DHUM 25A43 - 01 Investigating with AI

Welcome
January 28, 2025

Discord

- **#sciencespo-DHUM25A43**
- https://discord.gg/DDbh5AyHYH

Meet the teachers

Andreï Mogoutov

- PhD UPMC 93'
- Data scientist, researcher, consultant
- andrei.mogoutov@sciencespo.fr

Alexis Perrier

- PhD TelecomParis 95'
- Data scientist, author, teacher
- alexis.perrier@gmail.com

The Course

What is this course about?

Data Science for Social Sciences: Applying data science techniques to social science research.

- **Data Search and Collection**: Searching, gathering, and transforming diverse datasets to create corpora and heterogeneous data collections.
- **Exploratory Data Analysis**: Exploring textual content, categories, temporal evolution, extracting names, topics, and their relationships.
- **Qualitative and Quantitative Results**: Combining qualitative insights with quantitative analysis and visual representation of findings.
- **Result Communication**: Presenting outcomes through interactive documents and websites.

Digital investigations with AI for social sciences

- **Powerful LLMs**: Revolutionizing data analysis, interpretation, and automation.
- **Impact on Data Science**: Enhancing efficiency, accuracy, and scalability of workflows.
- **Code Generation with LLM :** Quickly develop data analysis pipelines
- **Use of LLMs as an Autonomous Data Analysis Tool:** extract patterns, trends, and insights from raw data. Analysis of large datasets or documents
- **Accessible Methods**: Simplifying complex data processes with user-friendly tools.

Course Organization

- **Project-Oriented Approach**: Focus on applying concepts through real-world projects.
- **Workshop-Style Sessions**: Interactive, hands-on learning environment.
- **Collaborative Work**: Students work in groups to tackle selected themes and datasets.
- **Practical Focus**: Emphasis on skills development through guided exercises and demos.
- **Professor Assistance**: Direct support and feedback during project work.

Session Structure

Theoretical Introduction: Present existing methods (no advanced math or equations) and new AI-assisted approaches.

Technical Demo: Showcase practical applications of the methods introduced.

Practical Student Work:

- Groups of 3-4 students collaborate on selected themes and datasets.
- Hands-on project development with guidance from professors.

Course Evaluation

- group note (80%)
- public presentation
- individual : (20%)
- critical reflexion on your work and experience with IA
- write a personal SWOT report on your work with AI and propositions

