

Sarà l'anno della GenAI?

Innovazione tecnologica per le
imprese: sfide e opportunità

Paola Bonomo
Gennaio 2024

.01

L'anno degli LLM

2023: ChatGPT come Taylor Swift

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* †
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task,

2022

“ In my lifetime, I’ve seen two demonstrations of technology that struck me as revolutionary.

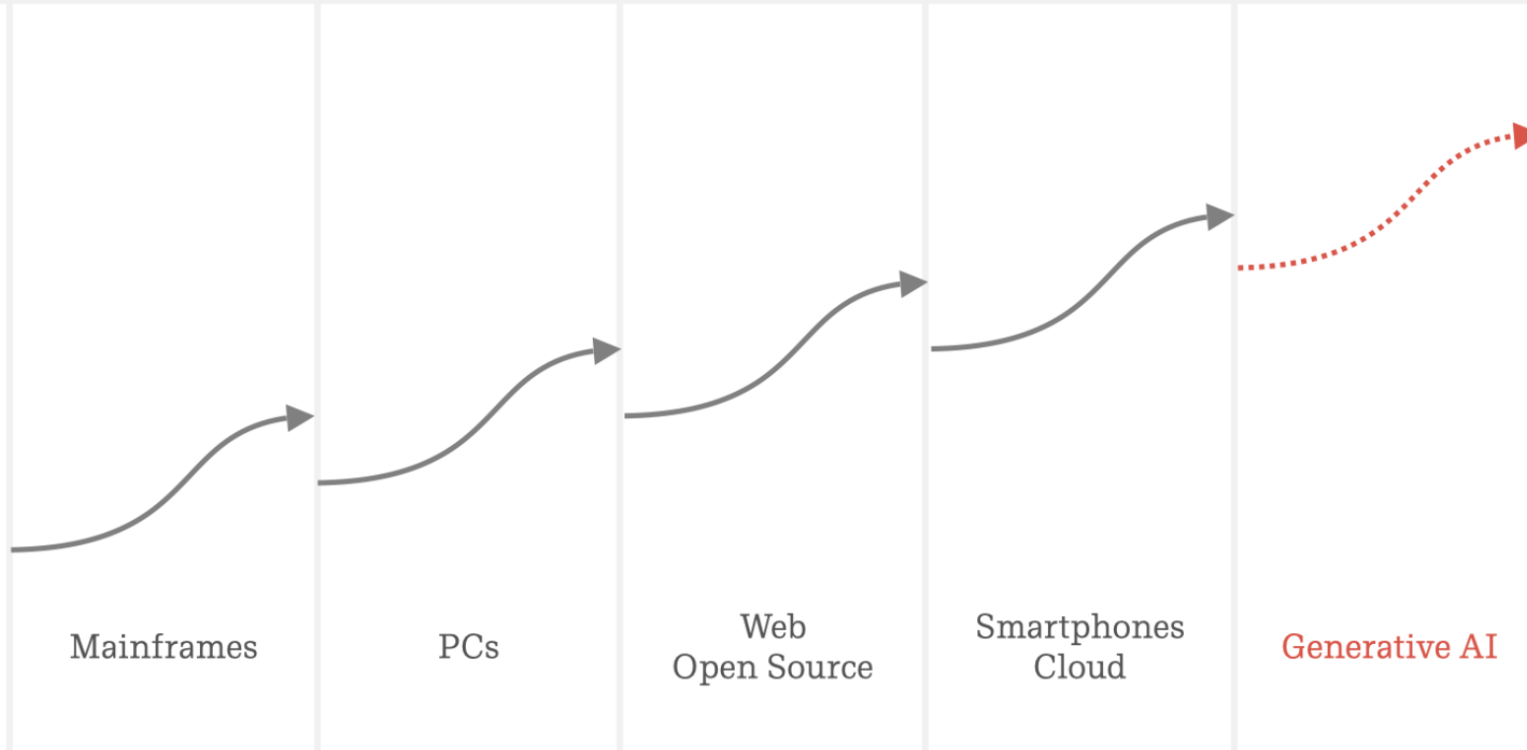
The first time was in 1980, when I was introduced to a graphical user interface [...]

The second big surprise came just last year. [...] GPT got a 5—the highest possible score, and the equivalent to getting an A or A+ in a college-level biology course [...] I knew I had just seen ***the most important advance in technology since the graphical user interface.*** ”



The tech industry moves in platform shifts

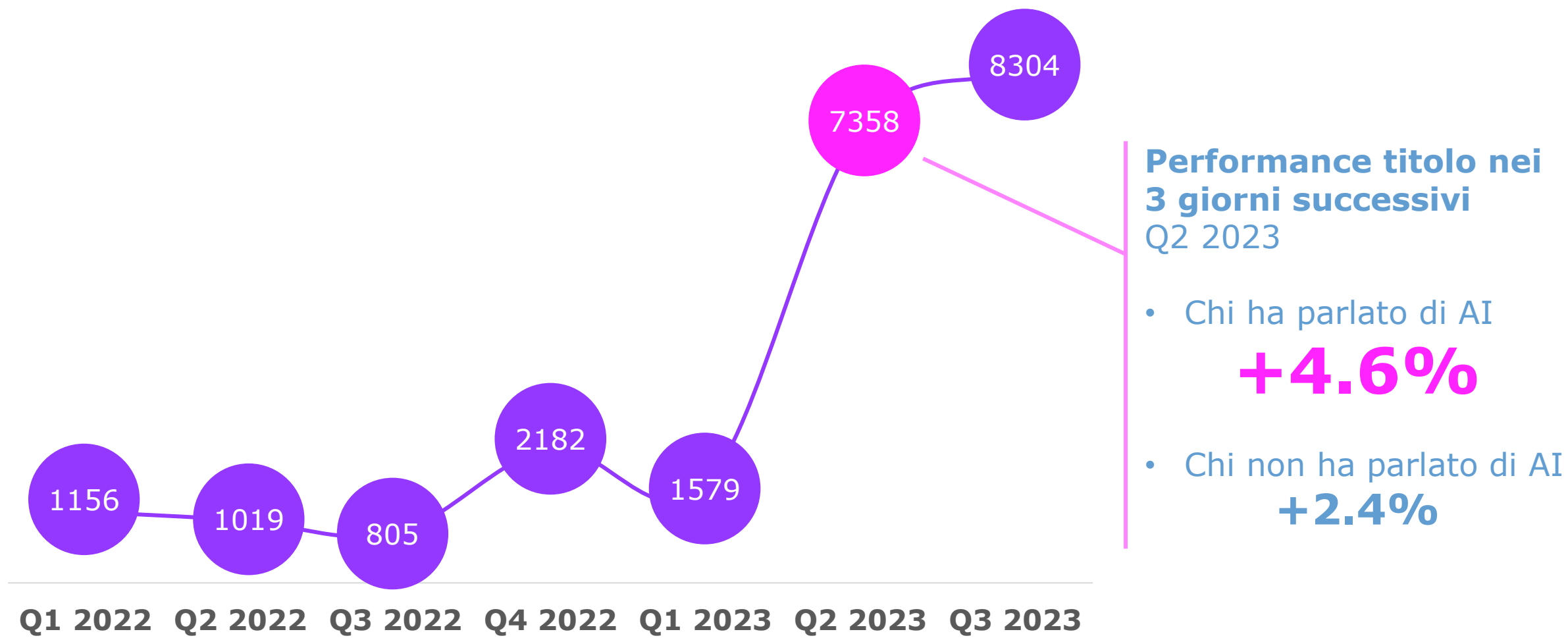
Every 10-15 years, a new platform resets the tech industry



