

E4

COMPUTER
ENGINEERING

WHEN PERFORMANCE MATTERS

STATO DELL'ARTE DI AI & HPC E TOOLBOX

Evento «Supercalcolo e Intelligenza Artificiale», Unindustria RE

Daniele Cremonini, Head of Aerospace & Defence BU
7 Novembre 2024

CONFIDENTIAL

www.e4company.com

E4 Computer Engineering progetta e realizza soluzioni altamente tecnologiche per cluster HPC, Cloud, Data Analytics, Intelligenza Artificiale e infrastrutture **Hyper-Converged** per il mercato accademico e industriale. Da anni collaboriamo con i principali centri di ricerca a livello nazionale e internazionale (CINECA, CERN, ECMWF, LEONARDO) e siamo coinvolti in **progetti nazionali ed europei** in ambito HPC e AI.

Ogni soluzione E4 è **UNICA**, come ogni nostro cliente; **TESTATA** in ogni singolo componente; **VALIDATA** per verificare le effettive prestazioni di ogni sistema e **SERVITA** da tecnici che forniscono assistenza nelle più estese e complesse infrastrutture informatiche italiane ed europee.

E4 ANALYTICS

LET YOUR DATA PAY YOUR GROWTH

Attraverso la consociata E4 Analytics, E4 lavora per integrare l'**Intelligenza Artificiale** e la **Data Science** nelle organizzazioni che intraprendono la **Trasformazione Digitale** del proprio business per migliorare prodotti/processi e ottimizzare le risorse. Operiamo all'intersezione tra business e tecnologia, supportando il cliente nell'adozione di **soluzioni GenAI** personalizzate e sicure: con E4 Analytics, i dati aziendali diventano una **risorsa per creare valore**, migliorare l'**innovazione** e la **competitività** sul mercato.

E4 IS THE ONLY PREMIUM SCIENTIFIC AND TECHNOLOGICAL SYSTEMS INTEGRATOR
WITH A HIGH LEVEL OF EXPERTISE TO SOLVE THE LATEST DIGITAL CHALLENGES

LA NOSTRA MISSION

Anticipiamo la trasformazione dirompente e sempre più rapida della nostra epoca, fornendo soluzioni mature e affidabili in ambienti tecnologici sofisticati con un approccio innovativo

LA NOSTRA VISION

Esploriamo scenari futuri per trovare soluzioni a richieste computazionali particolarmente complesse in aree applicative finora inimmaginabili

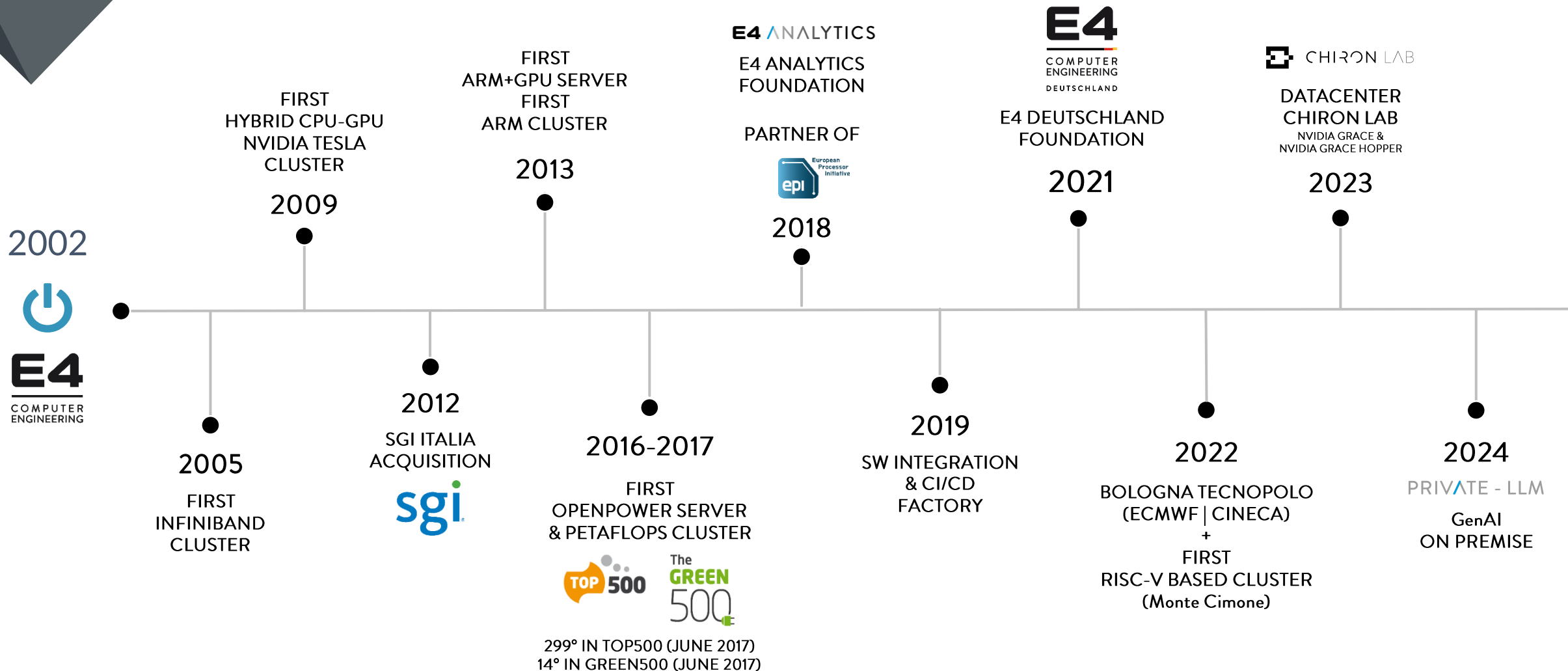
WE DELIVER SOLUTIONS **E**NGINEERED **4** HIGH
PERFORMANCE COMPUTATIONAL INFRASTRUCTURE

WHEN PERFORMANCE MATTERS

E4

COMPUTER
ENGINEERING

UNA STORIA DI INNOVAZIONE



E4 IN NUMERI

+91%



Crescita del fatturato dal 2021

>280



Clienti attivi (ultimi 3 anni)

