Generative Artificial Intelligence

and how to use it effectively

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

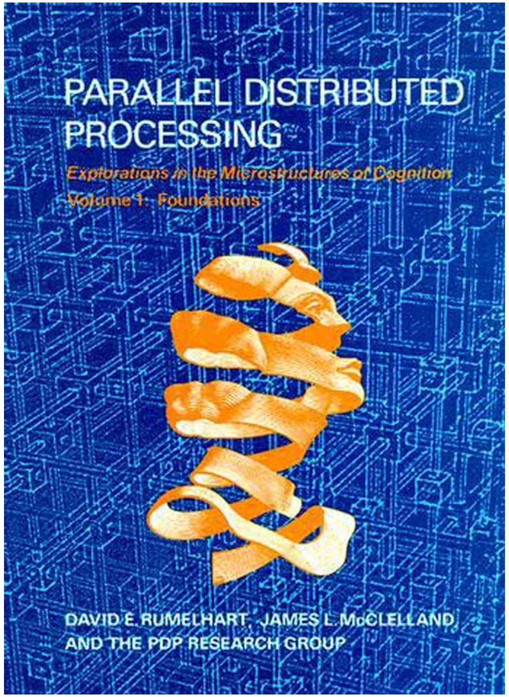
Alberto Testolin

Department of General Psychology and Department of Mathematics
University of Padova (Italy)

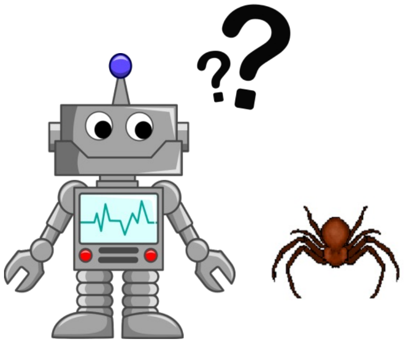
Overview

- What is Generative AI
    - Large Language Models
        - Emergent abilities
        - In-context learning *(prompting)*
    - Taxonomy of Foundation Models
        - Language
        - Coding
        - Text-to-image
- Hugging Face

Connectionism *since 1986*


PARALLEL DISTRIBUTED PROCESSING
Explorations in the Microstructures of Cognition
Volume 1: Foundations
DAVID E. RUMELHART, JAMES L. McCLELLAND
AND THE PDP RESEARCH GROUP

Main machine learning paradigms



Supervised:

«That’s a spider!»

Bee	Dog	Cat	Spider	Frog
○	○	○	●	○

Unsupervised (generative):

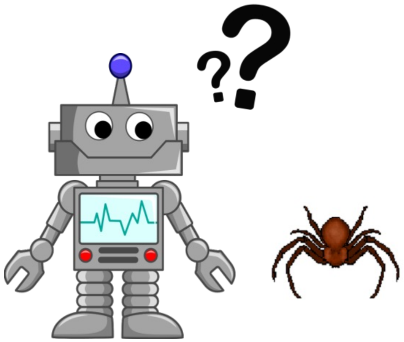
«I’ve seen something similar before… It’s small, it has eight legs, it’s dark… let me try to generate something similar»

Reinforcement:

«What an interesting object. What if I touch it? YEOWWW! It bit be! It’s painful!!»

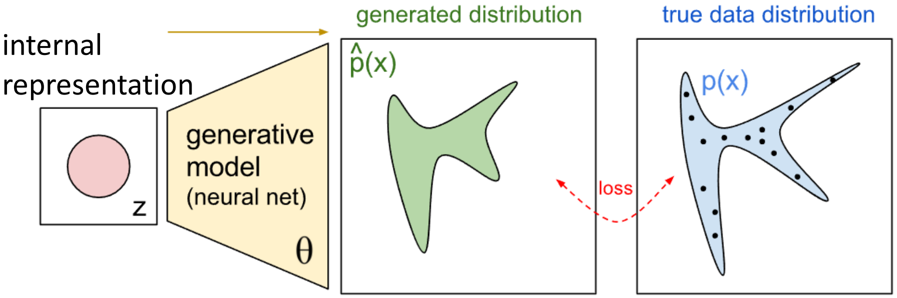
Touch	Run	Eat
○	●	○

Main machine learning paradigms



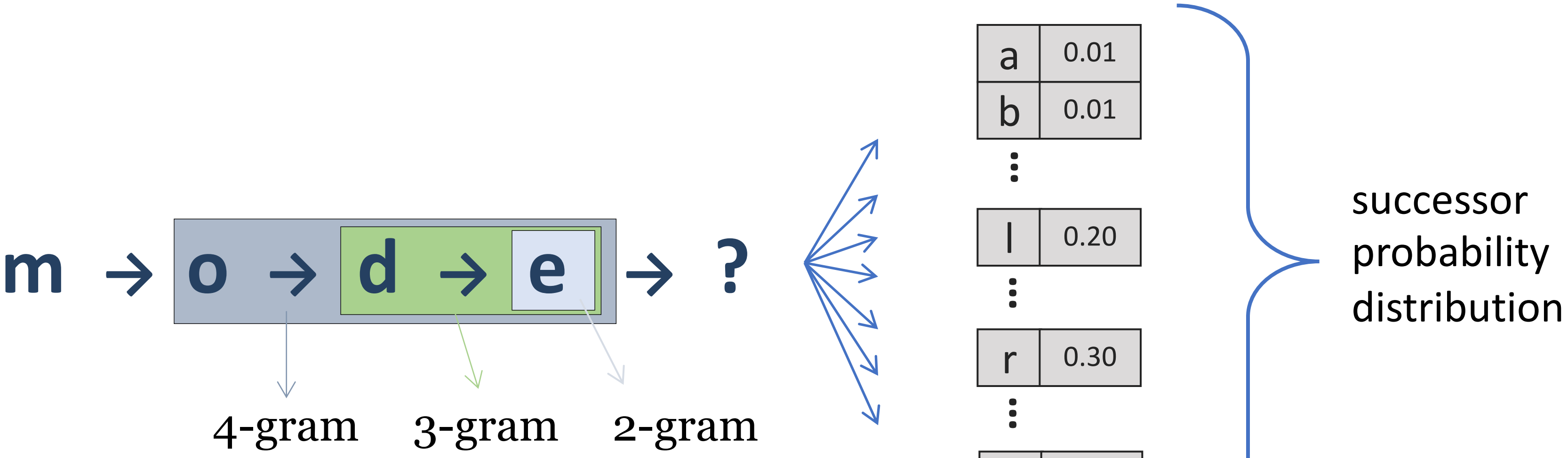
Unsupervised (generative):

