>>AI CONF 2026

Cost Engineering per la Generative AI

Leonardo Cruciani
Senior Data Scientist

Federica Pampurini
Data Scientist Lead

Softlab S.p.a.

improove

Kudos++

Executive

beSharp. we make IT run.

LEXROOM

REPLY STORM

Gold

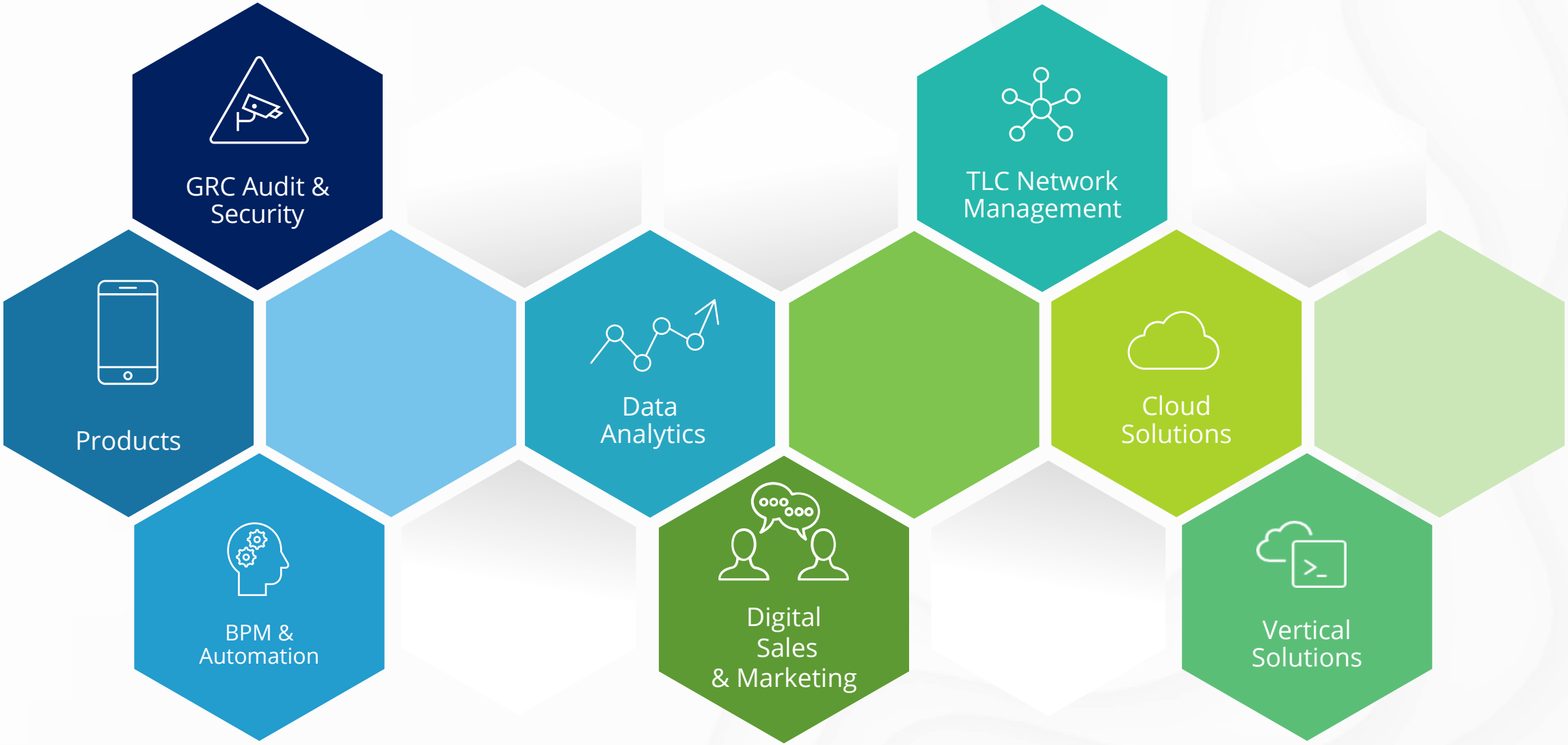
seeweb

FASTWEB + vodafone

Softlab S.p.A. is a Consulting Company listed at Milano Stock Exchange, working in Business Advisory & ICT Consulting.

