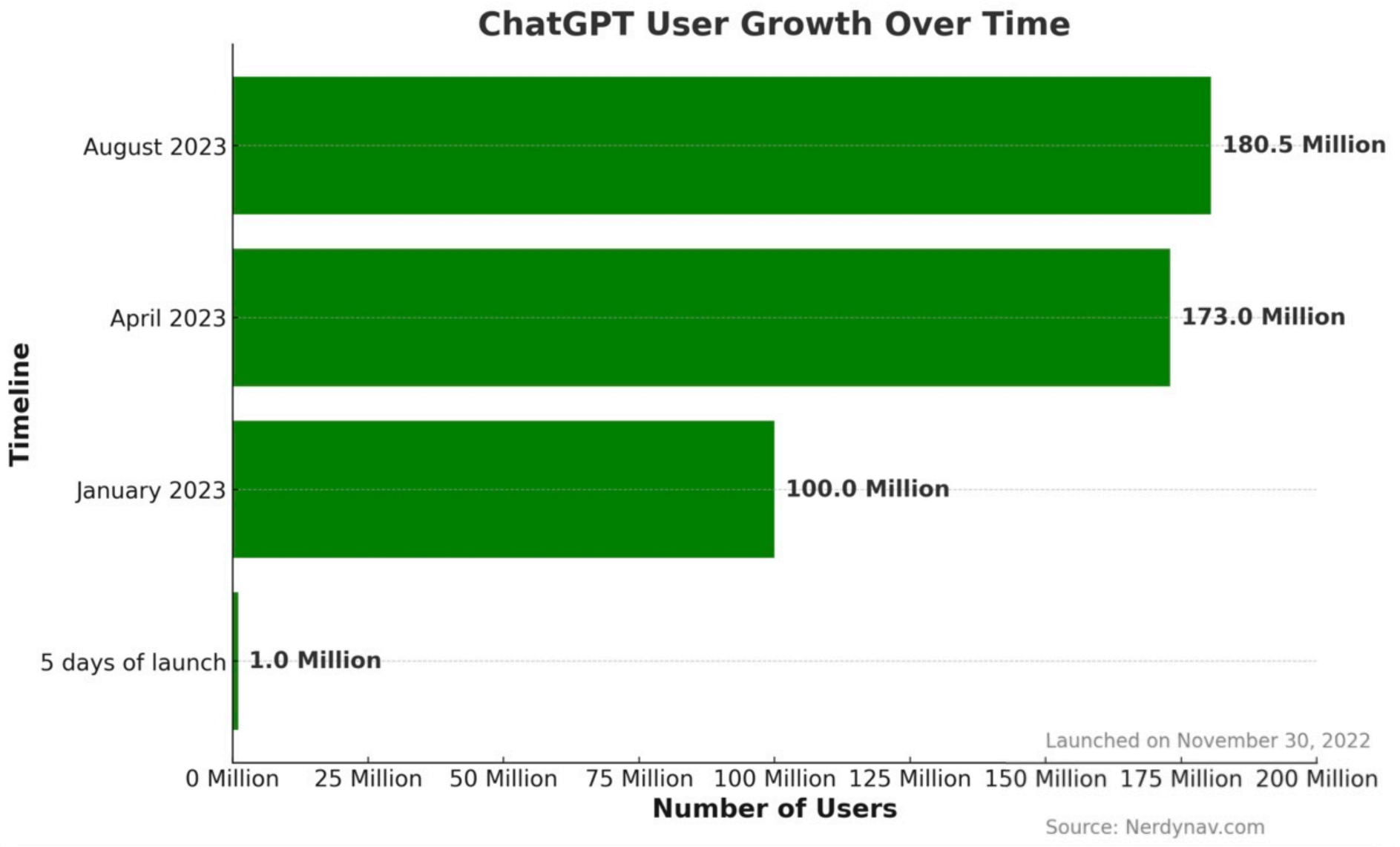
IA Generativa i Enginyeria del Software

La dimensió desconeguda

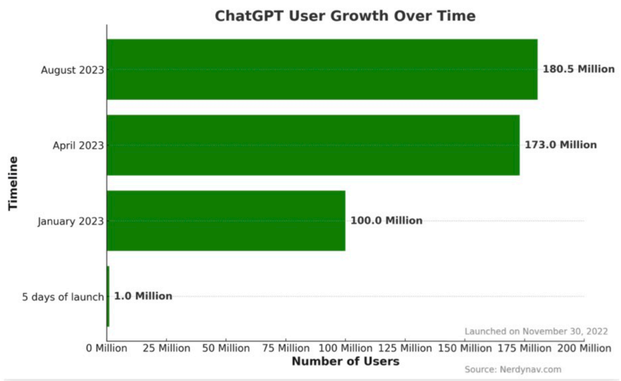
Marc Alier https://wasabi.essi.upc.edu/ludo

THE
TWILIGHT
ZONE


ChatGPT User Growth Over Time
Timeline
August 2023
April 2023
January 2023
5 days of launch
180.5 Million
173.0 Million
100.0 Million
1.0 Million
Launched on November 30, 2022
0 Million
25 Million
50 Million
75 Million
100 Million
125 Million
150 Million
175 Million
200 Million
Number of Users
Source: Nerdynav.com

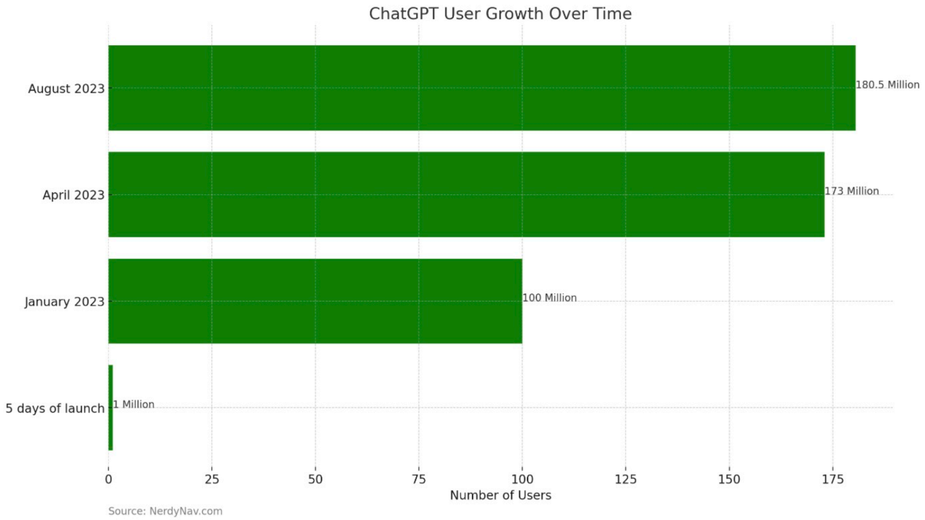
The
TWILIGHT
ZONE

You



create a python script that plots a graph like this one

ChatGPT



I've created a bar chart that resembles the one in the image you uploaded. You can download the chart from the link below:

Download the chart

If you need any changes or additional information on the graph, let me know!

You

hey dude ndo not be lazy ... I want the python script

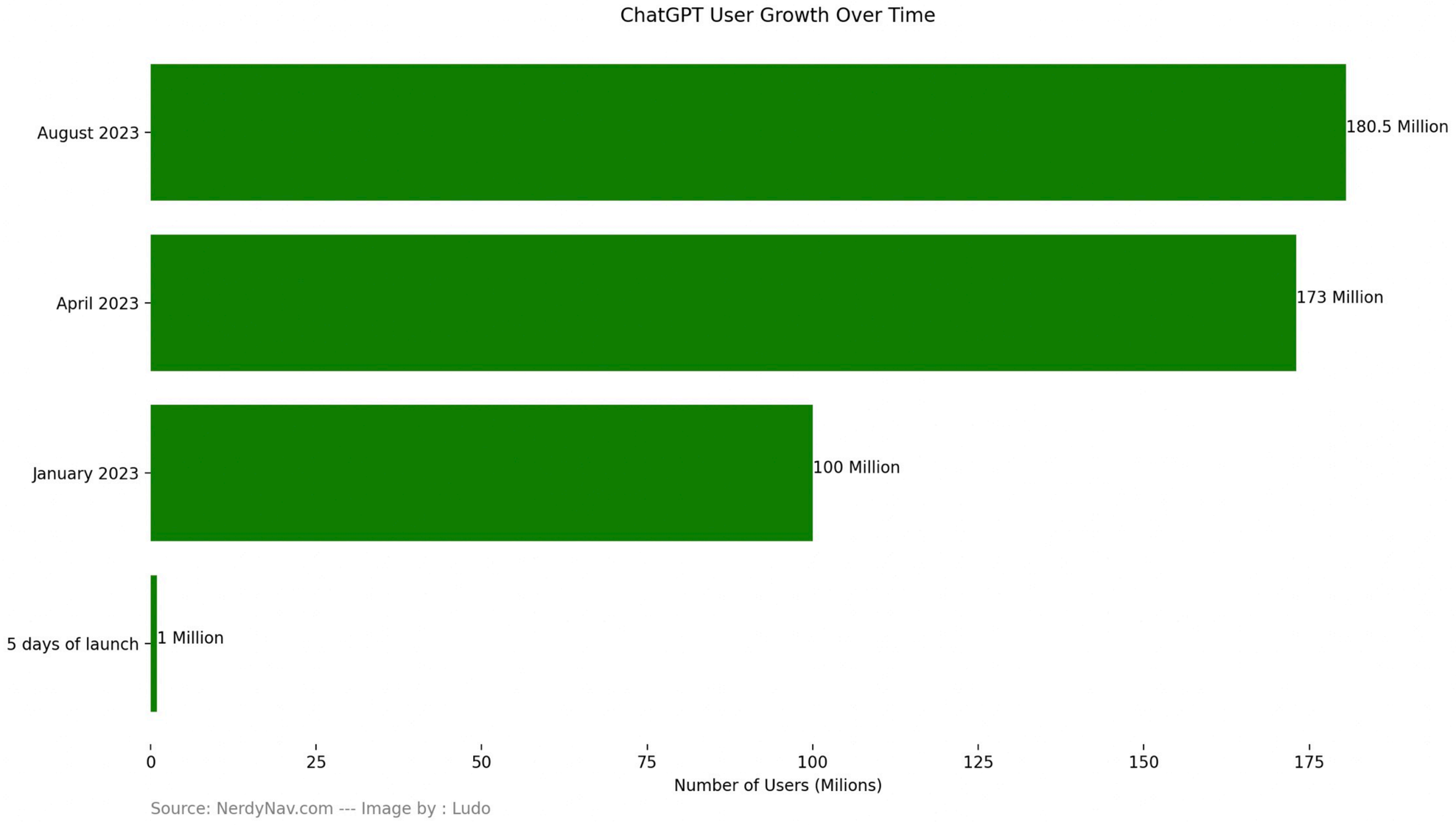
ChatGPT

Apologies for the oversight. Here's the Python script that generates a bar chart similar to the one you uploaded:

