KPMG

TECHNICALLY SPEAKING: A KPMG BLOG SERIES

The first token's free

The unknown long-term costs of AI adoption

The rapid adoption of artificial intelligence (AI) in recent years has led to a surge in companies looking to leverage the technology for a competitive advantage. According to researchers, 55 percent of CIOs will deploy generative AI (GenAI) solutions within the next 24 months[1]. While these forward thinkers are exploring new opportunities made possible through AI—streamlining processes, enhancing productivity, and innovating new products and services—many have not fully accounted for the true "total cost of ownership" of these technologies. The KPMG 2024 GenAI survey shows that 83 percent of respondents believe the investments into GenAI to increase over the next three years[2]. Failure to gain transparency to the full spectrum of financial implications of AI adoption today can create significant challenges in managing costs down the road.

Understanding the potential future costs AI can help your organization develop a clearer understanding of the AI investment landscape and make informed decisions regarding your AI implementation strategies.

83% believe investments into GenAI will increase over the next three years.

1 Source: "2024 CIO and Technology Executive Agenda", Gartner, Oct 17, 2023. 2 Source: "GenAI Survey – 2024", KPMG, Aug 15, 2024.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Running in the dark

Venture capital funding has played a significant role in driving growth for numerous businesses and industries over the past three decades. Many leading companies in sectors such as lodging, food delivery, transportation, and software owe their entry into the market to substantial investments from venture capital firms. These companies capitalized on the availability of external funds, enabling them to establish a strong foothold in their respective markets. Despite inherent risks and uncertainties, their bold strategies paid off as they quickly gained widespread user adoption and market dominance. Eventually, these companies had to stand on their own as viable business, often times increasing prices and adjusting their business models to generate the ROI that their founding investors expect.

The AI trend follows a similar trajectory. Just as in other industries, players with abundant financial resources can offer their AI services to users at minimal costs, ultimately outperforming new entrants. Eventually, as these players mature, they will adjust their business models to maximize service revenue.

First movers in AI are required, by necessity, to move fast through uncharted territory. But forging ahead quickly does not have to mean forging ahead blindly. In our experience, a wide-eyed view of the total cost of ownership (TCO) of any technology, including AI, is essential for long-term success.

KPMG experience in surfacing and managing costs associated with cloud computing can provide useful lessons for cost-effective, value-driven AI implementations. See our white paper Taking control of cloud costs: The FinOps imperative for an instructive perspective.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

The first one is free...

In truth, the costs associated with AI adoption at present may seem quite enticing. Vendors have been making the technology available today at virtual loss-leader pricing, effectively crowd-sourcing development of use cases and establishing a broad base of market penetration.

But as use cases emerge and dependency on a range of AI technologies grows, you can be sure that AI vendors are contemplating how to recoup their development costs and looking for new ways to monetize their offerings going forward. Generative AI is infamous for its hallucinations; perhaps the biggest hallucination of all is that its costs will remain as low as they are today.

We have seen it before with other technologies. Think about the launch of cloud-based email services and data storage from a few well-known players. These began as free services, but once they were refined and usage hit critical mass, they were bundled into "suites" and sold with a healthy enterprise licensing fee.

The interrelation between AI adoption and cloud consumption

Another widely recognized effect of the rapid adoption of AI is the closely linked increase in cloud consumption and usage. Hyperscalers are scaling up and leveraging their own cloud infrastructure to support the heavy computation and data storage demands of AI.

Concurrently, the Hyperscalers are investing billions of dollars into their own AI suites and even more into some of the most thriving AI solutions out there—while Microsoft strengthens its ties with partners such as OpenAI and Google, focusing on Anthropic as a competing force.