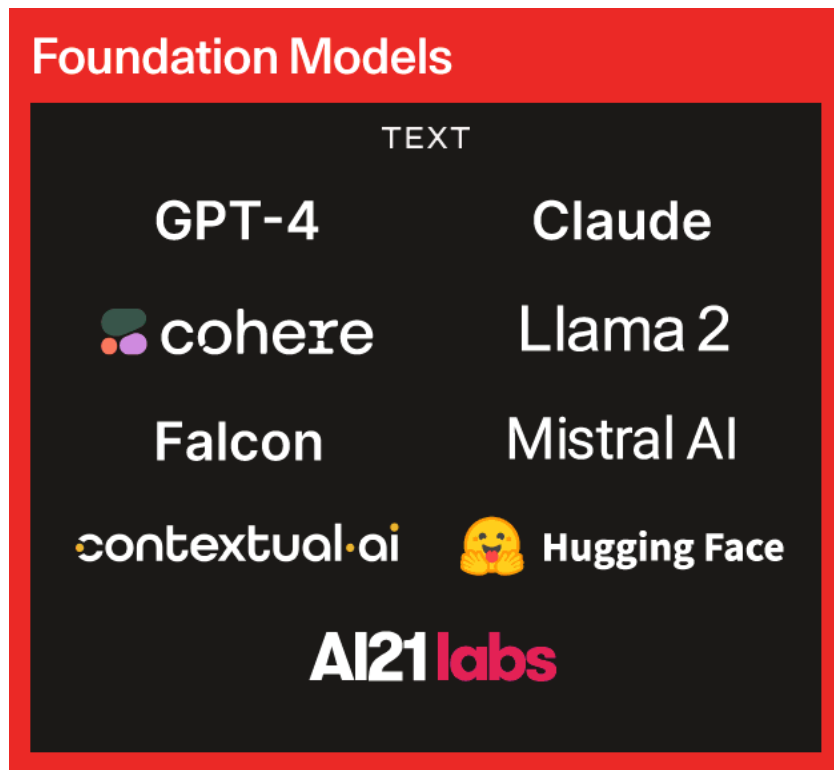


AI e mercati finanziari: citazioni nelle call sui risultati trimestrali

Mercati USA



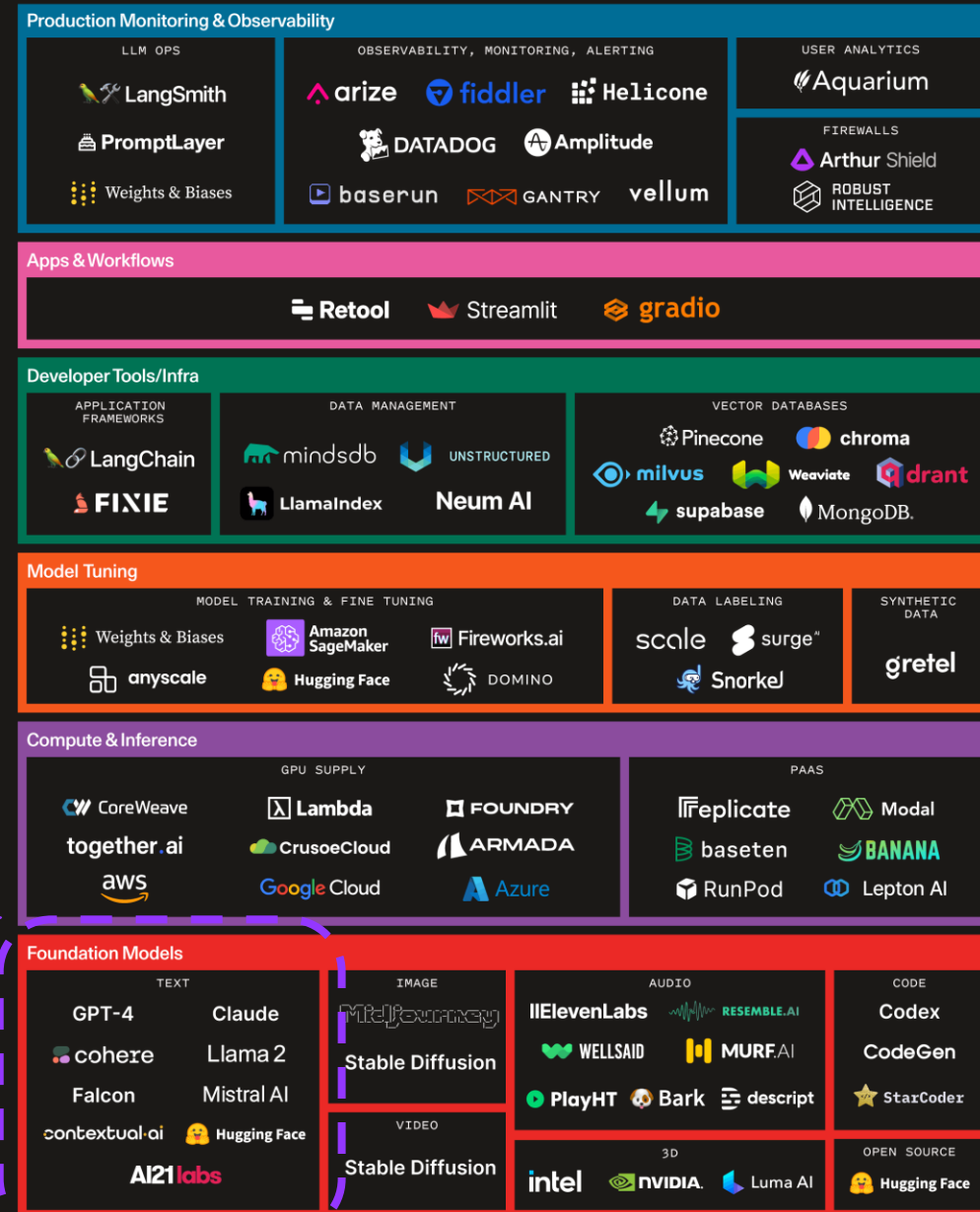
Ma non c'è solo ChatGPT...