*2025

Offering



Quanti di voi sono coinvolti nella progettazione e sviluppo di progetti di AI generativa?

Quanti sanno esattamente quanto spende quel sistema?

Softlab

Cos’è il FinOps?

La disciplina nata nel cloud per gestire una **spesa variabile, distribuita e generata in autonomia da molti team**

Il problema

Nel cloud ogni sviluppatore può attivare risorse a consume
La spesa diventa continua e frammentata

L'approccio

Mettere engineering, finance e business a lavorare sugli stessi dati di costo, attribuendo ogni spesa a team, progetto o feature.

I principi operativi

Inform Optimize Operate

Softlab

I principi operativi

Tre fasi che si alimentano a vicenda e si ripetono nel tempo

01

- Misurare la spesa in modo granulare
- Attribuirla a team, progetto, prodotto
- Baseline, budget e unit economics

02

- Rate optimization: pagare meno
- Usage optimization: consumare meno
- Prioritizzare per ritorno atteso

03

- Processi e responsabilità chiare
- Alert e gestione delle anomalie
- Verificare e **riattivare il ciclo**

Softlab

Il FinOps applicato alla Generative AI

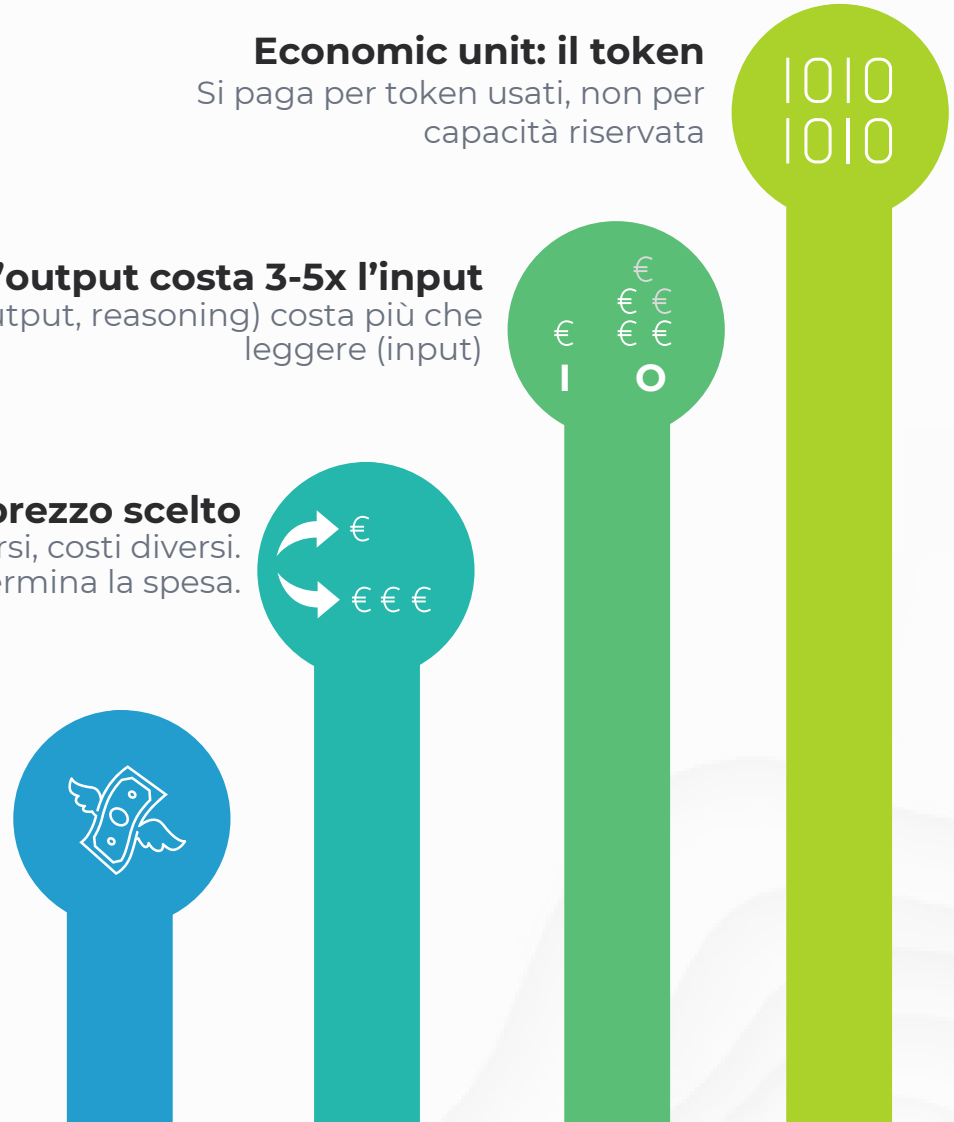
Cosa cambia rispetto al cloud tradizionale quando usi un LLM come servizio?