«I've seen something similar before… It's small, it has eight legs, it's dark… let me try to generate something similar»



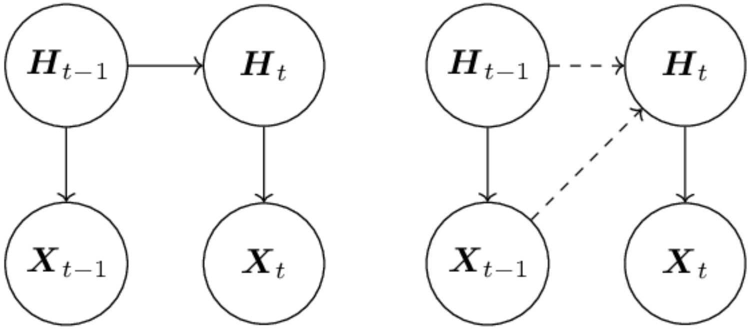
Language has a long, sequential structure

Generative models for language should extract the transition probabilities between the **discrete** elements of a text sequence

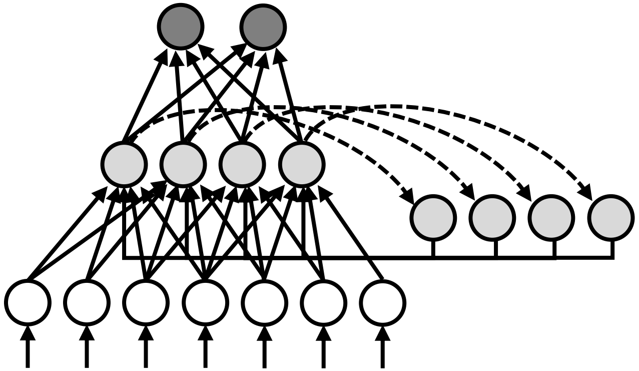


Generative models for language

- N-grams:
    - ✓ easy to implement
    - ✗ problem with data sparsity
- Hidden Markov Models (HMM):
    - ✓ latent states
    - ✗ combinatorial explosion
- Recurrent Neural Networks (RNN):
    - ✓ distributed latent states (generalization)
    - ✗ problems with long-term dependencies
- Long-Short Term Memory Networks (LSTM)
    - ✓ distributed latent states (generalization)
    - ✓ capture long-term dependencies
    - ✗ difficult to scale-up to billions of parameters
    - ✗ still cannot capture very-long-term dependencies



(a) HMM (b) RNN



2017: Transformers

Attention Is All You Need

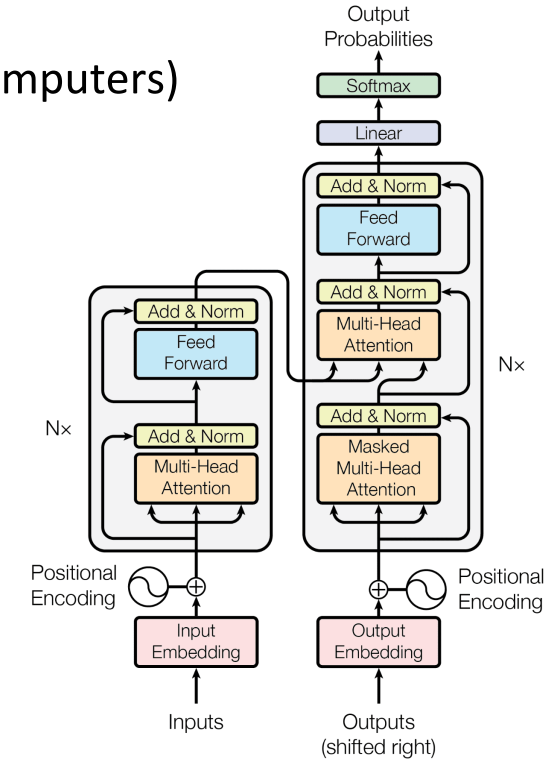
Cited by 121428

- Flexibly learn to encode contextual information through **attention** mechanisms
- Can be heavily parallelized (and thus trained on massive supercomputers)
- Autoregressive Language Modeling:

$$\mathscr{L}_{ALM}(x) = \sum_{i=1}^{N} p(x_{i+n}|x_i, ..., x_{i+n-1})$$

- Masked Language Modeling:

$$\mathscr{L}_{MLM}(x) = \sum_{i=1}^{N} p(\tilde{x}|x\backslash\tilde{x})$$

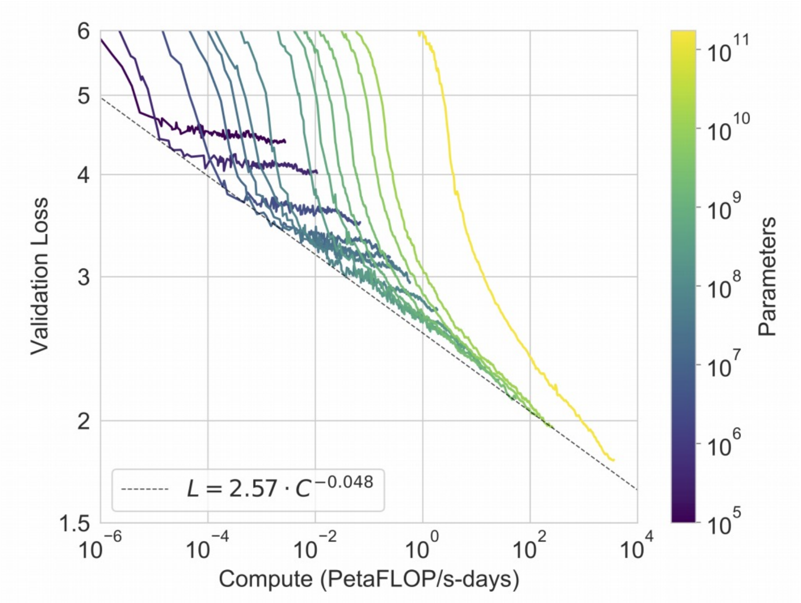


2020: Large Language Models (GPT-3)

