Ignite the Future of AI: Transforming Industry through AI Integration and Best Practices

KPMG. Make the Difference.

kpmg.com/cn

July 2026

Preface

Victor Hou
Managing Director and CTO of
Cisco Greater China

Connectivity and Protection – Building Resilient Foundations

Artificial intelligence (AI) is being taken up at an unprecedented pace–AI agents coalescing into a digital workforce; robotics transitioning from experimental prototypes to production lines; intelligent vehicles emerging as the next generation of mobile computing platforms… This proliferation of Agentic AI is acting as a catalyst for growth in upper-layer applications. Enterprises increasingly recognise that AI is no longer merely a generic tool unattached to core operations, but an essential productivity component deeply embedded within the operational core. This raises a more fundamental question: where must we lay the foundations for this burgeoning intelligence?

The release of *Ignite the Future of AI: Transforming Industry through AI Integration and Best Practices* addresses this fundamental challenge. Building upon our 2025 publication, *AI Readiness White Paper: Enterprise Digital Transformation Pathways and Assessment Framework*, which assessed organisational AI preparedness across the dual aspects of tangible infrastructure and organisational capabilities, this new research empowers enterprises to conduct a comprehensive assessment of their preparedness across five critical dimensions: strategy, data, technology, organisation, and governance. Current observations already indicate a significant shift: businesses are increasingly moving their focus beyond foundational AI prerequisites and toward developing and executing scalable implementation strategies. Through rigorous analysis of AI's evolutionary trajectory and deployment pathways by Cisco and KPMG, this whitepaper delivers a systematic and actionable framework to tackle organisations' most pressing pain points: how to implement a strategic vision into micro-level capabilities; how to seamlessly integrate foundational infrastructure (including energy) that powers a "silicon-based workforce"; how to consistently convert technological applications into tangible commercial value.

This year's research includes a dedicated analysis of regional ecosystems, examining how AI infrastructure, industrial applications, and commercial value form a closed-loop system through regional collaboration. This focus is driven by the emergence of AI as a key engine for developing "new quality productive forces" under the 15th Five-Year Plan's strategy for high-quality development. Consequently, AI transformation initiatives ultimately find their proving ground within broader regional testbeds – requiring robust ecosystems characterised by resource aggregation, industrial chain coordination, and institutional innovation to achieve scalable implementation. The Guangdong-Hong Kong-Macau Greater Bay Area exemplifies such regions, boasting four core strengths: concentrated supply chains for hardware manufacturing, diverse industrial application scenarios, mature data circulation mechanisms, and advantages in the opening-up of institutions. Together, these strengths create an optimal environment for validating and deploying advanced AI architectures at scale.

Through rigorous research and thorough analysis, we conclude that in large-scale AI implementation, collaborative AI agents, multi-tool orchestration, and ongoing interactions between automated workflows and business systems result in AI processes that increasingly resemble the lifecycle of AI tokens – spanning generation, routing, precision scheduling, and consumption. These tokens encapsulate intelligent outputs while simultaneously quantifying resource efficiency and business value. Consequently, interconnectivity and trust assurance are foundational enterprise capabilities in the AI era: interconnectivity governs whether computing power, data, models, and applications can interoperate efficiently, while trust assurance ensures intelligent behaviours operate within verified boundaries. This insight leads us to propose reframing AI infrastructure through integrated token factories–positioning them as core engineering foundations for building organisational resilience in the AI epoch.

In this report, we outline a systematic pathway for transforming infrastructure into integrated token factories, supported by actionable frameworks based on global best practices. We aim to take our insights and expertise and collaborate with enterprise leaders, technical architects, and ecosystem partners in a way that strengthens connectivity, secures digital perimeters, builds resilience, and advances toward a new era of deeply integrated AI and business operations.

Richard Zhang
Head of Technology Advisory,
KPMG China

The Silicon Era – Reshaping Growth Dynamics

By 2026, China will be entering the initial phase of its 15th Five-Year Plan period. At the same time, artificial intelligence is evolving ever faster, from a technological innovation into a key driver of industrial transformation and economic growth. It is this profound shift that is defining the emerging AI era: Large Language Model (LLM) inference capabilities are maturing into deployable engineering solutions; ever-growing ecosystems containing AI agents and skillsets are beginning to transform software delivery and business operations; and AI infrastructure itself is evolving from a traditional IT resource into an industrial-grade intelligent production platform. Simultaneously, supportive macroeconomic policies, industry capital, and regional innovation ecosystems are accelerating the integration of AI into high-value sectors such as the real economy, financial services, and public governance. As these forces come together, enterprises must redefine their understanding of the fundamental drivers that underpin sustainable growth.

For many years, KPMG has partnered with enterprises in identifying practical transformation challenges, value opportunities, and implementation pathways. During our long-term collaboration with Cisco, we jointly researched AI readiness, as an AI-ready status fundamentally determines an organisation's ability to strategically integrate AI scenario value, data infrastructure, process mechanisms, governance frameworks, and core capabilities. Today, fully embracing Agentic AI has become a strategic imperative for businesses in navigating existing competition and reshaping growth engines. Consequently, our research now focuses on embedding these technologies into core business operations – transforming AI into quantifiable, governable, and sustainable growth drivers.

In *Ignite the Future of AI: Transforming Industry through AI Integration and Best Practices*, we first apply a three-way analytical framework covering technological evolution, industry development, and regional dynamics. Our analysis indicates that once macroeconomic conditions and foundational technology capabilities reach maturity, industrial transformation efforts will inevitably shift from generic infrastructure development to highly complex, heterogeneous vertical sectors. We then explore industrial-AI integration, examining how AI is transforming mission-critical industries and core business scenarios. Building upon this foundation, Chapter 3 systematically outlines transformation pathways and essential actions for enterprise AI adoption, to help executives prioritise key initiatives across six domains: strategic decision-making, scenario implementation, data governance, organisational transformation, innovation of applications, and infrastructure enhancement.

To pioneer is to lead; to innovate is to succeed. To thrive in the silicon-driven era, enterprises must strategically define AI transformation goals and boundaries, target high-value scenarios, embed human-AI collaboration into operational workflows, activate the latent value of intelligent assets using data and knowledge systems, and, fundamentally, reconfigure operational models by restructuring organisations and talent pools. Achieving these depends entirely on end-to-end infrastructure capabilities. Building systems optimised for massive token throughput and real-time inference is the foundation for enterprise AI deployment at scale.

KPMG's joint research with Cisco integrates executive vision alongside foundational architecture perspectives to identify value opportunities, implementation barriers, and organisational constraints in AI transformation. This dual focus clarifies the core capabilities required for AI operationalisation at scale. We wrote this whitepaper to help enterprises cut through technological hype to capture tangible value and establish systematic pathways through complex transitions. Looking ahead, KPMG will continue collaborating with clients and ecosystem partners to deepen the integration of AI into industrial scenarios and usher in a new chapter of sustainable competitive advantages.

Contents

Chapter 3

Chapter 4

Chapter 1

Insights into AI-Driven Industrial Transformation

2026 has ushered in a new phase of transition from digitisation to intelligence. The official launch of the national 15th Five-Year Plan signifies that the development of AI now extends beyond mere algorithmic iteration. It has evolved into a systemic force driving macroeconomic growth, reshaping total factor productivity, and optimising industrial structures. Within this new cycle, the definitive breakthroughs in technology supply, the macroeconomic manifestation of new quality productive forces, and the proactive coordination of regional policies collectively shape the macro landscape of enterprise intelligent integration. The fundamental driver underpinning this macro environment stems from the “US$1 trillion compute tsunami” concept introduced at NVIDIA’s GTC 2026 conference: as global computational demand increases exponentially, intelligence transcends technological iteration to become an industrial-scale infrastructure transformation. Adopting a top-down strategic perspective, this chapter presents a panoramic view of industrial transformation logic driven by the intelligence wave. It examines three key dimensions: the endogenous technological evolution trends within Agentic AI, structural opportunities within macroeconomic cycles, and the exemplar role of institutional openness in the Guangdong-Hong Kong-Macau Greater Bay Area (GBA).

1.1

Technological evolution: A paradigm shift from generative assistance to deterministic business workflows

2026 marked the advent of a new stage in enterprise-grade AI applications, transitioning from probabilistic auxiliary tools to deterministic business execution entities. Whereas the preceding stage was characterised by an explosion of generative content, the defining feature of this new cycle is the pervasive integration of Agentic AI into core business scenarios. This represents a fundamental shift: AI is evolving from a basic information assistant into accountable agents directly responsible for business outcomes and delivery quality. Concurrently, AI is undergoing a profound evolution from System 1 to System 2 capabilities. This transition extends beyond mere linear scaling of model parameters, advancing instead towards greater cognitive architecture complexity, standardisation within skill ecosystems, and infrastructure industrialisation.

1.1.1 Cognitive architecture reconstruction: Chain-of-Thought and inferential economy

Early 2026 witnessed the establishment of deep thinking as standard enterprise-grade AI functionality, catalysed by next-generation flagship models including Claude Opus, GLM-5, Qwen3.5-397B-A17B, Gemini 3 Pro, and GPT-5.3-Codex.

Critical engineering breakthroughs in endogenous Chain-of-Thought: Contemporary Large Language Models (LLMs) now fully internalise logical inference processes, transitioning the underlying computational paradigm from single-sequence prediction to complex state search. Within this evolutionary leap, inference mechanisms have become deeply integrated with large-scale reinforcement learning. Models autonomously generate implicit reasoning tokens prior to output, effectively embedding complex task decomposition and logical construction into fundamental operational behaviour. The resultant architecture establishes dynamic optimisation and error-tracing capabilities, enabling parallel construction of multiple logical branches with prospective evaluation mechanisms. Upon detecting business compliance conflicts or logical gaps, models automatically initiate backtracking procedures, dynamically reallocating computational resources to high-confidence solution pathways. This multi-step deductive framework, reinforced by rigorous self-correction, elevates business output accuracy and achieves substantial improvements in determinacy. Collectively, these engineering advances systematically mitigate hallucination risks, providing core safeguards for AI agent deployment in zero fault-tolerance scenarios. Consequently, future competitive differentiation extends beyond model parameters to encompass systems' pre-execution capacity for deep logical competition and autonomous error correction.

1.1.2 Transforming the atomic unit of software: Fostering Skills and novel operating models

The proliferation of Agentic AI is driving a fundamental restructuring of software delivery and consumption models – shifting from SaaS applications towards AI agent Skills. Consequently, enterprise digitisation now prioritises developing flexible, interoperable skill sets as core strategic assets.

Skills have emerged as critical enterprise digital assets: Aligned with the industry's cutting-edge Model Context Protocol (MCP) architecture, Skills transcend conventional API interfaces through standardised encapsulation of semantic descriptions, data schemas, and execution logic. This framework resolves how agents interact with operational environments. Consider enterprise supply chain management: an MCP-compliant supply chain query skill eliminates traditional standalone application development. As Skills usually integrate natural language database schema descriptions with secure execution code, any authorised model can load them in real time to gain expert-level query capabilities.

Skill engineering supersedes prompt engineering: Enterprise IT research priorities are realigning accordingly. Unlike unstable prompt tuning, skill engineering establishes robust agent action boundaries through rigorously defined skill specifications and deterministic tool-code. IT departments now construct private skill repositories to manage version control, conduct permission audits, and ensure compliance certification. This Skills-as-a-Service paradigm significantly reduces internal tool development costs and reuse barriers, enabling agile architectures with small front-offices (agents) and a big middle-office (skill hub).

1.1.3 Reconstructing the physical layer: Advanced perceptual infrastructure and achieving full-stack observability

To support the high-concurrency, long-chain, and deeply integrated agent systems referenced earlier, enterprise data centres are evolving from **computational containers** into **perceptual AI factories**. Reflecting the industry consensus established at both GTC 2026 and the Cisco AI Summit 2026, data centres now transcend static IT asset status to function as core production facilities for intelligence as a new quality productive force. Within this paradigm, networks undergo fundamental redefinition as the "neural systems" of AI factories. Through AI-Ready Infrastructure, these neural systems enable end-to-end real-time perception and self-healing capabilities – spanning from physical-layer signals to business-level intelligence.

Machine data-driven perception loop: AI agents require capabilities beyond processing documents and data; they must also continuously monitor the physical state of the IT infrastructure. Foundational datasets for AI systems comprise vast volumes of machine data, including server logs, network telemetry, API call chains, and environmental sensor readings. By implementing an intelligent data weaving layer, enterprises can transform these underlying binary signals into business contexts comprehensible to AI. Consequently, infrastructure transcends its traditional role as a passive traffic conduit. Instead, it actively generates natural language alerts, complete with root cause analysis, for security agents. This capability enables the triggering of automated self-healing processes within milliseconds of detecting network jitter or anomalous access attempts.

High-speed optical interconnect and optical module upgrades: As AI workloads evolve from single-point model invocation towards distributed training, real-time inference, and multi-step task execution, networks increasingly become bottlenecks ahead of compute resources. This shift underscores the critical importance of high-speed optical modules (such as 800G and 1.6T) and novel optical interconnect solutions like Linear Drive Pluggable Optics (LPO). Their core value lies in minimising communication latency, enhancing cluster collaboration efficiency, and supporting both the training/inference phases and the stable operation of multi-agent collaboration within larger-scale, highly scalable infrastructure.

Reconfiguration of liquid cooling and thermal management: The evolution towards 400G, 800G, and higher-rate switching platforms has driven a continuous increase in thermal load for switch ASICs, optical interconnect components, and high-density cabinets. Traditional air-cooled architectures for network devices are approaching the efficiency limits. Novel thermal management solutions, typified by direct-to-chip liquid cooling, now deliver cooling capacity precisely to switch silicon and high-power components. These systems integrate with data centres' thermal management protocols, risk monitoring frameworks, and operational management platforms to facilitate comprehensive upgrades. This technology ensures stable operating conditions for high-density AI networks and sustained inference workloads. Critically, it prevents thermal dissipation, energy consumption, and availability bottlenecks during enterprise bandwidth upgrades and cluster scaling initiatives.

Redefining full-stack observability: As agentic AI systems advance from single-turn chat to multi-step execution, the boundaries of enterprise observability must extend beyond traditional IT metrics to encompass the intelligent processes themselves. For organisations prioritising business continuity, comprehensive observability must span the underlying hardware along with the network, model, and application layers. This allows for a causal mapping between infrastructure fluctuations and business experience impacts, while simultaneously monitoring shifts in AI agent quality during intermediate stages such as task decomposition, tool invocation, context utilisation, and result generation. Enterprises cannot rely solely on binary operational indicators like 'success/failure', 'fast/slow', or 'on/off'. They must additionally implement continuous evaluation to detect whether agent execution deviates from business objectives, generates invalid reasoning, or triggers hazardous pathways. Only with a unified view of the operational status, process quality, and business outcomes can IT teams transition from reactive troubleshooting to proactive monitoring, root cause analysis, and continuous optimisation. This holistic approach is essential for ensuring both the output quality and economic viability of AI-driven operations.

Inference economics and state-based storage architecture: With AI workloads shifting towards inference, unit token cost and task response efficiency have become critical considerations. During extended operational cycles involving multi-turn dialogues, memory recall, and skill invocation, agentic AI systems persistently generate and reuse vast quantities of intermediate states. This evolution transitions inference systems from stateless question-answer modes towards cacheable, reusable, and schedulable state-based operations. Consequently, enterprise infrastructure must now support not only models and knowledge bases, but also enable efficient data flow - including KV caches, task contexts, and execution states - across distributed computation nodes. Architectural innovations centred on pre-filling/decoding separation, cross-node cache reuse, and hierarchical resource pooling will critically affect inference latency, GPU utilisation, and per-token costs. These advancements are fundamental for enhancing sustained operational efficiency of AI factories.

1.1.4 Digital twins: From visual mirroring to digital native systems

As AI extends from virtual to physical domains, its technological evolution accelerates and increasingly converges with that of digital twin, which is poised to serve three critical functions: providing a structured framework for AI to comprehend real-world systems, offering a validation environment for testing action strategies, and acting as a coordination hub for implementing seamless feedback loops. The integration of these technologies will fundamentally transform complex system operations, enabling systems to not only "visualise states", but also perform dynamic simulation, intelligent decision-making, and closed-loop optimisation, and ultimately become digital-native.

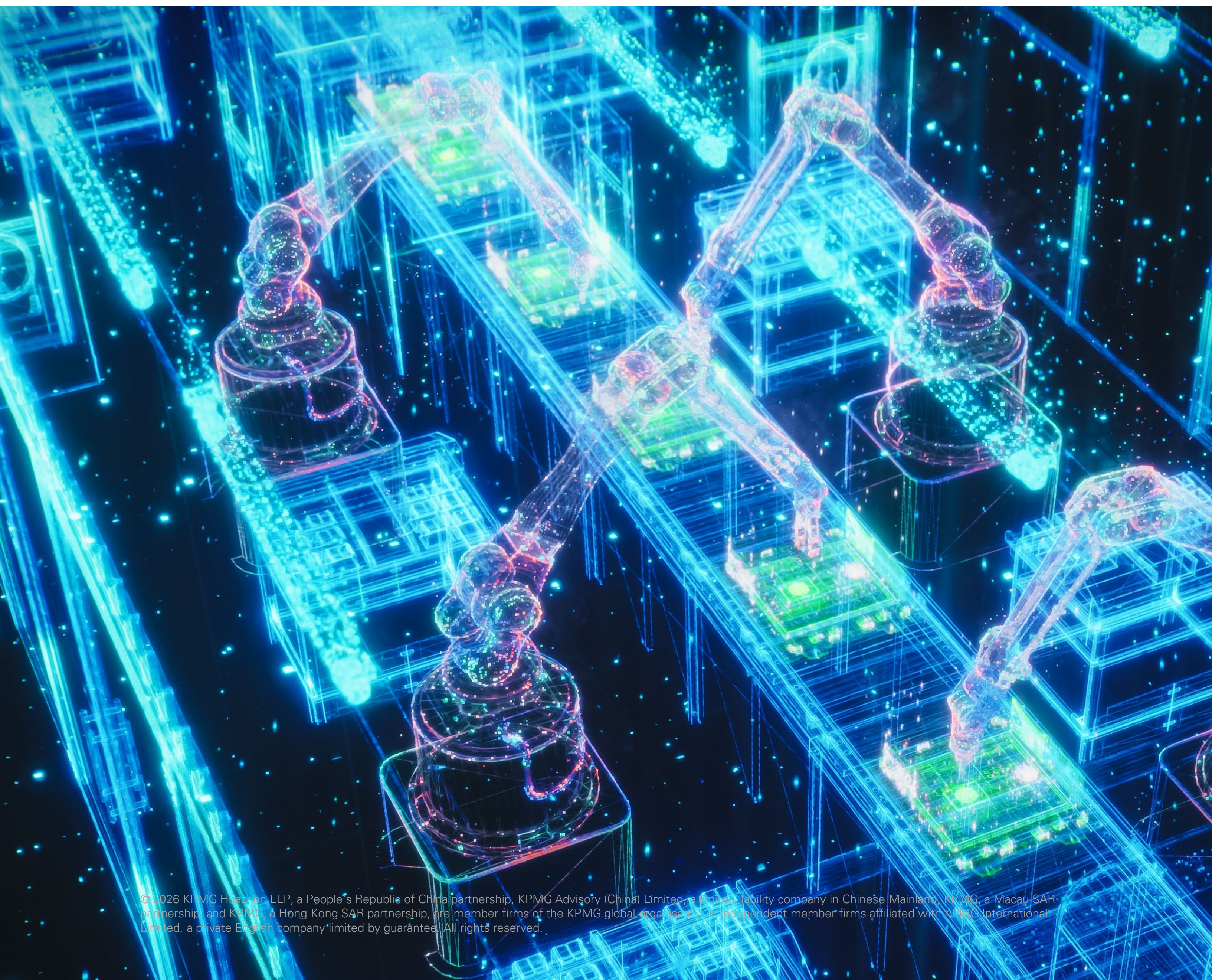
Digital intelligence-driven dynamic prediction and intelligent decision-making are transforming digital twins from 'static mapping' towards 'dynamic simulation'. As AI technology progresses beyond perceptual recognition and predictive analysis into generative reasoning, intelligent planning, and autonomous execution, AI systems can leverage real-time data gathered by digital twins to conduct predictive analysis and decision-making. This evolution brings virtual models closer to replicating the operational rules of real-world systems, thereby endowing them with the four predictive capabilities: predictive monitoring, early warning systems, simulation-based rehearsal, and contingency planning. This achieves a decision-making upgrade, effectively substituting computational power for human resources.

A novel paradigm is emerging for the intelligent transformation of existing infrastructure amidst urban renewal and industrial transformation, with digital twins accelerating their integration into legacy assets. By constructing digital twins of infrastructure elements such as rail networks, road systems, and utility corridors, the management model shifts from 'reactive repair' to 'predictive maintenance'. This transition significantly enhances the operational resilience and safety standards of physical systems.

Digital twin engineering for whole-lifecycle management represents a significant advancement, targeting large-scale, highly complex system-level applications. By integrating digital twins with systems engineering methodologies, this approach constructs 'digital twin engineering' frameworks encompassing the entire process from research and development through production to operations and maintenance. For complex sectors such as manufacturing, energy, transportation, and urban governance, digital twin engineering will emerge as a critical technological foundation, seamlessly connecting R&D, production, operations, and governance functions. For individual organisations, its value lies in bridging the collaborative relationships between assets, processes, data, and organisational structures. This integration addresses information gaps and collaborative inefficiencies inherent in whole-lifecycle management, while supporting cross-departmental, cross-hierarchical, and cross-system data interoperability.

The evolution from virtual-real mapping towards digital-native systems signifies a fundamental shift driven by the deep integration of AI and digital twins. This convergence propels complex systems towards inherent digital capabilities. Real-world systems continuously generate operational data, which digital twins transform into structured digital mapping and simulation. AI then leverages this data and the underlying models to perform understanding, prediction, and optimisation. Validated AI-derived strategies are then fed back to the physical systems. This closed-loop process ultimately enables complex systems to achieve continuous self-learning, self-optimisation, and self-adaptive operation.