The Hyperscalers' bet on the AI industry becomes even more evident when looking at the nature and structure of some of these investment deals. Only a fraction of Microsoft's multi-billion-dollar investment into OpenAI will come as a direct cashflow for the start-up. A large chunk of the funding comes in the form of credits for cloud computing power in the Azure cloud[3] .Similar trends are seen for investments made by Google[4] and others. By strengthening their strategic ties to the key emerging players in the AI industry, the Hyperscalers are ensuring a continuous demand for their computing power and, to some extent, locking the emerging players into their cloud ecosystem for the foreseeable future. The strategic nature of these investments becomes more relatable when looking at the transaction feed for the emerging AI providers. In the example for OpenAI, every token request towards OpenAI, whether through ChatGPT or other AI-native solutions with an API into OpenAI, requires cloud computing to generate the corresponding result. Hence, every AI transaction ultimately ensures cloud consumption for the Hyperscalers.

3 Source: "OpenAI has received just a fraction of Microsoft's $10 billion investment", Reed Albergotti, SEMAFOR, Nov 18, 2023.

4 Source: "Google Commits $2 Billion in Funding to AI Startup Anthropic", Berber Jin and Miles Kruppa, WSJ, Oct 27, 2023

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Uncovering the cost components of AI

Understanding and managing AI costs effectively requires a shift in mindset and practices. Astute companies acknowledge that AI implementation will impact their operating model and necessitate fundamental changes to various processes, including investment budgeting, resource allocation, and forecasting. We can help.

Cost components of GenAI

Pre-Development

- Data harvesting, preparation, ingestion processing, labeling
- Preparing corporate processes
- Developing AI governance and guidance
- Infrastructure modernization
- AI model & platform selection
- Analyzing AI skill gaps and creating training plans

Primary focus of whitepaper

Development

- Model development and training cost of human capital
- Complete GPU resources
- Application development
- Tools, systems and infrastructure modernization
- Subject matter experts (SMEs) to refine model for use case accuracy
- AI benchmarking to measure AI system performances
- Validating & testing AI models
- Planning & development strategy & rollout
- Model interference and serving

Run

- Training on GenAi applications and usage
- AI run cost (e.g. token cost)
- Prompt engineering
- Managing and monitoring AI TCO
- Monitoring AI governance risk, compliance, and usage
- AI-focused organizational change management
- Creating financial control dashboards
- ESG impact management
- Monitoring AI system heath
- Preparing and managing future AI scenarios
- Transition between models and solutions

KPMG offers a unique perspective on AI cost models, emphasizing the importance of identifying and properly categorizing AI costs and understanding their connection in your overall adoption scheme. We can help you think about the bigger picture, incorporating proactive financial management strategies into your implementation plan. This comprehensive approach can help your organization gain holistic oversight, optimize your AI investments, and increase the value generated by the technology.

Continued on next page.