Economic unit: il token
Si paga per token usati, non per capacità riservata

L'output costa 3-5x l'input
Generare (output, reasoning) costa più che leggere (input)

Il modello scelto è il prezzo scelto
Modelli diversi, costi diversi.
Il routing della richiesta determina la spesa.

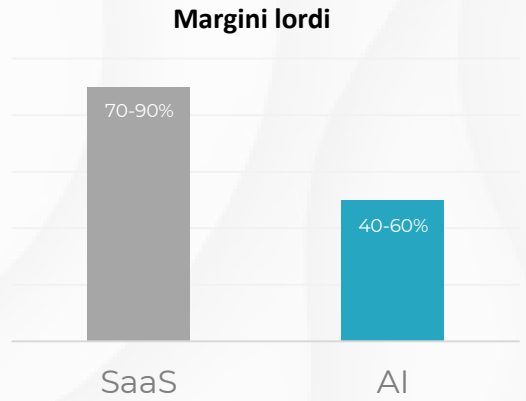
La spesa cresce in silenzio:
difficile da monitorare, granulare e diffusa



COSTO MARGINALE ≠ 0

Ogni inferenza si paga
Il costo scala con l'uso, non si azzera.

Inference ceiling
Ad alti volumi la crescita si ferma per ragioni economiche prima che tecniche