82

20



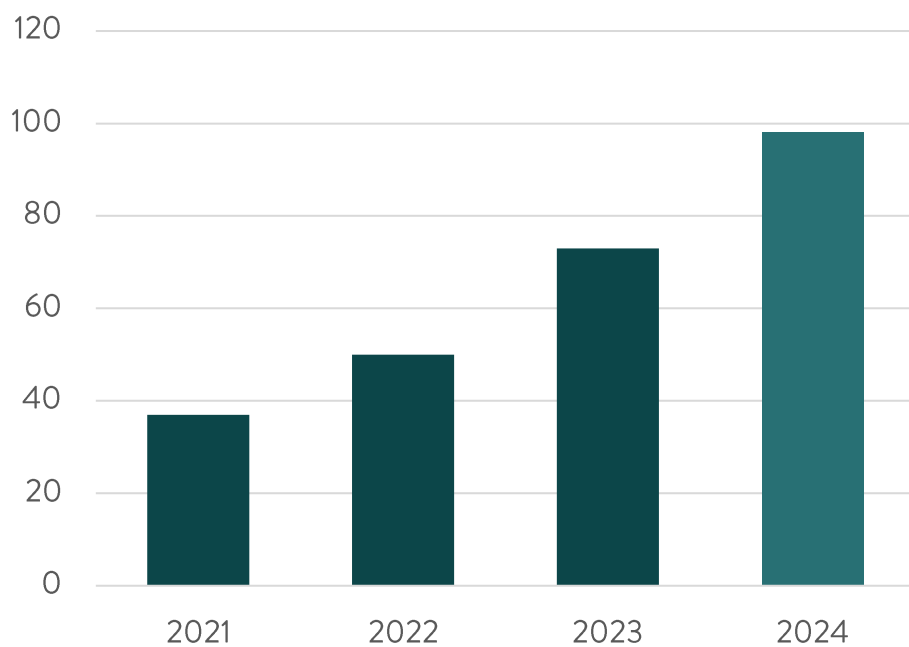
Dipendenti (20 previsti entro il 2024)

22



Progetti Europei con contributo di E4

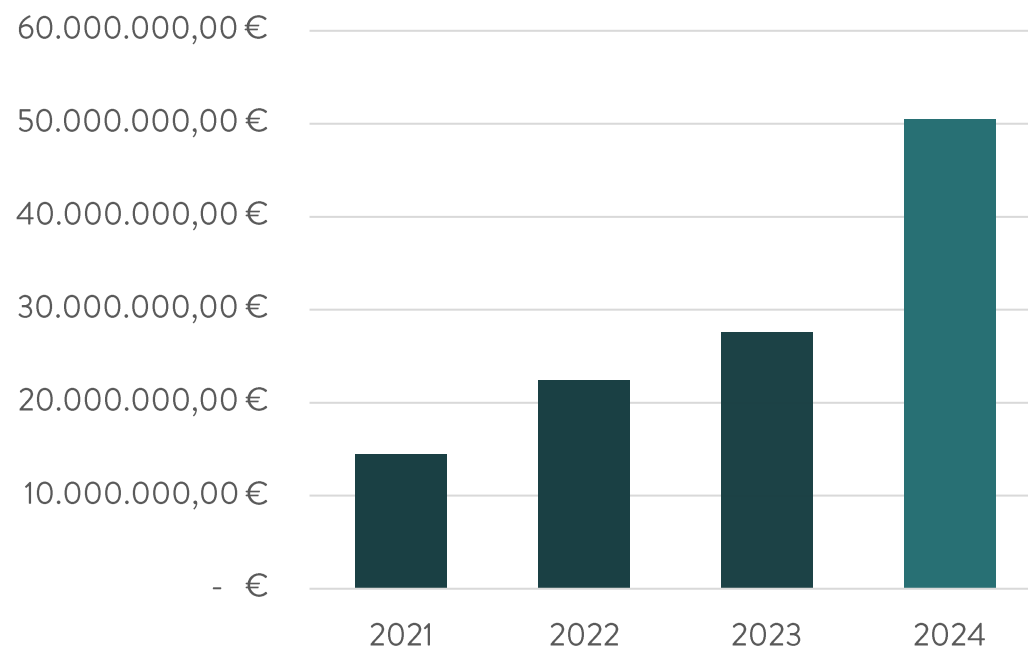
MEMBRI DEL TEAM



>85%

Impiegati in Tech/R&D

FATTURATO



AREE DI INTERESSE

HPC & HARDWARE

Solide competenze
costruite in oltre 20
anni di attività in ambito
di alte prestazioni

AI & PRIVATE LLM

Sfruttare il potenziale dell'AI
per trasformare i dati aziendali
in vantaggi competitivi, con
attenzione a sicurezza e privacy

SW & OPEN SOURCE

Professionisti esperti nella
valutazione di software
opensource o nell'integrazione di
classe enterprise

