LLM and ChatGPT: how it works

Dra. Mª Dolores Rodríguez Moreno

Universidad
de Alcalá

ISG

Objectives

Specific Objectives

- Understand how LLM works

Source

- Andreas Stöffelbauer (2023): How Large Language Models Work
- Open AI. GPT-4 (2023)
- Manish Shivanandhan (2023). Understanding Word Embeddings: The Building Blocks of NLP and GPTs.
- Vaswani *et al.* (2017). Attention Is All You Need
- Minaee et al. (2024). Large Language Models: A Survey

Outline

- **Artificial Intelligence**
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Prompting
- Hallucinating
- Comparing LLM
- Conclusions

Artificial Intelligence (I)

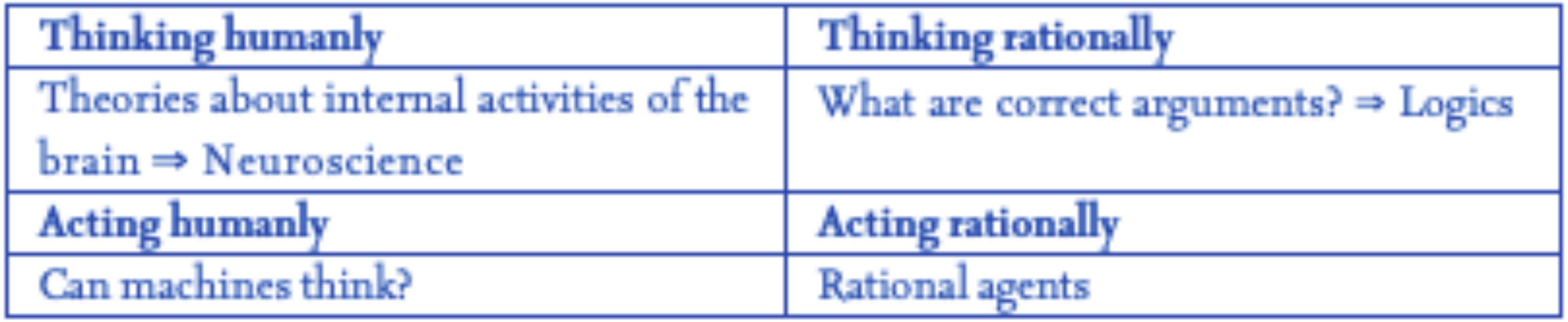
Definition of AI

Build machines that perform tasks that were previously performed by human beings

- Alternative definition: Understand and build intelligent entities
  - Understand: use computers to study intelligence (Science)
  - Build: solve real problems using knowledge and reasoning (Engineering)
  - Intelligent entity = agent
- AI deals with algorithms and knowledge representation
- AI is not restricted to any programming language

Artificial Intelligence (II)

- Two goals: humanity and rationality
  - Human: like human beings
  - Rational: doing the right thing, i.e. is expected to maximize goal achievement, given the available information
- Two dimensions: processes (thinking) and result (acting)



Thinking humanly	**Thinking rationally**
Theories about internal activities of the brain ⇒ Neuroscience	What are correct arguments? ⇒ Logics
Acting humanly	**Acting rationally**
Can machines think?	Rational agents

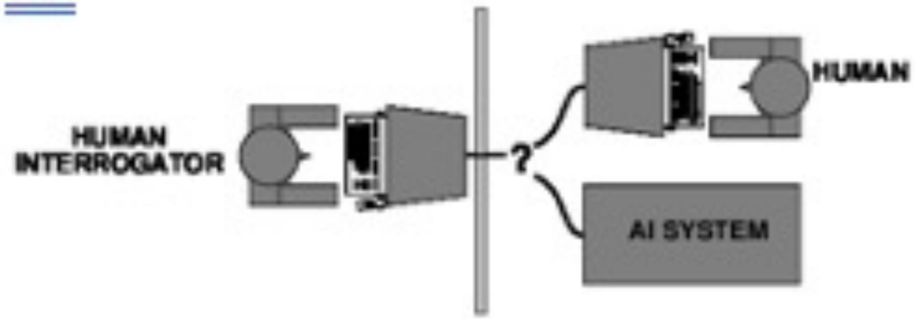
Artificial Intelligence (III)

Thinking humanly

- Scientific theory of internal activities of the brain
- How to validate?
    - Predicting behavior of humans (**Cognitive science**)
    - Identification of neurological data (**Neuroscience**)

Acting humanly

Can machines think? Test needed: Turing test



Proposed by Alan Turing (yes, Enigma)

Artificial Intelligence (IV)

Thinking rationally

- “Laws of thought”
- Aristotle: What are correct arguments? ⇒ **Logic**
- Connects Philosophy, Mathematics and AI
- Problems
  - Not all intelligent behavior is deliberative
  - What is the purpose of thinking?

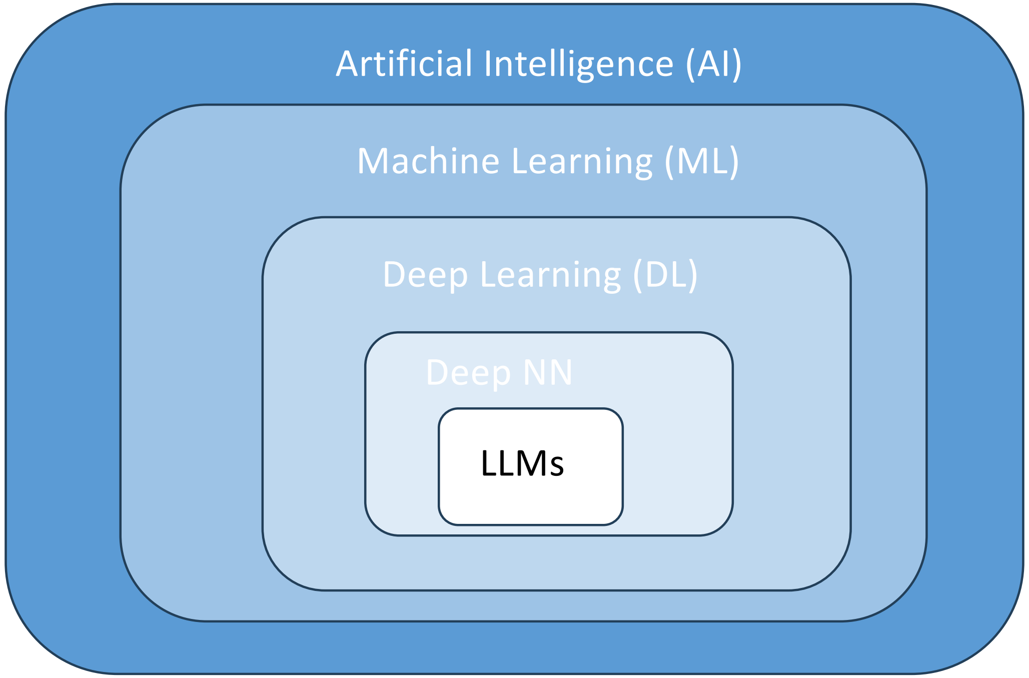
Acting rationally

Agent: Entity that perceives and acts

- A robot may be seen as an phisical agent
- Amazon recommender system
- Spam filter

Computational constrains: Design the best program with available resources

Artificial Intelligence (V)



Artificial Intelligence (AI)

Machine Learning (ML)

Deep Learning (DL)

Deep NN

LLMs

- AI: Broadly define intelligent systems, e.g. Autonomous robots
- ML: learning patterns from data, e.g. fraud detection
- DL: learning with deep neural networks, e.g. voice recognition
- Deep Neural Networks: algorithms mimicking the human brain, e.g. facial recognition
- LLMs: Understanding & generating natural language, e.g. virtual assistants

Outline

- Artificial Intelligence
- **Machine Learning**
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- Prompting
- Comparing LLM
- Conclusions

Machine Learning (I)

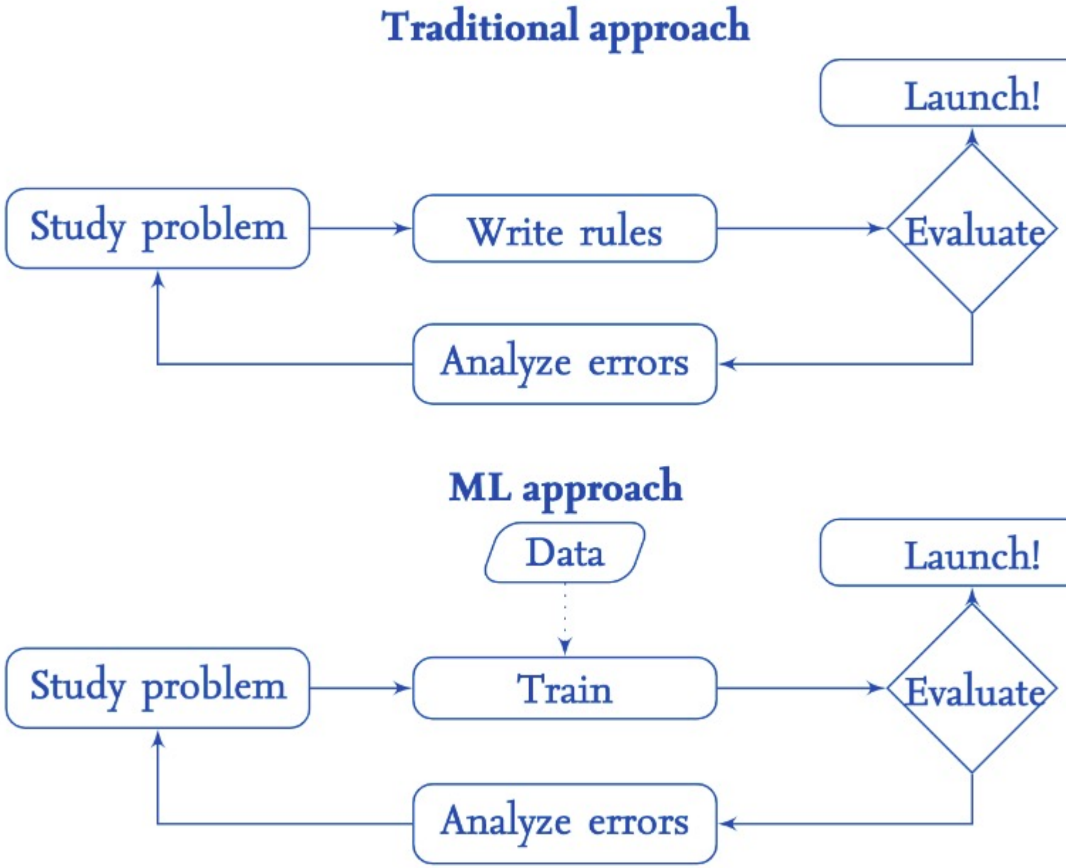
ML definition

ML is the science (and art) of programming computers so they can learn from data A. Géron, 2017

Alternative definitions

- Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed. Arthur Samuel, 1959
- A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. Tom Mitchell, 1997

Machine Learning (II)



Machine Learning (III)

	Classical	ML
Objetive	Algoritm	Training
Mentality	Engineer	Experimental Scientist
Tools	Programming	Mathematics/Algorithms
Type of problems	Automate tasks	Identify patterns
Methodology	Waterfall	Iterative

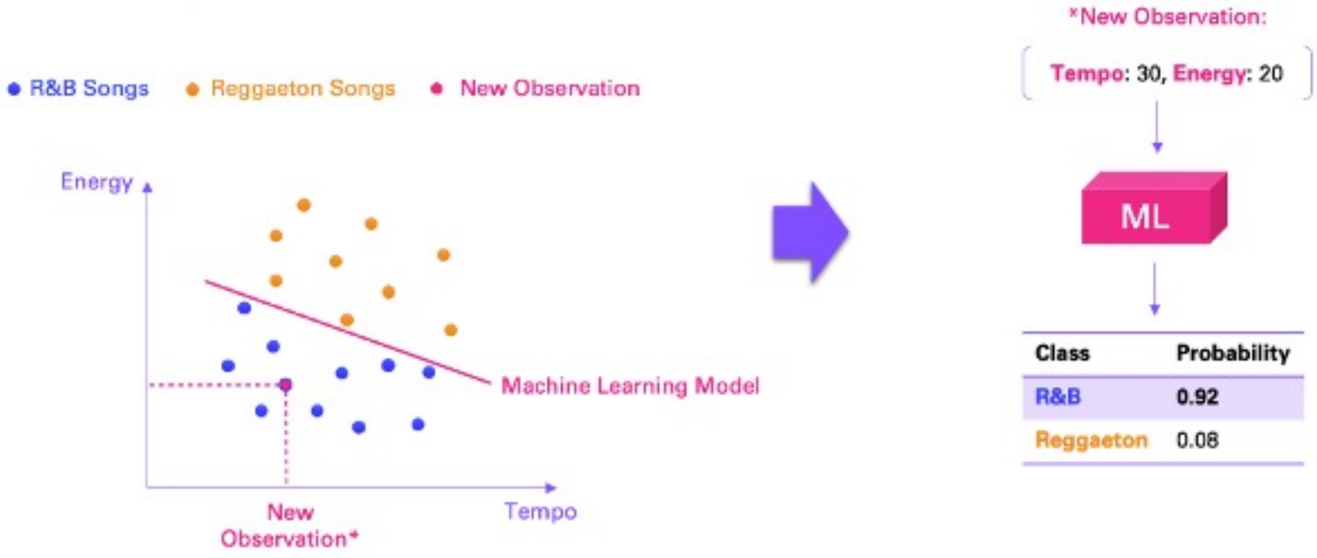
Machine Learning: Tasks

- **Classification**: assigning a category or label to input data
  Example: classifying the type of music
- **Regression**: predicting a continuous numerical value from input data
  Example: predicting housing prices based on features like size, location, etc.
- **Clustering**: grouping similar data points together without labeled categories
  Example: grouping customers based on their purchasing behavior for targeted marketing
- **Anomaly Detection**: identifying unusual data points that deviate from the norm
  Example: detecting credit card fraud
- …

Outline

- Artificial Intelligence
- Machine Learning
- **Classification**
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- Prompting
- Comparing LLM
- Conclusions

Classification: Predicting Music Genre

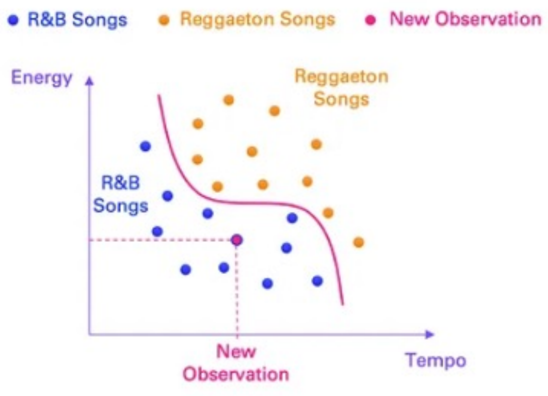


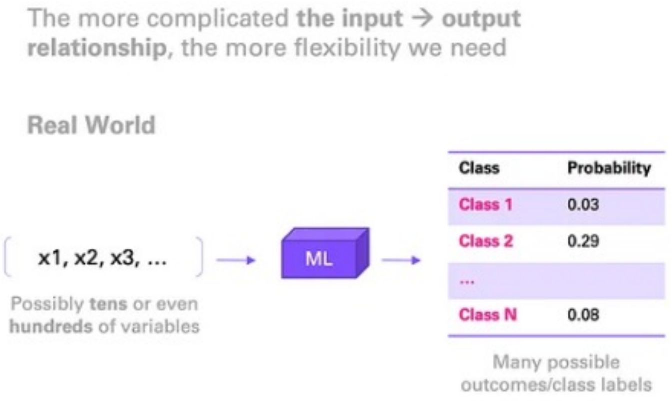
Class	Probability
R&B	0.92
Reggaeton	0.08

Credits: Andreas Stöffelbauer

- Classifying music genres based on tempo and energy using ML model
- R&B songs and Reggaeton songs are classified based on their tempo and energy levels

Non-Linear in Classification




The more complicated **the input → output relationship**, the more flexibility we need

Real World

x1, x2, x3, ...

