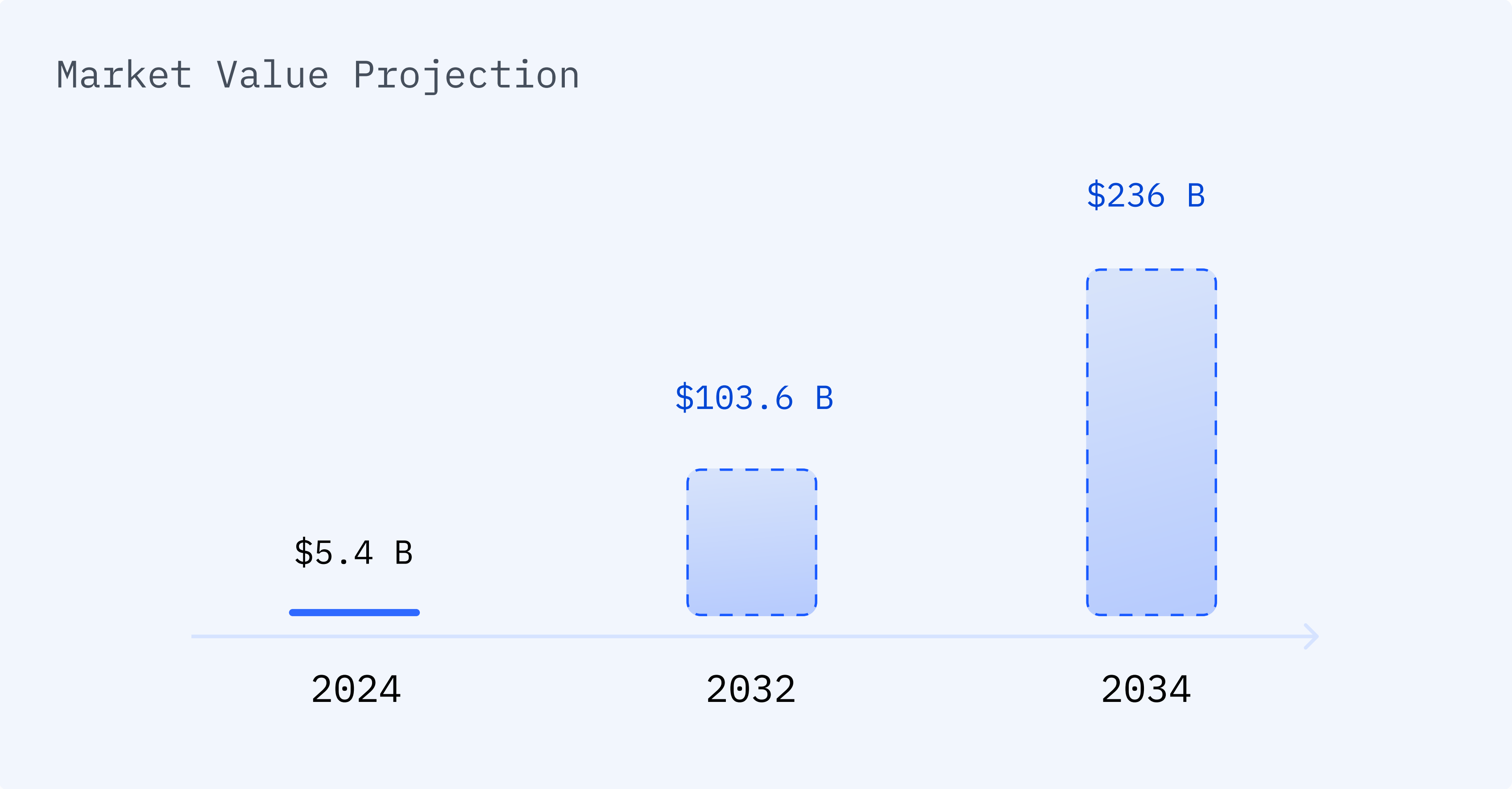
## Beyond Generative, Towards Agentic

The AI landscape is evolving at an unprecedented pace. To truly harness its power, executives must understand this progression, particularly the pivotal shift from traditional AI and even early Generative AI to the emerging paradigm of Agentic AI.

## The Evolving AI Landscape

For years, AI applications involved static models, like those for computer vision (CV) or natural language processing (NLP), with predictable training and inference loads. The advent of Generative AI, especially large language models (LLMs), introduced inference-heavy models consuming significant compute for real-time content generation—the first major wave of infrastructure strain. Now, we stand at the precipice of the Agentic AI era. This is a fundamental transformation: Agentic AI refers to autonomous systems that can understand complex goals, plan multi-step actions, interact with various tools, and adapt dynamically. Unlike earlier models that perform a single function, agents can reason through tasks, utilize multiple models and tools, and make decisions in pursuit towards the user's end goal.

