



Together makes progress

FinOps for AI: From cost complexity to measurable returns

August 2026

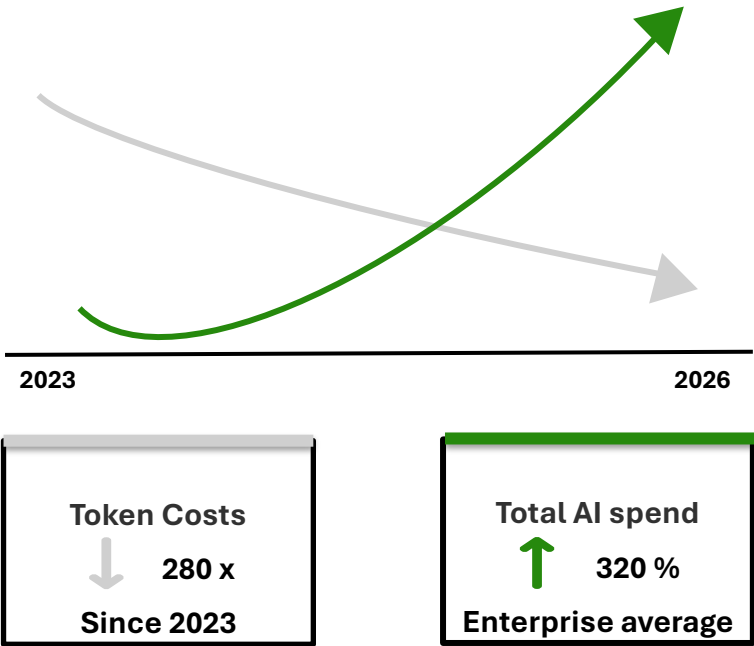


AI spend is scaling faster than financial control

As AI adoption accelerates, organisations need new cost visibility, governance, and value metrics to avoid budget surprises and value leakage

The Situation

While the token prices are falling, the total consumption is rapidly rising due to higher competition, model optimisation and open-source models



Source: www.oplexa.com

The Challenges

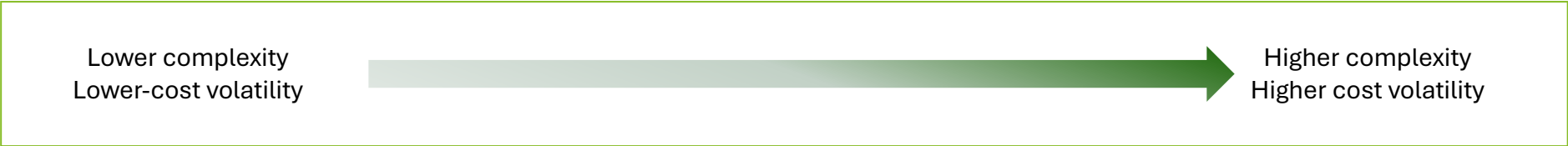
- Runaway experimentation:** Dozens of AI pilots and PoCs across the business, often using different providers and models, with limited visibility & control over it.
- Cost-value disconnect:** AI spending is growing, but organisations lack visibility into what they're spending and what they're getting in return.
- Widespread adoption:** Variety of different business user groups make use of AI without consistent governance and cost controls
- Regulators demanding control:** Heightened expectations on responsible AI, financial discipline, and sustainability are converging
- Vendor management:** Fragmented AI vendor landscape and varying pricing models limit organisations' ability to manage cost commitments effectively

Enterprises are accelerating AI and Agent pilots quickly, but lack a coherent economic and governance model, leading to budget surprises, shadow AI, and value leakage