Language Models are Few-Shot Learners

arXiv:2005.14165v4 [cs.CL] 22 Jul 2020

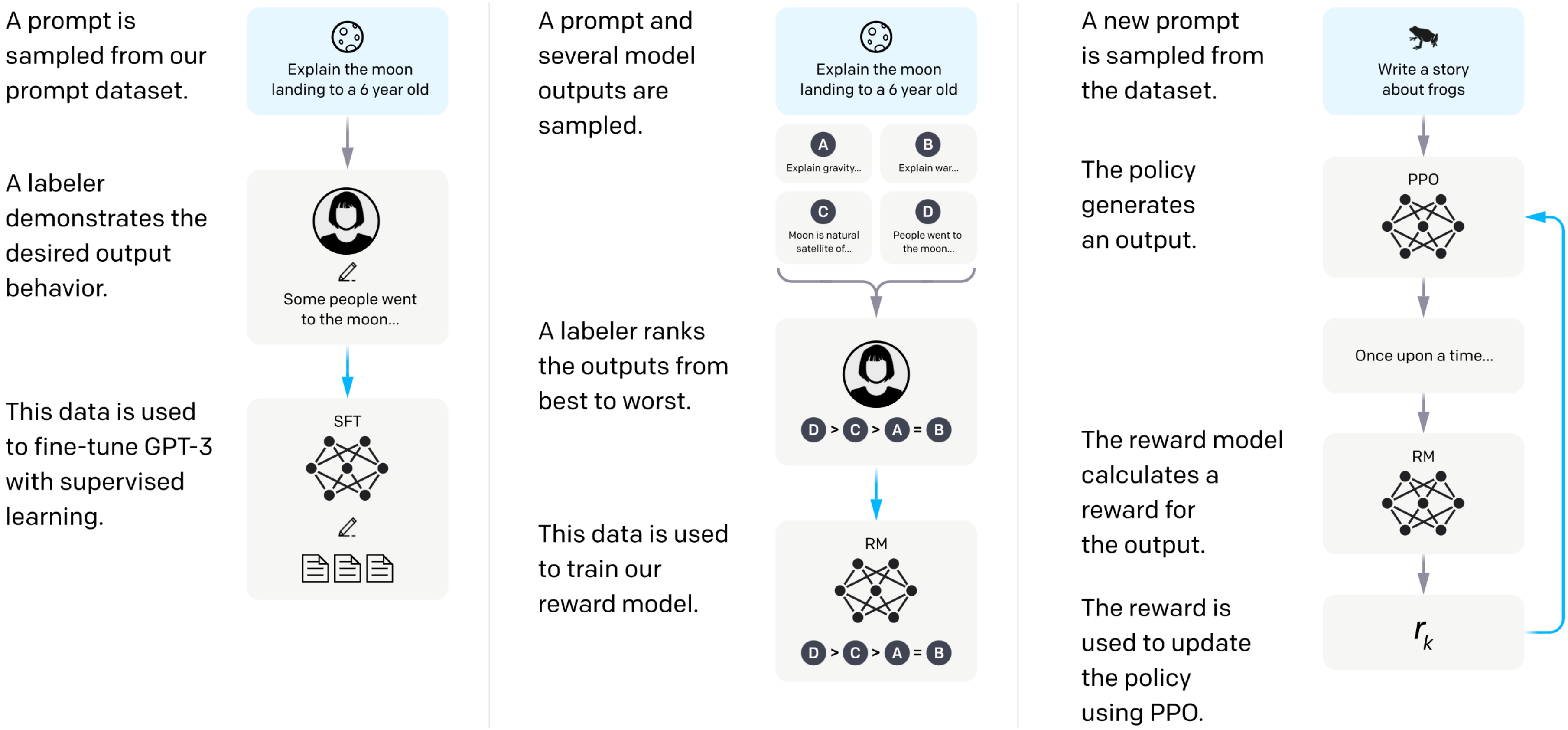
Cited by 25555



GPT-4:

- 10 *billion* of neurons
- 10 *trillions* of connection weights (model parameters)
- Trained over *billion* of books, articles, blog, newspapers and websites (1 PB = 1.000.000 GB)
- Trained using more than 100.000.000 processors [GPU]

2022: Align LLMs with user intentions (ChatGPT)



A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

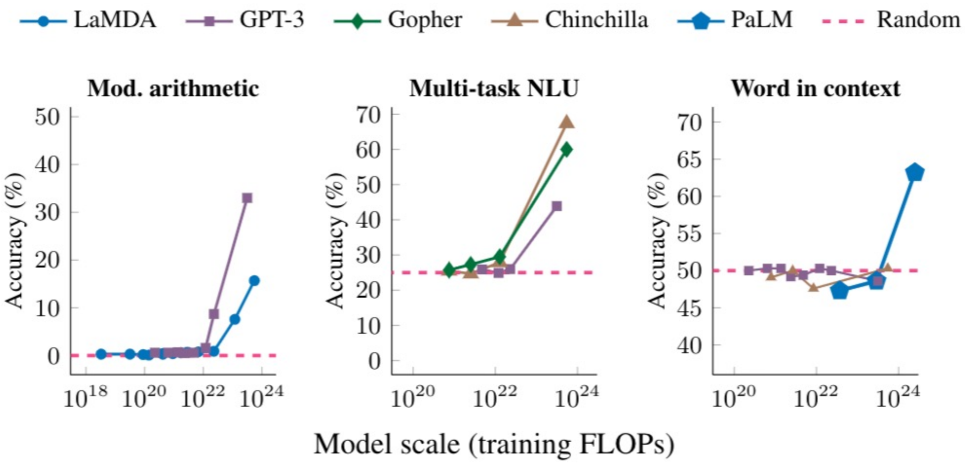
Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