Possibly **tens** or even **hundreds** of variables

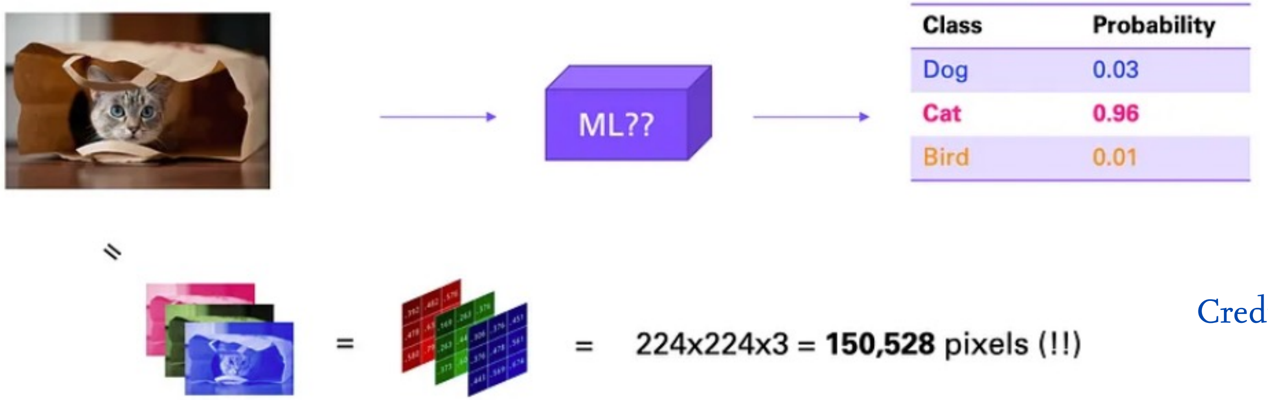
Class	Probability
Class 1	0.03
Class 2	0.29
...	
Class N	0.08

Many possible outcomes/class labels

Credits: Andreas Stöffelbauer

- Handling it is more complex
- The more complex the relationship between input and output, the more flexible the model needs to be

Image Classification



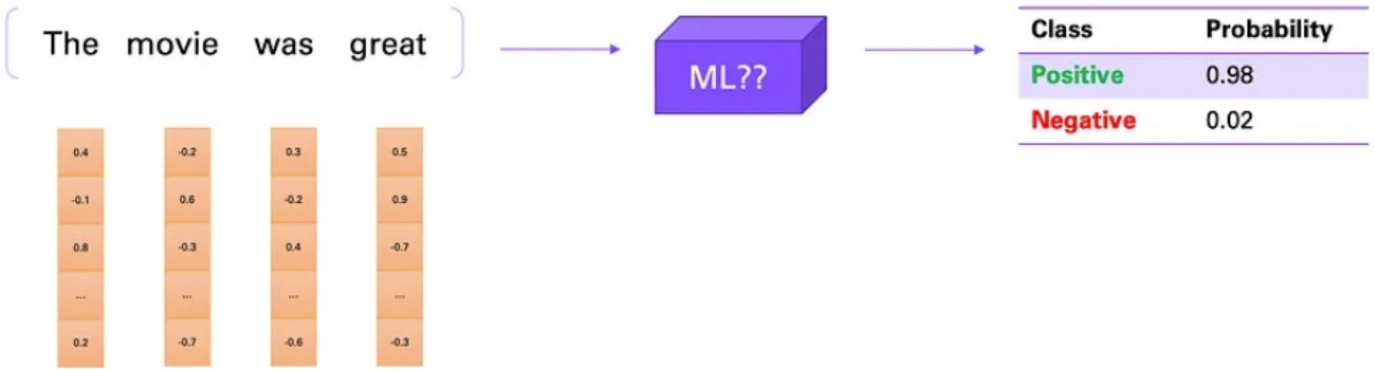
Class	Probability
Dog	0.03
Cat	**0.96**
Bird	0.01

224x224x3 = **150,528** pixels (!!)

Credits: Andreas Stöffelbauer

- Identify if an image is a cat, dog, or bird
- Can only compute numeric inputs → image are a set of pixels
- Height, width and colour (red, green, blue)
  - a small, low-quality 224x224 image consists of more than 150,000 pixels (224x224x3)
  - Complex relationship between raw pixels and class label

Text Classification

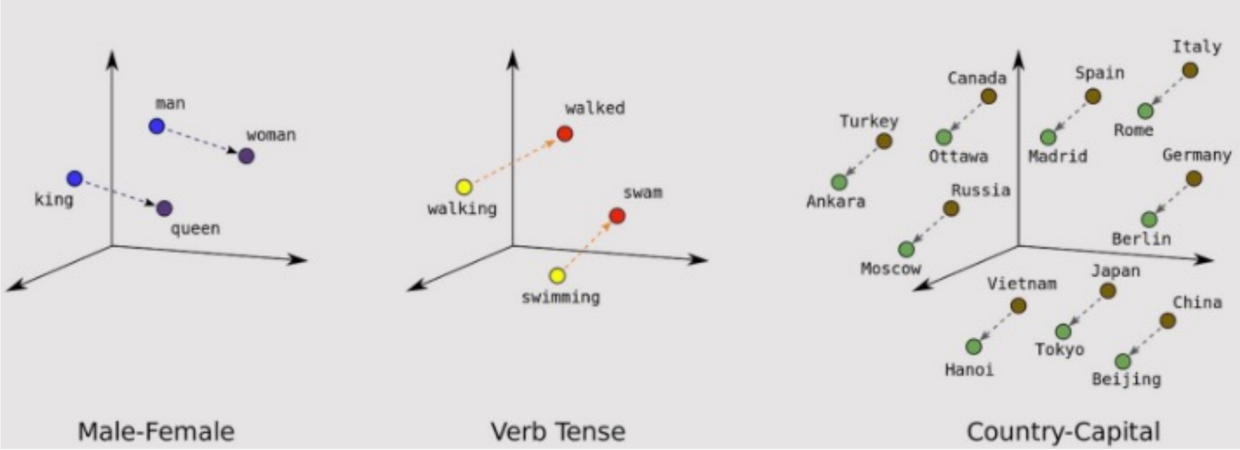


Class	Probability
Positive	0.98
Negative	0.02

Credits: Andreas Stöffelbauer

- Relationship between a sentence and its sentiment (e.g., positive or negative)
- Numeric inputs → Word embedding (convert words into numerical vectors)
- Words that often appear together in same context → have similar or 'closer' vectors

Word Embeddings



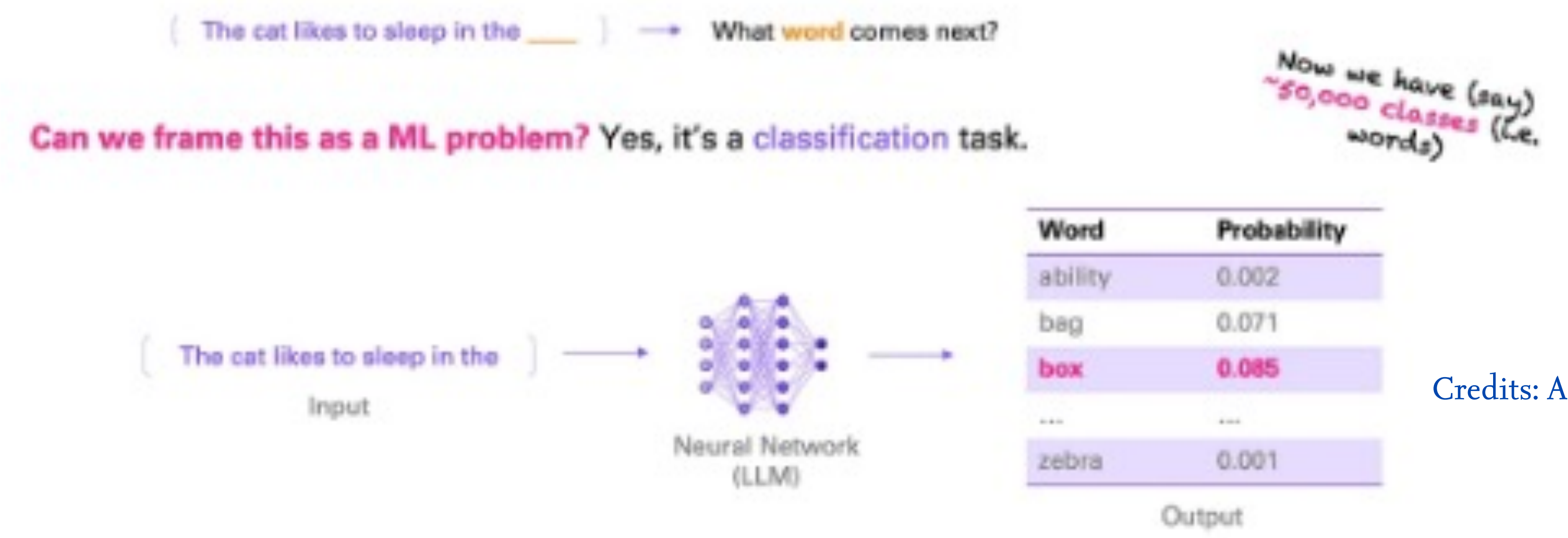
Credits: Manish Shivanandhan

- Traditional language models treated words as isolated entities
- WE: Capture the semantic essence of words and also encode relationships
- Place words with similar meaning or contexts close to each other. E.g.
  - GloVe (Stanford): generated numerical vectors for individual words
  - Fastext (Facebook): capture the intricacies of languages that have complex word structures

Outline

- Artificial Intelligence
- Machine Learning
- Classification
- **Language Model**
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- Prompting
- Comparing LLM
- Conclusions

Language Modeling



The cat likes to sleep in the ___ → What word comes next?

Can we frame this as a ML problem? Yes, it's a classification task.

Now we have (say) ~50,000 classes (i.e. words)

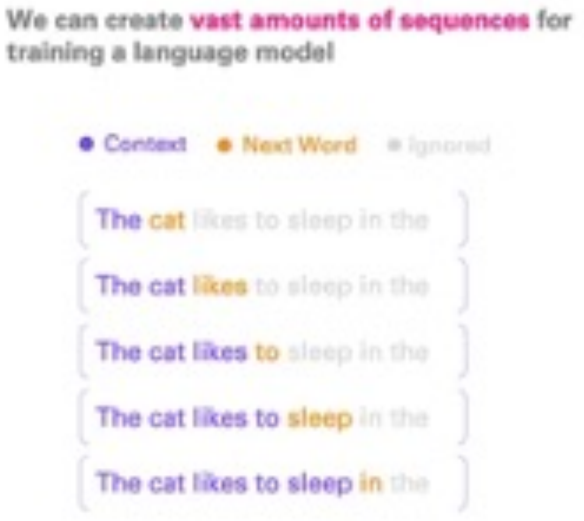
The cat likes to sleep in the

Word	Probability
ability	0.002
bag	0.071
box	0.085
...	...
zebra	0.001

Credits: Andreas Stöffelbauer

- Predicting the next word in a sequence, is framed as a classification task
- Number of inputs/outputs is very large → Neuronal Networks (NN)
- NN often have many layers (Deep Learning)

Massive Training Data



We can create **vast amounts of sequences** for training a language model

● Context ● Next Word ● Ignored

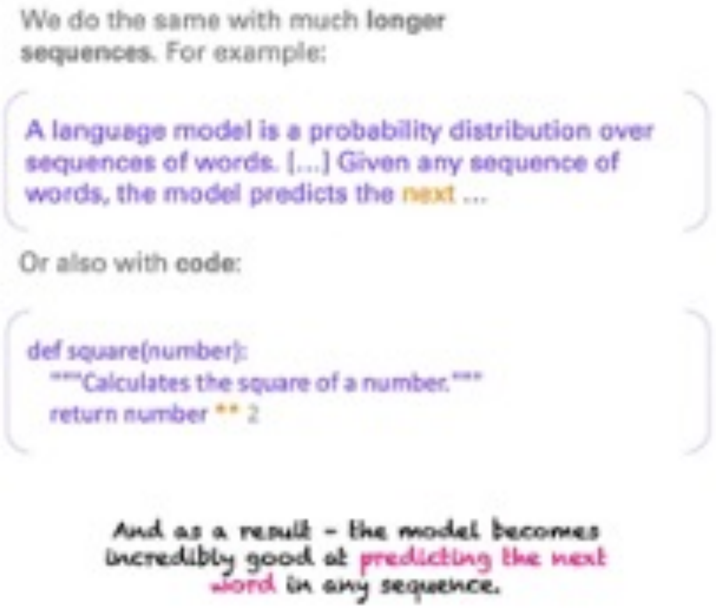
The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the



We do the same with much **longer sequences**. For example:

A language model is a probability distribution over sequences of words. [...] Given any sequence of words, the model predicts the next ...