**This isn't mere hype; it's a strategic reality.**

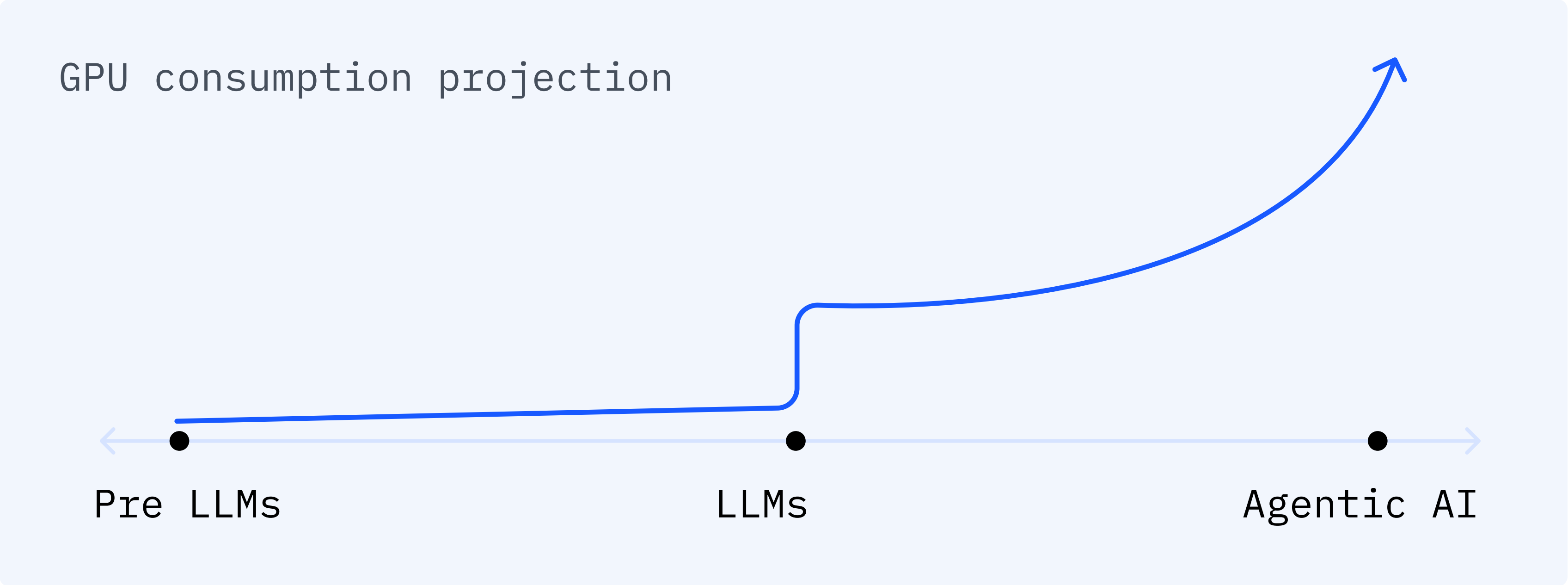


The market data unequivocally confirms this. After being valued at approximately $3.7 billion to $5.4 billion in 2023-2024, the AI agent market is projected to undergo a meteoric rise, with forecasts ranging from $103.6 billion by 2032 to an astonishing $236 billion by 2034. This dramatic expansion is driven by a consistent Compound

Annual Growth Rate (CAGR) of 40% to 46%, signaling a technology rapidly moving to the core of enterprise strategy. As of early 2025, a commanding 88% of organizations are either exploring or actively piloting AI agents, with a pioneering 12% already deploying them at scale. This widespread adoption, observed across both tech and non-tech companies, underscores the universal applicability and accelerating integration of this technology.

## Strategic Imperatives for the Agentic AI Era

The rise of Agentic AI has profound implications for your compute infrastructure, creating two critical strategic imperatives that demand immediate attention:



1. **Addressing the Consumption-Heavy Nature of Advanced AI**
   Unlike traditional AI models that often complete an inference request in a single pass, LLMs and Agentic AI are far more compute-intensive. Generating a single sentence or paragraph requires iterative processing, where the model performs multiple passes for what appears to be one continuous output. This fundamental difference leads to a substantial jump in GPU cycles and energy consumption per logical request, directly translating into higher operational costs, as well as unpredictable and varying compute usage. Efficient control and management of these resources are absolutely essential for financial sustainability.
2. **Managing Unpredictable and Dynamic Workloads**
   The evolution to Agentic AI dramatically increases workload unpredictability. While traditional models have stable resource demands, Agentic AI's ability to dynamically call upon multiple tools, models, and agents (e.g., an LLM for reasoning, then a vision model, then a data retrieval tool) creates highly erratic and unpredictable spikes in compute utilization. This dynamic behavior necessitates a truly elastic and dynamic infrastructure that can effortlessly scale up to meet high traffic and scale down during low periods. This dual capability is crucial for balancing peak performance with stringent cost efficiency.

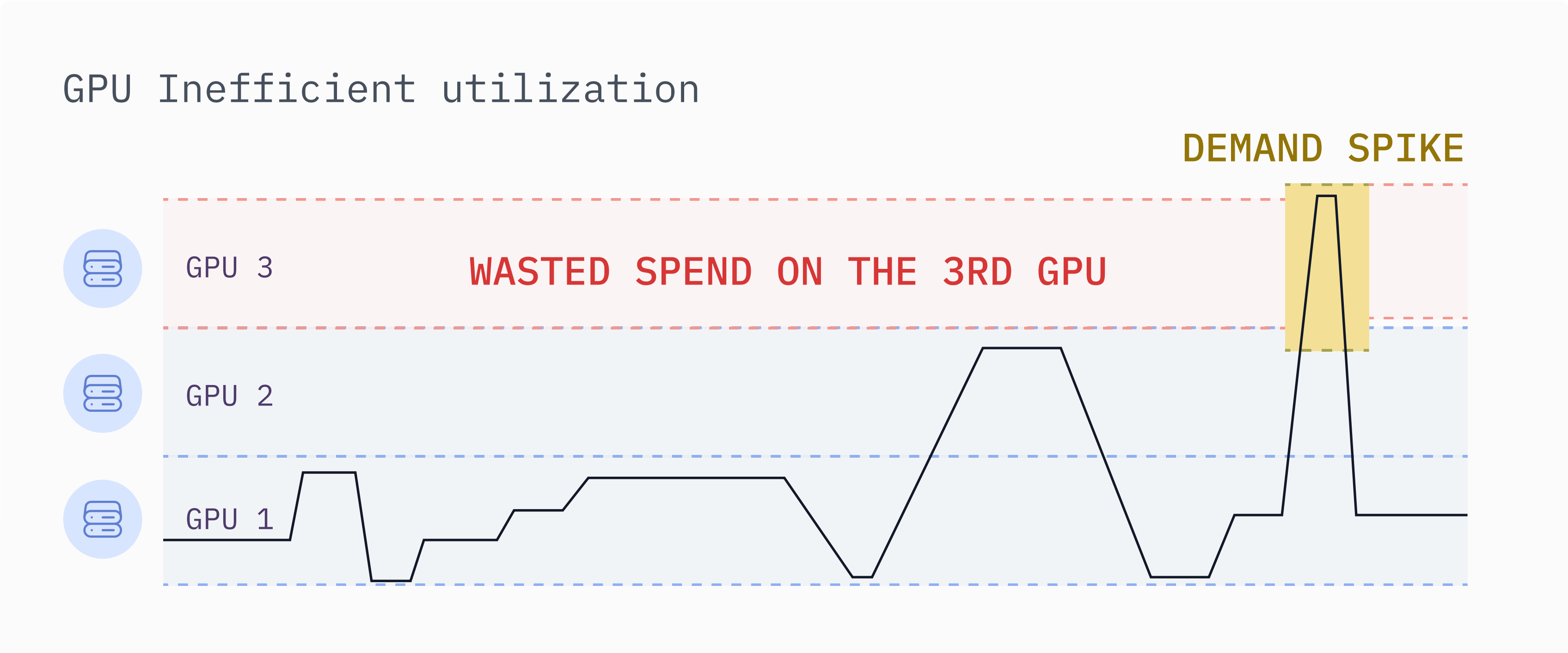
## The Current Infrastructure Gap

*Why Today's Infrastructure Can't Meet Tomorrow's AI Demands*

Most existing enterprise AI infrastructures, even those adapted for early Generative AI, are not built for this level of dynamism and unpredictable token consumption. While many organizations leverage cloud-native MLOps platforms or custom solutions for managing AI/ML pipelines, **these approaches often fall short when confronted with agentic AI's unique, variable inference demands.**

These existing methods typically focus on individual model life cycles, providing tools for training, deployment, and basic monitoring. However, they lack the inherent capability for dynamic, multi-model orchestration across diverse compute environments that agentic AI requires. Their auto-scaling mechanisms, while helpful for predictable traffic, often cannot react with the necessary granularity or speed to handle an agent's erratic, inter-model calls. Furthermore, relying heavily on a single cloud provider can lead to vendor lock-in, and integrating disparate tools for multi-cloud or hybrid environments still requires significant manual effort. This creates a critical infrastructure gap, leading to:

- **Wasted Spend:** Idle GPUs waiting for unpredictable agent spikes.
- **Performance Lags:** Agents waiting for resources to become available.
- **Operational Complexity:** Teams manually juggling resources to keep pace.
- **Production-Level Availability:** Ensuring consistent availability for complex AI workloads remains a substantial challenge, even for major AI service providers.