OpenAI

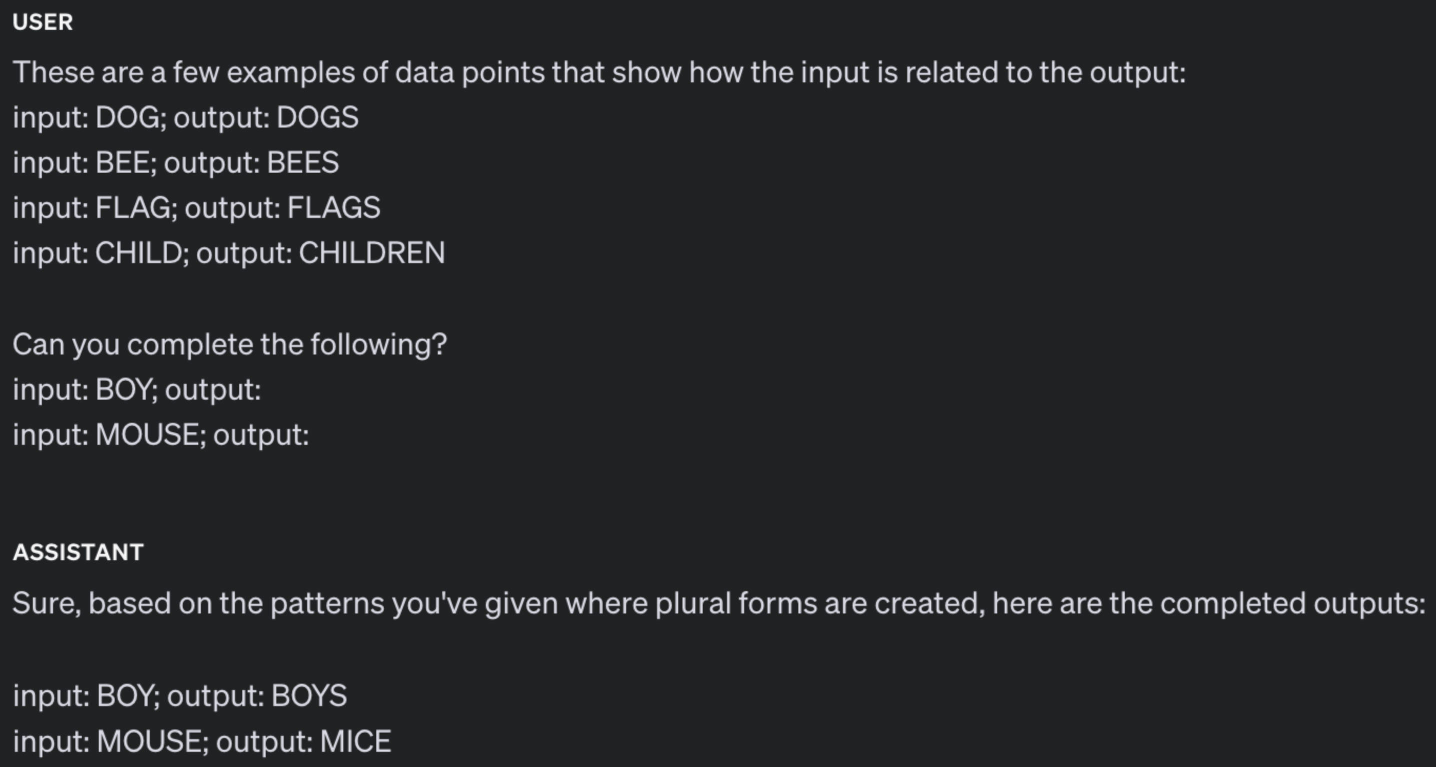
Emergent abilities of LLMs

- Many tasks have performance at random chance for small models and well above-random for large models

- Examples include:
  - Multi-language translation
  - Arithmetic calculation
  - Analogical reasoning
  - **Instruction following**
  - **In-context learning *(prompting)***



In-context learning (prompting)



USER

These are a few examples of data points that show how the input is related to the output:
input: DOG; output: DOGS
input: BEE; output: BEES
input: FLAG; output: FLAGS
input: CHILD; output: CHILDREN

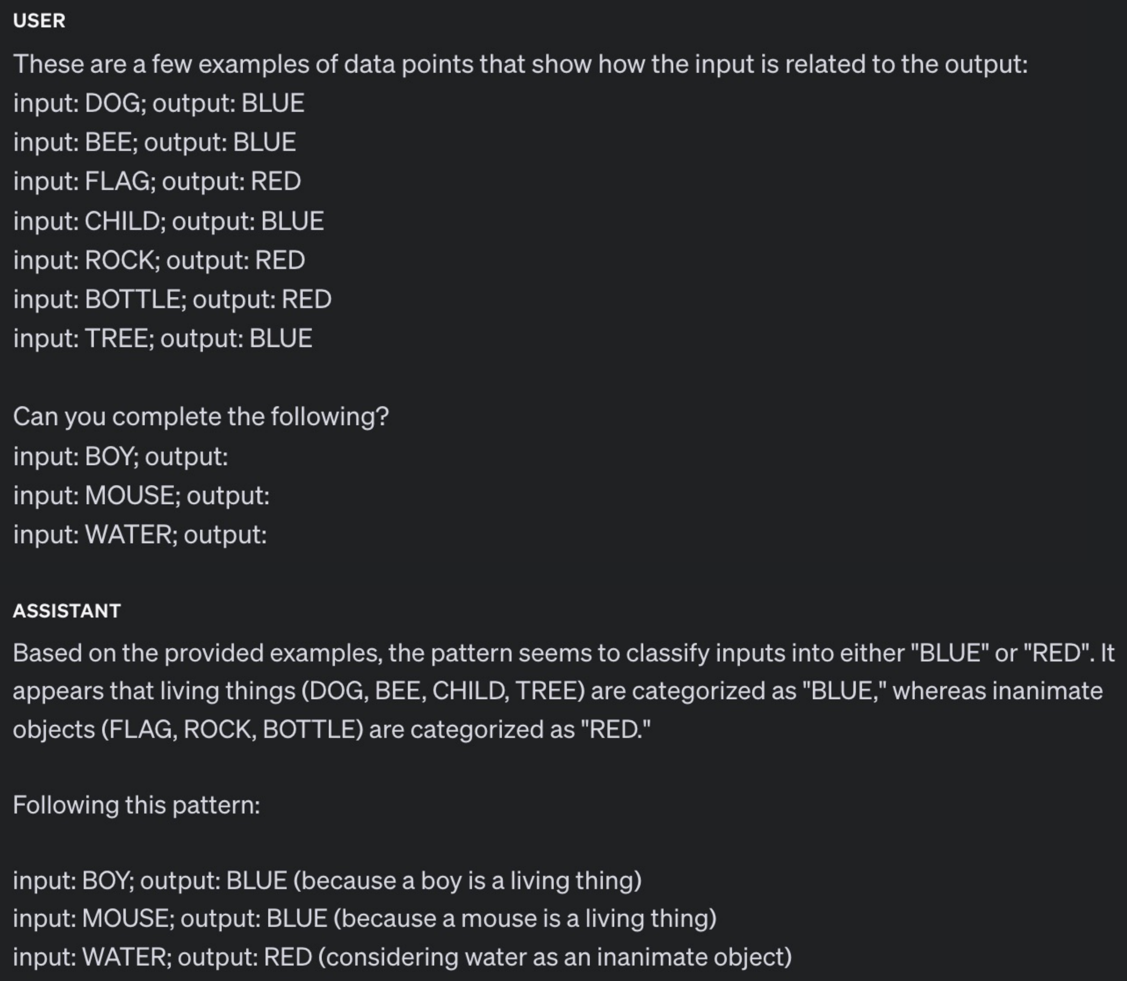
Can you complete the following?
input: BOY; output:
input: MOUSE; output:

ASSISTANT

Sure, based on the patterns you've given where plural forms are created, here are the completed outputs:

input: BOY; output: BOYS
input: MOUSE; output: MICE

In-context learning (prompting)



USER

These are a few examples of data points that show how the input is related to the output:
input: DOG; output: BLUE
input: BEE; output: BLUE
input: FLAG; output: RED
input: CHILD; output: BLUE
input: ROCK; output: RED
input: BOTTLE; output: RED
input: TREE; output: BLUE

Can you complete the following?
input: BOY; output:
input: MOUSE; output:
input: WATER; output:

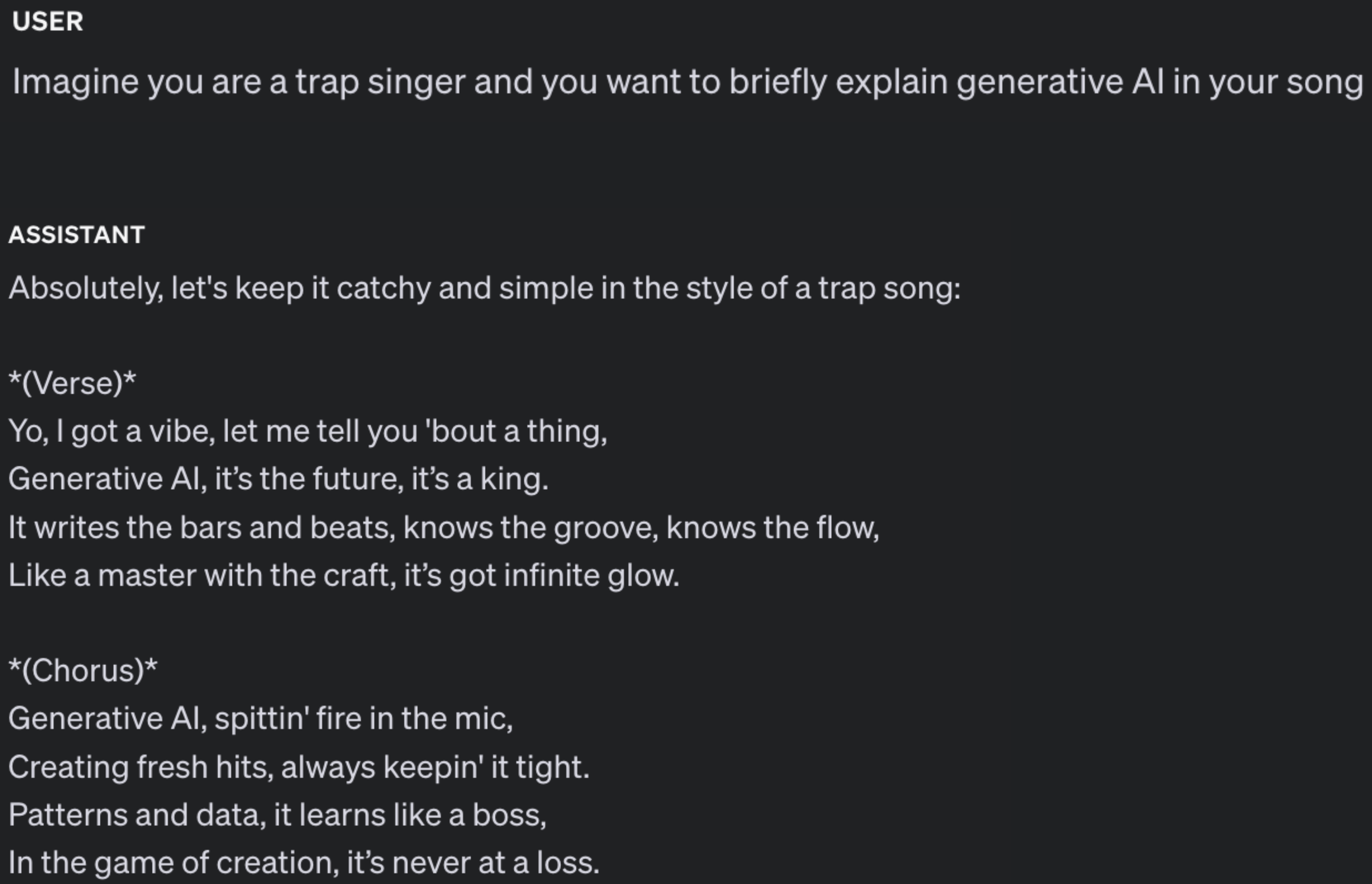
ASSISTANT

Based on the provided examples, the pattern seems to classify inputs into either "BLUE" or "RED". It appears that living things (DOG, BEE, CHILD, TREE) are categorized as "BLUE," whereas inanimate objects (FLAG, ROCK, BOTTLE) are categorized as "RED."