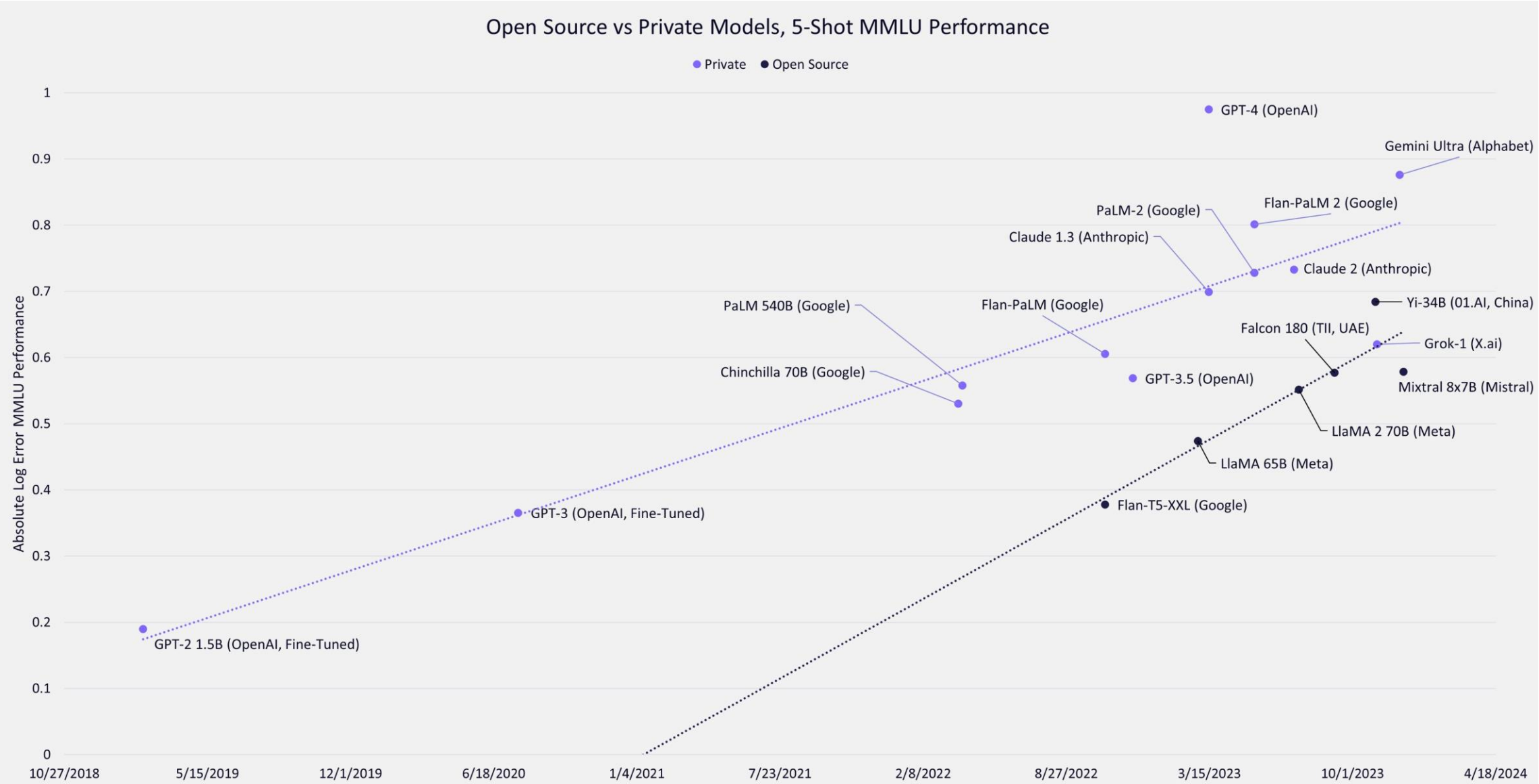
Fonte: <https://www.sequoiacap.com/article/generative-ai-act-two/>

The Generative AI Infrastructure Stack v1

A work in progress



Modelli open source sempre più performanti...



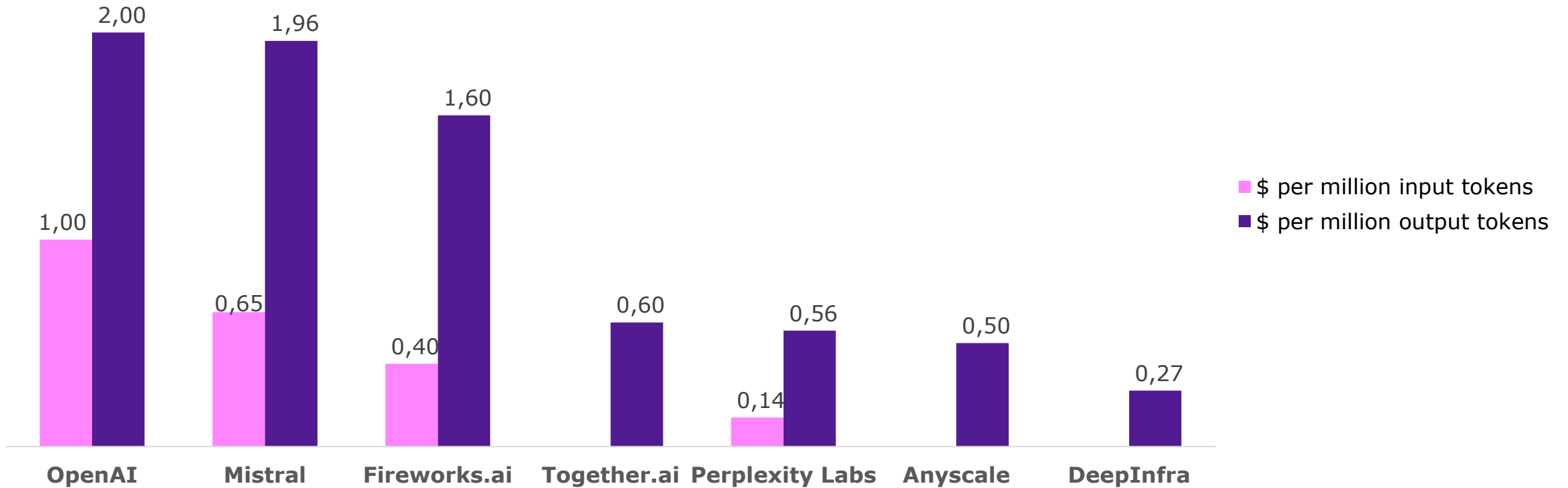
Fonte: ARKInvest, <https://twitter.com/JozefARK/status/1735085757566484493>

... e forte competizione sui costi di utilizzo

Subtitle

Inference costs (12/23)