The urgency to adapt your AI infrastructure is paramount. Organizations that fail to bridge this gap will find themselves outmaneuvered, burdened by unsustainable costs, and unable to fully capitalize on the transformative power of Agentic AI.

# 3. Challenges for Enterprise AI

## Hidden Costs and Strategic Bottlenecks

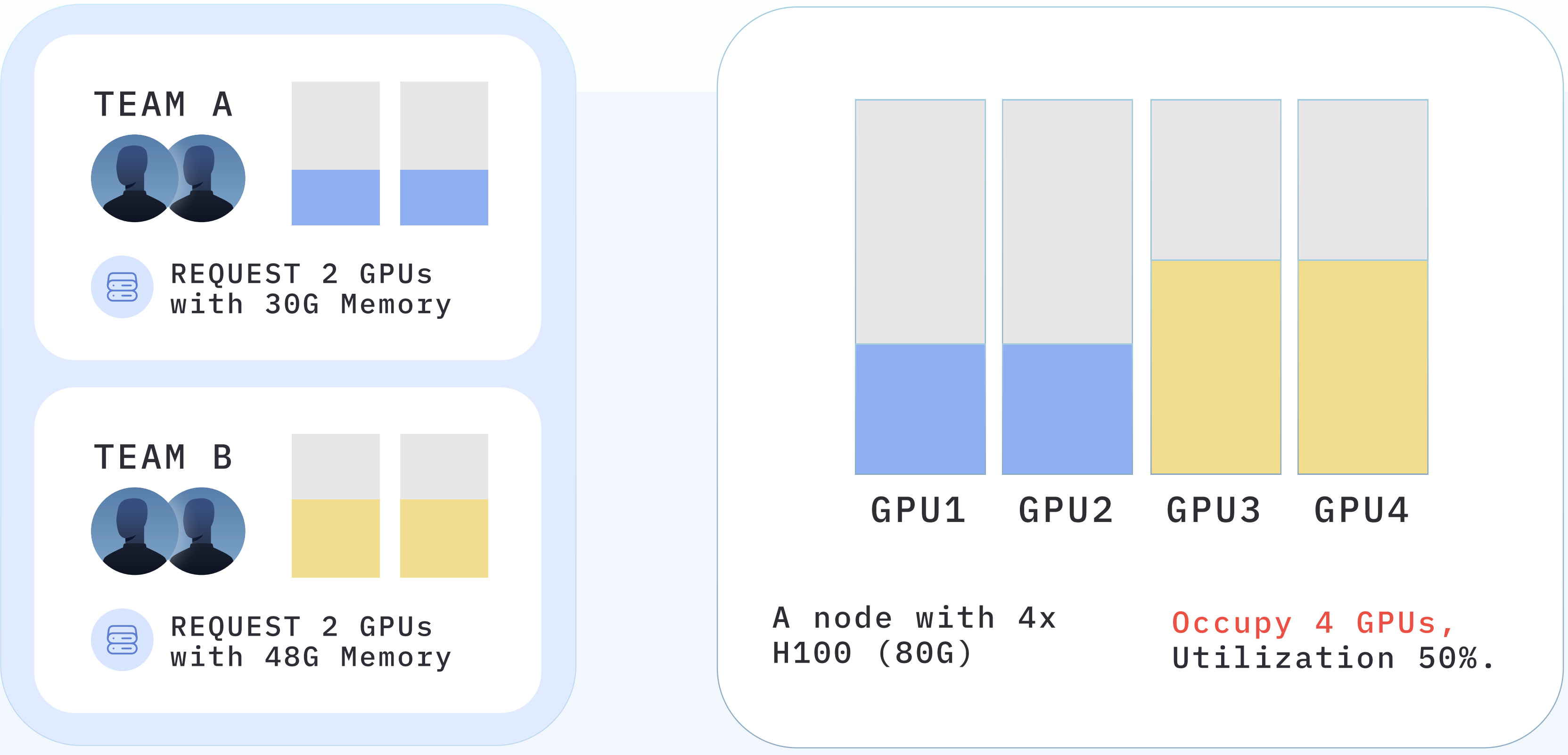
While the promise of AI is clear, its implementation often comes with significant, often hidden, costs and strategic bottlenecks that can derail even the most ambitious initiatives. Decision-makers must be acutely aware of these challenges to effectively navigate the AI landscape.

## Skyrocketing Costs & Poor Utilization

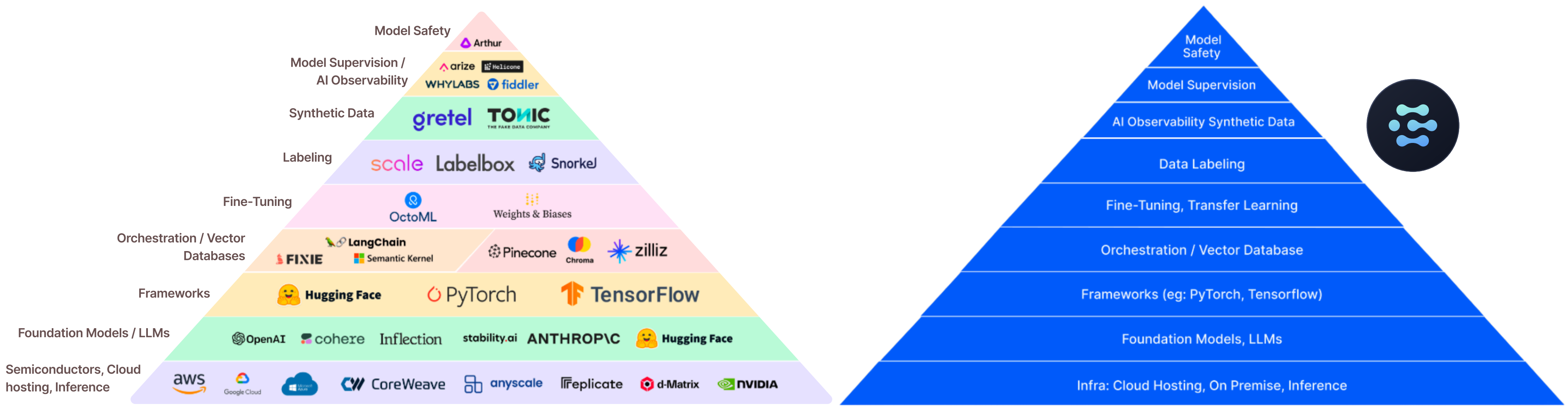
The most immediate and impactful challenge is the escalating cost of AI compute. The proliferation of AI, particularly the consumption-heavy generative models, is driving an unprecedented surge in cloud expenditures. This escalating cost is compounded by a fundamental lack of predictable or fixed costs, making budget forecasting and resource management a constant challenge. Our data shows a staggering 48% waste in compute resources, meaning nearly half of your valuable AI infrastructure budget is underutilized. Furthermore, a disproportionate 60% of AI/ML budgets are consumed by infrastructure alone, leaving insufficient funds for talent acquisition, model development, and core innovation. This is largely due to:

- **Under-utilized GPUs:** Expensive GPU resources often sit idle or are inefficiently allocated, leading to significant capital expenditure waste.
- **Static Provisioning:** Over-provisioning to meet peak demands, resulting in idle capacity during off-peak times.
- **Lack of Visibility:** Inability to accurately track and attribute usage across teams.

This unchecked expenditure directly impacts your ability to scale AI initiatives and diverts critical funds from areas that could drive competitive advantage.


TEAM A
REQUEST 2 GPUs with 30G Memory
TEAM B
REQUEST 2 GPUs with 48G Memory
GPU1
GPU2
GPU3
GPU4
A node with 4x H100 (80G)
Occupy 4 GPUs, Utilization 50%.

## AI Sprawl / Numerous Toolchains and Toolkits



The rapid proliferation of AI models, frameworks, and open-source toolkits has led to a chaotic "AI sprawl" within many enterprises. The average organization uses 10+ tools to create AI applications, each with its own learning curve, integration challenges, and management overhead. This fragmented ecosystem results in:

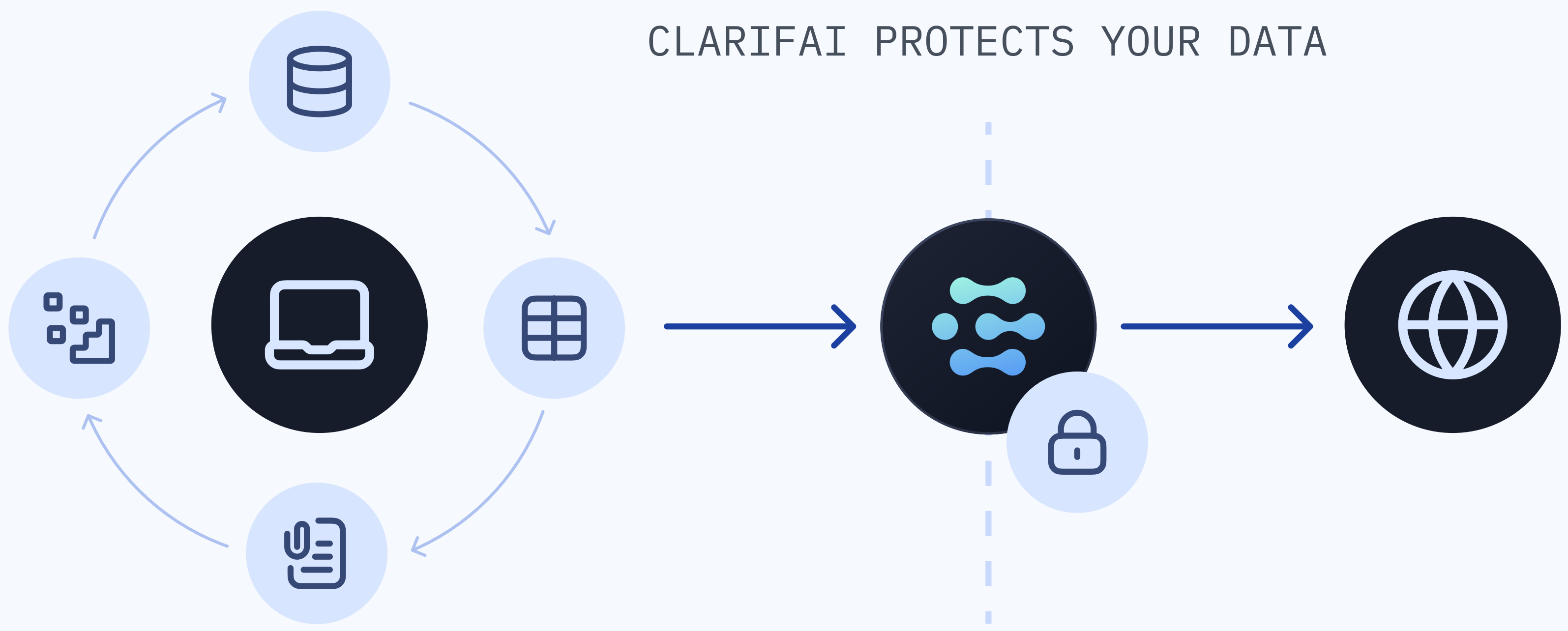
- **Immense Operational Complexity:** Managing a diverse and rapidly evolving AI/ML stack is immensely complex, including the intricate challenges of wrangling Kubernetes clusters for AI workloads, optimizing library and dependency management across frameworks, and implementing sophisticated multi-cloud autoscaling strategies.
- **Critical Talent Scarcity:** The highly specialized skills required for these complex operational tasks are possessed by very few DevOps and MLOps professionals. This severe talent scarcity exacerbates operational burdens, slows down development, and increases the likelihood of costly errors.
- **Developer Frustration & Reduced Velocity:** Engineers are often forced to spend valuable time on toolchain integration and maintenance, diverting them from core AI development. This fragmented environment directly impacts development velocity and leads to significant frustration.
- **DevOps Burden**: Inconsistent environments and disparate tools make it difficult to reproduce results, hindering model iteration and deployment. This sprawl places immense pressure on infrastructure and operations teams, with "64% of DevOps teams desiring consolidated toolchains" to manage their constantly expanding, disconnected environment.

## Security Risks & Governance Gaps

The distributed and often ad-hoc nature of AI development and deployment introduces significant security vulnerabilities and governance challenges. Disconnected tools and disparate deployments create vast attack surfaces, making it incredibly difficult to ensure compliance with internal policies and external regulations. Key concerns include:

- **Data Security**: Ensuring sensitive data used for training and inference remains protected across various environments, with particular attention to data sovereignty—the concept that data is subject to the laws and governance structures of the nation in which it is collected or processed.
- **Model Integrity**: Preventing unauthorized access or tampering with proprietary models.
- **Compliance & Auditability:** The struggle for centralized control, comprehensive audit logs, and fine-grained access management across the entire AI lifecycle. Without robust governance, enterprises face not only financial and operational risks but also reputational damage and regulatory penalties.