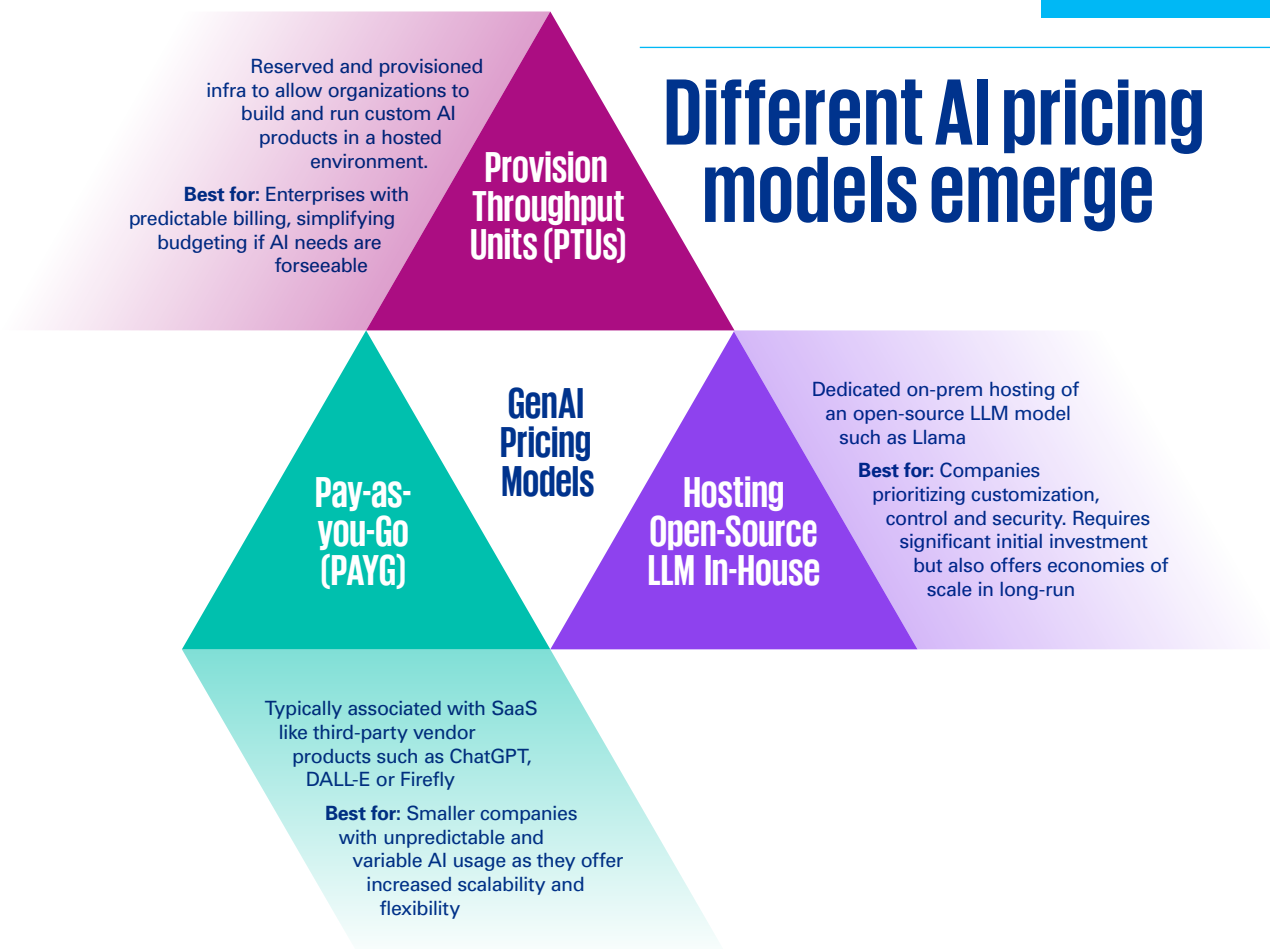
© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Uncovering the cost components of AI *continued*

To help enterprises better understand and manage AI costs, KPMG has categorized cost components into three different maturity stages: **pre-development**, **development**, and **run** costs. While most corporate customers will see most of their cost occur in the run stage, other selective costs from the pre-development and development stages, such as data harvesting, preparations, and labeling, will slowly bulk up and become a financial burden in the long run. Your firm's costs will be specific to your unique business case for AI and your implementation strategy.

The importance of managing data quality and its impact on the value of AI output cannot be overstated. See our white paper A Data-Driven Culture Will Differentiate the Winners from the Losers: What Businesses Should Do to Stay Ahead for valuable guidance on deriving the most advantage from your AI implementations.

Different AI pricing models emerge



Over the last year, the industry has seen a few different pricing models emerge. Understanding the differences and characteristics of each model will be crucial for organizations to select the ideal model based on their individual needs:

Additionally, recent months have seen various smaller providers emerge that offer competitive pricing and better service level agreements (SLAs) compared to the established providers.

Evaluating the different pricing models while considering your businesses strategic objectives, operational needs, and scalability needs is essential for a sustainable AI strategy. As the race for the AI market continues, new models will appear regularly and will require some flexibility and adjustment on the consumers end.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Tallying and managing the "true" cost components of AI

Unlike other software technologies, where costs to develop an application may be high, but costs to distribute and run it fairly low, AI is a different animal.

Consider widely adopted GenAI applications such as ChatGPT and the large language models (LLM) on which they depend. The high cost of training and inferencing these models introduces a novel structural expense that differs from other technologies. And once built and trained, models can require huge amounts of computing power to run the potential billions of calculations required in response to prompts—calculations with no attendant economies of scale as usage increases. Robust calculating capacity may require specialized and expensive hardware, such as graphics processing units (GPUs). This can increase operating costs associated with energy usage, bringing environmental and sustainability expense into the equation as well.

