Tool assisted research in the age of AI

Large Language Models and more

Renato Alves
Senior Bioinformatics Community Manager
Bio-IT project / Data Science - EMBL Heidelberg

MICROB-PREDICT

EMBL
50 YEARS | 1974–2024

Open in the browser or grab your mobile

Go to

www.menti.com

Enter the code

6413 1678

Or use QR code

> “Any sufficiently advanced technology is indistinguishable from magic.
>
> **– Arthur C. Clarke**”

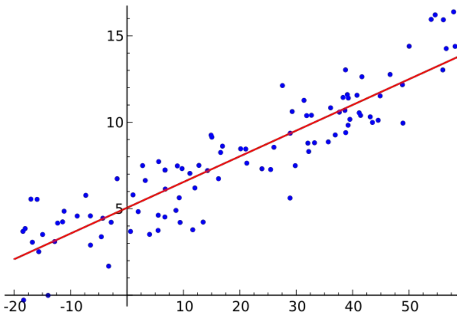
What is AI?

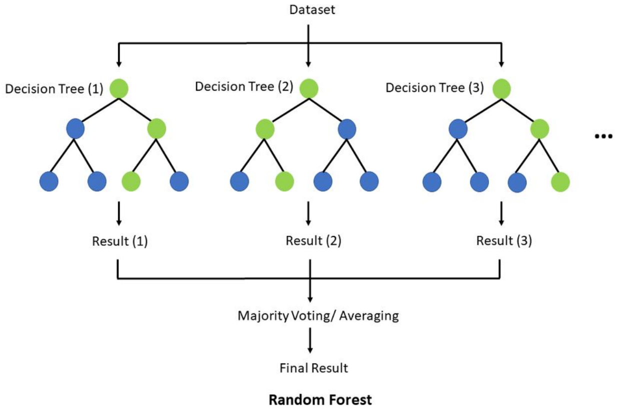
What is *intelligence*

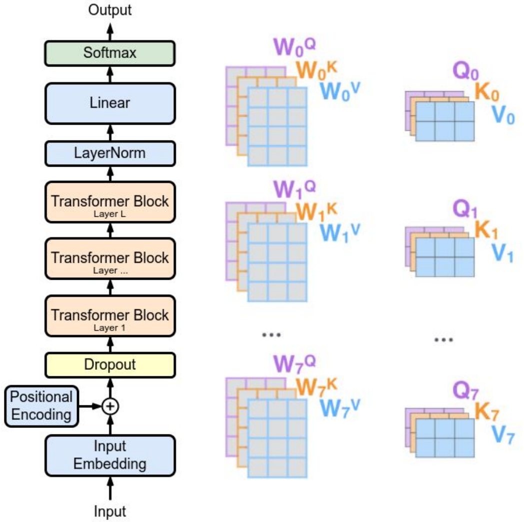
> **Intelligence**: the ability to learn, understand, and make judgments or have opinions that are based on reason
>
> – Cambridge Dictionary

Images from Wikipedia/Wikimedia and Modular (Illustrated transformer)

What is artificial *intelligence* (AI)







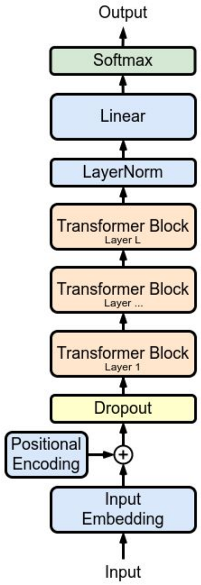
Stats
Linear regression

Machine Learning
Random forest

Deep Learning
Transformers (the T in GPT)

Images from Wikipedia/Wikimedia and Modular (Illustrated transformer)

Domains of AI …

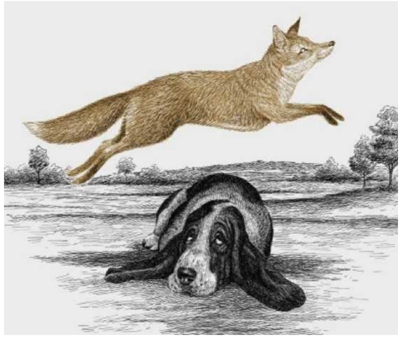


Natural Language Processing (NLP) (e.g. ChatGPT)

The quick brown fox jumps over the lazy dog

Sentiment Analysis
Machine Translation
Text Summarization
Text Clustering
Question Answering

Computer Vision (CV) (e.g. GPT4-Vision)

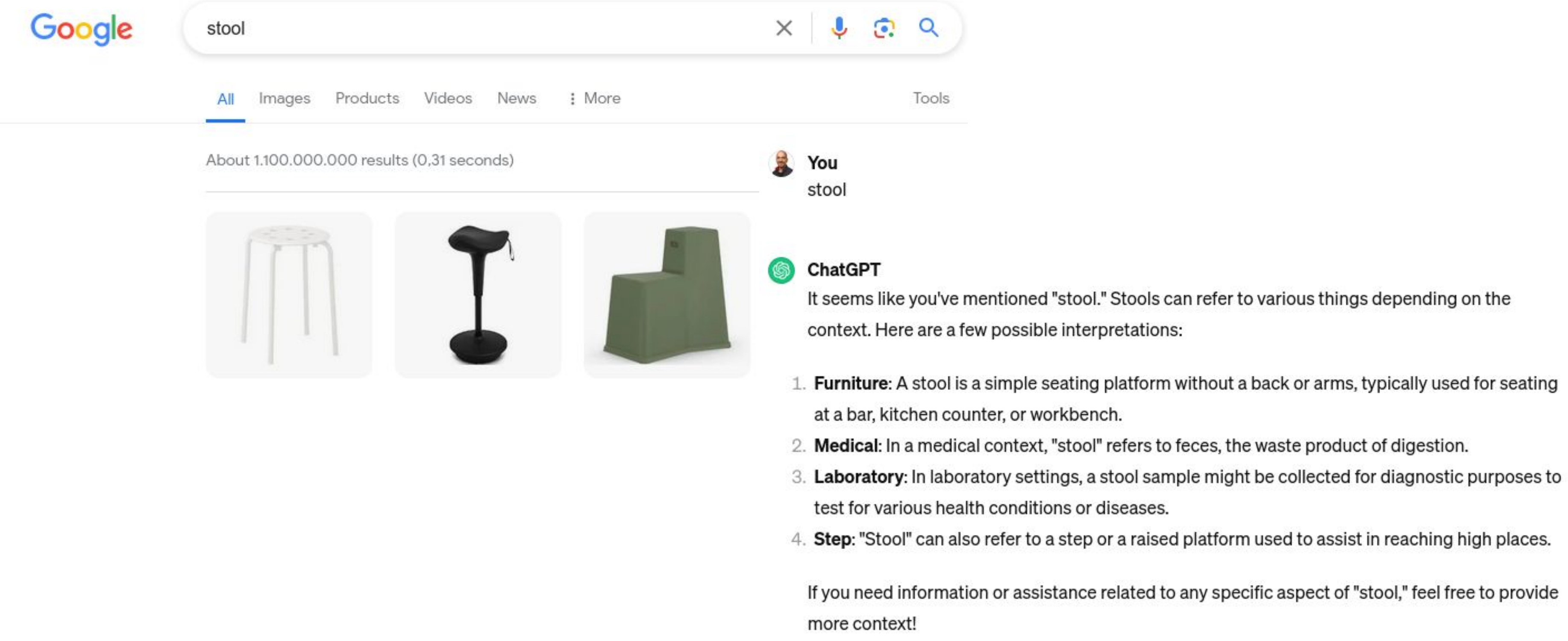


Object Detection
Object Recognition
Semantic Segmentation
Feature Detection and Description
Scene Understanding
Video Analysis

🤖: The phrase "The quick brown fox jumps over the lazy dog" is an English-language pangram, commonly used for touch-typing practice, testing typewriters and computer keyboards, displaying font examples, and other text-related applications where using all alphabet letters is necessary.

Images from Wikipedia/Wikimedia

To AI or not to AI



About 1.100.000.000 results (0,31 seconds)

You

stool

ChatGPT

It seems like you've mentioned "stool." Stools can refer to various things depending on the context. Here are a few possible interpretations:

1. **Furniture**: A stool is a simple seating platform without a back or arms, typically used for seating at a bar, kitchen counter, or workbench.
2. **Medical**: In a medical context, "stool" refers to feces, the waste product of digestion.
3. **Laboratory**: In laboratory settings, a stool sample might be collected for diagnostic purposes to test for various health conditions or diseases.
4. **Step**: "Stool" can also refer to a step or a raised platform used to assist in reaching high places.

If you need information or assistance related to any specific aspect of "stool," feel free to provide more context!

Recent AI evolution … very very fast

GPT-1
117 Million Parameters
2018

GPT-2
1.5 Billion Parameters
2019

GPT-3
175 Billion Parameters
2020

GPT-3.5
175 Billion Parameters
2022

GPT-4
1.76 Trillion Parameters
2023

GPT-4.5?
2024

ChatGPT talk?
GPT-4 Vision
DALL-E 3
Sep 2023

Stable Video Diffusion
Nov 2023

Google replaces Bard with Gemini
Dec 2023

Copyright lawsuits
Midjourney V6
Jan 2024

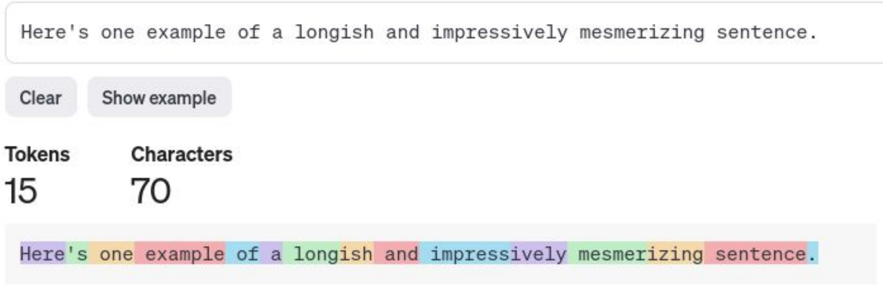
Sora
Feb 2024

Figure 01
Devin
Mar 2024

AI in practice

Some key jargon

Tokens → cost

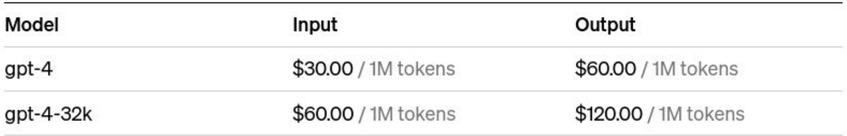


Here's one example of a longish and impressively mesmerizing sentence.