MERCATI DI RIFERIMENTO



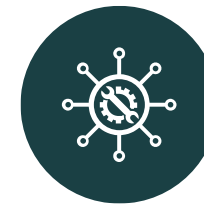
EDUCAZIONE



FINANZA



FARMACEUTICA



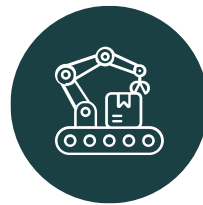
FORNITORI
DI SERVIZI



GOVERNO



AEROSPAZIO
& DIFESA



MANIFATTURIERO

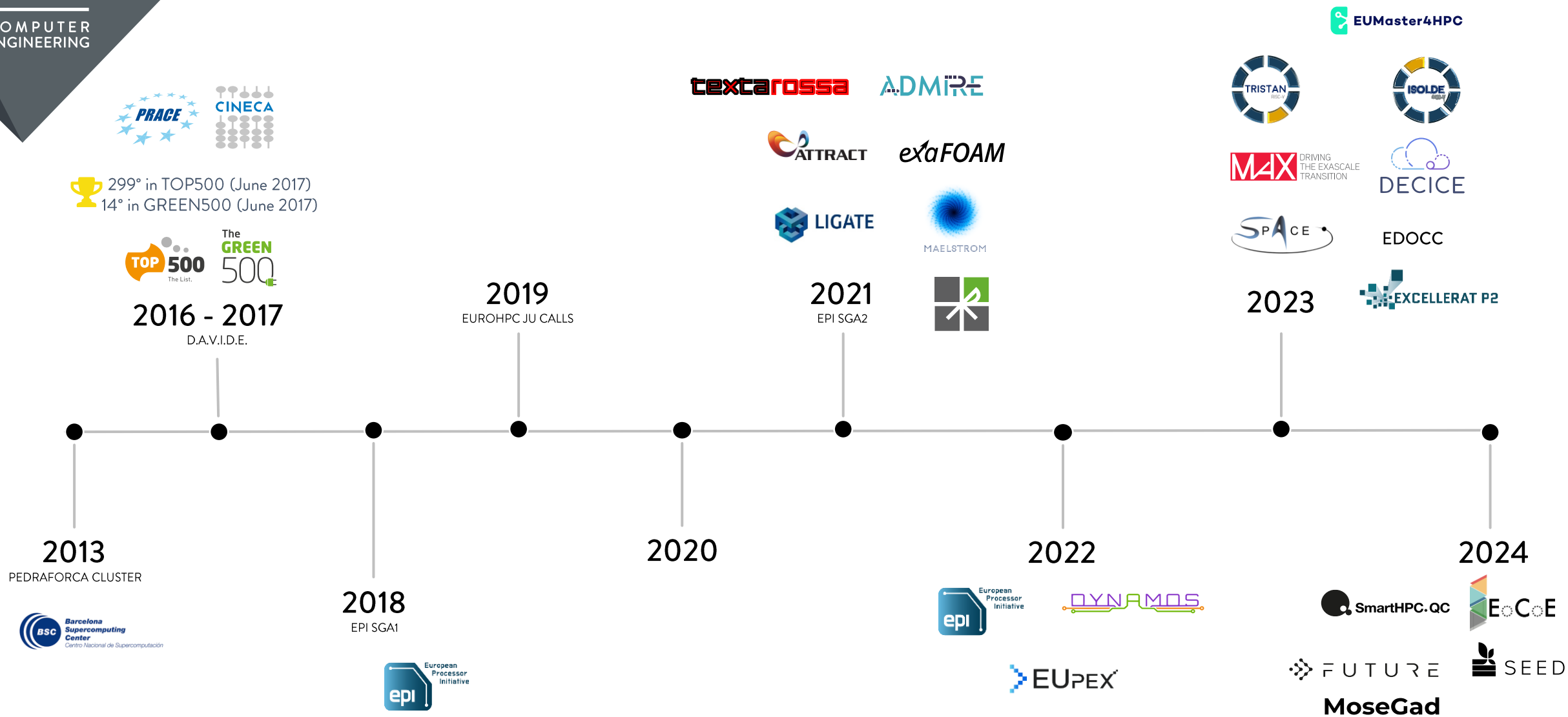


LEGALE

E4

COMPUTER
ENGINEERING

PROGETTI EUROPEI – DOVE CONTRIBUIAMO



Q4_2024 - E4_GAIA_E4_AL SOLUTION.IT

WHEN PERFORMANCE MATTERS



MEMBERSHIPS & PARTNERSHIPS

E4 è membro e partner di diverse organizzazioni/associazioni nel settore dell'innovazione e ha stabilito solide relazioni con fornitori chiave in Europa e nel mondo per accedere a tecnologie all'avanguardia



CERN openlab è un partenariato pubblico-privato, attraverso il quale il CERN collabora con aziende tecnologiche leader e altre organizzazioni di ricerca



QuIC è un'associazione no-profit fondata da grandi imprese, PMI, investitori e startup di tutta Europa. La missione di QuIC è quella di dare impulso all'industria europea delle tecnologie quantistiche



ETP4HPC è un'associazione privata, guidata dall'industria e senza scopo di lucro, la cui missione è promuovere la ricerca e l'innovazione HPC in Europa



HIPEAC è una rete di circa 2.000 ricercatori di sistemi informatici di livello mondiale, rappresentanti dell'industria e studenti



IFAB è una fondazione dedicata al dibattito scientifico e culturale sullo sviluppo umano, la sostenibilità e le scienze emergenti che si svolge a livello globale



OpenFOAM è un toolbox C++ per lo sviluppo di solutori numerici personalizzati e di utilità di pre/post-processing per la soluzione di problemi di meccanica dei continui



ISC-V comprende un'ampia organizzazione di membri che sta costruendo la prima comunità aperta e collaborativa di innovatori del software e dell'hardware che promuovono l'innovazione



The European Alliance for Industrial Data, Edge and Cloud è un'associazione di soggetti pubblici e privati coinvolti nell'archiviazione e nell'elaborazione dei dati che mira a promuovere lo sviluppo e la diffusione di tecnologie edge e cloud di nuova generazione



Il QuEra Quantum Alliance Partner Program mira ad accelerare lo sviluppo, la diffusione e l'utilizzo dei computer quantistici ad atomi neutri per risolvere i problemi di calcolo più impegnativi del mondo

Grazie a un ecosistema ben consolidato e a solide relazioni con i principali fornitori di tecnologie hardware e software, E4 gode di un accesso privilegiato alle tecnologie più recenti





- Scouting, selezione, progettazione, test, integrazione, configurazione, ottimizzazione, validazione e verifica e installazione dell'intero stack di componenti HW e SW e delle infrastrutture
- Servizi di consulenza e di expertise verso i clienti target europei e i prospect sulla soluzione ottimale per le loro esigenze
- Consideriamo lo **sviluppo del prodotto** come un processo senza fine, accettando i fallimenti ma puntando al successo
- Appliciamo il **co-design** in uno scenario in continua evoluzione
- **Approccio agnostico**: supporta la più ampia gamma possibile di componenti e sistemi operativi hw/sw

IL NOSTRO PORTAFOGLIO DI SOLUZIONI



MEDOOZA



ICE4HPC

HIGH
PERFORMANCE
COMPUTING



FLUCTUS

HIGH
PERFORMANCE
CLOUD



GAIA

GENERATIVE
ARTIFICIAL
INTELLIGENCE
APPLIANCE



URANIA

CLOUD-NATIVE
DATA
SCIENCE



KAPTAIN

HIGH
PERFORMANCE
KUBERNETES
CLUSTER



STONES

VIRTUALIZATION
SOLUTIONS

OUR SOLUTIONS



E4

COMPUTER
ENGINEERING



GAIA

GENERATIVE
ARTIFICIAL INTELLIGENCE
APPLIANCE

GAIA 3.0: LA SOLUZIONE DI E4 PER L'INTELLIGENZA ARTIFICIALE GENERATIVA

Q4 2024

www.e4company.com

L'APPLIANCE DISEGNATA PER PORTARE LA GENAI ON-PREMISE



GAIA

GAIA è la soluzione che offre servizi on-premise di **Intelligenza Artificiale Generativa**, ad alte prestazioni, scalabili, sicuri e basati su Large Language Model open-source stato dell'arte

INGEGNERIZZATA, NON ASSEMBLATA

COS'È LA GENERATIVE AI?

ChatGPT



Examples

"Explain quantum computing in simple terms" →

"Get any creative ideas for a 10 year old's birthday?" →

"How do I make an HTTP request in Javascript?" →



Capabilities

Remembers what user said earlier in the conversation

Allows user to provide follow-up corrections

Trained to decline inappropriate requests



Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge - it won't answer questions about events after 2021



HRDIVE

Deep Dive Opinion Library Events Press Releases

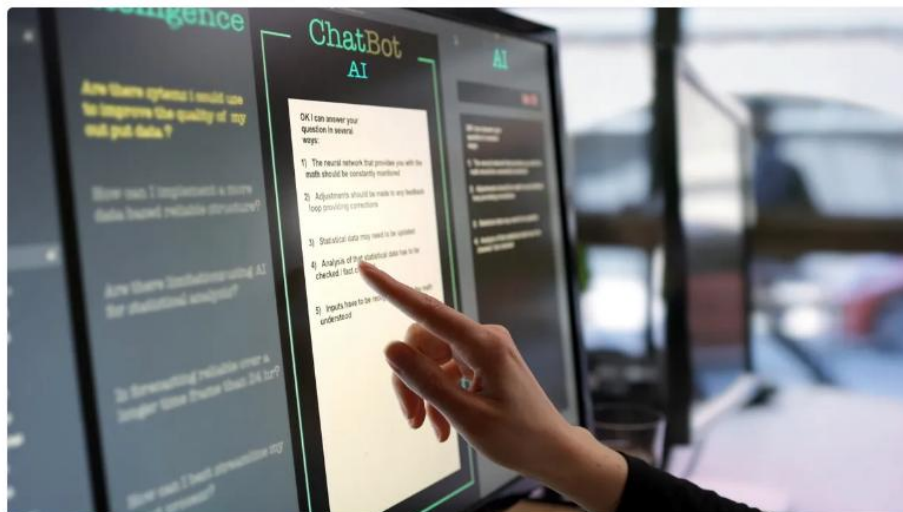
Compliance Talent Learning Comp & Benefits Diversity & Inclusion HR M

ChatGPT usage at work has more than doubled in a year

Now more than half of professionals say they use ChatGPT or other AI tools at work, a Glassdoor report shows.