By 2026, technological transformation will have progressed beyond incremental upgrades to constitute a systematic reconfiguration of productivity's organisational form. Within the software ecosystem, Agent Skills are emerging as fundamental capability units, redefining the delivery boundaries and collaborative patterns of digital functions. At the physical infrastructure level, AI Factories are becoming novel computational platforms, ensuring deep environmental perception and precise execution fidelity. This convergence enables the accelerated evolution of technologies like embodied intelligence and digital twin, which further catalyse the integration of software capabilities with hardware systems while enabling bidirectional mapping between virtual and physical domains. This paradigm shift signifies a new era of enterprise digital transformation, transitioning from connection-driven to autonomy-driven operations. For forward-thinking enterprises navigating cyclical challenges, constructing an intelligent production system that integrates hardware and software, blends virtual and physical elements, and maintains continuous evolutionary capacity represents a strategic imperative for establishing sustainable competitive advantages over the coming five-year horizon.



1.2 Industrial development: Aligned with the macroeconomic cycle and creation of new productive forces

As China commences its 15th Five-Year Plan period, macroeconomic development is shifting from factor-driven to innovation-driven. Within this new paradigm, AI has surpassed its role as a mere efficiency tool within individual sectors. It is now evolving into a systemic force that reshapes total factor productivity, restructures cross-industry value chains, and fundamentally redefines organisational dynamics within enterprises.

1.2.1 Macroeconomic analysis: Total factor productivity and capital paradigm transition

Under the top-level design framework of the 15th Five-Year Plan, the deeper integration of the digital economy with the real economy has been defined as a critical pathway for achieving high-quality development. Confronted with the macroeconomic constraints stemming from diminishing marginal returns on traditional capital and labour factors, AI technology is unlocking fresh growth dividends. This is achieved by reshaping production functions and fundamentally altering the logic of capital allocation.

The intelligent multiplier effect on total factor productivity stems from the inherent characteristics of data and algorithms: non-rivalry and near-zero marginal cost. By 2026, as Agentic AI permeates core enterprise business processes, AI's contribution to the macroeconomy will have shifted from linear assistance to exponential empowerment. On the supply side, AI directly reduces carbon emissions and energy costs per unit of GDP through the global optimisation of supply chain scheduling and energy management. On the demand side, highly personalised intelligent services generate novel consumption scenarios and create new value. This qualitative leap in efficiency, driven by increased technological penetration, constitutes the core manifestation of new quality productive forces at the macroeconomic level.

Computational infrastructure is becoming a public service and attracting significant capital. Regional governments are actively implementing policies such as AI and robotics industry funds while collaborating to establish next-generation computing infrastructure, thereby forging regional competitive advantages in industry. Concurrently, global capital is accelerating deployment, with large-scale intelligent computing centres evolving into essential public utilities analogous to electricity and water networks. In 2026, the valuation logic for AI in capital markets will be undergoing a rational refinement, pivoting focus from technological vision to commercial viability. This necessitates companies demonstrating sustainable margins post-inference costs for their products.

Token economics has emerged as a critical analytical lens connecting technological supply, industrial demand, and capital returns. At the 2026 GTC Conference, Jensen Huang introduced the concept of "Token Economics", along with a 5-tier token pricing model. It signals a strategic shift in AI infrastructure competition—from the provision of computational power towards maximising the effective intelligent output per resource unit. That same month, Liu Liehong, Director of China's National Data Administration, stated at the China Development Forum Annual Meeting that in the intelligent era tokens serve not only as value anchors but also as the fundamental "settlement unit", bridging technological supply and commercial demand and enabling quantifiable business models. This reflects the accelerating evolution of value systems centred on token invocation, distribution, and settlement—demonstrated by the widespread adoption of pricing in mainstream model services for input tokens, output tokens, and cache tokens. Consequently, enterprises are increasingly evaluating the performance of AI applications through comprehensive metrics that include token consumption, invocation frequency, latency, throughput, error rates, and operational expenditure.

1.2.2 Pan-industry chain analysis: From linear transmission to cognitive network reconstruction

The impact of AI technology has transcended the technological domain itself, driving fundamental reconfiguration of operational paradigms across automotive, consumer goods, financial services, and energy sectors.Meanwhile, Linear transmission is being replaced by cognitive network construction in shaping cross-industry value chains.

The physical expansion of the real economy: LLM is endowing industrial chains with global cognitive capacity. Through real-time analysis of worldwide shipping data, raw material price fluctuations, and end-consumer sentiment, AI functions beyond individual enterprise boundaries to enable dynamic cross-supply-chain collaboration. Concurrently, the global humanoid robotics market demonstrates accelerated growth, with Chinese manufacturers leading the expansion. Annual shipments from top producers are projected to achieve significant breakthroughs in 2026. Selected humanoid platforms now exhibit full autonomy in executing high-precision operations within complex industrial environments. This convergence of digital intelligence capabilities with physical entities will substantially enhance the real economy's resilience against external shocks.

Industrial delivery and value chain shift of the service economy: Within the high-end service sector, AI is liberating businesses from the traditional constraint of non-replicable, highly skilled human expertise. By encapsulating specialist knowledge into standardised capabilities, the service industry has, for the first time, attained the capacity for replication on a scale comparable to industrial manufacturing. Significant financing rounds across the sector indicate that enterprises possessing core industry-specific scenario data and closed-loop operational capabilities are positioned to command premium pricing within their value chains. This advantage is achieved by establishing definitive industry operating standards.

The structural premium associated with compute infrastructure and resource reallocation: AI servers exhibit memory demands 8-10 times higher than traditional servers. This escalating demand for AI data centre construction is transforming memory chips from general commodities into strategically scarce resources. Major storage manufacturers, including Samsung and SK Hynix, are consequently redirecting production capacity and capital expenditure towards high-margin, high-demand products like High Bandwidth Memory (HBM) and DDR5. This strategic shift is causing a contraction in the supply of mature DRAM products such as DDR4. Compounded by factors including increased AI server storage requirements and persistently low inventory levels, this supply constraint is exerting upward pressure on general-purpose memory prices. Concurrently, Solid State Drive (SSD) price increases are being driven more significantly by tightening NAND supply, expanding storage demands from AI data centres, and long-term supply agreements securing volume. As these rising hardware costs – encompassing memory, flash storage, and GPUs – propagate downstream, certain cloud service providers have begun implementing structural price adjustments for services with high underlying hardware costs, notably AI compute and parallel file storage solutions.

1.2.3 Enterprise analysis: restructuring organisations, improving governance, and innovating operations

The large-scale deployment of AI capabilities, driven by LLM developments, constitutes a fundamental organisational transformation that is reshaping the core characteristics of enterprises. Businesses must develop comprehensive adaptation frameworks across three critical dimensions, including organisational structure, governance mechanisms, and operational paradigms.

- **Organisational restructuring for asset synergy:** By 2026, Digital Employees will be formally integrated into corporate balance sheets and human resource planning. These digital agents operate under clearly defined permissions and performance metrics, undertaking comprehensive workloads in domains such as financial auditing, customer service, and code testing. Consequently, the functional emphasis for human employees has shifted towards skills development, exception management, and ethical oversight. Corporate structures are evolving towards agile formations combining human expertise with cluster networks, while Human Resources departments have expanded their remit to encompass the full lifecycle management of digital labour.

- **Skill sovereignty through deepened governance:** As application thresholds lower, unaudited and uncontrolled AI assets – such as Skills – have emerged as a critical enterprise data security risk. Therefore, corporate governance must move from passive compliance verification to proactively establishing skill sovereignty. Creating a unified private skill repository has thus become the central governance objective, ensuring all invoked Skills undergo rigorous certification for security, version control, and permission auditing. Governance boundaries now extend beyond data privacy to encompass ethical scrutiny of generated content accuracy and algorithmic decision-making processes.

- **Operational innovation and unit economics:** Enterprise AI expenditure is shifting focus from costly model pre-training towards sustained post-training and inference applications. This transformation necessitates establishing an AI FinOps system to meticulously manage asset input-output ratios. Consequently, the core operational metric has transitioned from model performance to a unit economics framework, precisely calculating the inference cost and business value of each agent task. Through strategic implementation of heterogeneous compute combinations and tiered storage solutions, enterprises can maintain operating costs within commercially sustainable parameters while guaranteeing service level agreement (SLA) compliance.

The industrial transformation anticipated by 2026 simultaneously exhibits the characteristics of macroeconomic advancement and micro-level operational refinement. At the macro scale, AI is fundamentally restructuring total factor productivity dynamics and industrial value chain logic. Concurrently, at the enterprise level, organisations are implementing digital employees through structural reorganisation, establishing data sovereignty through enhanced governance frameworks, and optimising unit economics through operational innovation. This dual evolution – integrating top-down systemic change with bottom-up enterprise adaptation – marks the transition of industrial intelligence transformation from conceptual validation into tangible value realisation.

1.3

Regional development: Institutional opening-up and leading industrialisation in the GBA

The broader macro landscape of LLMs and industrial-level infrastructure ultimately requires regional ecosystems with comprehensive factor aggregation capabilities, in order to achieve commercial viability. Currently, the GBA leverages its distinctive geographical advantages and industrial depth to accelerate the development of a globally leading AI innovation hub. According to the Global Innovation Index published by the World Intellectual Property Organisation, the Shenzhen-Hong Kong-Guangzhou innovation cluster continues to strengthen its position as one of the world's premier innovation centres. The GBA possesses not only a complete hardware manufacturing supply chain and extensive application scenarios but also establishes leading benchmark standards in cross-border data flows and institutional openness. This dual advantage of institutional frameworks and physical infrastructure creates an optimal testing environment for the large-scale deployment of advanced AI architectures.

1.3.1 Cross-border data collaboration and interconnected heterogeneous computing power

The GBA is transcending the geographical constraints of traditional administrative boundaries to establish an inter-regional network for compute coordination and core data resource allocation. The seamless integration of data and computing resources constitutes the basis for ensuring the operational integrity and efficiency of enterprise-grade AI agents.

Interconnection of geographically dispersed heterogeneous computing infrastructure: The GBA is advancing its regional compute collaborative scheduling mechanism. In Hong Kong (SAR), China, the Cyberport Artificial Intelligence Supercomputing Centre has announced it will provide 3,000 PFLOPS of computing capacity, accessible to sector-specific industries. This will significantly enhance the capability of local universities and commercial banks to support high-intensity intelligent computing operations. On the Chinese Mainland, leveraging the hub nodes of the Qianhai (Shenzhen), Nansha (Guangzhou), and Shaoguan data centre clusters, a collaborative framework has been established between the national compute internet service node and the China Computing Power Network Guangdong-Hong Kong-Macau Greater Bay Area Dispatch Centre. This deeply interconnected ecosystem of heterogeneous computing resources enables enterprises to dynamically allocate optimal computing capacity according to task complexity. Consequently, this significantly reduces per-unit operational costs for models during both local deployment and daily inference operations.

Institutional breakthroughs in data special zone development: To facilitate data element circulation, relevant governmental and regulatory bodies have systematically advanced the establishment of Data Special Zones within the GBA. This initiative explores the creation of bidirectional cross-border data flow mechanisms and trusted data exchange frameworks across Hengqin, Qianhai (Shenzhen), and Nansha (Guangzhou). Such developments are anticipated to enable multinational enterprises and financial institutions to leverage compliant multi-jurisdictional data effectively, thereby reducing regulatory approval costs for developing sector-specific digital finance models. Furthermore, the implementation of standardised whitelisting protocols and trusted data compliance infrastructure empowers organisations to achieve real-time monitoring and closed-loop management of cross-functional business data while maintaining strict data sovereignty safeguards.

1.3.2 Institutional opening-up and agile governance system

In the face of the rapid evolution of AI technologies, the GBA has demonstrated remarkable policy agility. Regional governance frameworks are evolving beyond passive compliance review towards proactive empowerment and strategic oversight.

Democratising compute access & guiding R&D paradigms: To lower AI adoption barriers for the real economy, Guangdong province has institutionalised innovative policy instruments such as compute vouchers through standardised mechanisms. These financial subsidies directly reduce intelligent computing centre usage costs for SMEs, facilitating model refinement and business transformation. Concurrently, the Hong Kong SAR government has strengthened institutional support for AI research institutes, enhancing regional capabilities in foundational algorithm innovation and security evaluation at the strategic level. This dual approach ensures regional AI development aligns with rigorous international compliance requirements.

The large-scale provision of AI capabilities across public services and government service scenarios: Regional administrative authorities function not only as regulators but also as leading adopters of AI technology. The Hong Kong SAR government has established explicit targets for AI deployment: the implementation of AI tools across 100 government processes by 2026. By the end of 2027, this will be further expanded to encompass no fewer than 200 processes, alongside the introduction of flagship AI projects designed to expedite the processing of diverse public service applications. This initiative of opening core government operational processes to AI Agents delivers a rigorous zero-fault-tolerance testing environment for foundational AI models. Simultaneously, it establishes operational standards for digital labour within the public service domain, thereby providing a robust reference paradigm for the large-scale implementation of AI in highly regulated sectors such as digital finance and healthcare.

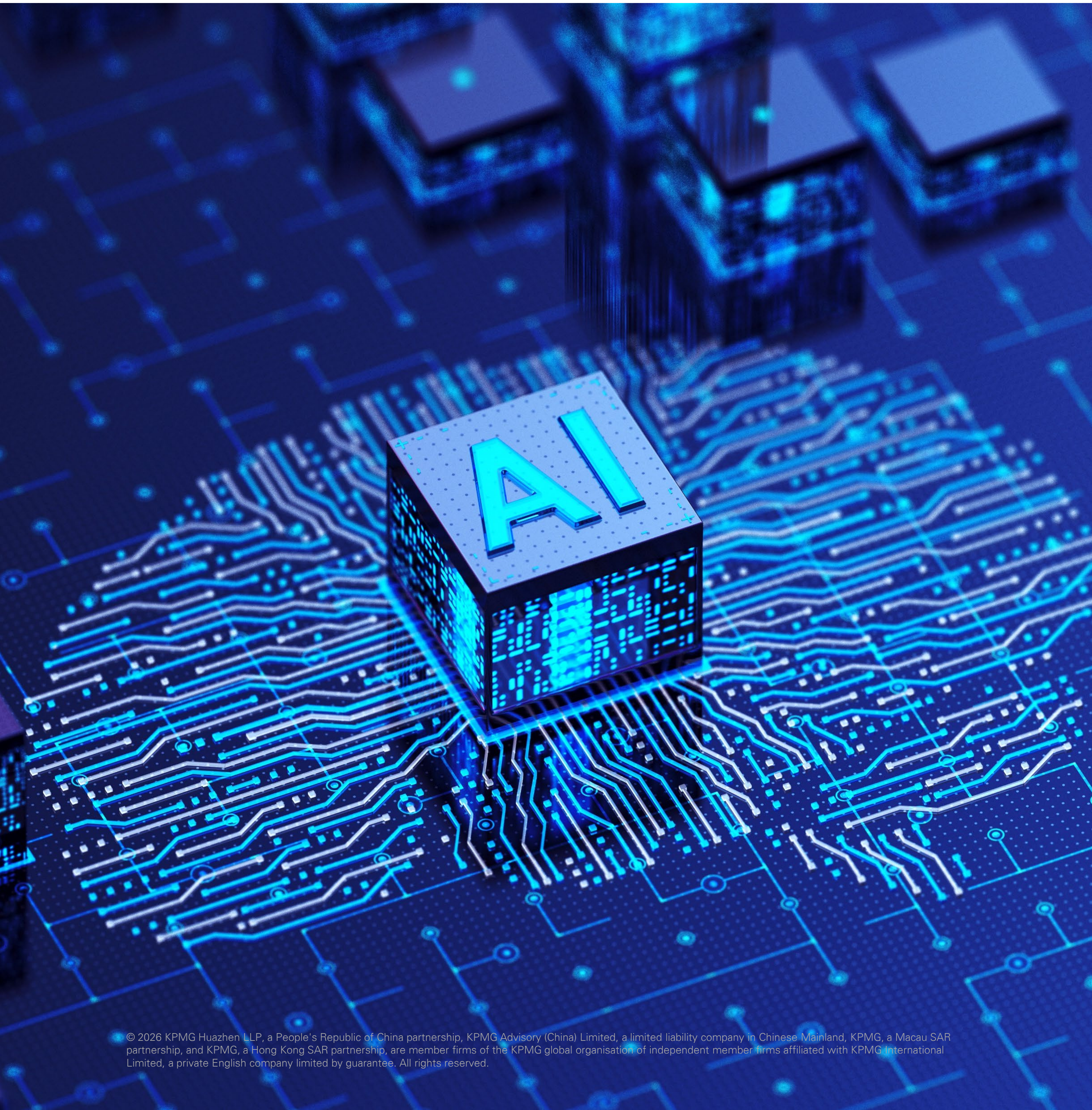
1.3.3 Total factor productivity aggregation and core scenario commercialisation

The GBA's distinct competitive advantage resides in its accelerated pathway for transforming foundational computing capacity into tangible economic output. The region has cultivated a comprehensive ecosystem – spanning research and development of core semiconductor technologies, through to terminal hardware manufacturing, and culminating in high-value vertical applications – thereby establishing a seamless operational continuum.

Large-scale industrial AI enablement: In the high-end manufacturing sector, the GBA is comprehensively advancing its initiative to harness AI for enhancing manufacturing quality. By directing regional industries towards reinforcing dominant supply chains – including electronic information, intelligent vehicles, and robotics – the GBA prioritises cultivating specialised vertical domain models and high-value industrial AI agents with sector-leading impact. This strategy bridges the gap between general-purpose software capabilities and industrial on-site operational functionalities. Through driving digital supply chain transformation and establishing national pilot cities for SME digitalisation, the GBA has shifted AI implementation from the software application layer towards adaptive production lines, intelligent equipment, and production control systems. This helps the region's advanced manufacturing enterprises develop competitive advantages rooted in total factor productivity gains amid global supply chain restructuring.

Integrated digital finance ecosystem and Industry-Academia-Research synergy: Within the modern services sector, Hong Kong as an international financial centre closely partners Qianhai, an institutional innovation area in China. Leveraging the aforementioned interconnected computing power infrastructure and data zones, the region's premier commercial banks and securities advisory institutions are pioneering the integration of expert knowledge, business processes, and AI agent capabilities. This is transforming digital financial planning expertise into standardised AI agent competencies compliant with Multi-layer Compliance Protocol (MCP) regulations. Meanwhile, enhanced business models are revealing strong complementarity between world-leading university research capabilities in Hong Kong and industrialisation capacities across Chinese Mainland. This tripartite collaboration between industry, academia, and research institutions is substantially narrowing the technological gap between theoretical advancement and commercial deployment. The GBA continues to attract a significant influx of globally-minded, multidisciplinary professionals spanning technology and business management disciplines. This talent convergence is furnishing substantial strategic capacity for the region's engagement in global AI sector competition.

The 2026 regional transformation demonstrates a compound effect arising from institutional opening-up and total factor agglomeration. In terms of foundational infrastructure, cross-regional integration of heterogeneous computational capacity and trusted cross-border data flows have fundamentally dismantled physical and administrative barriers. In terms of governance, agile regulatory sandboxes and standardised computational capacity democratization policies have established a proactively enabling innovation environment. Regarding commercial implementation, the extension of intelligent high-end manufacturing and the evolution of digital finance capabilities have achieved high-value scenario commercialisation. This profound synergy across computational resources, institutional frameworks, and application scenarios signifies the GBA's transition from an early-stage technology validation zone to a globally pre-eminent AI innovation entity. For enterprises operating within the region, strategic integration into this first-mover ecosystem that integrates hardware and software is critical for converting the region's policy advantages into enduring competitive advantages within global industrial value chains.



1.4

Strategic convergence: Converting macroeconomic insights into industry-wide applications

A systemic leap in the technological foundation, the restructuring of macroeconomic value chains and the emergence of a leading exemplar ecosystem within the GBA are collectively driving strategic intelligent transformation of industries. AI is no longer a peripheral business tool; it has become a core productivity factor deeply embedded within the core operations of enterprises. With both the macro-environment and the underlying technology supply now matured, the focus of industrial transformation will inevitably shift from general infrastructure development towards highly complex, heterogeneous vertical sectors.

The key focus of large-scale AI implementation has shifted towards economic validation. While the maturity of intelligent solutions and scenario penetration are foundational conditions for AI adoption across industries, they can hardly address critical questions of post-deployment cost efficiency and investment return. As AI integrates into enterprise-level production systems, unit token cost emerges as a pivotal operational efficiency KPI, though this metric cannot be simplistically regarded as model API price. A more practical enterprise accounting framework is summarised as follows: Unit Token Cost = (Model Call or Inference Compute Cost) + (Networking and Memory Cost) + (Energy and Cooling Cost) + (Platform Software and Scheduling Cost) + (O&M, Security, and Governance Cost).

Dedicate token management is directly linked to AI competitiveness. Token is being translating from a technical metric into a financial one, and further into an infrastructure management indicator. Enterprise leaders must understand the strategic significance of token as a unit of intelligent capacity measurement, while also delving into the specifics of its production, circulation, consumption, and value realisation, so as to implement robust management. Given the continuous improvements in the inference efficiency, open-source ecosystems, pricing strategies, and localised services of Chinese models, Chinese enterprises are more likely to have access to competitive models and indigenous infrastructure. This enables them to more quickly convert token management into cost and innovation advantages.

The absence of granular token accounting fundamentally obscures AI competitiveness measurement. Currently, most organisations lack clear token accounting to distinguish between the tokens for high-value scenarios, the tokens for redundant context, the tokens for costly models, and the tokens affected by network, memory, scheduling, or cooling bottlenecks. These cause the absence of objective measurement standards and data validation. Without this visibility, companies cannot determine whether AI investments can translate into tangible process efficiency, decision quality, and operational benefits, or only result in hidden costs within complex call chains.