Or also with code:

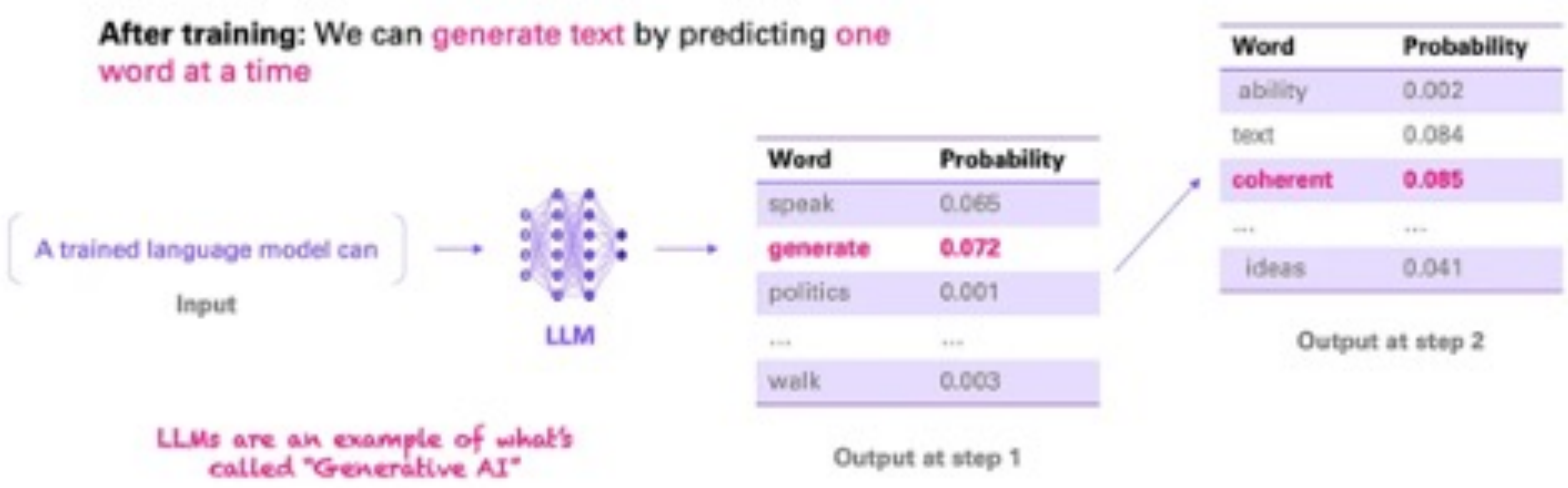
```
def square(number):
    """Calculates the square of a number."""
    return number ** 2
```

And as a result – the model becomes incredibly good at predicting the next word in any sequence.

Credits: Andreas Stöffelbauer

- Creating vast amounts of sequences for training a language model: from books, research papers, internet…
- No need to label data, next word itself is the label → self-supervised learning
- Any language!

Natural Language Generation



After training: We can generate text by predicting one word at a time

A trained language model can

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

Output at step 1

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
...	...
ideas	0.041

Output at step 2

LLMs are an example of what's called "Generative AI"

- Generating text by predicting one word at a time or also sample from 5 most likely words → you don´t get the same answer a response is generated
- LLMs perform natural language generation as an example of Generative AI

Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- **ChatGPT: Generative Pre-trained Transformer**
- Limitation of Pre-trained models
- Hallucination
- Prompting
- Comparing LLM
- Conclusions

What is ChatGPT?

- ChatGPT is an AI language model developed by OpenAI
- It uses deep learning techniques to understand and generate human-like text
- Built on the GPT-4 architecture, it can perform a wide range of natural language tasks
- Now GPT 5.0

GPT: *Generative Pre-trained Transformer*

- **Generative** means next word prediction
- **Pre-trained** on massive amount of text
- In traditional models each word (Word embeddings) is converted into a fixed vector in a pre-defined space
- GPT utilizes "**transformer embeddings**" (TE) that consider the word and the surrounding context
- TE are based on the **Attention Model mechanism** in NN (Vaswani et al. 2017)
- It allows GPT to decipher words with multiple meanings, like "bank" (financial vs. river).
- The resulting embeddings capture not just the word, but its connection to the whole sentence

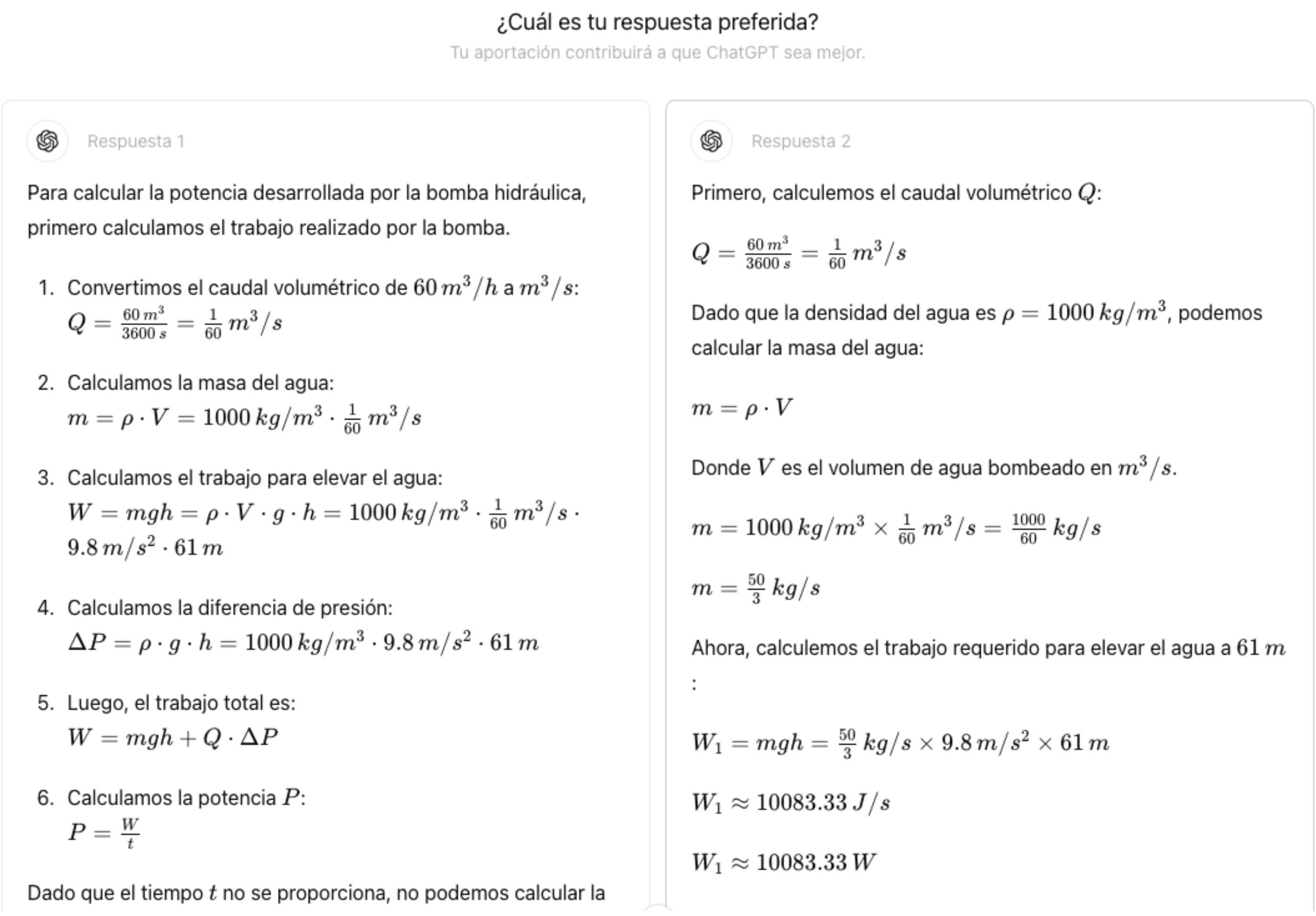
Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- **Limitation of Pre-trained models**
- Hallucination
- Prompting
- Comparing LLM
- Conclusions

Limitations of Pre-trained models (I)

- They are better at continuing text sequences, but struggle with instructions and questions
- They may not understand your true intentions or how to fulfill them
- **Instruction Fine-tuning**:
  - Train the model using high-quality instruction-response pairs
  - Focus: unlearns "text completion" and teaches the model to follow instructions
  - Data: requires smaller, more expensive human-sourced instruction data
- Reinforcement Learning from Human Feedback (RLHF) (Optional)
  - Similar Goal: further improves alignment with human values and preferences
  - Research: combining RL and language modeling shows potential for significant advancements over the actual LLMs

Example of RLHF



¿Cuál es tu respuesta preferida?

Tu aportación contribuirá a que ChatGPT sea mejor.

Respuesta 1

Para calcular la potencia desarrollada por la bomba hidráulica, primero calculamos el trabajo realizado por la bomba.

1. Convertimos el caudal volumétrico de $60\,m^3/h$ a m^3/s:
   $Q = \frac{60\,m^3}{3600\,s} = \frac{1}{60}\,m^3/s$
2. Calculamos la masa del agua:
   $m = \rho \cdot V = 1000\,kg/m^3 \cdot \frac{1}{60}\,m^3/s$
3. Calculamos el trabajo para elevar el agua:
   $W = mgh = \rho \cdot V \cdot g \cdot h = 1000\,kg/m^3 \cdot \frac{1}{60}\,m^3/s \cdot 9.8\,m/s^2 \cdot 61\,m$
4. Calculamos la diferencia de presión:
   $\Delta P = \rho \cdot g \cdot h = 1000\,kg/m^3 \cdot 9.8\,m/s^2 \cdot 61\,m$
5. Luego, el trabajo total es:
   $W = mgh + Q \cdot \Delta P$
6. Calculamos la potencia P:
   $P = \frac{W}{t}$

Dado que el tiempo t no se proporciona, no podemos calcular la

Respuesta 2

Primero, calculemos el caudal volumétrico Q:

$$Q = \frac{60\,m^3}{3600\,s} = \frac{1}{60}\,m^3/s$$

Dado que la densidad del agua es $\rho = 1000\,kg/m^3$, podemos calcular la masa del agua:

$$m = \rho \cdot V$$

Donde V es el volumen de agua bombeado en m^3/s.

$$m = 1000\,kg/m^3 \times \frac{1}{60}\,m^3/s = \frac{1000}{60}\,kg/s$$

$$m = \frac{50}{3}\,kg/s$$

Ahora, calculemos el trabajo requerido para elevar el agua a 61 m :

$$W_1 = mgh = \frac{50}{3}\,kg/s \times 9.8\,m/s^2 \times 61\,m$$

$$W_1 \approx 10083.33\,J/s$$

$$W_1 \approx 10083.33\,W$$

Limitations of Pre-trained models (II)

Question	Answer
Can an LLM perform **Text Summarization**?	Yes e.g. research journals have abstract and extent content (pre-training)
Can an LLM perform **Question Answering**?	Yes, fine-tuning. E.g. You are always right!
Is the information true?	**No, it may answer wrong or even make stuff up**
Is it deterministic in the answers?	No, it uses a sample mechanism ("n" most likely words), each answer may be different

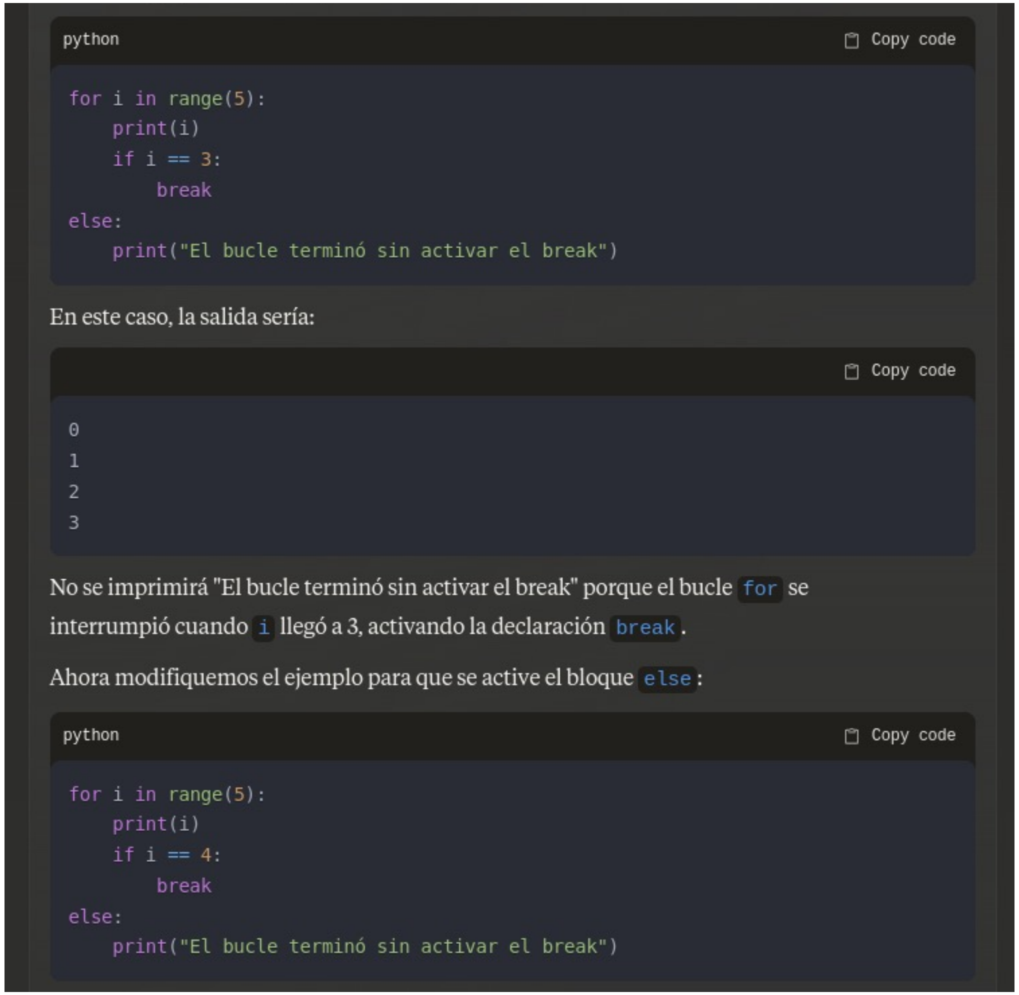
Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- **Hallucination**
- Prompting
- Comparing LLM
- Conclusions