Course outline - timeline

We have 12 classes, 2h each

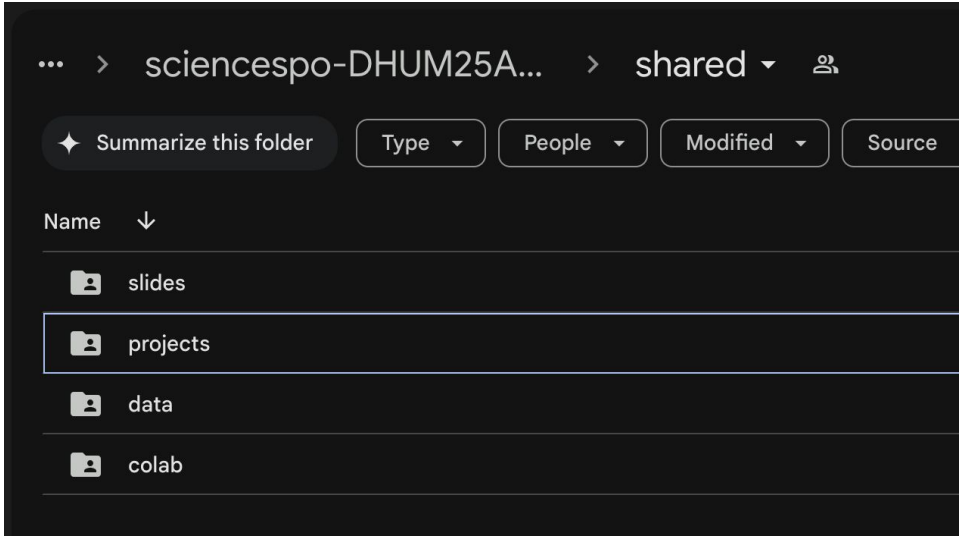
We'll cover

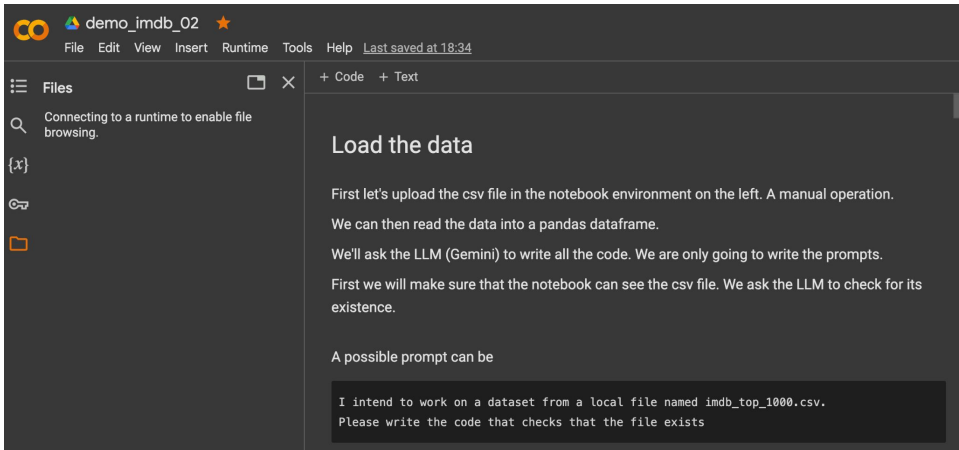
- Background on Web and API
- Data: how to collect and process
- Machine Learning : training models for prediction
- NLP : analyzing text to extract information
- AI & LLMs
- Agents : combining LLMs for agency

Tools

- discord for conversations
- google drive for documents
  - course material
    - ./slides
    - ./data
    - ./colab
  - student projects
    - ./projects
- Google colab notebooks for work

All course material is available in google drive and on the discord channel





Course Project

Project Presentation

Themes

- Climate change, energy
- AI - robotics
- Brain–computer interface
- Other issues…
- Data sources : media, social media, web, scientific publications, specific websites (COP, IMDB, wikipedia, Kaggle datasets)

Comment 1 : These topics focus on future-oriented challenges and opportunities

Comment 2 : The issues are very broad; we need to refine them into precise research questions or subtopics to ensure project feasibility and a reasonable dataset size.

Project Organization

- Find and Formulate Research Questions
- Initial Validation of Projects - Feasibility Check, Relevance
- Goal: Create a Website: Design an engaging platform to showcase your report.
- Creative Formats: Present findings as a report or in unique forms, such as a movie script or other innovative approaches.
- Publish the website
- Organization: 3 to 4, groups with complementary skills
- Evaluation, expectations
- Primary exposé after a few classes
- Last class: final exposé in front of class and experts

Project

Start thinking about your project

Share your project definition

Create your team

Getting to know you

Please fill out this form

We'd like to know a bit more about you

So that we can adapt the course to meet your expectations

All questions are optional (except your email)

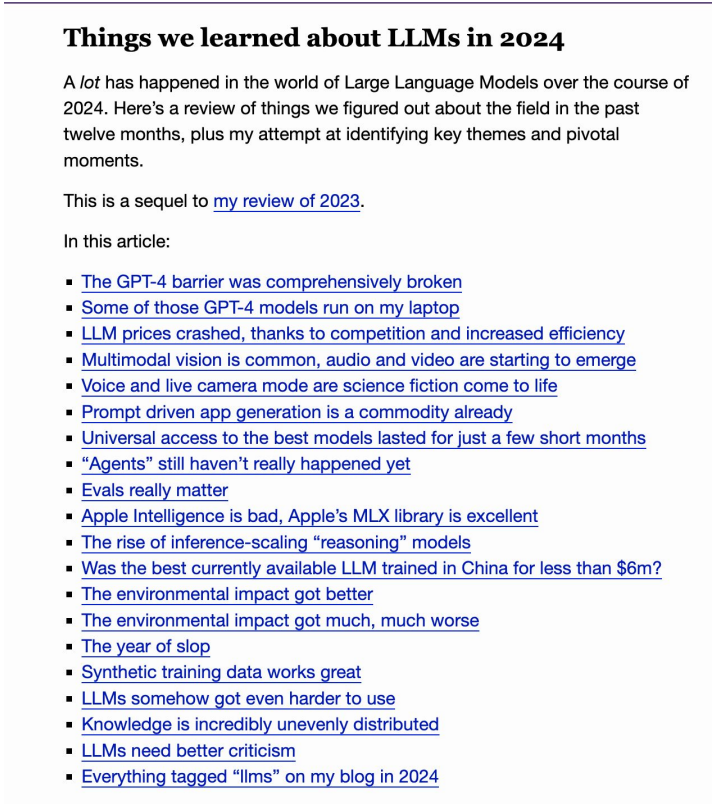
https://forms.gle/1ksP3qtAk2F5N2gT8

State of AI

A Thorough Recap of 2024

[Things we learned about LLMs in 2024]

by [Simon Willison].



