

WHITE PAPER

Understanding Inference and the Economics of Enterprise AI

A Framework for Investors:

What Is Inference? Why Does It Matter? How Does It Impact Value Creation?

MONTI SAROYA

Senior Managing Director,
Co-Head of Flagship Fund

LILY PEREZ

Operating Senior Vice President,
Flagship Fund





A New Economic Reality for AI and Software Companies

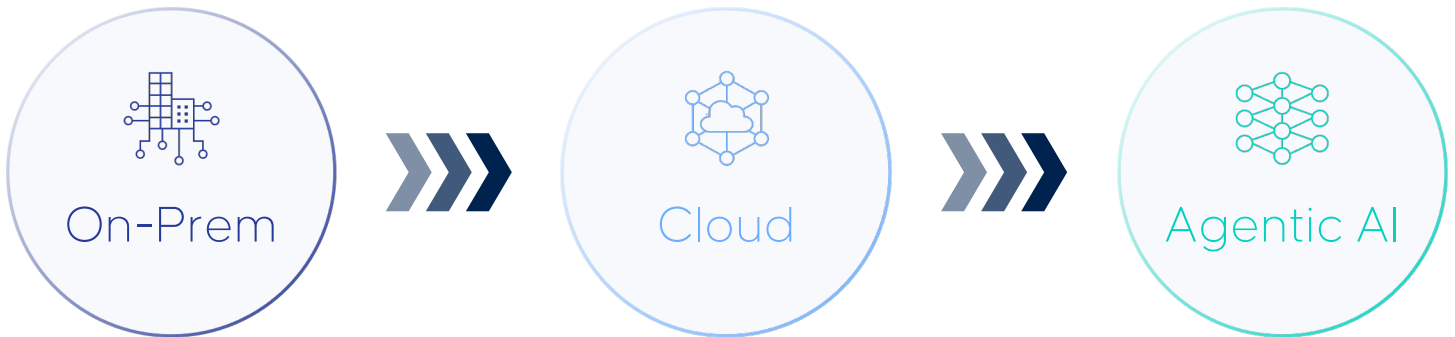
Prior technological transitions have reshaped how value gets created and captured by software. We believe the transition to Agentic AI could be the most consequential value unlock to date.

Unlike traditional software, which automates predefined workflows, Agentic AI systems can reason, make decisions and execute multistep tasks with minimal human direction. As a result, software is evolving from a tool that supports work into a digital worker that executes, and the addressable market may expand accordingly. Third-party research estimates a \$3 trillion opportunity for software as AI agents take on tasks previously performed by people.¹

However, this revenue opportunity arrives with a new structural cost. AI requires compute at every stage of its lifecycle, from model training and tuning as AI labs develop new models, to the inference required to run AI models once launched. Every action taken by an AI agent requires inference, and that inference has an associated cost. For software companies delivering Agentic AI products at scale, inference is emerging as a significant line item in the P&L.

We believe the software companies that thrive in this era will need to adopt agentic AI in the way it creates and expands value for its end customers, while also managing the cost of running those agents – inference is not a fixed input.

This paper explains what inference is, why it matters and how it can be managed.



SOFTWARE AS A

Product

Pricing: Licensing
Growth: Episodic

Service

Pricing: Seat-based
Growth: Linear

Worker

Pricing: Usage-based
Growth: Exponential

¹ Google Cloud and BCG, "The Transformative Potential of Agentic AI and the Strategic Imperative for Google Cloud Partners," March 2026. Analysis reflects a bottom-up sizing model across approximately 330 task categories spanning Financial Services / Insurance (banking, insurance, capital markets), Healthcare / Life Sciences (providers, payers, pharma, biotech), TMT (technology, telecommunications, media, entertainment, and gaming), Auto / Industrial (automotive, manufacturing, construction, utilities, transportation), Retail / CPG (retail, restaurants, food and beverage), and Public Sector (education, government institutions).