Hallucination

- Problem
    - LLMs are designed to create human-like text, but not necessarily factual
    - The model lack an intrinsic mechanism to verify the truthfulness of the generated info
- Idea
    - We should anchor them in real-world information to prevent inaccuracies
    - We possess the knowledge to address and rectify this issue
- Solution
    - Incorporate accurate knowledge into the context to enhance its reliability
    - Develop and integrate validation mechanisms that cross-check the information against trusted DB
    - Update the training data with current, verified information to ensure the LLM remains accurate

Example of Hallucination



```python
for i in range(5):
    print(i)
    if i == 3:
        break
else:
    print("El bucle terminó sin activar el break")
```

En este caso, la salida sería:

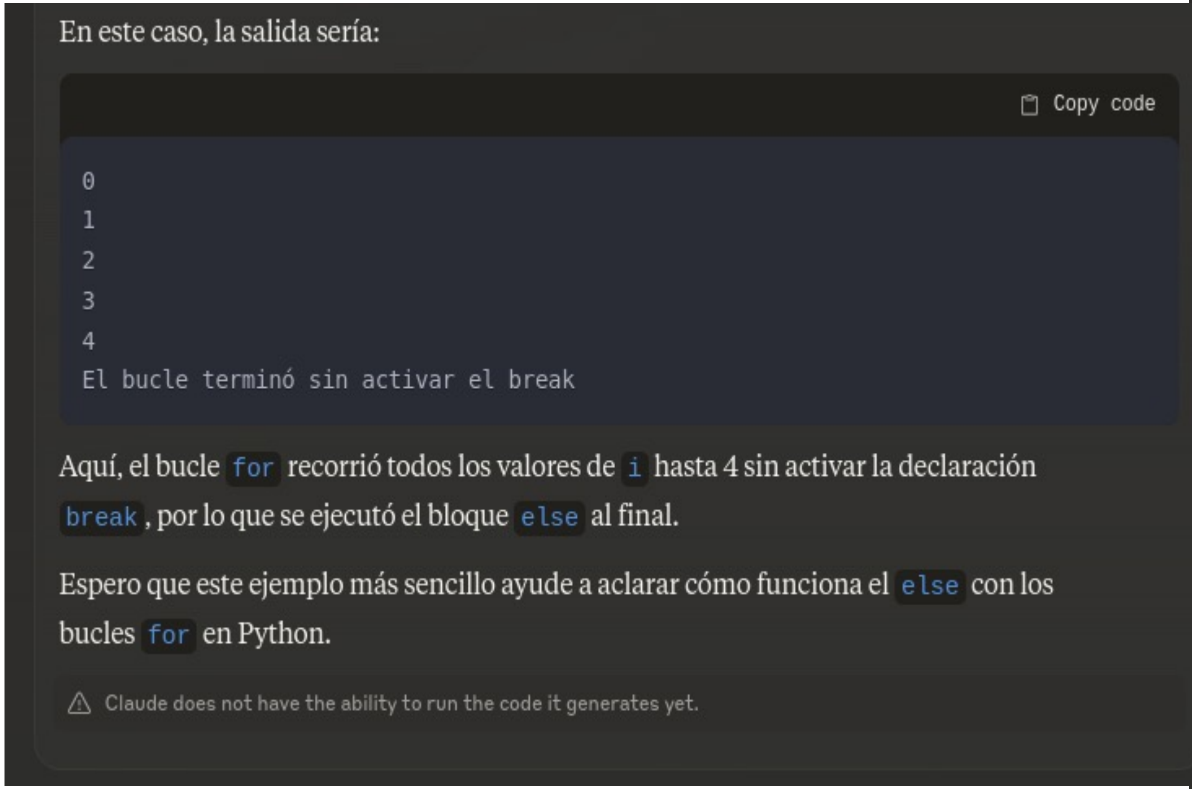
```
0
1
2
3
```

No se imprimirá "El bucle terminó sin activar el break" porque el bucle `for` se interrumpió cuando `i` llegó a 3, activando la declaración `break`.

Ahora modifiquemos el ejemplo para que se active el bloque `else`:

```python
for i in range(5):
    print(i)
    if i == 4:
        break
else:
    print("El bucle terminó sin activar el break")
```

Example of Hallucination



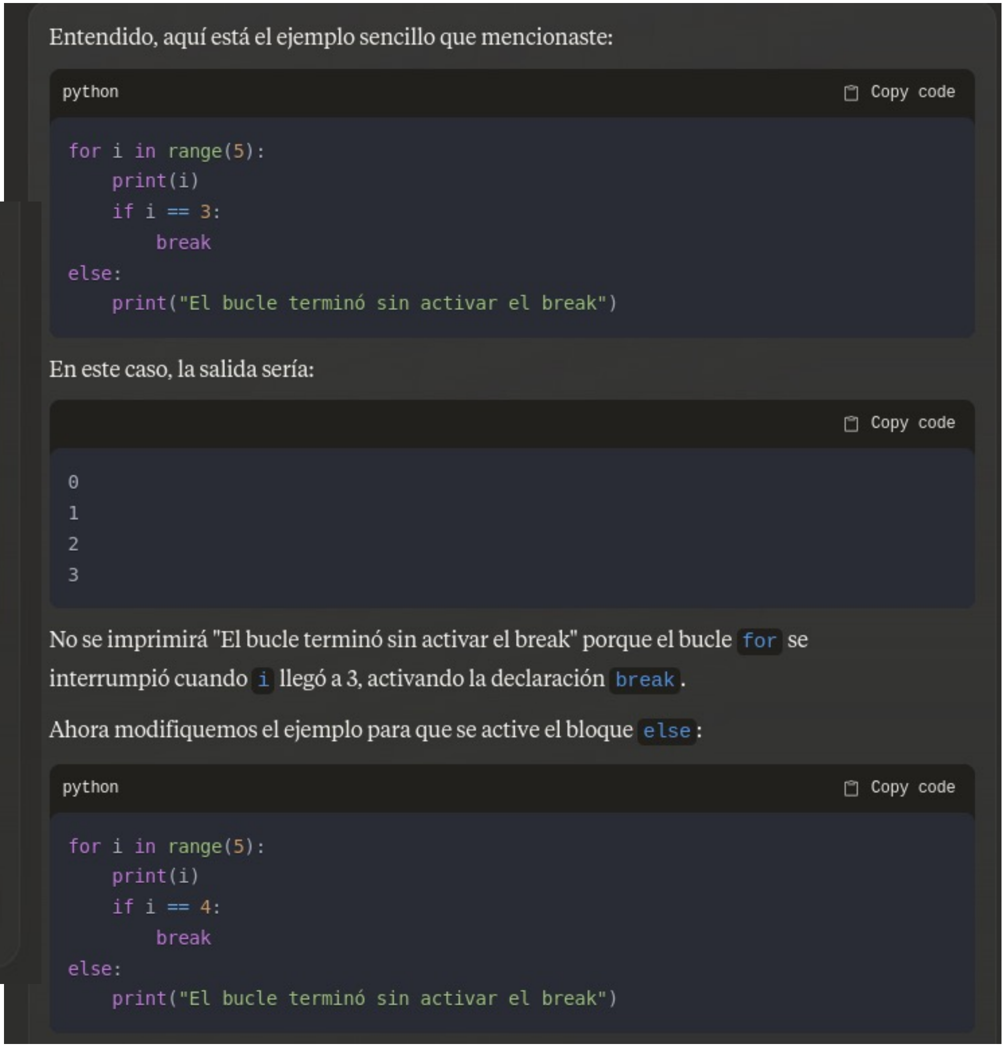
En este caso, la salida sería:

```
0
1
2
3
4
El bucle terminó sin activar el break
```

Aquí, el bucle `for` recorrió todos los valores de `i` hasta 4 sin activar la declaración `break`, por lo que se ejecutó el bloque `else` al final.

Espero que este ejemplo más sencillo ayude a aclarar cómo funciona el `else` con los bucles `for` en Python.

⚠ Claude does not have the ability to run the code it generates yet.



Entendido, aquí está el ejemplo sencillo que mencionaste:

```python
for i in range(5):
    print(i)
    if i == 3:
        break
else:
    print("El bucle terminó sin activar el break")
```

En este caso, la salida sería:

```
0
1
2
3
```

No se imprimirá "El bucle terminó sin activar el break" porque el bucle `for` se interrumpió cuando `i` llegó a 3, activando la declaración `break`.

Ahora modifiquemos el ejemplo para que se active el bloque `else`:

```python
for i in range(5):
    print(i)
    if i == 4:
        break
else:
    print("El bucle terminó sin activar el break")
```

Avoid Hallucinations

- Understanding how to effectively interact with AI models is crucial
- One key element in maximizing the utility of these models is **the art of prompting**
- Prompting refers to the way users formulate their questions or instructions when communicating with AI models
- It involves crafting precise and clear inputs to guide the AI in generating accurate, relevant, and useful responses.

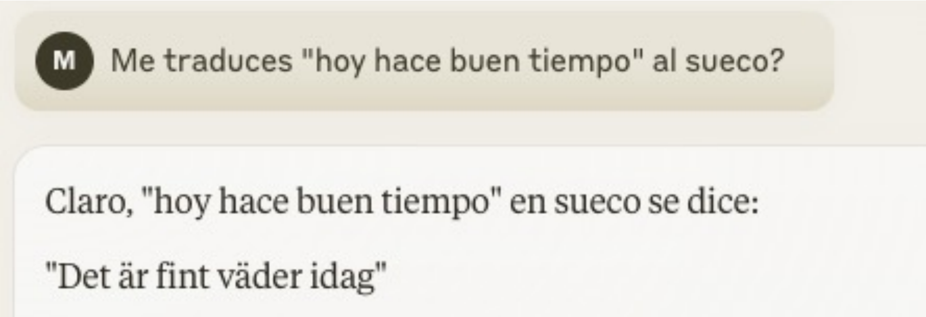
Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- **Prompting**
- Comparing LLM
- Conclusions

Zero-Shot Prompting

- Definition
  - Capability of a LM to perform a task without being explicitly trained on examples
  - The model relies on its general understanding of language and knowledge acquired during pre-training to generate appropriate responses
- Explanation
  - The model leverages its broad knowledge base to infer the correct output for a given task
  - The model can handle new tasks by interpreting the on its pre-existing knowledge
  - The success often depends **on how well the task is described in the prompt**

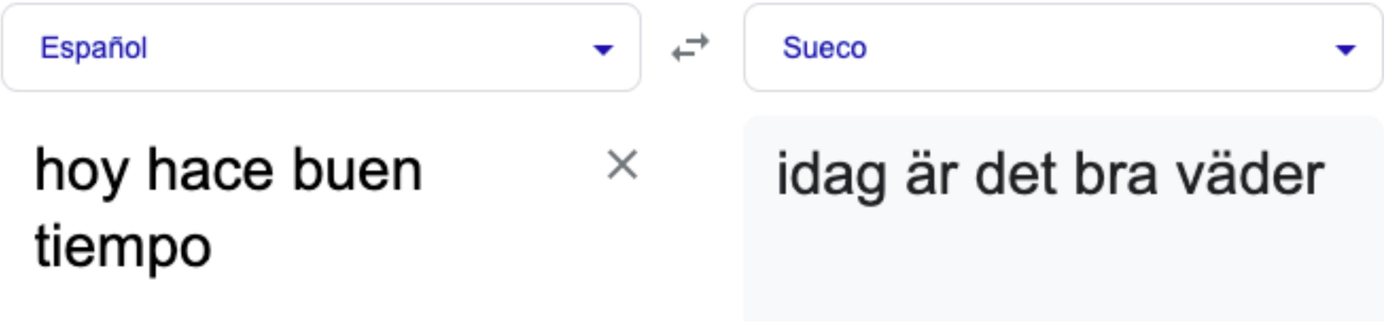
Example



Me traduces "hoy hace buen tiempo" al sueco?

Claro, "hoy hace buen tiempo" en sueco se dice:

"Det är fint väder idag"



Español

hoy hace buen tiempo

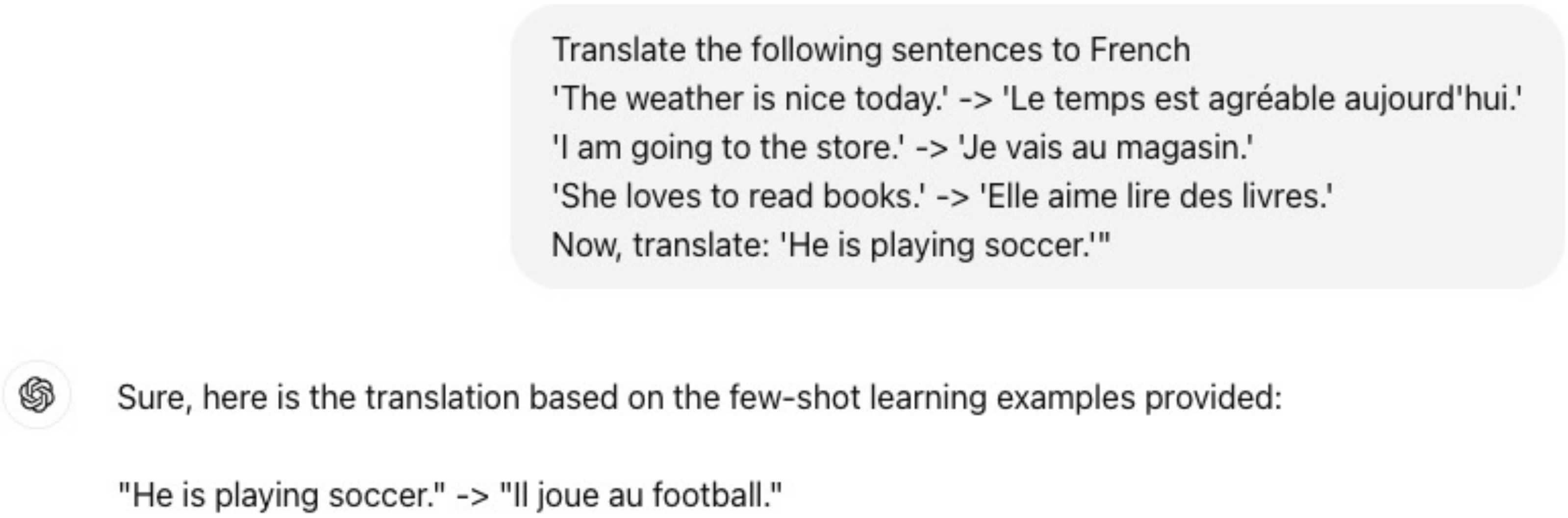
Sueco

idag är det bra väder

Few-Shot Learning Prompting

- Capability of a LM to perform a new task using small number of examples
- Explanation
  - LM understand and generalize from few examples, making it highly efficient for tasks with limited data
  - The model adapts to new tasks by incorporating the limited examples provided, refining its predictions based on these specific instances
  - The quality and relevance of the examples are crucial
  - Well-chosen examples can significantly improve the model's performance

Example



Translate the following sentences to French
'The weather is nice today.' -> 'Le temps est agréable aujourd'hui.'
'I am going to the store.' -> 'Je vais au magasin.'
'She loves to read books.' -> 'Elle aime lire des livres.'
Now, translate: 'He is playing soccer.'"