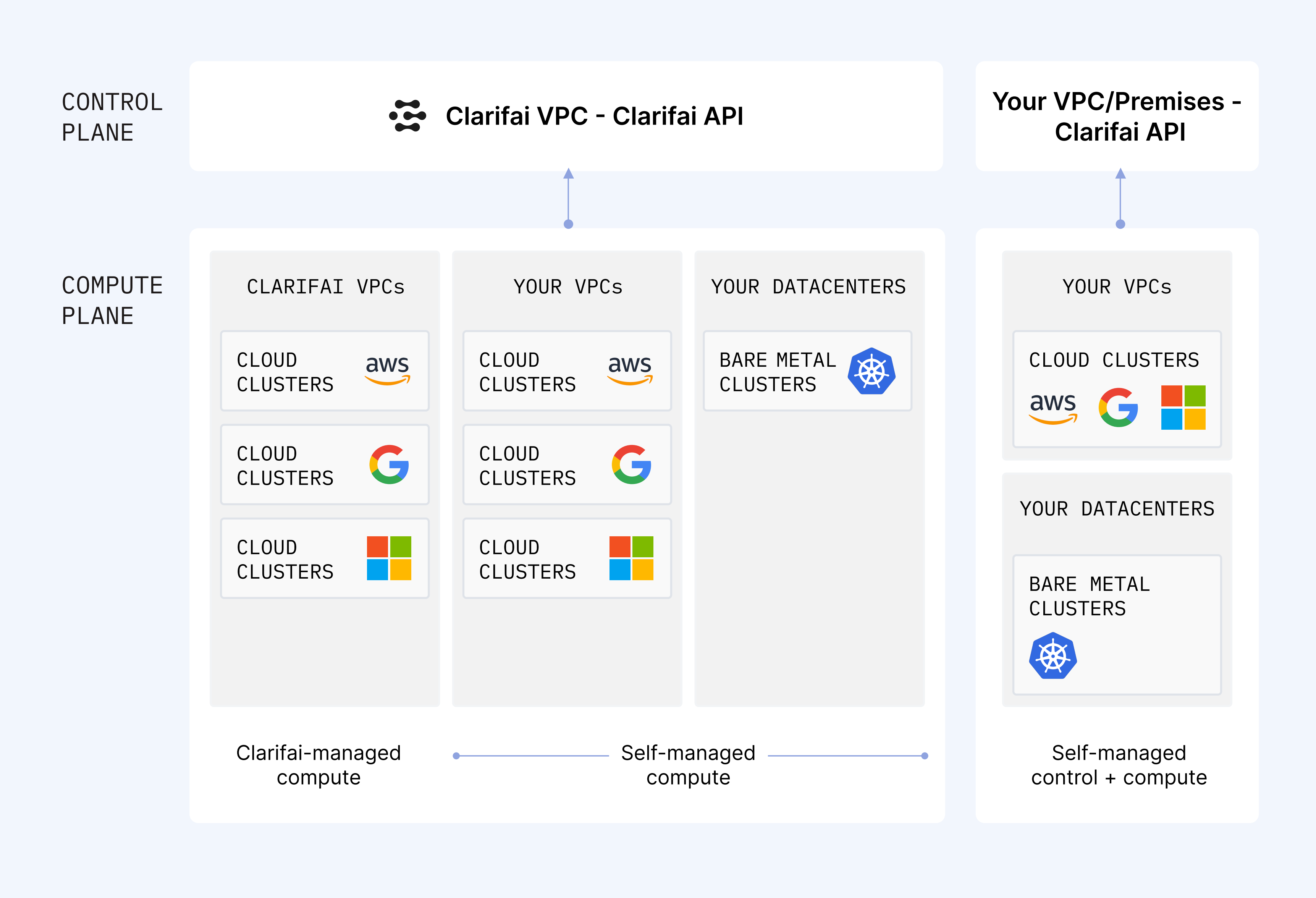
These challenges are not limited to large enterprises; agile startups also face them, albeit with different manifestations. While startups prioritize speed to production, they often hit scaling walls and face increasing costs as their AI initiatives mature, underscoring the universal need for intelligent infrastructure management.



# 4. Compute Orchestration - The Foundation for Modern AI

The complexities and costs of enterprise AI demand a new strategic approach: Compute Orchestration. This is not merely a tool; it's a foundational shift in how organizations manage, optimize, and scale their AI capabilities.

## Introduction to Compute Orchestration



Compute Orchestration is the intelligent, automated management and optimization of AI workloads across diverse and distributed compute infrastructures. It provides a unified control plane that abstracts away the underlying complexity of hardware, cloud providers, and AI frameworks, allowing organizations to deploy, monitor, and scale AI models with unprecedented efficiency and control. CO directly addresses the challenges of skyrocketing costs, AI sprawl, and governance gaps by **ensuring that every AI workload is matched with the most appropriate, available compute resource at the optimal cost.**

## How Clarifai Optimizes Compute for Your Enterprise

Clarifai's Compute Orchestration platform is designed to deliver tangible benefits, ensuring strategic ROI and operational excellence for executives and InfraOps teams.

## Benefits at a Glance

| Feature / Capability | Executive Benefits | InfraOps Benefits |
|---|---|---|
| Cost Optimization<br>• GPU Fractioning<br>• Autoscaling<br>• Batching Inference | Dramatically reduces AI infrastructure spend, freeing budget for innovation. | Maximizes GPU utilization, automates scaling to prevent waste, and optimizes GPU allocation. |
| Unified AI Platform & Model Agnosticism | Accelerates time-to-market for AI initiatives by streamlining the AI lifecycle. | Simplifies management of diverse models across any infrastructure. |
| Enterprise-Grade Governance & Visibility | Ensures compliance, mitigates security risks, and provides clear oversight. | Centralizes control, offers granular RBAC, comprehensive audit logs, and real-time analytics. |
| Leading Performance & Reliability | Ensures responsive AI applications, enhancing user experience and critical operations. | Ensures high throughput and uptime under load. |
| Flexible Hybrid Cloud Architecture | Future-proofs AI strategy by enabling deployment across any environment, optimizing spend by choosing the most cost-effective GPU providers and leveraging existing hardware investments. | Simplifies connecting and managing compute resources wherever they reside, enabling dynamic selection of optimal providers and maximixing utilization of owned infrastructure, while eliminating vendor lock-in. |
| Agentic AI Readiness<br>OpenAI Compatibility<br>Toolkit Support | Positions the organization at the forefront of innovation, enabling rapid adoption of agentic AI | Seamless integration with leading agent development tools and managed hosting for agentic workloads. |