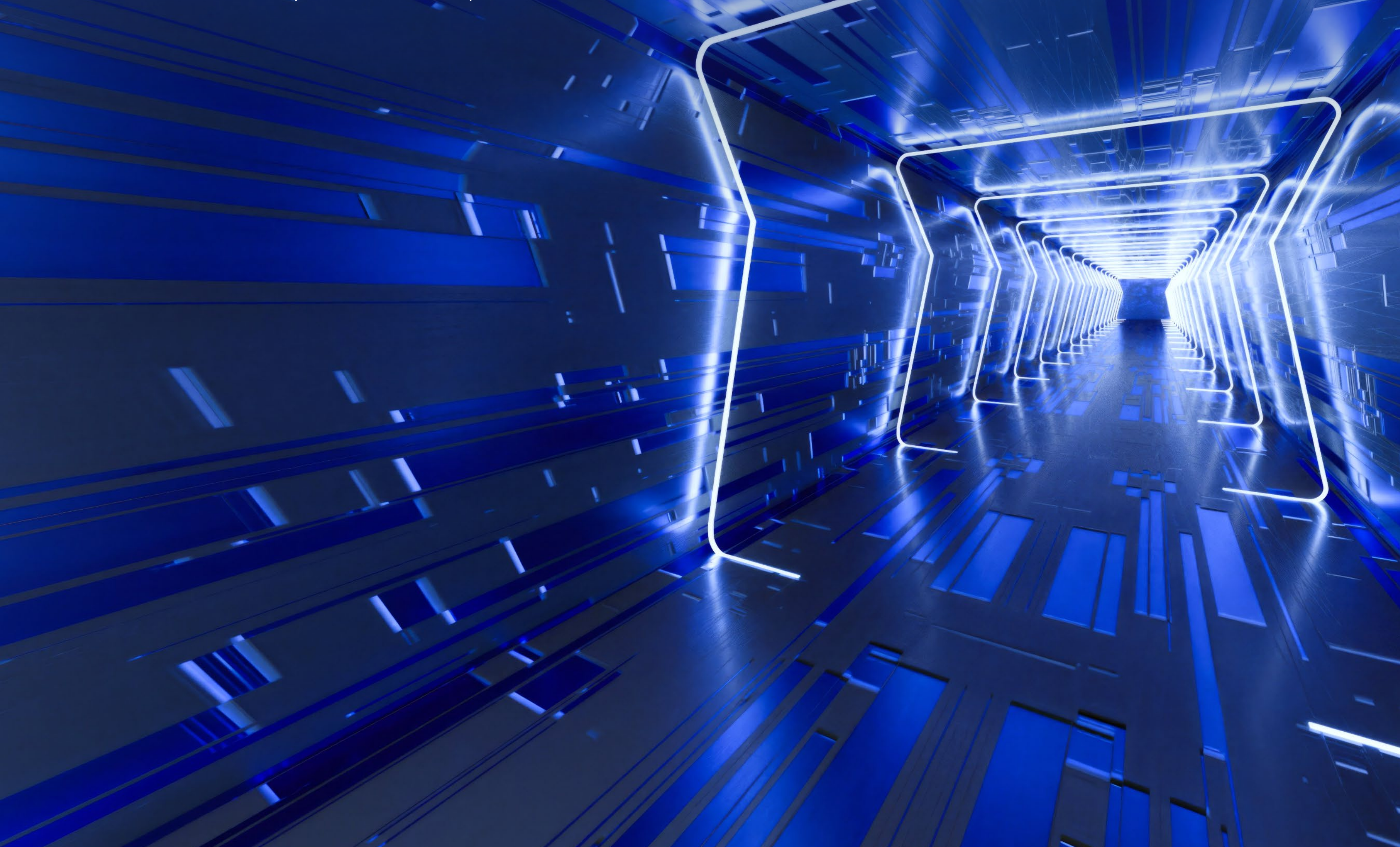
A token cost scale should be established via enterprise infrastructure decisions. The cost of each token is defined as it moves through each layer across the entire infrastructure chain. Infrastructure investments and operational efficiency collectively determine the production and delivery cost of tokens. Though routing strategies, budget controls, or inference optimisation can be used to reduce token consumption, they cannot help reveal the root cause of cost and efficiency bottlenecks. Only with a cost scale can organisations understand where token costs originate and escalate, and pinpoint the areas that require optimisation. This requires comprehensive restructuring from the management system to the foundational infrastructure.

The above analysis indicates that as sectors pursue deeper enterprise AI integration and transformation, they will face substantive challenges concerning economic viability, measurability, and value realisation. Deeper AI integration into specific industrial scenarios and business operations requires enterprises to move faster from fragmented applications and isolated scenario-based pilots to comprehensive, full-stack transformation. In this process, the token economics will provide a pivotal guidance for token factory transformation, enabling organisations to support intelligent manufacturing capacity through sustainable infrastructure; calibrate cost efficiency through transparent token accounting; and unlock the value of growth through achievable closed-loop business operations.

Chapter 2

The integration of AI across Industries in the new economic cycle

Enterprise intelligent transformation has fully shifted to the key components of the real economy and digital economy. Unlocking commercial value no longer just rely on scaling model parameters, but on sector-specific data assets, integration of intelligent systems with business scenarios, and autonomous execution capabilities of the systems.



To empower organisations across industries to seize transformation opportunities, KPMG has conducted a year-long study supported by its global network. We analysed over 500 AI implementation cases and surveyed 1,390 decision-makers, aiming to providing insights to help enterprises formulate future-proof strategies. Our key insights include:

AI maturity: Over half of the respondents from different industries have adopted AI at varying levels of sophistication. 62% of the manufacturers and 56% of the retailers have utilised AI for more than three years;

Technology adoption: AI agents are widely deployed by 63% of the healthcare organisations, 67% of the manufacturers, and 85% of the life sciences companies;

ROI: Nearly all surveyed enterprises reported measurable cost reductions and efficiency gains. 62% of the manufacturers and 60% of the energy firms have achieved over 10% return on investment;

Future investment: Over 60% of the respondents plan to increase AI spending. 63% of the technology firms, 63% of the energy companies, and 71% of the manufacturers expect to increase their budget by over 10%, and 66% of the insurers expect to increase their budget by over 20%;

Risk management: Data continues to sit at the heart of risk control, with quality, security, privacy, and siloed data cited as top concerns. Regulatory compliance is also regarded as a key challenge..

2.1 LLM's value matrix for and penetration in key sectors

The progress of AI implementation is sector-specific. The inherent characteristics of industries define the operational boundaries and commercial viability of cutting-edge technologies. When formulating intelligent strategies, corporate leadership must establish a rigorous cross-dimensional evaluation framework to prevent substantial sunk costs arising from misalignment with operational realities. Currently, industry-specific AI maturity and scenario penetration determine the maximum value of AI implementation, while enterprises' prudent ROI assessments and sustained investment commitment define the minimum.

2.1.1 Multidimensional quantitative analysis of AI application potential across industries

A systematic evaluation of AI application potential in a specific industry requires a quantitative analysis matrix comprising four dimensions, including AI maturity, scenario penetration, ROI assessment, and corporate investment commitment. This matrix reveals both the current status of digital intelligent infrastructure and the trajectory of technological assets allocation over the next 3–5 years.

- **AI maturity:** This dimension evaluates the infrastructure readiness, historical business data assets, and experience in algorithm and model deployment. An industry with a higher AI maturity means it has a better underlying data governance system and a greater volume of high-quality sector-specific corpora necessary for fine-tuning models.

- **Scenario penetration:** This dimension reveals the proportion of AI-powered core business processes and the frequency of AI usage by business personnel. A higher scenario penetration rate suggest a greater likeliness of standardised application and large-scale deployment. This means enterprises can make a profit from technology application more quickly and significantly reduce the marginal cost of business logic inference.

- **ROI assessment:** This dimension reveals the value of technologies from a financial perspective. It tracks objective comments from the industry on AI project ROI to help determine if the technology can truly help slash operational costs, improve service delivery efficiency, and directly create revenue.

- **Corporate investment commitment:** This dimension reflects decision-makers' long-term strategic commitment to and investment in intelligent transformation. If a company maintains a strong willingness to make additional investments after receiving a high initial return on investment, it indicates that the relevant application has an excellent potential to generate long-term commercial value. If it is not willing to make additional investments, it usually means that the sector is facing profound structural challenges such as stringent external data regulation, privacy compliance barriers, or underlying business system security risks.

Figure 2 **LLM application potential analysis matrix (by industry)**

Industry	Application potential	AI maturity	Scenario penetration	ROI assessment	Corporate investment commitment
Internet	Very high	Very high	High	Very high	Very high
Finance	Very high	Very high	Very high	High	High
Manufacturing	High	Medium	Medium	High	Very high
Automotive	High	High	High	Medium	High
Pharmaceutical	Medium	High	Medium	Medium	High
Government	Medium	Low	Medium	Low	High
Energy	Low	Low	Low	Low	Medium
Education	Low	Low	Low	Low	Low

Very high High Medium Low

Source: KPMG analysis

2.1.2 Characteristics of intelligent transformation in frontrunner industries

Internet: Transition from a digital-native to an AI-native sector

AI maturity: As an industry with data as its core factor of production, "data-as-capital" is the nature of the sector and it has been leading China's industrial digital transformation. Interaction through natural language are fundamentally disrupting traditional search, advertising, and recommendation models, driving comprehensive end-to-end workflow and productivity system transformation. This is steadily changing the digital-native industry to an AI-native industry.

Scenario penetration: Per CNNIC's January 2026 statistical report, China's Generative AI user base surpassed 600 million by December 2025, indicating a major leap in AI adoption by internet users. These technologies are now widely used for content creation, operations, user engagement, and product development lifecycles. The internet industry possesses extensive cloud infrastructure, making AI deployment via highly elastic public clouds is highly suitable for this industry.

ROI assessment: Computing infrastructure has entered a large-scale monetisation phase. Major cloud providers are achieving exponential growth of revenue from computing power and model API invocation, which proves the commercial value of foundational technologies. At the same time, AI startups targeting market segments are attracting significant early-stage funding, which shows the capital market's confidence in agile business models.

Corporate investment commitment: Leading enterprises are making unprecedented strategic investments in foundational technology ecosystems. IDC forecasts that China's total AI investment will reach USD 100 billion by 2028, sustaining a five-year CAGR of 35.2%.

Finance: Transforming unstructured data assets into core growth drivers

AI Maturity: China's financial sector is moving from fragmented digitisation toward systematic AI integration. It should address the internal need for driving intelligent front-middle-back office synergy, and the external challenge due to macroeconomic structural shifts and the imperative to redesign customer journeys. As a data-intensive industry, it is likely to more quickly use AI to extract value from unstructured data, thereby turning its data assets into a growth engine. Banks can generate as much as 820GB of data per million-dollar assets—twice the volume of that of the pharmaceutical and retail sectors. Chinese financial institutions currently possess data assets worth more than 100 billion.

Scenario Penetration: Since the industry deeply engages in customer-facing services and faces heightened regulatory scrutiny, it currently focuses AI adoption mainly on internal operational enhancement, covering precision client demand identification, AI-powered business decision-making, complex credit risk control model upgrade, and automated advisory services. Critical systems processing core ledger data and sensitive PII maintain exclusively private cloud deployment, while peripheral non-critical functions leverage secure hybrid cloud frameworks.

Figure 3 Generative AI use cases in financial services

Analysis and operation at the middle and back ends

More dependent on manual labour

Low efficiency

High error rate

- Data collection and sorting
- Programming assistance
  Code testing
- Auto report generation
- In-depth analysis of customer behaviour
  Product portfolio research

Risk management

Difficulty in integrating unstructured data

Slow response for real-time and concurrent transactions

Slow identification of novel risks

- Automatic credit review and approval
  Automatic compliance review
- Organising multi-modal risk data
- Credit report interpretation
- Market trend research
  Risk assessment and forecast

Marketing and customer acquisition

Slow response of human customer service

Lack of precision and personalisation

Static tags cannot capture real-time user needs

- Empowerment service advisor (Indirect customer facing)
- Smart customer service (Direct customer facing)
- Real-time insight into user intentions
- Customised customer service

Can be more effectively handled by Generative AI

Cannot be effectively handled by Generative AI

Source: Public information, KPMG analysis

Banking as an example: As a prototypical data-intensive industry, banks significantly outpace peers in both generative and Agentic AI implementation. However, under prudential regulatory frameworks, deployment remains primarily internal with cautious approaches to customer-facing applications.

Figure 4 **Banking - Generative/Agentic AI implementation scenarios**

Improvement of front-end business

- Customer service assistant (Internal)
- Virtual customer service agent
- Smart customer service (External)
- Personalised service agent
- Sales assistant
- Chat insight agent
- Smart investment advisor
- Investment advisory agent
- Transaction Q&A
- Asset pricing agent
- Marketing content generation
- AI native mobile banking
- Dynamic product pricing

Lean management in the middle-end

- Smart investment research
- Compliance and quality inspection agent
- Smart risk control
- Credit review and approval agent
- Smart due diligence
- Asset custody agent
- Smart operation and maintenance
- RPA workload reduction for basic-level staff agent
- Credit review and approval
- Customer report generation
- Cyber fraud identification

Supporting business decision-making in the back-end

- Encoding assistant
- Data insight agent
- Office assistant
- Agent Collaborative development platform
- Knowledge assistant
- Documentation assistant
- Data and analytics
- Synthesised data
- Smart internal audit

Typical Generative AI application scenarios

Across front, middle, and back-end scenarios, the core value proposition centres on efficiency enhancement. This encompasses leveraging existing information and data assets to deliver deeper insights into customer needs, enhancing business process efficiency, strengthening risk control mechanisms, and supporting informed decision-making.

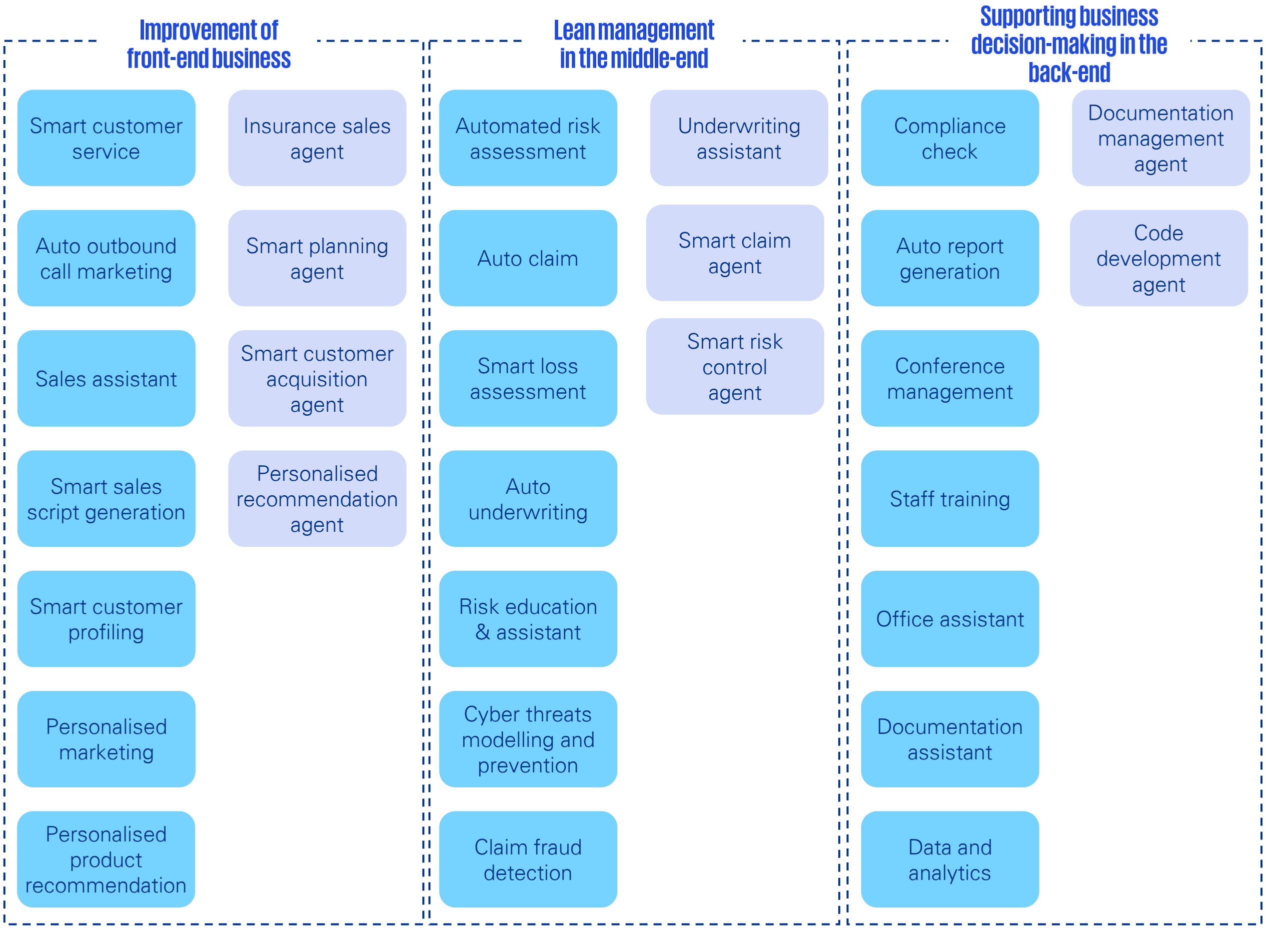
Typical Agentic AI application scenarios

Benefiting from enhanced autonomy and deeper integration within core business processes, multiple leading banks are developing dedicated internal platforms for collaborative intelligent agent development. This evolution is anticipated to catalyse widespread adoption of intelligent agent ecosystems across the banking sector.

Source: Bank websites, annual reports, KPMG analysis

Insurance as an example: As a representative data-intensive industry, insurance demonstrates prominent Generative/Agentic AI adoption, prioritising direct customer engagement through experience optimisation and service innovation.

Figure 5 **Insurance - Generative/Agentic AI implementation scenarios**



Typical Generative AI application scenarios

Across front, middle, and back-end functions, the core value proposition centres on cost reduction and efficiency enhancement. This is achieved by leveraging AI technology to optimise processes and deliver tailored services. Further, advanced data processing and intelligent modelling enhance decision support, provide deeper insights into customer needs, and strengthen risk exposure assessment.

Typical Agentic AI application scenarios

Driven by significantly enhanced autonomy and deep integration within core operational workflows, leading institutions are actively establishing frameworks for intelligent agent ecosystems. This strategic development is comprehensively advancing automated value chain management across the insurance sector.

Source: Public information, KPMG analysis

- **ROI assessment:** Financial institutions and asset managers exhibit exceptionally high confidence in ROI. Applications in banking and securities have progressed from concept validation to pilot projects in core business activities and have demonstrated clear financial benefits through process optimisation and cost reduction.
- **Corporate investment commitment:** Entering 2026, financial institutions are shifting digital finance strategies from scale expansion to ecosystem building. Leading players are embedding AI technologies across customer journey value chains, continuously exploring AI's redefined role and critical integration points amid uncertainties.

Manufacturing: Transition from process-driven to process-data convergence model

- **AI maturity:** China's manufacturing sector is accelerating its shift from traditional linear growth models toward precision management enabled by process-data convergence. Recent MIIT reports indicate continuous acceleration in digital transformation among large-scale industrial enterprises.

- **Scenario penetration:** Generative AI is rapidly penetrating R&D, production, supply chain, sales, and service operations, revitalising traditional manufacturing workflows. Notably, due to fragmented manufacturing scenarios and rule-intensive processes in high-knowledge domains, small models remain relevant in the short-to-medium term. Consequently, large and small models will coexist and complement each other. Per Gartner projections, AI penetration in China's manufacturing sector will grow at a CAGR of approximately 10% from 2025-2027. Given fragmented scenarios and physical constraints affecting latency-sensitive operations and system availability in production environments, edge-cloud hybrid architectures will dominate—as high-compute R&D simulations leverage central clouds, while latency-critical industrial control and quality inspection nodes require private edge deployments.

Figure 6 Manufacturing - Generative/Agentic AI implementation scenarios

R&D	Production	Supply	Sale	Service
Generative design	Intelligent products	Intelligent management	Sales assistant	Smart customer service
Software functionality enhancement	Programming assistance	Quality control	Marketing innovation assistance	Predictive maintenance
Process optimisation assistance	Work instruction plan generation	Supply chain cost analysis	Product pricing	Personalised services
Simulation analysis assistance	Exception analysis	Demand forecasting agent	Smart order generation	Remote technical support agent
Demand insights agent	Fault diagnosis and fix agent	Inventory management and optimisation agent	Smart sales forecasting agent	Digital analysis agent
Innovative idea generation agent	Smart production scheduling agent	Supplier coordination and management agent	Targeted customer mining agent	Decision-making support agent
Process simulation agent	Flexible production agent	Risk management agent	Personalised marketing agent	Knowledge library construction and update agent
Smart optimisation agent	Adaptive control agent	Logistics and distribution optimisation agent	Cross-channel integration management agent	Service process optimisation agent

Typical Generative AI application scenarios

Currently, the application of generative AI is primarily concentrated on scenarios characterised by frequent human-machine interaction and high task repetition, aiming to enhance operational efficiency. Looking ahead, its deployment will progressively extend beyond operational execution to encompass decision support and strategic management levels. This evolution will optimise end-to-end manufacturing processes, ultimately enabling enterprises to achieve cost reduction and efficiency gains.

Typical Agentic AI application scenarios

Agentic AI, distinguished by its advanced intelligence, possesses the capability for real-time autonomous decision-making and automated execution. This effectively addresses the complex challenges inherent in manufacturing scenarios, accelerating enterprise digital transformation. Consequently, the manufacturing sector is actively promoting small-scale Agentic AI trials to expedite the large-scale implementation of scenario-specific prototypes. This approach facilitates the construction of a robust closed-loop model encompassing "technical modelling - practical validation - decision refinement".

Source: Manufacturer websites, news media, research publications and other public information, KPMG analysis

- **ROI assessment:** Manufacturing leaders prioritise tangible production line requirements, rigorously targeting strategic investments toward capacity utilisation improvements and verifiable financial value. Complex industrial data architectures and deeply embedded tacit knowledge necessitate particularly rigorous ROI validation due to high trial-and-error costs during initial system deployment.

- **Corporate investment commitment:** Under national smart manufacturing and new-quality productive force mandates, more manufacturers are transitioning from pilot projects to deep implementation. However, leading enterprises and SMEs diverge significantly: leaders with mature digital infrastructure maintain transparent ROI pathways and heightened investment willingness, while SMEs face systemic constraints in cost structures, talent acquisition, and technology accessibility, leading to widespread investment hesitancy.

Automotive: Transition from hardware-centric to data-compute defined smart terminals

- **AI maturity:** Competition within China's automotive sector has fully transitioned from hardware component upgrades to intelligent system evolution centred on data, algorithms, and foundational compute power. Vehicles are evolving into mobile spatial computing platforms generating massive multimodal data flows.