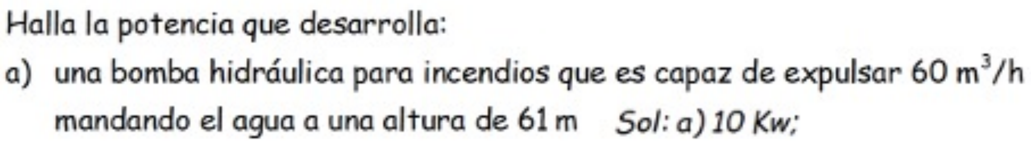
Sure, here is the translation based on the few-shot learning examples provided:

"He is playing soccer." -> "Il joue au football."

Step-by-Step Prompting

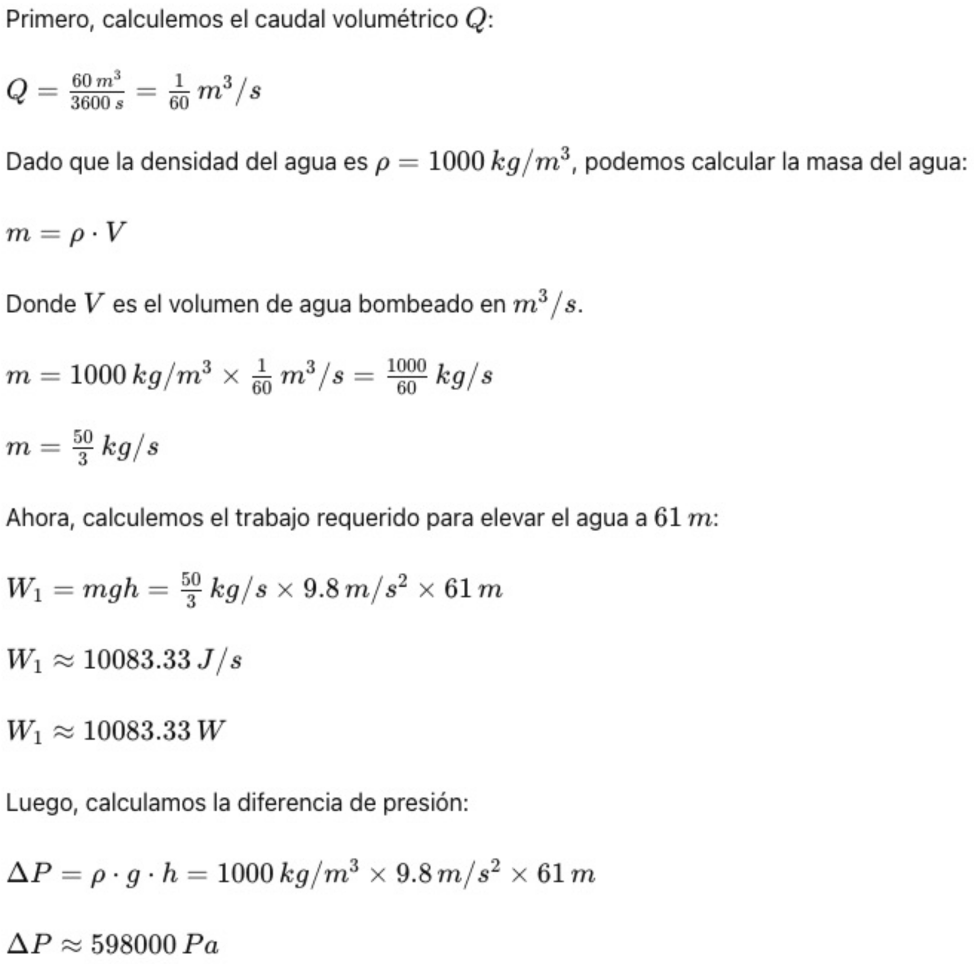
- Technically "Chain-of-Thought Prompting"
- It gives the model a "working memory" as humans
- Involves guiding a model through complex tasks by breaking them down into simpler, sequential steps
- **Purpose:** enhances the model's ability to reason and solve problems by providing it with a structured approach

Example (I)



Halla la potencia que desarrolla:

a) una bomba hidráulica para incendios que es capaz de expulsar 60 m^3/h mandando el agua a una altura de 61 m *Sol: a) 10 Kw;*



Primero, calculemos el caudal volumétrico Q:

$$Q = \frac{60\,m^3}{3600\,s} = \frac{1}{60}\,m^3/s$$

Dado que la densidad del agua es $\rho = 1000\,kg/m^3$, podemos calcular la masa del agua:

$$m = \rho \cdot V$$

Donde V es el volumen de agua bombeado en m^3/s.

$$m = 1000\,kg/m^3 \times \frac{1}{60}\,m^3/s = \frac{1000}{60}\,kg/s$$

$$m = \frac{50}{3}\,kg/s$$

Ahora, calculemos el trabajo requerido para elevar el agua a 61 m:

$$W_1 = mgh = \frac{50}{3}\,kg/s \times 9.8\,m/s^2 \times 61\,m$$

$$W_1 \approx 10083.33\,J/s$$

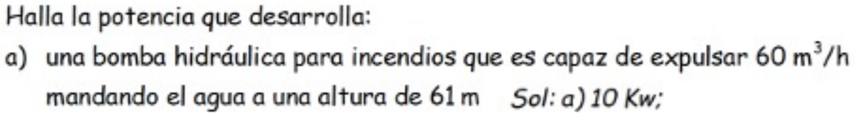
$$W_1 \approx 10083.33\,W$$

Luego, calculamos la diferencia de presión:

$$\Delta P = \rho \cdot g \cdot h = 1000\,kg/m^3 \times 9.8\,m/s^2 \times 61\,m$$

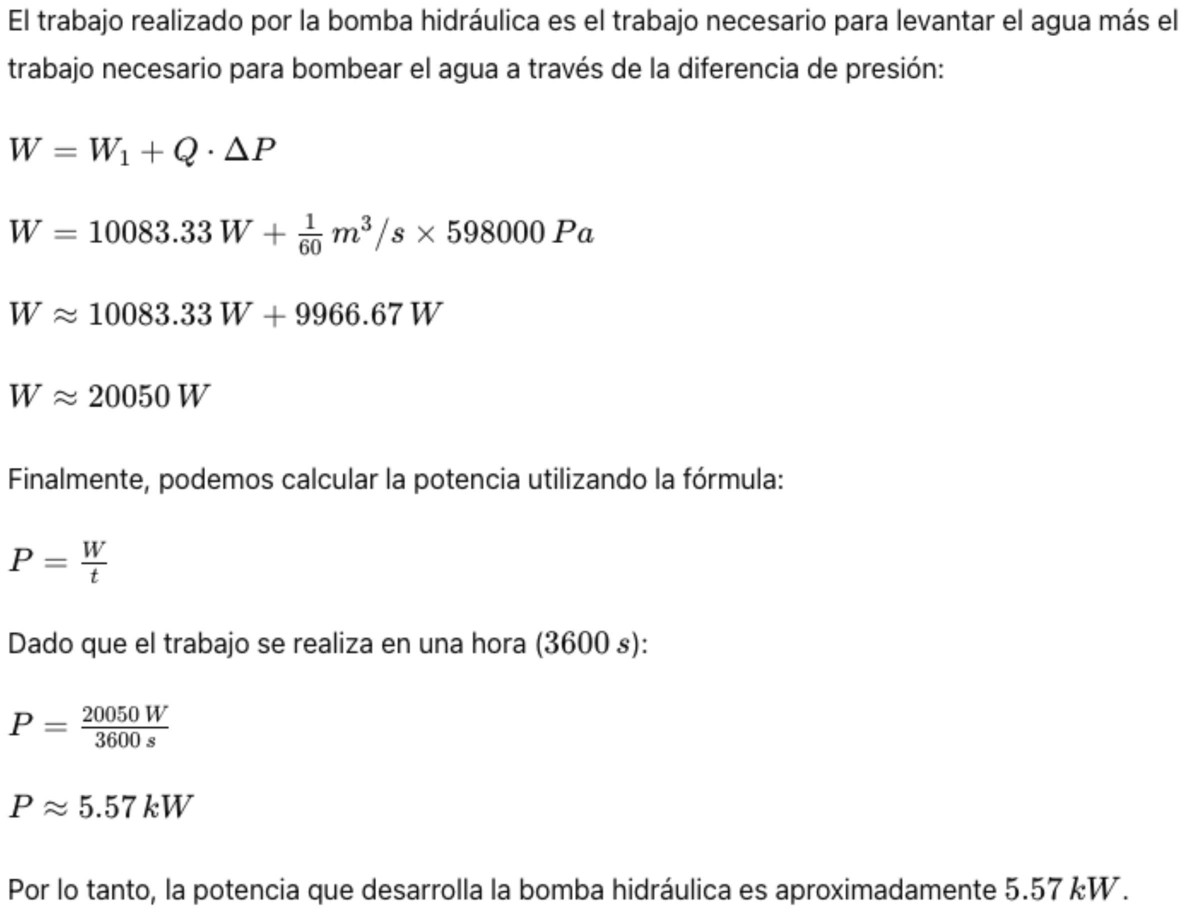
$$\Delta P \approx 598000\,Pa$$

…Example (I)



Halla la potencia que desarrolla:

a) una bomba hidráulica para incendios que es capaz de expulsar 60 m^3/h mandando el agua a una altura de 61 m *Sol: a) 10 Kw;*



El trabajo realizado por la bomba hidráulica es el trabajo necesario para levantar el agua más el trabajo necesario para bombear el agua a través de la diferencia de presión:

$$W = W_1 + Q \cdot \Delta P$$

$$W = 10083.33\,W + \tfrac{1}{60}\,m^3/s \times 598000\,Pa$$

$$W \approx 10083.33\,W + 9966.67\,W$$

$$W \approx 20050\,W$$

Finalmente, podemos calcular la potencia utilizando la fórmula:

$$P = \tfrac{W}{t}$$

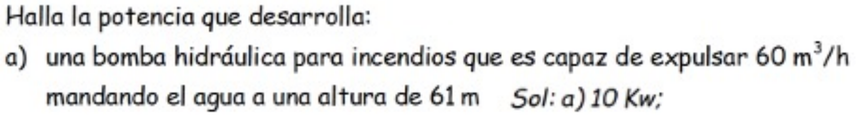
Dado que el trabajo se realiza en una hora ($3600\,s$):

$$P = \tfrac{20050\,W}{3600\,s}$$

$$P \approx 5.57\,kW$$

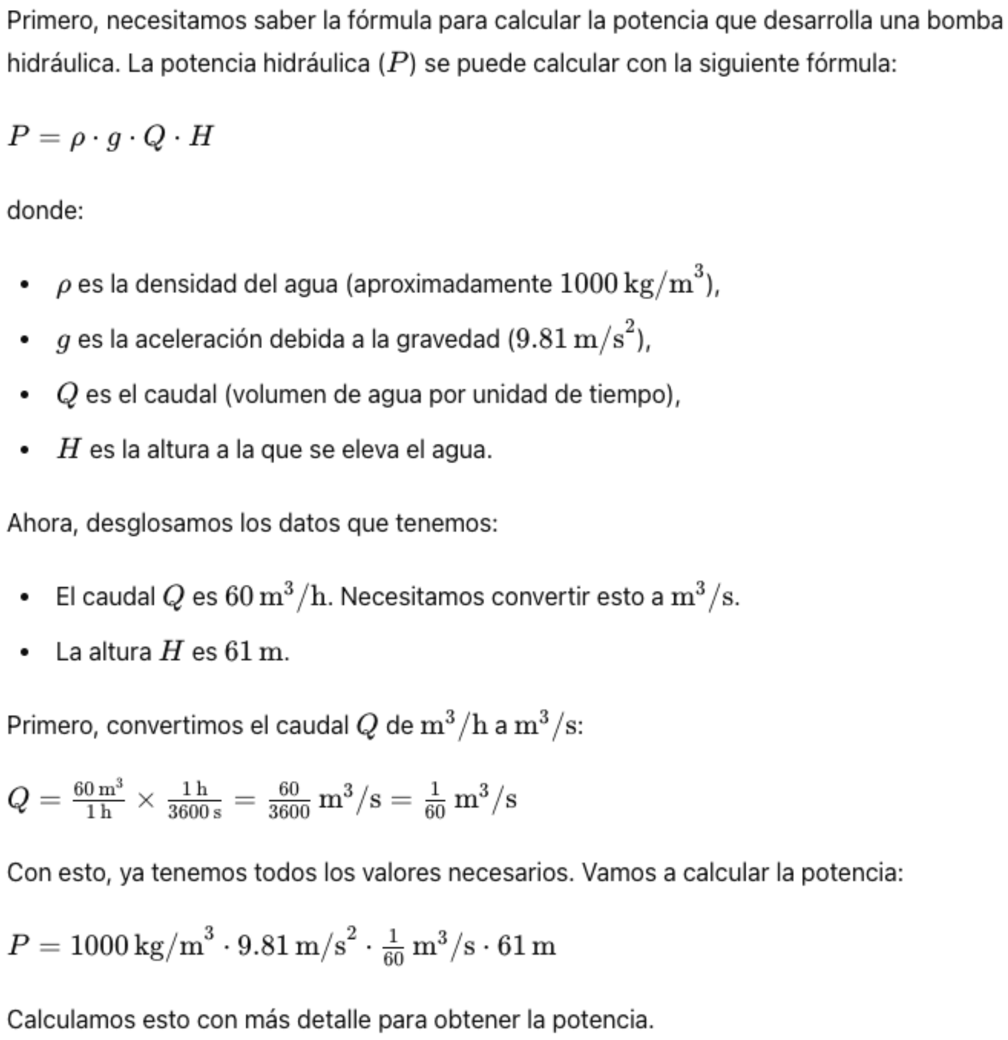
Por lo tanto, la potencia que desarrolla la bomba hidráulica es aproximadamente $5.57\,kW$.

