

FINOPS FOR AI

**AI-INVESTITIONEN EXPLODIEREN — WAS UNS DIE
ANFÄNGE VON FINOPS FÜR DEN AI-BOOM LEHREN**

September 2026



AI-Kosten wachsen schneller, als Unternehmen sie verstehen

Was heute für die Cloud gilt, gilt für AI in noch größerem Maßstab — und deutlich schneller.

+76%

Wachstum globaler GenAI-Ausgaben
2025 ggü. 2024 — auf 644 Mrd. \$

Quelle: [Gartner Pressemitteilung 31.03.2025](#)

75%

der AI-Initiativen verfehlen den ROI, den
CEOs erwarten – **nur 16%** skalieren
unternehmensweit.

Quelle: [IBM Institute of Business Value, CEO Study 2025 \(Mai 2025\)](#)

79%

der Unternehmen hatten 2025 AI-
Kostenüberschreitungen

Quelle: [Sapio Research / DoiT International, Feb 2026](#)



Warum AI-Kosten schwieriger zu steuern sind als Cloud-Kosten



Nutzungs Ebene

End-User
Nutzung

In-App
Embedding

Agentic
AI

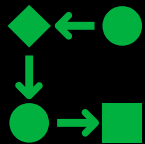
APIs &
Integrationen

Endprodukt
Integration



Cloud

Klar messbar
(vCPU-Stunde, GB/Monat,...)



Orchestrierungs Ebene

Model-
auswahl

Prompt
Engineering

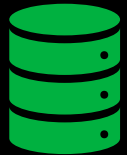
Prozesse
und
Workflow

Policies und
Security

Realtime
Monitoring



Einheitliche Service-Typen und
kalkulierbare Laufzeiten



Daten Ebene

Strukturierte
Daten

Unstrukturi
erte Daten

Sensitive
Daten

Aktualität und
Relevanz

Value hängt an
Datenqualität



Klare Kostenzuordnung und
Messgrößen
(GB, IOPS & Transferrate)



Modell Ebene

Kommerziell
vs Open

3rd Party
Abhängigkei
t

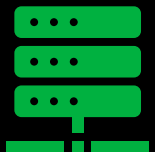
Training
und
Tuning

Intransparent
e
Performance

Spezialisierun
g



Bezogene Leistung zu 100%
quantifizierbar



Infrastruktur Ebene

Cloud
Infrastruktur

Verfügbarkei
t & Kosten

SaaS AI
Services

On-Prem
Infrastruktur

Scalability &
Performance



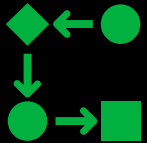
Standardisiert und vergleichbar



AI KOSTENTRANSparenZ

TRANSPARENZ SCHAFFT EIN GEMEINSAMES VERSTÄNDNIS FÜR INVESTITIONEN IN AI

Worüber wir Einblick benötigen



AI-Nutzung

Welche Applikationen, Modelle und Provider werden wo genutzt?



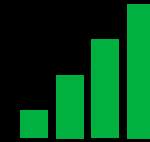
Verantwortlichkeiten

Wer ist verantwortlich für welche Ausgaben?



Ausgaben

Was sind die Ausgaben pro Applikation, Cost Center, Token, Verbrauch... ?



Verbrauch

Aktuelle und historische Token-Nutzung, Anfragen, Aufrufe, aktive Nutzer?



Use Cases

Was ist der Zweck, der Anwendungsfall und die Initiative der Ausgaben für AI?



Business Impact

Was ist der erwartete Nutzen? Gibt es einen messbaren ROI?

Transparenz ist die Grundlage für Verantwortlichkeit und fundierte Entscheidungen



Was auf dem Spiel steht, wenn wir jetzt nicht handeln

1

Kostenintransparenz

AI-Spend versteckt sich in Cloud-Rechnungen ohne Zuordnung zu Teams, Projekten oder Use Cases.