Published Jan. 31, 2024

By Carolyn Crist



Generation Z leads the way in AI tool use, a Glassdoor survey shows. Laurence Dutton via Getty Images

Forbes

FORBES > LEADERSHIP > CAREERS

EDITORS' PICK

3 Career Benefits Of ChatGPT Use, Based On Age And Education Trends

Tracy Brower, PhD Senior Contributor @

Dr. Tracy Brower writes about happiness and the future of work.

[Follow](#)

0

Mar 29, 2024, 06:32am EDT

Updated Mar 29, 2024, 12:30pm EDT



There is tremendous power and promise in ChatGPT for your career. GETTY

The use of ChatGPT is exploding—both on the job and outside of work—and you are wise to join the wave, so you can positively impact your work and advance your career.

Fonte: HRDive, Forbes



Attività scritte:

- Comunicati stampa
- Rapporti brevi
- Piani di analisi
- E-mail sensibili

Risultati:

- Produttività **+40%**
- Qualità **+18%**
- Adozione di ChatGPT
dopo l'esperimento **x2**

Fonte: MIT,
<https://www.science.org/doi/10.1126/science.adh2586>

WHEN PERFORMANCE MATTERS

Experimental evidence on the productivity effects of generative artificial intelligence

SHAKKED NOY  AND WHITNEY ZHANG  [Authors Info & Affiliations](#)

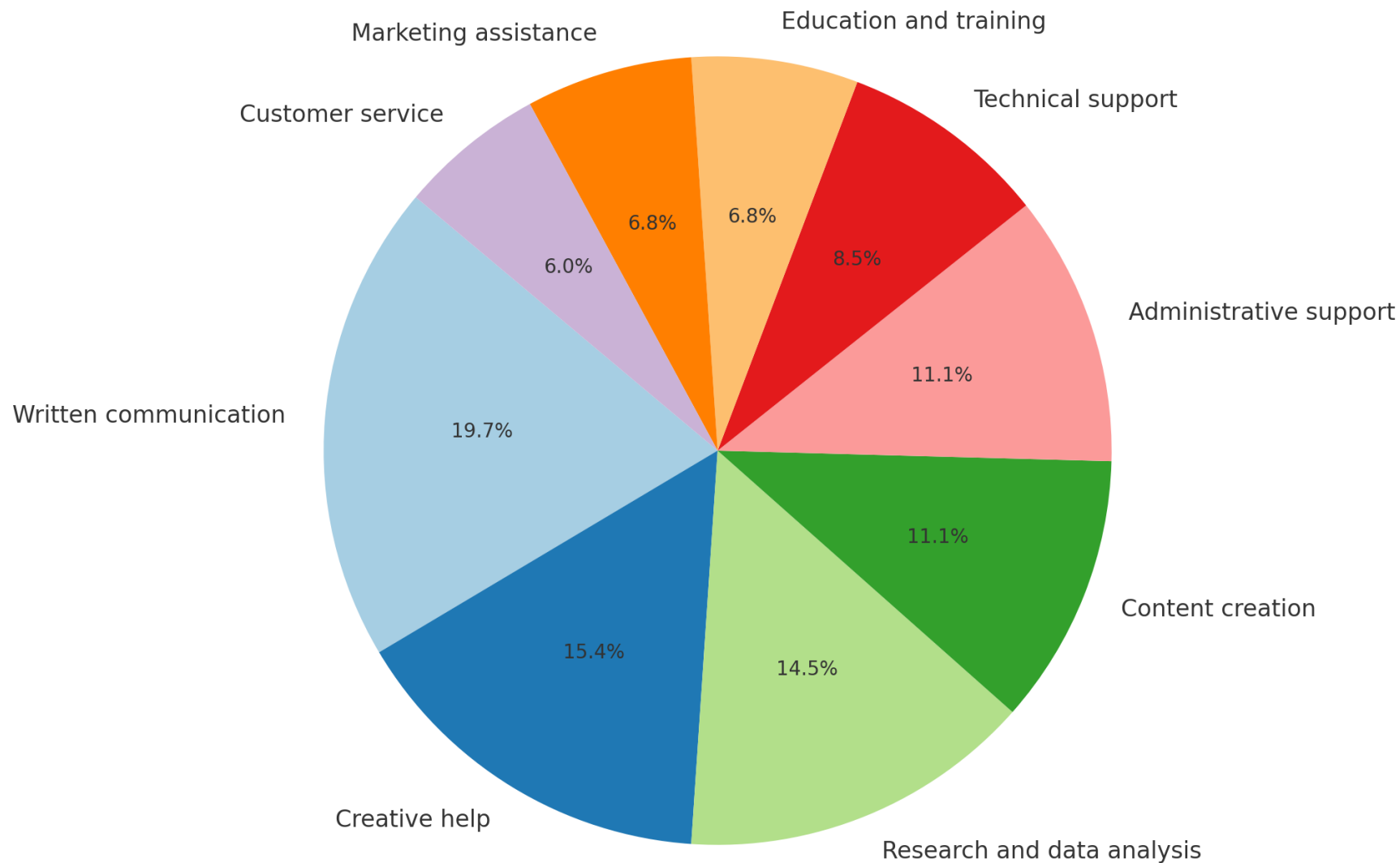
Abstract

We examined the productivity effects of a generative artificial intelligence (AI) technology, the assistive chatbot ChatGPT, in the context of midlevel professional writing tasks. In a preregistered online experiment, we assigned occupation-specific, incentivized writing tasks to 453 college-educated professionals and randomly exposed half of them to ChatGPT. Our results show that ChatGPT substantially raised productivity: The average time taken decreased by 40% and output quality rose by 18%. Inequality between workers decreased, and concern and excitement about AI temporarily rose. Workers exposed to ChatGPT during the experiment were 2 times as likely to report using it in their real job 2 weeks after the experiment and 1.6 times as likely 2 months after the experiment.