- **Scenario penetration:** OEMs and technology firms are actively pursuing high-value domains including autonomous driving systems, hyper-personalised vehicle design, intelligent marketing services, and smart cockpit experiences. Given the industry's complex product architecture, extended supply chains, and stringent regulatory requirements for privacy and safety, scenario penetration follows phased validation protocols. Per CAAM's latest analysis, China's NEV passenger vehicle market penetration surpassed 50% in 2025. Concurrently, Level 2+ ADAS-equipped vehicle penetration rates continue rising, with advanced driver assistance transitioning from premium-tier options to standard NEV features.

Figure 7 Four main directions of Generative/Agentic AI implementation in the automotive industry

Autonomous driving R&D	Automotive design customisation	Marketing and customer services	Smart cockpit experience
Autonomous driving pre-labelling	Automotive design assistance	Sales script generation	Cockpit image generation
Autonomous driving scenario reconstruction	Coding assistance	Customer service robots	Car mounted Q&A assistant
Multi-modal data generation	Automotive quality inspection assistant	User preference analysis	Vehicle owner health monitoring
Autonomous driving perception prediction	Component fault analysis assistant	Auto after-sales assistant	Car mounted chatbot

Source: public information KPMG analysis

- **ROI assessment:** Foundational automotive deployments follow two primary paths: domain-specific lightweight solutions enabling rapid standardisation and expedited ROI validation, versus end-to-end unified deployments facing multifaceted challenges in heterogeneous data integration and proprietary chip development. The latter presents significantly higher validation complexity.

- **Corporate investment commitment:** Major OEMs sustain consistent strategic investments in frontier technologies, targeting defined objectives: elevating vehicle intelligence, shortening development cycles, and reducing foundational development barriers. Governments nationwide are concurrently establishing autonomous driving demonstration zones through policy support including financial subsidies and tax incentives.

2.2 Key scenario analysis for Agentic AI deployment

The characteristics of a scenario fundamentally shape its pathway to commercial value realisation. When enterprise decision-makers identify and prioritise high-implementation-value business scenarios, they must establish a systematic and rigorous evaluation framework. Within specific business lines, Agentic AI's value proposition manifests most prominently across four critical dimensions: automating high-quality content generation to supplant human-centric tasks; deriving actionable intelligence through sophisticated reasoning across vast information sets; activating latent data value through systematic curation and synthesis; and implementing multimodal adaptive interfaces to optimise end-to-end workflows. Through strategic integration of these capabilities, application systems evolve into accountable drivers of direct business outcomes.

2.2.1 Scenario evaluation and decision pathways

To identify high-potential application directions, enterprises must adopt a three-dimensional evaluation model based on synergies among business, technology, and organisational factors:

Dimension I

Value imperative (business application necessity).
Quantification of technological contributions to overall business strategy and direct financial outcomes.

- Strategic positioning value: Assess whether specific scenario applications can help establish technical barriers or drive core business model innovation.
- Operational demand intensity: Evaluate frontline business units' urgency in resolving bottlenecks and implementing the scenario.
- Direct financial & experience contribution: Quantify potential value through operating cost reductions, new revenue generation, and cross-functional customer/employee experience enhancement.

Dimension II

Technical feasibility (engineering foundations for scalable deployment).
Constrained jointly by external foundational model maturity and internal data infrastructure readiness.

- Technical compatibility: Evaluate whether strict requirements for result accuracy, system response speed, and model explainability align with current technological capabilities.
- Engineering implementation complexity: Assess internal data quality, model fine-tuning costs, and practical integration challenges with core business systems.
- Scenario scalability potential: Evaluate the reusability feasibility of core algorithm components and architectural logic across business lines after successful deployment.

Dimension III

Risk controllability (operational safety baseline).
Requires enterprises to define comprehensive risk inventories and corresponding business continuity plans.

- Inherent technical vulnerability mitigation: Address risk such as prompt injection, training data poisoning, and core algorithm asset theft.
- External regulatory compliance constraints: Ensure strict adherence to industry-specific regulations, data privacy laws, and socio-ethical mandates.
- Internal security safeguards: Establish isolation mechanisms against system failures, malicious attacks, governance gaps, and employee misconduct.

Based on cross-dimensional analysis within the aforementioned three-tier assessment framework, deployment pathways for internal application scenarios demonstrate rigorously logical progression.

- **Back-office functions: Supporting operations feature well-defined, highly standardised business rules.** Within the three-dimensional assessment model, this domain achieves the highest scores in technical feasibility and risk controllability, exhibiting negligible systemic operational risk. Consequently, the back-office represents the optimal domain for large-scale deployment of autonomous AI systems and organisational capability-building.

- **Front-office functions: Directly tied to revenue generation and market expansion,** front-office operations rank highest in value potential within the assessment framework. However, their customer-facing nature imposes stringent risk controllability requirements. Deployment commences at peripheral interaction points, progressively extending into core transaction and pricing functions following substantial human-AI collaboration experience.

- **Middle-office functions: This layer demonstrates high value potential and technical feasibility. Nevertheless, given its custodianship of core business rules and enterprise-wide data assets, it operates under rigorous risk controllability constraints;** deployment here is contingent upon formalised compliance frameworks and air-gapped data environments.

The cross-dimensional calculation also directly determines the architecture options for system deployment.

- In high value scenarios with stringent risk control requirements, the current trend inevitably leads toward rigorous private deployment.

- For peripheral business contexts where technical feasibility is high and fault tolerance is relatively favourable, a hybrid deployment model can be flexibly employed to leverage external public compute. This approach achieves an optimal balance in the overall technical operational costs for enterprises.

This evaluation logic systematically categorises key application scenarios into three core operational pillars: front-office execution, middle-office agile decision-making, and back-office lean operations.

2.2.2 Examples: Scenarios in core segments

01 Front Office Business Units: Redefining End-User Interaction & Automating Business Expansion

Front-office functions engage directly with end clients, demanding exceptionally high standards for system response latency and output precision. The strategic deployment of Agentic AI within this domain targets the creation of personalised interaction paradigms and the achievement of direct commercial growth objectives. Aligned with risk control requirements, the enterprise currently prioritises implementation through higher-fault-tolerance content generation modules, progressively extending toward complex, high-stakes core transactional activities. Regarding deployment architecture, front-office large-scale interactions typically face high-concurrency traffic spikes. A hybrid deployment approach within non-sensitive operational environments emerges as the optimal solution, optimising elastic scaling capabilities against routine operational expenditures.

Business domain	Strategic scenario deployment	Solutions to core operational bottlenecks and business evolution pathways
Marketing strategy execution	Automated marketing journey orchestration & direct strategy resource deployment	Replace manual single-channel orchestration. Agentic systems handle multi-channel campaign deployment, dynamically reallocating budgets based on real-time exposure data to maximise marketing ROI.
Personalised interactions with end-users	Highly tailored end-user interaction experiences & conversion funnel guidance	Move beyond one-way broadcasting. The system automatically adapts messaging and next-step guidance based on real-time user engagement signals, creating fully customised end-to-end marketing workflows.
Sales conversion & lead follow-up	Automated full lifecycle lead management & business tracking systems	Free sales teams from tedious CRM data entry. AI agents automate email outreach, meeting scheduling, and objective lead scoring, surfacing high-intent prospects directly to sales.
Commercial demand forecasting & pricing	Dynamic end-market demand & automated pricing strategy adjustment	Transform manual, lagging-report-based price adjustments. The system autonomously calculates optimal price points using real-time competitor data and conversion rate fluctuations, executing changes directly in front-office systems.
End-to-end customer service support	Conversational multi-modal query handling & complex complaint resolution automation	Overcome human agent concurrency limits. After precisely identifying complex customer intents, agents directly invoke core system APIs to process returns/exchanges or billing adjustments, significantly boosting first-contact resolution (FCR) rates.

02 Middle office operations: Establishing a data-driven agile decision-making and automated risk control hub

The Middle Office system, primarily comprising Operations Management, Supply Chain Coordination, Risk Control, and Data Security departments, constitutes the critical support backbone ensuring enterprise stability. Its operational characteristics feature high rule density, strong data interdependencies, coupled with significant system complexity and stringent regulatory compliance constraints. The high-value imperative for Agentic AI in this area lies in achieving deep data penetration analysis and automated, pre-emptive risk intervention. As it directly governs the core business rules engine and enterprise-wide data asset pools, Middle Office scenarios face exceptionally stringent business continuity and regulatory security requirements. Its automation deployment is highly dependent on robust permission isolation. Private cloud deployment or strictly controlled hybrid cloud models are the typical choices for this domain.

Business domain	Strategic scenario deployment	Solutions to core operational bottlenecks and business evolution pathways
Supply chain scheduling and collaborative control	Proactive identification of external risks and automated order flow optimisation	Mitigate supply chain disruptions caused by macro-level emergencies. When detecting supplier anomalies, the system autonomously generates contingency plans and initiates automated procurement requests to alternative partners, ensuring uninterrupted core business continuity.
Anomaly risk assessment and proactive intervention	Real-time disruption prevention and adaptive risk pattern reconstruction	Risk management has evolved from post-event compliance reviews to in-process anomaly interception. AI agents identify suspicious transaction patterns within milliseconds, then directly execute account freezes and other security interventions through system-level permissions.
Regulatory compliance automation	Dynamic tracking of regulatory changes and automated internal remediation	Address delays in manual interpretation of complex and frequently updated legal requirements. The system automatically compares new regulatory mandates against existing internal controls, generating compliance gap reports and dispatching time-bound corrective action instructions to designated roles.
Advanced data analytics and self-service modelling	Intelligent data relationship mapping and enterprise-wide query response system	Reduce inefficiencies where business personnel depend on technical staff to provide analytical support. After accurately understanding natural language instructions from business units, the agent programme autonomously writes retrieval code, accesses relevant databases, and generates professional visual reports.
Cybersecurity and vulnerability testing	Dynamic malicious traffic interception and defence with automated system-level vulnerability scanning and patching	In environments characterised by frequent and covert cyberattacks, the entity responsible for security defence has shifted from human engineers to autonomous agents. When subjected to attacks, the system can autonomously isolate compromised microservice nodes while simultaneously deploying and repairing underlying security patches.

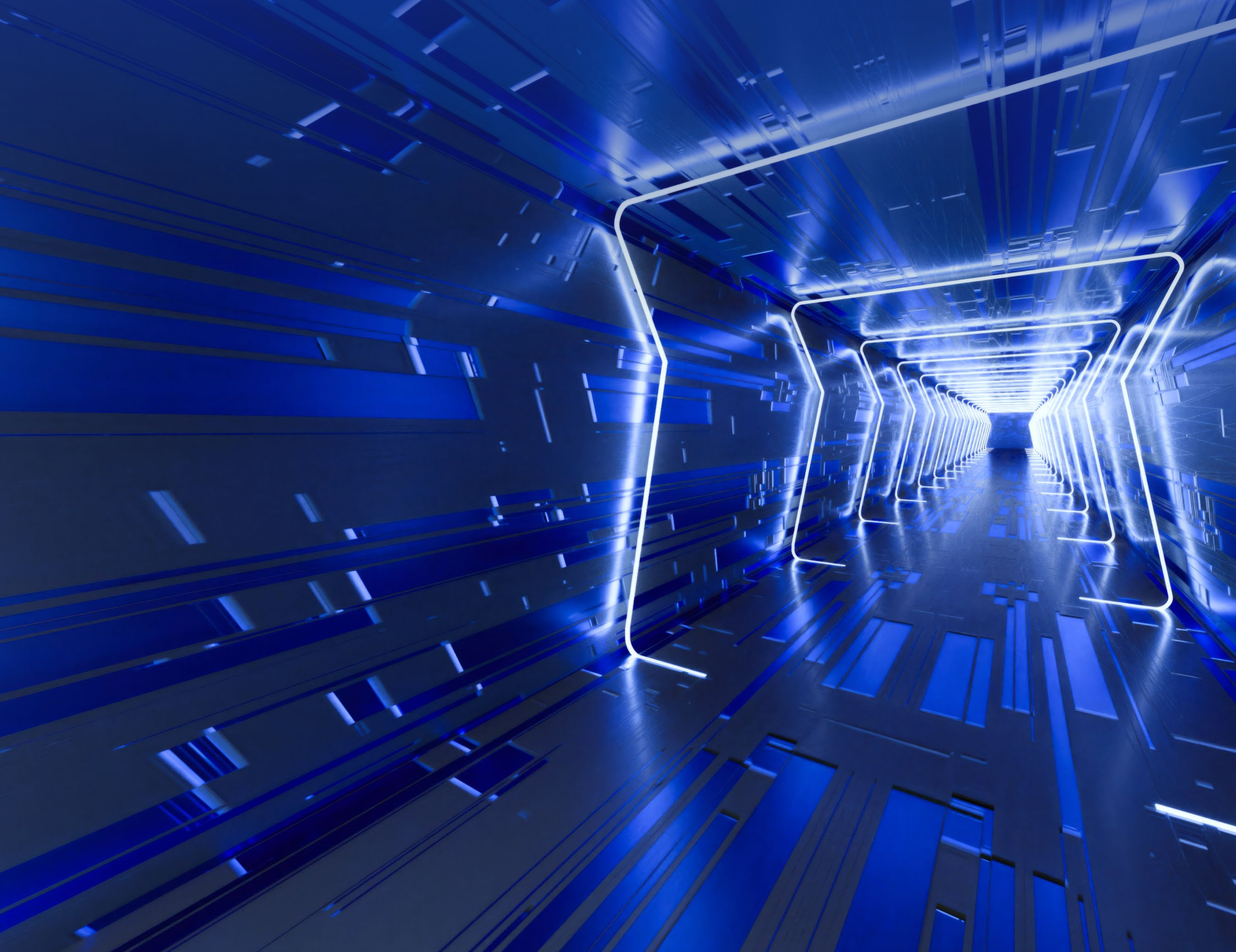
03 Back-office management: End-to-end process automation

Back-office departments exhibit highly standardised workflows, repetitive transactional tasks, and massive structured document processing volumes. According to the three-dimensional assessment model, these scenarios demonstrate moderate complexity with well-defined error tolerance boundaries and tightly controlled operational risks, achieving the highest technical feasibility scores across all domains. The ultimate implementation goal is to fully automate baseline operational tasks traditionally performed by entry-level positions.

Business domain	Strategic scenario deployment	Solutions to key operational bottlenecks and business evolution pathways
Organisational talent management & employee development strategy	End-to-end automation of talent recruitment processes and dynamic career pathway planning for employees	The recruitment pipeline—from automated parsing and screening of high-volume resumes to sequential execution of preliminary interview invitations—is fully managed by intelligent systems. For talent development, the platform autonomously calculates training needs based on historical performance data, enabling precise distribution of online professional resources and real-time competency tracking.
Financial operations & audit compliance	Dynamic predictive analysis of core financial KPIs and automated high-volume invoice verification	The financial operating model has evolved from reactive accounting to proactive profitability forecasting. The system automatically extracts cross-departmental expenditure streams for real-time reconciliation, directly generating financial optimisation proposals with decision-support value for executive review.
IT infrastructure support and R&D	Assisted refactoring of core application code and automated execution of software test cases	Development tooling has advanced from coding assistance to automated development units responsible for quality assurance of specific functional modules. This transformation reduces software iteration cycles by 40% while enabling preliminary self-diagnosis and recovery from routine system anomalies.
Internal administrative collaboration & knowledge asset governance	Autonomous construction of enterprise-wide knowledge bases and automated tracking of cross-departmental meeting resolutions	AI agent systems dismantle interdepartmental communication silos by automatically documenting transnational business decisions. These systems decompose resolutions into actionable task lists while maintaining 24/7 visibility into departmental implementation progress.
Asset documentation management & legal contract review	Lifecycle monitoring of fixed commercial assets and automated review of standard commercial contracts	Back-office legal and asset management teams are liberated from manual document verification. The system automatically identifies inconsistencies and risk imbalances in procurement contracts and generates revised compliant versions ready for formal execution.

Chapter 3

Roadmap and Critical Measures for AI Transformation



3.1 Systematic construction and governance of enterprise

The new generation of AI technology, epitomised by LLMs, is evolving from discrete efficiency tools into systemic engines that reshape enterprises' global business architecture. Translating technological capabilities into sustained commercial value remains contingent upon enterprises' disciplined execution of comprehensive transformation planning.

3.1.1 Baseline for change and strategic positioning: Clarifying the relationship between AI and business strategies

Comprehensive adoption of AI agents represents an inevitable strategic imperative for enterprises navigating the new competitive cycle. To this end, organisations must first implement the critical initial measure: establishing precise strategic positioning for AI, clarifying its relationship to core business strategy, and developing a clearly defined AI strategic blueprint.

Table 1 Enterprise maturity progression demonstrates distinct phases when advancing its technology strategy.

Definition of evolution stage	Core business characteristics and implementation status	Main limiting factors faced when transitioning to the next stage	Expected business impact
Awareness and exploratory stage	Ad hoc deployment of basic language models by individual employees for discrete tasks occurs without centralised compliance oversight or dedicated budget allocation.	An absence of enterprise-wide coordination mechanisms is compounded by insufficient executive sponsorship for targeted funding.	Impact remains confined to marginal task-level productivity gains within individual employees' daily desktop operations.
Structured experimentation stage	Structured implementation of automation tools within designated pilot business units, governed by rigorous financial and operational KPIs.	Underlying data infrastructure proves inadequate, with no established governance framework or cross-departmental access protocols.	Expedites throughput velocity while reducing error rates within individual business workflows.
Departmental scaling stage	The end-to-end workflow within a single functional department achieves comprehensive intelligent agency and automated flow.	Persistent absence of cross-departmental governance structures and enterprise-wide compliance standards.	Substantial optimisation of operational cost structures within functional departments.
Enterprise-wide horizontal integration stage	Multi-agent collaborative ecosystems operating across geographical regions and business units.	Leadership models misaligned with top-tier business strategies and lacking strategic coherence.	Marked enhancement of cross-departmental collaboration efficiency and strengthened enterprise-wide market responsiveness.
Native AI operation stage	Algorithmic logic fully embedded within the enterprise's highest-level decision-making frameworks, core business processes, and strategic planning cycles.	Continuous maintenance of technological agility and compliance resilience amidst volatile regulatory landscapes.	Enterprises reconfigure business models to establish distinct and sustainable competitive advantages.

Source: KPMG analysis

Business strategy driven and core resource allocation principles:

To ensure significant technical-level capital expenditures translate into substantial commercial returns, enterprises must anchor their strategic planning in a core business-value-driven mechanism. This approach necessitates that management first clarifies the core business development objectives for the current financial year. Through quantitative analysis, they must then identify the area of greatest deviation between existing business performance and strategic targets, precisely pinpointing the highest operational friction points within both customer interaction journeys and employee-internal workflow processes.

Our experience demonstrates that in any successful AI transformation initiative, only a marginal proportion of commercial value originates from performance optimisation of the underlying algorithmic models themselves. Similarly, limited value derives from deploying computational clusters and constructing technical infrastructure for data pipeline integration. The overwhelmingly predominant source of core business value resides in reshaping internal leadership structures, comprehensively upgrading human resource capabilities, fundamentally re-engineering business operation workflows, and restructuring cross-departmental collaboration mechanisms. A genuinely future-ready organisation will inevitably allocate over half its budgetary resources and organisational energy towards reskilling employees and driving structural evolution. This commitment is essential to establish a new operational paradigm adapted to the era of AI agents.

3.1.2 High-value AI scenario mining: Assessing return on investment and prioritisation

With the rapid iteration of full-stack AI agent technologies, the application boundaries accessible to enterprises are expanding at an unprecedented pace. Faced with an extensive array of potential implementation opportunities, organisations lacking a structured prioritisation framework risk becoming ensnared in a management predicament characterised by highly dispersed resources and diminished investment returns. To mitigate investment risks and validate business models expediently, enterprises should adopt an incremental strategy progressing from lightweight implementation to heavyweight expansion.

Consequently, organisations must implement the second critical measure: comprehensive identification of high-value AI application scenarios. This entails conducting precise return on investment (ROI) calculations for each scenario to establish implementation priorities.

- **The identification of high-value scenarios must adhere to rigorous, standardised analytical processes.** The enterprise's transformation steering committee should undertake in-depth consultations with business unit heads, conducting systematic analyses of operational challenges. During the preliminary phase of scenario discovery, direct engagement from senior business sponsors is essential. This ensures selected use cases possess both adequate financial resourcing and clearly defined business ownership within the enterprise's overarching strategic framework.
- **When evaluating a business scenario's transformative potential, fundamental task attributes must be examined**: Processes characterised by high repetition, data-intensive operations, routine multi-source content integration, and well-defined decision pathways are prime candidates. For such workflows, AI agent implementation typically achieves exceptional productivity conversion rates, translating directly into measurable expense reduction. Conversely, scenarios involving risk-sensitive undertakings requiring profound interpersonal engagement, highly ambiguous decision-making, or minimal error tolerance rarely demonstrate immediate quantifiable expense reductions in financial statements.
- **Evaluation matrix and ROI quantification:** Following initial scenario identification, enterprises should employ a structured, multi-dimensional scoring matrix to rigorously quantify and prioritise all candidate use cases. Concurrently, a standardised ROI quantification framework must be established for each shortlisted use case, enabling precise financial evaluation against defined business objectives.

Reference logic for ROI quantification of enterprise AI scenarios

ROI=direct financial benefits (including operational cost savings, increased revenue, and reduced working hours-equivalent compensation)/total cost of implementation and maintenance (including compute leasing, data procurement, cross system integration, and model fine-tuning)

Table 2 **Core dimensions of enterprise AI scenario evaluation**

Core dimensions of evaluation	Evaluation indicators	Direct impact on ROI and project success
Business value expectation	Contributions to revenue growth, reductions in direct labour costs, and decreased customer churn rates.	The core underlying factors directly determine the ceiling for project ROI.
Data readiness	Internal data integrity, degree of structuring, real-time retrieval latency, and clarity of ownership.	The most critical threshold criterion for project initiation viability directly impacts the avoidance of unnecessary expenditure.
Technical feasibility	Required model development complexity, integration compatibility with existing systems, and estimated computational resource consumption.	Core development costs within the corresponding denominator determine the initial capital expenditure burn rate.
Value realisation cycle	Proof-of-concept completion timeline, pilot phase duration, and estimated months to full-scale deployment.	Maintenance of capital's time value and sustained commitment from internal corporate governance structures.