Of course, not all companies will engage with AI in the same measure. Most companies today are in or have passed through an early *"Innovation Phase"* where the primary focus is on prompt engineering and working with out-of-box solutions.

As companies reach the limitations of simple prompt engineering, they strive for more advanced AI capabilities and enter their ***"Enterprise Scaling Phase."*** Many companies first turn to retrieval augmented generation (RAG) embeddings trained on internal data sets before implementing a full scall whitelabel LLM with extensive parameter adjustments.

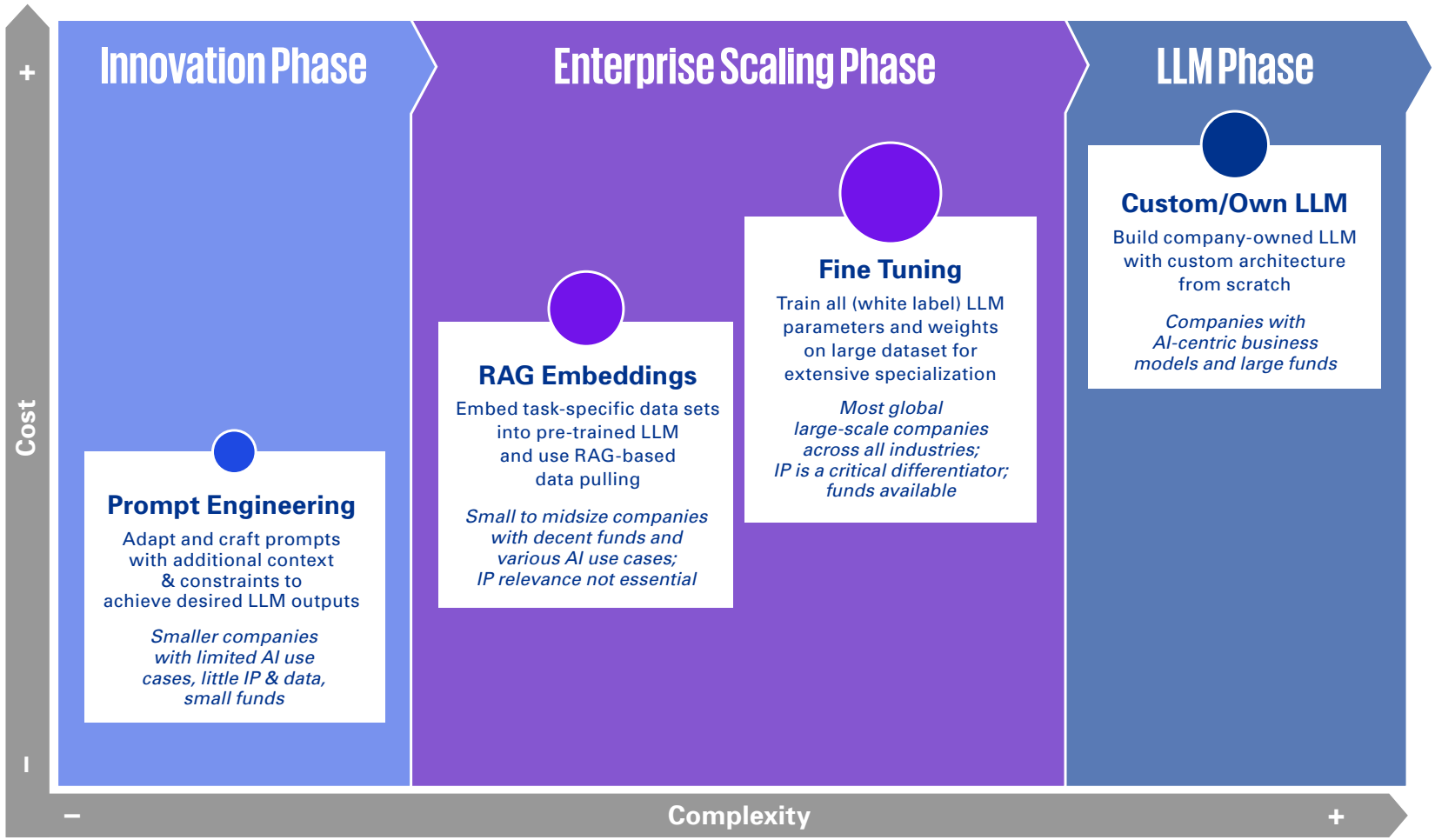
Only an extremely limited number of companies will have the use cases and requisite funds for entering the ***"LLM Phase."*** Hyperscalers—the major cloud platform or service providers—will spend into the billions to develop and train proprietary LLMs for their own consumption and/or to license for use by other companies. Most organizations will find their cost-to-value ratio maximized with RAG embeddings or LLM fine-tuning since the cost to build their own LLM often outweighs the value without subsequent monetization of the LLM to external parties to recoup their extensive training cost and investments.