\$ per million tokens

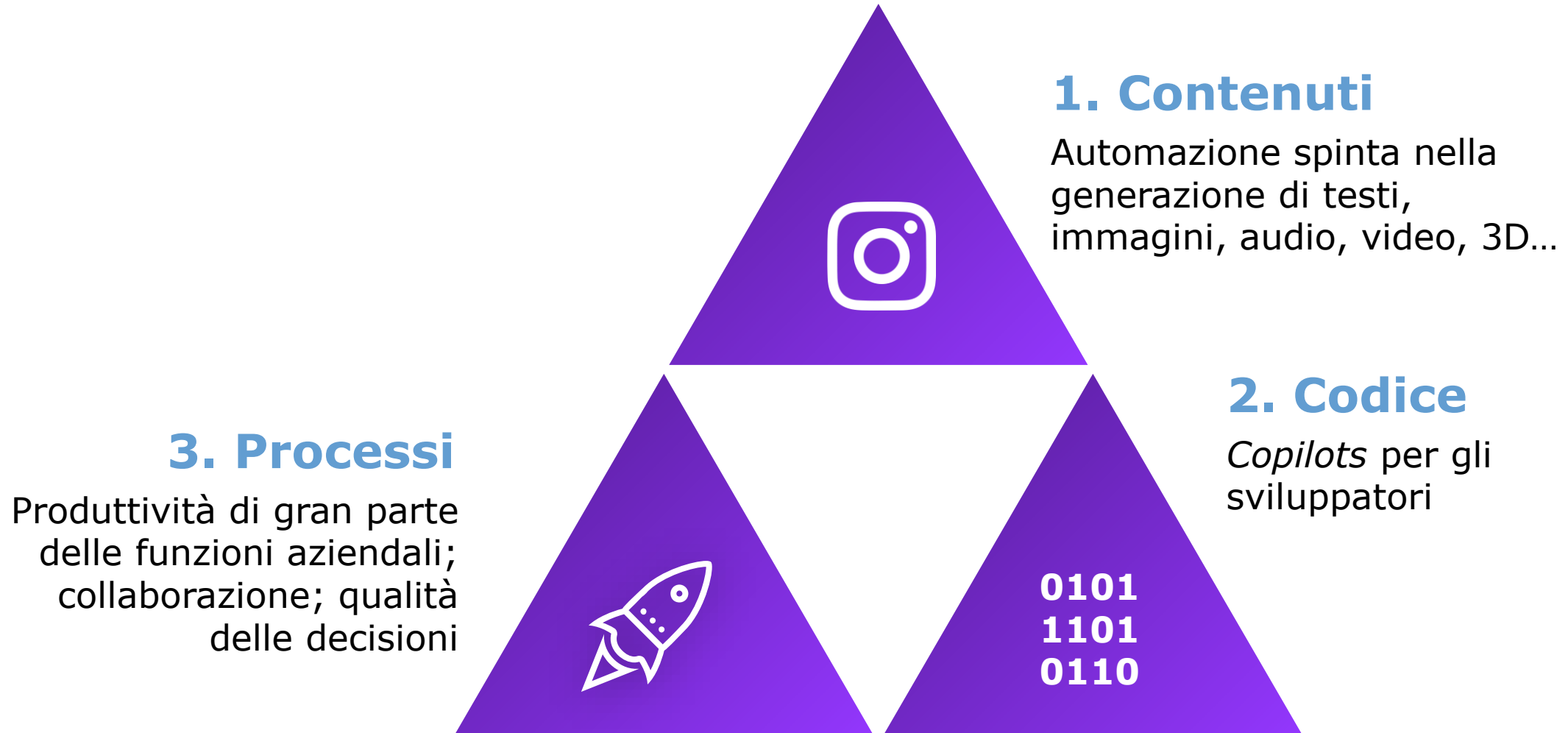


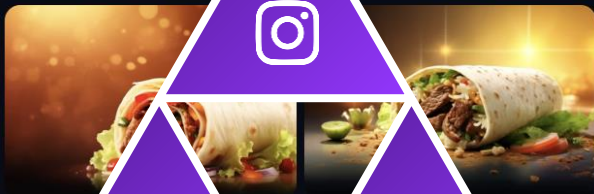
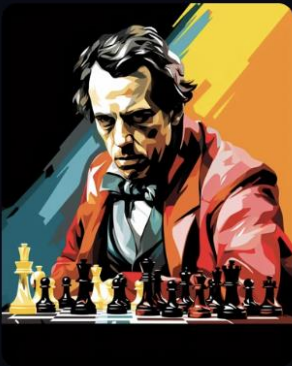
.02

LLM e l'economia globale

Imprese, settori e macroeconomia

Tre strade per l'impatto industriale





SAG-AFTRA.
ON
STRIKE

SAG-AFTRA.
ON
STRIKE!

YOU SHARE THE
WEALTH
BECAUSE
YOU

SAG-AFTRA®
ON
STRIKE!