UTILIZZO DI CHATGPT AL LAVORO



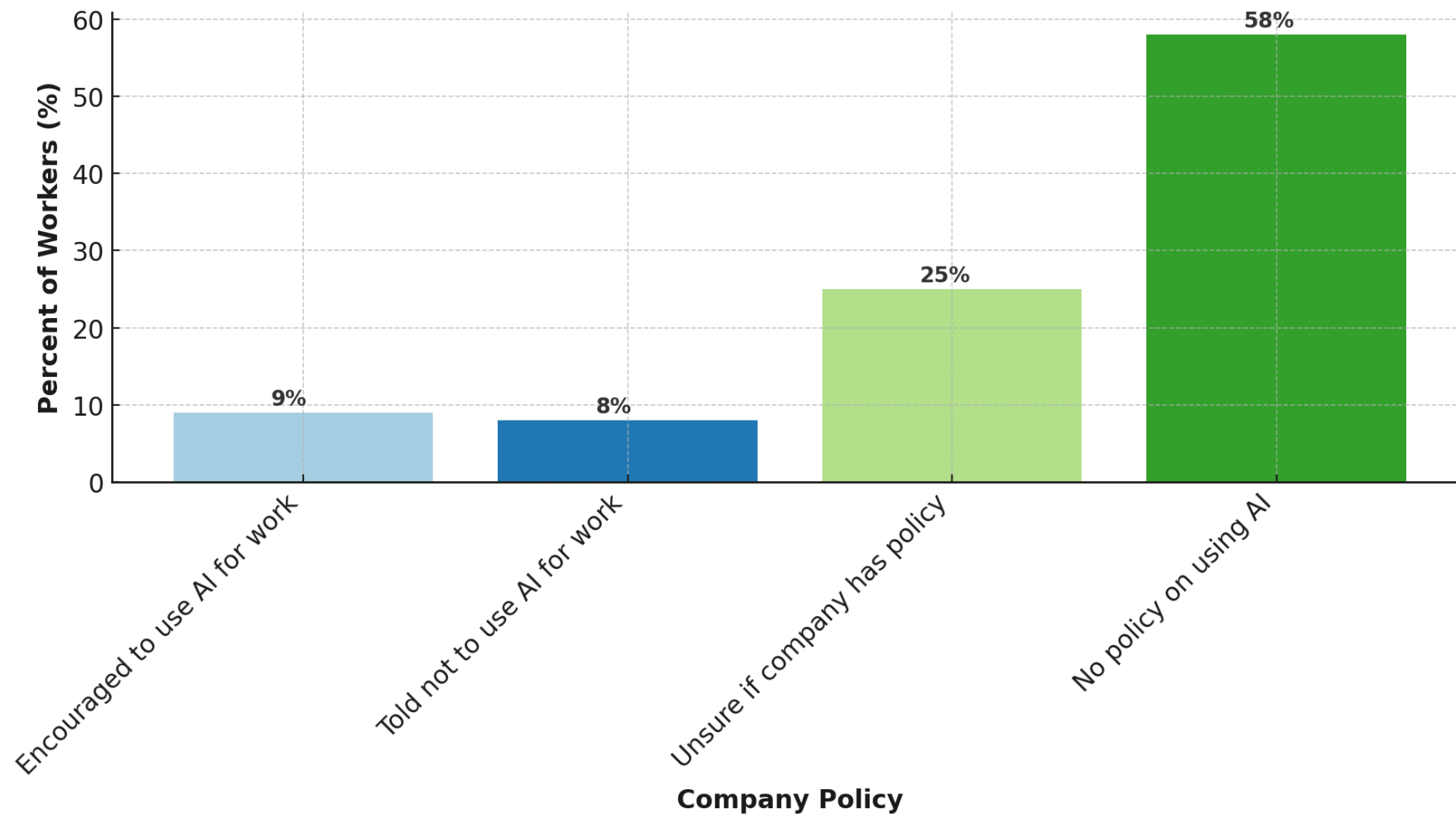
GAIA

GENERATIVE
AI APPLIANCE

LA POLICY NELLE AZIENDE OGGI



GAIA

GENERATIVE
AI APPLIANCE

Source: Business.com,

<https://www.business.com/technology/chatgpt-usage-workplace-study/>

WHEN PERFORMANCE MATTERS





[Home](#) [News](#) [Sport](#) [Business](#) [Innovation](#) [Culture](#) [Travel](#) [Earth](#) [Video](#) [Live](#)

The employees secretly using AI at work

25 October 2023

Share 

Some employers are either tacitly or outright banning access to generative AI tools like ChatGPT. But employees who love them are finding ways to discreetly backchannel.

Since OpenAI launched ChatGPT in November 2022, companies have scrambled to keep its workplace use under control.



L'11% dei dati che i dipendenti copiano su ChatGPT sono confidenziali!

Fonte: Cyberhaven

<https://www.cyberhaven.com/blog/4-2-of-workers-have-pasted-company-data-into-chatgpt>