Things we learned about LLMs in 2024

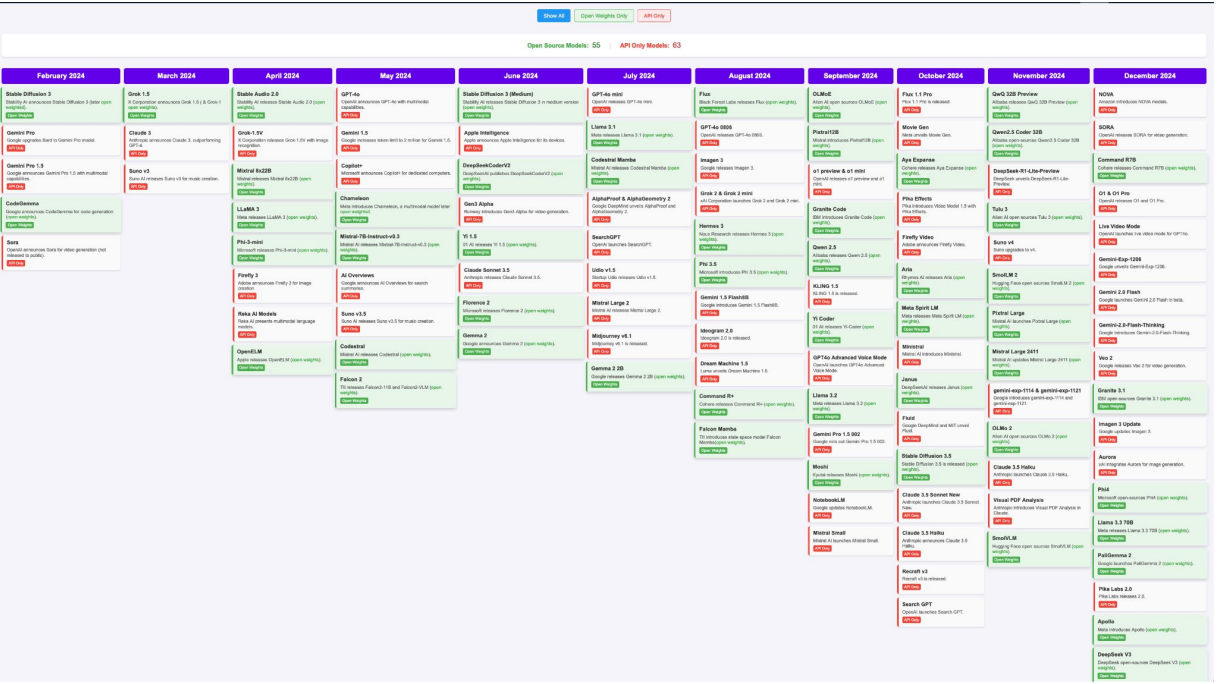
A *lot* has happened in the world of Large Language Models over the course of 2024. Here's a review of things we figured out about the field in the past twelve months, plus my attempt at identifying key themes and pivotal moments.

This is a sequel to my review of 2023.

In this article:

- The GPT-4 barrier was comprehensively broken
- Some of those GPT-4 models run on my laptop
- LLM prices crashed, thanks to competition and increased efficiency
- Multimodal vision is common, audio and video are starting to emerge
- Voice and live camera mode are science fiction come to life
- Prompt driven app generation is a commodity already
- Universal access to the best models lasted for just a few short months
- "Agents" still haven't really happened yet
- Evals really matter
- Apple Intelligence is bad, Apple's MLX library is excellent
- The rise of inference-scaling "reasoning" models
- Was the best currently available LLM trained in China for less than $6m?
- The environmental impact got better
- The environmental impact got much, much worse
- The year of slop
- Synthetic training data works great
- LLMs somehow got even harder to use
- Knowledge is incredibly unevenly distributed
- LLMs need better criticism
- Everything tagged "llms" on my blog in 2024

Over 120 new models released in 2024!