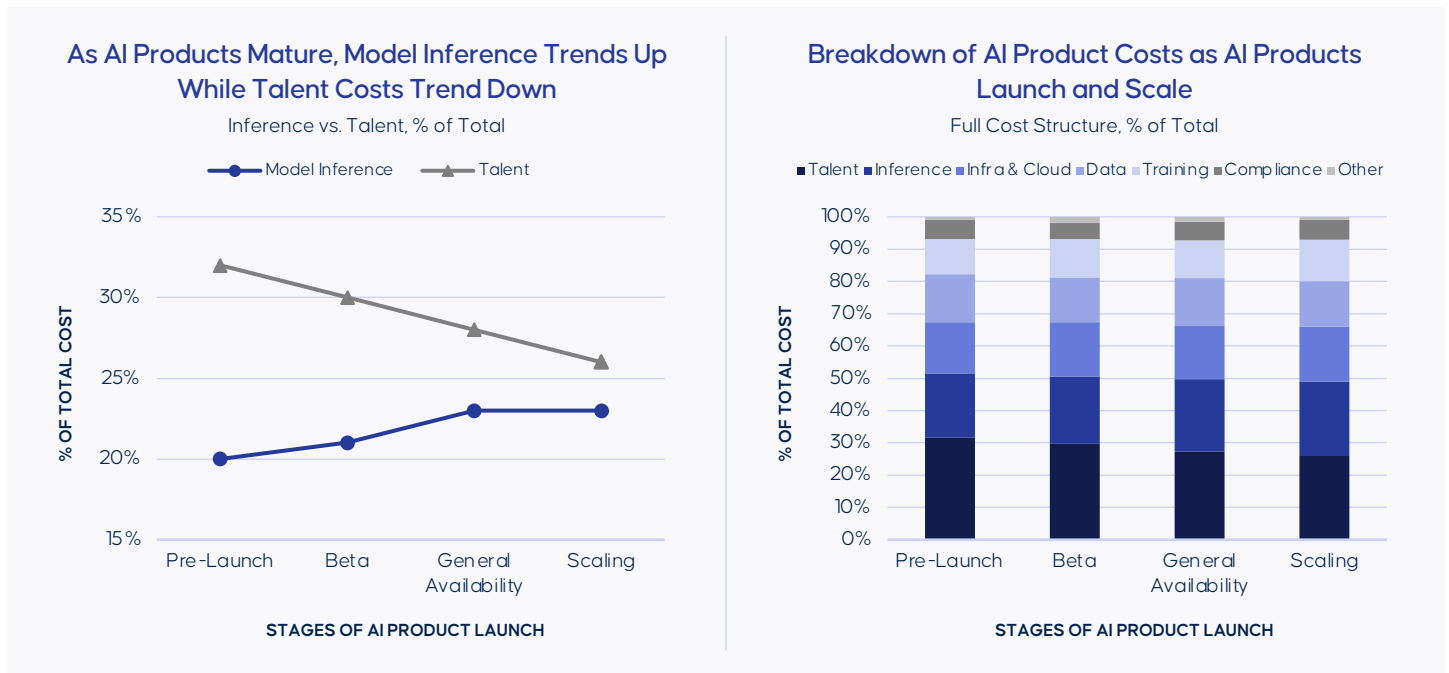
Inference Is the Variable Cost of Running AI

Agentic AI systems are powered by AI models, the foundational technology that enables these systems to reason, generate language and take action. These models are developed by specialized companies, including Anthropic, Google and OpenAI. Just as SaaS companies pay cloud providers like Amazon Web Services for hosting, software companies that build AI products pay these providers for their models and intelligence. Understanding how models are built and run is essential to understanding the cost structure of Agentic AI software.

AI models have two cost components. The first cost is fixed and paid by the company that builds the model. The second is variable and paid by everyone who uses the model.