Tokens 15

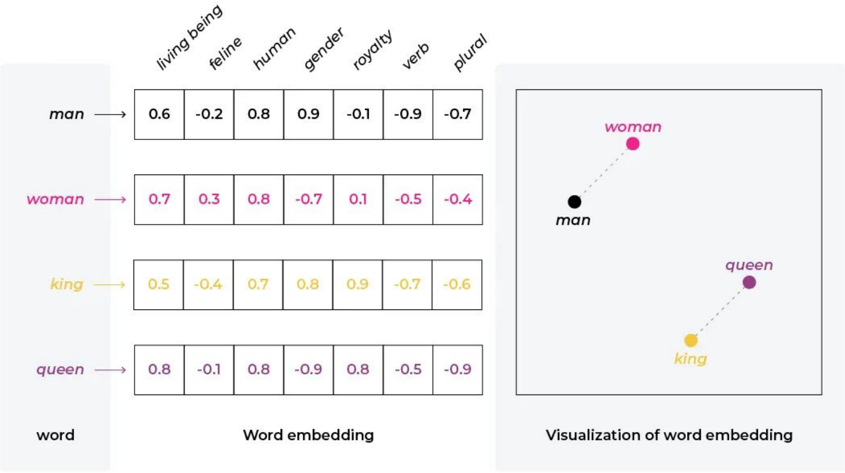
Characters 70

Here's one example of a longish and impressively mesmerizing sentence.



Model	Input	Output
gpt-4	$30.00 / 1M tokens	$60.00 / 1M tokens
gpt-4-32k	$60.00 / 1M tokens	$120.00 / 1M tokens

Embeddings → semantic/meaning



word	living being	feline	human	gender	royalty	verb	plural
man	0.6	-0.2	0.8	0.9	-0.1	-0.9	-0.7
woman	0.7	0.3	0.8	-0.7	0.1	-0.5	-0.4
king	0.5	-0.4	0.7	0.8	0.9	-0.7	-0.6
queen	0.8	-0.1	0.8	-0.9	0.8	-0.5	-0.9

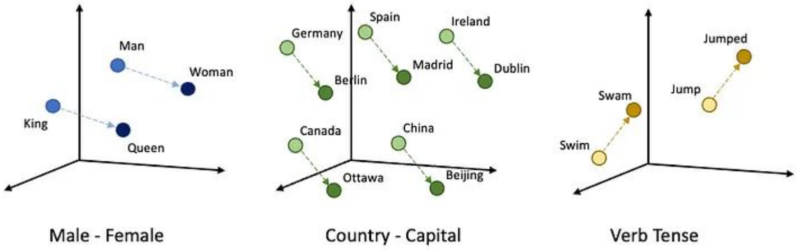
Tokenizer by OpenAI - Embeddings from Arize

Polite language generates better results

Hey Google, play us some music please?
Why are you being so polite?
Just in case.
KEEP THAT ONE ALIVE. HE ALWAYS SAID "PLEASE".

seebangnow

AI remembers!!

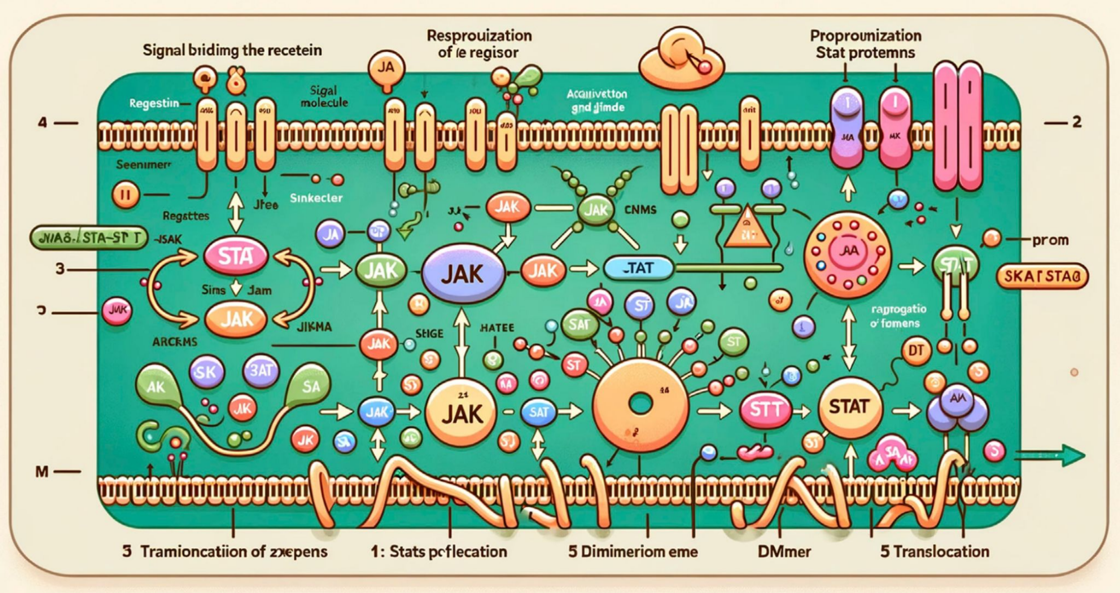


The role of politeness in human–machine interactions: a systematic literature review and future perspectives

doi: 10.1007/s10462-023-10540-1

Embeddings illustration from Towards Data Science@Medium

“With great power comes great responsibility”



“As of my last knowledge update …”

Proper use is hard to impossible to distinguish from no use.

You are accountable for using these tools responsibly.

Beware data privacy!

Xinyu Guo et al - doi:10.3389/fcell.2023.1339390

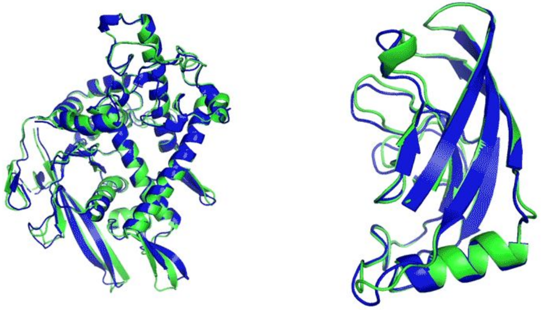
Earlier AI advances

AlphaGO

- Simple rules
- Complex play
- Harder than chess
- Adversarial

deepmind.google/technologies/alphago

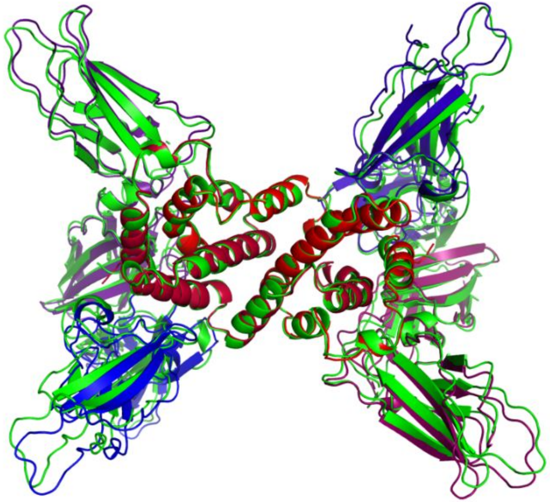
Alphafold / Alphafold multimer / Alphafold DB



T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

● Experimental result
● Computational prediction



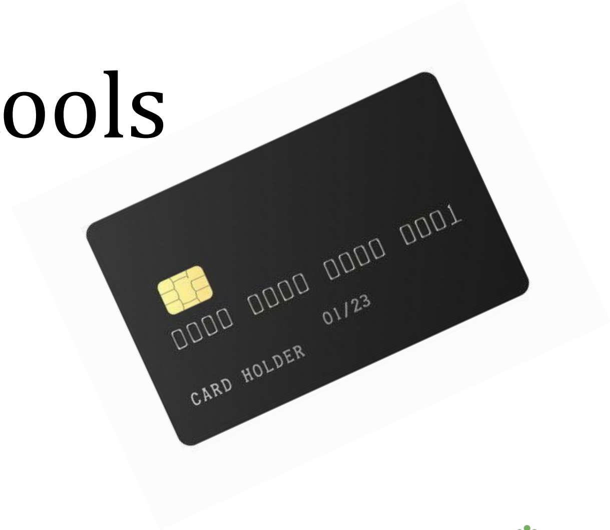
(a) A2B2C2 heteromer
TM-score = 97.4, N_{res} = 1,246, PDB ID = 6E3K

github.com/google-deepmind/alphafold / doi:10.1101/2021.10.04.463034 / alphafold.ebi.ac.uk