Different AI architectures create different cost dynamics

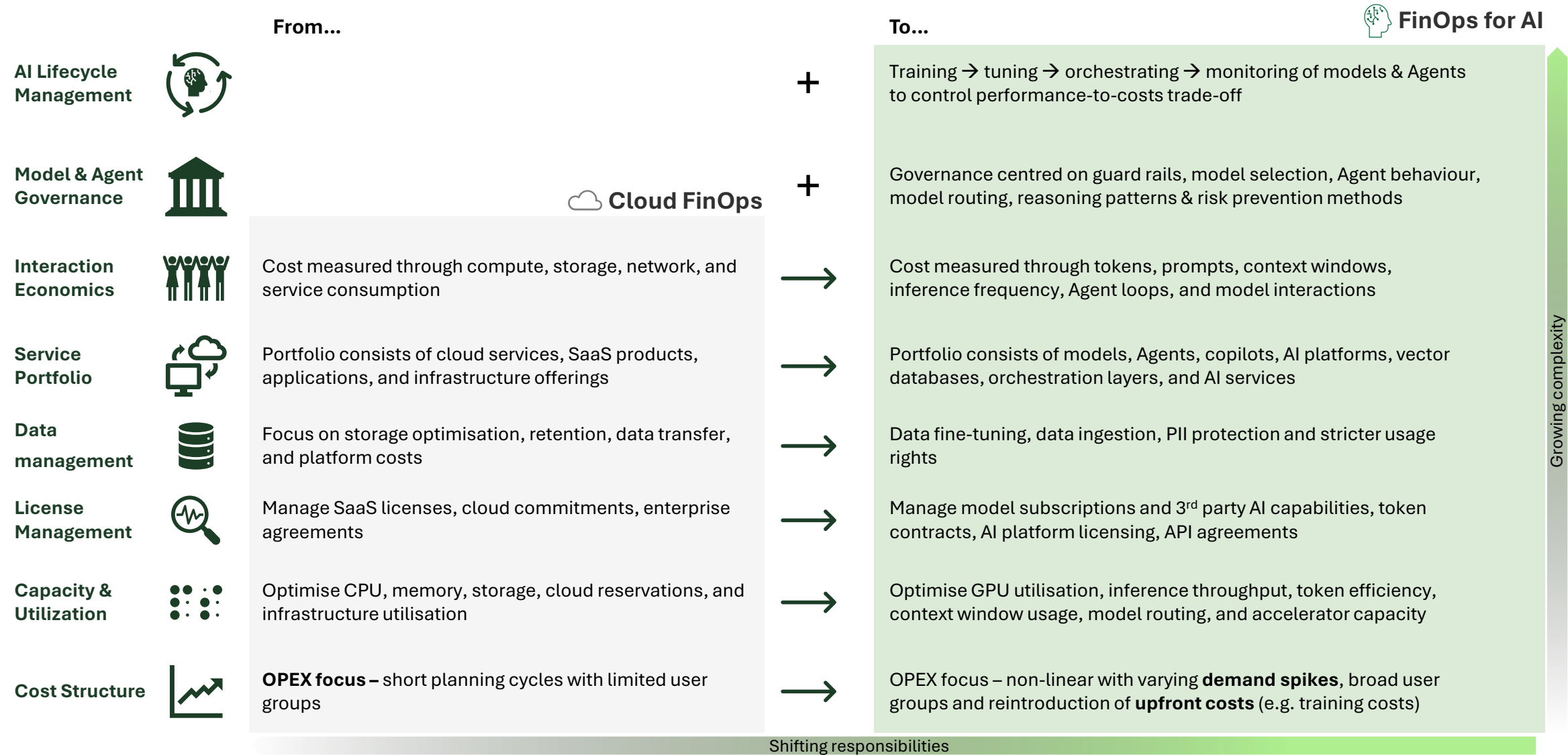
Understanding how different AI architectures create value, consume resources, and drive costs is the baseline for successful FinOps for AI

	 Predictive AI	 Generative AI	 Agentic AI
Description	Predicts outcomes based on historical patterns and data	Creates new content such as text, code, images, or audio	Reasons, plans, and executes tasks autonomously using tools and workflows
Use cases	Forecasting, fraud detection, recommendations, churn prediction	Chatbots, copilots, content creation, summarisation, code generation	Customer service Agents, research assistants, workflow automation, multi-step business processes
Optimisation Focus	Accuracy, training efficiency, model performance	Token efficiency, model selection, prompt optimisation	Task success rate, orchestration efficiency, Agent governance, cost per completed task
Unit Economics	Cost per prediction	Cost per token/inference	Cost per completed task



Traditional FinOps remains essential, but AI changes the focus

New AI workloads introduce new economics and risks, requiring capabilities beyond cloud FinOps



A large, dark green circle is centered on the page, surrounded by a bright green glow that fades into the background. The text "Spotlight Agentic AI" is centered within the dark circle.