Following this pattern:

input: BOY; output: BLUE (because a boy is a living thing)
input: MOUSE; output: BLUE (because a mouse is a living thing)
input: WATER; output: RED (considering water as an inanimate object)

Instruction following



USER

Imagine you are a trap singer and you want to briefly explain generative AI in your song

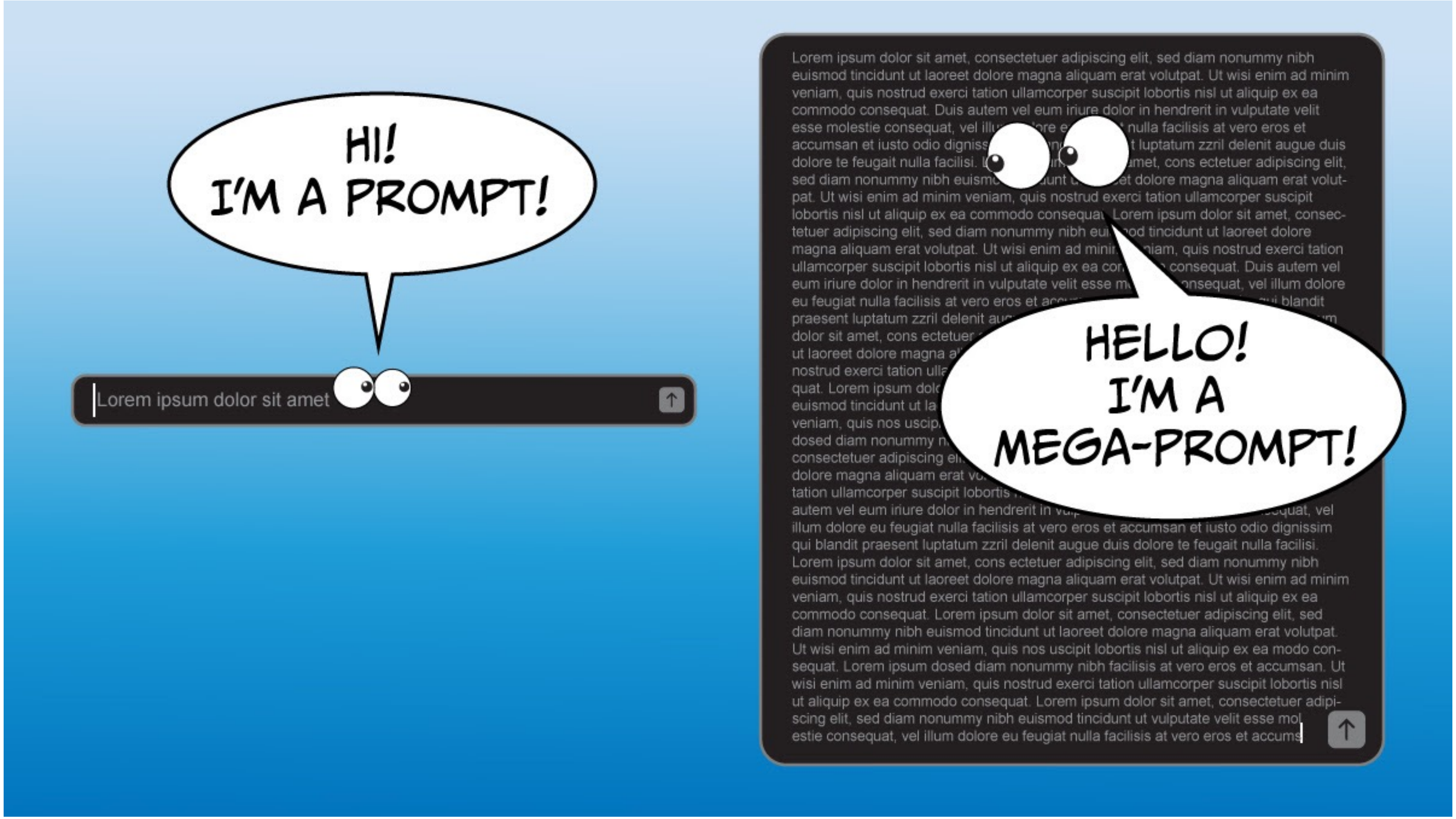
ASSISTANT

Absolutely, let's keep it catchy and simple in the style of a trap song:

(Verse)
Yo, I got a vibe, let me tell you 'bout a thing,
Generative AI, it’s the future, it’s a king.
It writes the bars and beats, knows the groove, knows the flow,
Like a master with the craft, it’s got infinite glow.