Source: KPMG analysis

3.1.3 Scenario-related process transformation and reshaping: Designing human-machine collaboration mechanisms

A single agent faces inherent limitations in memory retention and processing capacity when managing cross-domain, multi-node operations. Consequently, establishing multi-agent collaborative networks has become essential for implementing complex business processes. Architecturally, this necessitates moving beyond traditional linear workflows to incorporate parallel processing and state machine management systems.

Organisations should therefore implement the third critical measure: comprehensive process transformation with corresponding mechanism redesign, including provisions for both integrated personnel and external stakeholders.

Enterprises must adopt context-appropriate multi-agent coordination frameworks—such as hierarchical, pipelined, handover, or debate-based models—to enable decomposition, distribution, and cross-validation of complex operational instructions. Concurrently, organisations should establish sophisticated human-AI collaboration protocols, clearly delineate operational boundaries, and implement mandatory human review checkpoints for business functions where error tolerance costs are significant.

The "human-in-the-loop" mechanism fundamentally involves the deliberate integration of human oversight at critical decision nodes within automated workflows. This approach operates on the premise that potential error costs in AI systems substantially exceed human intervention expenses. The design philosophy intentionally avoids requiring micro-management of every AI operation, instead focusing robust oversight on high-risk scenarios.

Table 3 **Comparison of human-machine collaborative mechanisms**

Collaborative mechanism	Business execution permissions and system settings	Applicable business scenarios and risk characteristics	Implementation restrictions and human workload considerations
Human in the Loop (HITL)	The system exclusively generates draft recommendations, mandating workflow suspension. Only authorised personnel may sequentially execute approval instructions to effectuate valid implementation.	Scenarios characterised by low fault tolerance, infrequent concurrency, and irreversible decision-making outcomes; for instance, risk classification or decision-support formulation.	To compensate for human reviewers' cognitive processing constraints, this mechanism prioritises decision security through throughput efficiency optimisation.
Human on the Loop (HOTL)	Within predefined policy parameters, AI agents perform automated closed-loop operations by default, while human operators conduct parallel state monitoring and retain supreme authority for immediate termination through single-point intervention.	Medium-to-low risk scenarios with clearly defined rule boundaries and high throughput requirements; exemplified by routine IT fault resolution or standardised content modification.	Robust human-AI collaborative monitoring frameworks and capability development are imperative to prevent supervisory personnel desensitisation and fatigue from alert overload.
Human out of the Loop (HOOTL)	AI agents execute fully autonomous closed-loop operations, with human oversight limited to non-real-time quality sampling audits of macro-level statistical distribution characteristics.	Ultra-low risk scenarios featuring completely reversible operations and exceptionally high-volume data processing; for example, internal foundational Q&A retrieval systems.	This is strictly bound by data protection regulations, and prohibited to be applied on automatic rating activities exerting material legal consequences upon natural persons.

Source: KPMG analysis

Dynamic routing and compliance policy constraints:

When integrating diverse supervision modalities into operational business pipelines, organisations must implement dynamic routing configuration mechanisms predicated on model confidence metrics and business loss thresholds. During request processing, the AI agent synchronously generates a quantifiable confidence score. Where system confidence exceeds established thresholds and the commercial risk profile remains within safe parameters, requests are automatically routed to HOOTL pathways to ensure rapid response. Should confidence levels fall below baseline thresholds or encounter financially sensitive areas, workflows are immediately diverted to HITL review queues for domain expert verification and decision-making.

Local regulatory frameworks stringently constrain process redesign security boundaries. Decisions materially impacting natural persons are subject to rigorous cross-jurisdictional oversight. Certain sectors expressly prohibit fully automated decision-making without substantive human intervention. For high-risk applications, organisations must maintain comprehensive technical documentation of risk assessments and activity logs to demonstrate adherence during unannounced compliance audits.

3.1.4 Knowledge management and systems: Building next-generation intelligent data asset frameworks

High-fidelity contextual data constitutes the foundational prerequisite for AI agent decision-making accuracy. As enterprise data management evolves towards multimodal dimensions, organisations must prioritise knowledge management and governance as central pillars of data assetisation initiatives.

Consequently, enterprises must implement the fourth critical measure: establishing comprehensive knowledge management and engineering infrastructures.

The next-generation architecture’s core mission is the complete dissolution of physical data silos across business systems, transforming vast unstructured datasets into machine-readable, granularly access-controlled asset networks. Knowledge engineering systems address this challenge through three core components: a semantic abstraction layer, domain-specific ontology modelling and dynamically evolving knowledge graphs. By converting business knowledge, defined rules, and entity relationships into machine-interpretable structures, this framework enables cross-domain logical reasoning beyond isolated fact retrieval. The integrated semantic layer and knowledge graphs precisely encode business logic, furnishing AI agents with globally contextualised decision-making environments. This deep structural encoding establishes knowledge engineering as the dynamic core that continuously enhances ROI through operational cost reduction, decision velocity improvement and assurance of AI operations reflecting enterprise-specific domain expertise.

Standardised knowledge production pipeline and dynamic permission awareness:

Enterprises are establishing standardised knowledge engineering pipelines encompassing the entire lifecycle of intelligent information collection, data cleansing, quality verification, and final asset deployment. Through a rigorous knowledge governance responsibility matrix, business units, domain experts, compliance departments, and IT divisions possess clearly delineated responsibilities for knowledge creation, review, and archiving. This effectively mitigates governance gaps and prevents management vacuums.

Within highly regulated sectors, next-generation knowledge engineering systems must embed granular access control lists at every architectural query interaction point. These systems require real-time permission awareness and validation capabilities, integrated with data desensitisation and substitution technologies. This enables comprehensive evaluation of the request initiator’s digital identity, physical environment risk profile, and target data confidentiality level during retrieval – facilitating complex, millisecond-scale authorisation adjudication. To address divergent compliance requirements across business lines, enterprises should implement hybrid infrastructure configurations. This approach achieves optimal equilibrium between cloud-native agility and data sovereignty imperatives.

3.1.5 Organisational governance and talent capabilities: Reconstructing enterprise operational models

The comprehensive replacement of enterprise technology stacks represents merely the surface manifestation of intelligent transformation. The ultimate determinant of success resides in the complex core domain: the fundamental restructuring of internal organisational frameworks, profound evolution of human capital competencies, and reconfiguration of risk governance architectures.

Enterprises must therefore implement the fifth critical measure: advancing organisational governance and talent capability maturation.

Corporate structures are transitioning from vertically centralised, monolithic command-and-control models toward collaborative architectures integrating universal technology enablement platforms with distributed agile execution units. To optimise the equilibrium between centralised security-compliance oversight and decentralised operational responsiveness, industry leaders are decisively adopting federated operating frameworks. Within this hybrid paradigm, enterprises establish group-level Centres of Excellence to orchestrate foundational computing resources and mandate unified security protocols. Meanwhile, the end-to-end delivery cycle—from deep-mining specialised application scenarios through to operational deployment—remains wholly owned by cross-functional agile units constituted within respective business divisions.

In terms of management structure, enterprises are expediting the creation of Chief Artificial Intelligence Officer (CAIO) positions. This role bears core responsibility for orchestrating business-technology integration and leading enterprise-wide digital transformation initiatives. To facilitate effective human-AI collaboration, organisations have reconstituted their performance management frameworks, instituting a tripartite assessment methodology encompassing: Outcome Metrics: Evaluating direct commercial value generated by AI agents;Adoption Metrics: Measuring diffusion rates of successfully deployed AI tools across business units;Capability Metrics: Tracking workforce transition to strategic cognitive tasks and accretion of high-value intellectual capital.

Large-scale talent capability enhancement and perpetual upskilling ecosystems

The integration of agentic AI necessitates embedding AI collaboration competencies into core workflows and position specifications across organisations. Concurrently, progressive enterprises are fundamentally redefining promotion criteria and performance assessment frameworks, prioritising AI literacy and benchmarking proficiency in agentic AI utilisation.

Forward-thinking institutions are committing substantial investment to talent skill renewal programmes, developing explicit capability transformation roadmaps. To ensure effective execution of talent evolution strategies, industry leaders have implemented structural reforms within performance management systems. These include mandating protected upskilling hours with compensated study leave, while actively cultivating organisation-wide cultures of continuous skill regeneration. This foundational approach provides sustainable intellectual capital underpinning enterprise transformation initiatives.

Trusted AI systems and compliance control architecture

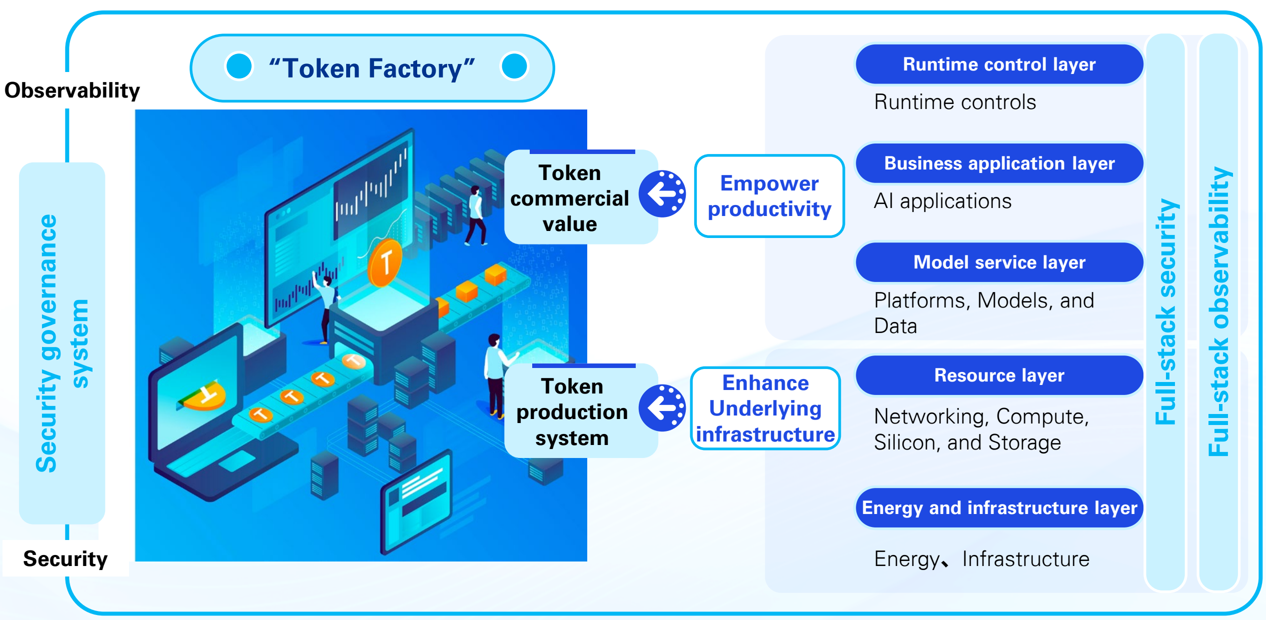
The foundational prerequisite for enterprise-wide adoption of new AI systems – by both employees and customer groups – rests upon establishing robust institutional trust. Constructing a trusted AI framework constitutes an essential undertaking for mature organisations. This necessitates implementing mandatory verification protocols throughout the entire system lifecycle, from research and development through to operational deployment. Confronting challenges posed by algorithmic opacity, enterprises must enforce rigorous risk stratification mechanisms. For high-stakes scenarios directly impacting customer rights, systems should mandate perpetual oversight by dedicated ethics committees. These bodies conduct concurrent technical and compliance audits to enable pre-emptive mitigation of regulatory breaches and associated penalties.

3.2 Full stack enhancement, empowerment, and security of enterprise infrastructure

Under the Agentic AI paradigm, AI has transitioned from single-round Q&A to continuous execution. This shift drives exponential growth in per-process token consumption, call chain complexity, context volume, and operational intricacy. Should enterprises persist in evaluating AI investment solely through computational scale, model parameters, or application quantity, they will fundamentally misapprehend the relationships between AI expenditure, output yield, and business value. Crucially, in an era where tokens constitute the universal metric for AI capability, enterprise AI operations now fundamentally comprise the continuous production, circulation, scheduling, and delivery of tokenised value.

Consequently, tokens are emerging as the pivotal reference point for re-evaluating AI infrastructure. Organisations must now establish "token factories" capable of supporting large-scale token production, high-quality delivery, and secure, stable operations. As tokens accelerate their evolution into core business variables—quantifying intelligent output, systemic costs, and commercial value—token economics will correspondingly form the foundational logic governing enterprise AI infrastructure development.

Figure 10 The mapping relationship between traditional enterprise infrastructure and token factories



Source: Public information, KPMG analysis

3.2.1 Enhancing underlying infrastructure: Breaking through hardware performance limits to build token production systems

Within the token factory paradigm, the foundational layer constitutes the systemic capability bedrock enabling token supply. Power provisioning, thermal management, spatial configuration, reliability engineering, and edge deployment capabilities collectively determine the physical production ceiling for token throughput. Simultaneously, the coordinated performance of computational resources, network infrastructure, storage systems, and orchestration frameworks governs both the quality of token production capacity and operational yield conversion efficiency.

01 Energy and infrastructure layer: Shifting from IT support to token production base

- **Energy and infrastructure represent hard constraints on the supply capacity of enterprise AI systems, making it progressively more challenging to support the rapidly expanding AI workload.** On the one hand, the continuous growth of AI workloads will impose significantly higher demands on the resilience of power, cooling, space, and supporting infrastructure. On the other hand, the construction cycle for physical infrastructure—including power supply access, transformation and distribution systems, refrigeration systems, and data centre refits—is typically far longer than the procurement cycle for computing hardware. This will inevitably create an increasingly acute structural contradiction: fast-growing AI workloads constrained by slow energy and facility upgrades within enterprises. According to the Cisco 2025 AI Readiness Index[1], based on the average level among surveyed companies, 51% of organisations anticipate their AI workloads increasing by more than 50% over the next three to five years. However, only 43% have established purpose-built infrastructure to optimise power consumption for AI deployments.

- **Improvements in data centre energy efficiency have reached a plateau phase, intensifying pressures for infrastructure upgrades and retrofits.** Historically, many facilities achieved significant reductions in Power Usage Effectiveness (PUE) from higher baselines through methods including hot/cold aisle containment, refrigeration system optimisation, and uninterruptible power supply (UPS) efficiency enhancements. However, the Uptime Institute's 2025 Global Data Centre Survey[2]indicates that respondents' weighted average annual PUE stands at 1.54 – a level maintained for six consecutive years. This demonstrates that the readily attainable energy efficiency gains of the past are diminishing. Consequently, the pressure from additional AI workloads will translate more directly into demands on power supply capacity, cooling capacity, cabinet power density, spatial configuration, redundant architecture design, and facility refurbishment investment.

- **The positioning of the energy and infrastructure layer has shifted towards serving as the token production base.** The core value of this base lies in its ability to effectively support AI workloads and stabilise token supply, while continuously optimising cost, latency, and reliability throughout the process. Specifically, power supply capacity determines the foundational power available and the scale of the load that can be supported. Cooling capacity dictates the system's continuous operational capability. The power density and spatial layout of server cabinets govern deployment efficiency and expansion flexibility per unit of space. Furthermore, redundant architecture and reliability engineering determine operational stability. As large-scale model training, real-time inference, and Agentic AI applications move into manufacturing sites, financial institutions, and industrial parks, these workloads will also face constraints related to low latency, data localisation, privacy compliance, and on-site business continuity. This necessitates collaborative deployment strategies across centralised clouds, enterprise data centres, and edge nodes.

- **The foundational infrastructure will exhibit stronger infrastructure-driven and operational focus attributes.** 'Infrastructure-driven' denotes that enterprises must proactively incorporate engineering prerequisites – including power supply, cooling systems, data centre space, networking, redundancy provisions, and edge node integration – thereby planning token production capacity with the same rigour as industrial manufacturing capacity. 'Operational focus' signifies that enterprises require ongoing monitoring, scheduling, and optimisation centred on power utilisation, cooling headroom, thermal efficiency, system availability, task prioritisation, and per-token operational costs.

1 How AI leaders architect differently,Cisco

https://www.cisco.com/c/dam/m/en_us/solutions/ai/readiness-index/2025-m12/documents/Cisco-AI-Readiness-Index_Infrastructure-Focus.pdf

2 Uptime Institute Global Data Center Survey 2025, Uptime Institute

https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/2025.Annual.Survey.Report.pdf?mkt_tok=NzExLVJJQS0xNDUAAAGcNPnjzIRtJDIpo9-Khi9n18G0DO03Sio3lnUg39J2C4NwajSneDRbDbQS68NjcylaxC-tv-Wxs8gNe8yLCsRPInSsJBBujdGUMIQQZ66IJcvg&version=0&utm_source=chatgpt.com

02 Resource layer: Transitioning from parallel resources to collaborative production systems

- **The resource layer (encompassing compute, networking, storage, and scheduling) governs the effective release of token capacity, where deficiencies in any component diminish overall efficiency.** Specifically: Compute determines the throughput of model training and inference; Networking governs data flow efficiency within clusters and across environments; Storage dictates responsiveness in data retrieval, vector search, and context construction; Scheduling ensures appropriate resource allocation based on task priority, latency requirements, data sensitivity, and cost constraints. Critically, resource availability does not guarantee resource controllability—particularly when training, inference, knowledge retrieval, and agent execution operate concurrently on shared infrastructure. The capacity to monitor and regulate resource consumption, task efficiency, and security risks directly impacts system longevity. Consequently, enterprises must concurrently enhance observability and security capabilities at this layer, enabling continuous monitoring, dynamic scheduling, and stable operation within defined security parameters.

- **Networking represents a frequently underestimated component within the resource layer, yet potentially constitutes the earliest constraint on AI's large-scale deployment.** As AI applications advance into real-time inference, multi-step reasoning, and multi-agent execution phases, network transmission efficiency directly impacts GPU utilisation, task responsiveness, and service stability. The cost implications of computational idle time during network-induced delays become increasingly significant. Furthermore, once data centre networks enter production environments, large-scale retrofits typically incur substantial downtime costs and extended implementation cycles. Consequently, enterprises must proactively design network capacity and cross-system integration capabilities before AI workloads accelerate. The Cisco 2025 AI Readiness Index underscores these challenges[3]: surveyed enterprises demonstrate merely 30% average integration between AI deployments and infrastructure stacks, indicating widespread system fragmentation. Simultaneously, only 15% of organisations possess fully adaptable network scaling capabilities – a limitation likely to bottleneck AI expansion before compute becomes constrained.

- **The advancement of AI networking capabilities manifests primarily through elevated transmission efficiency and sustained high-load operational capability.** High-speed optical modules, power-efficient optical interconnects, and high-bandwidth switching devices enable faster, more stable, and lower-power data flow between GPUs, servers, storage systems, and racks. This reduces computational efficiency losses caused by data transmission bottlenecks. Concurrently, emerging technologies such as liquid-cooled switches ensure continuous operation of high-bandwidth networking equipment under prolonged heavy workloads, preventing degradation of overall AI cluster performance due to inadequate thermal management of network components.

- **The coordinated integration of computation, networking, storage, and scheduling is pivotal for optimising large-scale inference costs.** Enterprise AI systems consistently generate substantial volumes of repetitive or similar contextual states. Processing these contexts from scratch per request leads to significant redundant GPU computation, elevating computational power consumption, queuing latency, and energy usage. Over time, finite GPU memory and system memory capacity will struggle to accommodate expanding context caches. Organisations may convert duplicate computation into more economical storage and network access through cache reuse, tiered storage, and high-speed networking – an approach known as "storage for computing". However, inadequate network speeds, sluggish storage, and prolonged GPU cache retrieval can degrade system performance. Consequently, a new cost trade-off emerges at the resource layer: repeated computation incurs GPU time, task queuing, and compute expenditure, whereas cache reuse introduces storage occupancy, network handling, and read latency costs. Enterprises must enhance coordination between computation, networking, storage, and scheduling. Intelligent scheduling systems should evaluate task types, cache hit rates, data locality, latency requirements, and resource pricing to determine the optimal balance between recomputation costs and cached data retrieval expenses.

- **Contemporary infrastructure architecture exhibits pronounced systemic collaboration characteristics.** Industry leaders including NVIDIA, Google Cloud, and Microsoft Azure are progressively de-emphasising isolated hardware component performance in AI infrastructure design, instead prioritising holistic coordination across chip interconnects, memory bandwidth, network architecture, storage throughput, thermal management, and resource orchestration. Consequently, enterprise infrastructure development now extends beyond computational scale to encompass continuous optimisation capabilities for compute utilisation efficiency, network transmission performance, task completion stability, and operational expenditure. The underlying principle remains consistent: as AI workloads grow in complexity, the final output becomes less dependent on any individual hardware component's performance. What fundamentally determines enterprise AI return on investment is the system's capacity to deliver sustained, stable, and cost-effective integration of data pipelines, model architectures, computational resources, and business workflows.

---

3 how AI leaders architect differently, Cisco
https://www.cisco.com/c/dam/m/en_us/solutions/ai/readiness-index/2025-m12/documents/Cisco-AI-Readiness-Index_Infrastructure-Focus.pdf

Table 4 **Critical initiatives to enhance underlying infrastructure from the perspective of Token economics**

	Objectives	Technical focus	Assessment mechanism
Energy and infrastructure layer	• Addressing the conflict between continuously growing AI workloads and traditional data centres' inadequate load capacity, with an aim of optimising per-token cost.	• **Load-Bearing Transformation:** Re-engineering power and thermal management for AI workloads, with liquid cooling established as the critical solution for high-power-density deployments – shifting facilities from static capacity reservation to dynamic load matching. • **Operational Synchronisation:** Integrating capacity alerts and failure response mechanisms into AI scheduling to minimise facility instability impacts on business continuity. • **Edge-Centre Orchestration:** Defining roles between central and edge nodes based on latency requirements and local data processing needs, preventing excessive AI workload centralisation.	• **Capability Validation:** Translating power, thermal, and spatial specifications into AI load-oriented metrics to verify foundational readiness for large-scale inference throughput. • **Efficiency Benchmarking:** Implementing continuous accounting of energy consumption, cooling efficiency, space utilisation, and maintenance costs – with per-token facility expenditure as the core KPI – transitioning data centres from capital expenditure assets to operationally optimised production units. • **Resilience Assurance:** Establishing operational safeguards through capacity monitoring, fault recovery protocols, and impact analysis to determine whether facility fluctuations propagate as service disruptions or degraded inference stability.
Resource layer	• Resolving the tension between complex AI workloads and fragmented resource provisioning, with inference efficiency optimisation as the core objective.	• **Unified Resource Pool:** Integrating heterogeneous compute, network, and storage assets through a centralised scheduling platform with AI workload observability. • **High-Speed Interconnect Optimisation:** Leveraging optical networking components to enhance intra-cluster data transfer, alleviating network constraints in distributed training and high-concurrency inference. • **Task-Centric Allocation:** Dynamically matching resource provisioning to workload fluctuations through intelligent scheduling of extended-duration AI task chains.	• **Unified Metrics Framework:** Establishing cross-resource performance indicators spanning token throughput, task latency, resource utilisation, and unit cost to eliminate siloed evaluation and attribution challenges. • **Integrated Cost Allocation:** Incorporating compute utilisation, network transmission, storage access, scheduling overhead, and task outcomes within a unified cost accounting framework. • **Continuous Improvement Cycle:** Feeding evaluation results into resource allocation, capacity planning, and cost distribution mechanisms to transform resource coordination from static configuration to perpetual optimisation.