Spotlight Agentic AI

Agentic AI: The value and cost multiplier

While LLMs have introduced human-guided queries with relatively transparent and per-request cost models, Agentic AI adds another layer of complexity with multi-step execution and fragmented cost drivers

The Value Multiplier

Autonomy: Agents execute 24/7 without human intervention, reducing operational headcount and enabling teams to redirect effort to more impactful work

Complexity: Agents handle multi-step tasks beyond the scope of human-assisting tasks of earlier models

Tool Integration: Agents connect across systems to enable E2E automation and closed-loop feedback

Continuous Learning: Agents adapt to patterns and optimise decisions over time, reducing the need for manual retraining

Scalability: Agents enable parallel execution without proportional headcount growth

But...

The Cost Multiplier

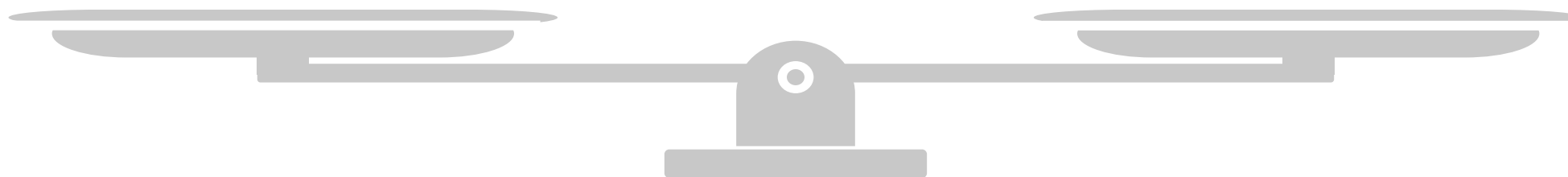
Autonomous reasoning gives less control over the execution of the tasks including iterative loops per task, pushing token consumption up

Complex queries require recursive model calls, and extensive context retrieval can increase per-decision costs by 10-100x

Orchestration complexity expands API call volumes, and introduces security/compliance overhead

Ongoing model retraining, drift monitoring, and emergent management introduce recurring and often **hidden costs**

Production scale introduces infrastructure strain, API concurrency costs, and data pipeline bottlenecks **driving non-linear cost increases**



Agentic AI unlocks **step-change value** through autonomous decision-making, end-to-end workflow automation, and continuous learning.

Without control the **costs can outpace the value** quickly. Organisations must actively manage cost-accuracy trade-offs to ensure measurable business value.

Spotlight Agentic AI: Comparison of cost variances

The economics of the same task can vary materially depending on model and architecture choices. Therefore, the choice of the right model for the right task is crucial for the creation of measurable business value

Model pricing comparison*

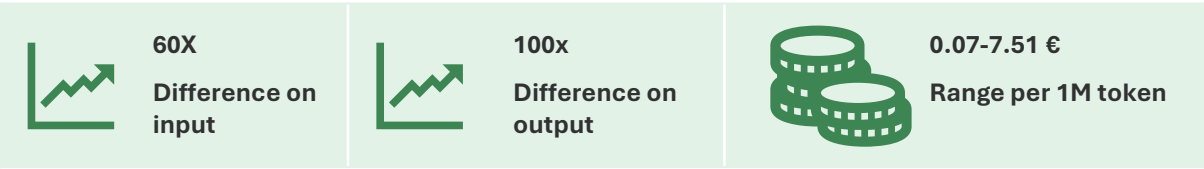
Meta Llama 4 Scout		€0.07
DeepSeek V4 Flash		€0.14
Google Gemini 3.5 Flash Lite		€0.22
Claude Haiku 4.5/ OpenAI GPT-5.6 Terra		€0.75
Claude Sonnet 4.6		€2.25
Claude Opus 5 / OpenAI GPT-5.6 Sol		€3.76
Claude Fable 5		€7.51

* Frontier list prices. EUR per **1M** input tokens, list prices **July 2026** via cloudprice.net.