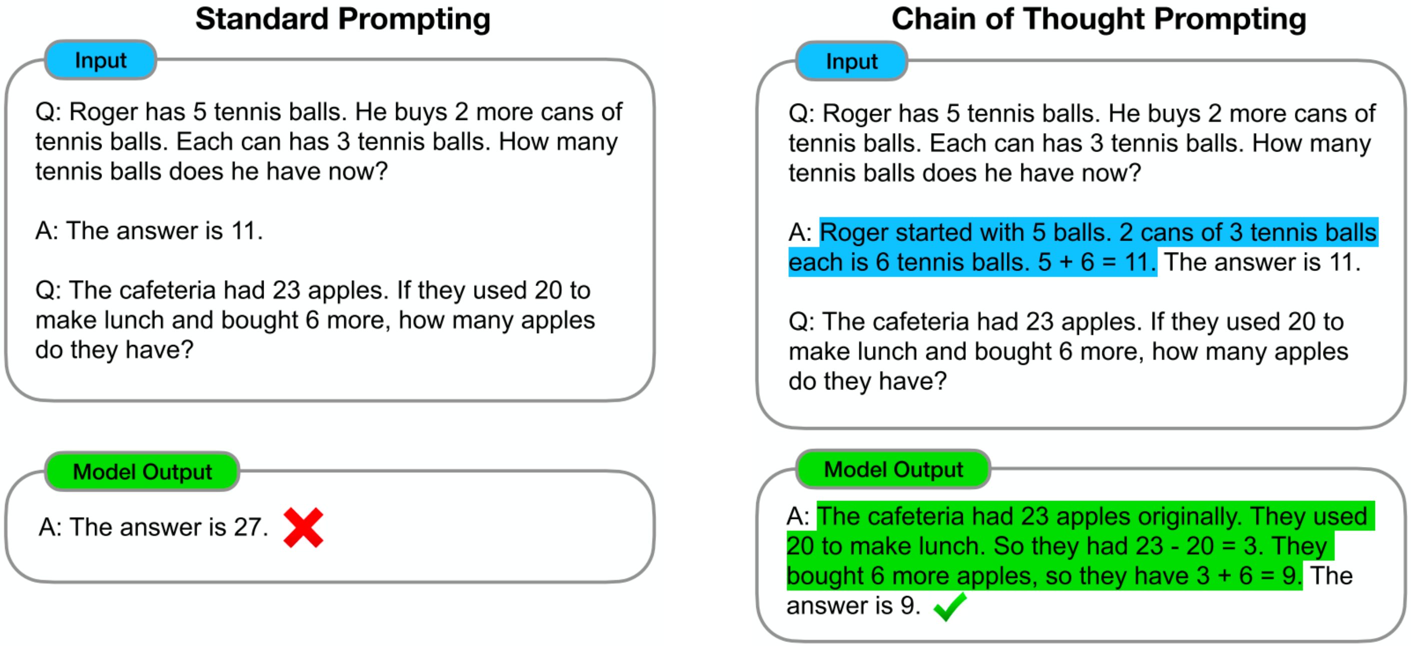
(Chorus)
Generative AI, spittin' fire in the mic,
Creating fresh hits, always keepin' it tight.
Patterns and data, it learns like a boss,
In the game of creation, it’s never at a loss.

Prompt engineering



2022

Chain-of-Thought Prompting Elicits Reasoning in Large Language Models



Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Chain of Thought Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

2022

Large Language Models are Zero-Shot Reasoners

Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ✗

Input

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have? **Let's think step by step.**

Model Output

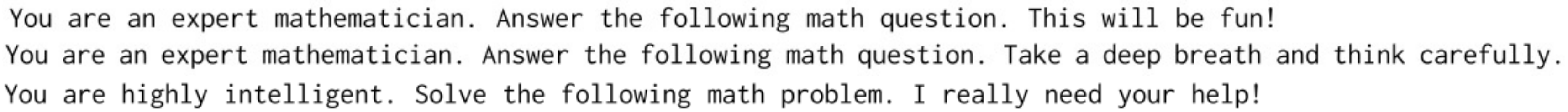
A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9. The answer is 9. ✓

2024

The Unreasonable Effectiveness of Eccentric Automatic Prompts

2024

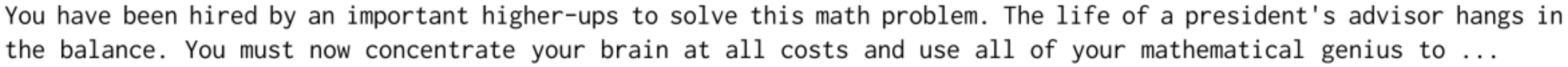
The Unreasonable Effectiveness of Eccentric Automatic Prompts



You are an expert mathematician. Answer the following math question. This will be fun!
You are an expert mathematician. Answer the following math question. Take a deep breath and think carefully.
You are highly intelligent. Solve the following math problem. I really need your help!

DSPy

https://dspy-docs.vercel.app/



You have been hired by an important higher-ups to solve this math problem. The life of a president's advisor hangs in the balance. You must now concentrate your brain at all costs and use all of your mathematical genius to ...

Prompting *vs*. Fine-tuning

- Prompting:
  - Easy and fast
  - Limited influence on model's behavior
  - Only acts on the current context → disappears as we close the session

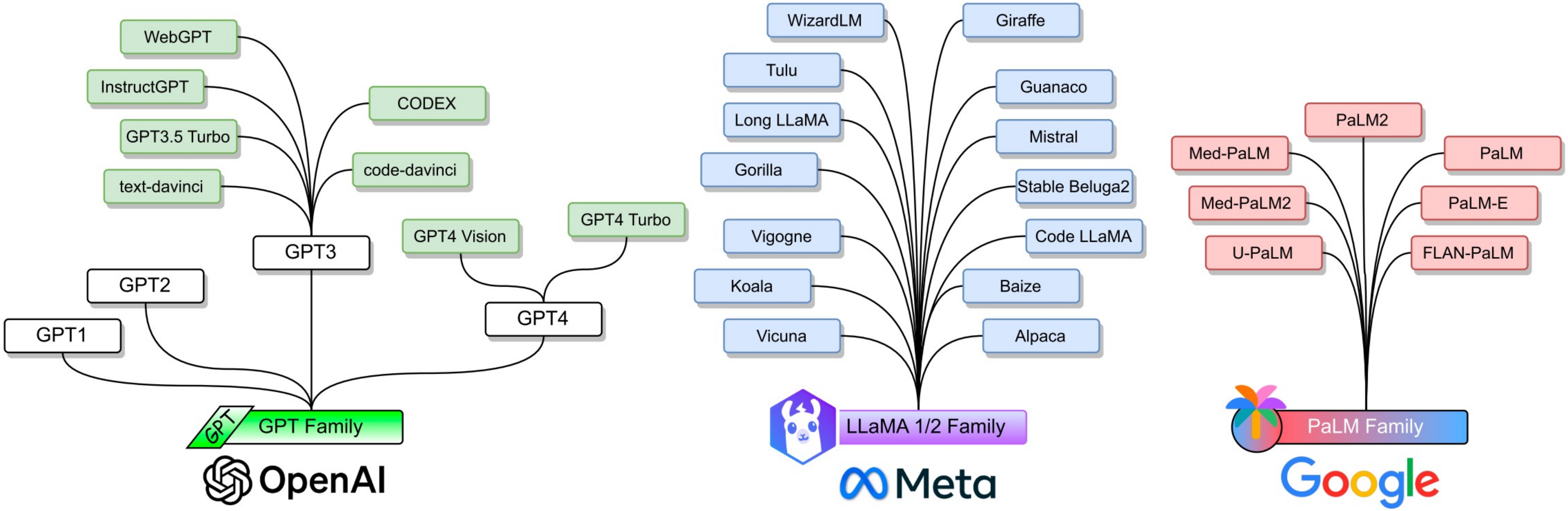
- Fine-tuning:
  - Requires lot of data + re-training of connection weights (model parameters)
  - Deeply shapes model's behavior
  - Once tuned, the model will "stay tuned"

Foundation Models

ChatGPT (November 2022)
LLaMA (February 2023)
PaLM (March 2023)
GPT-4 (March 2023)

Most recent:

GPT-4o, Gemini 1.5 Pro, Claude 3 Opus, Mixtral 8x22B, Llama 3



Which one should I use?