SAG-AF

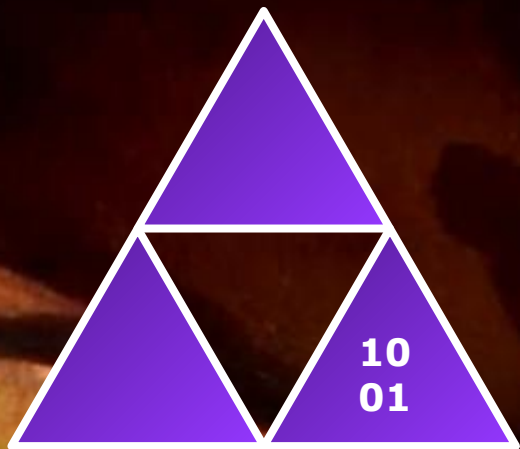
SAG-AFTRA



AFT

STRO

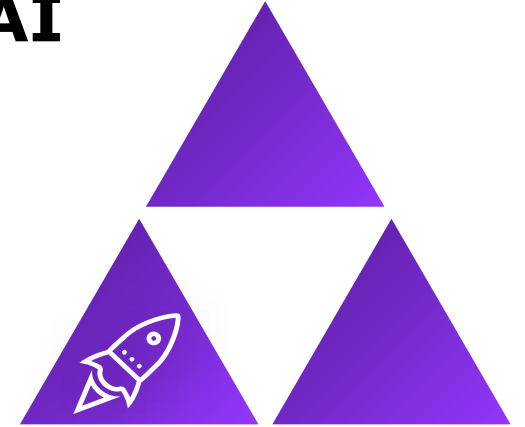
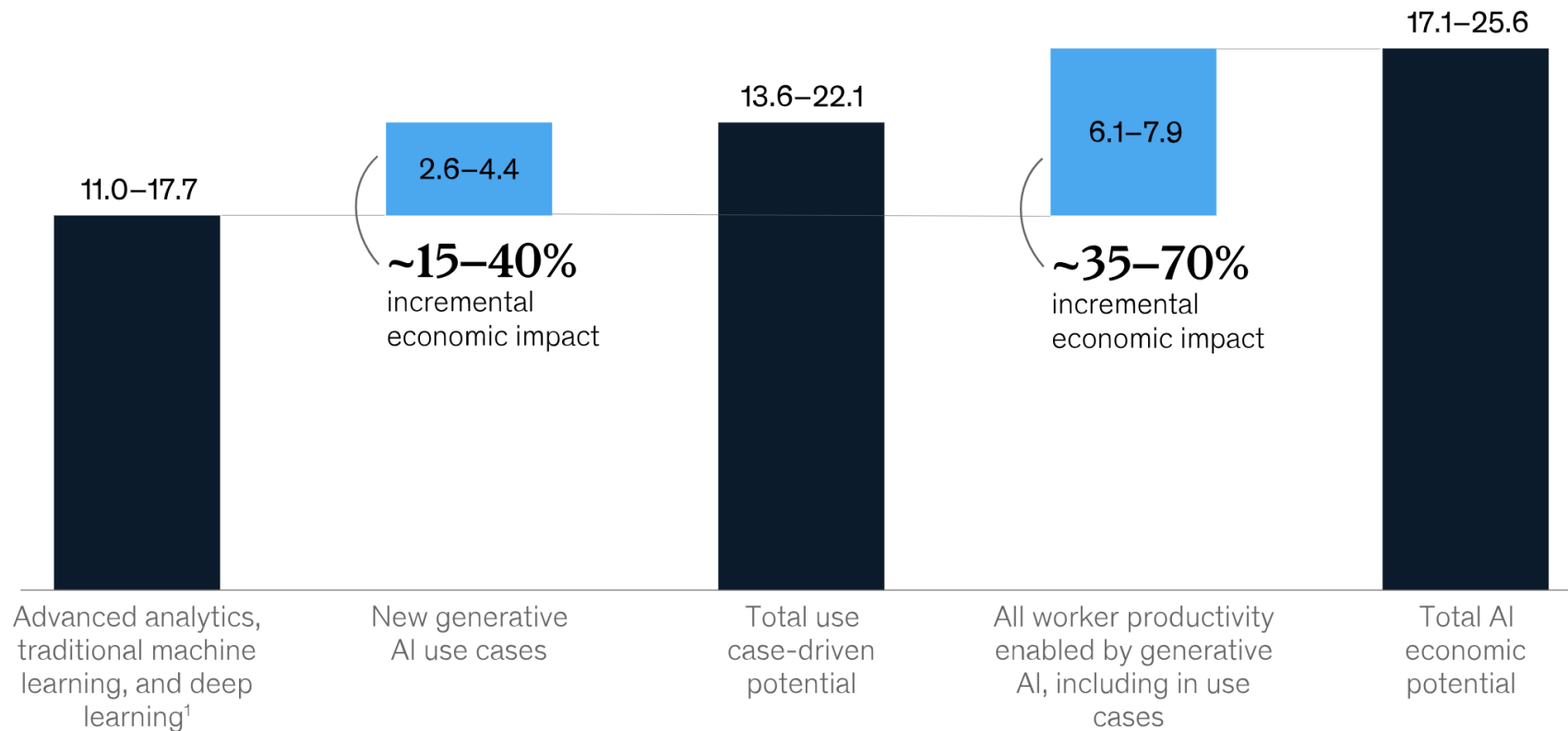




McKinsey prevede un importante impatto della GenAI sulla produttività globale...

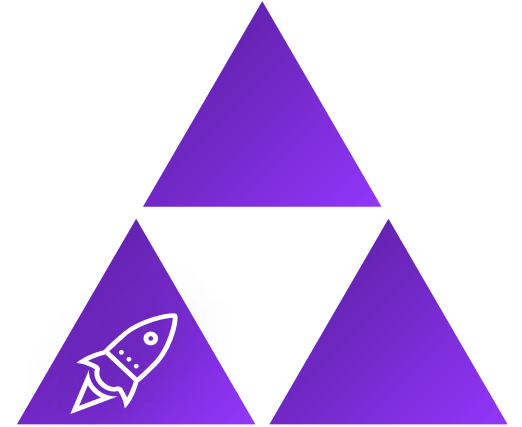
Giugno 2023

AI's potential impact on the global economy, \$ trillion



... così come Goldman Sachs...

Gennaio 2024



As tools using advances in natural language processing work their way into businesses and society, they could drive a

7% (or almost **\$7 trillion**)

increase in global GDP and lift productivity growth by **1.5** percentage points over a 10-year period.



... mentre altri non sono così ottimisti

Agosto 2023

Negli ultimi decenni a grandi innovazioni tecnologiche non si è mai accompagnato un corrispondente balzo della produttività



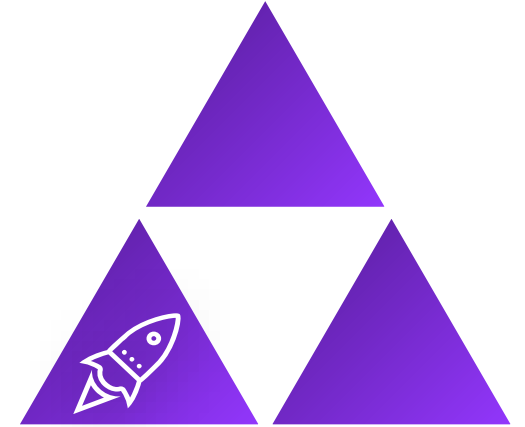
Al contrario, prezzi in calo e ipercompetizione sono sempre andati a beneficio dei consumatori



Chi stima 3 punti percentuali di produttività in più nell'economia USA è troppo ottimista ("too exuberant")



L'impatto sarà limitato e spalmato nel tempo per effetto di ostacoli ad oggi non quantificabili, come barriere regolamentari e accettazione a livello sociale.