Task pricing comparison*

Model	Capability	Cost per task
Claude Fable 5	Premium AI model for advanced reasoning & intelligence	€1.41 per task
OpenAI GPT-5.6 Terra	General purpose AI model with strong reasoning	€0.36 per task
Claude 4.5 Haiku	Lightweight AI model enhanced with RAG and caching	€0.14 per task
Meta Llama 4 x Maverick	Open-source model hosted internally	€0.02 per task

* Workload: research Agent summarising 50 internal docs ~120K input + 8K output tokens, 6 tool calls

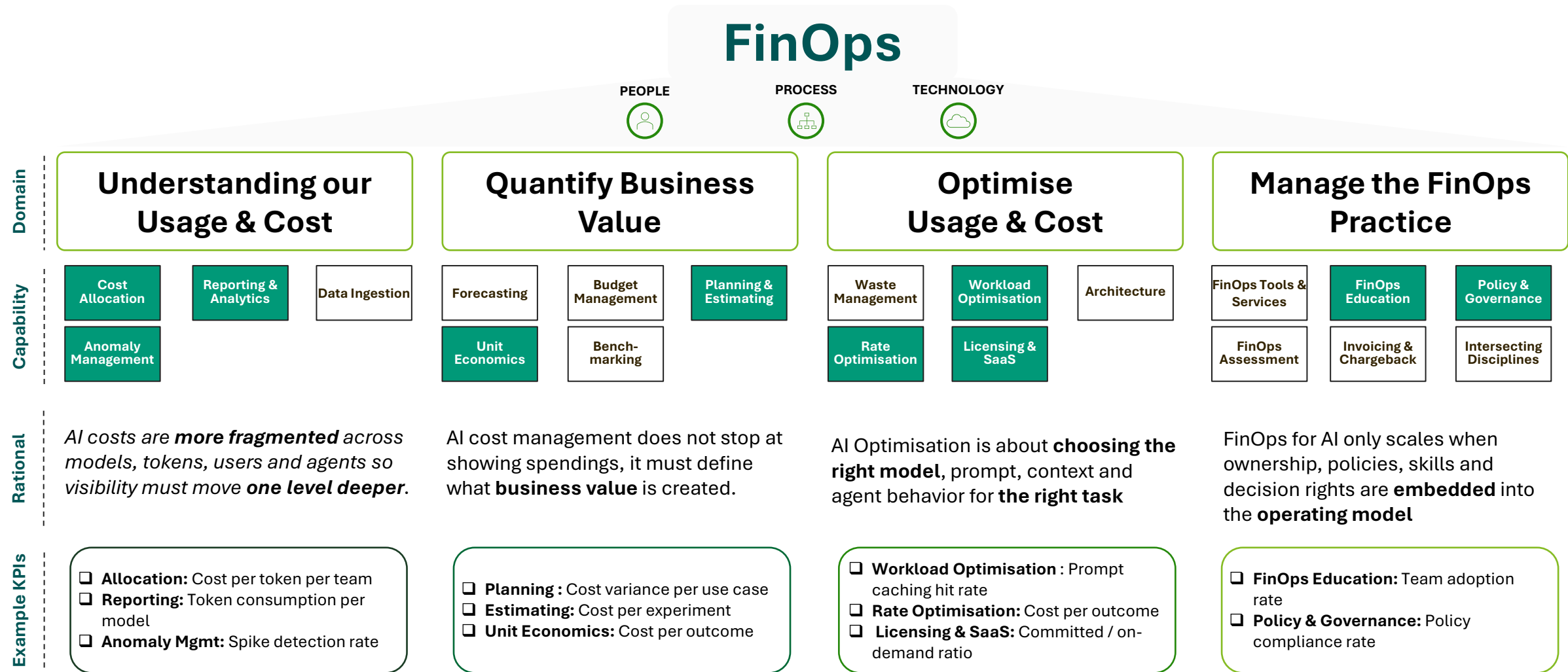


Making the right architectural choices and picking the right model for the right task will lead to significant cost differences, making cost control a critical optimisation capability