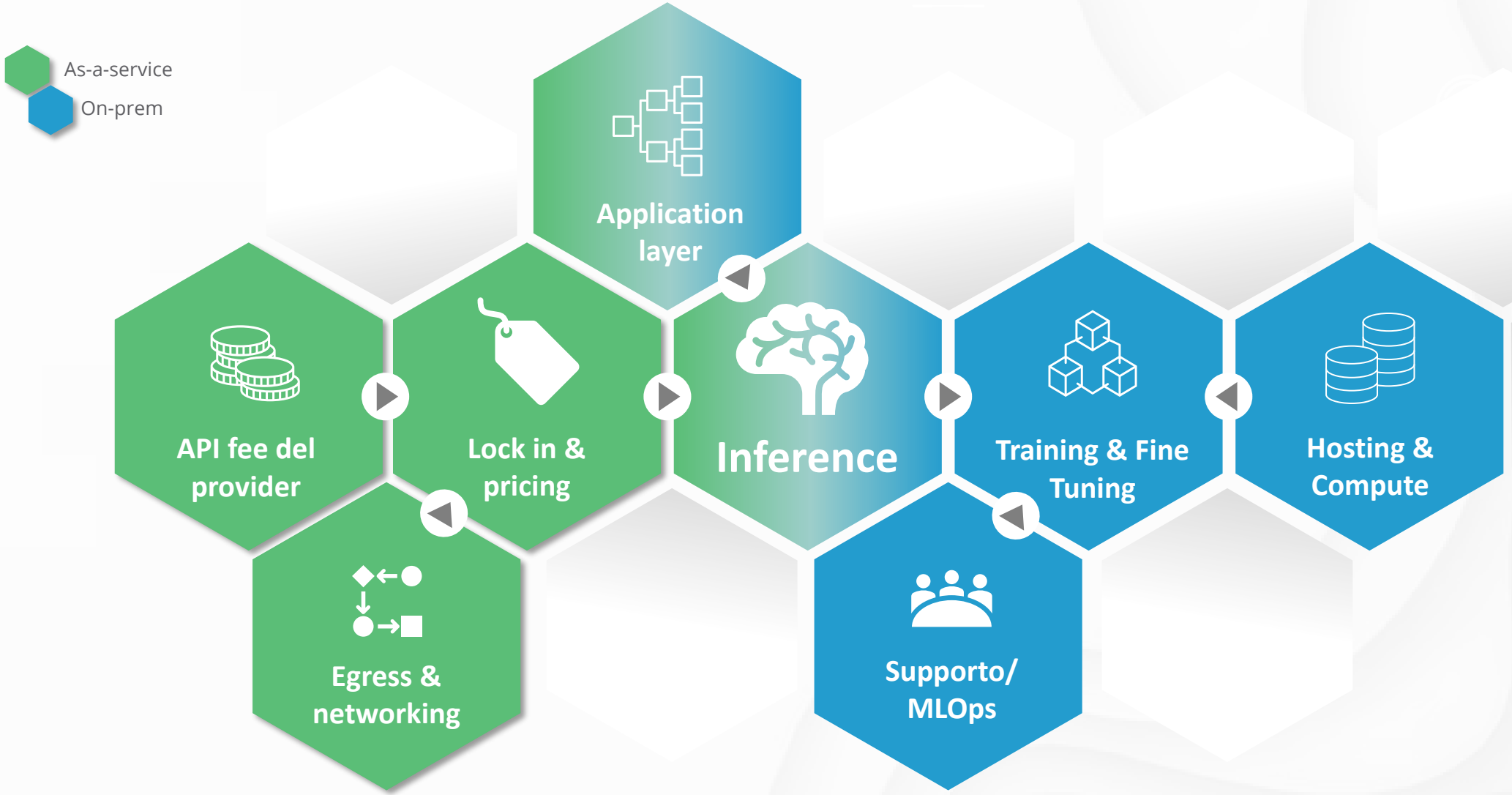


Fonte: SoftwareSeni, ICONIQ Capital

Softlab

Il costo di un sistema di GenAI

I principali fattori di costo (tecnici)



Softlab

Cost Engineering nel ciclo di vita di un progetto AI

Ottimizzazioni tecniche, scelte organizzative e profilo di costo

POC

PILOT

PRODUCTION

SCALING

Softlab

Cost Engineering nel ciclo di vita di un progetto AI

Ottimizzazioni tecniche, scelte organizzative e profilo di costo

POC

FinOps: Pre-Inform

Validare fattibilità e valore

Scelte organizzative

- Budget discrezionale
- Team ristretto, nessun SLA
- Stage gate sulla qualità, non sul costo.

Ottimizzazioni tecniche

- Uso di modello frontier
- Prompt esplorativo
- RAG prototipale
- Nessun caching né routing.

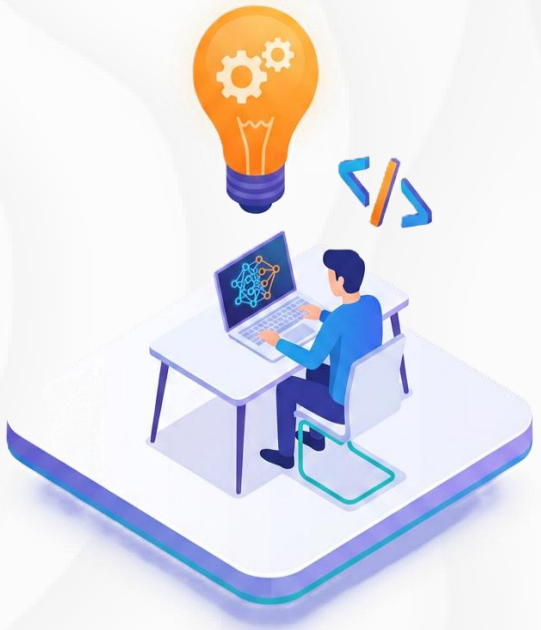
KPI

Focus: qualità dell'output

Task success rate =

$$\frac{\text{n. di volte in cui il sistema ha completato correttamente la task}}{\text{n. di tentativi}}$$

Manuale (o via LLM-as-a-judge) su casi campione. I volumi sono troppo bassi per metriche statistiche.



