




# **Una prospettiva aziendale su come integrare l'IA nei prodotti**

Lorenzo Pozzi   
Data Scientist  
PCO R&D

# UNA CLASSIFICAZIONE DELL'INTELLIGENZA ARTIFICIALE (IN BASE ALLE CAPACITA')



Il panorama dell'IA è ampissimo e ogni ramo rappresenta un micro-universo di modelli e applicazioni. Per fare un po' di chiarezza proviamo a creare una tassonomia in base alle CAPACITA' dei modelli IA.

Computer Vision, Computer Audition e Computer Linguistics comprendono competenze legate alla elaborazione di immagini, dati audio e linguaggio naturale.



La Robotica è legata alla guida e al controllo fisico di sistemi robotici, e.g. AI nel settore agroalimentare ([link](#))

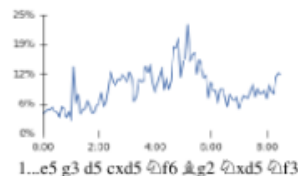
Forecasting si riferisce all'analisi dei dati delle serie temporali.

Discovery include tutte quelle analisi basata su cluster e ricerca di strutture nei dataset.

Planning include modelli ML che consentono di sviluppare strategie a lungo termine.

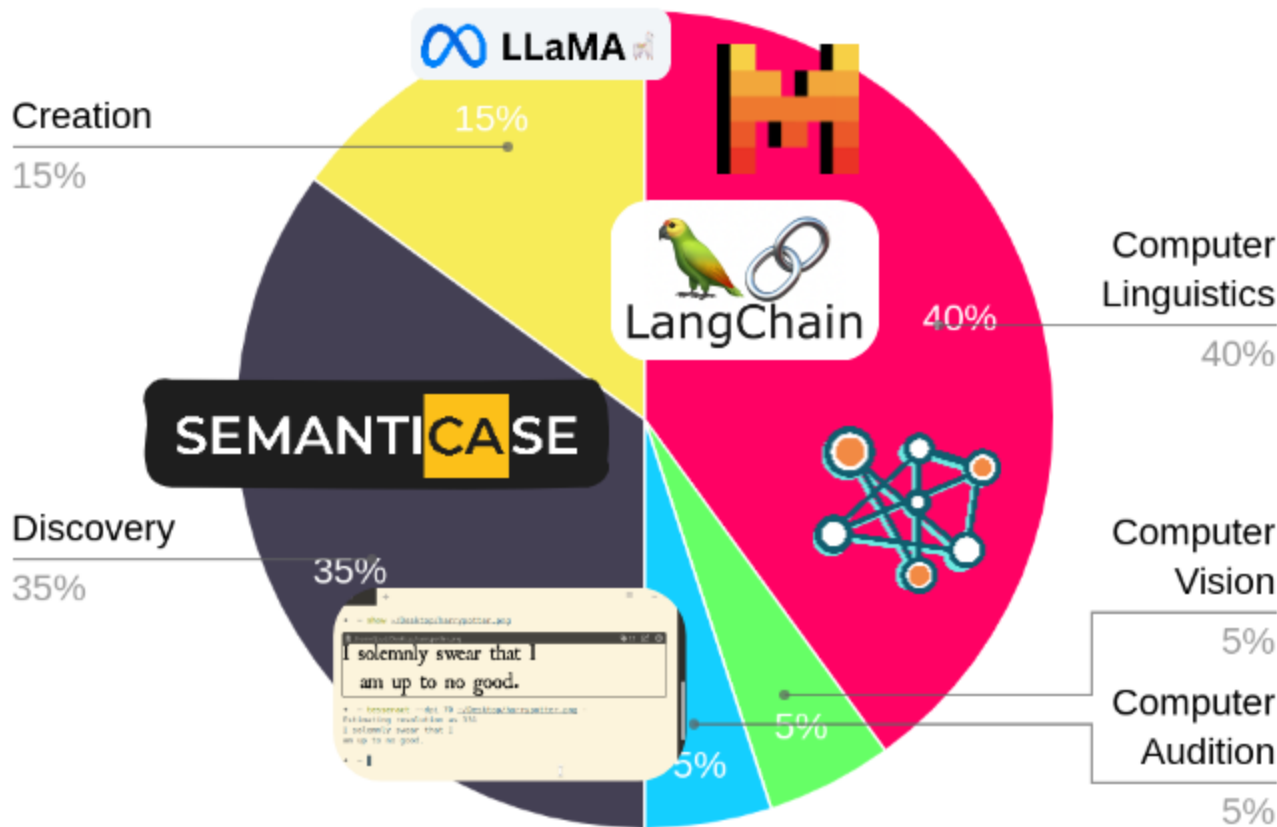


A10: English Opening



Creation, descrive la creazione di nuovi dati di varia natura:immagini, numerici, testuali, etc.