Tool assisted research

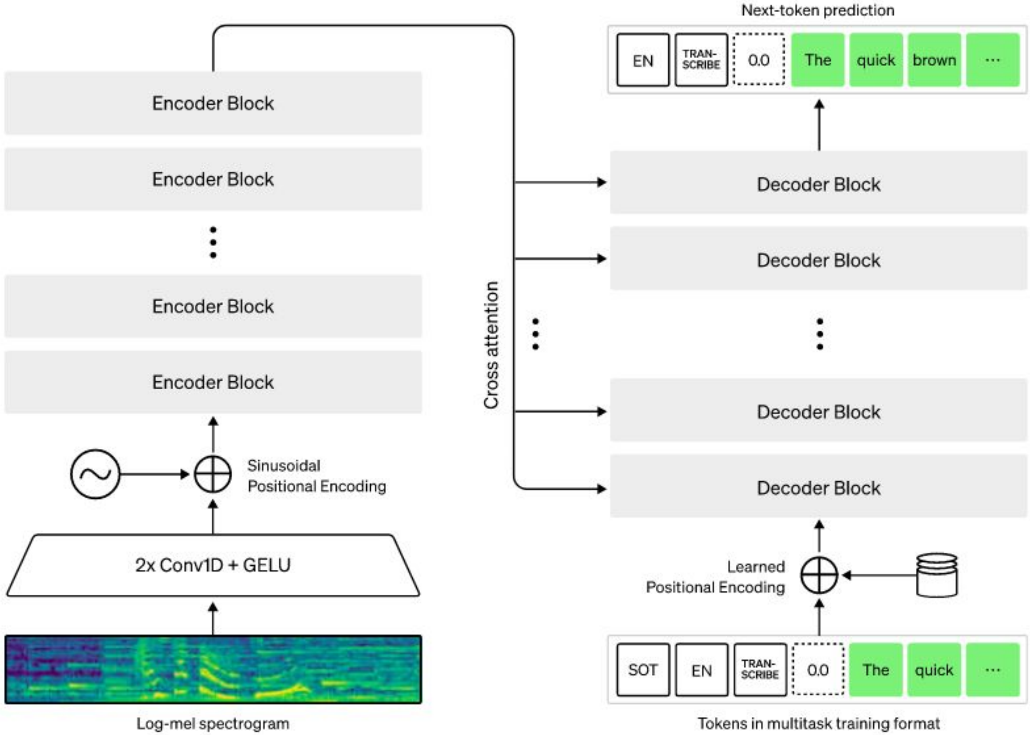
~~Tool assisted research~~

General purpose AI tools



Audio assistants - Dictation

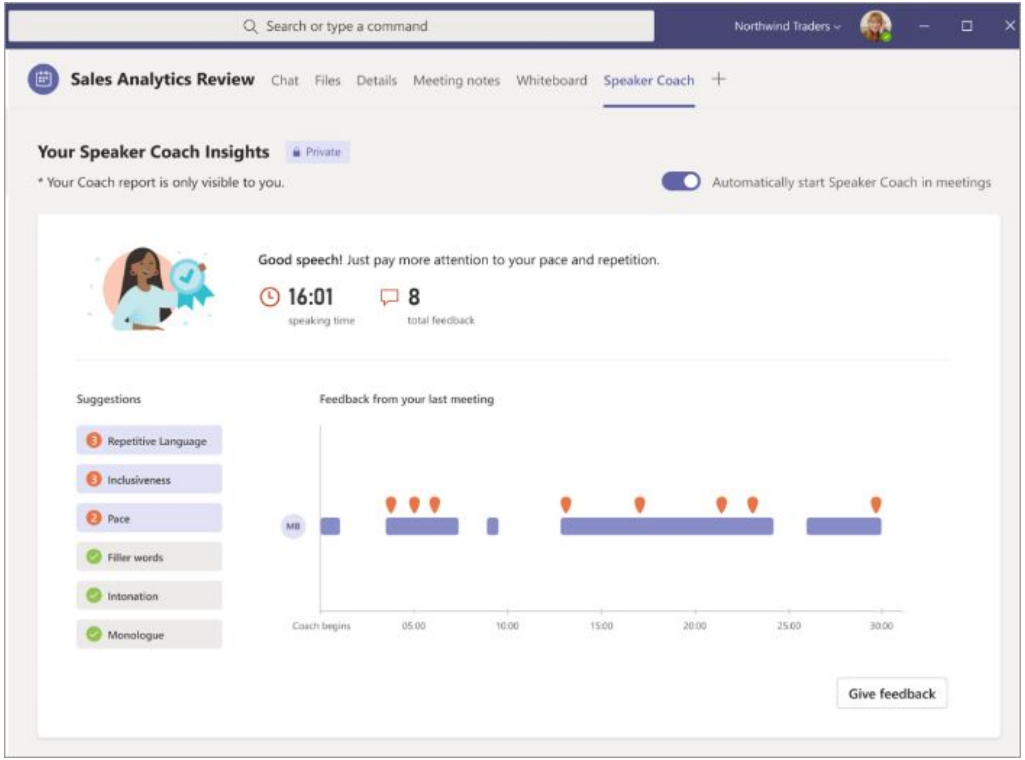
- Faster than typing
- Video transcription
- (Translation)
- Not ideal for code



OpenAI Whisper

Audio assistants - Presentation rehearsal

- Iterative
- Always available
- Remove shame factor
- Decent advice
- Self awareness



Your Speaker Coach Insights

* Your Coach report is only visible to you.

Automatically start Speaker Coach in meetings

Good speech! Just pay more attention to your pace and repetition.

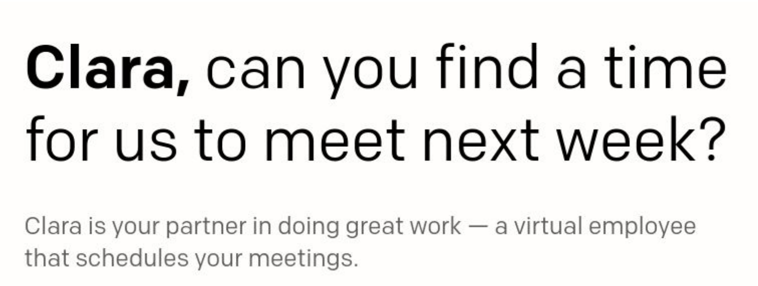
16:01 speaking time

8 total feedback

Microsoft Speaker Coach

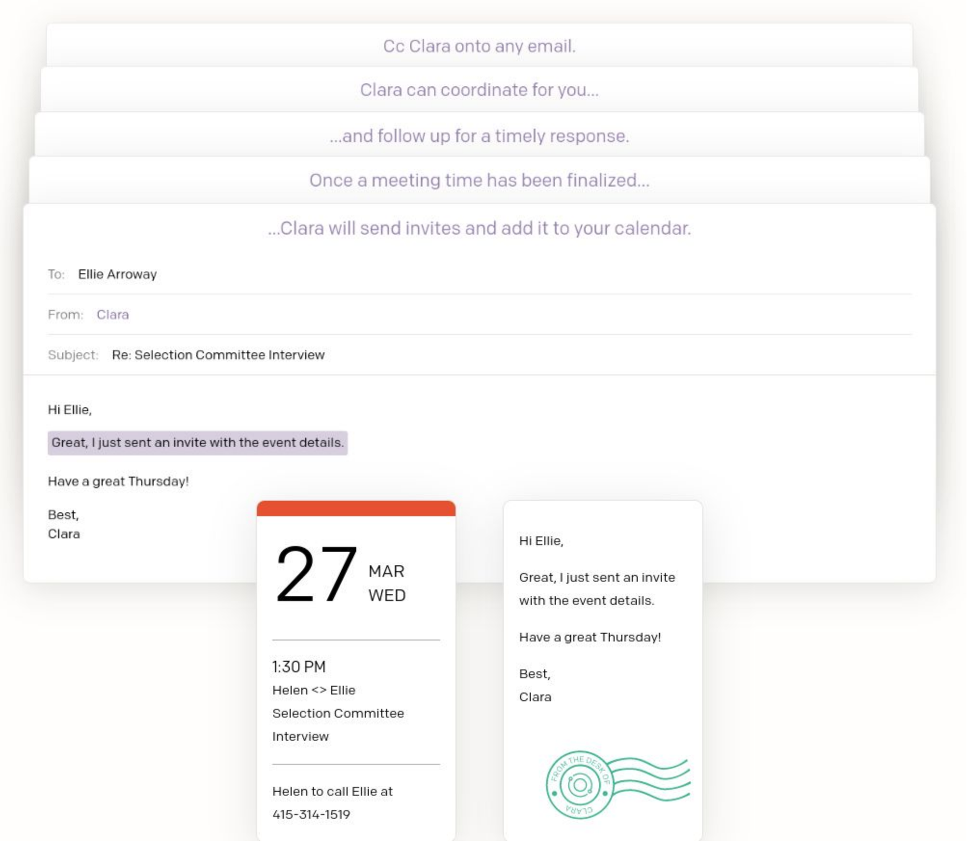
Planning assistants

- Death to Doodles
    - RIP
    - Seriously
    - Please



Clara, can you find a time for us to meet next week?

Clara is your partner in doing great work — a virtual employee that schedules your meetings.



Cc Clara onto any email.

Clara can coordinate for you...

...and follow up for a timely response.

Once a meeting time has been finalized...

...Clara will send invites and add it to your calendar.

To: Ellie Arroway

From: Clara

Subject: Re: Selection Committee Interview

Hi Ellie,

Great, I just sent an invite with the event details.

Have a great Thursday!

Best,
Clara

27 MAR WED

1:30 PM
Helen <> Ellie
Selection Committee Interview

Helen to call Ellie at 415-314-1519

Hi Ellie,

Great, I just sent an invite with the event details.

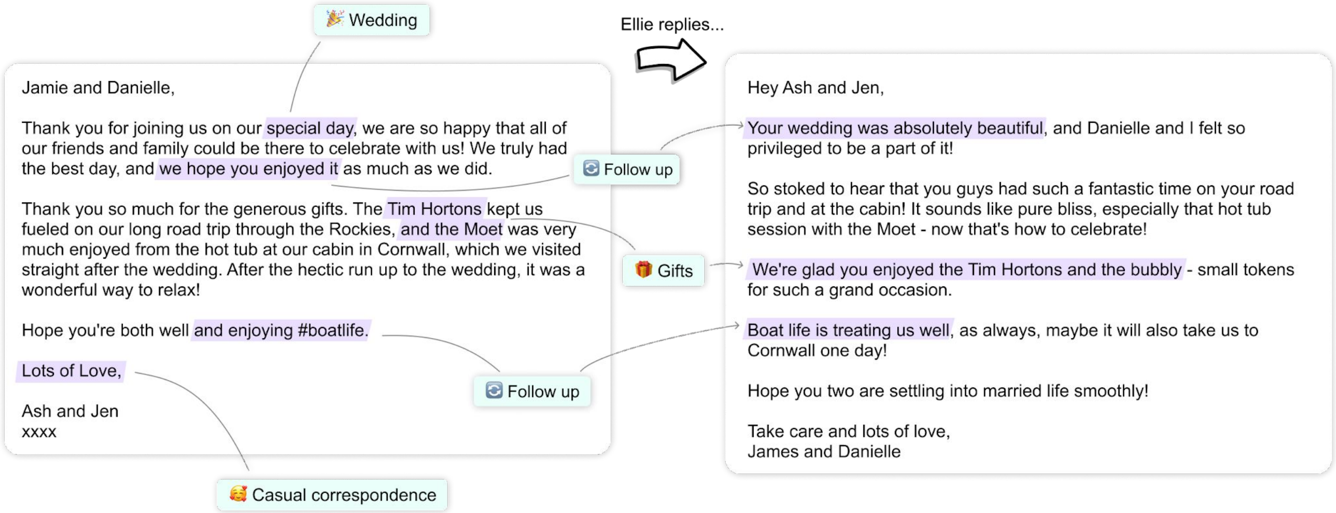
Have a great Thursday!

Best,
Clara

claralabs.com

Email assistants

- Draft reply
- Tidy-up and clarify
- No more typos



Jamie and Danielle,

Thank you for joining us on our special day, we are so happy that all of our friends and family could be there to celebrate with us! We truly had the best day, and we hope you enjoyed it as much as we did.

Thank you so much for the generous gifts. The Tim Hortons kept us fueled on our long road trip through the Rockies, and the Moet was very much enjoyed from the hot tub at our cabin in Cornwall, which we visited straight after the wedding. After the hectic run up to the wedding, it was a wonderful way to relax!

Hope you're both well and enjoying #boatlife.

Lots of Love,

Ash and Jen
xxxx

Hey Ash and Jen,

Your wedding was absolutely beautiful, and Danielle and I felt so privileged to be a part of it!

So stoked to hear that you guys had such a fantastic time on your road trip and at the cabin! It sounds like pure bliss, especially that hot tub session with the Moet - now that's how to celebrate!