Released Models 2024

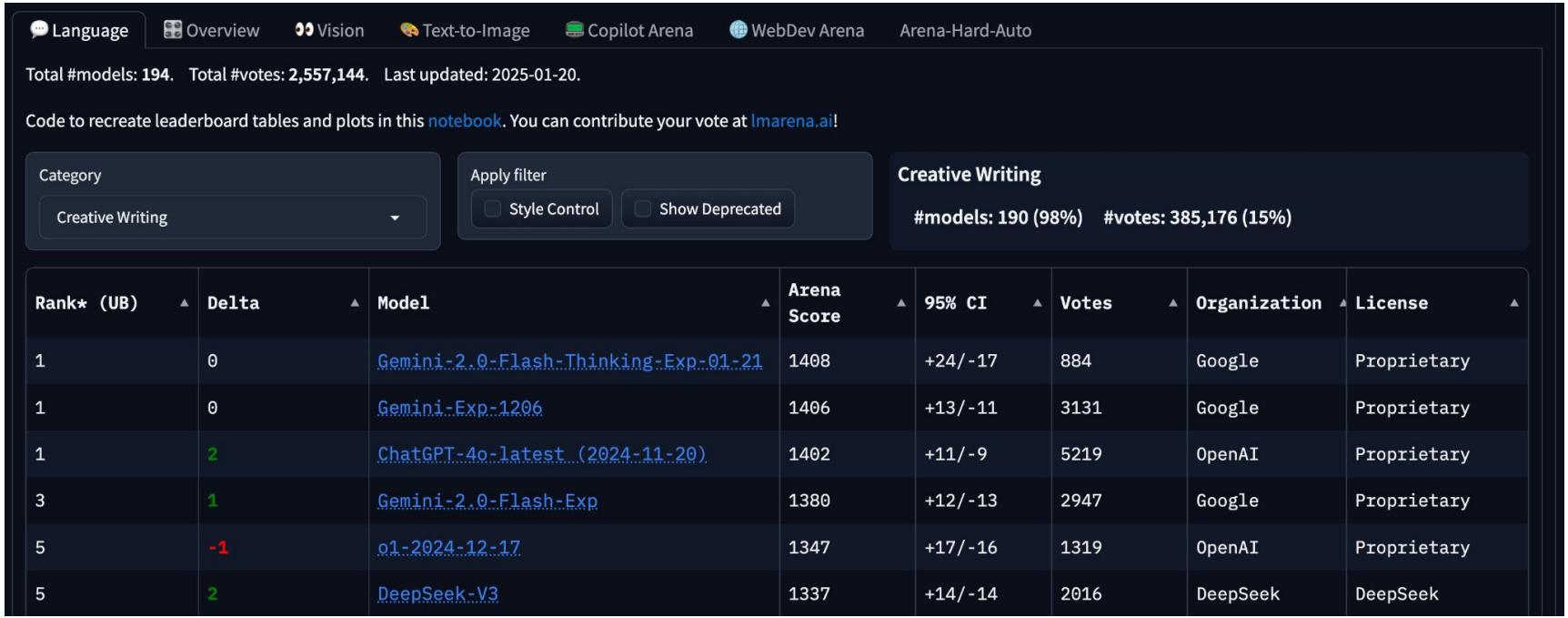
Open Source Models: 55

API Only Models: 63

[2024 AI Timeline a Hugging Face Space]

Chatbot Arena: LLMs vs LLMs

Where LLMs compete https://lmarena.ai/?leaderboard



Language | Overview | Vision | Text-to-Image | Copilot Arena | WebDev Arena | Arena-Hard-Auto

Total #models: **194**. Total #votes: **2,557,144**. Last updated: 2025-01-20.

Code to recreate leaderboard tables and plots in this notebook. You can contribute your vote at lmarena.ai!

Category: Creative Writing

Apply filter: Style Control | Show Deprecated

Creative Writing

#models: 190 (98%) **#votes: 385,176 (15%)**

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License
1	0	Gemini-2.0-Flash-Thinking-Exp-01-21	1408	+24/-17	884	Google	Proprietary
1	0	Gemini-Exp-1206	1406	+13/-11	3131	Google	Proprietary
1	2	ChatGPT-4o-latest (2024-11-20)	1402	+11/-9	5219	OpenAI	Proprietary
3	1	Gemini-2.0-Flash-Exp	1380	+12/-13	2947	Google	Proprietary
5	-1	o1-2024-12-17	1347	+17/-16	1319	OpenAI	Proprietary
5	2	DeepSeek-V3	1337	+14/-14	2016	DeepSeek	DeepSeek

2024

70 models from 18 organizations are now performing better than GPT4

- GPT-4 level models run on a laptop (not mine tho)
- Multimodal vision is common, audio and video are starting to emerge
- Prices have dropped => less energy
- EU AI act

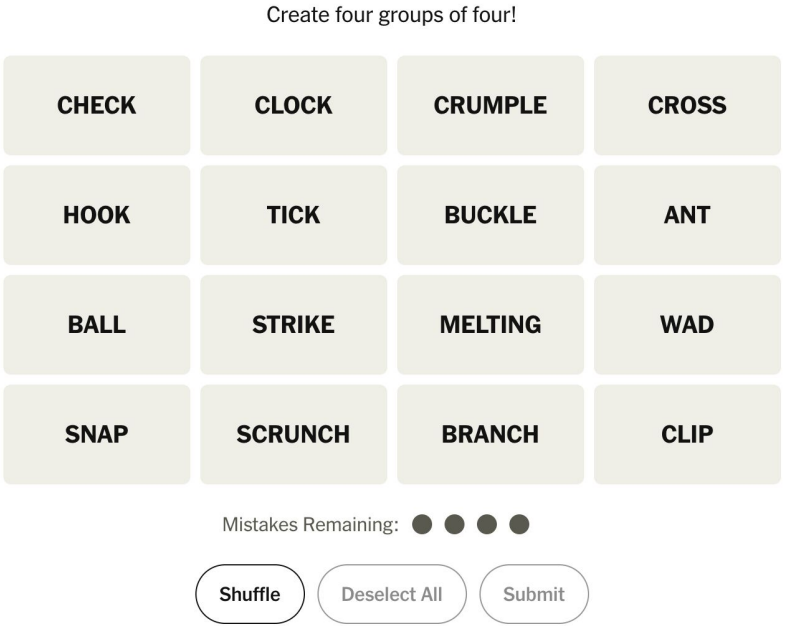
Benchmarks

- The Massive Multitask Language Understanding (MMLU) : range of exam questions on academic subjects.
- BIG-bench : Beyond the Imitation Game
- GPQA Diamond: harder still, google proof,
  - PhDs domain experts reach 74% accuracy
- LiveBench: avoid contamination