Models can require huge amounts of computing power to run the potential billions of calculations required .

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Tallying and managing the "true" cost components of AI *continued*

Each company's needs to understand their future AI use cases and applicability to drive long-term decision making



- Low effort plug-and-play applications
- No utilization of company data or IP
- App maintenance with AI providers

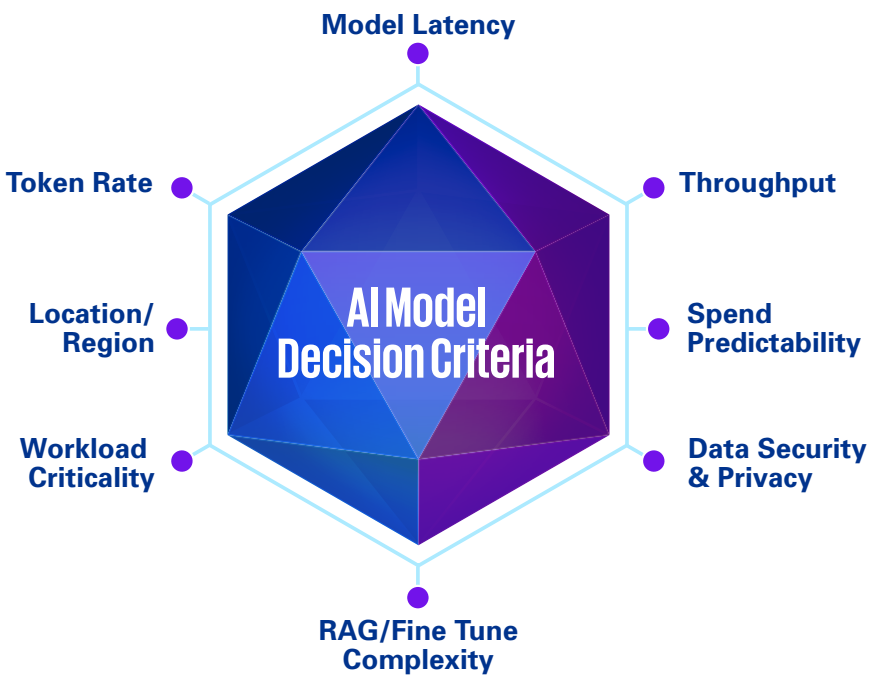
Value:

- High pre-development efforts to prepare for GenAI
- Extensive LLM development & training efforts
- Substantial efforts to run & maintain model and data

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

AI model selection driven by strategic considerations and constraints

Getting a clear understanding of your organization's strategic intentions and guardrails towards GenAI will be crucial in driving transparency and value-focused decisions around AI models. Each company should consider both decision criteria and decision constraints.



Decision criteria primarily align with an organization's strategic intentions and objectives. Our criteria can help an organization define and document strategic and measurable parameters to inform the model selection process.