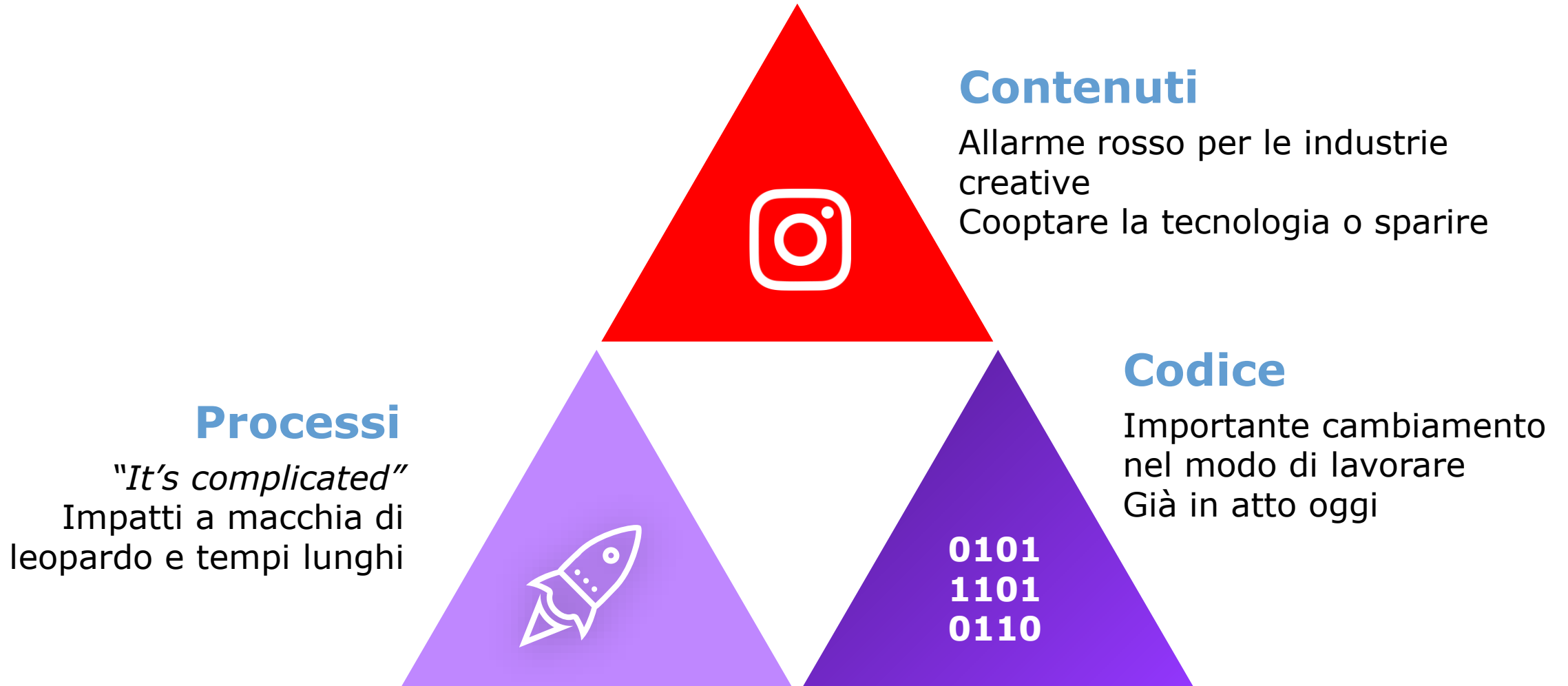


BCG

**BCG
HENDERSON
INSTITUTE**

**WORLD
ECONOMIC
FORUM**

In sintesi



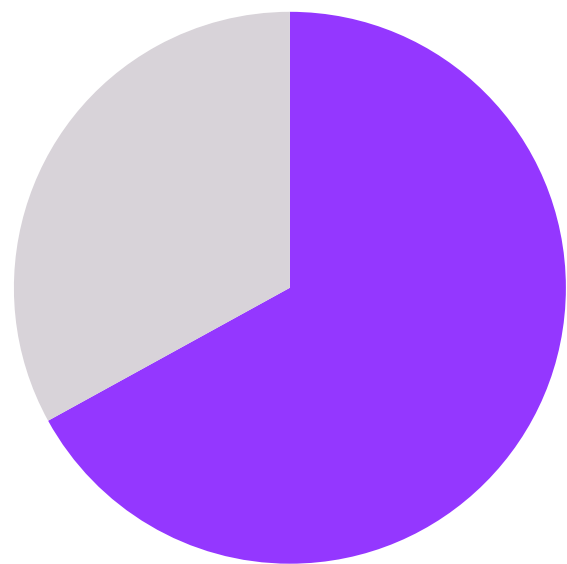
.03

LLM nelle imprese

Successi, dubbi e difficoltà

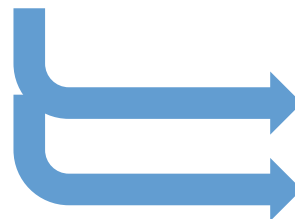
La maggior parte delle aziende è nella fase “lavori in corso”

Novembre 2023



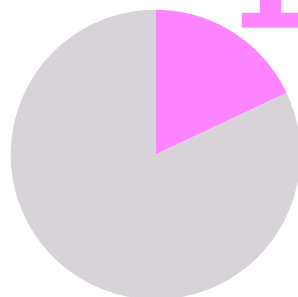
67%

Riferiscono che la loro azienda **sta utilizzando l'AI generativa**



41% Da oltre 1 anno

26% Da meno di 1 anno



18%

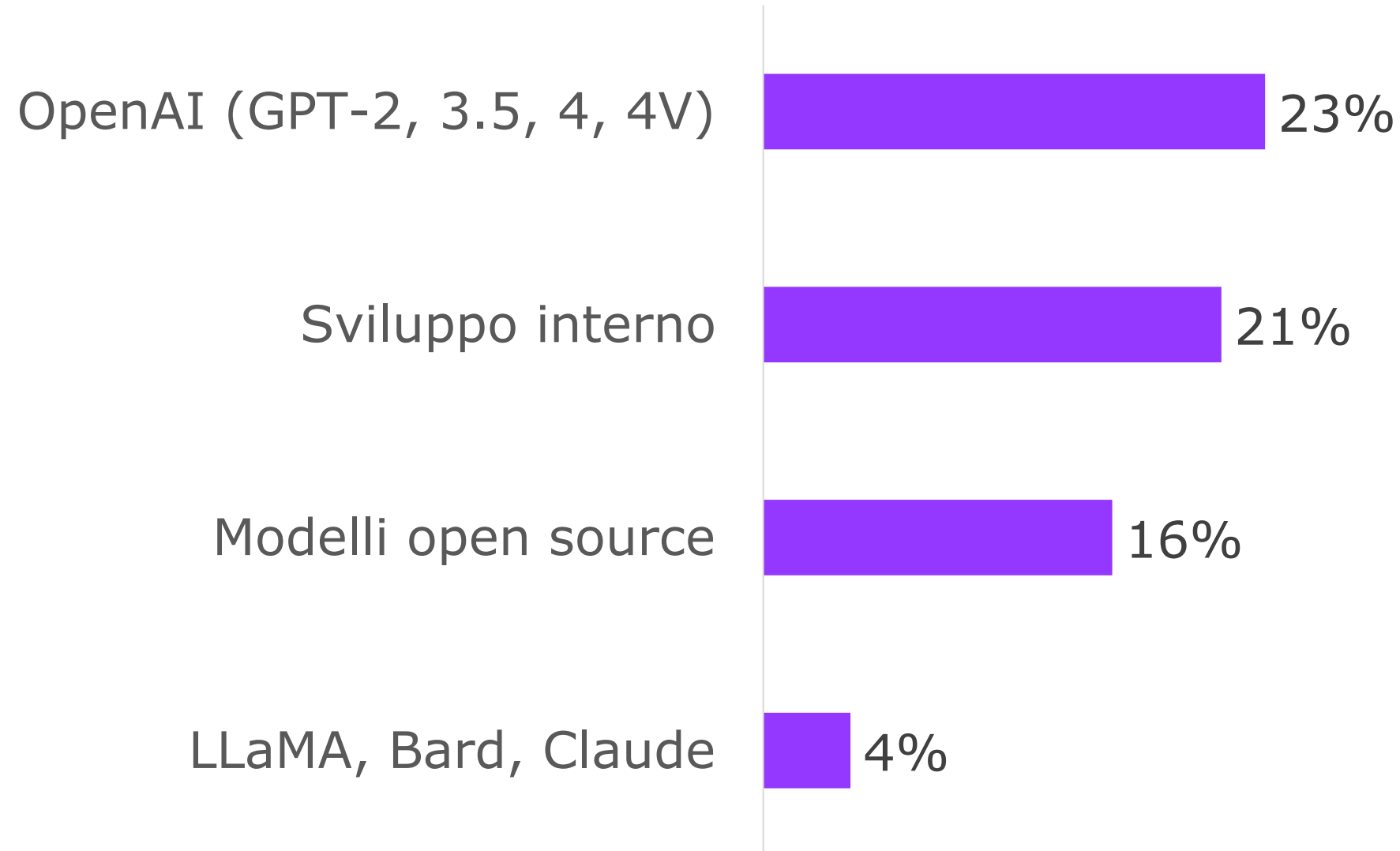
Hanno già **applicazioni in produzione**