- **Training is the one-time cost of building a model.** Training the most advanced models – what the industry calls frontier models – now costs hundreds of millions of dollars, borne by the companies that build them.²
- **Inference is the variable cost incurred every time a model is used.** Unlike training, it is paid by whoever runs the model, whether that be an individual user or a business.

To understand why inference is becoming such an important line item for software companies, it is helpful to understand the mechanics of how inference is priced.



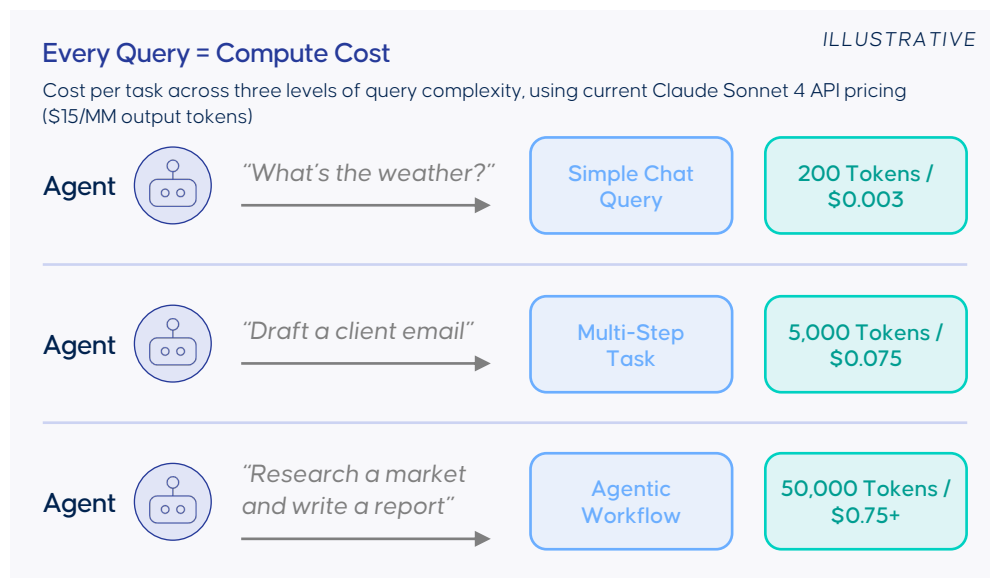
Source: ICONIQ Capital, *State of AI: Bi-Annual Snapshot – The Execution Era of AI* (January 2026). Based on survey of 202 enterprise AI leaders across product stages. "Other" includes compliance and miscellaneous costs. Model inference rises from 20% of total AI product cost at pre-launch to 23% at scaling stage in this sample; talent declines from 32% to 26% over the same progression. Findings reflect early-stage data on a rapidly evolving cost structure.

² Ben Cottier et al., Epoch AI, "The Rising Costs of Training Frontier AI Models," updated October 22, 2025.

Inference Is Priced and Billed in Tokens

Model providers charge for inference based on usage, measured in units called tokens. A token is a small piece of text, roughly three-quarters of a word. The phrase, "Hello, how can I help you?" is about seven tokens. Every time a model is used, it processes input tokens (such as a question and any context) and generates output tokens (the answer). The model provider charges for both.

Today, the most cutting-edge, advanced models charge between \$1 and \$75 per one million tokens, depending on the model and whether the tokens are input or output.³ Older or smaller models cost a fraction of that. As a point of reference, a single ChatGPT-style query currently consumes roughly half a cent of compute. Individually, the cost is small, but as consumption rises – particularly among enterprise customers – it can become consequential.



Enterprise Inference Consumption Is Rising Rapidly

As an individual consumer, you may pay an AI model provider (like Anthropic or OpenAI) a monthly subscription fee. Enterprise pricing works differently.

When an enterprise software company embeds an AI model within its platform, the software company pays the model provider for every inference call its customers generate. The cost is variable and grows with every user, every query and every action an agent takes.