## Unlocking Value: Clarifai's Core Capabilities

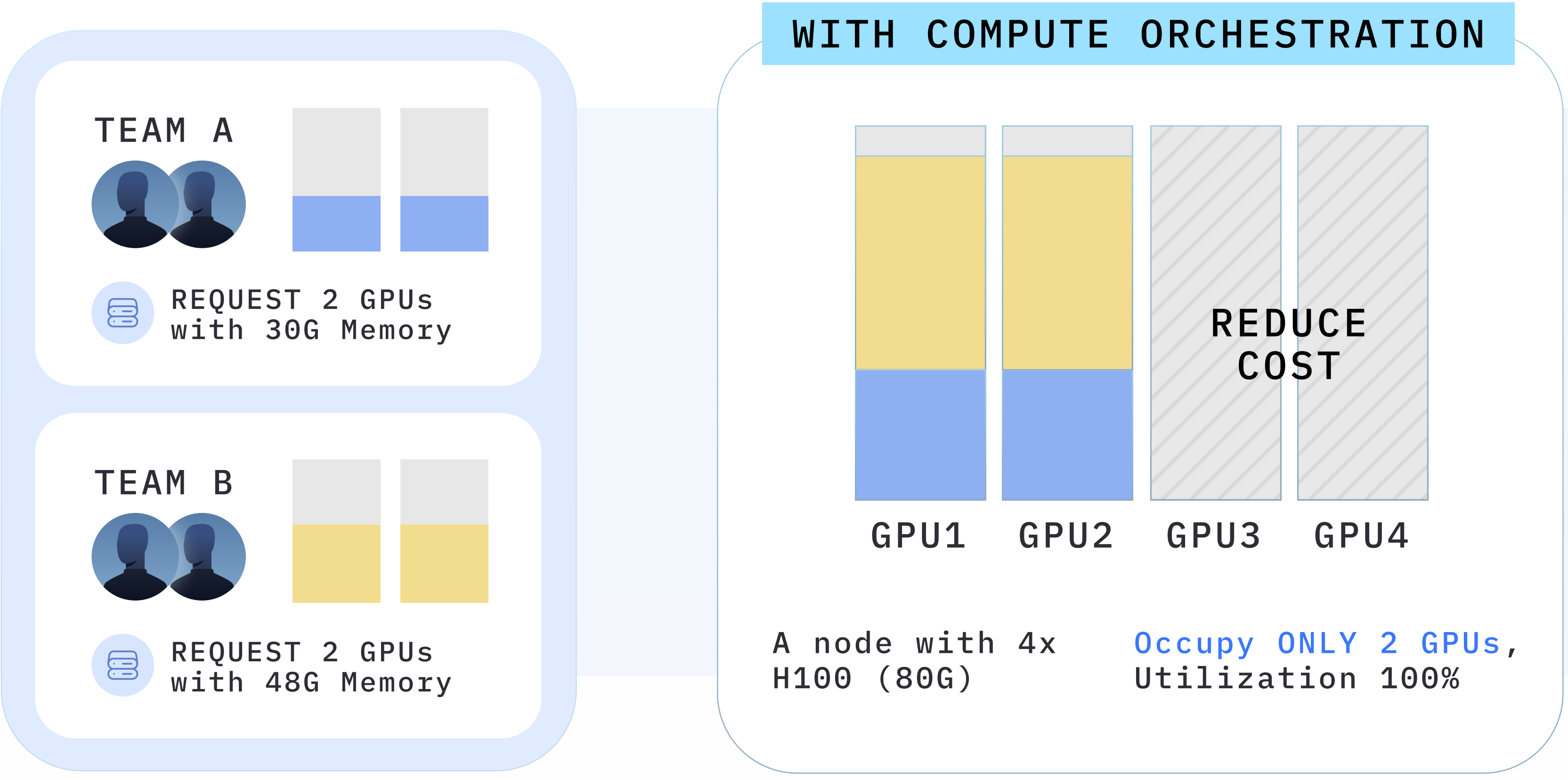
Clarifai’s Compute Orchestration delivers a comprehensive suite of capabilities designed to overcome the most pressing AI infrastructure challenges, translating directly into strategic value for executives and operational efficiency for InfraOps.

## Controlling Skyrocketing Costs & Maximizing Utilization

Clarifai's advanced optimization strategies are engineered to deliver dramatic savings, directly impacting your bottom line and freeing up budget for innovation. Our GPU Fractioning technology allows you to divide and share expensive GPUs among multiple models and teams, ensuring near maximized utilization. Traffic-based autoscaling, including the ability to scale-to-zero, dynamically adjusts resources based on real-time demand, collapsing idle nodepools to zero when not in use. Intelligent batch inferencing helps group multiple requests into optimized batches, further increasing utilization.

Through these innovations, Clarifai delivers significant financial benefits:

- GPU Fractioning: 2-4x savings
- Traffic-based Autoscaling & Scale-to-Zero: 2-10x savings
- Intelligent Batch Requests: 2-3x efficiency gains
- Total Compounded Cost Savings: 16-100x total cost savings
- Overall Compute Savings: Over 90% compute savings


TEAM A
REQUEST 2 GPUs with 30G Memory
TEAM B
REQUEST 2 GPUs with 48G Memory
WITH COMPUTE ORCHESTRATION
REDUCE COST
GPU1
GPU2
GPU3
GPU4
A node with 4x H100 (80G)
Occupy ONLY 2 GPUs, Utilization 100%

## Taming AI Sprawl & Navigating Open-Source Toolkits

Clarifai's unified AI Platform streamlines the entire data and model lifecycle, significantly reducing the complexity and overhead associated with managing disparate tools. Even if you have your own set of favorite tools, ours can help to fill any gaps or enhance your processes, with features such as automated data labeling, versioned dataset management, and streamlined model training and evaluation. Our focus on unmatched vendor & model agnosticism means you can deploy any AI workload, at any scale, across any infrastructure – whether on the cloud or with any hardware vendor, allowing you to seamlessly manage third-party, open-source, and custom models all in one place, eliminating vendor lock-in and simplifying deployments.

## Ensuring Security Risks & Robust Governance

Our enterprise-grade platform features provide the comprehensive control and visibility executives demand, ensuring compliance and mitigating security risks. With a Unified Control & Governance plane, you gain a single-pane-of-glass view for comprehensive oversight of all AI resources. We offer robust role-based access controls (RBAC) for granular permissions, comprehensive audit logs to track actions across users, teams, and projects for full accountability, and analytics for insights into model usage, compute utilization, and spend.