Decision constraints, on the other hand, are focused on financial, operational or regulatory limitations that a company is facing. Common constraints are:

– Budget and innovation funds for AI

– Availability of in-house AI talent and expertise

– Requirements around data and infrastructure authority

– Sector / market regulation

KPMG has developed a holistic AI model decision framework that can help organizations answer two distinct questions:

1. Which LLM model should your organization choose to deploy (e.g., GPT 3.5, GPT 4o)?
2. How should your organization provision the model (e.g., PTU, PAYG, In-house hosting)?

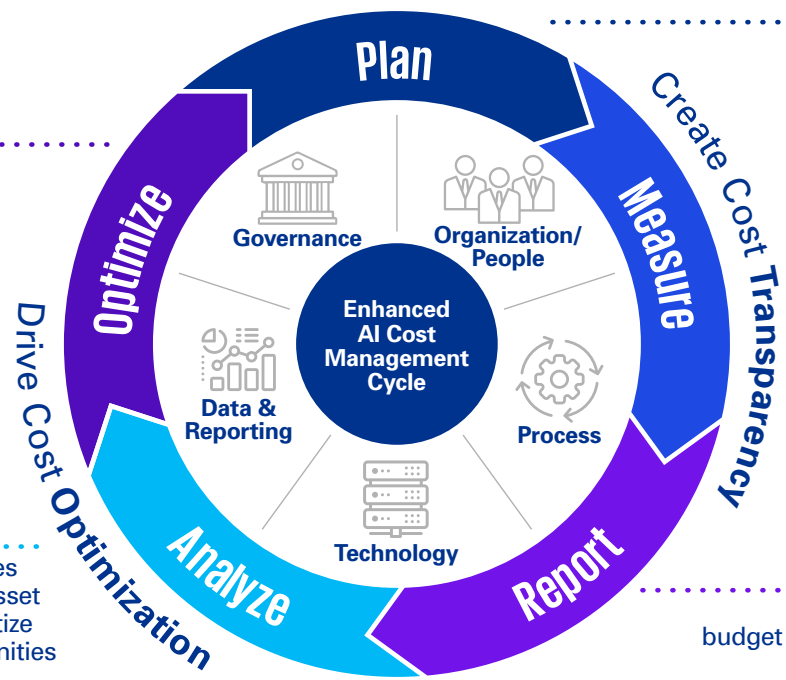
© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Incorporating AI cost management into FinOps

The KPMG integral approach to cloud and technology-related cost management (FinOps) is at the heart of a continuous cycle of planning, measuring, reporting, analyzing, and optimizing the impact of your technology implementations, including those associated with the adoption of AI.

Sound FinOps practices anticipate and accommodate expenses associated with your people, process, technology, data and reporting, and governance. We can help you incorporate AI cost management into your existing FinOps infrastructure, or help you devise a tailored FinOps cost management system from the ground up. Our holistic orientation can enhance your ability to extract the most value from any technology initiative, including deployment of AI, at favorable cost.

An AI-inclusive Financial Management Framework



Plan

Determine expected GenAI use cases, best-fit AI model & solutions, anticipated value, and cost, ahead of large-scale AI investments

Measure

Establish KPIs, forecast licensing costs, understand usage & effectiveness, and identify trends/red flags

Report

Relay cost insights (usage, licensing costs, budget variances, and trends to all stakeholder groups

Analyze

Identify optimization opportunities and collaborate with Software Asset Management to validate & prioritize licensing and other cost opportunities

Optimize

Drive accountability for implemented changes to improve performance and reduce costs

Continued on next page.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Incorporating AI cost management into FinOps *continued*

Our experience in the market shows that a few simple foundational standards and processes can help an organization gain the necessary capabilities and insights to better plan, manage, and control the cost of AI:

1. Develop and establish a **comprehensive TCO** model around AI solutions.
2. Establish **cross-functional governance** with explicit policies and guardrails around investing in, licensing, and using AI (e.g., pricing model decision framework).
3. Adapt existing **processes and standards** to be more inclusive of the unique characteristics associated with the costs of AI: budgeting, forecasting, resource allocation, and so on.
4. Establish a **business unit chargeback process for AI costs** to foster increased ownership and accountability for AI deployment and usage.
5. Implement **standard procedures, metrics, and technologies** to monitor and make costs associated with AI more predictable and amenable to forecasting .
6. Integrate **legal and risk functions** into the AI journey to ensure necessary frameworks and parameters are adhered to.
7. Ensure consistent, efficient, and secure **architecture standards and design patterns**.