3.2.2 Empowering upper-layer productivity: Anchoring business application efficacy and unlocking token-based value realisation

Having resolved foundational token supply constraints at the infrastructure layer, enterprises now face the more critical challenge of effectively orchestrating, reliably invoking, and ultimately translating these tokens into tangible business process outcomes. Through service models, closed-loop applications, and runtime control, each token unit becomes more readily convertible into demonstrable business value – directly enhancing organisational productivity.

01 Model service layer: From discrete tokens to encapsulated intelligence – meeting enterprise intelligent service demands

- **Enterprise demand for AI is progressively manifesting as requirements for intelligent service capabilities.** Based on real-world enterprise deployments of AI agents, organisational AI needs centre on intelligent services that facilitate cross-departmental collaboration and drive task execution. According to KPMG's Global AI Pulse Q1 2026[4], AI applications are extending beyond functional use cases towards cross-functional collaboration: amongst surveyed organisations actively deploying AI agents, 52% utilise AI to automate multi-functional workflows, 41% employ AI to establish shared knowledge environments, and 40% leverage AI to support cross-team collaborative decision-making. However, most organisations remain in the early transition phase between task automation and cross-functional coordination, having yet to develop genuine enterprise-level orchestration capabilities.

- **Model supply-side product offerings are evolving, transitioning AI from universal conversational interfaces towards structured workflows and enterprise-grade service encapsulation.** Per OpenAI's The State of Enterprise AI 2025[5], weekly active users for Custom GPTs and Projects have grown approximately nineteen-fold since year-start. In recent months, roughly 20% of user interactions within ChatGPT Enterprise were processed through Custom GPTs or Projects. The most widely deployed GPTs primarily serve to embed organisational knowledge within reusable assistants or achieve workflow automation through internal system integrations. Notably, certain enterprises have cultivated cultures promoting large-scale development and sharing of Custom GPTs.

- **Disparate departmental management inhibits organisational capability consolidation, creating an imperative for enterprises to establish unified service frameworks.** Permitting scattered model usage across departments fosters operational silos characterised by redundant development, fragmented capabilities, non-reusable knowledge, disconnected processes, and inconsistent security protocols. Consequently, while model output volumes may increase, genuinely valuable intelligent services that embed organisational capabilities remain limited. Enterprises increasingly require integration of models, data resources, toolchains, and security policies within a cohesive framework. This enables business systems to: invoke model capabilities with consistent standards; access enterprise knowledge and tools through reusable interfaces; and manage task workflows with traceable oversight.

4 Global AI Pulse: Q1 2026 report, KPMG
https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2026/04/global-ai-pulse.pdf.coredownload.inline.pdf
5 The state of Enterprise AI, OpenAI
https://openai.com/business/guides-and-resources/the-state-of-enterprise-ai-2025-report/

02 Business application layer: Establishing sustainable closed-loop systems to accelerate token-to-value conversion

- **Token value realisation within business applications critically depends on operational closed-loop systems.** Consider AI-assisted software development – a high-frequency iterative scenario. The 2025 DORA Report: State of AI-assisted Software Development, drawing upon 100+ hours of qualitative research and insights from nearly 5,000 global technical professionals[6], reveals AI primarily acts as an amplifier in development processes: intensifying performance advantages in high-functioning organisations while exacerbating inefficiencies in underperforming teams. Google Cloud's analysis[7] of the report further demonstrates that while AI adoption positively correlates with software delivery throughput and product performance, it retains an inverse relationship with delivery stability. Crucially, without automated testing, mature version control, and rapid feedback mechanisms, increased change volume amplifies downstream stability risks. This evidence confirms that whether directly invoking model tokens or utilising encapsulated intelligent services, integration within controllable operational frameworks is essential to transform invocation costs into consistent, sustainable business value.

- **The evolution from isolated assistance to sustainable closed-loop systems establishes real-time intelligence as the defining characteristic of the business application layer.** These operational frameworks enable AI agents to continuously comprehend task contexts during execution, determine subsequent actions informed by business knowledge, and dynamically adjust outcomes based on performance feedback as AI penetrates core enterprise processes. This transformation has shifted business intelligence from passive response to proactive collaboration, wherein real-time intelligence evolves from a value-added enhancement to a fundamental requirement for closed-loop value realisation. Consider these implementations: Customer Service: AI must continuously contextualise customer profiles, interaction histories, knowledge repositories, authorisation protocols, and ticket statuses; Operations Maintenance: AI requires integration of alerts, system logs, network topologies, change records, and resolution actions into unified diagnostics; Cross-System Automation: AI orchestrates information extraction, task delegation, status synchronisation, and exception handling across multiple business platforms. Ultimately, token value realisation pathways will extend to real-time perception, comprehension, response, and refinement cycles addressing business conditions, user requirements, process evolution, and execution outcomes.

03 Runtime control layer: Governing AI complexity for sustainable operations and process integrity

- **As enterprise AI applications accelerate their evolution towards Agentic AI, operational complexity will escalate substantially.** This complexity stems not merely from model inference requirements, but from extended task chains, expanding context scales, deepening execution actions, and increasing collaborative entities. When applications progress to large-scale deployment—potentially forming AI agent networks—concurrent requests, system integration demands, and continuous service obligations compound operational challenges. Consequently, enterprises face increasingly dynamic, unpredictable runtime issues with critical traceability requirements. Gartner[8] forecasts that beyond 40% of Agentic AI initiatives may face cancellation by late 2027 due to escalating costs, ambiguous business value, or inadequate risk governance.

- **The runtime control layer is emerging as a critical nexus for enterprise AI systems transitioning towards large-scale, stable operation.** This layer furnishes a unified control surface for AI task execution spanning models, tools, and business systems. Core requirements for enterprise-grade production scenarios encompass ensuring continuous operation and process governance. Continuous operation necessitates the stable progression of AI task chains, preventing state disruptions, context deviations, and execution anomalies that could impede task completion. Process governance mandates that AI execution workflows remain monitorable, continuously evaluable, and optimisable, thereby guaranteeing that model invocations, tool executions, and result feedback consistently align with business objectives. Consequently, the runtime control layer not only integrates token usage, task progression, and business outcomes within a unified control framework, but also establishes the procedural foundations for subsequent security auditing, accountability tracing, and policy refinement.

6 2025 DORA Report: State of AI-assisted Software Development, DORA
https://services.google.com/fh/files/misc/2025_state_of_ai_assisted_software_development.pdf
7 Announcing the 2025 DORA Report: State of AI-Assisted Software Development, Google Cloud
https://cloud.google.com/blog/products/ai-machine-learning/announcing-the-2025-dora-report
 Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, Gartner
https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027?utm_source=chatgpt.com

Table 5 Critical initiatives to empower upper-layer productivity from the perspective of Token economics

	Objectives	Technical focus	Assessment mechanism
Model service layer	• Establish a standardised, service-oriented, and governable intelligent service supply system focused on enhancing the efficiency of converting unit tokens into enterprise-grade intelligent services.	• **Standardisation:** Create a model registry, capability taxonomy, and unified evaluation benchmarks to define capability boundaries, applicable scenarios, quality metrics, cost tiers, and compliance requirements across models. Compile prompts, contextual parameters, and knowledge components for high-frequency tasks into reusable templates. • **Service Orientation:** Implement model routing, AI Gateway, RAG services, and caching mechanisms to encapsulate models, knowledge bases, tool interfaces, and contextual resources into intelligible services reliably invoked by business systems. • **Governability:** Incorporate permission controls, call auditing, data access governance, and policy enforcement within model service workflows. This ensures model invocations, knowledge retrievals, and tool usage operate within authorisation boundaries, with traceability and verifiability at critical operational stages.	• **Service Utilisation Efficiency:** Monitor model service reuse rates, task template adoption, knowledge component efficacy, and cross-scenario migration performance. Assess whether token expenditure has catalysed organisation-level capability development. • **Service Governance Efficiency:** Track call compliance, audit trail completeness, high-risk intervention efficacy, and policy optimisation feedback. Determine whether intelligent services operate within prescribed parameters while continuously refining invocation strategies.
Business application layer	• Establish a sustainable closed-loop system with real-time intelligence capabilities – including perception, comprehension, response, and feedback – centred on business value realisation.	• **Real-time business status perception:** Enhance real-time data ingestion and business event triggering capabilities, enabling AI systems to continuously interpret operational signals such as customer requests, work order statuses, R&D modifications, operational alerts, process milestones, and historical records. • **Real-time task context comprehension:** Structure business data, organisational knowledge, procedural rules, and historical context into model-interpretable task frameworks. This empowers AI to discern user intent, operational boundaries, and subsequent processing directions. • **Real-time process adaptation:** Integrate AI with business process systems, authorising it to generate recommendations, invoke tools, initiate process progression, or trigger manual confirmations within defined parameters. • **Real-time execution feedback:** Implement mechanisms for result validation, manual correction, anomaly fallback, and operational observability. Subsequently, channel execution outcomes, user feedback, and manual interventions back into process and knowledge systems to form a continuous optimisation cycle.	• **Business value delivery efficiency:** Monitor task completion rates, process acceleration metrics, manual intervention frequency, per-successful-task costs, and tangible business outcomes. Correlate real-time intelligence functions to verify whether token expenditure translates into measurable business value. • **Token operation & measurement framework:** Develop internal token accounting, cost allocation, and value attribution frameworks. Employ tokens as the primary quantifier for intelligent service pricing, resource distribution, and computational resource management.
Runtime control layer	• Prioritise dynamic equilibrium between marginal benefits and marginal costs as the core principle, preventing token expenditure from transitioning from value creation to non-value-adding expenditure.	• **Controlled execution environment:** Utilise Agent Harnesses, Agent Runtimes and similar mechanisms to constrain model invocations, tool execution, and context management boundaries. This prevents disorderly calls, context drift, and erroneous execution that inflate token consumption. • **Dynamic task orchestration:** Employ Agent Orchestration to dynamically determine the necessity of continued reasoning, tool invocation, agent collaboration, or manual intervention based on task progression. This ensures execution chains terminate optimally. • **Continuous observability & evaluation:** Implement tracing, evaluation frameworks, and AgentOps capabilities to monitor task states, token expenditure, failure root causes, and business outcomes. Identify unproductive consumption to iteratively refine model routing, orchestration strategies, and governance protocols.	• **Marginal benefit optimisation:** Focus on enhancing task completion rates, improving result accuracy, reducing process milestones, saving manual processing labour, and increasing anomaly resolution efficiency. • **Marginal cost containment:** Continuously quantify incremental token consumption, response latency, redundant invocations, tool execution failures, manual intervention costs, and unmitigated risk exposure. Identify hidden cost drivers within task workflows. • **Dynamic convergence protocol:** Trigger model switching, process termination, manual escalation, or task abandonment when marginal returns diminish and marginal costs escalate. This prevents token expenditure from devolving into economically unjustifiable overhead.

Source: KPMG analysis

3.2.3 Organisational governance and talent capabilities: Reconstructing enterprise operational models

The foundation of enterprise infrastructure security governance lies in identifying entities requiring protection, constraint, and monitoring. As AI evolves from passive tooling to autonomous agents actively advancing tasks, systems – including Agentic AIs, AI workflows, and multi-tool applications – now orchestrate data, models, tools, and business processes into continuous execution chains. Consequently, token flows as intelligent operational vectors directly impact business continuity.

Enterprises must therefore synchronously establish a. Security control points across infrastructure layers, governed by access, invocation, and execution policies implementing full-stack security principles; b. Observability collection points detecting operational states, risk indicators, and business impacts through full-stack observability. These complementary mechanisms constitute an engineering substrate enabling secure, reliable AI operation at scale.

01 Full-stack security: Implementing intelligent execution boundaries across the entire infrastructure stack

- **Enterprise AI security governance is evolving towards a layered control system that spans the full infrastructure stack. Tokens continuously traverse context concatenation, knowledge retrieval, model invocation, tool triggering, and business actions**, crossing multiple layers including energy and infrastructure, compute, networking, storage resources, model services, business applications, and runtime controls. Consequently, security controls must also be layered throughout the infrastructure stack. This capability directly impacts an enterprise's confidence in deploying AI within genuine business processes. The "Guidelines for the Construction of Artificial Intelligence Security Governance Standard System in the Industrial and Information Technology Field (2025)" incorporate governance capabilities, infrastructure security, cybersecurity, data security, algorithmic model security, and application security into a unified framework. This clearly indicates that AI security governance is adopting a fully stacked and systematic characteristic. Emerging AI applications, such as AI agents, further underscore the necessity for full-stack security. The "Implementation Opinions on the Standardised Application and Innovative Development of AI Agents" define AI agents as intelligent systems possessing autonomous perception, memory, decision-making, interaction, and execution capabilities. These opinions mandate the development of full-cycle security standards covering the development, deployment, application, and maintenance of AI agents. Furthermore, they require strengthened security management for model access, application programming interface (API) calls, and the use of extension tools.

- **The implementation of full-stack security hinges on defining precise security boundaries across infrastructure layers.** Specifically, the energy and infrastructure layer safeguards operational continuity by mitigating fundamental environmental risks—including power supply, cooling, physical space constraints, and edge node vulnerabilities—to minimise service disruption. The resource layer's core objective is to ensure collaborative resource operation within trusted parameters, preventing resource abuse, data commingling, and risk propagation. At the model service layer, security emphasis shifts to the model call chain itself; enterprises must regulate model access, context inputs, knowledge retrieval, and tool interfaces to preclude erroneous information or sensitive data from infiltrating inference processes. The business application layer must delineate scenarios where AI may execute tasks autonomously, assist in judgement, or require manual validation. Finally, the runtime control layer functions as a safety net, deploying approval workflows, human takeover protocols, and traceability mechanisms to maintain controllability, auditability, and accountability throughout continuous AI operations.

02 Full-stack observability: End-to-end visibility across resources, operations, and business functions

- **Full stack observability entails correlating disparate signals from diverse infrastructure layers within unified intelligent execution chains.** For enterprise AI systems, observable entities extend beyond conventional metrics like server status, interface calls, and application logs to encompass token flow efficiency, cost, quality, and risk exposure. Only by integrating signals from energy and infrastructure, compute-networking-storage resources, model services, business applications, and runtime controls can enterprises: a) identify critical factors impacting AI system performance, cost efficiency, and stability; b) evaluate whether tokens are effectively converted into tangible business value. Contemporary full-stack observability practice has evolved beyond single-point system monitoring towards cross-domain signal correlation. This approach unifies applications, networks, infrastructure, and external dependencies through integrated platforms, whilst explicitly mapping performance indicators, quality metrics, cost analyses, and security risks to business outcomes.

- **Full-stack observability necessitates the concurrent development of resource visibility, operational visibility, and business visibility.** Resource visibility provides the foundational capability for enterprises to assess AI efficiency, prioritising the identification of fundamental constraints governing token generation and delivery; crucially, in high-frequency calls and long-chain tasks, any resource-layer bottleneck manifests directly as increased token latency or wasted computational expenditure. Operational visibility reconstructs end-to-end AI task execution pathways to quantify token consumption across workflow stages, diagnose faults and performance bottlenecks, and determine critical factors affecting task quality and execution efficacy. Business visibility delivers the ultimate validation of AI's commercial contribution: where sustained growth in token consumption fails to correlate with improvements in task completion rates, process efficiency, or output quality, this signals systemic inefficiencies such as overextended call chains, inadequate scenario adaptation, or broken execution feedback loops.

Table 6 **Critical initiatives to build a security governance system from the perspective of Token economics**

	Objectives	Technical focus	Assessment mechanism
Full-stack security	• Define security boundaries throughout the infrastructure stack. Translate unified security policies into layered, proximate, and enforceable control points. This mitigates risk propagation and reduces governance overhead during large-scale AI deployment.	• **Security Asset Inventory:** Catalogue models, data, knowledge bases, tools, agents, and business actions to identify assets requiring protection, constraints, and traceability. • **Layered Control Points:** Embed appropriate mechanisms (access control, resource isolation, data protection) within relevant infrastructure layers (energy, infrastructure) according to security boundary requirements. Ensure security controls execute as close as possible to the source of risk, preventing the need for centralised remediation after escalation. • **Reusable Policy Templates:** Formalise security requirements for high-frequency scenarios into reusable policy templates. These support standardised online evaluation, exception handling, and compliance auditing.	• **Layered Security Boundary Assessment:** Verify clear definition of protected assets, constrained behaviours, and traceability stages at each level. • **Proximity Enforcement Capability Assessment:** Evaluate whether risks can be promptly identified and mitigated at the point of occurrence (AI workload execution, model invocation, tool triggering, business action initiation). • **Governance Efficiency Evaluation:** Measure reductions in: AI application launch evaluation cycles; frequency of security reassessments; proportion of manual approvals; incident resolution time; audit evidence collection duration; and compliance remediation expenditure.
Full-stack observability	• Advance observability capabilities from isolated system monitoring to end-to-end interpretation mechanisms spanning resources, operations, and business functions. This reduces the time and cost associated with risk localisation, cost allocation, and continuous optimisation during AI operations at scale.	• **Resource Observability:** Correlate key operational signals across infrastructure, model services, application systems, and security events to establish a unified observable framework. • **Operational Observability:** Trace the lifecycle of AI tasks from initiation through model invocation, tool triggering, human intervention points, to exception rollback. This enables precise fault attribution to resource provisioning, model responses, tool execution, or business processes. • **Business Observability:** Integrate operational status, cost fluctuations, risk incidents, and task outcome effectiveness. Establish automated alerting, root cause analysis, audit evidence capture, and continuous optimisation mechanisms.	• **Cross-Domain Signal Integration Assessment:** Determine whether key operational signals (infrastructure, networks, model services, applications, security events, cost consumption) are centrally aggregated and correlated. • **Intelligent Execution Diagnostics Assessment:** Evaluate if organisations can rapidly diagnose causes of AI task degradation, failure, cost escalation, or risk triggers using contextual data (model calls, token consumption, response latency, failure rates, tool execution, exception events). • **Optimisation Loop Evaluation:** Assess whether performance, cost, risk, and business outcomes can be analysed within unified workflows, while progressively reducing timelines for anomaly detection, root cause identification, manual intervention, and cost allocation.

Source: KPMG analysis

Chapter 4

Leading Practice Insights

Case Study 1

Strengthening AI-ready foundations for quick-wins in semiconductor manufacturing

Endpoint intelligence and escalating process complexity: Driving scenario planning and organisational capability development

By advancing end-point application demands and increasing the adoption of AI, semiconductor manufacturers are moving towards intelligent production. Imaging algorithms that serve end-point applications—including smartphones, automotive CIS (CMOS Image Sensors), smart homes, and industrial vision systems—are pushing continuous enhancement in areas such as image sensor resolution and power efficiency consistency, stability, and reliability. This means wafer fabrication processes face heightened requirements in process control precision, yield stability, defect identification, and production collaboration efficiency.

Take the example of a chip manufacturer, strategically expanding its mid-to-high-end product portfolio to consolidate its market position. It recognises the growing importance of leveraging AI for optimised production schedules, better yield rates, and quicker process development cycles. The company prioritised planning for potential scenarios and organisational capability development as the main routes for transformation, involving targeted AI implementation to secure quick wins.

- **Planning AI Application Scenarios:** The company focused on its core business functions—including R&D, manufacturing, quality, and supply chain—and adopted a top-down approach that involved analysing end-to-end processes to identify high-value scenarios, such as yield analysis, defect detection, process parameter optimisation, equipment predictive maintenance, production scheduling optimisation, and supply chain risk prediction. An AI implementation standard was established that was fully verifiable and which covered the full cycle of business decomposition, scenario identification, capability alignment, and value assessment. A standardised, tiered (L1-L5) business process framework was also implemented, with quantitative ROI analysis as the basis of decision-making. By making clear the requirements regarding data, systems, computation, processes, and organisational capabilities across AI scenarios, it provided distinct short-, medium-, and long-term implementation roadmaps.

- **Building AI Organisational Capacity:** A robust organisational capability is fundamental to the successful implementation of AI. A logical organisational structure defines key roles, including AI Product Managers, AI Architects, Algorithm Engineers, Data Engineers, and Business Scenario Leads. Meanwhile, by developing an AI cost-benefit model, personnel expenditure, tooling costs, computing resource investments, implementation timelines, projected returns, and risk control measures can be built into a unified evaluation framework. Formal mechanisms for scenario initiation, resource allocation, model validation, operational deployment, and benefits realisation review help establish the organisational foundations necessary for subsequent large-scale AI deployment.