Enterprise AI workloads consume far more tokens per task than individual use. An individual request to draft an email might use a few thousand tokens. But an enterprise-grade agent that processes an insurance claim by reasoning through the steps, looking up policy details and calling specialized sub-agents can use five to 30 times more tokens to complete the task.⁴ Multiply that across thousands of employees running agents, and inference can become a material variable cost in the business.

According to a Deloitte survey of 550 U.S. enterprise leaders, many organizations already exceed 10 billion tokens per month, and the share expecting to surpass 100 billion is projected to triple by 2028.⁵ The average enterprise spent roughly \$7 million on AI model usage in 2025, nearly triple the \$2.5 million spent in 2024,⁶ and some enterprises are now reporting monthly inference bills in the tens of millions of dollars.⁷

³ Anthropic, "API Pricing," accessed May 2026.

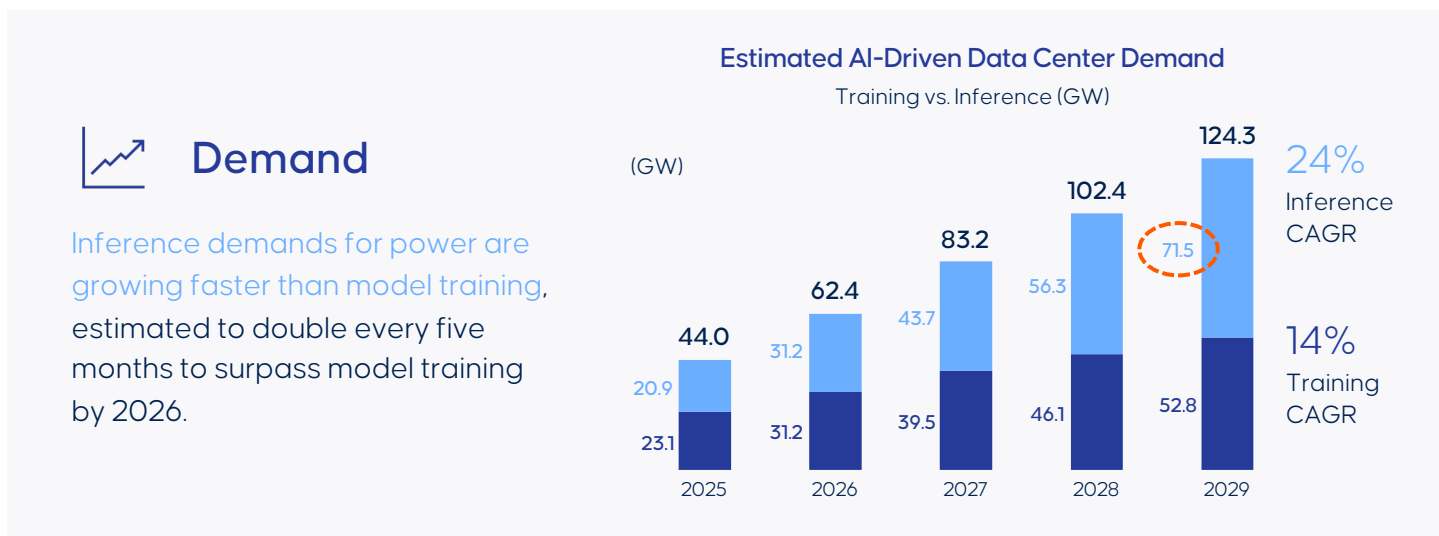
⁴ internal.ai, "AI Token Usage Guide (2026) – 10 Use Case Cost Profiles," March 29, 2026.

⁵ Deloitte Center for Integrated Research, "Deloitte's Enterprise AI Infrastructure Survey: A 2028 Outlook," March 30, 2026.

⁶ SoftwareSeni, "The AI Inference Market in 2025 – Hardware Consolidation, Pricing Wars, and What It Means for Buyers," March 18, 2026.

⁷ Deloitte, "The AI Infrastructure Reckoning: Optimizing Compute Strategy in the Age of Inference Economics," February 2026.