and many others

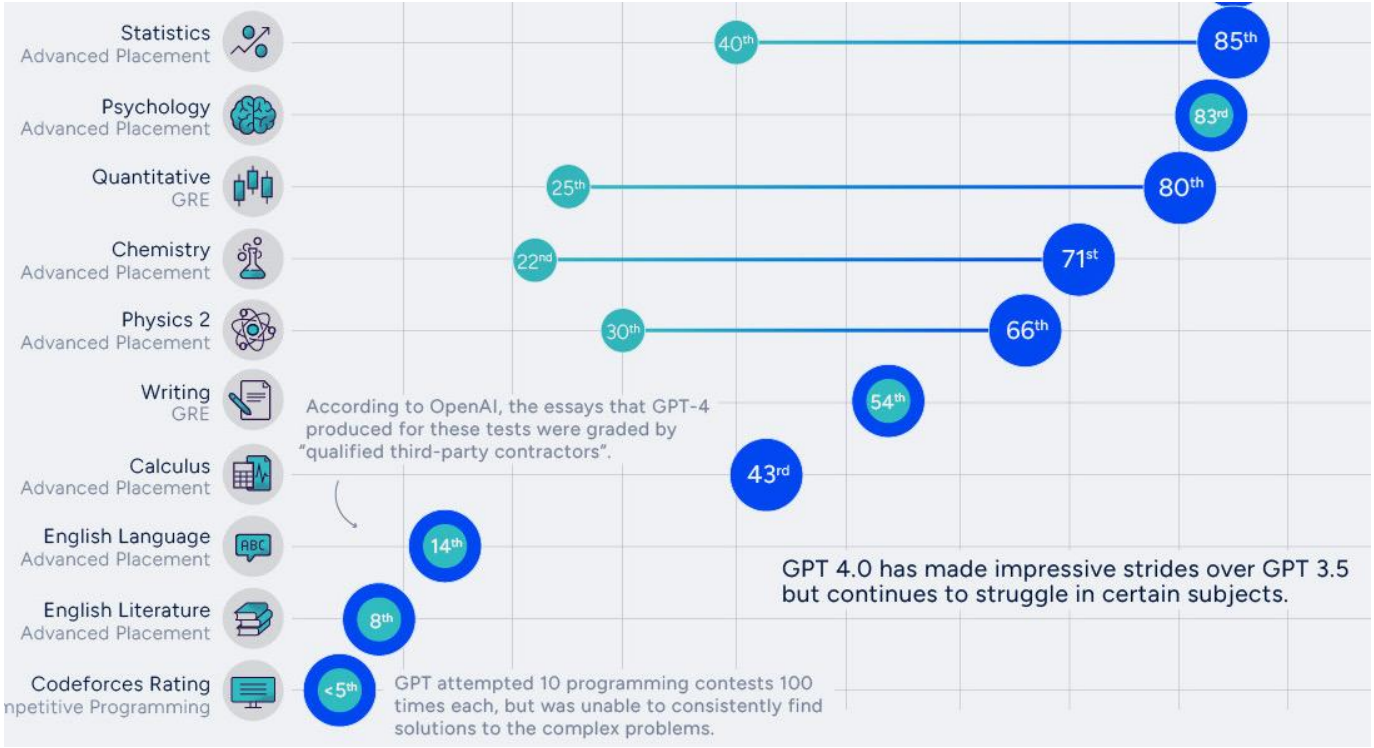
- Humanity's Last Exam
- CultureVLM
- …

It's becoming harder to find tasks that are difficult enough for the LLMs.



Connections — The New York Times

AI vs Humans (March 2023)



[How Smart is ChatGPT?]

Augmented LLMs

It's no longer just about the model

- 🦵 Web browsing
- 📚 File upload - projects - knowledge base
- 🧠 Dynamic memory
- 👁️ Streaming
- 🖐️ Function calling
- 👄 speech
- 👂 voice

2025 ?

- Models: DeepSeek-R1, OpenAI o3
- Stargate project ?!
- China vs US
- Massive deployment of AI in all Google and Microsoft services
- Platform evolutions : Anthropic, OpenAI, Gemini, …
  - more features, connections, memory, protocols

2025 ?