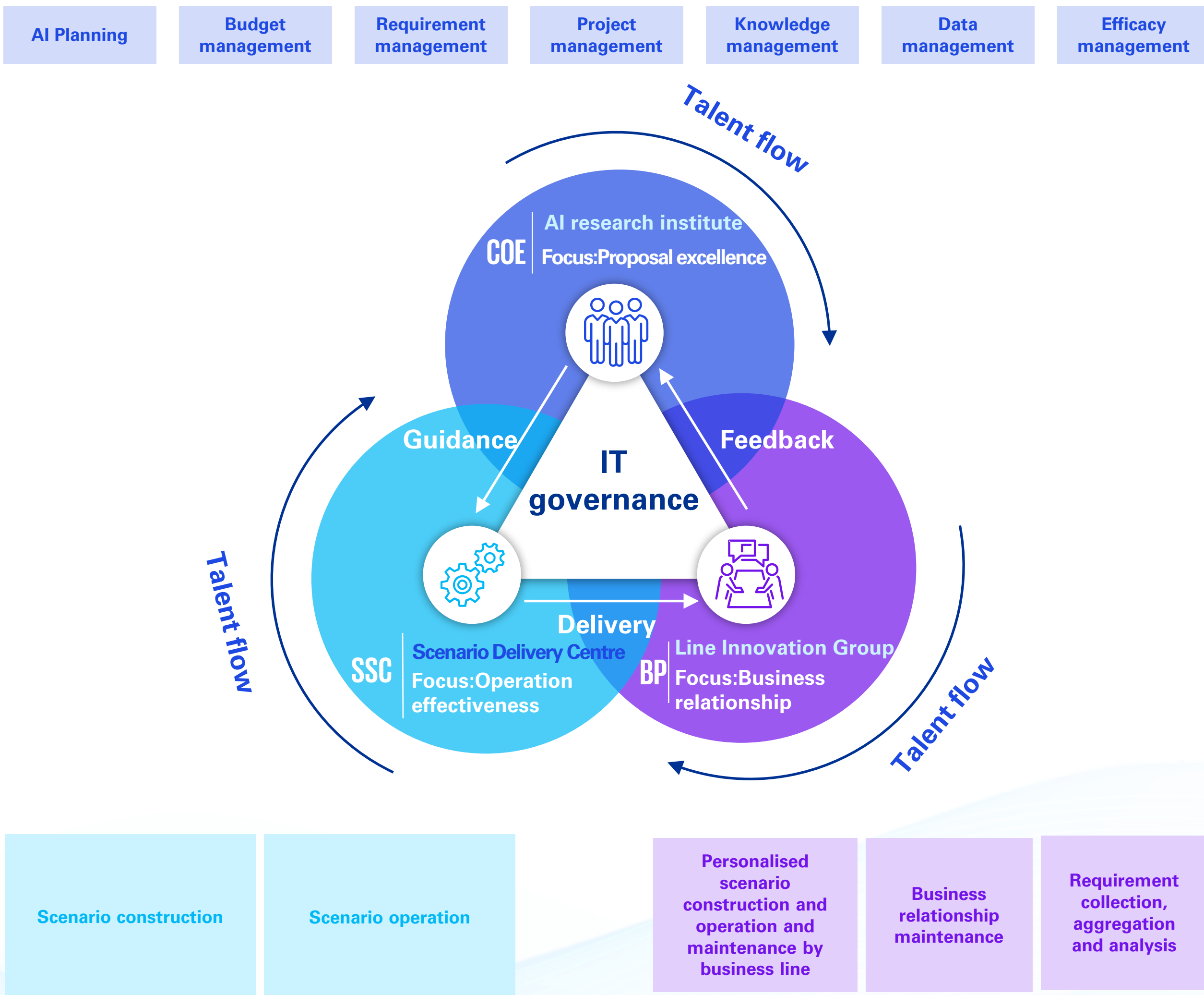
The enterprise established an AI Centre of Excellence (COE), jointly driven by the AI Research Institute, the Scenario Delivery Centre, and Line Innovation Groups to advance AI transformation. Within this framework:

The AI Research Institute is responsible for the enterprise's overarching AI strategic planning. This involves coordinating the governance and oversight of AI budgets, projects, requirements, data, knowledge, model efficacy, and security.

The Scenario Delivery Centre undertakes the construction and operational management of AI scenarios developed alongside the AI Research Institute, progressively assuming responsibility for scenario operations across all business lines.

Line Innovation Groups act as operational intelligence partners for regional and business-specific units. Their remit covers managing stakeholder relationships, gathering and analysing their requirements, and constructing and operating custom scenarios for the respective regions and business lines.

Figure 8 **AI COE organisational structure**



Source: Public information, KPMG analysis

How legacy technical architecture hinders full-scenario smart manufacturing; infrastructure modernisation is imperative

During implementation, the AI COE took business actions and converted them into AI tags based on AI capabilities, creating basic AI-enabled analytical units. These actions would be further reduced into executable AI scenarios using the AI tags, leading to applications with discrete business value. However, as AI scenario prioritisation and resource allocation decisions became fixed in place, the expert group realised that the current technical architecture was not capable of supporting the continuous evolution of full-scenario smart manufacturing.

From a scenario development perspective, the enterprise previously achieved basic automation in certain production and quality scenarios by using disparate computational resources and machine learning techniques. However, with the proliferation of vertical-domain small language models, AI agents, and multimodal algorithms, the existing architecture was impeding rapid algorithm iteration, efficient data flows, and cross-scenario reusability. A transition from traditional deep learning to a hybrid architecture – integrating general-purpose and specialised models – was urgently required.

From the perspective of improving manufacturing capabilities, semiconductor fabrication is capital-intensive, complex, and yield-sensitive. As technology advances, production environments face escalating data density, process variables, and equipment synchronisation complexity. Consequently, the IT department must establish higher-specification intelligent infrastructure that enables high-precision control of production processes.

As a result of all this, the AI COE had gained a clear understanding that existing data centres were not yet fully AI-ready due to deficiencies in GPU resource pooling, network bandwidth, and low-latency connectivity. These limitations constrain the parallel deployment of multiple AI scenarios – including R&D simulation, industrial visual inspection, production data governance, quality traceability, and intelligent workplace solutions. In addition, ongoing large-scale capacity expansion generates substantially greater volumes of data on production, equipment, and quality. The intelligent management of this data, through AI computing power and network infrastructure – integrated within a systematic framework for long-term AI evolution – is the critical path to sustaining manufacturing efficiency, yield performance, and maintaining a competitive advantage.

Enabling collaborative training, inference, and smart manufacturing scenarios through AI-ready infrastructure

Increasing the scale of AI training, inference, and smart manufacturing scenarios naturally leads to heightened demands on network bandwidth, low-latency interconnectivity, computational resource orchestration, data access efficiency, and operational stability. At the data centre level, this necessitates establishing a pre-emptive, scalable, observable, and governable AI-ready foundation.

As part of its solution to this problem, the enterprise adopts a collaborative approach to developing network, computational, and operational maintenance, and security capabilities. This is guided by the strategic objective: "efficient computation deployment, accelerated data flows, and sustained system controllability":

- **High-performance data centre networks** provide high-bandwidth, low-latency data paths between GPU clusters, servers, storage, and business systems. This significantly reduces the transmission latency experienced during large-scale training, parallel computing, and high-concurrency inference processes. Solutions such as Cisco Nexus deliver advanced network capabilities, forming the foundation for efficient intra-cluster data transfers and the evolution of potential high-speed interconnectivity within AI clusters.

- **Unified computation platforms** deliver a stable computational capacity for diverse AI tasks. This enables multiple scenarios—such as research and development, manufacturing, quality assurance, and office functions—to be incrementally deployed on top of a shared resource base. Utilising platforms such as Cisco UCS, the enterprise gains full access to standardised, scalable computing resources for AI training, inference, and data processing.

- **Cloud-based operations and maintenance (O&M) management capabilities** transform static infrastructure into always-on operational assets. This empowers the enterprise to better plan capacity, allocate resources, diagnose faults, and manage process lifecycles. Leveraging management solutions such as Cisco Intersight, allows for improved capabilities for cross-environment resource orchestration, system visualisation, and O&M automation.
- **Integrating cloud-native security and observability** into the infrastructure provides the enterprise with a comprehensive visibility and governance loop than spans network connectivity, workload operations, application interactions, and security incidents.

Key takeaways:

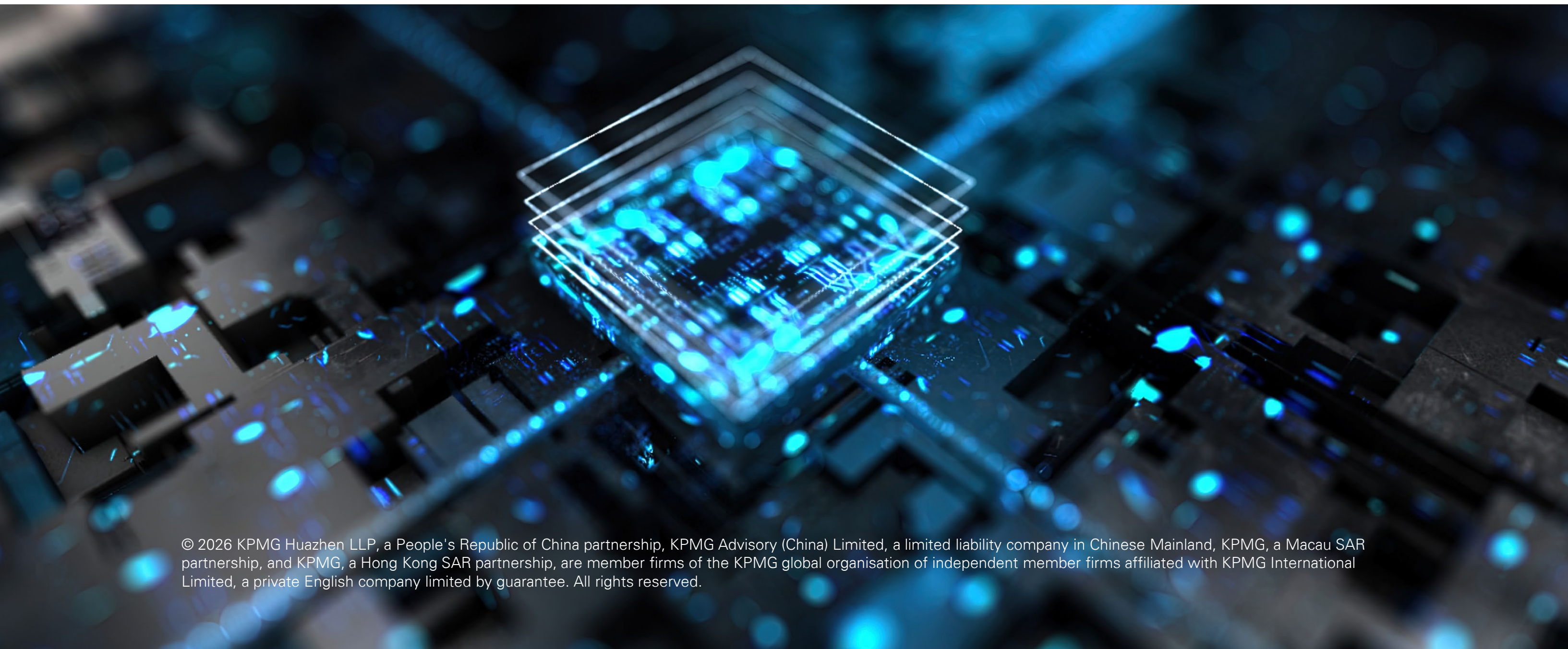
Achieving AI quick wins does not only require scenario breakthroughs but also synergy between scenarios, organisations, and infrastructure

An AI quick win strategy does not mean ignoring the development of a foundational infrastructure. For the semiconductor manufacturer in this case study, true transformation – where AI evolves from a point-solution tool for improving efficiency into an enduring, strategic capability basis for manufacturing enhancement and core competitiveness – only occurs when business scenarios, organisational mechanisms, and the AI technological base are closely interconnected.

By confronting increased product performance demands and heightened process complexity driven by end-point intelligence, the enterprise proactively advances AI scenario planning and organisational capability building. This prevents its AI transformation from stalling across fragmented pilot projects and during technical validation. Through collaboratively integrating its AI Research Institute, Scenario Delivery Centre, and production-line innovation teams, it embedded strategic planning, business innovation, scenario implementation, and operational reviews within a unified framework. This system provides robust support for full-scope intelligent manufacturing.

Systematic assessment by the AI COE that evaluates scenario requirements, data growth patterns, and infrastructure limitations revealed that the existing data centre architecture could neither support the continuous expansion of AI scenarios nor accommodate escalating manufacturing data volumes.

The main advantage of upgrading the AI-ready infrastructure lies not in being able to carry out hardware replacements, but in establishing a sustainable foundation for subsequent AI scenario expansion, model service deployment, industrial data circulation, and intelligent operations. AI applications are evolving from localised quick-win scenarios to mission-critical processes, including in production, R&D, and quality management. This means having an infrastructure that possesses high-performance interconnectivity, unified resource management, continuous operational oversight, and security observability is becoming indispensable. This is fundamental for the enterprise to maximise the value of AI , mitigate large-scale implementation risks, and sustain a long-term competitive advantage.



Case Study 2:

Systematic capability development – From an AI application blueprint to resource foundation in asset management

Differentiated development demands drive AI transformation amid persistent LLM implementation constraints

As part of the continuous optimisation of computing resources and dynamic advancements in AI technologies, large language model (LLM)-centric innovations within the financial sector are beginning to emerge, evolving and iterating at unprecedented speeds. The deployment of AI scenarios by financial institutions is based on strategic imperatives and value-driven approaches. This is advancing LLM capabilities beyond point solutions and towards embedded business processes, enhanced professional competencies, and organisational efficiency gains.

Unlike banks and securities firms, which undertook earlier systematic exploration, asset management institutions (including funds, wealth managers, and trusts) focus on implementing scenario-specific tools as part of their LLM initiatives. They are behind their peers in establishing comprehensive capability frameworks that cover strategy, scenarios, data, models, computation, and governance. One asset management company – having navigated an institutional restructuring, volatility in its assets under management (AUM), intensifying homogeneous product competition, and deepening industry pressures, over the course of five years – confronted the differentiated development requirements of its next phase of strategic development. Using a top-down strategy, it sought to crystallise its digital transformation trajectory.

To ensure strategic alignment, the company first defined the AI application maturity stages, potential opportunities, and implementation constraints within asset management by using sector benchmarking, trend analysis, and cross-functional stakeholder engagement. By examining industry best practices of leading financial institutions, three primary AI transformation vectors were identified:

- **Make us of professional expertise:** Take legacy capabilities based on specialist experience – including investment research analysis, asset allocation strategies, product interpretation, and client engagement – and turn them into reusable knowledge assets via LLMs. This democratisation of expertise operationalises intelligence across business functions and client touchpoints, elevating the company's overall professional services level.
- **Embed Augmented Intelligence:** Deeply integrate AI within financial operations and management workflows to assist personnel in generating material, retrieving information, pulling together research, analysing portfolios, identifying risks, checking compliance, and processing transactions. This reduces workloads by minimising repetitive tasks, freeing core personnel to focus on high-value work and strategic intervention.

- **Transform Client Engagement Models:** As client expectations evolve from standardised services towards personalised, dynamic, real-time services, leverage AI to enhance customer understanding, precision targeting, product-fit assessment, and sustained relationship-building.

As part of this process, 'data refinement' and 'data collaboration' are progressively emerging as the core competitive advantages of AI applications. At the same time, advancements in regulatory technology ('RegTech') and enhancements to governance frameworks provide crucial support for the ongoing deployment of LLMs within financial scenarios.

However, the company also realised that the development of LLM applications continued to face multiple constraints:

- **Data fragmentation and insufficient corpus supply:** High-value data resources remained scattered and siloed over the long term across multiple business functions, such as investment research, product development, channels, risk control, compliance, and operations. This fragmentation leads to a deficiency in the supply of a large-scale, high-quality corpus essential for effective model training and scenario implementation.

- **Unclear strategic direction and business value verification:** Strategic planning for LLM adoption remained ambiguous, and the pathway for validating its business value is often undefined. This lack of clarity resulted in uncoordinated scenario selection, duplicated resource investments, and misaligned development priorities.

- **Low business fault tolerance requires high model constraints:** The asset management business is characterised by low fault tolerance, stringent regulatory oversight, and high professional standards. Scenarios such as investment research analysis, product sales, customer service, and risk control demand exceptionally high levels of accuracy, interpretability, and controllability in model outputs. Therefore, LLMs implemented should exhibit low hallucination, strong constraint, and full traceability.

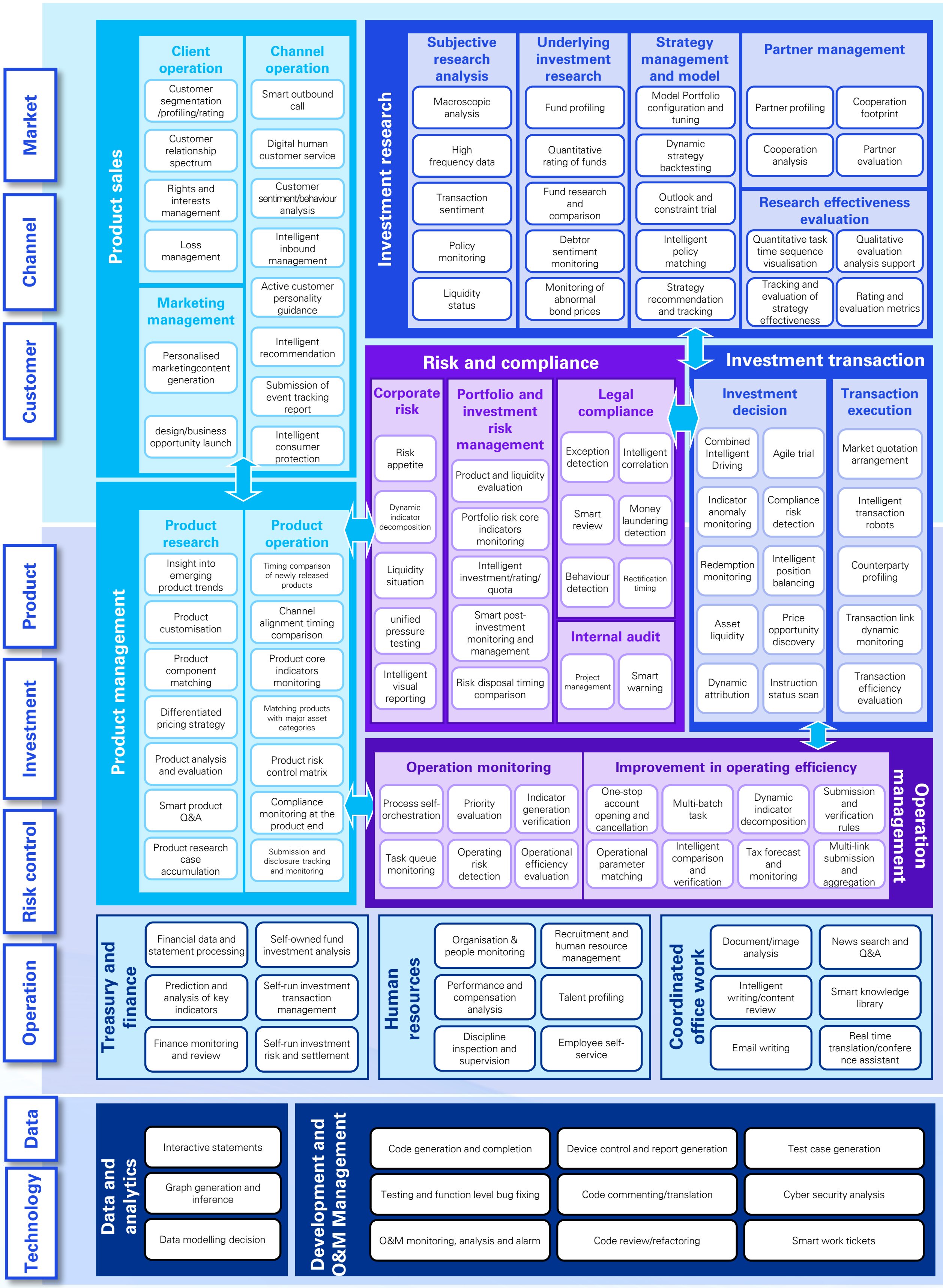
- **Rapid technological iteration necessitates continuous organisational evolution:** LLM technology is undergoing rapid iteration. This pace imposes demands on organisational structures, talent composition, technology selection strategies, and governance capabilities.

Formulating the AI application blueprint: Strengthening systematic development from business scenarios to foundational resources

Based on a comprehensive assessment of industry trends, sector practices, and its own organisational capabilities, the company resolved to roll out an AI application blueprint for systematic AI capability development, from business scenarios through to foundational resources. Measures included conducting digital maturity assessments and AI application blueprint design across core business processes; creating the technical and data architecture designs required to operationalise the AI blueprint; and implementing complementary digital governance frameworks alongside AI ethics and security protocols.

During blueprint formulation, input from business departments regarding operational pain points, process bottlenecks, and capability gaps served a dual purpose: they provided critical insight for identifying high-impact AI opportunities and acted as a key validation benchmark for the blueprint's practical ability to be implemented. The company employed a rigorous 30-dimensional screening framework across key functions – including products, investment research, investment management, risk control, operations, and technology – to evaluate potential AI applications; the prioritisation of scenarios based on strategic value, technological feasibility, return on investment, and specific LLM risk mitigation capabilities; and to finalise the blueprint based on the above efforts.

Figure 9 AI Application scenarios construction blueprint



Source: Public information, KPMG analysis

When deploying resources for the AI application blueprint, the asset management company identified scenarios such as investment research analysis, risk identification, customer insight, knowledge retrieval, and intelligent operations, that were more demanding in terms of model service responsiveness, data access stability, and task completion efficiency. Furthermore, as the sector required low fault tolerance, high regulatory compliance, and high professionalism, model capabilities cannot be the only foundation for successful AI implementation. The depth to which AI can penetrate core business processes also depends on whether models can comprehend authentic business contexts; whether their outputs are auditable, interpretable, and constrained; and whether robust computational, network, and storage infrastructure are available in support of applications.

Scaling GPU clusters, increasing storage access frequency, and elevating training task parallelism invariably intensifies network demands. This directly impacts computational resource utilisation, task completion latency, and overall system stability. Rapidly expanding storage systems makes this pressure even worse, as substantial data I/O and model training traffic converge. Therefore, network bandwidth, low-latency transmission, effective congestion control, and intelligent traffic scheduling become critical determinants of AI task efficiency.

As a result, the company's existing 200G network architecture was no longer sufficient is demonstrably insufficient to accommodate future AI data centre expansion requirements. Upgrading this network infrastructure was essential for the company to move has therefore become an essential infrastructure initiative for the company's transition from AI blueprint design to seamless large-scale implementation.

Enabling financial-grade model training and intelligent service scaling with high-performance AI network infrastructure

In building its next-generation AI data centre network, the company selected 800G high-performance Ethernet for network capacity expansion, so as to enhance data flow efficiency between GPU clusters, storage systems, and model services through optimised bandwidth, reduced latency, and granular traffic scheduling capabilities.