UN'ALTRA CLASSIFICAZIONE DELL'INTELLIGENZA ARTIFICIALE  
(IN BASE ALLA GENERALITA')



Position: Levels of AGI for Operationalizing Progress on the Path to AGI

Meredith Ringel Morris<sup>1</sup> Jascha Sohl-Dickstein<sup>2</sup> Noah Fiedel<sup>2</sup> Tris Warkentin<sup>2</sup> Allan Dafoe<sup>3</sup>  
Aleksandra Faust<sup>2</sup> Clement Farabet<sup>3</sup> Shane Legg<sup>3</sup>



LANGUAGE AGENTS ACHIEVE SUPERHUMAN SYNTHESIS OF  
SCIENTIFIC KNOWLEDGE

**FutureHouse Team**  
Michael D. Skarlinski<sup>1</sup> Sam Cox<sup>1,2</sup> Jon M. Laurent<sup>1</sup>  
James D. Braza<sup>1</sup> Michaela Hinks<sup>1</sup> Michael J. Hammerling<sup>1</sup>  
Manvitha Ponnampati<sup>1</sup> Samuel G. Rodrigues<sup>1,3\*</sup> Andrew D. White<sup>1,3\*</sup>  
<sup>1</sup>FutureHouse Inc., San Francisco, CA  
<sup>2</sup>University of Rochester, Rochester, NY  
<sup>3</sup>Francis Crick Institute, London, UK  
\*These authors jointly supervise technical work at FutureHouse.  
Correspondence to: {sam, andrew}@futurehouse.org

ABSTRACT

Language models are known to “hallucinate” incorrect information, and it is unclear if they are sufficiently accurate and reliable for use in scientific research. We developed a rigorous human-AI comparison methodology to evaluate language model agents on real-world literature search tasks covering information retrieval, summarization, and contradiction detection tasks. We show that PaperQA2, a frontier language model agent optimized for improved factuality, matches or exceeds subject matter expert performance on three realistic literature research tasks without any restrictions on humans (i.e., full access to internet, search tools, and time). PaperQA2 writes cited, Wikipedia-style summaries of scientific topics that are significantly more accurate than existing, human-written Wikipedia articles. We also introduce a hard benchmark for scientific literature research called LitQA2 that guided design of PaperQA2, leading to it exceeding human performance. Finally, we apply PaperQA2 to identify contradictions within the scientific literature, an important scientific task that is challenging for humans. PaperQA2 identifies  $2.34 \pm 1.99$  (mean  $\pm$  SD,  $N = 93$  papers) contradictions per paper in a random subset of biology papers, of which 70% are validated by human experts. These results demonstrate that language model agents are now capable of exceeding domain experts across meaningful tasks on scientific literature.