PILOT PRODUCTION SCALING

Softlab

Cost Engineering nel ciclo di vita di un progetto AI

FinOps: Inform

Ottimizzazioni tecniche, scelte organizzative e profilo di costo

PILOT

Validare a volume reali e stabilire la baseline economica

Scelte organizzative

- Definizione owner del costo
- Ingresso di Finance
- KPI e alert threshold

Tracing & Cost allocation OpenTelemetry/Langfuse		*Cost Allocation Coverage* *Cost per prompt*
Business task definition		*Cost per Task*
Prompt optimization		*I/O Ratio* *Avg. Output Token*
Financial guardrails & Alerting		*Forecast accuracy*

POC — PRODUCTION — SCALING

Softlab

Cost Engineering nel ciclo di vita di un progetto AI

Ottimizzazioni tecniche, scelte organizzative e profilo di costo

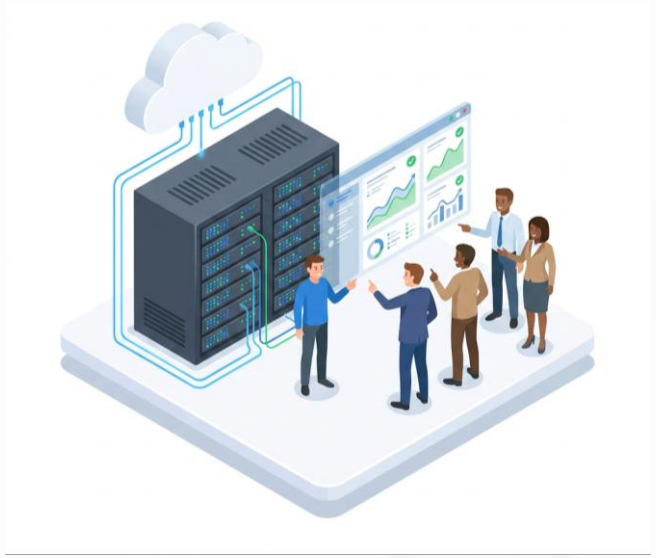
FinOps: Optimize

PRODUCTION

Servire in modo affidabile e ottimizzato, sotto SLA

Scelte organizzative

- SLA e ownership del costo
- Governance dei budget
- Quota management



Ottimizzazioni tecniche

- Rate optimization- Prompt caching, batch API per flussi async
- Model routing con complexity classifier
- Usage optimization – Context pruning, re-ranking, riduzione retry
- Deterministic control layer – Semantic cache, intent routing