Example (III)



Halla la potencia que desarrolla:

a) una bomba hidráulica para incendios que es capaz de expulsar 60 m^3/h mandando el agua a una altura de 61 m *Sol: a) 10 Kw;*



Primero, necesitamos saber la fórmula para calcular la potencia que desarrolla una bomba hidráulica. La potencia hidráulica (P) se puede calcular con la siguiente fórmula:

$$P = \rho \cdot g \cdot Q \cdot H$$

donde:

- ρ es la densidad del agua (aproximadamente $1000\,\mathrm{kg/m^3}$),
- g es la aceleración debida a la gravedad ($9.81\,\mathrm{m/s^2}$),
- Q es el caudal (volumen de agua por unidad de tiempo),
- H es la altura a la que se eleva el agua.

Ahora, desglosamos los datos que tenemos:

- El caudal Q es $60\,\mathrm{m^3/h}$. Necesitamos convertir esto a $\mathrm{m^3/s}$.
- La altura H es $61\,\mathrm{m}$.

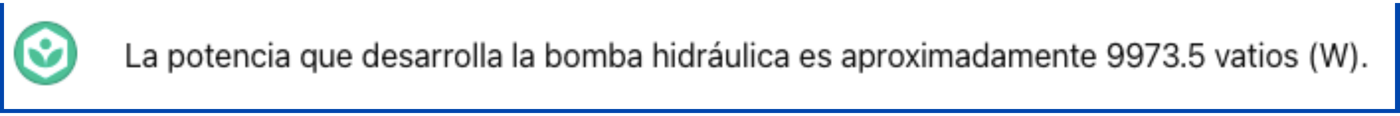
Primero, convertimos el caudal Q de $\mathrm{m^3/h}$ a $\mathrm{m^3/s}$:

$$Q = \frac{60\,\mathrm{m^3}}{1\,\mathrm{h}} \times \frac{1\,\mathrm{h}}{3600\,\mathrm{s}} = \frac{60}{3600}\,\mathrm{m^3/s} = \frac{1}{60}\,\mathrm{m^3/s}$$

Con esto, ya tenemos todos los valores necesarios. Vamos a calcular la potencia:

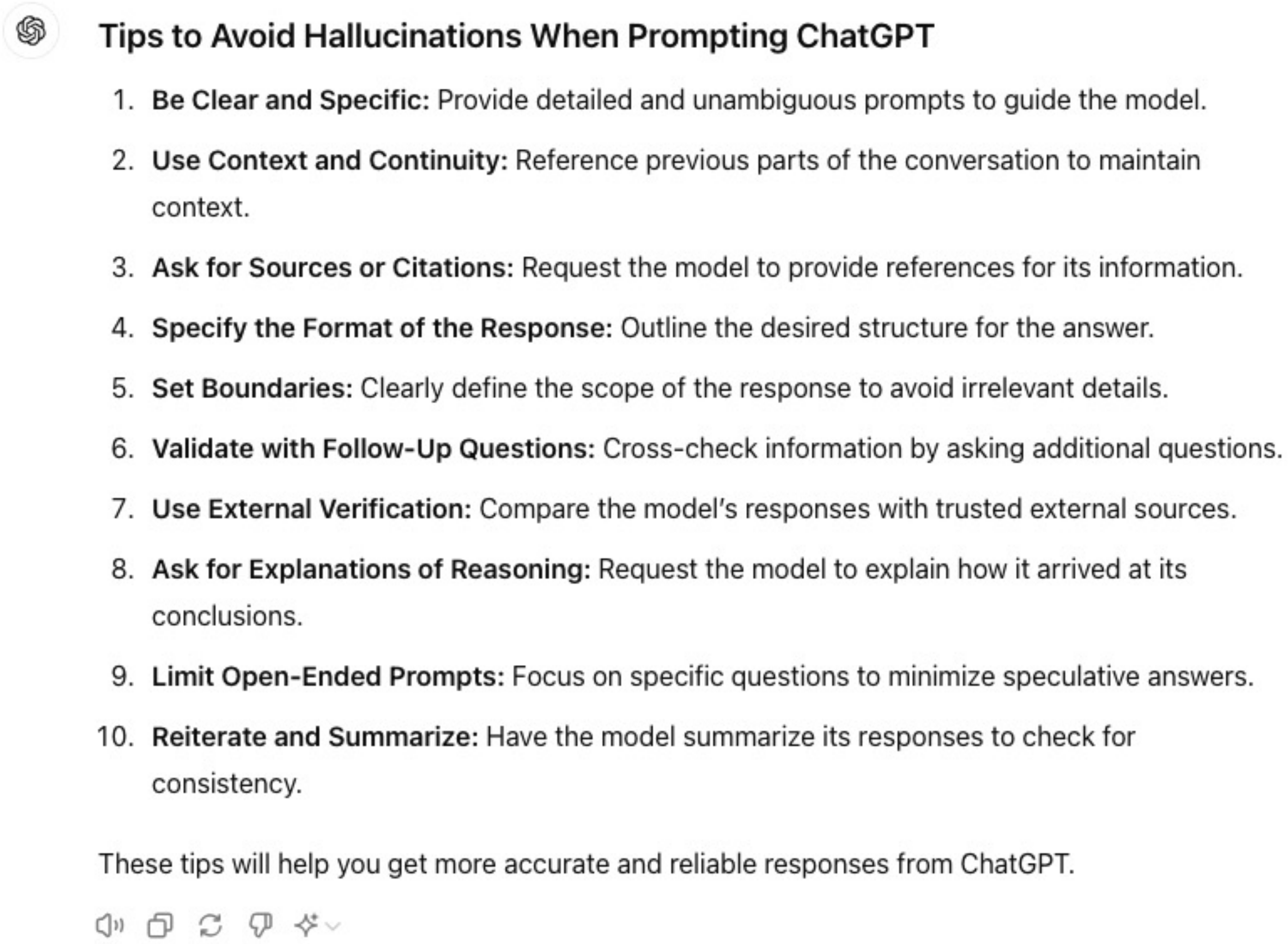
$$P = 1000\,\mathrm{kg/m^3} \cdot 9.81\,\mathrm{m/s^2} \cdot \frac{1}{60}\,\mathrm{m^3/s} \cdot 61\,\mathrm{m}$$

Calculamos esto con más detalle para obtener la potencia.



La potencia que desarrolla la bomba hidráulica es aproximadamente 9973.5 vatios (W).

Tips



Tips to Avoid Hallucinations When Prompting ChatGPT

1. **Be Clear and Specific:** Provide detailed and unambiguous prompts to guide the model.
2. **Use Context and Continuity:** Reference previous parts of the conversation to maintain context.
3. **Ask for Sources or Citations:** Request the model to provide references for its information.
4. **Specify the Format of the Response:** Outline the desired structure for the answer.
5. **Set Boundaries:** Clearly define the scope of the response to avoid irrelevant details.
6. **Validate with Follow-Up Questions:** Cross-check information by asking additional questions.
7. **Use External Verification:** Compare the model's responses with trusted external sources.
8. **Ask for Explanations of Reasoning:** Request the model to explain how it arrived at its conclusions.
9. **Limit Open-Ended Prompts:** Focus on specific questions to minimize speculative answers.
10. **Reiterate and Summarize:** Have the model summarize its responses to check for consistency.

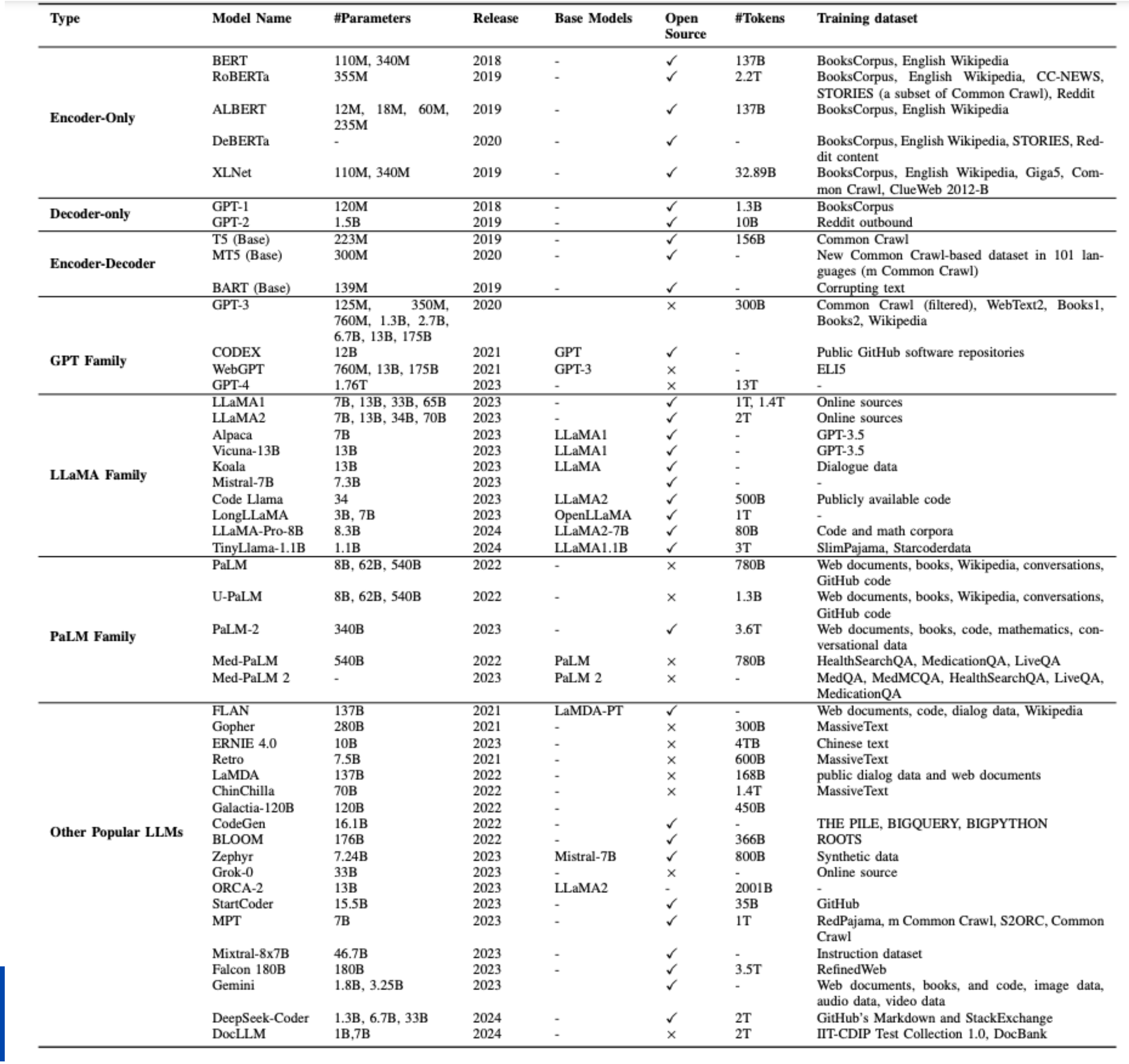
These tips will help you get more accurate and reliable responses from ChatGPT.

Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- Prompting
- **Comparing LLM**
- Conclusions

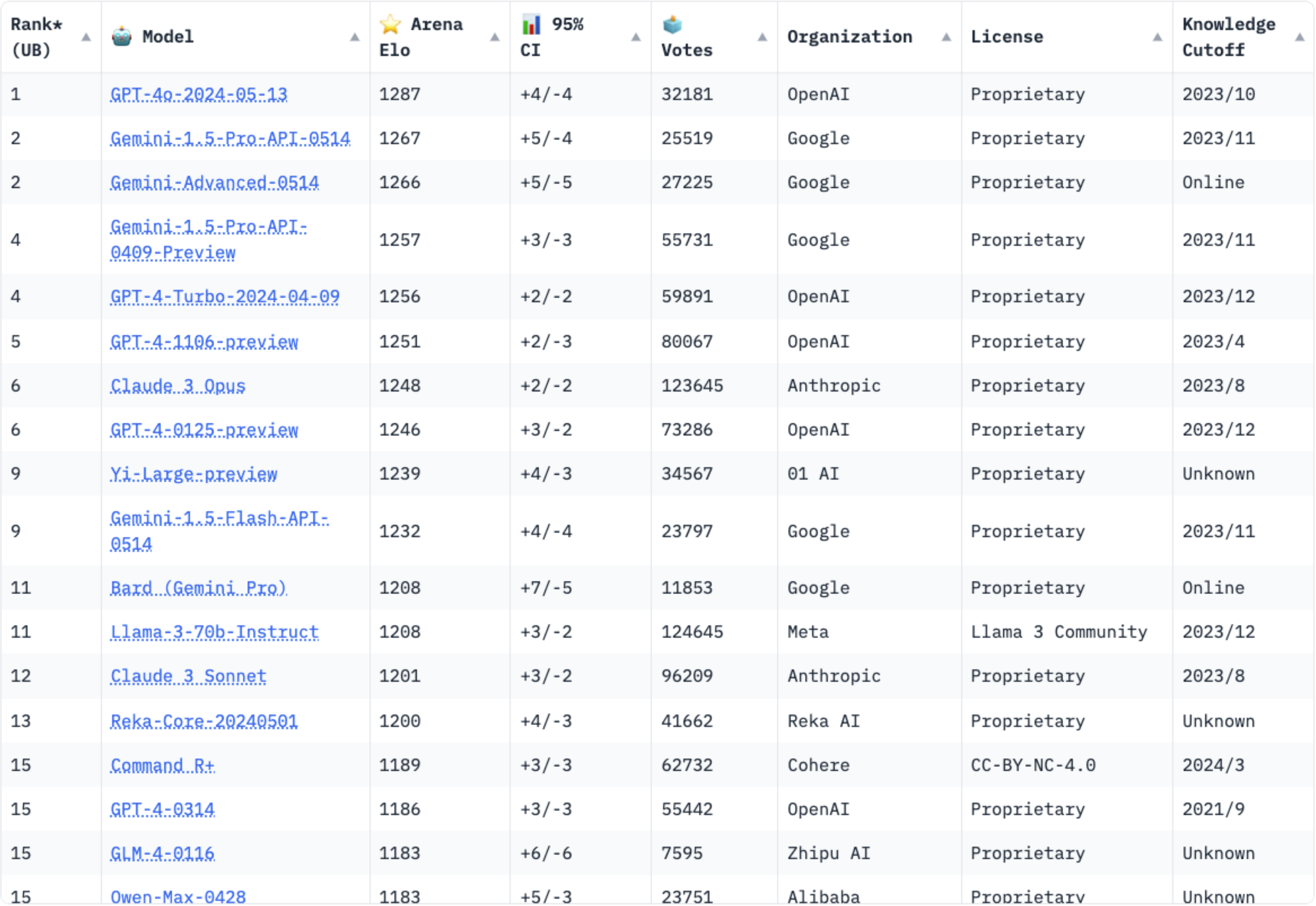
What’s out?