Ethan Mollick : [Which AI to Use Now: An Updated Opinionated Guide]

Service	Best Model	Live Mode	"Reasoning"	Web Access	Generates Images	Executes Code	Data Analysis	Sees images	Sees video	Reads Docs	Personality	Superpower
OpenAI ChatGPT	GPT-4o	✓ Full multimodal	✗	✓	✓ DALL-E3	✓	✓	✓	In Live Mode	✓	Polished and efficient in text. In live mode, expressive and adaptive.	Live mode, most versatile set of features and capabilities
	o1/o3 family	✗	✓	✗	✗	✓	✗	✓	✗	✗	Methodical and analytical	Very powerful model for complex reasoning tasks, particularly in science, coding, and mathematics
Microsoft Copilot	"Copilot"	Voice only	✓	✓	✓ DALL-E3	Limited	✗	✓	✗	✓	Since it uses different models behind the scenes, a little inconsistent	Works well with Microsoft products and services
Anthropic Claude	Claude 3.5	✗	✗	✗	✗	✓	Limited	✓	✗	✓	Clever and friendly	Often the most creative and socially engaging model
Google Gemini	Gemini family	Voice only	✓	✓	✓ Imogen-3	Limited	Limited	✓	✓	✓	Helpful and a bit bland	Wide variety of features, good connections with search
X.ai Grok	Grok-2	✗	✗	✓ Mostly X	✓ Aurora	✗	✗	✓	✗	✓	Sarcastic and "fun" (though you can tone that	Powerful model integrated tightly with X
DeepSeek	DeepSeek v3	✗	✓	✓	✗	✗	✗	✓	✗	Limited	Neurotically helpful, warm	Remarkably cheap and powerful model out of China

Emergence

Emergence

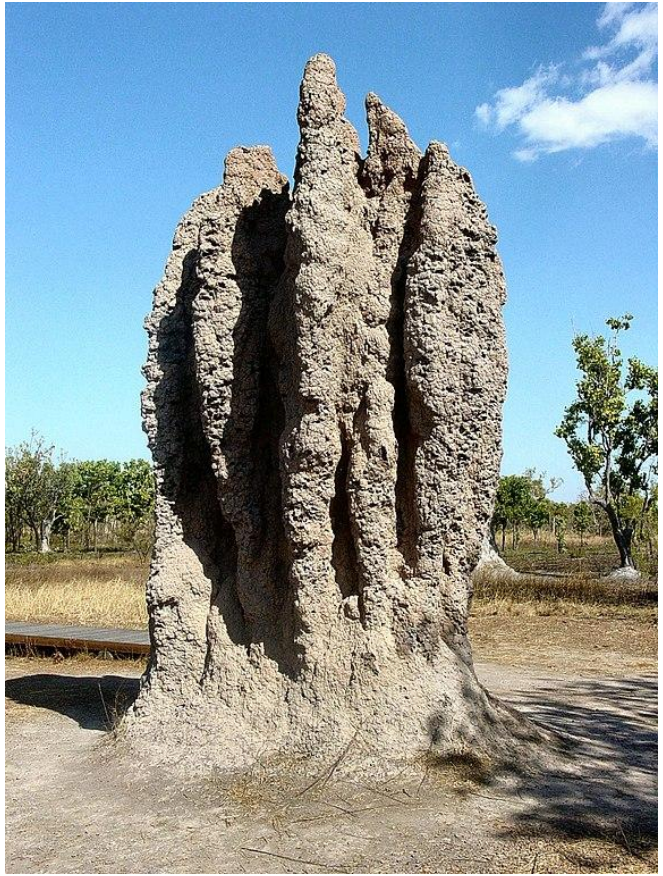
Complex systems exhibit properties or behaviors that are not reducible to their individual components.

Strong Emergence: raises the possibility of truly intelligent machines!

- The observed abilities are genuinely novel, irreducible to the individual components or algorithms of the LLM.
- the wetness of water isn't found in individual water molecules.

Weak Emergence: sophisticated but fundamentally different from human intelligence

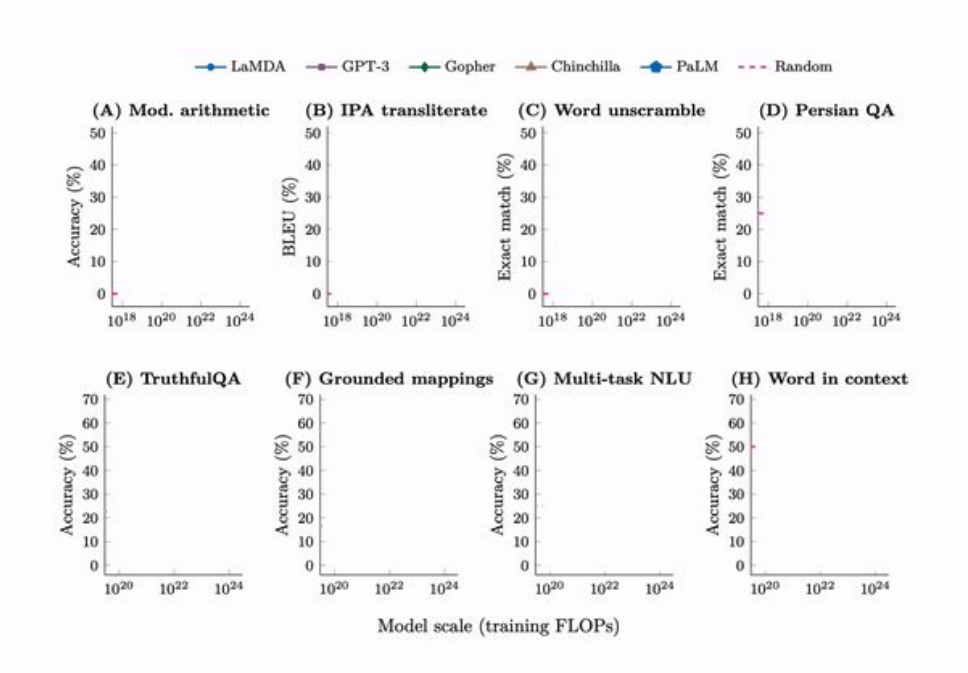
- The observed abilities can still be explained by the underlying mechanisms,
- But their appearance is unexpected and simply difficult to predict.
- ~ modeling traffic patterns using the behavior of individual cars.