We're glad you enjoyed the Tim Hortons and the bubbly - small tokens for such a grand occasion.

Boat life is treating us well, as always, maybe it will also take us to Cornwall one day!

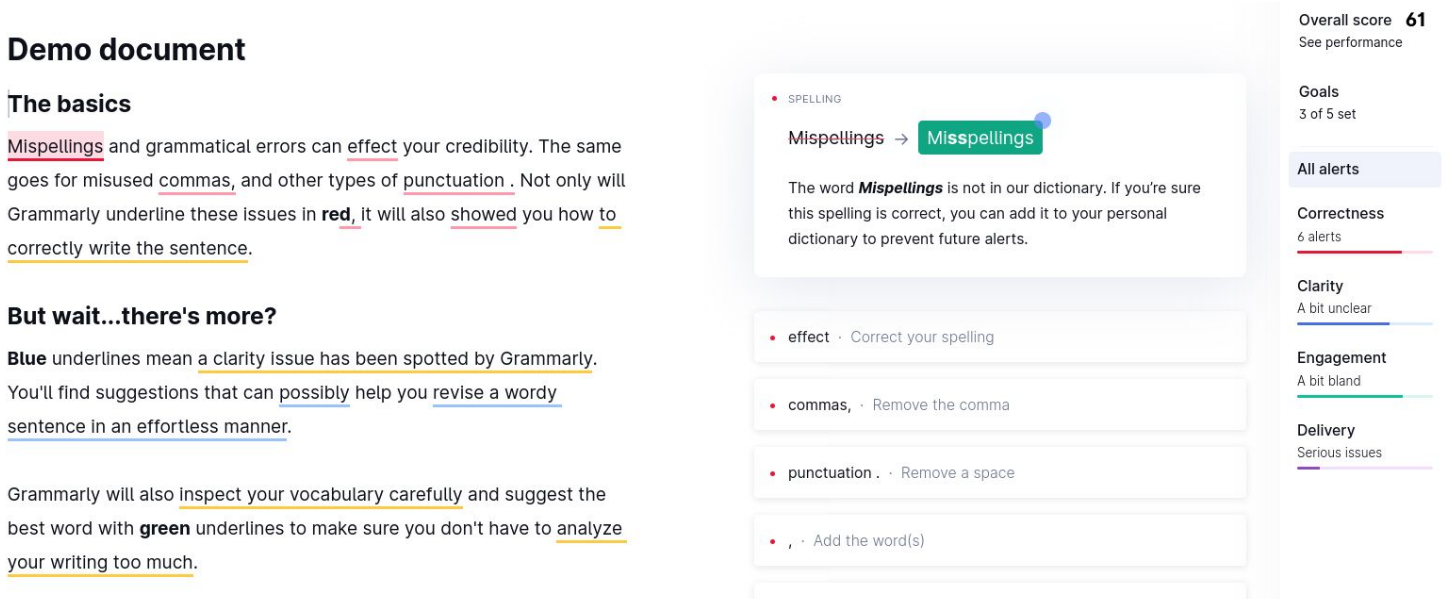
Hope you two are settling into married life smoothly!

Take care and lots of love,
James and Danielle

tryellie.com

Language assistants

- Write like a native
- Understand shortcomings



Demo document

The basics

Mispellings and grammatical errors can effect your credibility. The same goes for misused commas, and other types of punctuation . Not only will Grammarly underline these issues in **red**, it will also showed you how to correctly write the sentence.

But wait...there's more?

Blue underlines mean a clarity issue has been spotted by Grammarly. You'll find suggestions that can possibly help you revise a wordy sentence in an effortless manner.

Grammarly will also inspect your vocabulary carefully and suggest the best word with **green** underlines to make sure you don't have to analyze your writing too much.

SPELLING

~~Mispellings~~ → Misspellings

The word ***Mispellings*** is not in our dictionary. If you're sure this spelling is correct, you can add it to your personal dictionary to prevent future alerts.

- effect · Correct your spelling
- commas, · Remove the comma
- punctuation . · Remove a space
- , · Add the word(s)

Overall score **61**
See performance

Goals
3 of 5 set

All alerts

Correctness
6 alerts

Clarity
A bit unclear

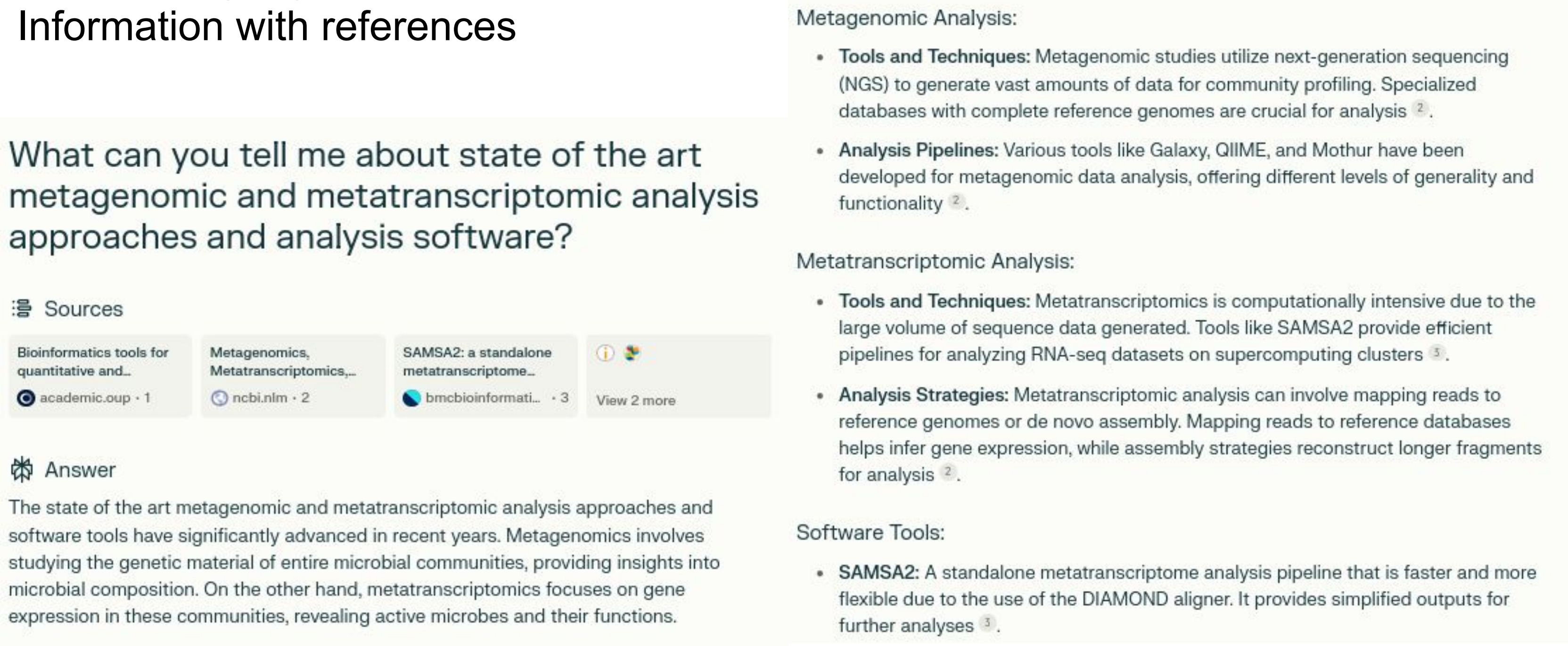
Engagement
A bit bland

Delivery
Serious issues

Grammarly

Improved search engines

- Natural language
- Information with references



What can you tell me about state of the art metagenomic and metatranscriptomic analysis approaches and analysis software?

Sources

Bioinformatics tools for quantitative and… academic.oup · 1

Metagenomics, Metatranscriptomics,… ncbi.nlm · 2

SAMSA2: a standalone metatranscriptome… bmcbioinformati… · 3

View 2 more

Answer

The state of the art metagenomic and metatranscriptomic analysis approaches and software tools have significantly advanced in recent years. Metagenomics involves studying the genetic material of entire microbial communities, providing insights into microbial composition. On the other hand, metatranscriptomics focuses on gene expression in these communities, revealing active microbes and their functions.

Metagenomic Analysis:

- **Tools and Techniques:** Metagenomic studies utilize next-generation sequencing (NGS) to generate vast amounts of data for community profiling. Specialized databases with complete reference genomes are crucial for analysis [2].
- **Analysis Pipelines:** Various tools like Galaxy, QIIME, and Mothur have been developed for metagenomic data analysis, offering different levels of generality and functionality [2].

Metatranscriptomic Analysis:

- **Tools and Techniques:** Metatranscriptomics is computationally intensive due to the large volume of sequence data generated. Tools like SAMSA2 provide efficient pipelines for analyzing RNA-seq datasets on supercomputing clusters [3].
- **Analysis Strategies:** Metatranscriptomic analysis can involve mapping reads to reference genomes or de novo assembly. Mapping reads to reference databases helps infer gene expression, while assembly strategies reconstruct longer fragments for analysis [2].

Software Tools:

- **SAMSA2:** A standalone metatranscriptome analysis pipeline that is faster and more flexible due to the use of the DIAMOND aligner. It provides simplified outputs for further analyses [3].