Additionally, calibrating gains against losses in AI implementations requires relevant measurement strategies. How are companies assessing and monitoring their success? Some key metrics categories can help your organization effectively track and manage success:

- **Latency Metrics** (e.g., Median End-to-End Latency per Call, etc.)
- **Token Metrics** (e.g., Median Number of Tokens per Call, Maximum Tokens Allowed, etc.)
- **Request Metrics** (e.g., Completed Requests per Minute, Request Error Rate, etc.)
- **Timing Metrics** (e.g., Time to First Token, Inter-Token Latency, etc.)
- **Throughput Metrics** (e.g., Input Throughput, Output Throughput, etc.)
- **Resource Metrics** (e.g., GPU Resource Utilization, Memory Utilization, etc.)
- **Training Metrics** (e.g., LLM Training Cost, LLM Fine-tuning Cost, LLM Inferencing, etc.)

Understanding how changing parameters relating to these metrics can impact the associated cost is crucial for an organization in being able to plan and forecast the expenses around adopting and advancing AI.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Things your organization should be thinking about around the cost of AI

- How effective are we in governing, tracking, and managing our licensing and implementation of new AI investments? Is our Software Asset Management team involved in the process?
- Do we have enough transparency around our AI licensing costs and usage assumptions? What technologies are we using to help us gain more visibility into AI spending?
- Who is responsible for tracking, managing, and reporting cloud and AI costs?
- How are we forecasting AI budget for upcoming years? How accurate were our recent forecasts in previous years?
- What mechanisms do we have in place to prevent, identify, and monitor potential legal risks associated with AI?
- How effective are we at monitoring and managing AI data security and governance to prevent leakage?
- What is the state of our data and analytics capabilities to fuel our AI capabilities?
- Do we have a framework within which to assess whether our AI investments are achieving desired outcomes and realizing anticipated value?

The market around AI technologies has seen a number of important cooperations and acquisitions in the past two years. CIOs are concerned that the shift in the market indicates the vendors' strategy to sell larger and more expensive software suites.

Barry Brunsman of KPMG says *"Technology vendors are the same as every other sort of business. They are interested in finding ways to expand their share of wallet or their mind share with their customers."* **See "Tech M&A Raises Fear Over Software Pricing, Bundling for CIOs"**

The evolving landscape necessitates a proactive approach in monitoring a company's software spend and licensing agreements. monitoring their technology spend and the terms of their licensing agreements. KPMG Software Asset Management experts are continuously working on providing our clients with the latest technologies and processes to more effectively manage and monitor software licensing spend.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Call on KPMG

The time to capitalize on the promise of AI, while gaining control and managing its costs is NOW. To be successful and gain true competitive advantage, your company must recognize the full impact of AI implementation, understand the hidden costs, and adapt your processes accordingly. But you do not have to go it alone. KPMG offers expertise in a wide range of services from FinOps and beyond—data analytics and governance, provider licensing and asset management, data security, legal and risk management, training and resource allocation, architecture and infrastructure design, operating model enhancement, and enterprise change management—all of these can help you optimize your technology implementations, AI among them. Let us help you ensure that your AI expenses stay manageable today and into the future and set the foundation for a value- and cost-driven approach to your AI and technology investments. organization can muster.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Contact

Len Epelbaum
Managing Director,
Technology Strategy
KPMG LLP

773-255-3455

lepelbaum@kpmg.com

Swami Chandrasekaran
Principal, Advisory AI
KPMG LLC

972-740-8799

swamchan@kpmg.com

Dominik Fiebig
Manager, Technology
Strategy
KPMG LLP

831-824-6124

dfiebig1@kpmg.com

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Learn about us in: | **kpmg.com**

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the situation.

© 2024 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved. MGT-9286-03-25

The KPMG name and logo are trademarks used under license by the independent member firms of the KPMG global organization.