2

Unvorhersehbare Ausgabenspitzen

Token-intensive Prompts oder ein viraler interner Use Case erzeugen Rechnungsüberraschungen ohne Vorwarnung.

3

Shadow AI

Teams beschaffen AI-Tools und APIs am zentralen Procurement vorbei – unkontrolliert und dupliziert.

4

Fehlende Wertkopplung

Kosten und Business Value werden in getrennten Systemen erfasst – Investitionen werden über Aktivität statt Ergebnis gerechtfertigt.

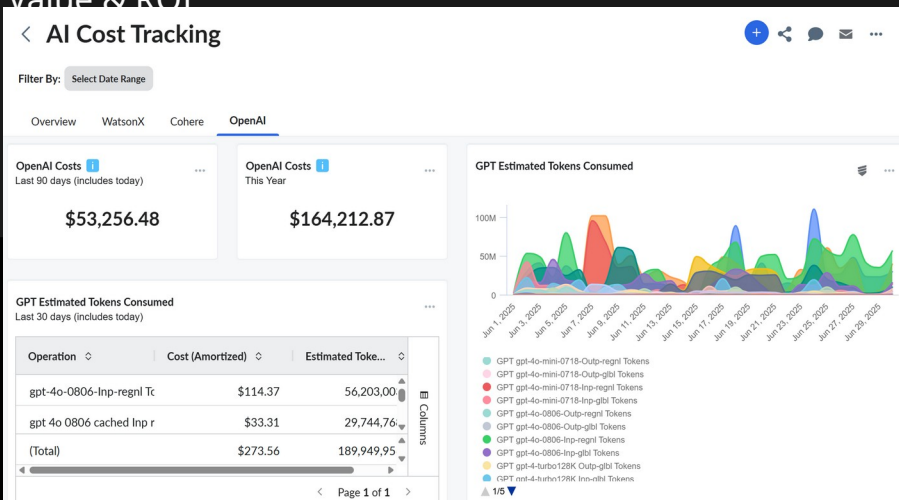


Apptio bringt AI TCO und Value zusammen

IBM Apptio AI TCO & Usage

Umfassende AI-Kosten- und Nutzungsanalyse

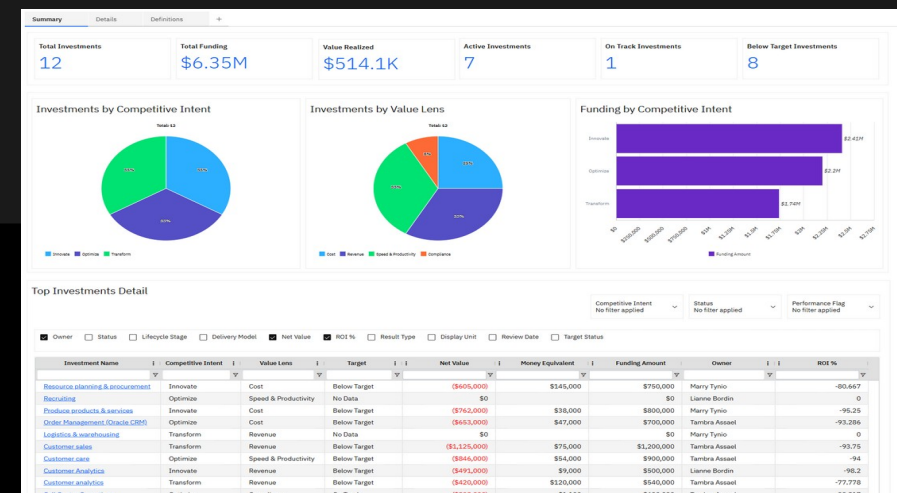
- Detailanalyse von AI-Kostenfaktoren, Personalstruktur, Nutzungsmustern und Stückkosten
- Tracking des Verbrauchs zur Unterstützung von Kostenverteilung und -optimierung
- Erkennung von Einführungs-, Nutzungs- und Kostensignalen, die governance-relevante Aufmerksamkeit benötigen
- Lieferung modellierter AI-Kostendaten an IBM Apptio AI Value & ROI