POC

PILOT

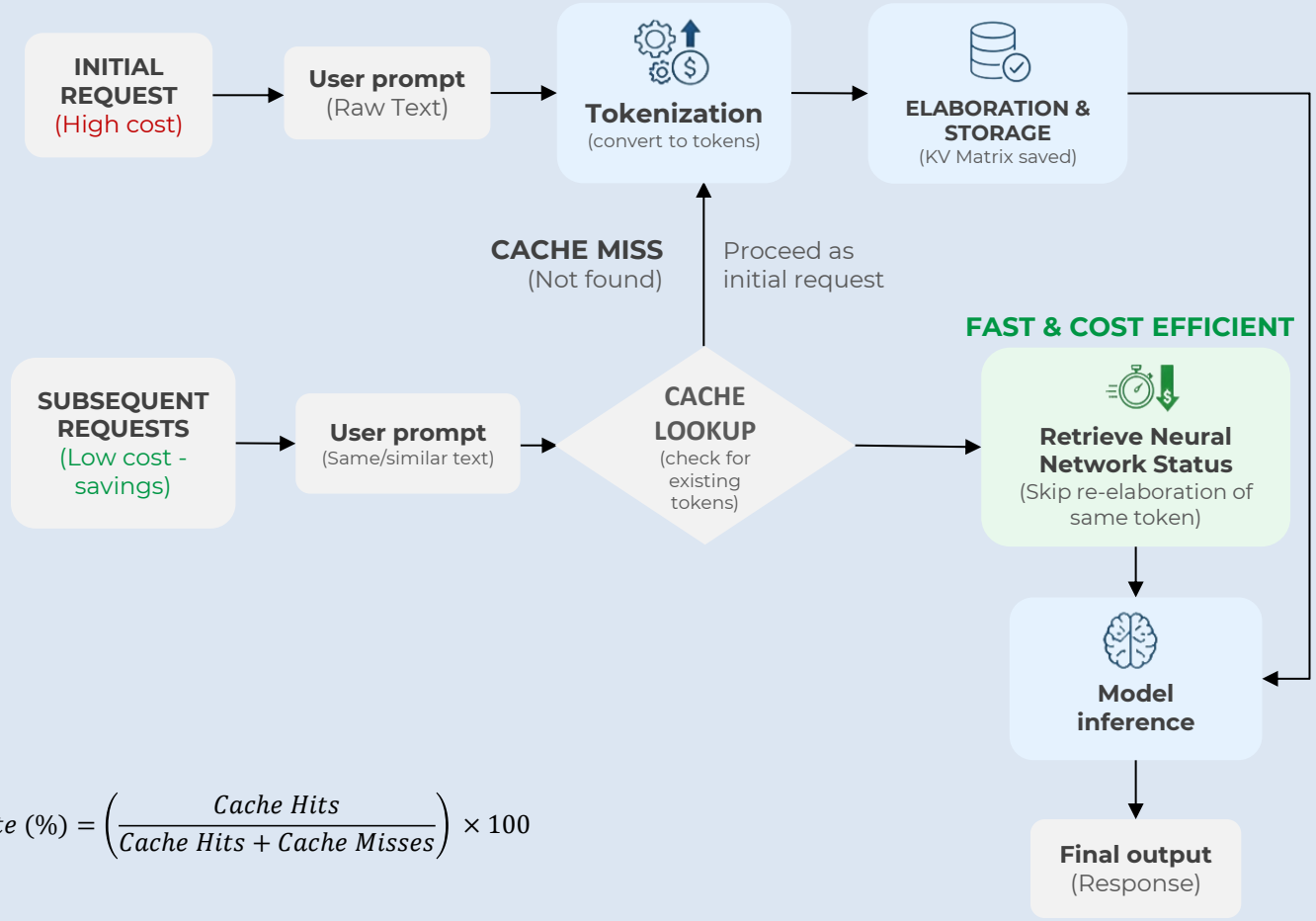
SCALING

Softlab

PROMPT CACHING

Latenza più bassa
Abbattimento dei tempi di risposta del modello

Taglio dei costi
Tariffe inferiori per i token input in quanto non rielaborati dal modello.



$$Cache\ Hit\ Rate\ (\%) = \left(\frac{Cache\ Hits}{Cache\ Hits + Cache\ Misses}\right) \times 100$$

SEMANTIC CACHING

(cache su testi *semanticamente* simili)

Salvataggio di token di input e testo in output per evitare re-run di query di input molto simili che prevedono stessa risposta