1 Introduction

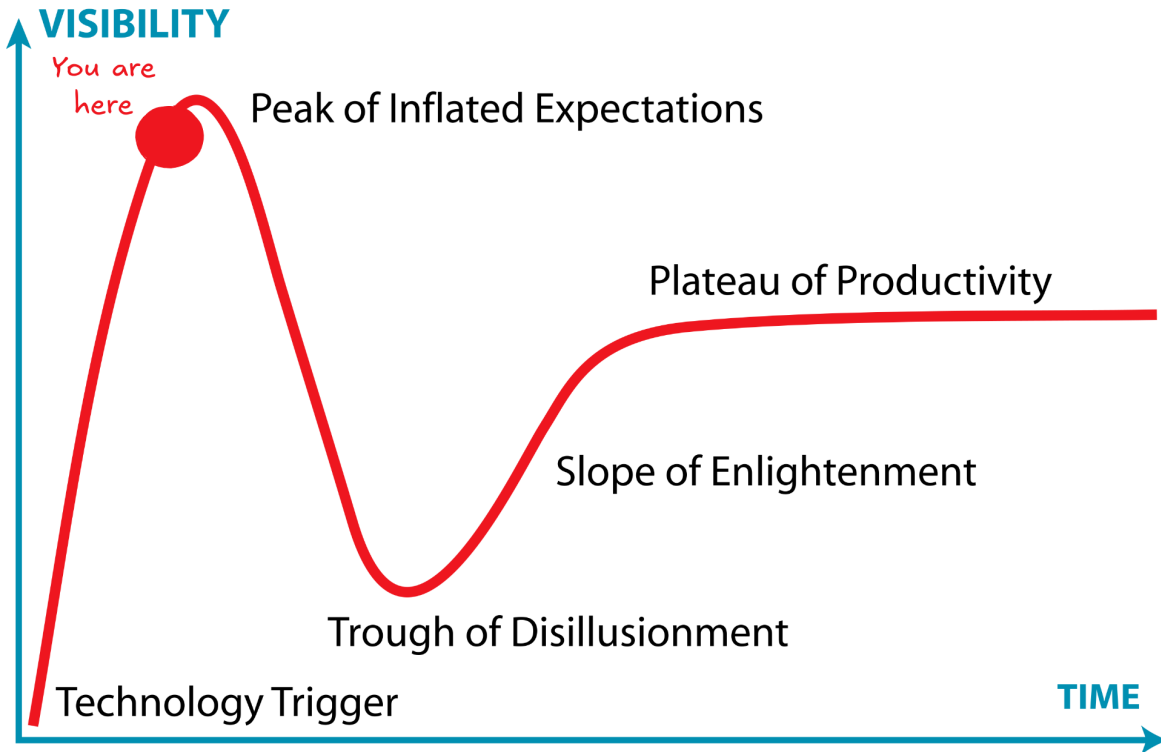
Large language models (LLMs) have the potential to assist scientists with retrieving, synthesizing, and summarizing the literature<sup>1,2</sup>, but still have several limitations for use in research tasks. Firstly, factuality is essential in scientific research, and LLMs hallucinate<sup>3</sup>, confidently stating information that is not grounded in any existing source or evidence. Secondly, science requires extreme attention to detail, and LLMs can overlook or misuse details when faced with challenging reasoning problems<sup>4</sup>. Finally, benchmarks for retrieval and reasoning across the scientific literature today are underdeveloped. They do not consider the entire literature, but instead are restricted to abstracts<sup>5</sup>, retrieval on a fixed corpus<sup>6</sup>, or simply provide the relevant paper directly<sup>7</sup>. These benchmarks are not suitable as performance proxies for real scientific research tasks, and more importantly, often lack a direct comparison to human performance. Thus, it remains unclear whether language models and agents are suitable for use in scientific research.

arXiv:2409.13740v1 [cs.CL] 10 Sep 2024

Performance (rows) x Generality (columns)	Narrow <i>clearly scoped task or set of tasks</i>	General <i>wide range of non-physical tasks, including metacognitive tasks like learning new skills</i>
Level 0: No AI	Narrow Non-AI calculator software; compiler	General Non-AI human-in-the-loop computing, e.g., Amazon Mechanical Turk
Level 1: Emerging <i>equal to or somewhat better than an unskilled human</i>	Emerging Narrow AI GOF AI (Boden, 2014); simple rule-based systems, e.g., SHRDLU (Winograd, 1971)	Emerging AGI ChatGPT (OpenAI, 2023), Bard (Anil et al., 2023), Llama 2 (Touvron et al., 2023), Gemini (Pichai & Hassabis, 2023)
Level 2: Competent <i>at least 50th percentile of skilled adults</i>	Competent Narrow AI toxicity detectors such as Jigsaw (Das et al., 2022); Smart Speakers such as Siri (Apple), Alexa (Amazon), or Google Assistant (Google); VQA systems such as PaLI (Chen et al., 2023); Watson (IBM); SOTA LLMs for a subset of tasks (e.g., short essay writing, simple coding)	Competent AGI not yet achieved
Level 3: Expert <i>at least 90th percentile of skilled adults</i>	Expert Narrow AI spelling & grammar checkers such as Grammarly (Grammarly, 2023); generative image models such as Imagen (Saharia et al., 2022) or Dall-E 2 (Ramesh et al., 2022)	Expert AGI not yet achieved
Level 4: Virtuoso <i>at least 99th percentile of skilled adults</i>	Virtuoso Narrow AI Deep Blue (Campbell et al., 2002), AlphaGo (Silver et al., 2016; 2017)	Virtuoso AGI not yet achieved
Level 5: Superhuman <i>outperforms 100% of humans</i>	Superhuman Narrow AI AlphaFold (Jumper et al., 2021; Varadi et al., 2021), AlphaZero (Silver et al., 2018), StockFish (Stockfish, 2023)	Artificial Superintelligence (ASI) not yet achieved