python Copy code

```python
import matplotlib.pyplot as plt

# Data for plotting
months = ['5 days of launch', 'January 2023', 'April 2023', 'August 2023']
users = [1, 100, 173, 180.5]

# Plotting the bar chart
plt.figure(figsize=(14, 8)) # Size to match the provided aspect
plt.barh(months, users, color='green')

# Adding the data labels
for index, value in enumerate(users):
    plt.text(value, index, f"{value} Million")

# Setting labels and title
plt.xlabel('Number of Users')
plt.title('ChatGPT User Growth Over Time')

# Source text
plt.text(0, -1, 'Source: NerdyNav.com', fontsize=10, color='gray')

# Remove the frame of the chart.
for spine in plt.gca().spines.values():
    spine.set_visible(False)

# Save the plot as a file
plt.savefig('/mnt/data/user_growth_chart.png', bbox_inches='tight', dpi=150)

# Show the plot
plt.show()
```

This script uses the `matplotlib` library to create a horizontal bar chart. It sets up the data, adds the bars to the plot, labels each bar with its corresponding user number, and adds titles and axes labels. The source text is added at the bottom, and the chart frame is removed for a cleaner look. The script saves the chart to a file and also ws it in an output window.


ChatGPT User Growth Over Time
August 2023
April 2023
January 2023
5 days of launch
180.5 Million
173 Million
100 Million
1 Million
0
25
50
75
100
125
150
175
Number of Users (Milions)
Source: NerdyNav.com --- Image by : Ludo

Juanan propietari

Un hilo para enmarcar y reflexionar (cómo no, de Simon Willison)

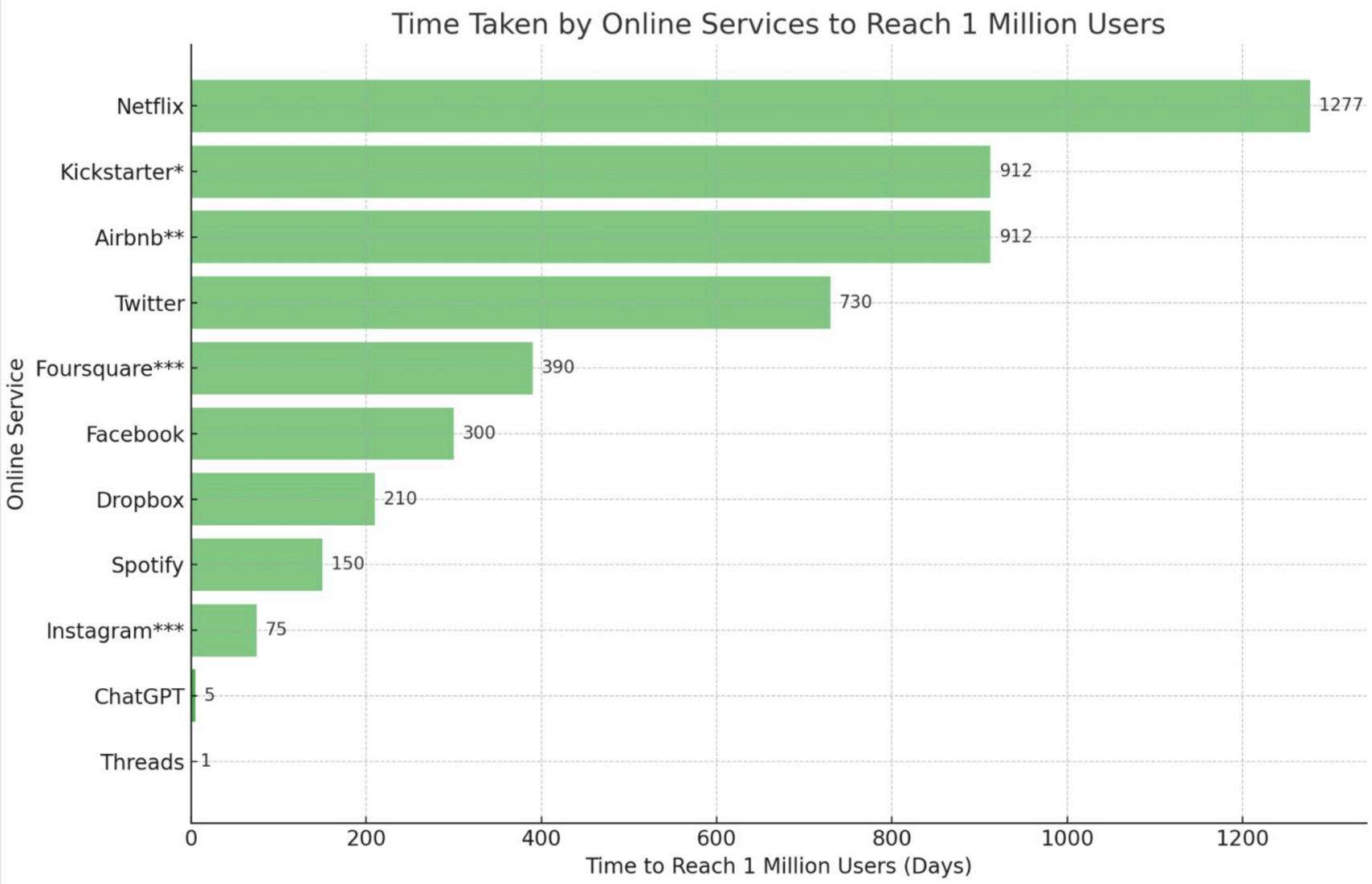
https://twitter.com/simonw/status/1768636336247107664

*** Hay que seguir aprendiendo a programar ahora que la IA es tan buena programando?**

- Los prompts que enviamos a la IA para que hagan lo que queremos tendrán más o menos éxito en base a lo que controlemos de programación e ingeniería software. Si sabes programar o tienes unos conceptos sólidos, tendrás superpoderes para manejar la IA.

- Pero ya no es necesario aprender los detalles. Ya no es necesario saber de memoria la sintaxis de un bucle for en diferentes lenguajes (lo que antes nos daba ventajas como desarrollador frente a otros ya no lo hará).

Lo importante ahora es aprender a leer y entender con fluidez código en diferentes lenguajes. Esa es y seguirá siendo una competencia clave durante mucho tiempo.


Time Taken by Online Services to Reach 1 Million Users
Online Service
Netflix
Kickstarter*
Airbnb**
Twitter
Foursquare***
Facebook
Dropbox
Spotify
Instagram***
ChatGPT
Threads
1277
912
912
730
390
300
210
150
75
5
1
0
200
400
600
800
1000
1200
Time to Reach 1 Million Users (Days)

ChatGPT
No està sol

THIS IS SPARTA!!

Spaces | lmsys / chatbot-arena-leaderboard | like 1.5k | Running | App | Files | Community 19

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **200,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **56**. Total #votes: **244024**. Last updated: Jan 26, 2024.

Contribute your vote at chat.lmsys.org! Find more analysis in the notebook.

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo	1249	+13/-13	30268	OpenAI	Proprietary	2023/4
2	Bard (Gemini Pro)	1215	+16/-15	3014	Google	Proprietary	Online
3	GPT-4-0314	1189	+14/-12	18062	OpenAI	Proprietary	2021/9
4	GPT-4-0613	1161	+13/-13	27441	OpenAI	Proprietary	2021/9
5	Mistral Medium	1150	+15/-15	11480	Mistral	Proprietary	Unknown
6	Claude-1	1150	+13/-13	17630	Anthropic	Proprietary	Unknown
7	Claude-2.0	1131	+14/-13	11698	Anthropic	Proprietary	Unknown
8	Mixtral-8x7b-Instruct-v0.1	1123	+14/-13	14165	Mistral	Apache 2.0	2023/12
9	Gemini Pro (Dev API)	1122	+15/-16	6332	Google	Proprietary	2023/4
10	Claude-2.1	1119	+14/-13	23323	Anthropic	Proprietary	Unknown
11	GPT-3.5-Turbo-0613	1117	+14/-13	30326	OpenAI	Proprietary	2021/9
12	Gemini Pro	1114	+16/-15	6981	Google	Proprietary	2023/4
13	Yi-34B-Chat	1112	+13/-15	5807	01 AI	Yi License	2023/6
14	Claude-Instant-1	1110	+13/-12	17623	Anthropic	Proprietary	Unknown
15	WizardLM-70B-v1.0	1106	+13/-13	8079	Microsoft	Llama 2 Community	2023/8
16	Tulu-2-DPO-70B	1105	+16/-15	5127	AllenAI/UW	AI2 ImpACT Low-risk	2023/11

Spaces | lmsys / chatbot-arena-leaderboard | like 1.5k | Running | App | Files | Community 19

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **200,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **56**. Total #votes: **244024**. Last updated: Jan 26, 2024.