A termite "cathedral" mound produced by a termite colony offers a classic example of emergence in nature.
https://en.wikipedia.org/wiki/Emergence

Emergence - Jason Wei - OpenAI

[137 emergent abilities of large language models — Jason Wei]



As the size of the model increases, we see sudden improvements in its performance on certain tasks

for instance 3 digit addition

gradually increase size of model

- errors, errors, …, errors, …

and then suddenly

- correct

LLMs you can use

- Gemini (Colab)
- Gemini (aistudio.google.com)
- chatGPT (with memory)
- DeepSeek
- Claude.ai, (create a project)
- Gemini
- LLama
- …

Google Colab

Google Colab

Like google docs but also for executing code

Shareable, collaborative work

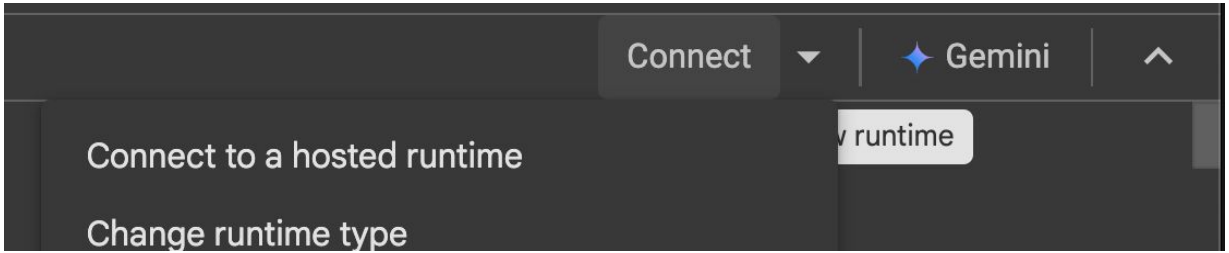
A notebook is a series of executable cells