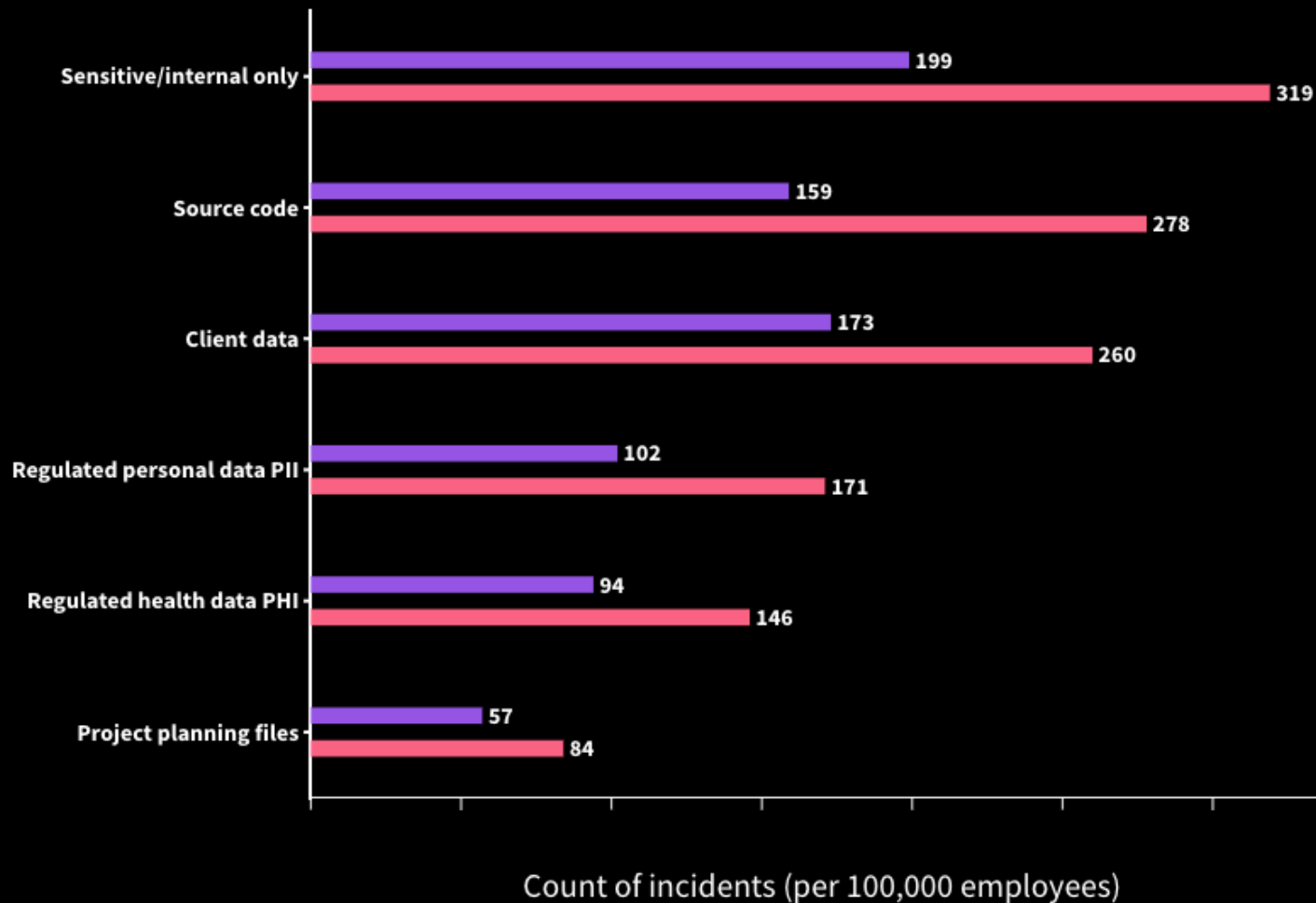
WHEN PERFORMANCE MATTERS

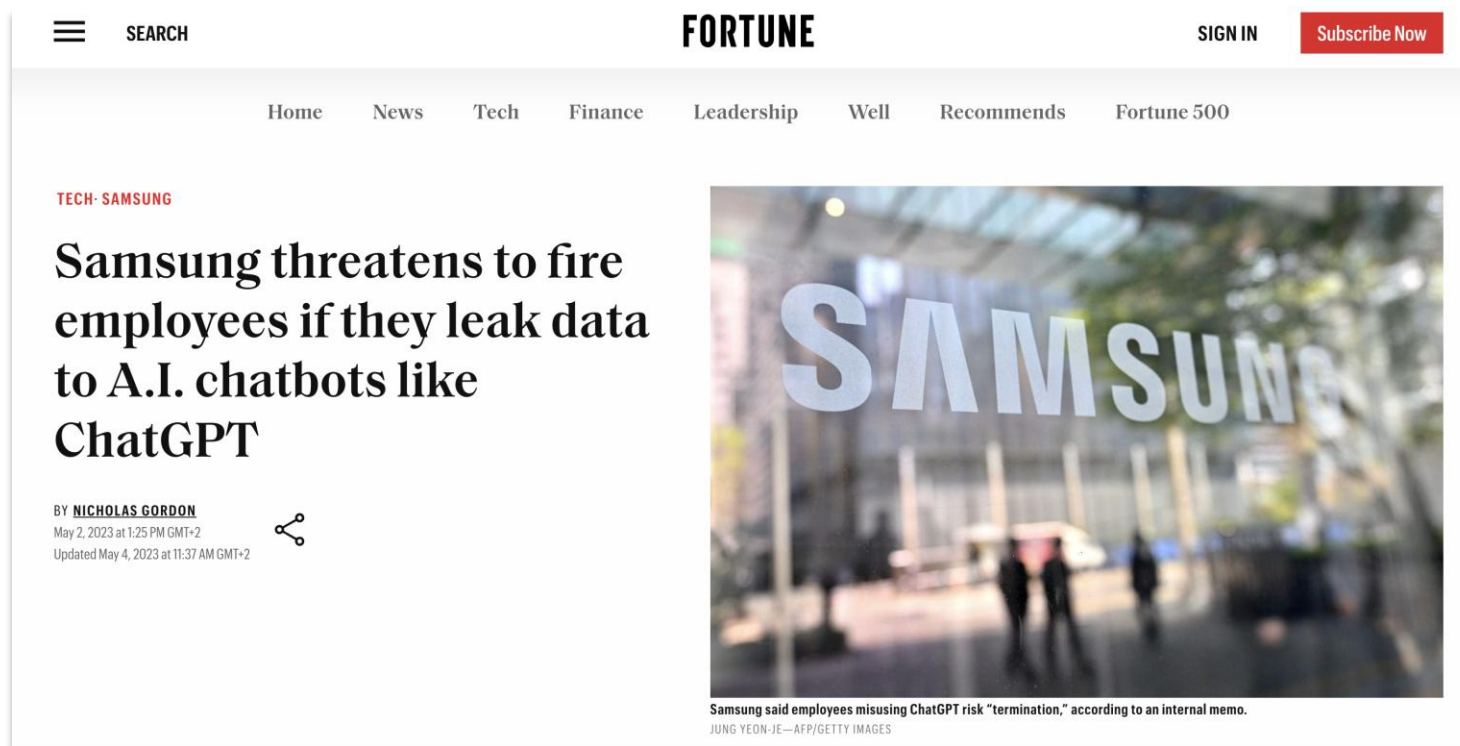
How much sensitive data goes to ChatGPT

(Incidents per 100,000 employees)

Source: Cyberhaven.com

Feb 26 - Mar 4 Apr 9 - Apr 15





- **Condivisione involontaria:** Il caricamento di documenti e dati sensibili su piattaforme cloud comporta sempre il rischio che tali informazioni possano essere consultate e divulgate da terzi non autorizzati.
- **Un esempio:** I PDF aziendali caricati per l'analisi o i riepiloghi contenenti informazioni finanziarie, strategiche o personali possono essere esposti a rischi se non gestiti in modo appropriato.

Sicurezza dei Dati

Vulnerabilità del cloud: Anche le migliori soluzioni cloud possono essere vulnerabili a violazioni della sicurezza, hacking o perdita di dati. Le aziende devono affidarsi alle misure di sicurezza implementate dai fornitori di servizi cloud, che potrebbero non essere sufficienti. Rischi legati al trasferimento dei dati: Durante i trasferimenti di dati tra l'azienda e il cloud, esiste il rischio di intercettazione o di attacchi man-in-the-middle.

Conformità alle Normative

GDPR e altre Norme: Molte aziende devono rispettare le severe normative sulla protezione dei dati, come il GDPR in Europa. L'utilizzo di soluzioni cloud può complicare la conformità a queste normative, poiché i dati possono essere archiviati in giurisdizioni diverse.

Controllo limitato dei Dati

Dipendenza dai Fornitori: Utilizzare i servizi cloud significa affidarsi completamente al fornitore per la gestione e la sicurezza dei dati. Questo comporta una perdita di controllo diretto e può essere problematico in caso di problemi tecnici o di governance.

Costi a lungo termine

Costi nascosti: Anche se inizialmente le soluzioni cloud possono sembrare convenienti, i costi possono rapidamente aumentare con l'uso continuo e la necessità di servizi aggiuntivi.

« È possibile avere
allo stesso tempo i
vantaggi della generative AI
e il
controllo totale dei propri dati? »

INIZIAMO CON UN MODELLO OPEN-SOURCE...

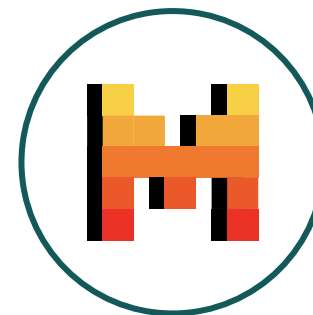


CHATGPT

VS



LLAMA3



MISTRAL

Modelli Proprietari:

Posseduti e gestiti da un'unica organizzazione. Questi modelli spesso offrono prestazioni elevate ma sono closed-source, richiedono un pagamento per l'accesso e limitano la personalizzazione.

Modelli Open-Source:

Sviluppati e mantenuti dalla comunità. Questi modelli sono liberamente disponibili per l'uso e la personalizzazione, garantendo flessibilità e trasparenza, ma possono richiedere maggiori competenze per essere implementati efficacemente in sede.