Contribute your vote at chat.lmsys.org! Find more analysis in the notebook.

Rank	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4-Turbo	1249	+13/-13	30268	OpenAI	Proprietary	2023/4
2	Bard (Gemini Pro)	1215	+16/-15	3014	Google	Proprietary	Online
3	GPT-4-0314	1189	+14/-12	18062	OpenAI	Proprietary	2021/9
4	GPT-4-0613	1161	+13/-13	27441	OpenAI	Proprietary	2021/9
5	Mistral Medium	1150	+15/-15	11480	Mistral	Proprietary	Unknown
6	Claude-1	1150	+13/-13	17630	Anthropic	Proprietary	Unknown
7	Claude-2.0	1131	+14/-13	11698	Anthropic	Proprietary	Unknown
8	Mixtral-8x7b-Instruct-v0.1	1123	+14/-13	14165	Mistral	Apache 2.0	2023/12
9	Gemini Pro (Dev API)	1122	+15/-16	6332	Google	Proprietary	2023/4
10	Claude-2.1	1119	+14/-13	23323	Anthropic	Proprietary	Unknown
11	GPT-3.5-Turbo-0613	1117	+14/-13	30326	OpenAI	Proprietary	2021/9
12	Gemini Pro	1114	+16/-15	6981	Google	Proprietary	2023/4
13	Yi-34B-Chat	1112	+13/-15	5807	01 AI	Yi License	2023/6
14	Claude-Instant-1	1110	+13/-12	17623	Anthropic	Proprietary	Unknown
15	WizardLM-70B-v1.0	1106	+13/-13	8079	Microsoft	Llama 2 Community	2023/8
16	Tulu-2-DPO-70B	1105	+16/-15	5127	AllenAI/UW	AI2 ImpACT Low-risk	2023/11

🏆 LMSYS Chatbot Arena Leaderboard

| Vote | Blog | GitHub | Paper | Dataset | Twitter | Discord |

LMSYS Chatbot Arena is a crowdsourced open platform for LLM evals. We've collected over **200,000** human preference votes to rank LLMs with the Elo ranking system.

Arena Elo | Full Leaderboard

Total #models: **64**. Total #votes: **299816**. Last updated: Feb 15, 2024.

Contribute your vote 🗳️ at chat.lmsys.org! Find more analysis in the notebook.

Rank	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4-1106-preview	1254	+5/-5	38745	OpenAI	Proprietary	2023/4
2	GPT-4-0125-preview	1253	+10/-8	6308	OpenAI	Proprietary	2023/4
3	Bard (Gemini Pro)	1218	+8/-7	10313	Google	Proprietary	Online
4	GPT-4-0314	1191	+6/-6	20430	OpenAI	Proprietary	2021/9
5	GPT-4-0613	1164	+5/-6	32941	OpenAI	Proprietary	2021/9
6	Mistral Medium	1152	+5/-7	17847	Mistral	Proprietary	Unknown
7	Claude-1	1150	+7/-5	19017	Anthropic	Proprietary	Unknown
8	Qwen1.5-72B-Chat	1147	+8/-8	5204	Alibaba	Qianwen LICENSE	2024/2
9	Claude-2.0	1132	+6/-8	12753	Anthropic	Proprietary	Unknown
10	Gemini Pro (Dev API)	1122	+7/-7	9024	Google	Proprietary	2023/4
11	Claude-2.1	1120	+6/-4	27723	Anthropic	Proprietary	Unknown
12	Mixtral-8x7b-Instruct-v0.1	1120	+5/-6	18410	Mistral	Apache 2.0	2023/12
13	GPT-3.5-Turbo-0613	1118	+5/-5	36704	OpenAI	Proprietary	2021/9
14	Gemini Pro	1115	+9/-9	6958	Google	Proprietary	2023/4
15	Yi-34B-Chat	1111	+7/-8	7734	01 AI	Yi License	2023/6
16	Claude-Instant-1	1109	+7/-5	20090	Anthropic	Proprietary	Unknown

New Chat

Modelfiles

Prompts

Documents

Search

"Mixtral: What Is It

New Chat

"Computing History 1950-1

Greetings in Spanish! (Note:

"Welcome Phrase for Chat"

Marc

"Mixtral: What Is It?" I understand that you'd like a concise phrase with a strict 3-5...

mixtral:latest (24.6GB)

Set as default

You

que es mixtral?

Ollama Mixtral:Latest 30/01/2024 16:01

Mixtral is not a commonly used word in English, and I'm assuming you are asking about its meaning. However, I was unable to find any reference to "mixtral" as a standalone word in any major English language dictionaries or linguistic resources.

It's possible that there might be a typo or error in the spelling, and you meant to ask about something else. If you could provide more context or clarify your question, I would be happy to try and help you further.

You

and mistral.ai?

Ollama Mixtral:Latest

Mistral AI is a cutting-edge company based in Paris, France, developing large language models. The company was founded in 2021 by several leading experts in artificial intelligence, including Criteo's former research director, Iadine Chades, and Mozilla's former head of AI, Pierre-Yves Oudeyer.

Mistral AI's mission is to create large language models that can understand and generate human-like text while ensuring alignment with human values and ethical considerations. The company has gained significant attention in the tech industry for its innovative approach to developing large language models, as well as its commitment to responsible AI development.

Mistral AI's technology has numerous applications across various industries, including customer service, content creation, translation, and more. By building advanced language models that can understand and generate human-like text, Mistral AI aims to unlock new possibilities for natural language processing and revolutionize how humans interact with artificial intelligence systems.

Send a message

LLMs can make mistakes. Verify important information.

ollama.ai

Blog Discord GitHub

Models Sign in Download

Get up and running with large language models, locally.

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

Download ↓

Available for macOS & Linux
Windows coming soon

Models

Find a model

llama2

The most popular model for general use.

273.6K Pulls 102 Tags Updated 4 weeks ago

mistral

The 7B model released by Mistral AI, updated to version 0.2.

174.0K Pulls 53 Tags Updated 4 weeks ago

llava

🌋 A novel end-to-end trained large multimodal model that combines a vision encoder and Vicuna for general-purpose visual and language understanding.

18.8K Pulls 33 Tags Updated 4 weeks ago

mixtral

A high-quality Mixture of Experts (MoE) model with open weights by Mistral AI.

40.0K Pulls 34 Tags Updated 4 weeks ago

starling-lm

Starling is a large language model trained by reinforcement learning from AI feedback focused on improving chatbot helpfulness.

6,868 Pulls 18 Tags Updated 2 months ago

neural-chat

A fine-tuned model based on Mistral with good coverage of domain and language.

8,584 Pulls 50 Tags Updated 5 weeks ago

codellama

A large language model that can use text prompts to generate and discuss code.

117.1K Pulls 199 Tags Updated 9 hours ago

dolphin-mixtral

An uncensored, fine-tuned model based on the Mixtral mixture of experts model that excels at coding tasks. Created by Eric Hartford.

97.7K Pulls 70 Tags Updated 4 weeks ago

mistral-openorca

Mistral OpenOrca is a 7 billion parameter model, fine-tuned on top of the Mistral 7B model using the OpenOrca dataset.

88.5K Pulls 17 Tags Updated 3 months ago

llama2-uncensored

Uncensored Llama 2 model by George Sung and Jarrad Hope.

56.3K Pulls 34 Tags Updated 2 months ago

orca-mini

A general-purpose model ranging from 3 billion parameters to 70 billion, suitable for entry-level hardware.

41.5K Pulls 119 Tags Updated 2 months ago

vicuna

General use chat model based on Llama and Llama 2 with 2K to 16K context sizes.

25.4K Pulls 111 Tags Updated 2 months ago

deepseek-coder

DeepSeek Coder is a capable coding model trained on two trillion code and natural language tokens.

22.8K Pulls 102 Tags Updated 4 weeks ago

wizard-vicuna-uncensored

Wizard Vicuna Uncensored is a 7B, 13B, and 30B parameter model based on Llama 2 uncensored by Eric Hartford.

22.6K Pulls 49 Tags Updated 2 months ago

Més enllà de ChatGPT en IA Generativa

- ChatGPT: Una part petita de la IA Generativa.
- Utilitza models transformadors i tècniques de difusió.
- Basat en LLM GPT-3 i GPT-4.
- Arquitectura transformador: Clau en la recerca d'IA.