- code (python)
- text with Markdown

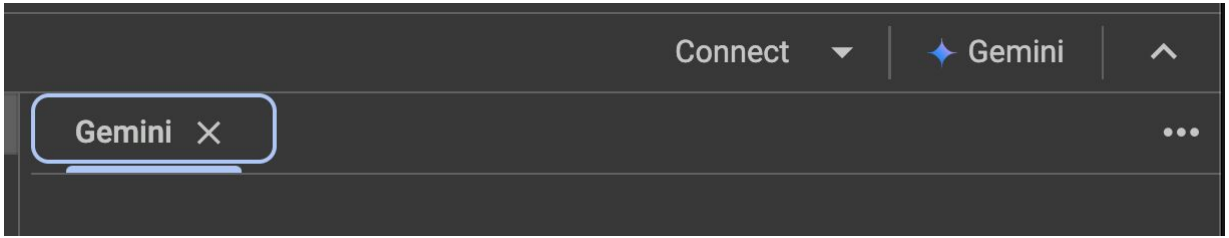
https://colab.research.google.com/

Google Colab

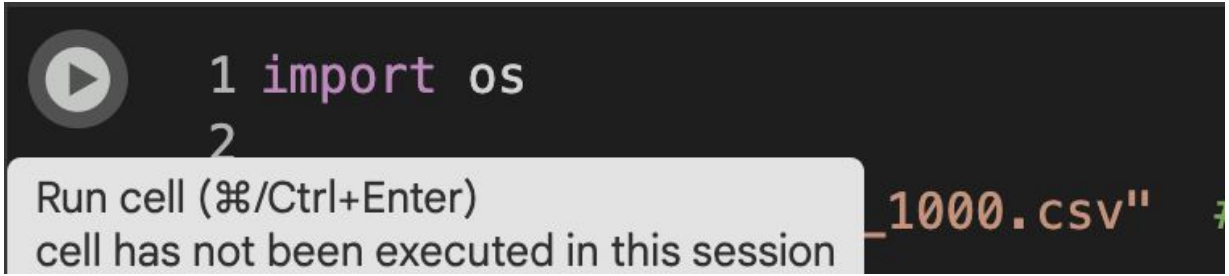
Connect to a runtime



Prompt Gemini to write the code for a given task



Click to run the code



Text with Markdown

What you write

```
# LLM

## Definition

A **large language model** (LLM) is a type of machine learning model designed
for [natural language processing](https://en.wikipedia.org/wiki/
Natural_language_processing) tasks such as language generation.

## History of NLP

- 1954s: The Georgetown experiment
- some years
- 2010, Tomáš Mikolov  applied a simple recurrent neural network with a single
hidden layer to language modelling.


## A quote:

> It is not the answer that enlightens, but the question
> Ionesco
```

What you get !

LLM

Definition

A **large language model** (LLM) is a type of machine learning model designed for natural language processing tasks such as language generation.

History of NLP

- 1954s: The Georgetown experiment
- some years
- 2010, Tomáš Mikolov applied a simple recurrent neural network with a single hidden layer to language modelling.

A quote:

> It is not the answer that enlightens, but the question Ionesco

Text with Markdown

What you write

```
Simple Syntax

# Header 1

**this is bold**, not bold

[a link](https://sciencespo.fr)
```

What you get !

Simple Syntax

Header 1

this is bold, not bold

a link

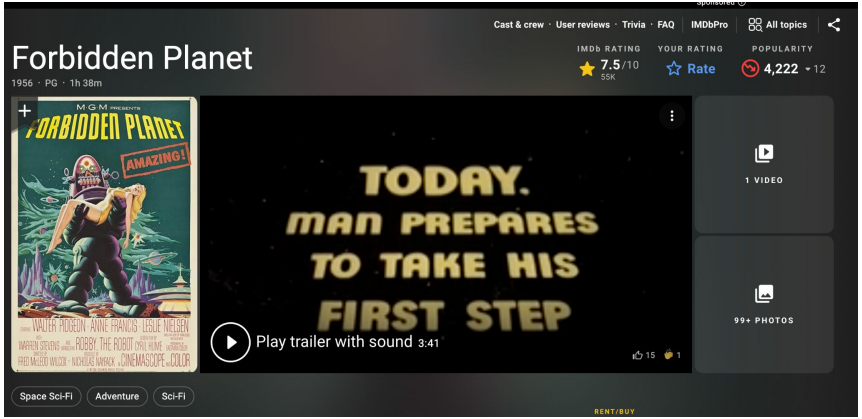
Demo

Movie Analysis

The dataset is available in the shared folder in google drive (csv file, google spreadsheet)

It contains information of a 1000 movies

- title
- description
- ranking (imdb, Meta)
- duration
- genres
- actors, director
- revenue



imdb_top_1000

File Edit View Insert Format Data Tools Extensions Help

M30 Lawrence A. Bonney

	B	C	D	E	F	G
1	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating
2	The Shawshank Redemption	1994	A	142 min	Drama	9.3
3	The Godfather	1972	A	175 min	Crime, Drama	9.2
4	The Dark Knight	2008	UA	152 min	Action, Crime, D	9
5	The Godfather: Part II	1974	A	202 min	Crime, Drama	9
6	12 Angry Men	1957	U	96 min	Crime, Drama	9
7	The Lord of the Rings: The Return of the King	2003	U	201 min	Action, Adventur	8.9
8	Pulp Fiction	1994	A	154 min	Crime, Drama	8.9
9	Schindler's List	1993	A	195 min	Biography, Dram	8.9
10	Inception	2010	UA	148 min	Action, Adventur	8.8
11	Fight Club	1999	A	139 min	Drama	8.8
12	The Lord of the Rings: The Fellowship of the Ring	2001	U	178 min	Action, Adventur	8.8
13	Forrest Gump	1994	UA	142 min	Drama, Romanc	8.8

Demo

1. create a new notebook
2. upload the csv file
3. ask Gemini to
    a. load the data
    b. analyze the data
    c. suggest & explain
    d. extract information from the movie description
    e. save the new data
4. share the notebook

Example of analysis notebook:
https://colab.research.google.com/drive/1KWmqZRSg7O2gJEWtY8NroUYX68TLkEL9#scrollTo=yoLD0yLL0Phc

Next time

Colab and AI

In Colab

- Load a dataset
- prompt Gemini on what questions you can ask on this dataset
- prompt Gemini to create the code to answer your questions

Create your Project

- Create your team
- Define your project
- Share your project definition
- Announce on the discord channel #sciencespo-dhum25a43 and we will create the related drive folders and subchannels

Need help ?

1. Ask an LLM
2. We are available on discord
    a. if possible, please post in the course channel #sciencespo-dhum25a43 not in private messages, so all can contribute