AWS QUALIFIED SOFTWARE · CMMC · SOC 2 TYPE 2 · GDPR

NIST AI RMF · NIST 800-171 · CLARIFAI TRUST CENTER

Our intuitive Org/Team management allows you to structure your AI initiatives efficiently, aligning with your organizational hierarchy. This critical "visibility" empowers InfraOps teams with real-time insights and control over compute consumption organization-wide, preventing runaway AI and ensuring efficient resource allocation.

## Ensuring Peak Performance & Adaptive Scalability

Beyond cost and governance, Clarifai's platform is engineered for unparalleled speed, stability, and adaptability, ensuring your AI applications perform optimally under any load and across any environment. We deliver proven 1.6M+ requests/sec and 99.999% reliability under load, with low-latency streaming inference and optimized kernels ensuring rapid response times critical for real-time agentic workflows and enhanced user experience. Our highly adaptable architecture allows you to easily connect any existing compute, whether in the cloud or on-premise, and dynamically group CPU/GPU types into separate node pools. This hybrid cloud-to-on-prem flexibility ensures optimal performance and data sovereignty, while providing the convenience of serverless compute anywhere, with the full control and governance you require.

1.6M+

Inference requests/sec supported

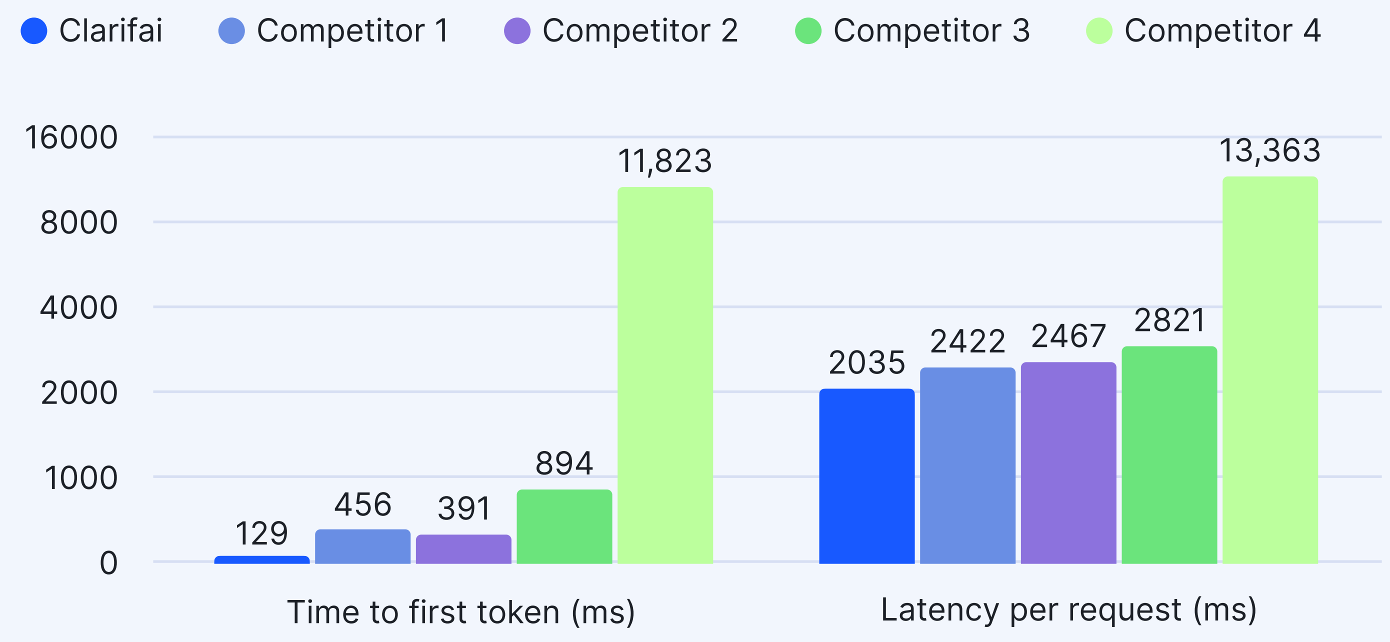
99.99%

Reliability under extreme load

-90%

Compute required

**Ultra-Low Latency:** For AI applications to be truly effective, especially in real-tie scenarios, minimal latency is paramount. Clarifai's platform is engineered to significantly reduce the time from an AI request's initiation to the delivery of the first response token. This rapid processing ensures applications operate smoothly and provide immediate feedback.



Performance comparison between Clarifai and Market leaders

(Lower means better for time and latency)

# 5. The Agentic AI Imperative

## Future-Proofing Your Enterprise with Clarifai

The shift to Agentic AI is not just an evolution; it's an imperative for future-proofing your enterprise. As agents become the dominant paradigm for complex AI applications, your infrastructure must be ready to support their dynamic, unpredictable, and resource-intensive nature. Clarifai's Compute Orchestration is uniquely positioned to be the essential infrastructure for this emergent landscape. CO specifically aligns with and enables the unique needs of Agentic AI, particularly its unpredictable and model-fluid workloads.

Our platform is built on the understanding that agents will dynamically switch between models, consume tokens erratically, and require seamless integration with various tools and data sources. We've developed cutting-edge features tailored for Agentic AI development and deployment:

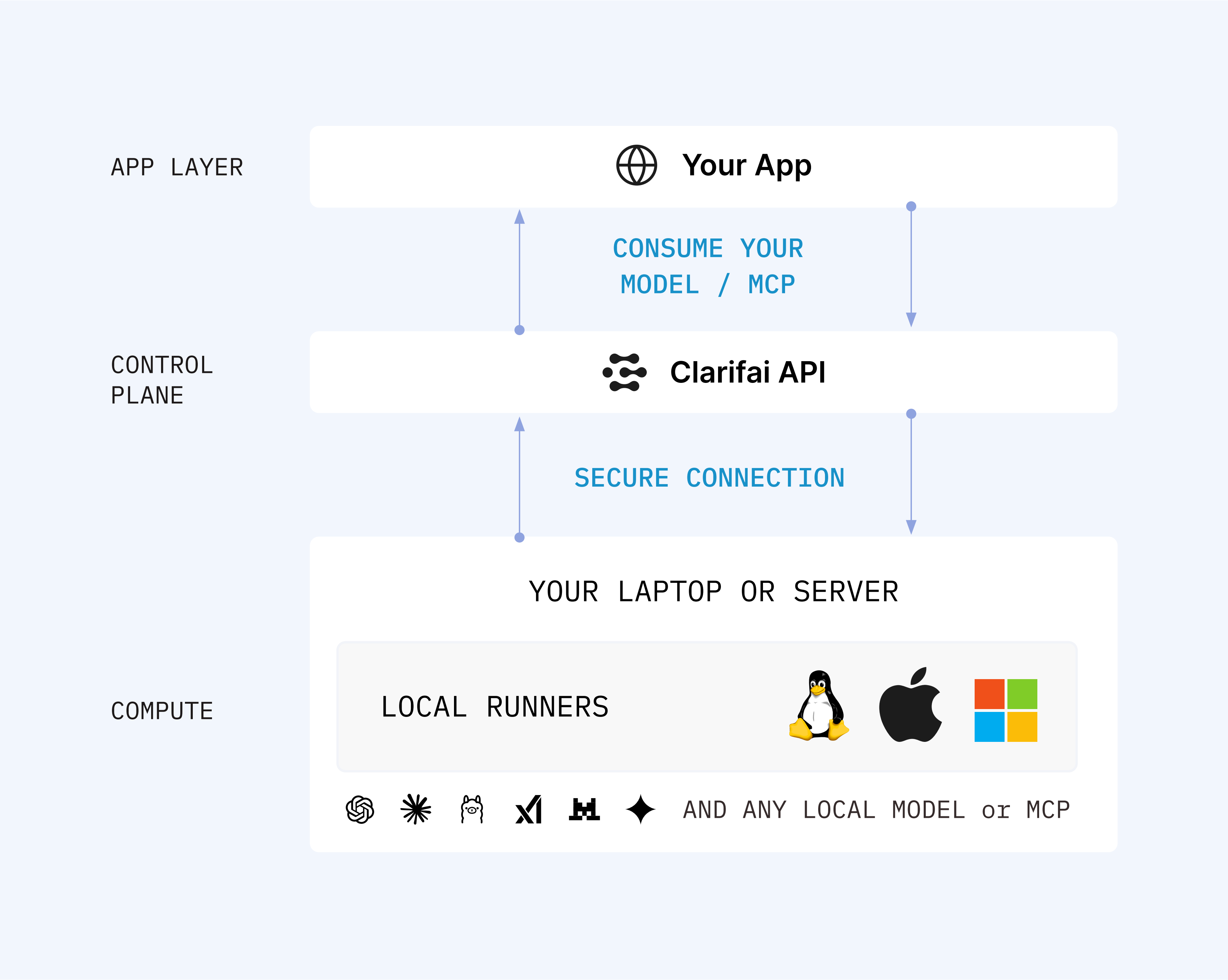
- **MCP (Model Context Protocol) Server and Client Hosting**
  Clarifai provides secure, autoscaled hosting for complex agent workflows. This allows agents to run efficiently, connecting seamlessly to your data and tools through flexible hybrid deployments, ensuring your agents are always operational and optimized.
- **OpenAI Compatible Outputs**
  Our LLMs and Generative AI models are fully compatible with OpenAI's specifications. With more and more agentic toolkits adopting the OpenAI spec, this ensures effortless integration into the broader AI ecosystem and existing agent frameworks, allowing your developers to leverage familiar tools while benefiting from Clarifai's optimized compute.

OpenAI ANTHROP\C Meta Vercel

Visual Studio Code Agent Development Kit LangChain

LlamaIndex All Hands cline crewai LiteLLM

- **Robust Support for Leading Agentic AI Toolkits**
  Clarifai integrates seamlessly with popular agent development frameworks such as LangChain, LiteLLM, CrewAI, Vercel, and others. This empowers your teams to build sophisticated agents using their preferred tools, with Clarifai handling the underlying compute orchestration.
- **Local Runners**
  We empower your developers with secure connections from their local development environments directly to Clarifai's APIs. This enables rapid iteration and quick feedback loops without constant redeployments, ensuring a seamless transition from prototype to production for your agentic AI initiatives.



The strategic advantage for executives is clear: by leveraging Clarifai, you are not just adopting a technology; you are future-proofing your AI strategy. This enables your teams to innovate at an accelerated pace, leading to "2-3x more AI experiments" and critically, unblocking prototypes to production specifically for Agentic AI initiatives. This ensures your enterprise remains at the forefront of AI innovation, turning agentic potential into tangible business value.

## 6. Take Control of Your AI Future

The journey into the age of Agentic AI presents both immense opportunities and significant infrastructure challenges. As this guide has demonstrated, Compute Orchestration is not merely a technological enhancement; it is a foundational strategy for navigating the complex, dynamic, and cost-intensive world of modern AI. Without intelligent orchestration, the promise of AI can be overshadowed by runaway costs, operational chaos, and missed opportunities for innovation.

Clarifai stands as your comprehensive, cost-effective, high-performance, and future-proof solution for enterprise AI. With a decade of pioneering AI Ops experience, we offer the unified platform necessary to control your compute, accelerate your development, and confidently scale your AI initiatives, particularly as you embrace the power of Agentic AI. We empower your teams to build more AI, faster, and with greater return on investment.

### Your Call to Action:

#### Schedule an AI Sprint

Experience a hands-on, immersive lab where you can bring your own workload, deploy fast, and modify your applications with Clarifai tools.

Schedule AI sprint →

#### Use Compute Orchestration

Dive deeper into the Clarifai platform's capabilities to run, optimize, and control your AI workloads across any infrastructure and at any scale.

Learn more →

#### Apply for free credits

Thank you for reading this ebook. You can visit our website and apply to get free credits and begin your journey with Clarifai.

Apply for free credits

# 7. About Clarifai

Clarifai is a leading AI platform company that enables organizations to create, deploy, and manage AI with unparalleled efficiency and control. Founded in 2013, our mission is to make AI accessible and impactful for every enterprise.

## Our Journey and Recognition

With over a decade of experience in AI operations, Clarifai has consistently been recognized as an industry leader. We have been positioned as a Leader in Computer Vision by Forrester Wave, and consistently recognized by Gartner in their Magic Quadrant for Cloud AI Developer Services (CAIDS) and Critical Capabilities reports. We are also proud recipients of the "Best Innovation in AI Development Framework" at the AI Dev Summit 2025 and recognized as a Challenger and Fast Mover in AI Infrastructure Solutions by GigaOm Radar.

## Our Commitment

We are committed to providing an enterprise-grade AI platform that prioritizes trust, security, and responsible innovation. Our cutting-edge technology brings industry-best AI directly to organizations securely, safely, and responsibly. We serve a diverse range of customers, from leading enterprises like Lowe's, Sony, and Walmart, to government agencies and innovative startups, helping them unlock the transformative power of AI across various industries including retail, media, manufacturing, and defense.

Clarifai empowers you to "Don't build tools to build AI, build the AI." We provide the robust, flexible, and cost-effective infrastructure that frees your teams to focus on innovation, accelerate experiments, and bring prototypes to production with confidence.

### Contact us

For a personalized consultation and a strategic discussion on how Clarifai can transform your AI infrastructure roadmap.

✉ Contact an AI Expert

### Connect with Clarifai

Visit our site more information and follow us to stay updated on the latest in AI innovation.

Visit site

in LinkedIn