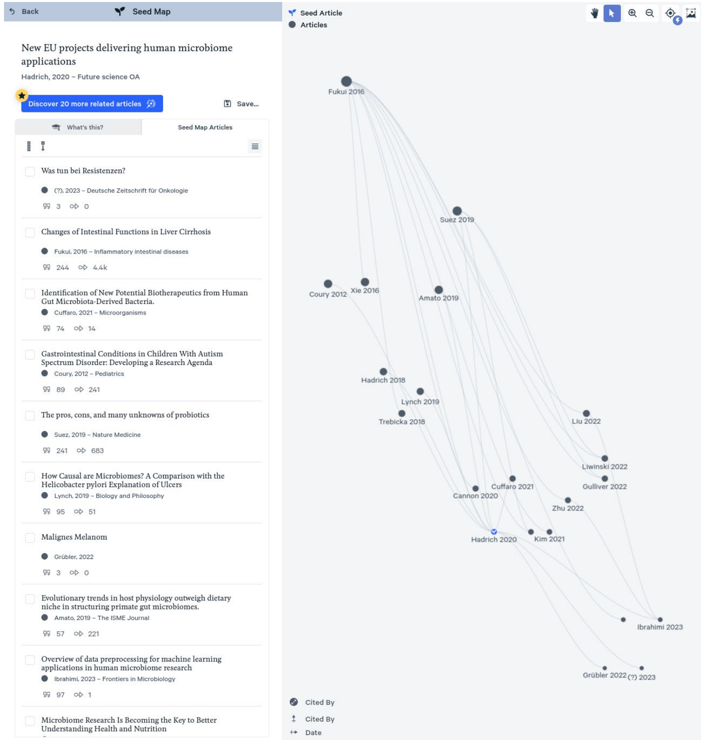
perplexity.ai

Tool assisted research

Litmaps / Research rabbit / Inciteful

Literature

- Article discovery
- Relatedness
- Citation map
- Contextualization



New EU projects delivering human microbiome applications
Hadrich, 2020 – Future science OA

☐ Was tun bei Resistenzen?
● (?), 2023 – Deutsche Zeitschrift für Onkologie
3 0

☐ Changes of Intestinal Functions in Liver Cirrhosis
● Fukui, 2016 – Inflammatory intestinal diseases
244 4.4k

☐ Identification of New Potential Biotherapeutics from Human Gut Microbiota-Derived Bacteria.
● Cuffaro, 2021 – Microorganisms
74 14

☐ Gastrointestinal Conditions in Children With Autism Spectrum Disorder: Developing a Research Agenda
● Coury, 2012 – Pediatrics
89 241

☐ The pros, cons, and many unknowns of probiotics
● Suez, 2019 – Nature Medicine
241 683

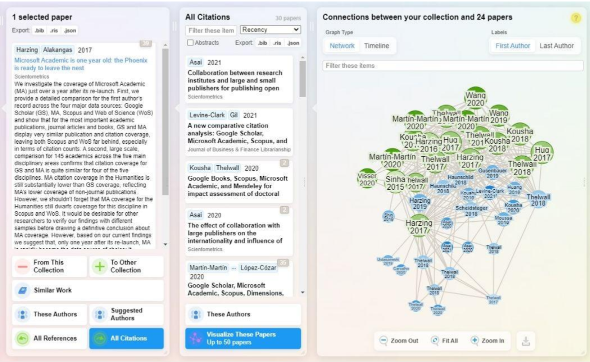
☐ How Causal are Microbiomes? A Comparison with the Helicobacter pylori Explanation of Ulcers
● Lynch, 2019 – Biology and Philosophy
95 51

☐ Malignes Melanom
● Grübler, 2022
3 0

☐ Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes.
● Amato, 2019 – The ISME Journal
57 221

☐ Overview of data preprocessing for machine learning applications in human microbiome research
● Ibrahimi, 2023 – Frontiers in Microbiology
97 1

☐ Microbiome Research Is Becoming the Key to Better Understanding Health and Nutrition



Harzing Alakangas 2017
Microsoft Academic is one year old: the Phoenix is ready to leave the nest
Scientometrics
We investigate the coverage of Microsoft Academic (MA) just over a year after its re-launch. First, we provide a detailed comparison for the first author's record across the four major data sources: Google Scholar (GS), MA, Scopus and Web of Science (WoS) and show that for the most important academic publications, journal articles and books, GS and MA display very similar publication and citation coverage, leaving both Scopus and WoS far behind, especially in terms of citation counts. A second, large scale, comparison for 145 academics across the five main disciplinary areas confirms that citation coverage for GS and MA is quite similar for four of the five disciplines. MA citation coverage in the Humanities is still substantially lower than GS coverage, reflecting MA's lower coverage of non-journal publications. However, we shouldn't forget that MA coverage for the Humanities still dwarfs coverage for this discipline in Scopus and WoS. It would be desirable for other researchers to verify our findings with different samples before drawing a definitive conclusion about MA coverage. However, based on our current findings we suggest that, only one year after its re-launch, MA

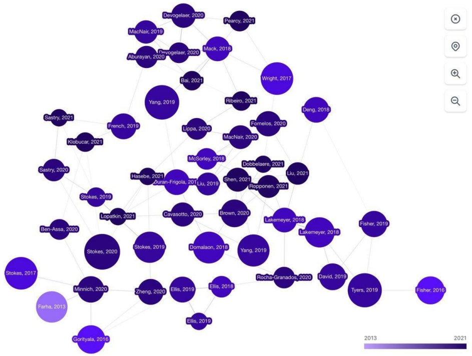
Asai 2021
Collaboration between research institutes and large and small publishers for publishing open
Scientometrics

Levine-Clark Gil 2021
A new comparative citation analysis: Google Scholar, Microsoft Academic, Scopus, and
Journal of Business & Finance Librarianship

Kousha Thelwall 2020
Google Books, Scopus, Microsoft Academic, and Mendeley for impact assessment of doctoral

Asai 2020
The effect of collaboration with large publishers on the internationality and influence of
Scientometrics

Martín-Martín … López-Cózar 2020
Google Scholar, Microsoft Academic, Scopus, Dimensions,

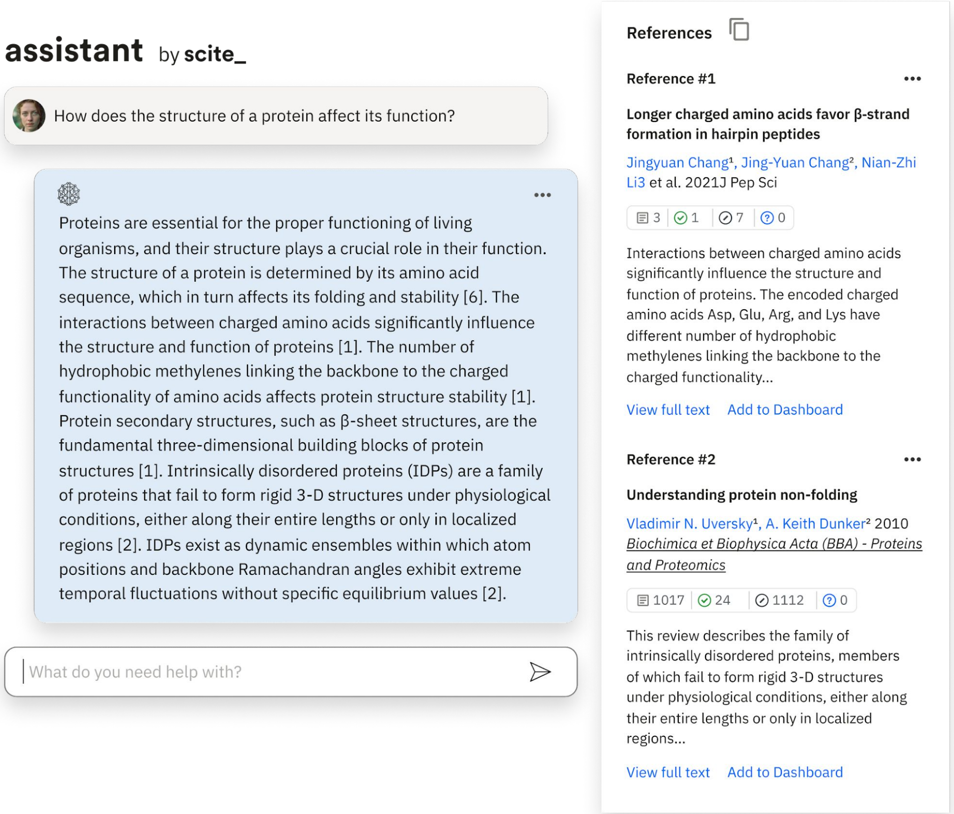


litmaps.com / researchrabbitapp.com / inciteful.xyz

Scite

Literature

- Review
- Summarization
- Reference verification
- Contextualization



How does the structure of a protein affect its function?

Proteins are essential for the proper functioning of living organisms, and their structure plays a crucial role in their function. The structure of a protein is determined by its amino acid sequence, which in turn affects its folding and stability [6]. The interactions between charged amino acids significantly influence the structure and function of proteins [1]. The number of hydrophobic methylenes linking the backbone to the charged functionality of amino acids affects protein structure stability [1]. Protein secondary structures, such as β-sheet structures, are the fundamental three-dimensional building blocks of protein structures [1]. Intrinsically disordered proteins (IDPs) are a family of proteins that fail to form rigid 3-D structures under physiological conditions, either along their entire lengths or only in localized regions [2]. IDPs exist as dynamic ensembles within which atom positions and backbone Ramachandran angles exhibit extreme temporal fluctuations without specific equilibrium values [2].

References

Reference #1

Longer charged amino acids favor β-strand formation in hairpin peptides

Jingyuan Chang[1], Jing-Yuan Chang[2], Nian-Zhi Li3 et al. 2021J Pep Sci

Interactions between charged amino acids significantly influence the structure and function of proteins. The encoded charged amino acids Asp, Glu, Arg, and Lys have different number of hydrophobic methylenes linking the backbone to the charged functionality...

Reference #2

Understanding protein non-folding

Vladimir N. Uversky[1], A. Keith Dunker[2] 2010
Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics

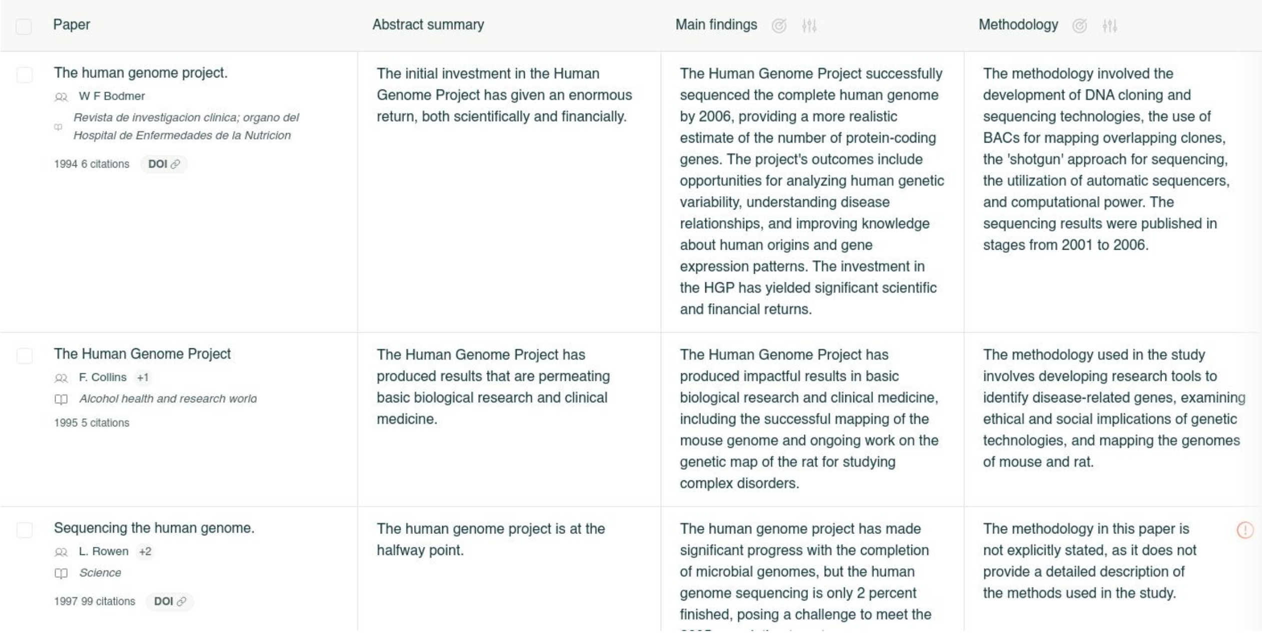
This review describes the family of intrinsically disordered proteins, members of which fail to form rigid 3-D structures under physiological conditions, either along their entire lengths or only in localized regions...