By 2030, Inference Is Expected To Surpass Training as the Dominant AI Workload in Global Data Centers



Source: McKinsey & Company, *The future of AI workloads* (February 2026), drawing on the McKinsey Data Center Demand Model. Supporting analysis in *The next big shifts in AI workloads and hyperscaler strategies* (December 2025). Global data center demand projected to grow from 82.3 GW in 2025 to 219 GW by 2030 (22% CAGR). Inference workload demand grows 35% annually from 20.9 GW to 93.3 GW, overtaking non-AI workloads by 2029. Training grows 22% annually to 62.2 GW. Forecast reflects a midrange “continued momentum” scenario; McKinsey notes that the precise growth trajectory of inference workloads remains uncertain given improving hardware efficiency curves.

Managing AI Cost of Goods Sold Is Critical to Capturing Value

Software’s Traditional Cost Structure

Software has commanded a premium for the last quarter-century because of an enviable P&L dynamic: build it once, sell it many times and incur near-zero incremental cost per customer.

A traditional SaaS company spends roughly 33 to 48% of revenue on sales and marketing, 23 to 30% on R&D (mostly engineering), 15 to 25% on cost of goods sold (COGS, primarily cloud hosting), and 10 to 15% on general and administrative costs (G&A).⁸ The result is gross margins of 75 to 80%, with the largest cost being people, and the largest variable cost being the cloud bill paid to hyperscalers.⁹

Agentic AI Changes Both Sides of the Ledger

We believe the shift to Agentic AI can impact both sides of the P&L: revenue has the potential to expand and operating costs can compress as AI drives operational efficiency. Inference is emerging as the line item that largely influences how much of the upside revenue reaches the profit line.

Operating costs compress. The functions that have historically consumed the most headcount in software companies – R&D, sales and marketing, and customer support – are where AI is already delivering measurable efficiency gains. Across Vista’s portfolio, R&D FTEs (full-time employees) per \$10 million of revenue have decreased approximately 9.5% since 2023, and sales and marketing FTEs per \$10 million of revenue have fallen by more than 30%.¹⁰

⁸ Blossom Street Ventures, “R&D Spend Should Be 24% of SaaS Revenue,” 2024; SaaS Capital, 2025 Spending Benchmarks for Private B2B SaaS Companies, April 2025; Benchmarkit, 2025 SaaS Performance Metrics, 2025.

⁹ Gross margin figure derived from the cost structure ranges cited in footnote 8. Actual figures will vary by company.

¹⁰ Vista internal portfolio analysis as of 12/31/2025.

Revenue expands. In SaaS, revenue potential was capped by headcount, or how many licenses a corporate customer bought for their employees. Agentic AI removes that constraint. When software performs work previously done by humans, the addressable market of potential revenue extends beyond the software budget and into the labor and services budget. Bain & Company projects \$5 to \$7 trillion in cumulative value flowing to AI-enabled software by 2030 – a more than 4x expansion from current levels and potentially the largest addressable-market expansion in software’s history.¹¹

A new variable cost emerges: AI COGS. Inference is the largest component of a new cost category: AI cost of goods sold. Unlike traditional cloud hosting, which scales modestly with users, AI COGS scales directly with usage intensity. Every agent action, every reasoning step and every model call adds cost. For agentic products, inference can already exceed hosting as a cost line, with one analysis finding inference costs average 23% of revenue at AI-native B2B companies.¹²

The Margin Question

This is the central economic tension of the agentic transition: revenue expands and operating costs compress, but a new variable cost emerges. Left unmanaged, the new revenue from agentic products is captured by model providers and hyperscalers rather than by the software company. Inference optimization is therefore one of the most consequential value creation levers in the transition from SaaS to Agentic AI, and margin advantage may accrue to companies that capture the revenue upside while treating AI COGS as a core cost discipline.