PUBLICATIONS ›

Attention is All You Need

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

NIPS (2017)

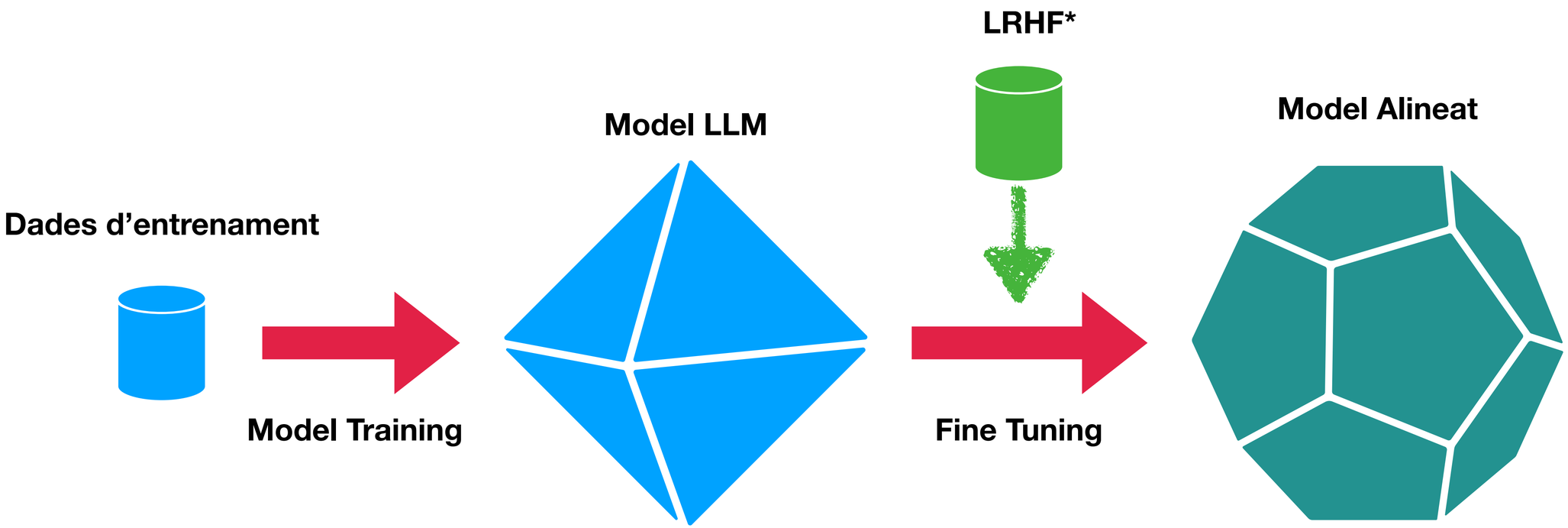
Agency [edit]

An LLM is a language model, which is not an agent as it has no goal, but it can be used as a component of an intelligent agent.

The ReAct ("Reason+Act") method constructs an agent out of an LLM, using the LLM as a planner. The LLM is prompted to "think out loud". Specifically, the language model is prompted with a textual description of the environment, a goal, a list of possible actions, and a record of the actions and observations so far. It generates one or more thoughts before generating an action, which is then executed in the environment.[48] The linguistic description of the environment given to the LLM planner can even be the LaTeX code of a paper describing the environment.[49]

The Reflexion method[50] constructs an agent that learns over multiple episodes. At the end of each episode, the LLM is given the record of the episode, and prompted to think up "lessons learned", which would help it perform better at a subsequent episode. These "lessons learned" are given to the agent in the subsequent episodes.

Wikipedia:LLM



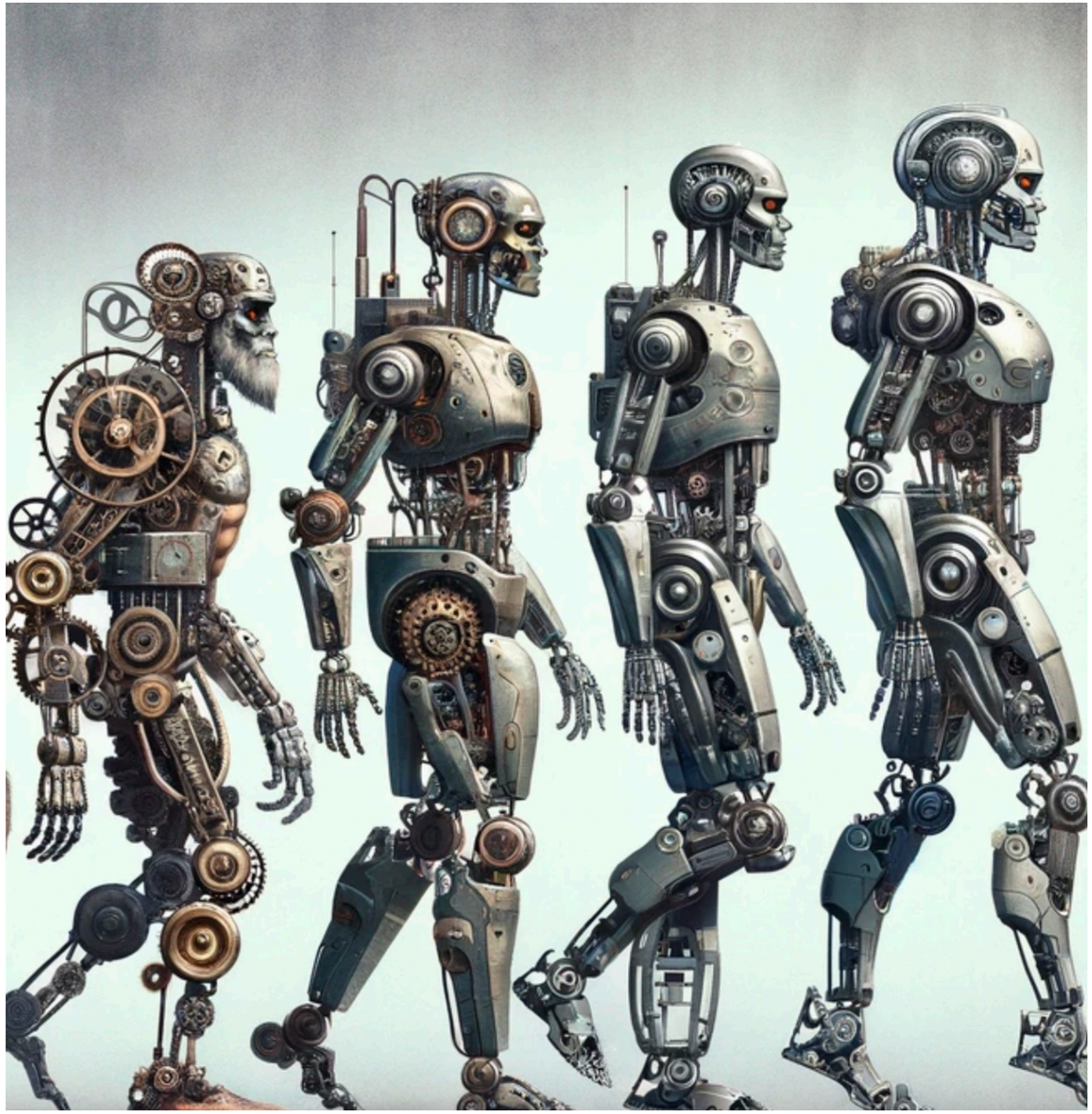
Learning Reinforcement through Human Feedback

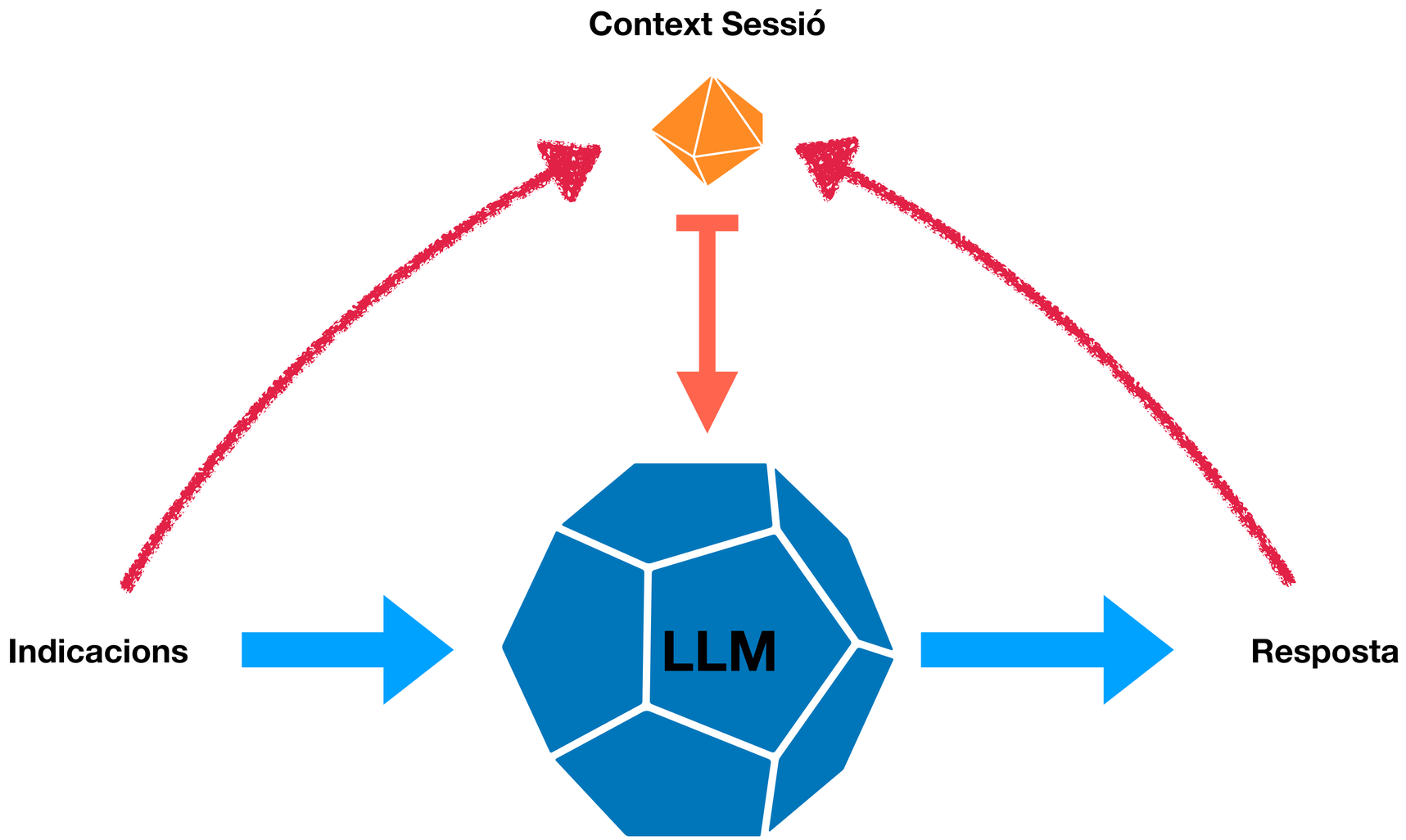
https://openai.com/research/learning-from-human-preferences

La quantitat té una qualitat en si mateixa

A mesura que es fan grans apareixen qualitats emergents

- Capacitat d'aprenentatge ràpid amb pocs exemples.
- Raonament en cadena de pensament en tasques complexes.
- Habilitats no anticipades en programació, aritmètica i correcció de conceptes erronis.
- Evidència de representacions internes per a raonament abstracte més enllà de l'estructura lingüística.
- Exemples:
  - Representació interna de colors alineada amb la percepció humana.
  - Diferenciació entre conceptes erronis i fets, i proves de raonament de sentit comú.