For large-scale distributed training and high-concurrency inference workloads, the company chose RoCEv2's RDMA capability to establish a high-throughput, low-latency data transmission foundation within the Ethernet ecosystem. This mitigates communication bottlenecks during model training, parameter synchronisation, and distributed storage access. It also leverages dynamic load balancing, microsecond-level traffic scheduling, and a multi-plane network architecture to significantly diminish localised congestion impacts on task efficiency – transforming the network from a connectivity layer into an efficiency adjustment layer for AI operations.

To address repetitive context storage demands from AI agent applications, the company has evolved its optimisation paradigm from "computation expansion" to "context reuse". This keeps computed context states within reusable memory hierarchies. Through tiered storage, shared access, and cache-locality-aware scheduling, repetitive computation overhead is converted into predictable storage-access and network-transport costs. The company's expected ROI analysis – benchmarked against public datasets and test cases – indicates VAST's KV Cache acceleration can increase TPS by 60%-130%.

Table 7 **Token throughput and cost efficiency improvement enabled by KV Cache acceleration**

Hit Rate（x）	Average Request Time (sec)	Effective System TPS	% Increase in TPS
0%	62.0	2,064	0%
40%	38.4	3,333	62%
60%	26.0	4,812	133%

Note: This calculation employs a prudent baseline assumption for enterprise AI workflows, with cache hit rates of 40-60%. Notably, AI agent tasks and code generation scenarios offer potential for further optimisation as their context reuse is inherently higher.

The core value of the new architecture lies in its comprehensive support capabilities, encompassing high-performance AI networking, ecosystem compatibility, validated architecture design, and observable operations:

- With the underlying Cisco Nexus 9000 series high-density 800G switching platform, the company established a high-speed network foundation for the horizontal scaling of GPU clusters.
- Complementary capabilities such as RoCEv2, dynamic load balancing, telemetry, and open APIs further enabled the network to accommodate AI traffic, and the operational team to more quickly identify states, analyse capacities, and locate problems.
- The proven reference architecture can work properly with mainstream GPUs, network interface cards, and the AI infrastructure ecosystem, helping reduce integration complexity and implementation risks during network upgrades.

By evolving from a 200G to an 800G network architecture, the company has established a more flexible, manageable, and cost-effective resource foundation for the expansion of AI applications over the next three years. This evolution supports the continuous scaling of diverse AI scenarios, underpinned by stable data flows, efficient computing release, and enhanced operational visibility.

Key takeaways:

Strategic AI transformation – Building enterprise competitiveness from an application blueprint to an operational foundation where small changes can have long-term impacts

AI is playing an increasingly important part in building competitive edge in the business world. For the asset management company introduced in this case study, pressures for competitive differentiation served as the primary driver for AI transformation. In order to translate AI into a sustainable competitive edge, the company should follow its comprehensive AI application blueprint, covering data, computation, network infrastructure, and governance architecture to systematically, scalably, and sustainably build capabilities.

The company doesn't view AI transformation only as the adoption of LLM tools or isolated proof-of-concept pilots. Its AI application blueprint not only identifies which scenarios merit investment but also enables the company to pinpoint weaknesses in the underlying data foundations, model capabilities, technical architectures, and resource bases required for sustained scenario operation. Growing demands for model training, fine-tuning, high-concurrency inference, and distributed storage access are leading to deficiencies in network, storage, compute, and other foundational resources, directly constraining the responsiveness, operational stability, and scalability of AI applications. Therefore, a high-performance, observable, and scalable network architecture is becoming essential for financial institutions seeking to unlock the value of AI applications, mitigate large-scale implementation risks, and foster differentiated competitiveness.

Case Study 3

A global wealth management group's overseas infrastructure modernisation under its "AI + Business Integration" strategy

As the external environment reshapes the competition landscape in the financial sector, AI transformation has become the Group's strategic priority

As the external environment reshapes the competition landscape in the financial sector, AI transformation has become the Group's strategic priority

The global financial sector has been navigating an increasingly complex business cycle since 2025. Intertwined factors – including trade uncertainty, geopolitical instability, asset price volatility, shifting interest rate trajectories, and sovereign debt pressures – have intensified challenges across investment management, risk identification, client servicing, and operational resilience. At the same time, a new wave of AI technological evolution, spearheaded by LLMs and AI agents, is enabling the financial sector to move beyond AI-informed decision-making to AI-powered autonomous actions.

In the sector, client expectations and competitive dynamics are changing as well. Clients increasingly demand real-time insights, personalised asset allocation, seamless cross-regional service experiences, and greater advisory service transparency. Meanwhile, market competition is transitioning beyond brands and channels towards comprehensive capabilities in technological prowess, customer data utilisation, product innovation, and risk governance.

Amidst accelerating AI advancements, profound shifts in client demographics, and expanding global operations, the Group's leadership recognised that AI transformation constitutes a strategic imperative essential for attracting high-net-worth clients, enhancing advisory capacity, strengthening global investment research capabilities, optimising operational costs, and fulfilling compliance obligations. They understood the urgency of a fundamental reassessment of AI's strategic role across business growth, organisational collaboration, service delivery models, and infrastructure readiness.

In advancing its latest intelligent transformation initiative, the group set "AI + Business Integration" at the heart of its AI capability development for the coming five-year period. The Group believes that it must comprehensively embed AI – particularly next-generation intelligent technologies epitomised by LLM – and position it as the core driver for reengineering business processes, innovating client services, optimising risk management, and increasing the efficiency of internal operations.

Based on this vision, the group has established a clear AI strategic blueprint, which focuses on consolidating three foundational pillars while enhancing the supporting mechanism.

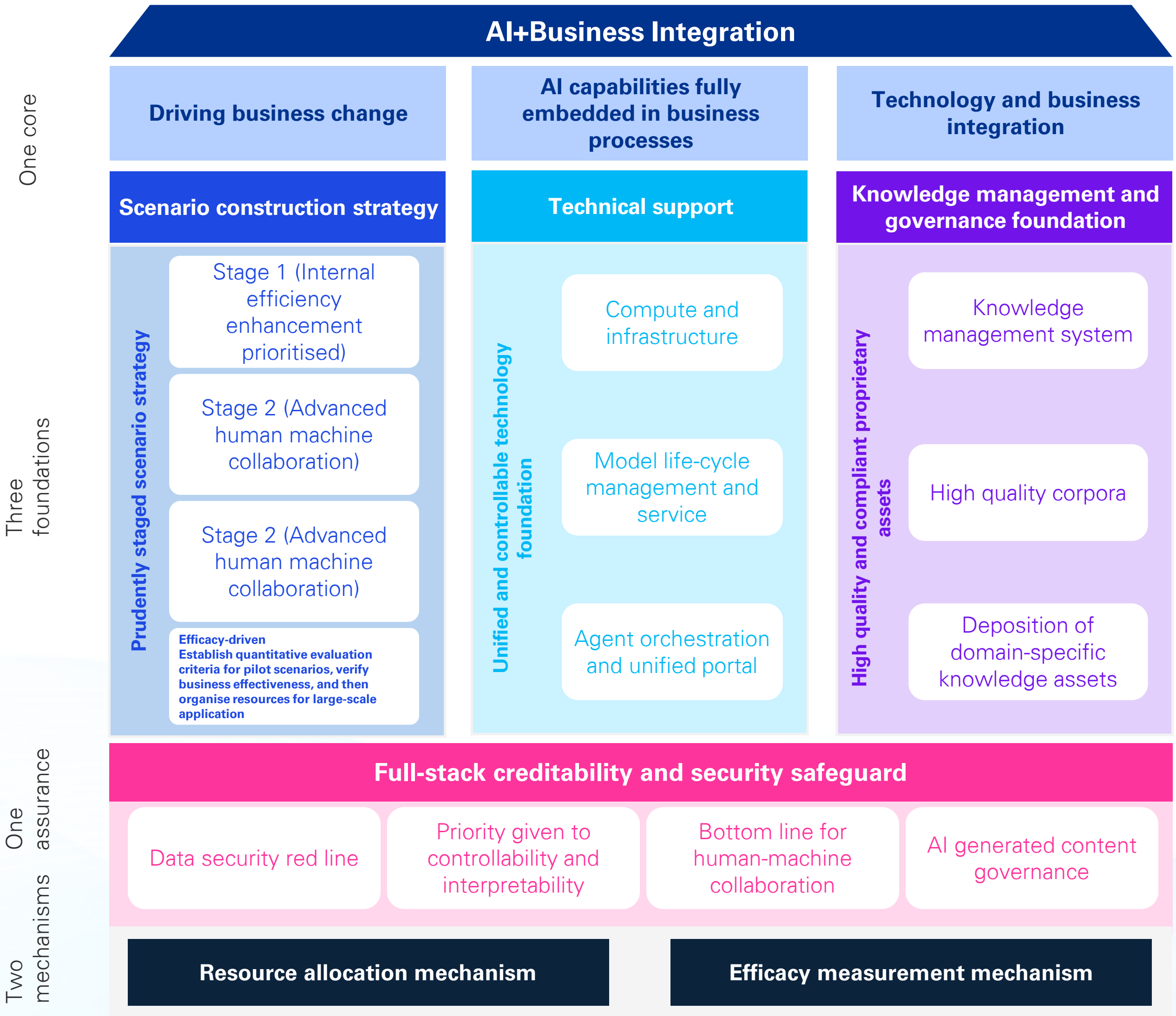
Scenario Development: Given the financial industry's high-risk, heavily regulated, and high-trust nature, the Group underlined scaling scenario development based on effectiveness. It adopted a phased approach, starting from internal efficiency enhancement scenarios before progressively moving to key scenarios such as human-machine collaboration and external service provision.

- **Technical Enablement:** The Group established an AI technology infrastructure to enable different business lines to share computing resources, model services, AI agent orchestration, and unified portals, thereby mitigating redundant developments, fragmented access, and operational cost inefficiencies.
- **Knowledge Management and Governance:** It also planned to set up a unified knowledge management system to facilitate robust mechanisms including access control, content auditing, and version management. This can strengthen the accumulation of domain-specific knowledge assets and the development of high-quality training corpora.

When building the supporting mechanism, the group considered full-stack trustworthiness and security protection as the cornerstone for AI capability development.

- It implemented a tiered governance framework addressing data security, model risk management, content quality assurance, human oversight, and accountability boundaries.
- It established resource allocation and benefit measurement mechanisms, incorporating computing resources, model utilisation, business scenarios, cost consumption, and application efficacy within a unified evaluation framework. This helps track the benefits of AI to customer service enhancement, operational cost reduction, and risk control optimisation.

Figure 11 Strategic blueprint for "AI+Business Integration"



Source: Public information, KPMG analysis

From an AI capability blueprint to computational infrastructure: Upgrading its overseas data centre as the implementation cornerstone

Guided by its AI strategic blueprint, the Group has defined several core initiatives and implementation measures covering computational infrastructure engineering, model management systems, high-quality corpus development, and AI agent orchestration frameworks. The Technical Expert Group is spearheading two critical actions under the computational infrastructure initiative:

- Enhance AI computational capacity and model service management capabilities and improve model access, invocation, scheduling, monitoring, and cost attribution, so as to enable standardised, reusable AI services across all Group business scenarios.
- Progressively upgrade data centre power systems and supporting infrastructure to provide resilient support for AI workload growth, centralised system deployment, and overseas expansion.

Infrastructure modernisation has emerged as a critical enabler for the Group's AI capability deployment. Recent acceleration of its global expansion strategy has driven increasing concentration of overseas business systems, as well as exponential growth in transaction volumes, system complexity, and real-time processing requirements. Establishing a future-proof, sustainably evolvable core infrastructure platform is now essential to support long-term overseas growth and intelligent capability scalability.

Within the Group's computational infrastructure architecture, the Americas Data Centre has become the core digital hub and computational platform for overseas operations. As this hub continues to deliver a greater portion of critical capabilities for core systems, cross-border services, and regional collaboration, the facility must satisfy multiple requirements, including financial-grade business continuity; cross-regional system interoperability; low-latency network access; centralised operational management; and scalability for future intelligent applications. Therefore, upgrading its power system and supportive infrastructure is not just for a routine data centre refresh, but to standardise the infrastructure to underpin long-term overseas growth.

From a service delivery perspective, the Americas Data Centre faces complex challenges:

Core business consolidation necessitates perpetual support for concurrent operation of diverse financial functions – trading, clearing, settlement, investment research, operations, and client servicing. Traditional project-based siloed deployments are now fundamentally misaligned with requirements for continuous expansion and unified governance.

- Core business concentration necessitates continued support for concurrent diverse financial functions, including trading, clearing, settlement, investment research, operations, and client servicing. Traditional project-based siloed deployments are now misaligned with requirements for continuous expansion and unified governance.
- Financial services highly depend on network latency stability, jitter mitigation, reliability assurance, and security perimeters. Infrastructure instability could immediately compromise overseas core business performance, operational resilience, and regulatory compliance.
- With the integration of data-intensive applications, intelligent analytics, and future AI services into business workflows, intra-facility traffic has increased, computing resource invocation intensified, and cross-system data access grown significantly. Therefore, infrastructure must be designed with the capability to support AI scenarios.
- The Group requires a standardised, visualised, and automated operational architecture with sustainable controls, in order to address overseas operational complexities arsing from multi-location team coordination, cross-jurisdictional regulatory obligations, and remote maintenance challenges.

Based on the above requirements, the Group has decided to upgrade its data centre in the Americas to build a unified computing platform, cloud management capabilities, and a high-performance data centre network so as to enable expansion of overseas nodes.

The **Cisco UCS X-Series modular computing platform** provides a standardised and scalable computing infrastructure for multiple types of business workloads. It supports the phased evolution of traditional enterprise applications, data-intensive systems, and subsequent intelligent workloads within a unified architecture.

- **Cisco Intersight** integrates the lifecycle management, policy consistency, operational visualisation, and automated operation and maintenance of computing resources onto a unified platform through its cloud management capabilities. This significantly reduces the complexity of remote operations and cross-site collaboration within overseas data centres.

- The **Cisco Nexus 9000 high-performance network** delivers foundational capabilities for high-speed interconnection, low-latency forwarding, and large-scale east-west traffic handling between core business systems. Furthermore, it provides headroom for the evolution towards the high-bandwidth networks required for future AI and data-intensive applications.

This architectural approach establishes the group's future-oriented technical baseline for overseas data centres through the integration of computing, network, and operational management capabilities. For current business needs, this upgrade delivers a more stable, high-performance, and readily governed operating environment for the Americas region's core financial systems. Simultaneously, for future evolution, its modular computing, high-performance networking, and centralised operational maintenance platform collectively form a sustainable and scalable AI-ready infrastructure foundation. This will seamlessly accommodate the phased implementation of subsequent intelligent applications – including investment research, risk analysis, customer insights, operational automation, and data-intensive workloads.

Key takeaways:

Comprehensive AI transformation in large organisations – A systematic approach grounded in full-stack infrastructure modernisation

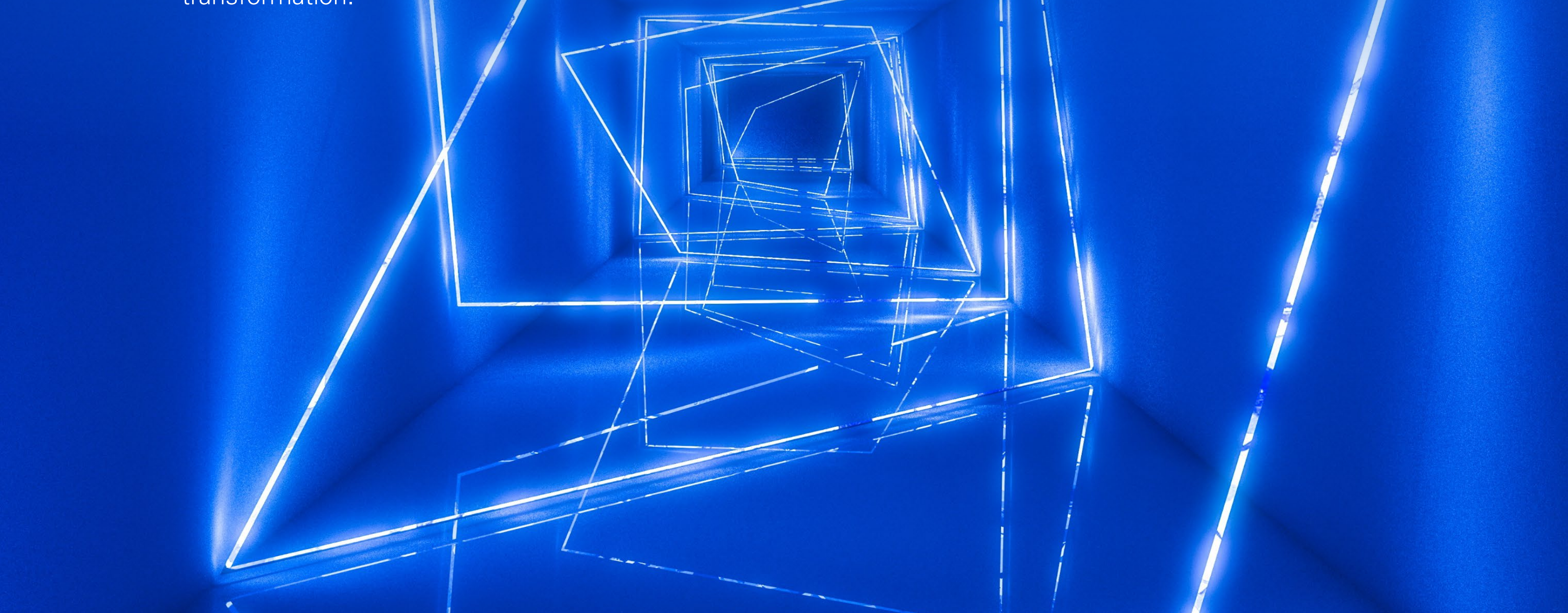
The essence of AI transformation lies in a systemic shift centred on redefining an enterprise's core competitive advantage. For the global wealth group in this case study, the true measure of successful AI transformation hinges on whether AI capabilities can be deeply integrated across five critical dimensions: business strategy, organisational mechanisms, knowledge assets, governance frameworks, and infrastructure capabilities. This integration must coalesce into a reusable, governable, and sustainable next-generation operational ecosystem.

This progression from limited AI pilot projects to large-scale implementation demands both innovation in application-layer solutions and underlying infrastructure capable of supporting high-concurrency processing, ultra-low latency, carrier-grade reliability, and continuous scalability. The group's Americas data centre has evolved from a regional support node into the primary deployment platform for the group's overseas digital and AI capabilities, establishing a replicable, evolvable, and governable blueprint for global infrastructure development. As the group's international business footprint expands, this infrastructure-first approach offers a critical reference framework for subsequent overseas regional expansion, standardised deployment of core systems, and systematic implementation of AI capability frameworks. Ultimately, this foundation delivers enduring, stable, and sustainable support for global operations and intelligent transformation initiatives.

Conclusion and outlook

Recognising emerging patterns and contextual dynamics is paramount. Enterprise intelligent transformation is progressing irreversibly towards scaled implementation, operationalised trustworthiness, and production-grade deployment. Unlocking intelligent productivity demands both comprehensive enhancement of foundational infrastructure – spanning energy systems, compute-network-stack resources, and model-service orchestration; and harmonised collaboration across model services, business applications, and runtime governance. Simultaneously, organisations must rigorously evaluate per-token computational economics, task completion fidelity, and business value realisation, while embedding security-by-design, regulatory compliance, auditability, and risk mitigation throughout the AI operational lifecycle. This white paper examines industrial transformation catalysed by AI technologies and explores next-cycle industry-AI integration pathways. It delineates essential frameworks and critical success factors for enterprise artificial intelligence transformation. Through pioneering implementation case studies, we provide actionable reference architecture for organisations to navigate transformation vectors, address capability gaps, and execute implementation roadmaps in the emerging AI agent paradigm.

Maintaining stability requires a focus 'towards the future'. Technologies such as embodied intelligence and digital twins will accelerate the transition of AI from the digital realm into the physical world. The path towards Artificial General Intelligence (AGI) necessitates extensive long-term exploration, while enterprise AI integration inevitably faces multifaceted challenges including deepening application scenarios, engineering implementation, organisational adaptation, and trustworthy governance. Looking ahead, we will continue to monitor emerging trends, scenarios, and issues in the integration and development of enterprise AI. Our focus will encompass AI readiness assessment, systematic enterprise AI architecture development, full-stack infrastructure modernisation, token economics applications, AI agent governance, and trustworthy AI engineering. We shall distil practical, implementable methodologies and experiential insights from these endeavours. We look forward to collaborating with stakeholders across sectors to advance the transformation of AI capabilities from technological applications into productive systems. This will assist enterprises in consolidating digital resilience, unlocking intelligent productivity in the era of AI agents, and progressing steadily towards a new stage of AI-enabled transformation.



KPMG

KPMG China has established offices in 31 cities with more than 14,000 partners and employees in Beijing, Changchun, Changsha, Chengdu, Chongqing, Dalian, Dongguan, Foshan, Fuzhou, Guangzhou, Haikou, Hangzhou, Hefei, Jinan, Nanjing, Nantong, Ningbo, Qingdao, Shanghai, Shenyang, Shenzhen, Suzhou, Taiyuan, Tianjin, Wuhan, Wuxi, Xiamen, Xi'an, Zhengzhou, Hong Kong SAR and Macau SAR. Operating under an efficient integrated management model, these offices collaborate seamlessly to deliver consistent premium services to clients.

Cisco (NASDAQ: CSCO)

Cisco (NASDAQ: CSCO), a global technology leader, is dedicated to transforming enterprise connectivity and security through AI. For over 40 years, Cisco has connected the world with security at its core. Through AI-driven solutions, it empowers customers, partners, and communities to unlock innovation, enhance productivity, and strengthen digital resilience. Guided by its vision of an "inclusive future," Cisco continues advancing global connectivity.

The company ranks 246th on the 2024 Fortune Global 500 list and has consistently featured among the world's best workplaces in rankings by Great Place to Work® Institute, the renowned global workplace research and certification body.

kpmg.com/cn/socialmedia

For a list of KPMG China offices, please scan the QR code or visit our website:
https://kpmg.com/cn/en/about/office-locations.html

Publication number: 1782456051207

Publication date: 2026.7