Source: Vista analysis as of February 2026. Provided for illustrative purposes only. Ranges provided such as revenue, COGS, gross profit margins, operating expenses, and free cash flow ranges are illustrative to reflect mature, high quality software companies in each respective period. The information presented herein is based upon Vista’s analysis and assumptions and reflect Vista’s beliefs. There can be no assurances that any plans, estimates or expectations noted herein will occur as described, if at all. Moreover, there can be no assurances that historical trends will continue. Statements regarding the impact of artificial intelligence represent opinions and are not statements of fact. Any correlations or relationships shown between market movements and AI-related developments are illustrative and do not imply causation. Please see the Important Disclosures for additional important information.

¹¹ Q4 2025 Bain & Company market study. Figures represent cumulative market capitalization value forecasted to be created between 2023 and 2030 (e.g., 2025 values represent total projected value accumulated to that point). Value creation was estimated through a combination of (A) assessing current market performance of ‘Generative AI Market Leaders’ (Hardware: Nvidia, Intel, AMD, Broadcom, HP; CSPs / Infra: Microsoft, Google, Amazon, IBM, Alibaba; Software: Adobe, Salesforce, ServiceNow, Oracle, SAP) to estimate value created to-date and (B) projecting Generative AI-driven revenue through 2030 using Bain’s AI market forecast to calculate incremental market capitalization gains.

¹² SaaStr, “Inference Costs Average 23% of Revenue at AI B2B Companies,” February 2026.

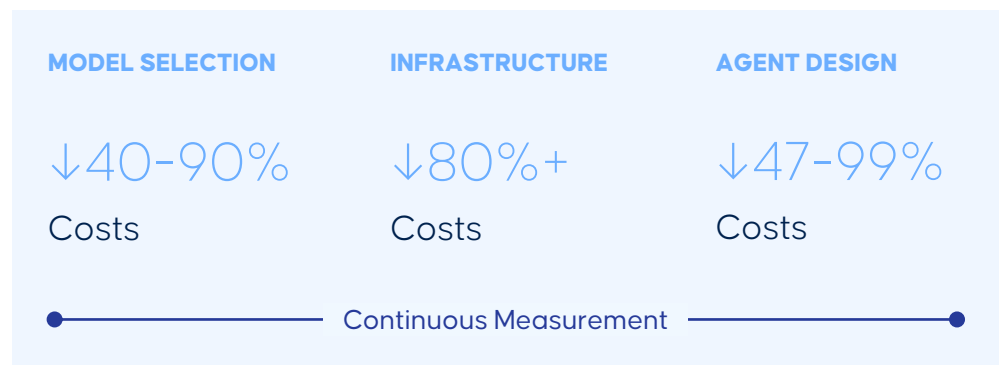
What Disciplined AI Cost Management Looks Like

If inference is the critical new cost line, what can actually be done about it? Our belief, drawn from the agents in production across over 50 of our portfolio companies, is that there are three key levers for driving value. We observe that the same workload, delivered with the same accuracy, can cost dramatically different amounts depending on how these levers are pulled.

Model selection. Not every task requires the most powerful and most expensive model. Models vary in cost per token by 10x or more.¹³ Frontier models are appropriate for tasks that genuinely require state-of-the-art reasoning that most enterprise workflows do not. For example, a claims-processing agent that extracts data from a standard form and checks it against policy terms does not need the same model that powers a research assistant parsing thousands of documents. A well-designed enterprise AI system uses a router that classifies each incoming request and directs it to the cheapest model that can handle it effectively. In many cases, the most cost-effective option is an open-source model – a model whose underlying code is publicly available, allowing companies to run it on their own infrastructure rather than paying a model provider per token. With proper tuning and guardrails, open-source models can perform at or near frontier benchmarks for many enterprise tasks while reducing costs by **40 to 90%**.¹⁴