# GARTNER HYPE CYCLE



Ogni ciclo di hype approfondisce le cinque fasi chiave del ciclo di vita di una tecnologia.

1. Innesco tecnologico: Una svolta tecnologica dà il via alle cose, provocando l'interesse dei media. Spesso non esistono prodotti utilizzabili e la fattibilità commerciale non è dimostrata. (pubblicazione GPT3)

2. Picco di aspettative: la pubblicità iniziale produce una serie di storie di successo, ma spesso accompagnate da decine di fallimenti. Alcune aziende agiscono; la maggior parte no.

3. L'inizio della disillusione: l'interesse diminuisce man mano che gli esperimenti e le implementazioni non danno risultati. L'investimento continua solo se i fornitori sopravvissuti migliorano i propri prodotti in modo soddisfacente per i primi utilizzatori.

4. Illuminismo: Altri esempi dei vantaggi della tecnologia iniziano a cristallizzarsi e ad essere più ampiamente compresi. I prodotti di seconda e terza generazione provengono da fornitori di tecnologia consapevoli.

5. Altopiano della produttività: l'adozione mainstream inizia a decollare. I criteri per valutare la fattibilità del fornitore sono definiti più chiaramente. L'ampia applicabilità e rilevanza del mercato della tecnologia stanno chiaramente dando i loro frutti.

# PROBLEMI DEGLI LLM

1. sono dei modelli probabilistici quindi se la nostra applicazione è fuori distribuzione performeranno male + possono allucinare
2. richiedono molte risorse
3. non sono applicabili a tutti i tipi di dati

## The Devil is in the Tails: How Long-Tailed Code Distributions Impact Large Language Models

Xin Zhou<sup>1</sup>, Kisub Kim<sup>\*1</sup>, Bowen Xu<sup>1,2</sup>, Jiakun Liu<sup>1</sup>, DongGyun Han<sup>5</sup>, David Lo<sup>1</sup>

<sup>1</sup>Singapore Management University, Singapore

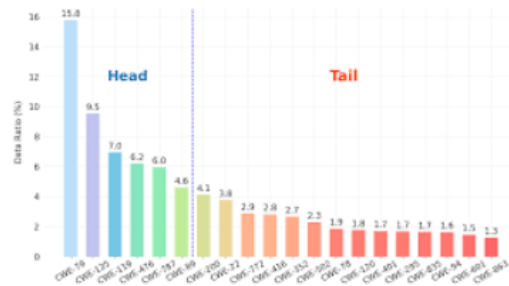
{xinzhou.2020, bowenxu.2017}@phdcs.smu.edu.sg, {kisubkim, jliu, davidlo}@smu.edu.sg

<sup>2</sup>North Carolina State University, USA

bxu22@ncsu.edu

<sup>5</sup>Royal Holloway, University of London, UK

donggyun.han@rhul.ac.uk



Our experimental results reveal that the long-tailed distribution has a substantial impact on the effectiveness of LLMs for code. Specifically, LLMs for code perform between 30.0% and 254.0% worse on data samples associated with infrequent labels compared to data samples of frequent labels

## Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools

Varun Magesh<sup>\*</sup>  
Stanford University

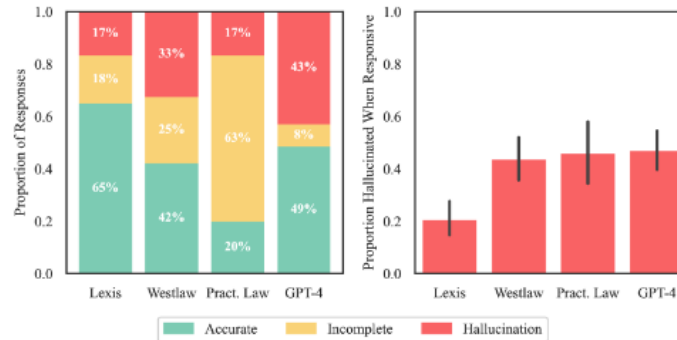
Faiz Surani<sup>\*</sup>  
Stanford University

Matthew Dahl  
Yale University

Mirac Suzgun  
Stanford University

Christopher D. Manning  
Stanford University

Daniel E. Ho<sup>†</sup>  
Stanford University



Nel Febbraio 2024, un esecutivo a Westlaw dichiara che il sistema RAG "dramatically reduces hallucinations to nearly zero". Anche a LexisNexis viene detto che la RAG "deliver accurate and authoritative answers that are grounded in the closed universe of authoritative content".

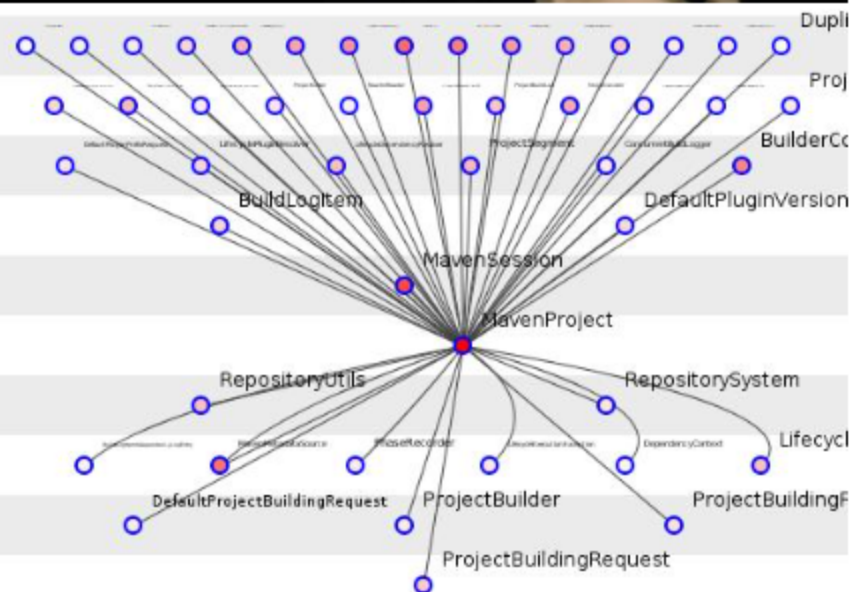
Dal paper possiamo invece vedere che Lexis+ AI ha fornito risposte accurate per il 65% delle domande, mentre Westlaw e Practical Law AI scendono a precisioni del 41% e 19%.

# The God Model



Il God Object è un termine usato nella progettazione software per descrivere un anti-pattern in cui un singolo oggetto di un programma diventa eccessivamente grande e complesso, assumendo troppe responsabilità.

Come in programmazione esiste il God Object, nel panorama dell'IA ci si affida sempre di più a modelli generali. Verso dei God Models... ma non è ancora il momento.



# COME INCLUDERE L'IA IN UN PRODOTTO/SOFTWARE?

Cominciamo smontando l'idea del God Model e pensiamo l'IA non come un prodotto ma come parte di esso.

Fatto questo, usiamo un concetto dall'ingegneria del software: il concetto di modularità.

In informatica la programmazione modulare è un paradigma di programmazione che consiste nella realizzazione di programmi suddivisi in moduli, ognuno dei quali svolge precise funzioni e il più possibili indipendenti.

Allo stesso modo i modelli IA devono essere dei moduli, in una pipeline più complessa.