Context Sessió
Indicacions
LLM
Resposta

explica que es un haiku

Un haiku és una forma de poesia breu que es va originar al Japó. Tradicionalment, un haiku està format per tres línies i un total de 17 síl·labes, distribuïdes de la següent manera:

- La primera línia conté 5 síl·labes.
- La segona línia conté 7 síl·labes.
- La tercera línia conté 5 síl·labes.

Els haikus solen centrar-se en temes relacionats amb la natura, les estacions, la bellesa, la tranquil·litat i els moments efímers de la vida. També és habitual que presentin una espècie de "salt" o "gir" en el pensament, conegut com a "kireji", que serveix per aprofundir la comprensió del lector o evocar emocions.

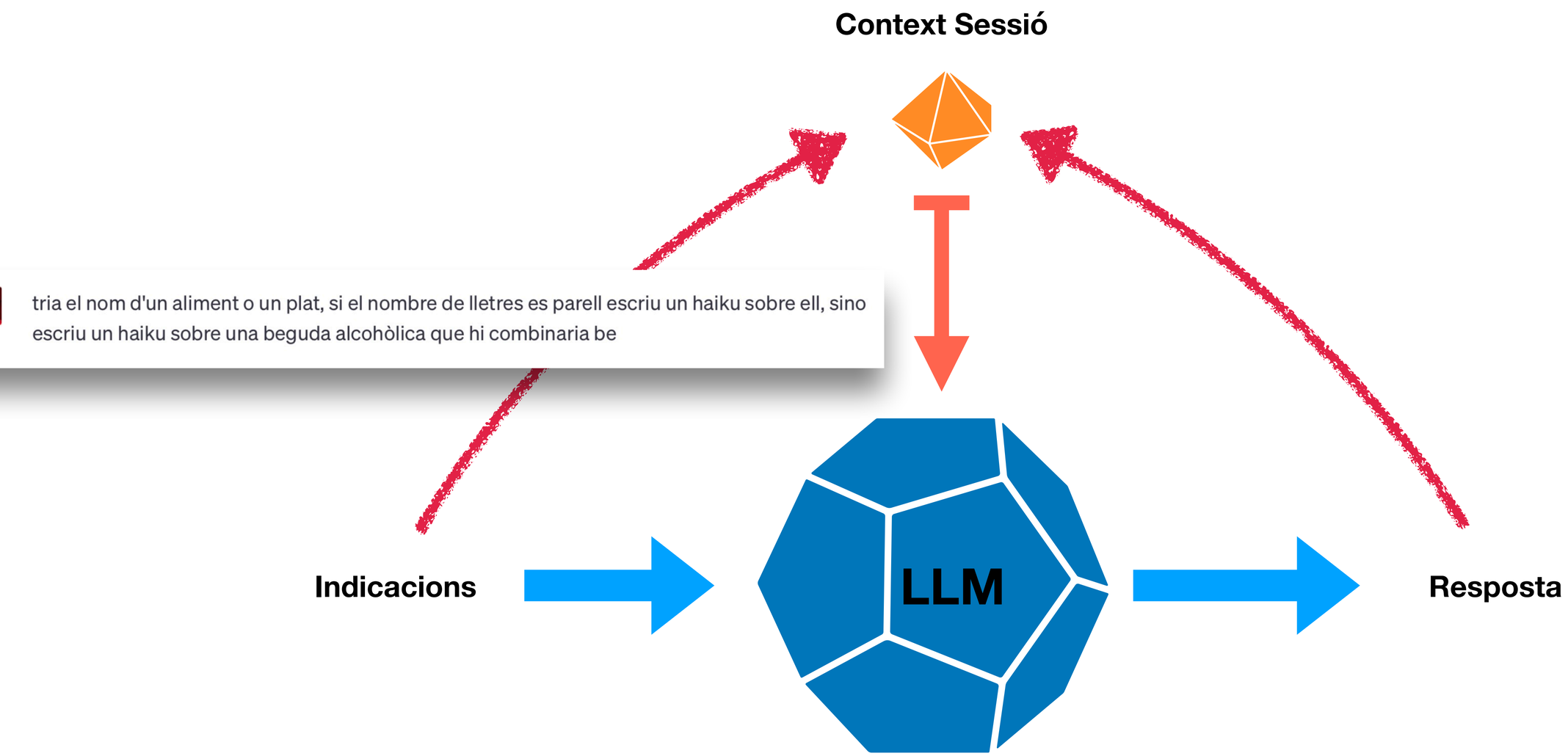
Però és important assenyalar que la forma tradicional del haiku s'ha adaptat i variat en diverses cultures i llengües, i per tant, es poden trobar haikus que no segueixen estrictament la regla de les 17 síl·labes.

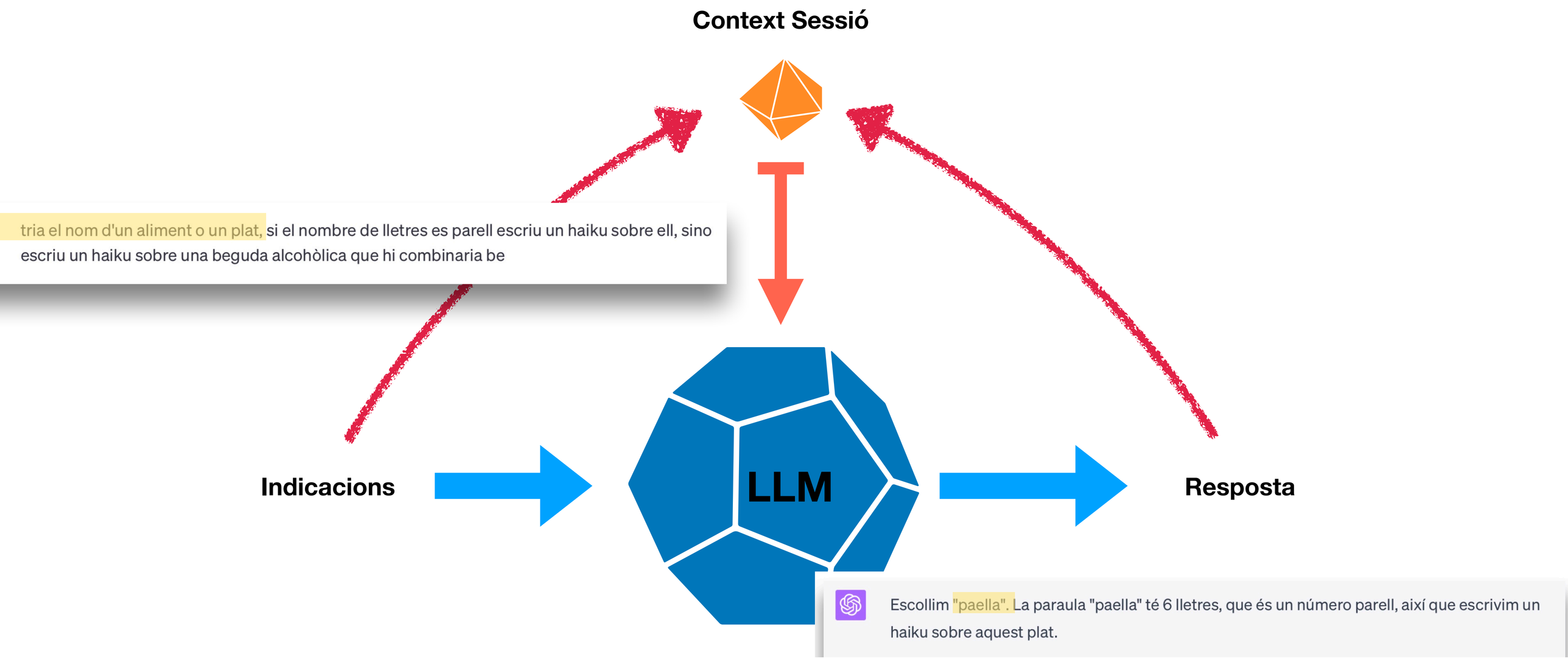
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be