Infrastructure. Models run on chips housed in data centers. Based on our analysis, frontier models running on general-purpose chips in a commercial cloud are among the most expensive options, because the chips require significant amounts of energy, liquid cooling and specialized physical infrastructure. For many enterprise workloads, it's our belief that simpler infrastructure that is purpose-built for optimized, lower-cost models can often deliver comparable quality outputs while significantly reducing costs. Our research on running Agentic AI workloads has found that **disciplined inference management strategies in aggregate can deliver cost reductions of 80% or more, with accuracy within one to two percent of the most expensive alternative**.¹⁵

Agent design. Every action an agent takes sends a package of instructions to the model, including context, history, tools and a task description. Poorly designed agents send bloated, redundant packages, while well-designed agents strip instructions to their essentials, reuse shared context and avoid unnecessary steps. Vista's internal research found that the way an agent reuses context can reduce inference costs by 47 to 99% depending on the workflow, with the optimization of agent prompts reducing inference by an additional 15 to 40%.¹⁶



Vista's Agentic Factory has built capabilities across each of these levers in partnership with our portfolio companies.

¹³ CloudZero, "LLM API Pricing Comparison in 2026: Every Major Model, Ranked by Cost."

¹⁴ Vista internal portfolio analysis as of 03/31/2026.

¹⁵ Performance estimates ("Estimates") included herein were generated by Vista and prepared to assist in Vista's review of the impact of using open-source models on SambaNova chips for Duck Creek's First Notice of Loss process for preparation for claims decisioning. Estimates are based on modeled assumptions regarding claim volumes (15-25M range), token usage, and system behavior, and are not based on production-scale deployment. Token counts reflect scaled benchmark data and may differ materially from full usage. Cost projections assume a uniform re-run rate, no context caching, and full reprocessing of multimodal inputs, which may overstate or understate actual performance. Model configurations, pricing, and node design are based on March 2026 data and are subject to change. The data inputs are believed to be reliable and are based on current expectations, estimates, projections, targets, opinions and/or beliefs of Vista. The Model involves known and unknown risks, uncertainties and other factors, and undue reliance should not be placed thereon. In addition, no representation or warranty is made with respect to the reasonableness of the Model, which should be regarded as illustrative only. Actual events, results or actual performance may differ materially from those reflected or contemplated in the Model.

¹⁶ Vista internal portfolio analysis as of 05/01/2026.

The Bottom Line for Investors

Three takeaways are worth keeping in mind as you evaluate Agentic AI investment opportunities:

- **The revenue opportunity represents a paradigm shift.** We believe the agentic expansion of the addressable market from software budgets into labor and services budgets presents a material opportunity for software companies that transform to deliver valuable enterprise agentic solutions.
- **Inference is becoming an important P&L line item in the AI era.** Inference costs can scale as companies utilize AI to enhance efficiency across their organization, and as they develop enterprise agentic solutions to expand value for their end-customers. The cost is already material.
- **Inference can be strategically optimized.** The same workload can cost dramatically different amounts depending on how it is architected. The capability to do this work, in our view, is becoming a real source of competitive differentiation. The companies that capture AI-driven operational efficiencies and TAM/revenue expansion from agentic enterprise solutions are poised to improve their P&L's. Those that combine that upside while managing the cost of inference are poised to capture more than their fair share of the economic rent-maximizing both revenue and profitability.

Important Disclosures

This document does not constitute an offer to sell any securities or the solicitation of an offer to purchase any securities. This document discusses broad market, industry or sector trends, or other general economic, market or political conditions and should not be construed as research, investment advice, or any investment recommendation.

Statements contained in this document (including those relating to current and future market conditions and trends in respect thereof) that are not historical facts are based on current expectations, estimates, projections, targets, opinions, beliefs, and/or assumptions Vista considers reasonable. Such statements involve known and unknown risks, uncertainties and other factors, and undue reliance should not be placed thereon. In addition, no representation or warranty is made with respect to the reasonableness of any estimates, forecasts, illustrations, prospects or returns, which should be regarded as illustrative only, or that any profits will be realized. Certain information contained herein constitutes “forward-looking statements,” which can be identified by the use of terms such as “may”, “will”, “should”, “expect”, “project”, “estimate”, “intend”, “continue”, “target” or “believe” (or the negatives thereof) or other variations thereon or comparable terminology. Due to various risks and uncertainties actual events or results may differ materially from those reflected or contemplated in such forward-looking statements. No representation or warranty is made as to future performance or such forward-looking statements.