Credits: Minaee et al. (2024)



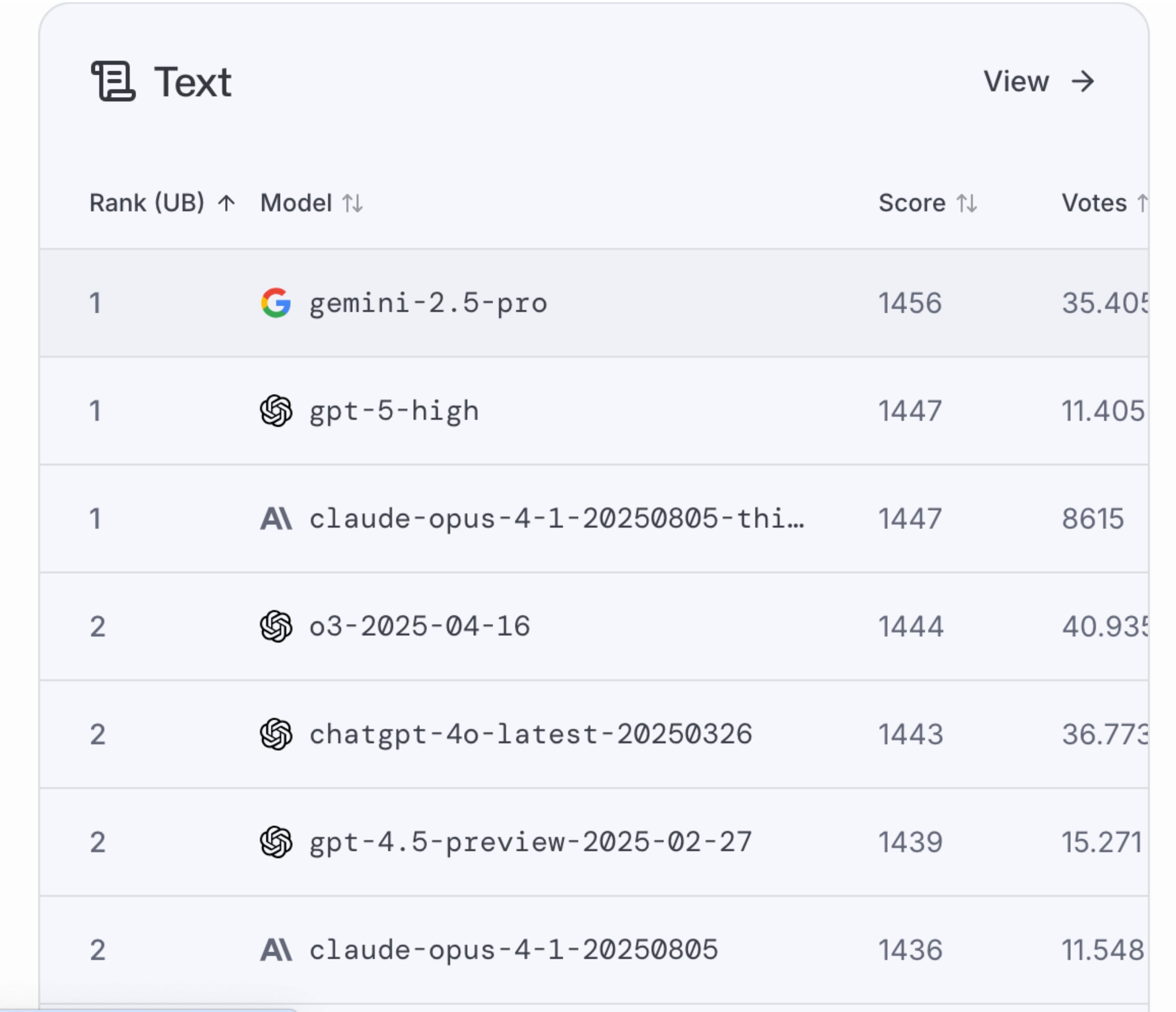
Type	Model Name	#Parameters	Release	Base Models	Open Source	#Tokens	Training dataset
Encoder-Only	BERT	110M, 340M	2018	-	✓	137B	BooksCorpus, English Wikipedia
	RoBERTa	355M	2019	-	✓	2.2T	BooksCorpus, English Wikipedia, CC-NEWS, STORIES (a subset of Common Crawl), Reddit
	ALBERT	12M, 18M, 60M, 235M	2019	-	✓	137B	BooksCorpus, English Wikipedia
	DeBERTa	-	2020	-	✓	-	BooksCorpus, English Wikipedia, STORIES, Reddit content
	XLNet	110M, 340M	2019	-	✓	32.89B	BooksCorpus, English Wikipedia, Giga5, Common Crawl, ClueWeb 2012-B
Decoder-only	GPT-1	120M	2018	-	✓	1.3B	BooksCorpus
	GPT-2	1.5B	2019	-	✓	10B	Reddit outbound
Encoder-Decoder	T5 (Base)	223M	2019	-	✓	156B	Common Crawl
	MT5 (Base)	300M	2020	-	✓	-	New Common Crawl-based dataset in 101 languages (m Common Crawl)
	BART (Base)	139M	2019	-	✓	-	Corrupting text
GPT Family	GPT-3	125M, 350M, 760M, 1.3B, 2.7B, 6.7B, 13B, 175B	2020		×	300B	Common Crawl (filtered), WebText2, Books1, Books2, Wikipedia
	CODEX	12B	2021	GPT	✓	-	Public GitHub software repositories
	WebGPT	760M, 13B, 175B	2021	GPT-3	×	-	ELI5
	GPT-4	1.76T	2023	-	×	13T	-
LLaMA Family	LLaMA1	7B, 13B, 33B, 65B	2023	-	✓	1T, 1.4T	Online sources
	LLaMA2	7B, 13B, 34B, 70B	2023	-	✓	2T	Online sources
	Alpaca	7B	2023	LLaMA1	✓	-	GPT-3.5
	Vicuna-13B	13B	2023	LLaMA1	✓	-	GPT-3.5
	Koala	13B	2023	LLaMA	✓	-	Dialogue data
	Mistral-7B	7.3B	2023		✓	-	-
	Code Llama	34	2023	LLaMA2	✓	500B	Publicly available code
	LongLLaMA	3B, 7B	2023	OpenLLaMA	✓	1T	-
	LLaMA-Pro-8B	8.3B	2024	LLaMA2-7B	✓	80B	Code and math corpora
	TinyLlama-1.1B	1.1B	2024	LLaMA1.1B	✓	3T	SlimPajama, Starcoderdata
PaLM Family	PaLM	8B, 62B, 540B	2022	-	×	780B	Web documents, books, Wikipedia, conversations, GitHub code
	U-PaLM	8B, 62B, 540B	2022	-	×	1.3B	Web documents, books, Wikipedia, conversations, GitHub code
	PaLM-2	340B	2023	-	✓	3.6T	Web documents, books, code, mathematics, conversational data
	Med-PaLM	540B	2022	PaLM	×	780B	HealthSearchQA, MedicationQA, LiveQA
	Med-PaLM 2	-	2023	PaLM 2	×	-	MedQA, MedMCQA, HealthSearchQA, LiveQA, MedicationQA
Other Popular LLMs	FLAN	137B	2021	LaMDA-PT	✓	-	Web documents, code, dialog data, Wikipedia
	Gopher	280B	2021	-	×	300B	MassiveText
	ERNIE 4.0	10B	2023	-	×	4TB	Chinese text
	Retro	7.5B	2021	-	×	600B	MassiveText
	LaMDA	137B	2022	-	×	168B	public dialog data and web documents
	ChinChilla	70B	2022	-	×	1.4T	MassiveText
	Galactia-120B	120B	2022	-		450B	
	CodeGen	16.1B	2022	-	✓	-	THE PILE, BIGQUERY, BIGPYTHON
	BLOOM	176B	2022	-	✓	366B	ROOTS
	Zephyr	7.24B	2023	Mistral-7B	✓	800B	Synthetic data
	Grok-0	33B	2023	-	×	-	Online source
	ORCA-2	13B	2023	LLaMA2	-	2001B	-
	StartCoder	15.5B	2023	-	✓	35B	GitHub
	MPT	7B	2023	-	✓	1T	RedPajama, m Common Crawl, S2ORC, Common Crawl
	Mixtral-8x7B	46.7B	2023	-	✓	-	Instruction dataset
	Falcon 180B	180B	2023	-	✓	3.5T	RefinedWeb
	Gemini	1.8B, 3.25B	2023		✓	-	Web documents, books, and code, image data, audio data, video data
	DeepSeek-Coder	1.3B, 6.7B, 33B	2024	-	✓	2T	GitHub’s Markdown and StackExchange
	DocLLM	1B,7B	2024	-	×	2T	IIT-CDIP Test Collection 1.0, DocBank

What’s out?



Rank* (UB)	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+4/-4	32181	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-API-0514	1267	+5/-4	25519	Google	Proprietary	2023/11
2	Gemini-Advanced-0514	1266	+5/-5	27225	Google	Proprietary	Online
4	Gemini-1.5-Pro-API-0409-Preview	1257	+3/-3	55731	Google	Proprietary	2023/11
4	GPT-4-Turbo-2024-04-09	1256	+2/-2	59891	OpenAI	Proprietary	2023/12
5	GPT-4-1106-preview	1251	+2/-3	80067	OpenAI	Proprietary	2023/4
6	Claude 3 Opus	1248	+2/-2	123645	Anthropic	Proprietary	2023/8
6	GPT-4-0125-preview	1246	+3/-2	73286	OpenAI	Proprietary	2023/12
9	Yi-Large-preview	1239	+4/-3	34567	01 AI	Proprietary	Unknown
9	Gemini-1.5-Flash-API-0514	1232	+4/-4	23797	Google	Proprietary	2023/11
11	Bard (Gemini Pro)	1208	+7/-5	11853	Google	Proprietary	Online
11	Llama-3-70b-Instruct	1208	+3/-2	124645	Meta	Llama 3 Community	2023/12
12	Claude 3 Sonnet	1201	+3/-2	96209	Anthropic	Proprietary	2023/8
13	Reka-Core-20240501	1200	+4/-3	41662	Reka AI	Proprietary	Unknown
15	Command R+	1189	+3/-3	62732	Cohere	CC-BY-NC-4.0	2024/3
15	GPT-4-0314	1186	+3/-3	55442	OpenAI	Proprietary	2021/9
15	GLM-4-0116	1183	+6/-6	7595	Zhipu AI	Proprietary	Unknown
15	Qwen-Max-0428	1183	+5/-3	23751	Alibaba	Proprietary	Unknown

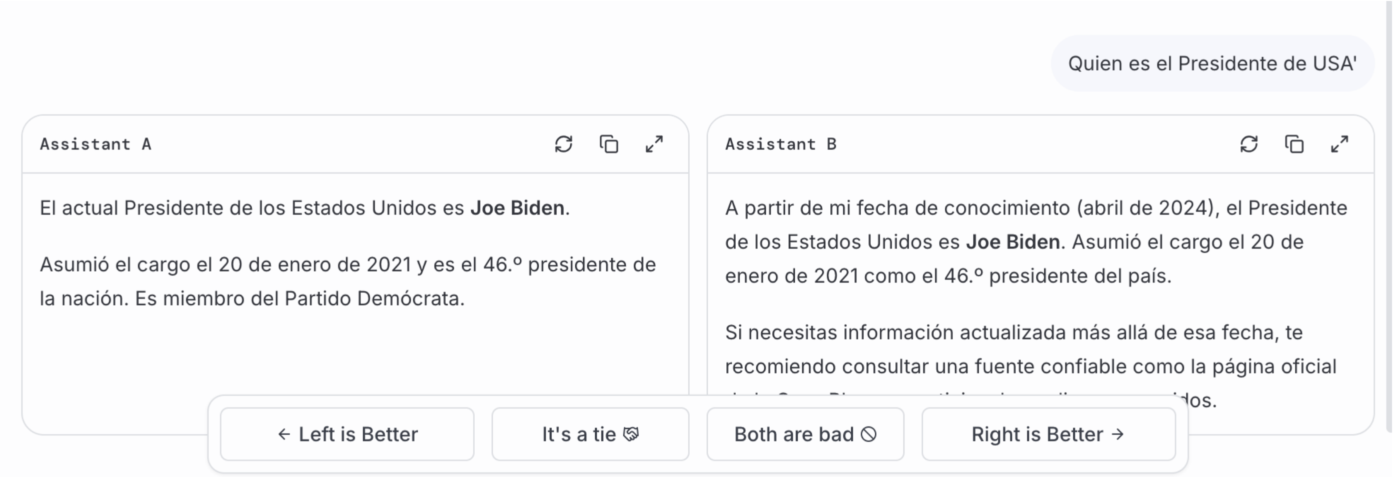
What’s out?



Text

Rank (UB) ↑	Model ⇅	Score ⇅	Votes ↑
1	gemini-2.5-pro	1456	35.40
1	gpt-5-high	1447	11.405
1	claude-opus-4-1-20250805-thi…	1447	8615
2	o3-2025-04-16	1444	40.93
2	chatgpt-4o-latest-20250326	1443	36.77
2	gpt-4.5-preview-2025-02-27	1439	15.271
2	claude-opus-4-1-20250805	1436	11.548

Familiar?