IBM Apptio AI Value & ROI

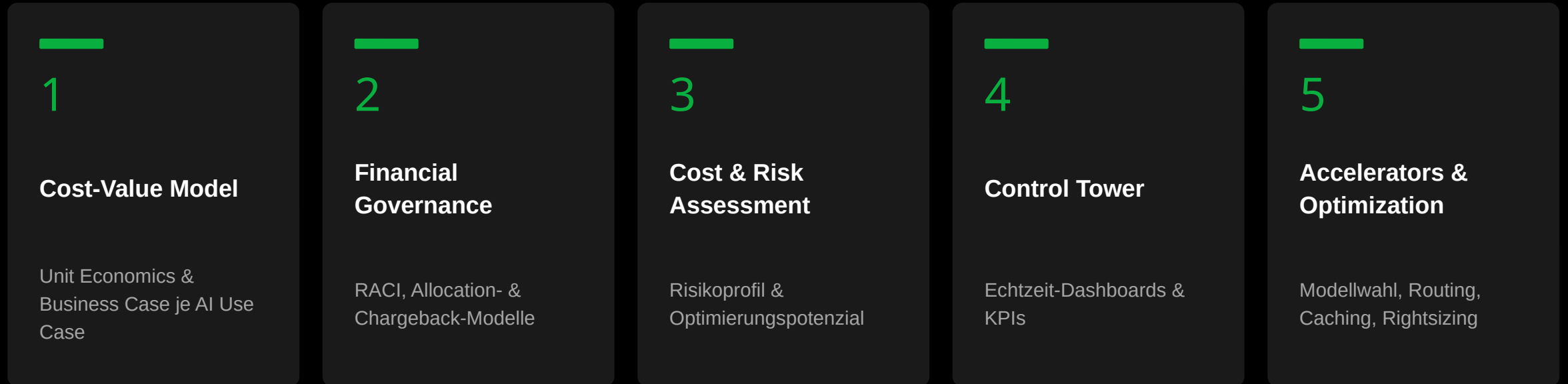
Perspektive der AI-Wertschöpfung

- Abgleich zwischen geplantem und tatsächlich realisiertem Wert mit entsprechenden Nachweisen
- Einbettung des AI-Initiativenwerts in den Gesamtinvestitionsrahmen,
- ROI-Bewertung auf Basis transparenter Kosten-Nutzen-Rechnung
- Fundierte Entscheidungsgrundlage für Finanzierung, Ausweitung, Anpassung, oder Beendigung von AI-Projekten



Der FinOps for AI Accelerator

Ein modulares Framework statt einer Einheitslösung – aufbauend auf der FinOps Foundation, erweitert um AI-spezifische Bausteine.



Zwei weitere Dimensionen steuern den Einsatz der Bausteine: der AI-Produkt-Lifecycle und das AI Maturity Assessment bestimmen, welches Paket wann und wie tief zum Einsatz kommt.



NÄCHSTER SCHRITT

Lassen Sie uns mit einem AI Maturity Assessment starten.

1

Kickoff & Scoping-Workshop

Ziele, Stakeholder und Datenzugriff gemeinsam festlegen

2

Datenzugriff & Interviews

Bestehende AI-Use-Cases, Provider und Kostentreiber erfassen

3

Ergebnis & Roadmap-Freigabe

Reifegrad, Business Case und nächste Phase gemeinsam beschließen

Cortex Reply



Vielen Dank für ihre Aufmerksamkeit



Josef Petzenhammer

Executive Value Lead bei IBM
Fokus auf FinOps und ITFM
Mit Sitz in München



Alexander Baier

Geschäftsführer von Cortex
Reply
Fokus auf FinOps und ITFM
Mit Sitz in München