FinOps Framework & Levers

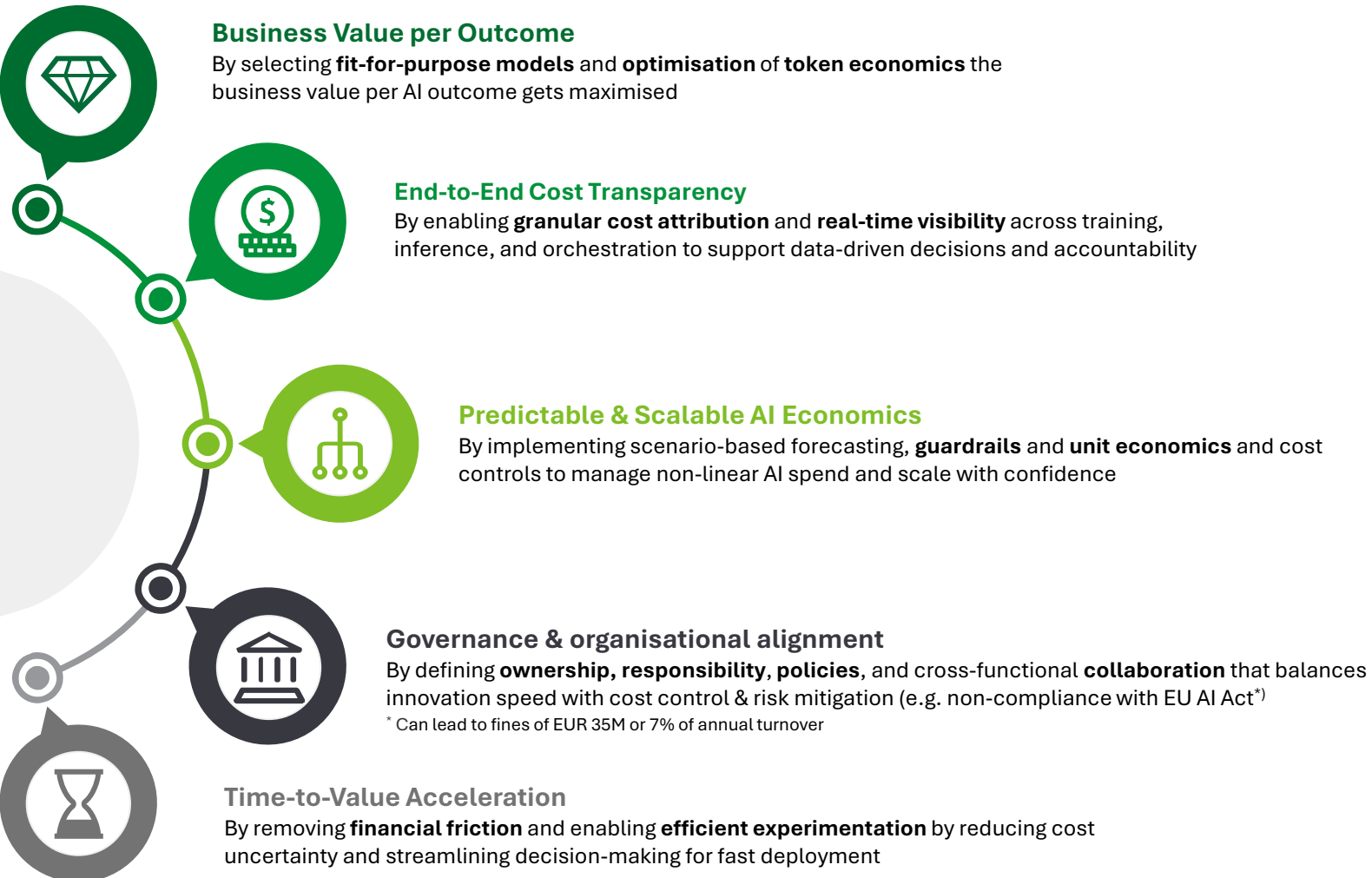
The FinOps for AI framework

FinOps for AI builds on the same core disciplines. However, a different focus must be applied to address the new cost dynamics and successfully generate business impact out of AI use cases



Five Levers for FinOps for AI

Five operational levers enable organisations to transform AI cost complexity into predictable, measurable business value and scale with financial discipline