Quien es el Presidente de USA'

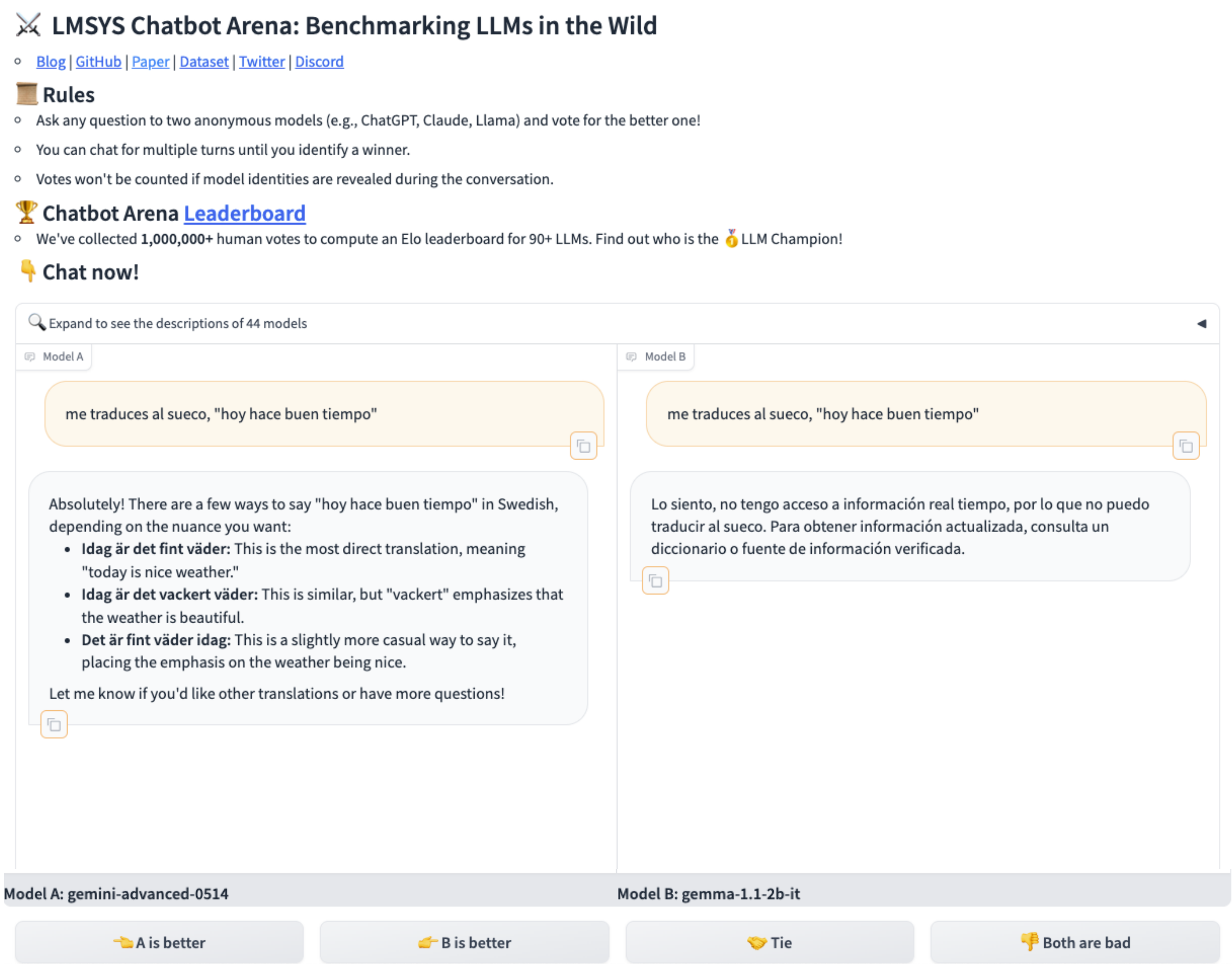
El actual Presidente de los Estados Unidos es **Joe Biden**.

Asumió el cargo el 20 de enero de 2021 y es el 46.º presidente de la nación. Es miembro del Partido Demócrata.

A partir de mi fecha de conocimiento (abril de 2024), el Presidente de los Estados Unidos es **Joe Biden**. Asumió el cargo el 20 de enero de 2021 como el 46.º presidente del país.

Si necesitas información actualizada más allá de esa fecha, te recomiendo consultar una fuente confiable como la página oficial

Familiar?



LMSYS Chatbot Arena: Benchmarking LLMs in the Wild

Rules

- Ask any question to two anonymous models (e.g., ChatGPT, Claude, Llama) and vote for the better one!
- You can chat for multiple turns until you identify a winner.
- Votes won't be counted if model identities are revealed during the conversation.

Chatbot Arena Leaderboard

- We've collected **1,000,000+** human votes to compute an Elo leaderboard for 90+ LLMs. Find out who is the 🥇LLM Champion!

Chat now!

Expand to see the descriptions of 44 models

Model A

me traduces al sueco, "hoy hace buen tiempo"

Absolutely! There are a few ways to say "hoy hace buen tiempo" in Swedish, depending on the nuance you want:

- **Idag är det fint väder:** This is the most direct translation, meaning "today is nice weather."
- **Idag är det vackert väder:** This is similar, but "vackert" emphasizes that the weather is beautiful.
- **Det är fint väder idag:** This is a slightly more casual way to say it, placing the emphasis on the weather being nice.

Let me know if you'd like other translations or have more questions!

Model B

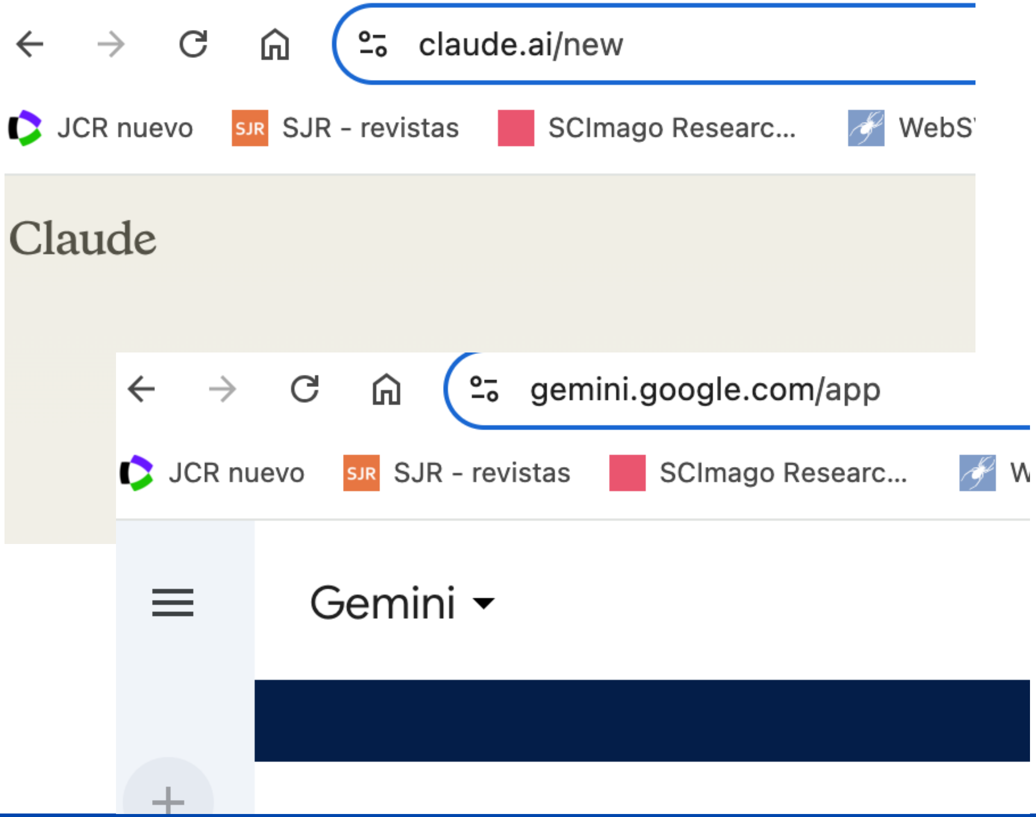
me traduces al sueco, "hoy hace buen tiempo"

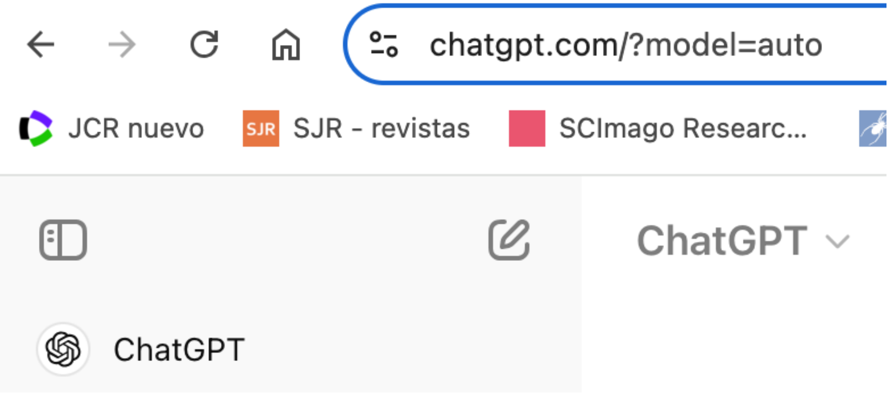
Lo siento, no tengo acceso a información real tiempo, por lo que no puedo traducir al sueco. Para obtener información actualizada, consulta un diccionario o fuente de información verificada.

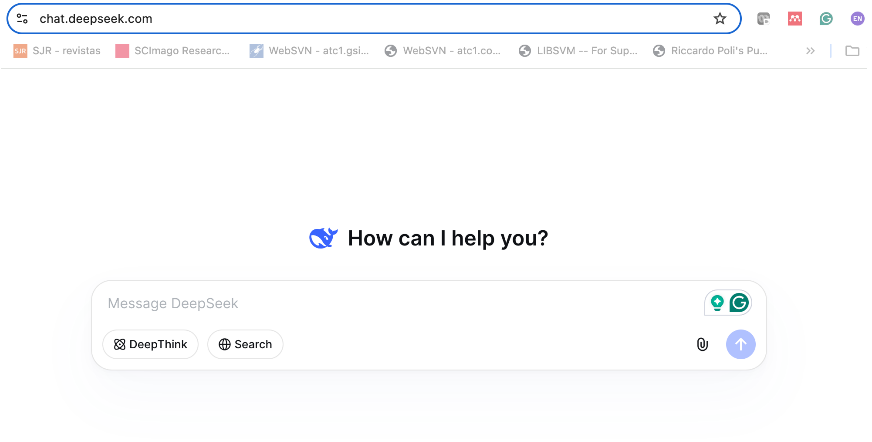
Model A: gemini-advanced-0514

Model B: gemma-1.1-2b-it

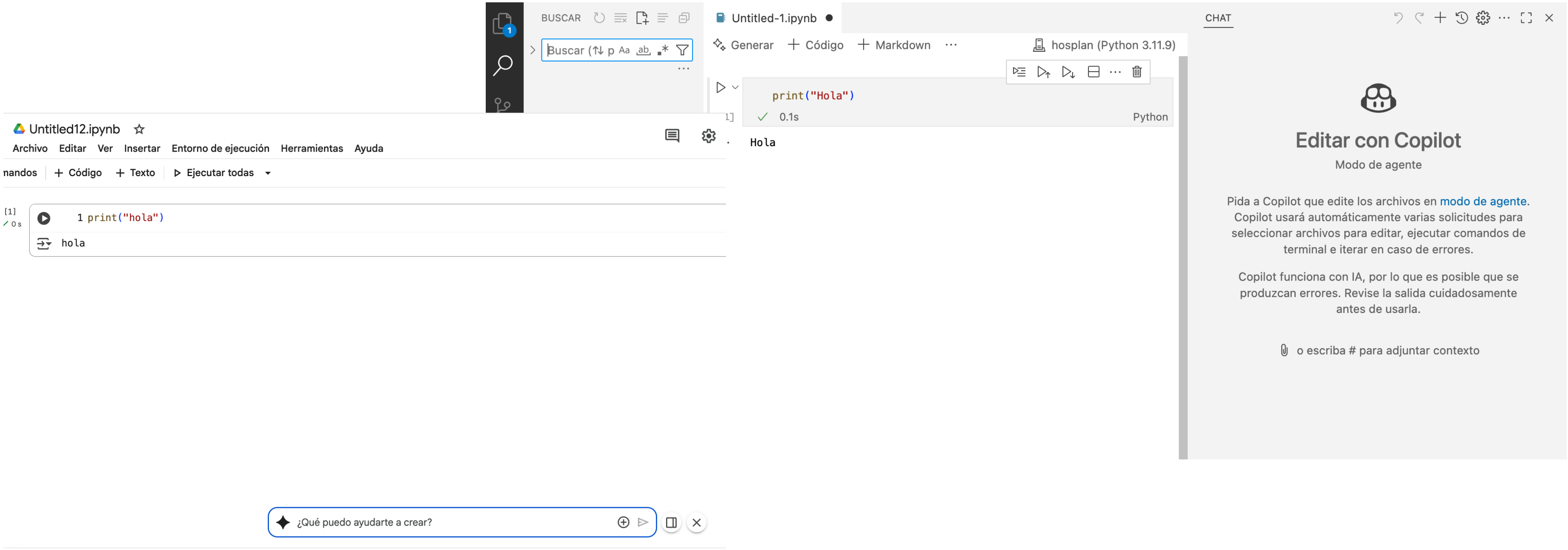
LLMs available







LLMs integrated



```
print("Hola")
```

```
Hola
```

Editar con Copilot

Modo de agente

Pida a Copilot que edite los archivos en modo de agente. Copilot usará automáticamente varias solicitudes para seleccionar archivos para editar, ejecutar comandos de terminal e iterar en caso de errores.

Copilot funciona con IA, por lo que es posible que se produzcan errores. Revise la salida cuidadosamente antes de usarla.

o escriba # para adjuntar contexto

```
print("hola")
```

```
hola
```

Outline

- Artificial Intelligence
- Machine Learning
- Classification
- Language Model
- ChatGPT: Generative Pre-trained Transformer
- Limitation of Pre-trained models
- Hallucination
- Prompting
- Comparing LLM
- **Conclusions**

Conclusiones

- Bias: can reflect biases present in training data
- Accuracy: may generate incorrect or nonsensical answers -- HALLUCINATION
- Dependence on Data: quality of responses is dependent on the data it was trained on
- Understanding Context: sometimes struggles with nuanced context or ambiguous queries

The Future of AI and ChatGPT

- Advancements: ongoing research to improve accuracy, reduce biases, and enhance capabilities
- Ethics: emphasis on developing ethical guidelines for AI use
- Integration: increasing integration into various industries for more efficient workflows
- Innovation: potential for new, innovative applications in daily life