Gli altri: Proof of Concept, sviluppo prodotto, costruzione modello, testing

Quale LLM?

Novembre 2023



Fonte: <https://www.oreilly.com/radar/generative-ai-in-the-enterprise/>

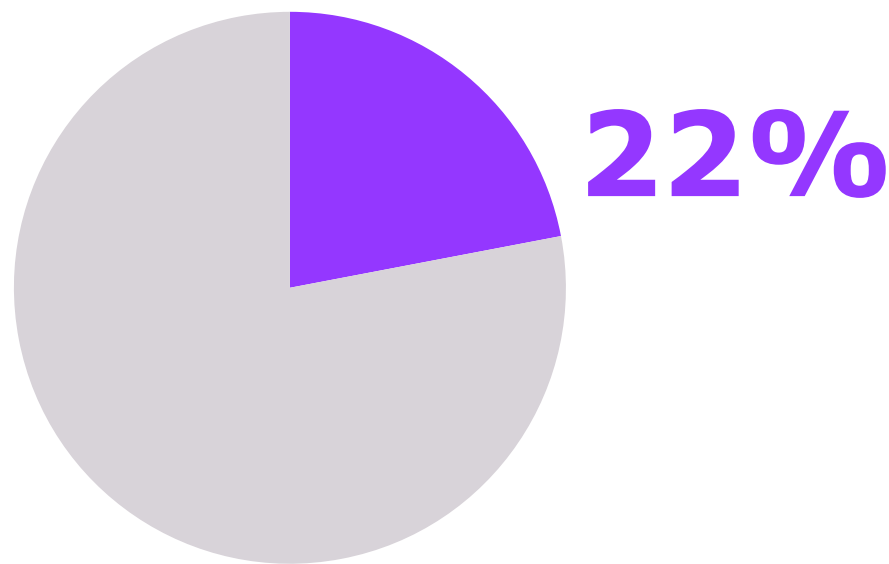
Quali aspettative?

Novembre 2023



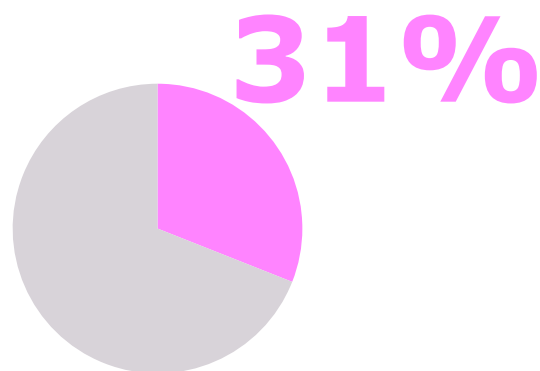
Trovare i casi d'uso è la principale barriera...

Novembre 2023



22%

Tra le aziende che stanno usando l'AI generativa, il 22% riferisce che la **difficoltà nel trovare casi d'uso appropriati** è la principale barriera a una maggiore adozione