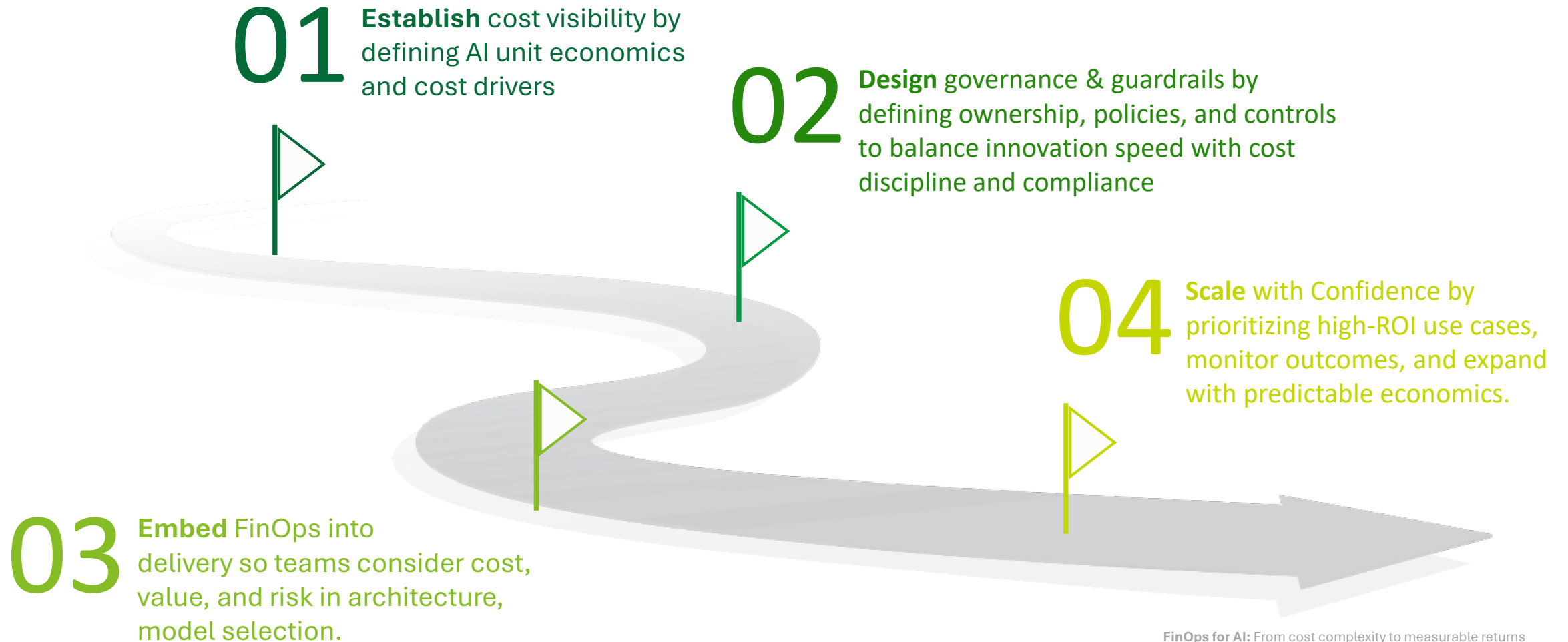


Questions to Define Your AI Setup

- ✓ **What values are we optimising for?**
Efficiency → lower-cost model, quick throughput, cost per task
Growth → stronger models, user adoption, validation layers
Quality → validation layers, monitoring, guardrails
- ✓ **Is this experimental or scaled production?**
Experiment → managed APIs, spend cap, cost tracking
Pilot → success metrics, monitoring, ownership
Production → production controls, observability, chargeback
- ✓ **How predictable is the usage?**
Ad hoc → pay-per-use, low/no capacity planning
Predictable → reserved capacity, batching, caching
Spikey → scaling policies, monitoring, tight guardrails
- ✓ **How autonomous should the AI be?**
Assistant → Fast models, optimised for cost per interaction
Automation → Mid-tier models with controls, cost per approval
Agent → Strong model orchestration, robust monitoring, cost-loop-controls
- ✓ **What level of compliance control is required?**
Internal → standard managed setup
Regulated → strong access control, logging, approved platforms
Highly restricted → dedicated environments / self-hosting, high governance focus

From AI ambition to AI accountability

Start your FinOps for AI journey by making AI economics visible before scaling high-value use cases



Deloitte brings the capabilities to operationalise FinOps for AI

We combine FinOps, AI engineering, cloud alliances, and operating model experience to help clients move from insight to measurable impact



Deep FinOps and cloud heritage

- Extensive experience implementing FinOps practices **across large enterprises and industries**.
- Proven methods to **integrate FinOps** into operating models, not just technology stacks.