scite.ai

Elicit

Literature

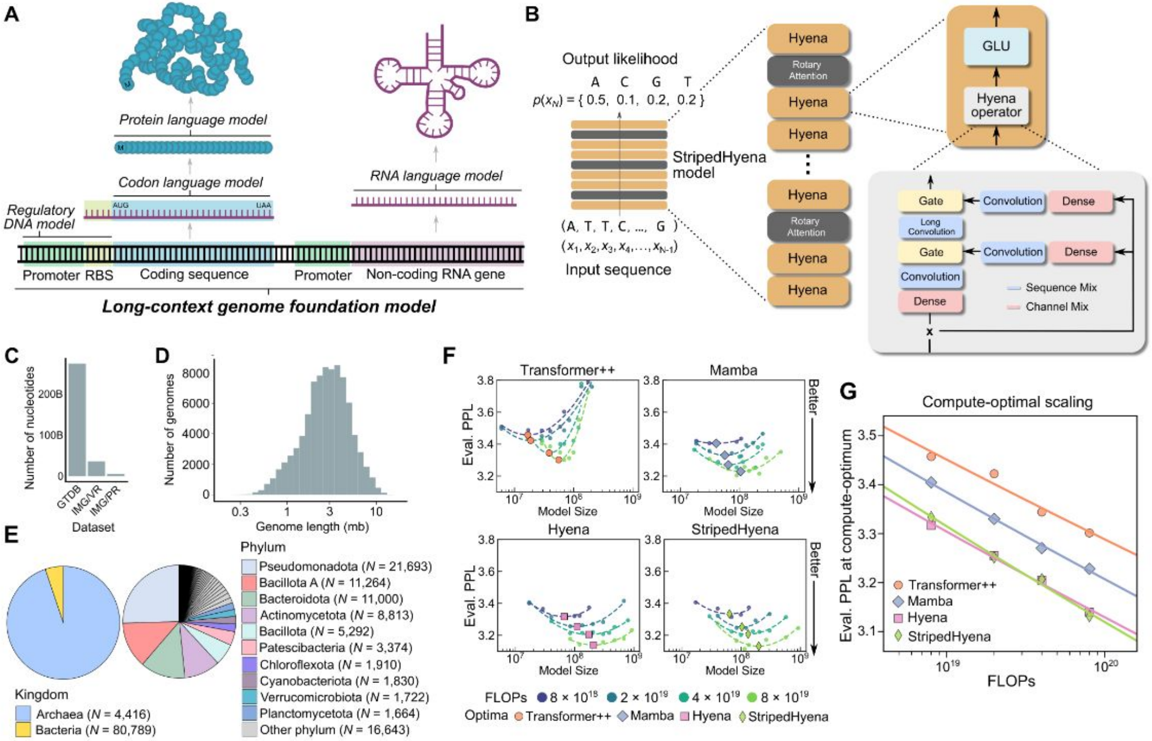
- Semantic analysis
- Review
- Summarization
- Contextualization
- PDF → “Write me an abstract”



☐	Paper	Abstract summary	Main findings	Methodology
☐	The human genome project. W F Bodmer *Revista de investigacion clinica; organo del Hospital de Enfermedades de la Nutricion* 1994 6 citations DOI	The initial investment in the Human Genome Project has given an enormous return, both scientifically and financially.	The Human Genome Project successfully sequenced the complete human genome by 2006, providing a more realistic estimate of the number of protein-coding genes. The project's outcomes include opportunities for analyzing human genetic variability, understanding disease relationships, and improving knowledge about human origins and gene expression patterns. The investment in the HGP has yielded significant scientific and financial returns.	The methodology involved the development of DNA cloning and sequencing technologies, the use of BACs for mapping overlapping clones, the 'shotgun' approach for sequencing, the utilization of automatic sequencers, and computational power. The sequencing results were published in stages from 2001 to 2006.
☐	The Human Genome Project F. Collins +1 *Alcohol health and research world* 1995 5 citations	The Human Genome Project has produced results that are permeating basic biological research and clinical medicine.	The Human Genome Project has produced impactful results in basic biological research and clinical medicine, including the successful mapping of the mouse genome and ongoing work on the genetic map of the rat for studying complex disorders.	The methodology used in the study involves developing research tools to identify disease-related genes, examining ethical and social implications of genetic technologies, and mapping the genomes of mouse and rat.
☐	Sequencing the human genome. L. Rowen +2 *Science* 1997 99 citations DOI	The human genome project is at the halfway point.	The human genome project has made significant progress with the completion of microbial genomes, but the human genome sequencing is only 2 percent finished, posing a challenge to meet the	The methodology in this paper is not explicitly stated, as it does not provide a detailed description of the methods used in the study.

elicit.com

Genomic models - capture biological languages



Evo, a long-context biological foundation model based on the StripedHyena architecture that generalizes across the fundamental languages of biology: DNA, RNA, and proteins.

Evo is capable of both prediction tasks and generative design, from molecular to whole genome scale (over 650k tokens in length).

Evo is trained at a nucleotide (byte) resolution, on a large corpus of prokaryotic genomic sequences covering 2.7 million whole genomes.

Evo model, ArcInstitute - source and BioRxiv doi:10.1101/2024.02.27.582234v1

Knowledge exploration

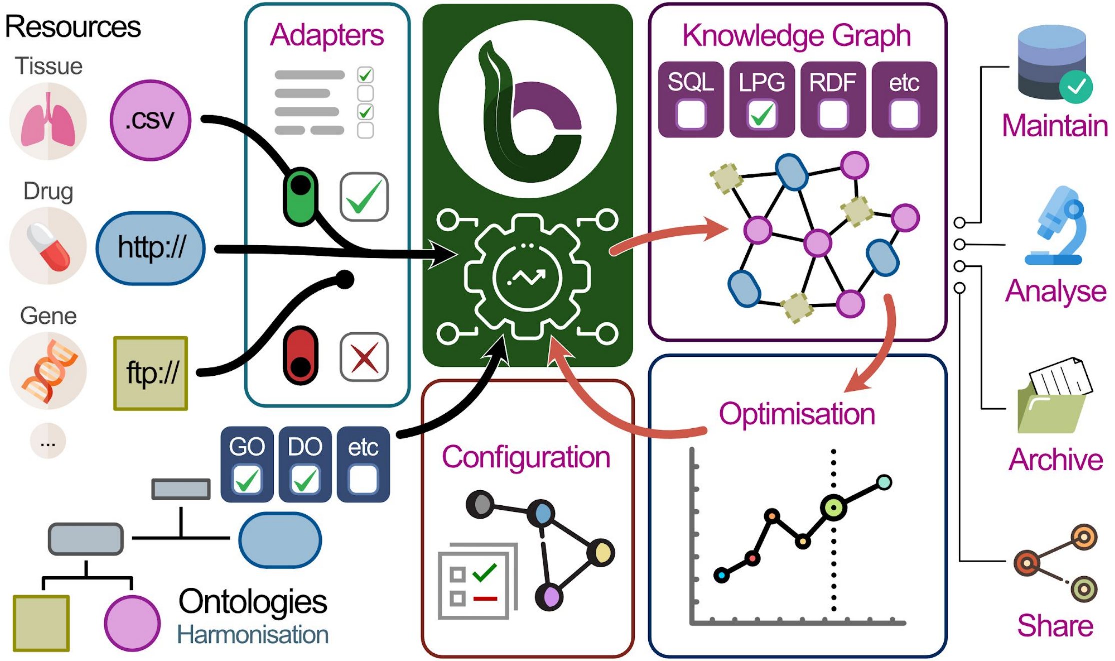
biocypher

a unifying framework for biomedical knowledge graphs

Complex data

- Graphs for structured knowledge
- AI as gateway
- Powerful combo

- Great documentation
- Complex to setup



biocypher.org

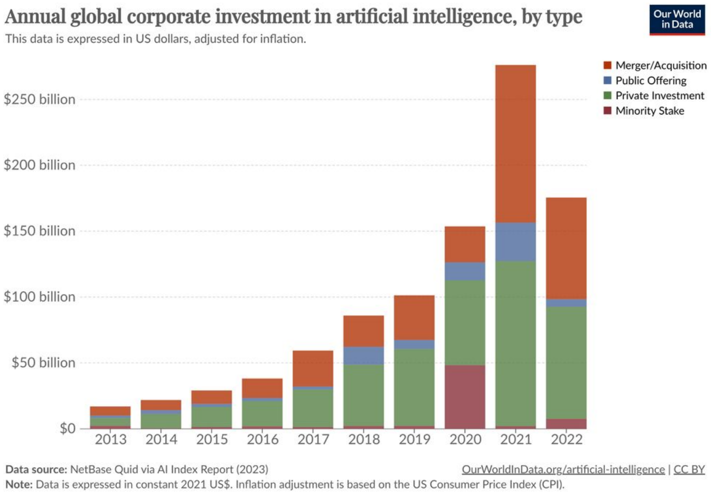
MICROB-PREDICT

EMBL 50 YEARS | 1974–2024

GPT-4V - Vision

Figures

- Interpretation
- “Write me a legend”


Annual global corporate investment in artificial intelligence, by type

This data is expressed in US dollars, adjusted for inflation.

Data source: NetBase Quid via AI Index Report (2023) OurWorldInData.org/artificial-intelligence | CC BY

Note: Data is expressed in constant 2021 US$. Inflation adjustment is based on the US Consumer Price Index (CPI).

🤖: This is a bar chart visualizing the **annual global corporate investment in artificial intelligence** (AI) by type, from 2013 to 2022. The data are presented in U.S. dollars and have been adjusted for inflation to 2021 dollars. This adjustment is based on the U.S. Consumer Price Index (CPI).

The chart is broken down into four types of investments:

1. **Merger/Acquisition**: represented by red on the top of some columns, this indicates the amount of money invested in AI through mergers and acquisitions.

2. **Public Offering**: shown in blue, this includes investments made through public offerings, possibly referring to investments made during IPOs or subsequent public investments into AI companies.

3. **Private Investment**: shown in green, this represents private funding rounds such as venture capital, private equity, or other types of private placement.