IL PERCORSO VERSO IL PRIVATE - LLM



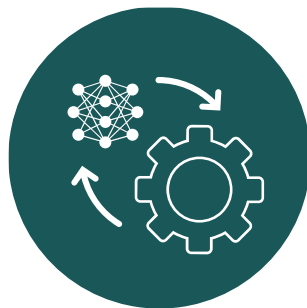
GAIA
GENERATIVE
AI APPLIANCE

... LO RENDIAMO DOMAIN-SPECIFIC ...



LLAMA3

+



RAG / FINE TUNING
ON-PREMISE

=



PRIVATE - LLM

- **RAG (Retrieval-Augmented Generation):** combina l'LLM con l'accesso in tempo reale ai database aziendali, migliorando l'accuratezza delle risposte con informazioni aggiornate e contestualizzate
- **Fine-Tuning:** adatta un modello linguistico pre-addestrato a dati specifici dell'azienda, ottimizzando le prestazioni e la pertinenza per esigenze particolari



... E IL TUTTO FUNZIONA ON-PREMISE.



LA TUA AI: Il modello diventa proprietà intellettuale dell'azienda, garantendone la completa proprietà e il controllo a tempo indeterminato.



I TUOI DATI: Utilizza la risorsa più preziosa dell'azienda, i suoi dati, per addestrare e personalizzare il modello, arricchendolo con le tue conoscenze aziendali specifiche.



LA TUA SICUREZZA: L'implementazione on-premise garantisce la totale sicurezza, protezione e privacy dei dati, nel rispetto degli standard più elevati.

E4

COMPUTER
ENGINEERING

GAIA 3.0

L'ALTERNATIVA
ON-PREMISE
A
CHATGPT



GAIA

GENERATIVE
ARTIFICIAL INTELLIGENCE
APPLIANCE

E4

COMPUTER
ENGINEERING

GAIA 3.0



GAIA

=



KAPTAIN

INFRASTRUTTURA
HIGH-
PERFORMANCE
KUBERNETES

+

PRIVATE - LLM

MICRO-SERVICES
ARCHITECTURE CHE OFFRE
UN'ALTERNATIVA ON-PREMISE
A CHATGPT



GAIA

GENERATIVE
ARTIFICIAL INTELLIGENCE
APPLIANCE

LE 3 CONFIGURAZIONI DI GAIA



GAIA

GENERATIVE
AI APPLIANCEGAIA
ALL-IN-ONE

MASTER/WORKER NODE

GAIA
CLUSTER

MASTER NODE



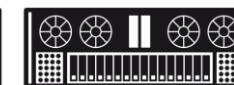
WORKER NODE



WORKER NODE



WORKER NODE



WORKER NODE

GAIA
HA CLUSTER

MASTER NODE



MASTER NODE



MASTER NODE



WORKER NODE



WORKER NODE



WORKER NODE



WORKER NODE



New Chat

Workspace

Search

Today

Nuova Missione GAIA

E4 Television Network Info

Nuova Gaia Spiegata

Yesterday

Ciao Response Summary

Previous 7 days

Agenda Conferenza Programma

Casual Greeting Example

Nuova Missione GAIA

Caratteristiche del Server

Caratteristiche del Server

Futuro della Generativa AI

Previous 30 days

GAIA Space Mission Update

Ottimizzazione del Workflow

E4 Success Stories Showcase

Legge di Bilancio 2023

Novità Legge Di Bilancio

Novità Legge di Bilancio

E Erionis

GAIA for E4

Ricordami cos'è la nuova GAIA.



GAIA for E4

GAIA 3.0 rappresenta l'evoluzione più recente della piattaforma di AI generativa di E4, progettata per adattarsi continuamente e integrare le tecnologie più avanzate, rispondendo alle esigenze dinamiche del mercato. GAIA è composta da KAPTAIN, un cluster Kubernetes ad alte prestazioni on-premise, e dalla componente PrivateLLM.

Ad oggi la versione di GAIA è la 3.0, ed offre funzionalità comparabili ai principali modelli di AI come ChatGPT, arricchite con capacità avanzate di AI basata su agenti. Con il progredire della piattaforma, le future iterazioni (GAIA 4.0 e successive) introdurranno funzionalità sempre più sofisticate, guidate dall'evoluzione tecnologica e dalle nuove tendenze di mercato.

Funzionalità di GAIA:

- Implementazione on-premise di LLM open-source con protezione avanzata dei dati.
- Capacità di RAG di base per l'accesso a documenti e conoscenze aziendali.
- Nuove capacità di agent AI per migliorare l'interazione e l'automazione.

Target

- Aziende che necessitano di una soluzione di AI generativa sicura e privata.
- Organizzazioni con elevate esigenze di protezione dei dati.

Benefici:

- Sicurezza e privacy dei dati garantite.
- Piattaforma evolutiva con aggiornamenti continui.
- Personalizzazione per specifiche esigenze aziendali.



+ Send a Message



LLMs can make mistakes. Verify important information.