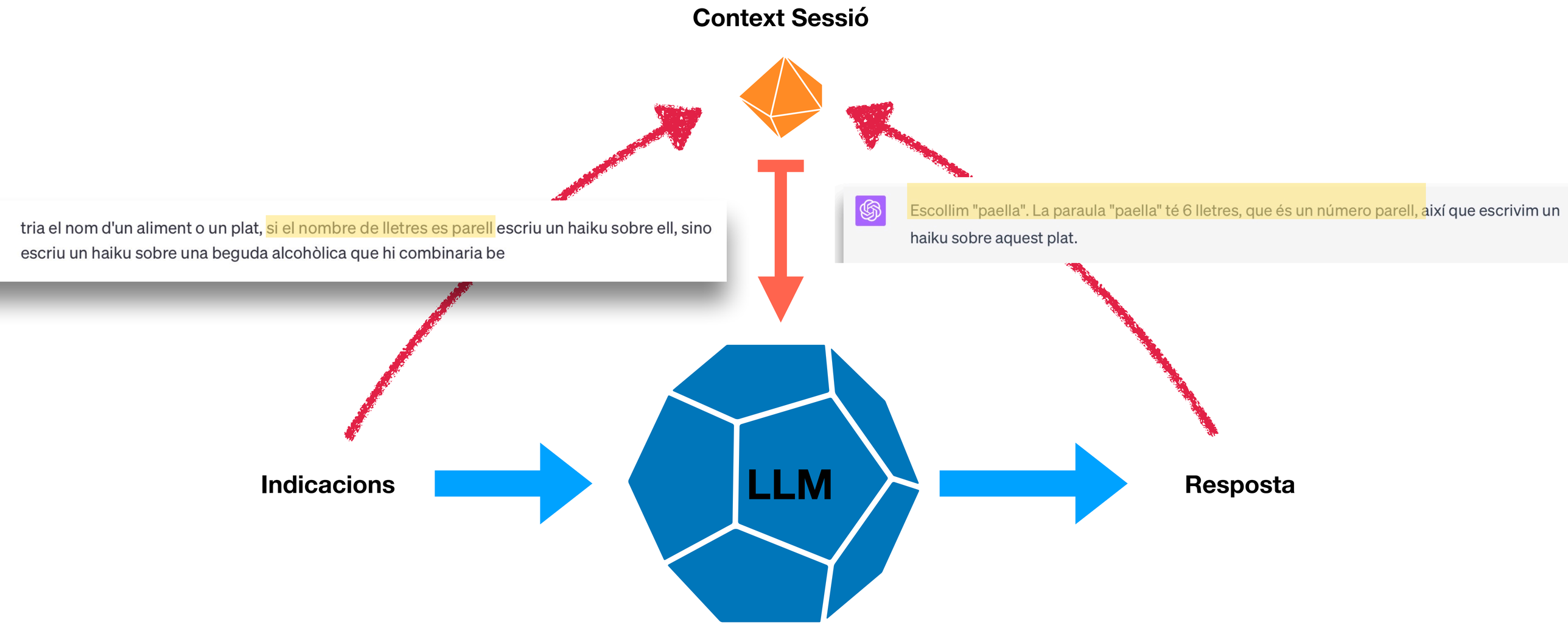
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be

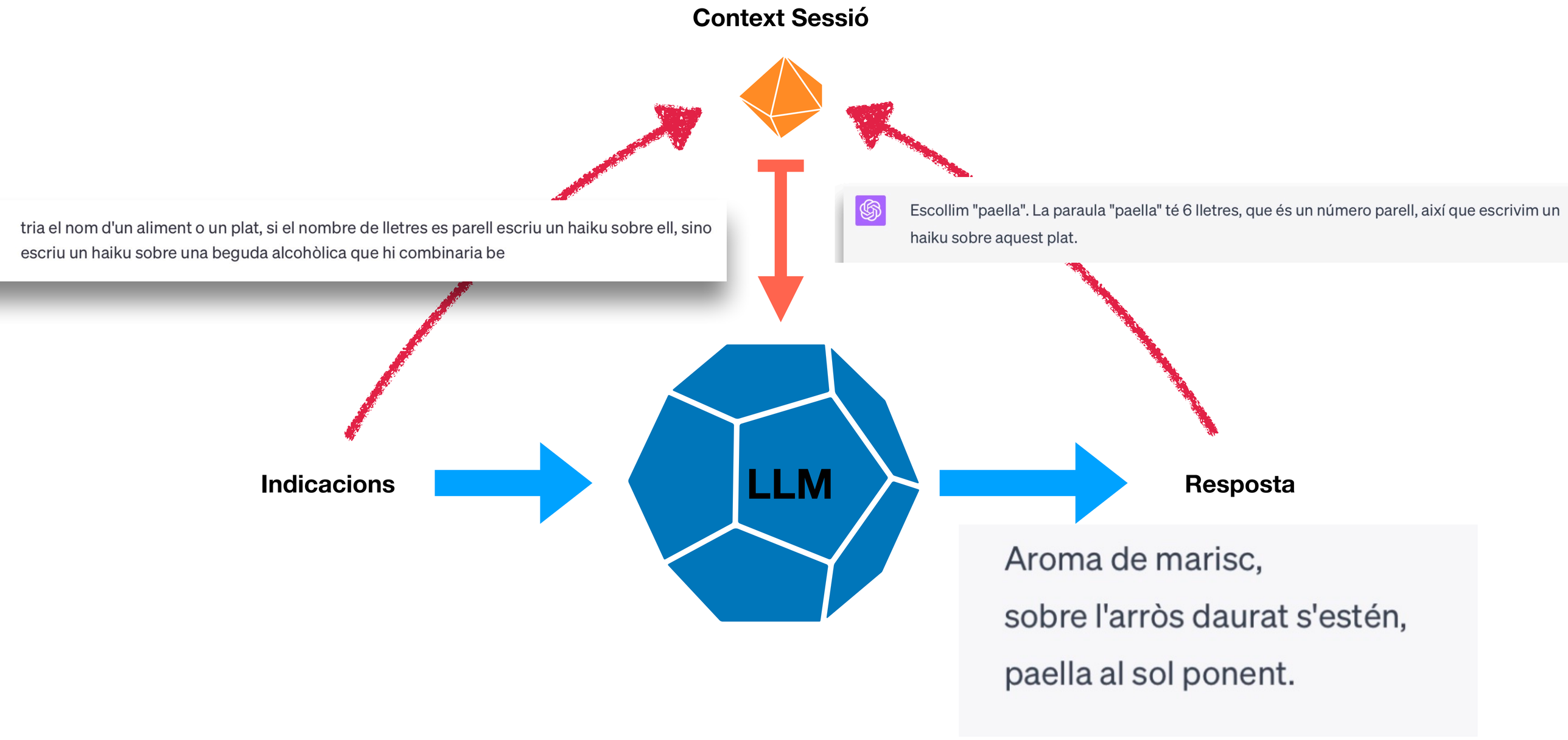
Escollim "paella". La paraula "paella" té 6 lletres, que és un número parell, així que escrivim un haiku sobre aquest plat.

Aroma de marisc,
sobre l'arròs daurat s'estén,
paella al sol ponent.


Context Sessió
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be
Indicacions
LLM
Resposta


Context Sessió
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be
Indicacions
LLM
Resposta
Escollim "paella". La paraula "paella" té 6 lletres, que és un número parell, així que escrivim un haiku sobre aquest plat.


Context Sessió
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be
Escollim "paella". La paraula "paella" té 6 lletres, que és un número parell, així que escrivim un haiku sobre aquest plat.
Indicacions
LLM
Resposta


Context Sessió
tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be
Escollim "paella". La paraula "paella" té 6 lletres, que és un número parell, així que escrivim un haiku sobre aquest plat.
Indicacions
LLM
Resposta
Aroma de marisc,
sobre l'arròs daurat s'estén,
paella al sol ponent.

tria el nom d'un aliment o un plat, si el nombre de lletres es parell escriu un haiku sobre ell, sino escriu un haiku sobre una beguda alcohòlica que hi combinaria be

Escollim "sushi". La paraula "sushi" té 5 lletres, que és un número senar, així que escrivim un haiku sobre una beguda alcohòlica que hi combinaria bé, el "sake".

Sake suau i net,
sushi fresc, màgia al paladar,
batega el cor d'Orient.

te voy a dar un texto entre triples tildes, trasnformalo en una lista de instrucciones del estilo :
Paso 1:
Paso 2:
...
Paso N:

Si no te doy ninguna lista de instrucciones coherente, simplemente di:
No hace falta hacer nada:
´´´Con 10 cañones por banda viento en popa a toda vela..
´´´

No hace falta hacer nada.

te voy a dar un texto entre triples tildes, trasnformalo en una lista de instrucciones del estilo :
Paso 1:
Paso 2:
...
Paso N:

Si no te doy ninguna lista de instrucciones coherente, simplemente di:
No hace falta hacer nada:
´´´Colocamos el choco en un colador para que pierda el líquido. Lo secamos con papel de cocina y lo metemos en una bolsa junto con la harina. Sacudimos la bolsa para que la harina se adhiera al pescado.Se puede usar solo harina especial para freír, pero le da un punto más crujiente añadir un tercio de harina de sémola de trigo duro.
Con las manos vamos sacudiendo para eliminar el exceso de harina y ponemos inmediatamente en el aceite caliente.
Freímos hasta que doren por fuera y servimos al momento. Podemos rociar los chocos fritos con unas gotas de limón.
´´´

te voy a dar un texto entre triples tildes, trasnformalo en una lista de instrucciones del estilo :
Paso 1:
Paso 2:
...
Paso N:

Si no te d
No hace f
´´´Colo
de cocina
harina se
punto má
Con las m
inmediat
Freímos h
fritos con
´´´

Aquí está el texto transformado en una lista de instrucciones:

Paso 1: Colocamos el choco en un colador para que pierda el líquido.
Paso 2: Lo secamos con papel de cocina y lo metemos en una bolsa junto con la harina.
Paso 3: Sacudimos la bolsa para que la harina se adhiera al pescado.
Paso 4: Se puede usar solo harina especial para freír, pero le da un punto más crujiente añadir un tercio de harina de sémola de trigo duro.
Paso 5: Con las manos vamos sacudiendo para eliminar el exceso de harina y ponemos inmediatamente en el aceite caliente.
Paso 6: Freímos hasta que doren por fuera.
Paso 7: Servimos al momento.
Paso 8: Opcionalmente, podemos rociar los chocos fritos con unas gotas de limón.