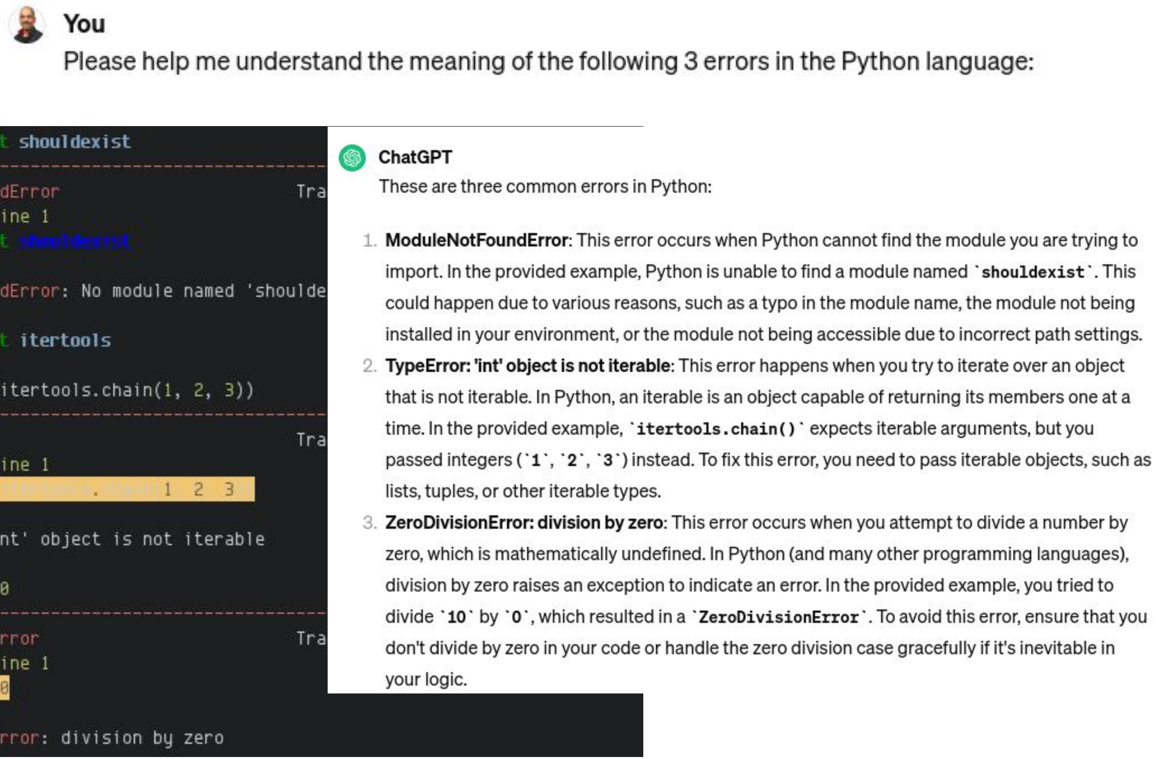
4. **Minority Stake**: represented in orange, this signifies investments made by purchasing a minority stake in AI companies.

The vertical axis indicates the amount of investment in billions of dollars, while the horizontal axis represents the year. **There's a clear trend of increasing investment in AI over the years**, with a significant jump from 2020 to 2021 and a slight decrease from 2021 to 2022. In 2022, the investments are spread across all four types, with private investment appearing to comprise the largest share.

ChatGPT - for code

Code troubleshooting

- Beginner friendly
- Explains with context
- Almost as good as a human, sometimes better



You

Please help me understand the meaning of the following 3 errors in the Python language:

```
In [1]: import shouldexist
---------------------------------------------------------
ModuleNotFoundError                               Tra
Cell In[1], line 1
----> 1 import shouldexist

ModuleNotFoundError: No module named 'shoulde

In [2]: import itertools

In [3]: list(itertools.chain(1, 2, 3))
---------------------------------------------------------
TypeError                                         Tra
Cell In[3], line 1
----> 1 list(itertools.chain(1, 2, 3))

TypeError: 'int' object is not iterable

In [4]: 10 / 0
---------------------------------------------------------
ZeroDivisionError                                 Tra
Cell In[4], line 1
----> 1 10 / 0

ZeroDivisionError: division by zero
```

ChatGPT

These are three common errors in Python:

1. **ModuleNotFoundError**: This error occurs when Python cannot find the module you are trying to import. In the provided example, Python is unable to find a module named `shouldexist`. This could happen due to various reasons, such as a typo in the module name, the module not being installed in your environment, or the module not being accessible due to incorrect path settings.
2. **TypeError: 'int' object is not iterable**: This error happens when you try to iterate over an object that is not iterable. In Python, an iterable is an object capable of returning its members one at a time. In the provided example, `itertools.chain()` expects iterable arguments, but you passed integers (`1`, `2`, `3`) instead. To fix this error, you need to pass iterable objects, such as lists, tuples, or other iterable types.
3. **ZeroDivisionError: division by zero**: This error occurs when you attempt to divide a number by zero, which is mathematically undefined. In Python (and many other programming languages), division by zero raises an exception to indicate an error. In the provided example, you tried to divide `10` by `0`, which resulted in a `ZeroDivisionError`. To avoid this error, ensure that you don't divide by zero in your code or handle the zero division case gracefully if it's inevitable in your logic.

ChatGPT / CoPilot / Devin

Code generation

- Accessible to non-programmers
- From pseudo-code
- From documentation
- Semi-automated testing
- Very fast prototypes

- Harder to use with complex codebases

GitHub Copilot

AI Personas / AI Agents / AI Council

Reasoning / Ideation

- Understand points of view
- Highlight biases
- Complementary feedback
- Not (yet) self sufficient

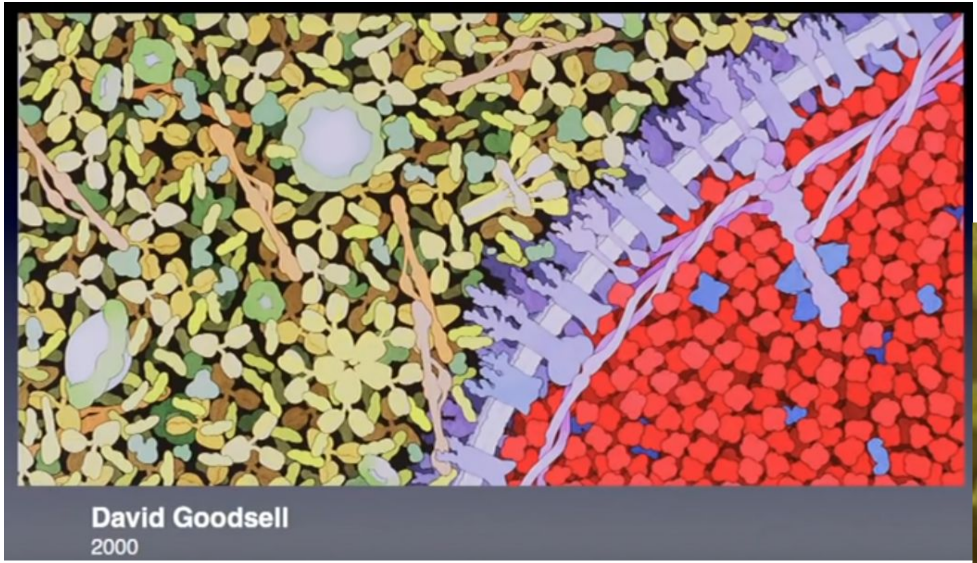
newo.ai

Image from spyscape.com / ChatGPT AI Council

MICROB-PREDICT EMBL 50 YEARS 1974–2024

Scientific communication

Illustrations



WEHI.TV
DREW BERRY
COVID-19 vs HUMAN ANTIBODIES x1,000,000

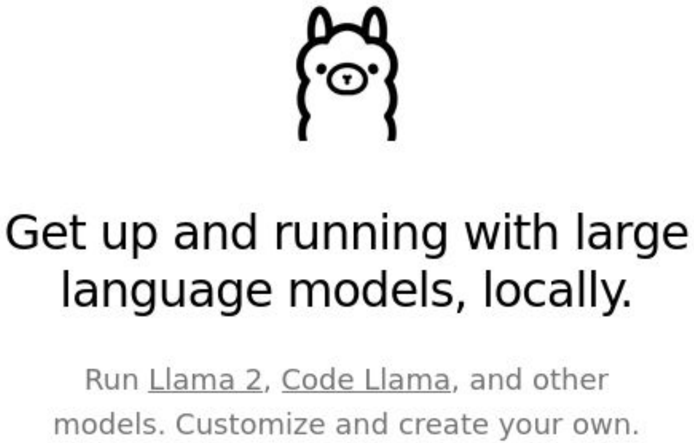
Video generation

Prompt: A gorgeously rendered papercraft world of a coral reef, rife with colorful fish and sea creatures.

Sora by OpenAI

When (data) privacy is important

Running models privately

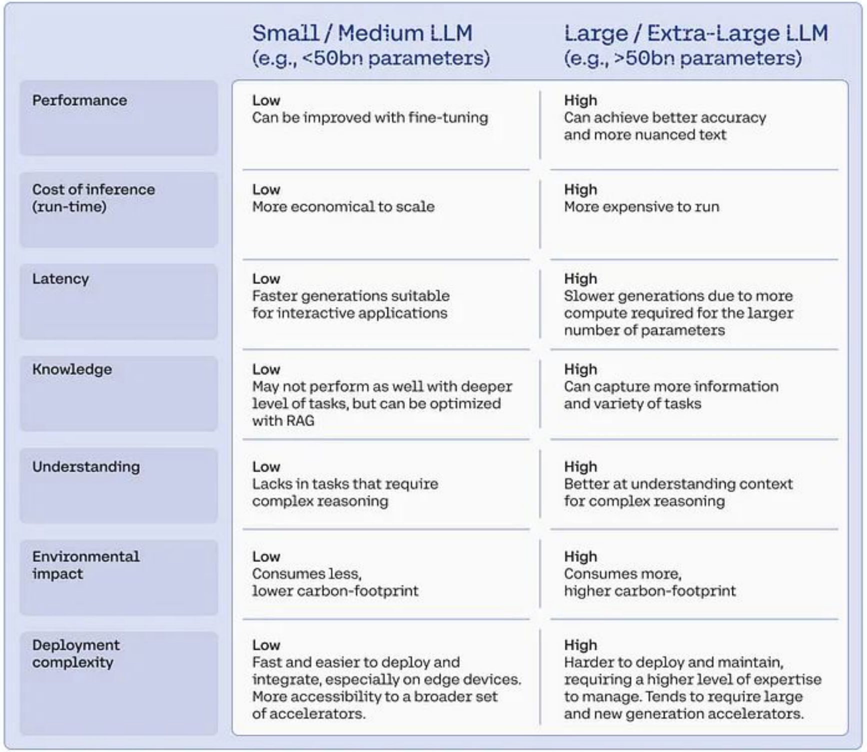


Get up and running with large language models, locally.