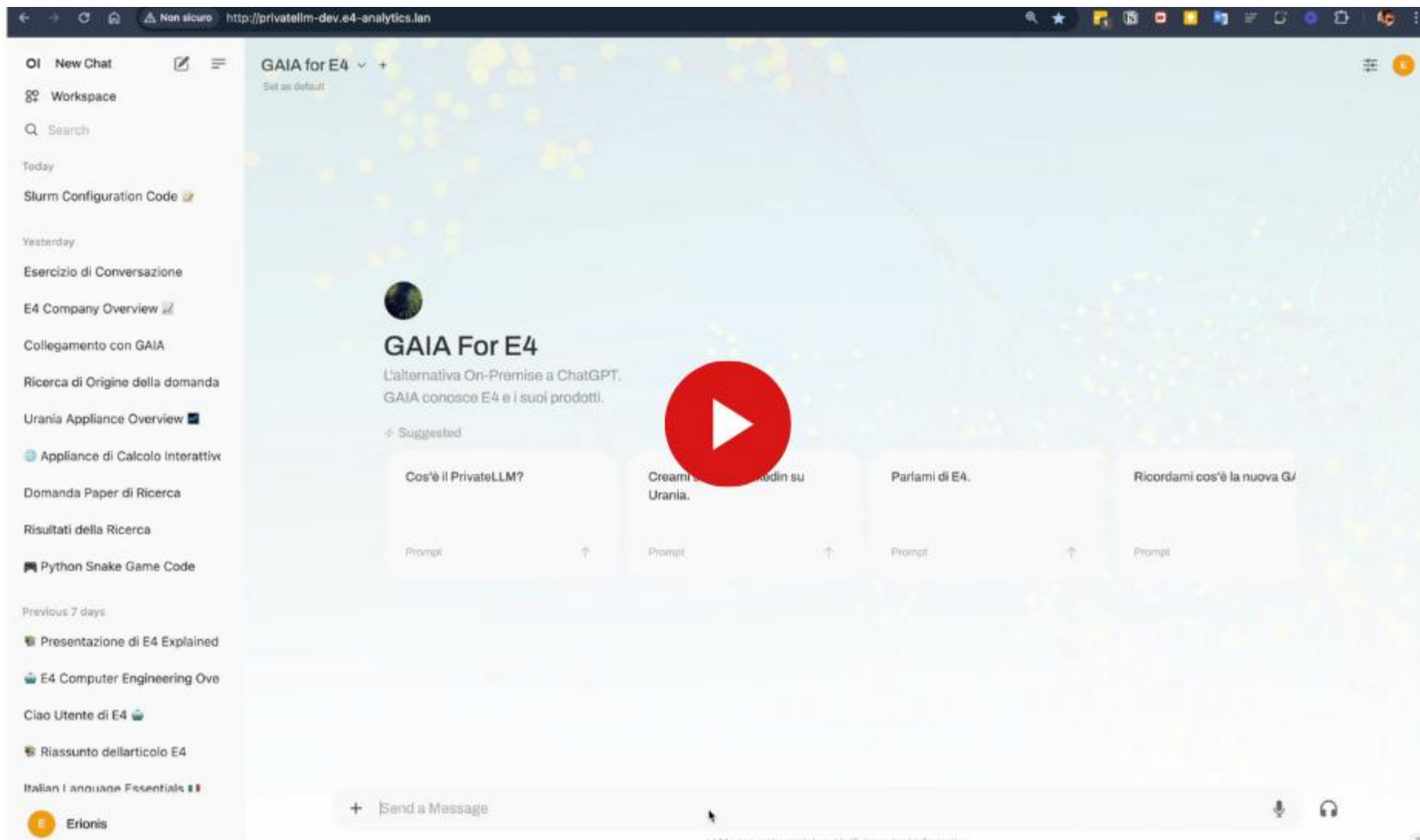
E4

COMPUTER
ENGINEERING



GAIA

GENERATIVE
AI APPLIANCE



VERSO GAIA 4.0

- GraphRAG
- MultiModal RAG
- Agentic Workflows

From Local to Global: A Graph RAG Approach to Query-Focused Summarization

ColPali: Efficient Document Retrieval with Vision Language Models

Manuel Faysse*^{1,3} Hugues Sibille*^{1,4} Tony Wu*¹ Bilel Omrani¹
Gautier Viaud¹ Céline Hudelot³ Pierre Colombo^{2,3}
¹Illuin Technology ²Equall.ai
³CentraleSunélec. Paris-Saclay ⁴ETH Zürich

2024-8-19

Automated Design of Agentic Systems

Shengran Hu^{1,2}, Cong Lu^{1,2} and Jeff Clune^{1,2,3}
¹University of British Columbia, ²Vector Institute, ³Canada CIFAR AI Chair

Researchers are investing substantial effort in developing powerful general-purpose agents, wherein Foundation Models are used as modules within *agentic systems* (e.g. Chain-of-Thought, Self-Reflection, Toolformer). However, the history of machine learning teaches us that hand-designed solutions are eventually replaced by learned solutions. We formulate a new research area, Automated Design of Agentic Systems (ADAS), which aims to *automatically create powerful agentic system designs, including inventing novel building blocks and/or combining them in new ways*. We further demonstrate that there is an unexplored yet promising approach within ADAS where agents can be defined in code and new agents can be automatically discovered by a meta agent programming ever better ones in code. Given that programming languages are Turing Complete, this approach theoretically enables the learning of *any possible agentic system*: including novel prompts, tool use, control flows, and combinations thereof. We present a simple yet effective algorithm named Meta Agent Search to demonstrate this idea, where a meta agent iteratively programs interesting new agents based on an ever-growing archive of previous discoveries. Through extensive experiments across multiple domains including coding, science, and math, we show that our algorithm can progressively invent agents with novel designs that greatly outperform state-of-the-art hand-designed agents. Importantly, we consistently observe the surprising result that agents invented by Meta Agent Search maintain superior performance even when transferred across domains and models, demonstrating their robustness and generality. Provided we develop it safely, our work illustrates the potential of an exciting new research direction toward automatically designing ever-more powerful agentic systems to benefit humanity.

<https://github.com/ShengranHu/ADAS>

arXiv:2404.16130v1 [cs.CL] 24 Apr 2024

:2407.01449v2 [cs.IR] 2 Jul 2024

8435v1 [cs.AI] 15 Aug 2024





GAIA

- **GAIA** è la tua soluzione pronta all'uso per i servizi di **AI generativa on-premise**
- **GAIA** offre ai suoi utenti **un'esperienza d'uso simile a quella di ChatGPT**
- **GAIA** è una appliance che utilizza esclusivamente **hardware di fascia alta**
- **GAIA** è una soluzione per l'AI Generativa orchestrata da **Kubernetes**
- **GAIA** è una soluzione aperta in cui è possibile integrare **container di terze parti**
- **GAIA** è **periodicamente aggiornata** per garantirne sempre lo stato dell'arte

GAIA: L'APPLIANCE PRONTA ALL'USO PER LA GENAI



E4

COMPUTER
ENGINEERING

GRAZIE
THANK YOU



GAIA

GENERATIVE
ARTIFICIAL INTELLIGENCE
APPLIANCE

The logo consists of the letters 'E4' in a bold, white, sans-serif font, positioned inside a dark grey diamond shape.

COMPUTER
ENGINEERING

CONTACTS

Email contacts

info@e4company.com

support@e4company.com

sales@e4company.com

E4 Computer Engineering SpA

Via Martiri della Libertà, 66 . 42019 Scandiano (RE) - Italy

Tel. +39 0522 991811





KAPTAIN

Kaptain è la soluzione Kubernetes disegnata per carichi di lavoro intensivi dal punto di vista computazionale, che include una web UI



RANCHER MANAGER – WEB UI



KUBERNETES TELEMETRY



STORAGE CLASSES FOR SHARED AND LOCAL BLOCK STORAGE



NVIDIA GPU OPERATOR



NVIDIA NETWORK OPERATOR



RANCHER KUBERNETES ENGINE 2

- **Rancher Manager** è la *web user interface* di Kaptain sia per l'amministratore che per l'utente finale
- **Prometheus Stack** è la soluzione per la telemetria per Kubernetes che supporta anche il sottosistema delle GPU
- **Data Persistence:** Kaptain è pre-configurato con 2 Storage Class, per garantire accesso sia a sistemi NAS che DAS
- **NVIDIA Operators** garantiscono ai container un accesso diretto alle GPU e al network RDMA presenti sul sistema
- **RKE2** è la distribuzione Kubernetes di nuova generazione certificata dalla CNCF

NB: il layout singolo nodo non include il Network Operator di Nvidia

Private-LLM è la micro-services architecture ad alte prestazioni che implementa i servizi di genAI per l'utente finale



PRIVATE-LLM USER INTERFACE



PRIVATE-LLM BACKEND MICRO-SERVICES



SCALABLE INFERENCE SERVICES FOR OPEN SOURCE LLMs



HIGH PERFORMANCE S3 OBJECT STORAGE



IDENTITY PROVIDER SERVICE

- **Identity Provider:** servizio di autenticazione e autorizzazione basato sullo standard openID connect
- **Object Storage:** servizio distribuito e resiliente basato su MinIO che offre un'interfaccia S3 per l'accesso ai dati
- **Inference Services:** sistemi scalabili configurati per offrire le massime performance nell'esposizione di LLM open-source
- **Private-LLM Backend:** Insieme di micro-servizi che implementa il back-end necessario alla UI
- **Private-LLM UI:** applicazione web che implementa il servizio di chatbot multi-utente, multi-chat e multi modello