AI and Agentic expertise

- Dedicated AI and GenAI practices with **hands-on** experience in building copilots and Agentic systems.
- Ability to translate AI architecture into **financial and risk terms** that CFOs and boards understand.



Multi-hyperscaler alliances

- Strategic alliances with all major **hyperscalers** (e.g., AWS, Azure, Google Cloud) and leading AI platform providers.
- Ability to **benchmark** pricing models and patterns across providers and design vendor-agnostic FinOps for AI strategies.



Tech-savvy, hands-on delivery

- Teams combining cloud engineers, data scientists, and FinOps practitioners **working together** on the same client outcomes.
- Focus on **tangible results**: measurable cost savings, dashboards, and governance artifacts—not just strategy slides.



165+

Clients Served Across The Globe



6

Key Industries Supported



260+

Certified FinOps Practitioners



15%-20%

Average Savings Identified Across Clients

“ A worldwide leader in Cloud Professional Services based on capability and strategy - IDC ¹ ”

Deloitte has also been recognised as Amazon Web Services Platinum Partner, Google Cloud Alliance Partner, and has a thriving Microsoft Azure Technology Practice



Critical Success Factors



Business-Led, Outcome-Focused Approach

We build and operate FinOps for AI workloads, combining engineering, data, and finance. Outcome-driven delivery ensures rapid value with controlled AI spend.



Proven Record of FinOps Success across Major CSPs

Deep FinOps experience across CSPs, extended to AI and GenAI use cases. We help avoid cost visibility gaps and uncontrolled usage growth.



Multi-Disciplinary Integrated Offerings

Expertise across AI & Data, FinOps, SAM, and TBM for full lifecycle cost governance. We link model, training, and usage decisions to financial outcomes.

¹ IDC MarketScape

Key contacts



Andrea Farnon
Partner
FinOps Lead Partner
afarnon@deloitte.nl



Janosch Glass
Manager
FinOps Lead
janglass@deloitte.nl



Kirti Yadav
Manager
Technical FinOps Lead
kirtyadav@deloitte.nl



Mahtab Mahdavifar
Consultant
FinOps Expert
mmahdavifar@deloitte.nl

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited (“DTTL”), its global network of member firms, and their related entities (collectively, the “Deloitte organization”). DTTL (also referred to as “Deloitte Global”) and each of its member firms and related entities are legally separate and independent entities, which cannot obligate or bind each other in respect of third parties. DTTL and each DTTL member firm and related entity is liable only for its own acts and omissions, and not those of each other. DTTL does not provide services to clients. Please see www.deloitte.com/about to learn more.

Deloitte provides leading professional services to nearly 90% of the Fortune Global 500® and thousands of private companies. Our people deliver measurable and lasting results that help reinforce public trust in capital markets and enable clients to transform and thrive. Building on its 180+ year history, Deloitte spans more than 150 countries and territories. Learn how Deloitte’s over 470,000 people worldwide work together every day to make an impact that matters at www.deloitte.com

This communication contains general information only, and none of Deloitte EMEA (EMEA), its shareholders and affiliates is, by means of this communication, rendering professional advice or services. Before making any decision or taking any action that may affect your finances or your business, you should consult a qualified professional adviser.

No representations, warranties, or undertakings (express or implied) are given as to the accuracy or completeness of the information in this communication, and none of EMEA, the EMEA shareholders, their respective affiliates, employees or Agents shall be liable or responsible for any loss or damage whatsoever arising directly or indirectly in connection with any person relying on this communication. EMEA, the EMEA shareholders and their respective affiliates are legally separate and independent entities.