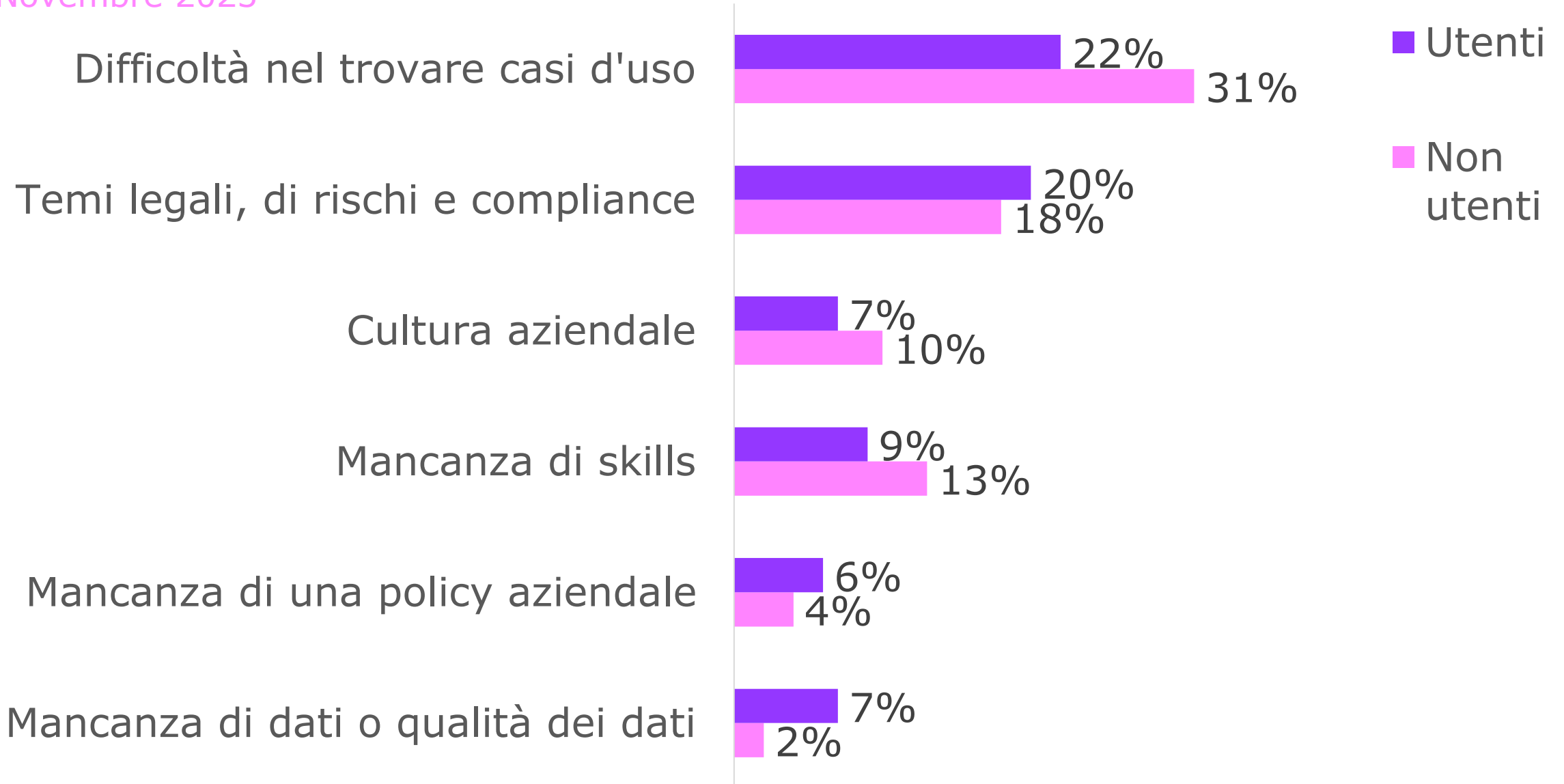


31%

Tra quelle che non la stanno usando, il 31% dice che la difficoltà nel trovare casi d'uso appropriati è il motivo principale

... ma non è l'unica difficoltà

Novembre 2023



Fonte: <https://www.oreilly.com/radar/generative-ai-in-the-enterprise/>

Molti dei temi sono gli stessi delle trasformazioni digitali degli ultimi 10 anni



In più: le “allucinazioni” degli LLM



AI was asked to create images of Black African docs treating white kids. How'd it go?

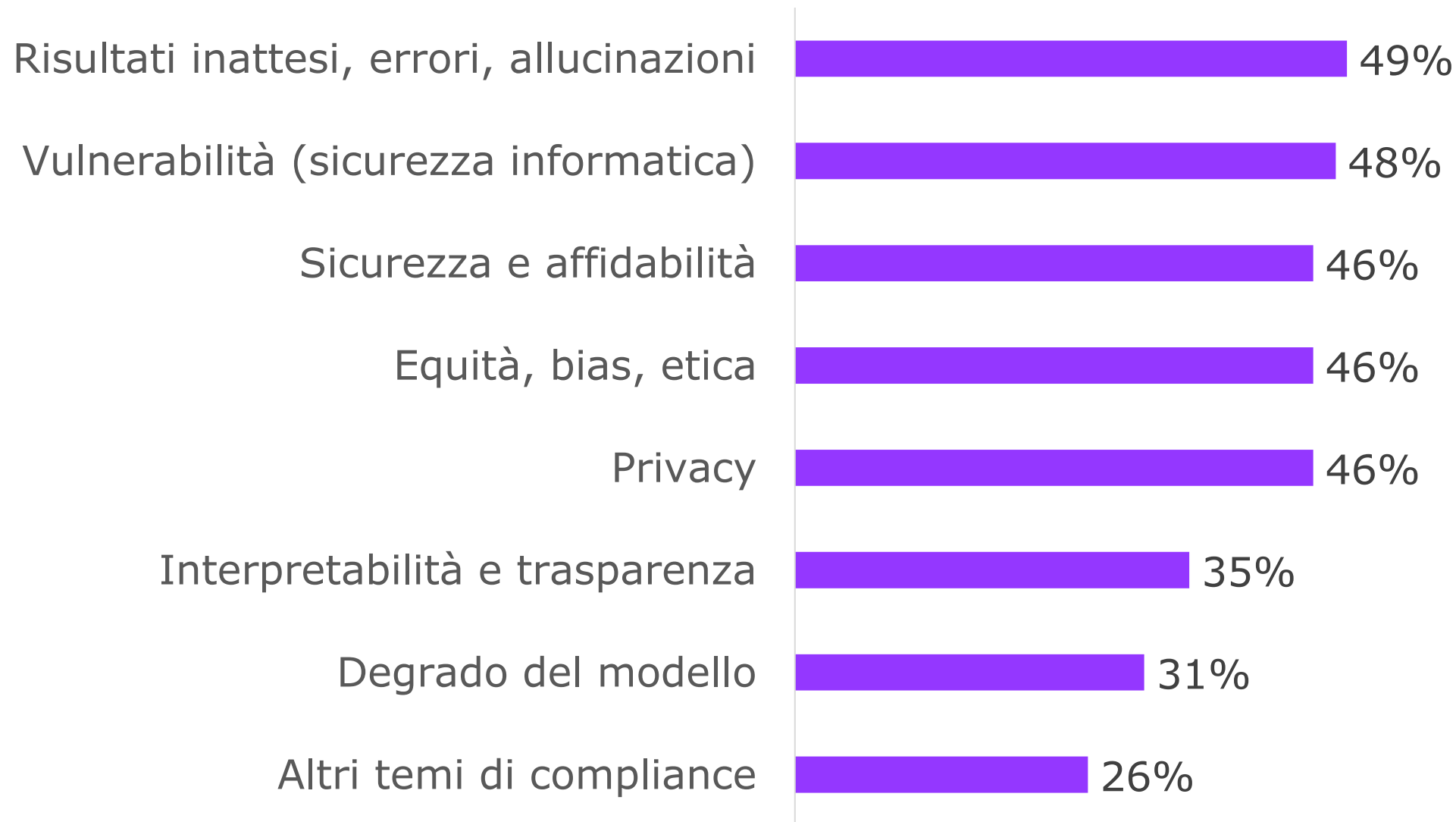
OCTOBER 6, 2023 · 7:44 AM ET



Fonte: Oxford-Johns Hopkins Global Infectious Disease Ethics Collaborative,
<https://www.npr.org/sections/goatsandsoda/2023/10/06/1201840678/ai-was-asked-to-create-images-of-black-african-docs-treating-white-kids-howd-it->

I principali rischi da identificare e mitigare in fase di test

Novembre 2023



.04

Fa' la cosa giusta

Governi, stakeholder, persone

Tecnologia problematica? Temi noti da sempre, ma in *overdrive*

Proprietà dei dati e
proprietà intellettuale

Regolamentazione e
moderazione dei
contenuti

Dominanza del mercato
e pratiche
anticoncorrenziali

Pratiche lavorative e
futuro dei posti di lavoro

Usi sensibili vs. libertà
civili (es.
riconoscimento facciale)

Tassazione: dove viene
creato il valore?