Certain information contained in this document has been obtained from published and non-published sources prepared by other parties, which in certain cases have not been updated through the date hereof. While such information is believed to be reliable, Vista does not assume any responsibility for the accuracy or completeness of such information and such information has not been independently verified by it. Except where otherwise indicated herein, the information provided in this document is based on matters as they exist as of the date of preparation of this document and not as of any future date and will not be updated or otherwise revised to reflect information that subsequently becomes available, or circumstances existing or changes occurring after the date hereof, or for any other reason.

No representation or warranty, either express or implied, is provided in relation to the accuracy or completeness of the information contained herein.

Operational Intelligence refers to Vista’s longstanding practice of pursuing operational intelligence in its investments, internal approach to deal execution including through Vista’s best practices and the use of Vista’s Value Creation Team.

Artificial intelligence technology models (“AI”), including generative artificial intelligence and similar technologies (“GenAI”), can pose risks to Vista, the Funds, and their investments. AI is an emerging and rapidly evolving technology and therefore it is difficult to fully assess the risks associated with it and those posed to Vista, the Funds, and/or the Funds’ investments. Vista endeavors to evaluate AI models and related risks before using them in its business. However, there can be no assurance that it will do so successfully, and the use of AI may adversely affect Vista and the Funds and/or the Funds’ investments. Vista is exposed to the risks of these developing and evolving technologies, including in situations where AI is used by third-party service, data, or information vendors, or by companies where the Funds have or are considering an investment. Use of AI implicates risks resulting from inaccuracies in data input and output or signals, modeling, and information security and related regulatory developments, among others. Vista and/or the Funds could incur liability or expenses in connection with claims of infringement or similar claims by third parties related to information which Vista receives through GenAI. As a result, these risks may subject Vista to potential litigation (particularly trademark, licensing terms of use, and copyright claims), conflicts of interest, and/or other legal or operational risks. It is possible that new regulations may emerge in this area which impedes or hinders Vista’s ability to use AI in the future. The adoption of proposed regulatory rules regulating AI and other similar systems may also impose additional obligations and expenses on Vista. Vista’s practices regarding the use of AI could potentially disadvantage Vista competitively and there can be no assurance that Vista’s anticipated use of AI will be able to continue without restrictive regulatory requirements. Any of the foregoing factors could have a material and adverse effect on Vista, the Funds and the portfolio companies. As referenced herein, “Agentic AI” refers to AI systems capable of understanding a broader goal and coordinating, to varying degrees, the steps and decisions needed to pursue it and “AI Agent” refers to an AI-powered component that can perceive context, reason about next steps, and take actions toward a specific task, either independently or as part of a larger agentic workflow.

Additional important disclosures can be found [here](#). ©2026 Vista



Invest in *Operational Intelligence*SM

AUSTIN

401 Congress Avenue
Suite 3100
Austin, TX 78701
512.730.2400

CHICAGO

2 Prudential Plaza
180 North Stetson Avenue
Suite 4000
Chicago, IL 60601
312.229.9500

NEW YORK

50 Hudson Yards
Floor 77
New York, NY 10001
212.804.9100

SAN FRANCISCO

4 Embarcadero Center
20th Floor
San Francisco, CA 94111
415.765.6500

HONG KONG

Cheung Kong Center
Suite 1902
2 Queen's Road
Central, Hong Kong

ABU DHABI

Office Unit 1, Floor 7, Al Khatem Tower
Abu Dhabi Global Market Square
Abu Dhabi, Al Maryah Island
United Arab Emirates



Please contact us for more information at vistaequitypartners.com/contact or 512.730.2400

© 2026 Vista