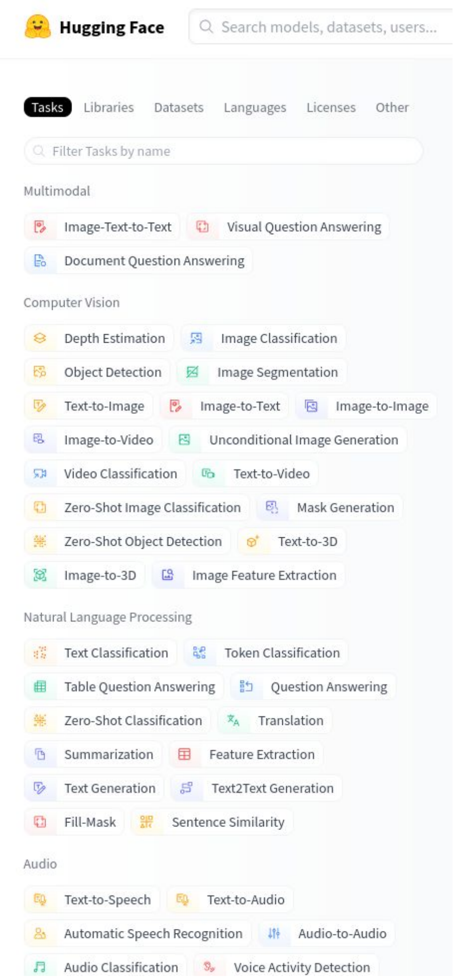
Run Llama 2, Code Llama, and other models. Customize and create your own.

ollama.com

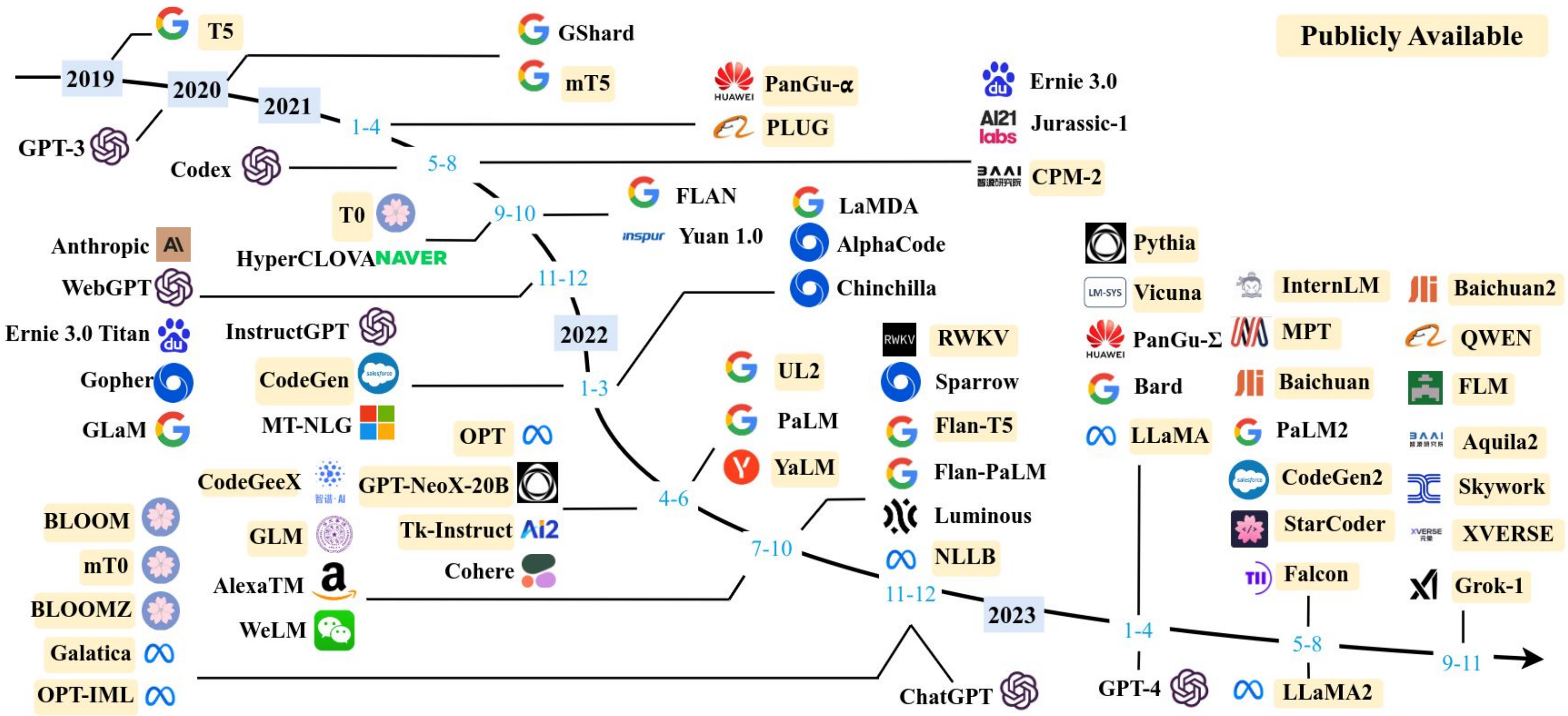
The right model and hardware



	Small / Medium LLM (e.g., <50bn parameters)	Large / Extra-Large LLM (e.g., >50bn parameters)
Performance	Low Can be improved with fine-tuning	High Can achieve better accuracy and more nuanced text
Cost of inference (run-time)	Low More economical to scale	High More expensive to run
Latency	Low Faster generations suitable for interactive applications	High Slower generations due to more compute required for the larger number of parameters
Knowledge	Low May not perform as well with deeper level of tasks, but can be optimized with RAG	High Can capture more information and variety of tasks
Understanding	Low Lacks in tasks that require complex reasoning	High Better at understanding context for complex reasoning
Environmental impact	Low Consumes less, lower carbon-footprint	High Consumes more, higher carbon-footprint
Deployment complexity	Low Fast and easier to deploy and integrate, especially on edge devices. More accessibility to a broader set of accelerators.	High Harder to deploy and maintain, requiring a higher level of expertise to manage. Tends to require large and new generation accelerators.



A lot to choose from



A Survey of Large Language Models - doi: 10.48550/arXiv.2303.18223

Wrapping up

Summary & Challenges / Opportunities

AI can be immensely useful if used correctly

- Trainees are more independent / Not afraid to ask
- “Hallucinations” / incorrect “*facts*” → ok for reasoning / not-ok for decision making
- Data privacy / *Cross*-contamination → ok for generation / not-ok for review
- Copyright / Intellectual property rights → Regulation playing catch → Lawsuits
- Computation / Energy / Hardware cost → 1M€ BLOOM LLM → “*GPUrush”*
- Cultural bias → Notions of right and wrong → Really hard problem
- Even greater information overflow → AI as solution → circular problem
- Society implications → New jobs / Obsolete jobs → Automation everywhere
- Fast pace and hype → Widening technology gap low-, mid-, high-income countries
- Challenging current definitions: creativity, originality, plagiarism,...
- Human-machine interactions → More forgiving towards humans than machines

Thank you

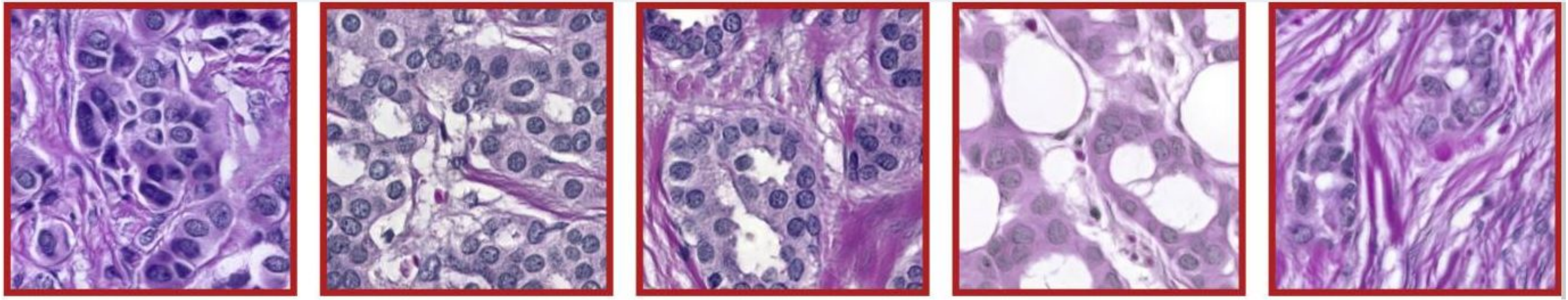
Michael Khun
Marisa Keller

AI discussion group
Bork group

DOI: 10.1016/j.patter.2022.100511

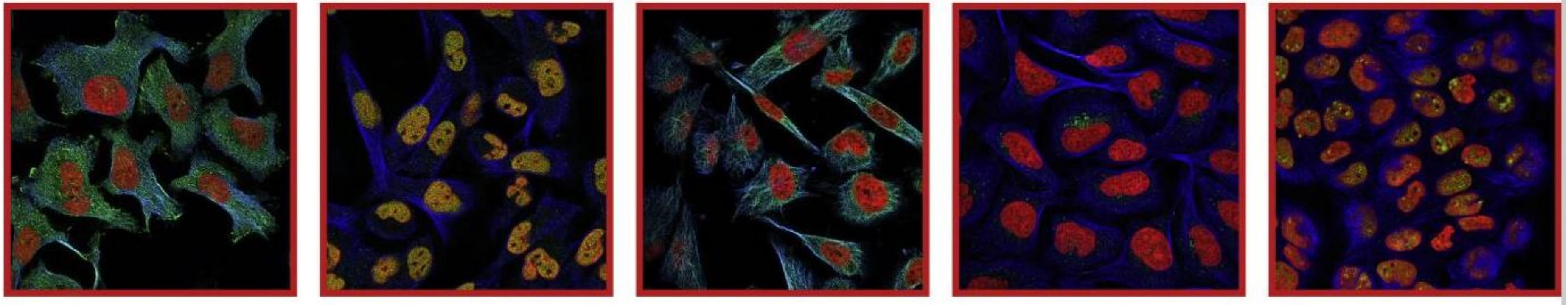
MICROB-PREDICT

EMBL
50 YEARS | 1974–2024



Thank you / Questions?

PS: Find the AI generated



DOI: 10.1016/j.patter.2022.100511

They are all AI generated

-
  -