# 1+4 Criteri per Usare l'IA

0. Dobbiamo usare un'IA?

1. fissare dei requisiti funzionali

2. fissare dei requisiti non-funzionali

3. pianificare in base alle risorse disponibili

4. riusabilità

# DOBBIAMO USARE UN'IA? IL PRINCIPIO DEL RASOIO DI OCCAM

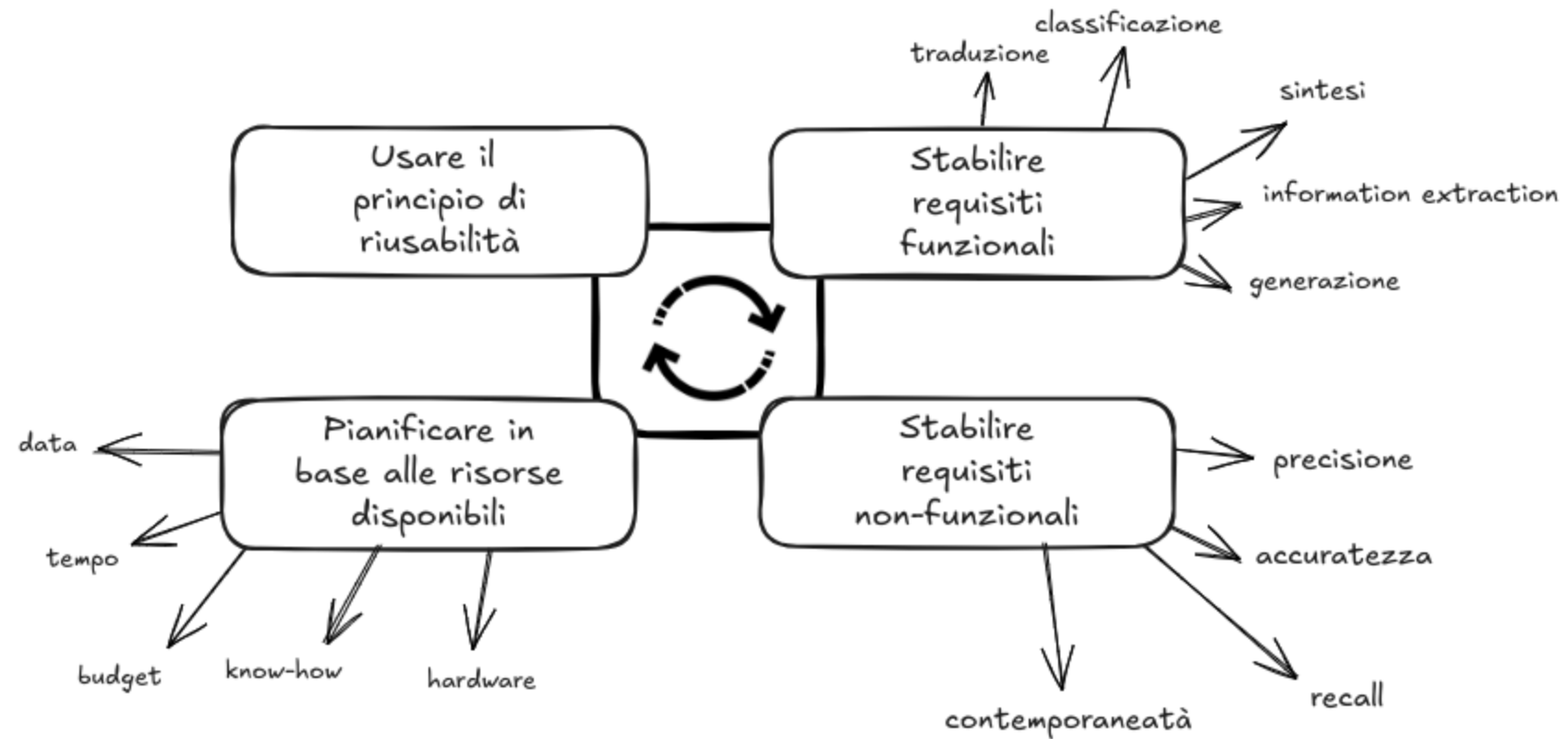


Il rasoio di Occam è un principio metodologico formulato nel XIV secolo dal frate francescano Guglielmo di Occam. Secondo questo principio è sempre meglio scegliere la soluzione più semplice tra più soluzioni egualmente valide di un problema.

In molte situazioni usare un modello complesso di IA, o l'IA di per se, non è la situazione più valida.



Open in Colab



# GRAZIE!



Pubblighiamo le nostre presentazioni e codice su un repository open e visibile a tutti.  
Siamo sempre alla ricerca di nuovi progetti e idee innovative. Veniteci a trovare!