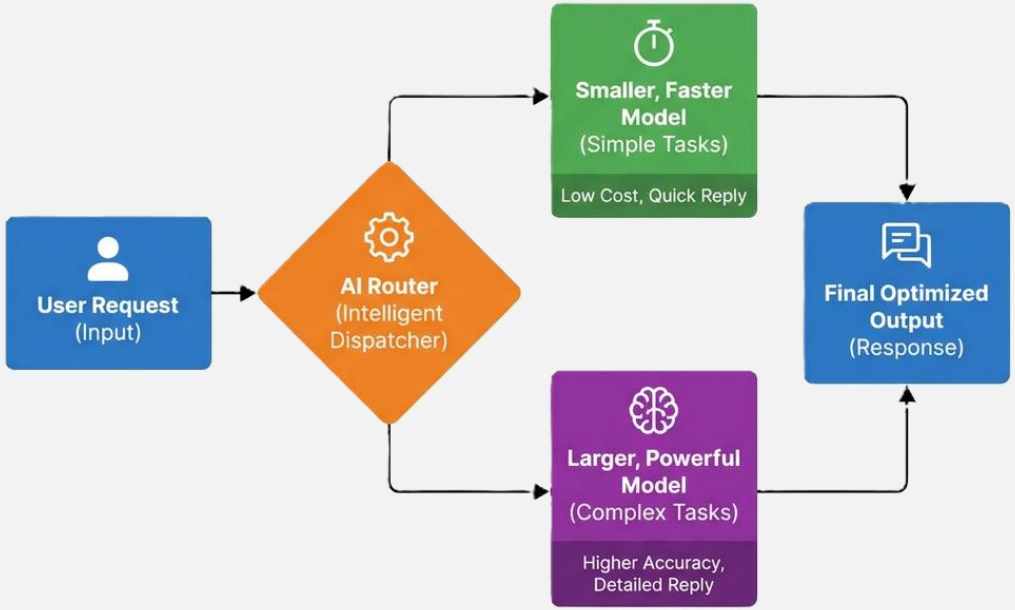
Softlab

MODEL ROUTING

Efficienza di costo
Evita lo spreco di risorse costose su attività banali

Velocità
Risposte più rapide grazie all'uso di modelli snelli quando possibile

Rapporto Performance/Prezzo:
Ottimizzazione dinamica per ogni singola chiamata.



$$\text{Tasso di Overshooting (\%)} = \left(\frac{\text{Numero di Query Semplici inviate a Modelli Complessi}}{\text{Numero Totale di Query Semplici}} \right) \times 100$$

BATCH API

Raggruppamento di grandi volumi di query non urgenti per un'elaborazione asincrona

Sconti sui token
Taglio dei costi fino al 50%

Rate limit più alti
Ottimizzato per l'elaborazione di dati massivi offline

RERANKING (RAG)

Step intermedio nei RAG che analizza e riordina i documenti estratti, passando all'LLM solo quelli davvero pertinenti.

Accuratezza
LLM riceve solo le informazioni più rilevanti

Meno allucinazioni e retry
Risposte corrette al primo colpo, riducendo i costi di calcolo.

CONTEXT PRUNING

Riduzione mirata del testo recuperato prima di inviarlo all'LLM, eliminando frasi o paragrafi ridondanti.
LLM Lingua

No *lost in the middle*
Evita che il modello si perda in contesti troppo lunghi.

Efficienza
Meno token in input (costo ridotto) e latenza minimizzata.

Cost Engineering nel ciclo di vita di un progetto AI

Ottimizzazioni tecniche, scelte organizzative e profilo di costo

SCALING

Sostenibilità economica ad alti volumi e decisione architetturale

FinOps: Operate

Approccio organizzativo

- Da progetto a portfolio (AI Cost Council)
- Procurement maturity per contratti AI
- Capacità di disinvestire ciò che non genera valore

Leve tecniche

- Reserved capacity & enterprise agreements
- Multi-vendor / multi-model
- Portfolio-level optimization (platform interna)
- Internalizzazione selettiva (fine-tuning / self-hosting/..)

KPI DI SCALING

Vendor concentration / lock-in

Break-even per opzione architetturale

Distribuzione contribution margin sul portfolio

POC

PILOT

PRODUCTION

Softlab

Considerazioni finali

L’approccio FinOps per la GenAI è un allineamento organizzativo prima che uno strumento tecnico

La fase del progetto determina la strategia di ottimizzazione

Le leve tecniche funzionano, ma vanno attivate con i volumi giusti

L'AI scalabile e gestibile non nasce da un modello migliore. Nasce da un'organizzazione che ha imparato a misurare, decidere e ottimizzare al momento giusto.

Softlab

Thank you!

👉 slides & videos: https://www.improove.tech/videos

>>AI CONF 2026

improove