Espero que esta lista de instrucciones te sea útil. ¡Disfruta de tus chocos fritos!

Trucs d'enginyeria de prompts

- Ser especific
- Indentificar que es cada part del prompt
- Especificar o donar un exemple del format de sortida
- Deixa que el model "pensi"

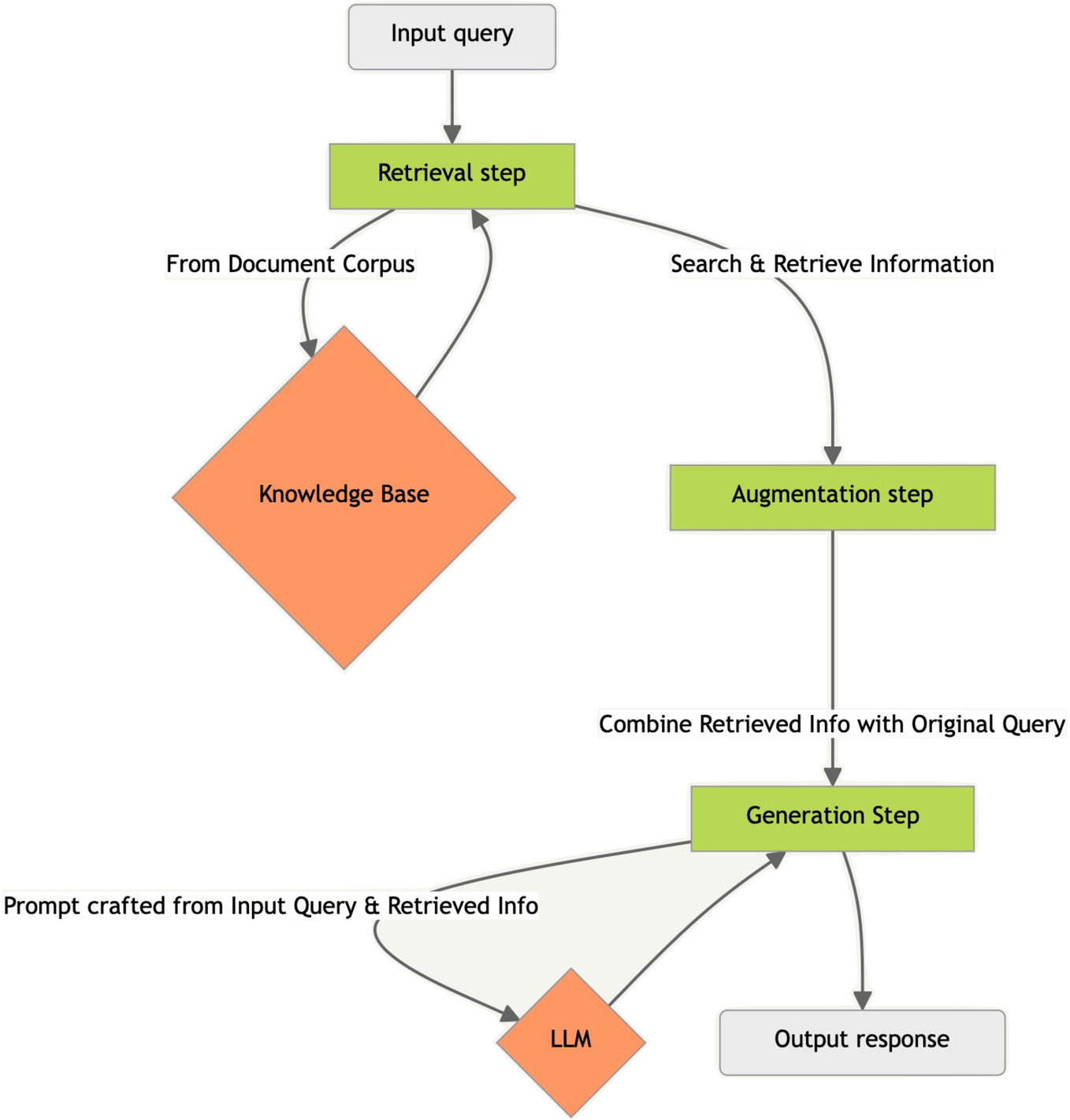
Demo - Jupyter Notebook

Superpoders de la IA Generativa

Retrieval Augmented Generation
Embeddings
Vector Databases
Huge context windows

Mermaid Split Copy Caption

```mermaid
graph TD
    classDef coloredNode fill:#f96;
    classDef rectangleNode fill:#bada55;
    A(Input query)
    A --> B[Retrieval step]
    B -->|From Document Corpus| B1{Knowledge Base}
    B1 --> B
    B -->|Search & Retrieve Information| D[Augmentation step]
    D -->|Combine Retrieved Info with Original Query| F[Generation Step]
    F -->|Prompt crafted from Input Query & Retrieved Info| G{LLM}
    G --> F
    F --> I(Output response)
    G:::coloredNode
    B1:::coloredNode
    B:::rectangleNode
    D:::rectangleNode
    F:::rectangleNode
```


Input query
Retrieval step
From Document Corpus
Search & Retrieve Information
Knowledge Base
Augmentation step
Combine Retrieved Info with Original Query
Generation Step
Prompt crafted from Input Query & Retrieved Info
LLM
Output response

Embeddings

https://simonwillison.net/2023/Oct/23/embeddings/

Storing and serving related documents with openai-to-sqlite and embeddings

I decide to upgrade the related articles feature on my TILs site. Previously I calculated these using full-text search, but I wanted to try out a new trick using OpenAI embeddings for document similarity instead.

My openai-to-sqlite CLI tool already provides a mechanism for calculating embeddings against text and storing them in a SQLite database.

I was going to add a command for calculating similarity based on those embeddings... and then I saw that Benoit Delbosc had opened a pull request implementing that feature already!

I took Benoit's work and expanded it. In particular, I added an option for saving the resulting calculations to a database table.

This meant I could find and then save related articles for my TILs by running the following:

```
wget https://s3.amazonaws.com/til.simonwillison.net/tils.db
```

This grabs the latest `tils.db` used to serve my TIL website.

```
openai-to-sqlite embeddings tils.db \
  --sql 'select path, title, topic, body from til'
```

This retrieves and stores embeddings from the OpenAI API for every row in my til table - embedding the `title`, `topic` and `body` columns concatenated together, then keying them against the `path` column (the primary key for that table).

The command output this:

```
Fetching embeddings  [####################################]  100%
Total tokens used: 402500
```

402,500 tokens at $0.0001 / 1K tokens comes to $0.04 - 4 cents!

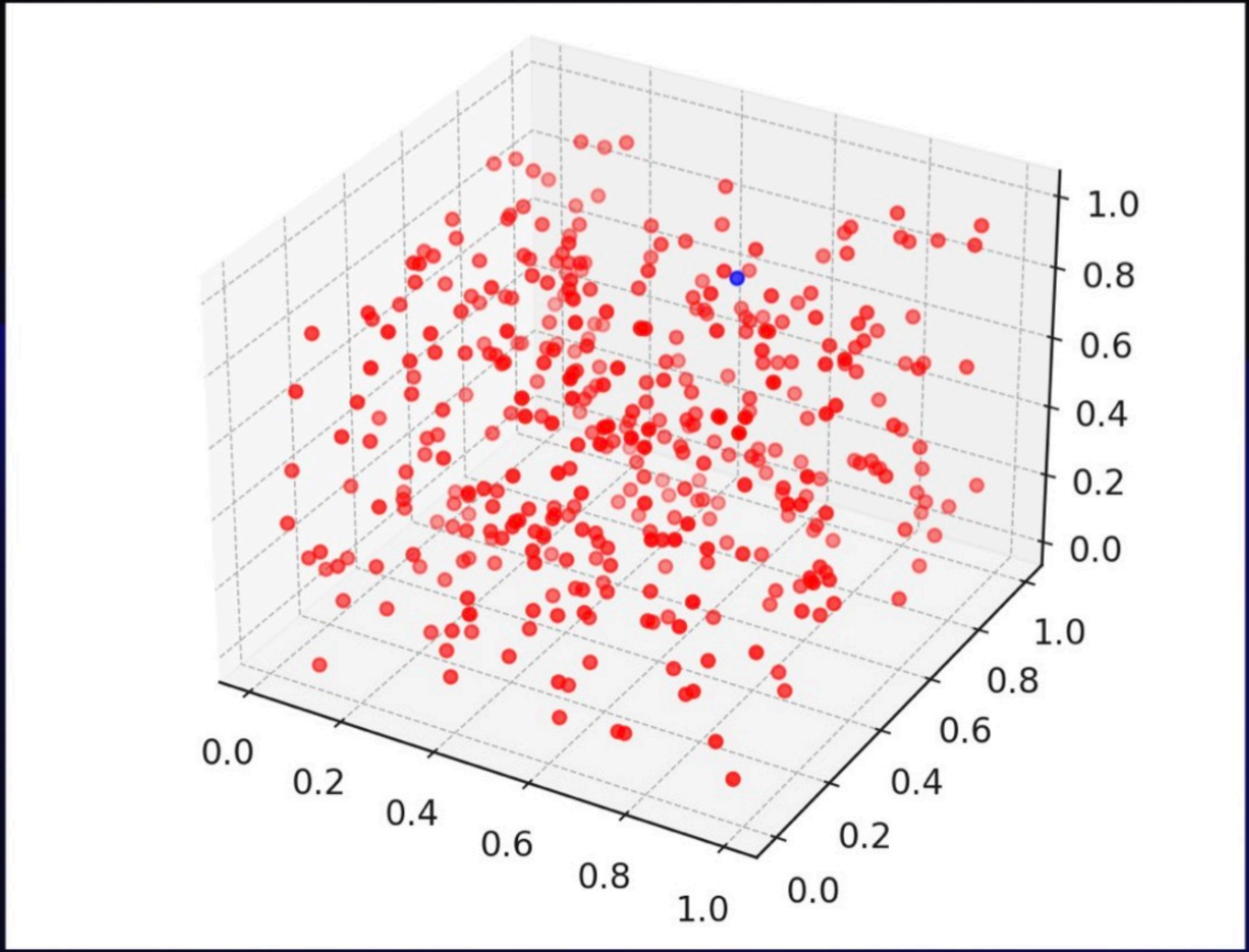
Now that I've embedded everything, I can search for the most similar articles to a particular article like this:

[
0.00171270314604043 96,
-0.00497547071427106 9,
0.0107359681278467 18,
-0.00793749280273914 3,
-0.0177794024348258 97,
0.0166018623858690 26,
-0.0177794024348258 97,
...
Array of floating point numbers
]

Fixed size: 300, 1000, 1536…

```
$ curl https://api.openai.com/v1/embeddings \
  -H "Content-Type: application/json" \
  -H "Authorization: Bearer $OPENAI_API_KEY" \
  -d '{
    "input": "What is shot-scraper?",
    "model": "text-embedding-ada-002"
  }'
{
  "object": "list",
  "data": [
    {
      "object": "embedding",
      "index": 0,
      "embedding": [
        -0.031646647,
        -0.0015310864,
        0.015132851,
        -0.009191913
```

A location in many-multi-dimensional space



word	similarity
france	1
french	0.7000748343471224
belgium	0.6933180492111168
paris	0.6334910653433325
germany	0.627075617939471
italy	0.6135215284228007
spain	0.6064218103692152

the AI-native open-source embedding database

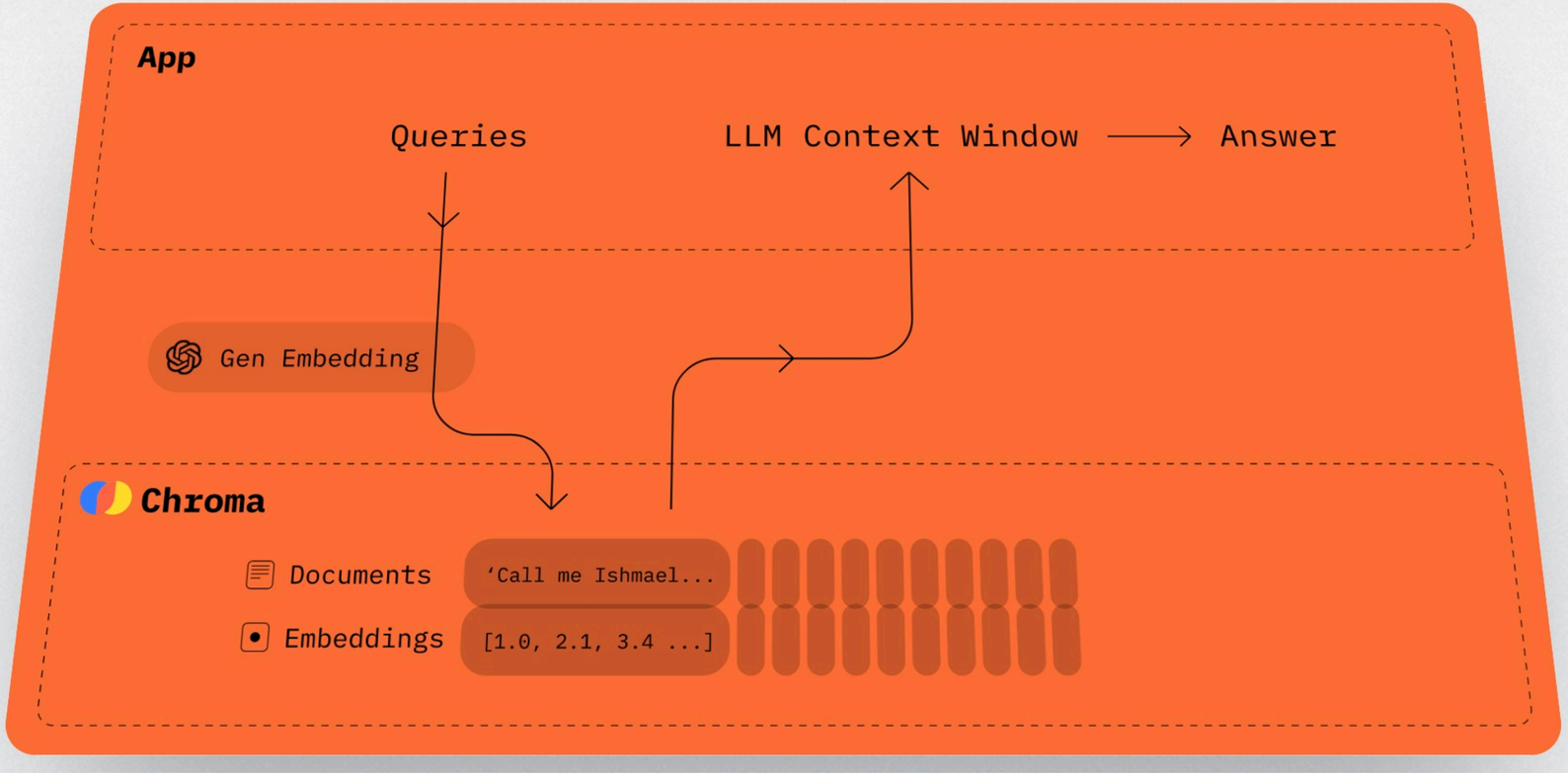
Getting Started

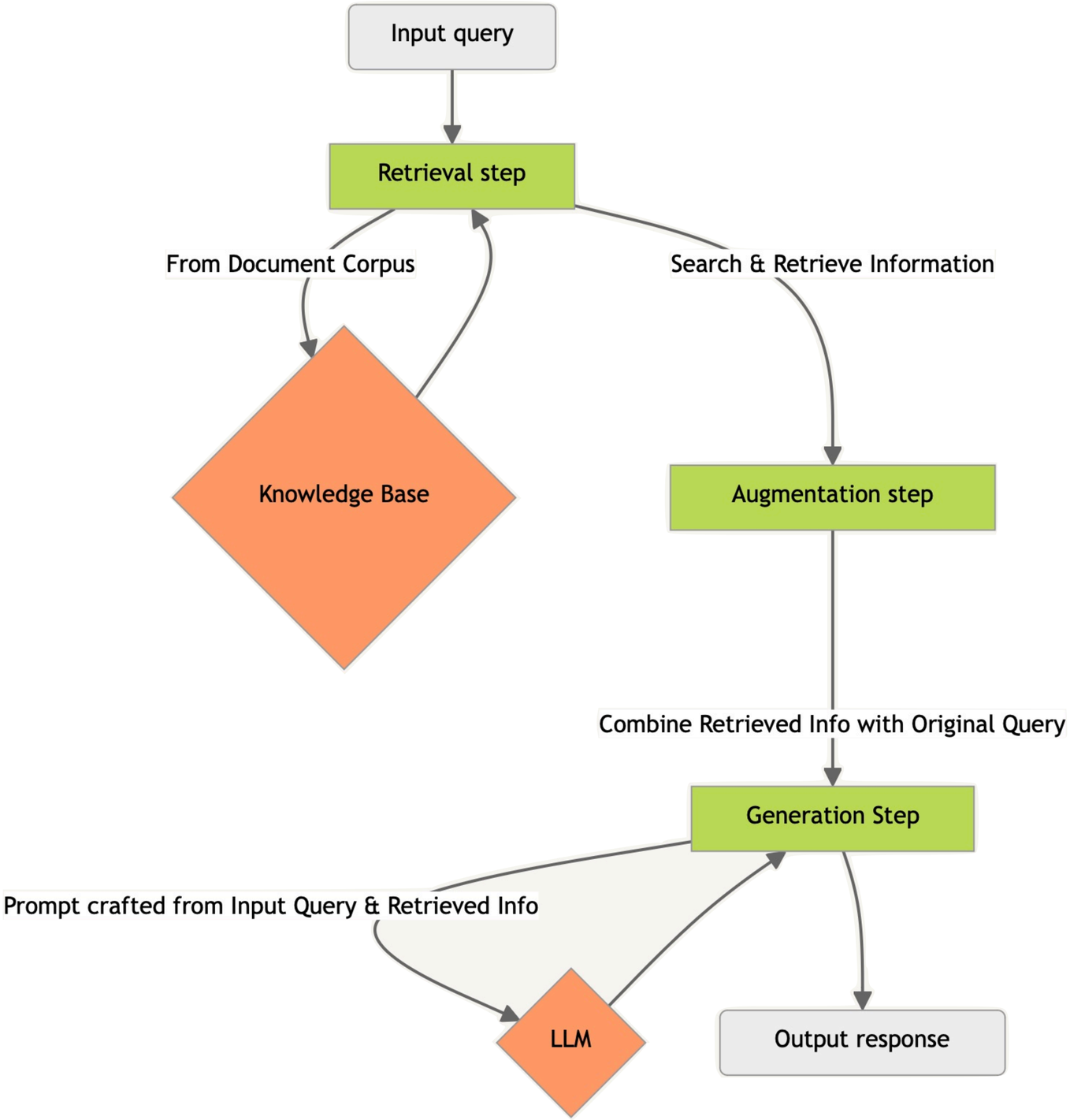
Colab Demo

chat 1008 online

Star 12k

Sign up for Chroma Cloud




Input query
Retrieval step
From Document Corpus
Search & Retrieve Information
Knowledge Base
Augmentation step
Combine Retrieved Info with Original Query
Generation Step
Prompt crafted from Input Query & Retrieved Info
LLM
Output response

Demos

https://semantic.ikasten.io/