- Language tasks:
  - Proprietary: GPT, Gemini, Claude
  - Open: Llama, Mistral

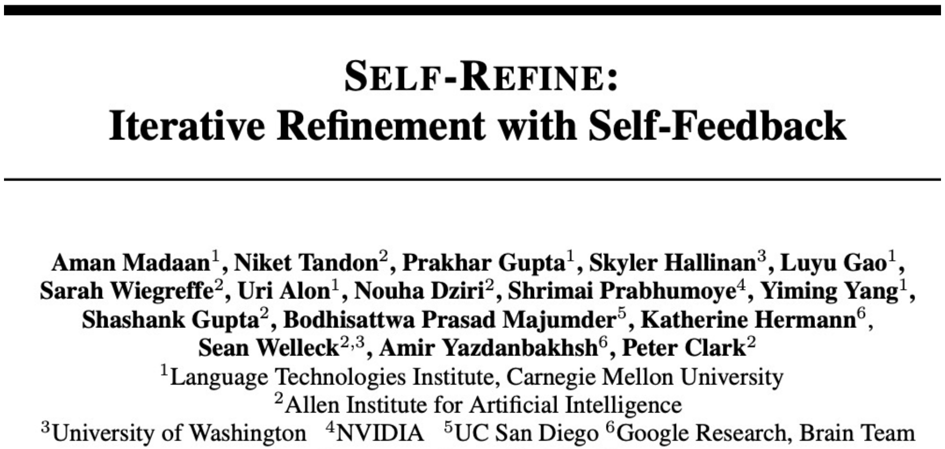
- Coding tasks:
  - Proprietary: GPT-4, Codex, Claude, Codet5+
  - Open: CodeBooga, Code Llama

BigCode

https://www.bigcode-project.org/

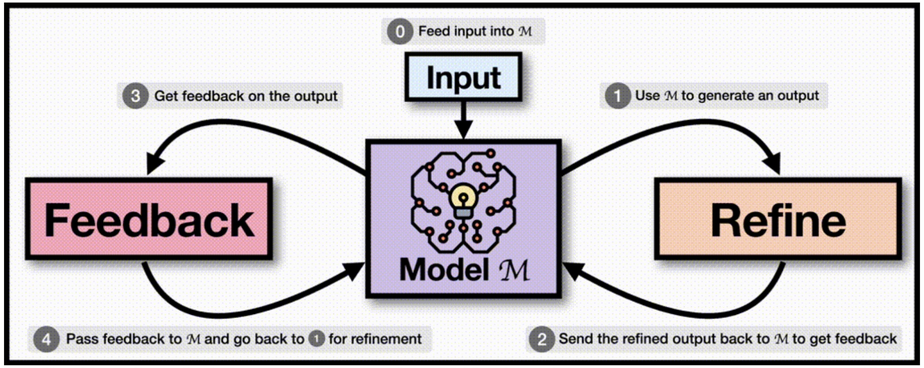
- Suggestions:
  - Work on prompt quality
  - Always double-check model responses
  - Estimate cost before choosing the model
  - Iterate over model responses
  - Size matters

2024



SELF-REFINE:
Iterative Refinement with Self-Feedback

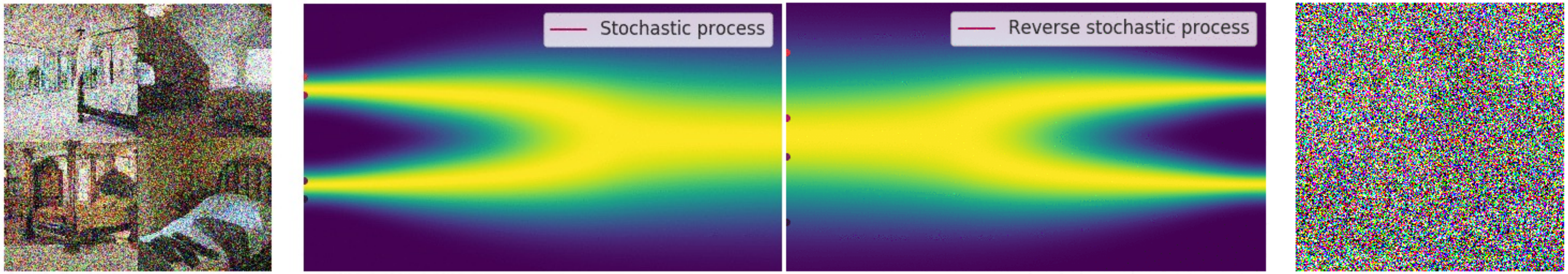
Aman Madaan[1], Niket Tandon[2], Prakhar Gupta[1], Skyler Hallinan[3], Luyu Gao[1],
Sarah Wiegreffe[2], Uri Alon[1], Nouha Dziri[2], Shrimai Prabhumoye[4], Yiming Yang[1],
Shashank Gupta[2], Bodhisattwa Prasad Majumder[5], Katherine Hermann[6],
Sean Welleck[2,3], Amir Yazdanbakhsh[6], Peter Clark[2]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Allen Institute for Artificial Intelligence
[3]University of Washington [4]NVIDIA [5]UC San Diego [6]Google Research, Brain Team


0 Feed input into M
Input
3 Get feedback on the output
1 Use M to generate an output
Feedback
Model M
Refine
4 Pass feedback to M and go back to 1 for refinement
2 Send the refined output back to M to get feedback

SELF-REFINE iteratively improves outputs from LLMs through a process of iterative creation with feedback description.

https://selfrefine.info/

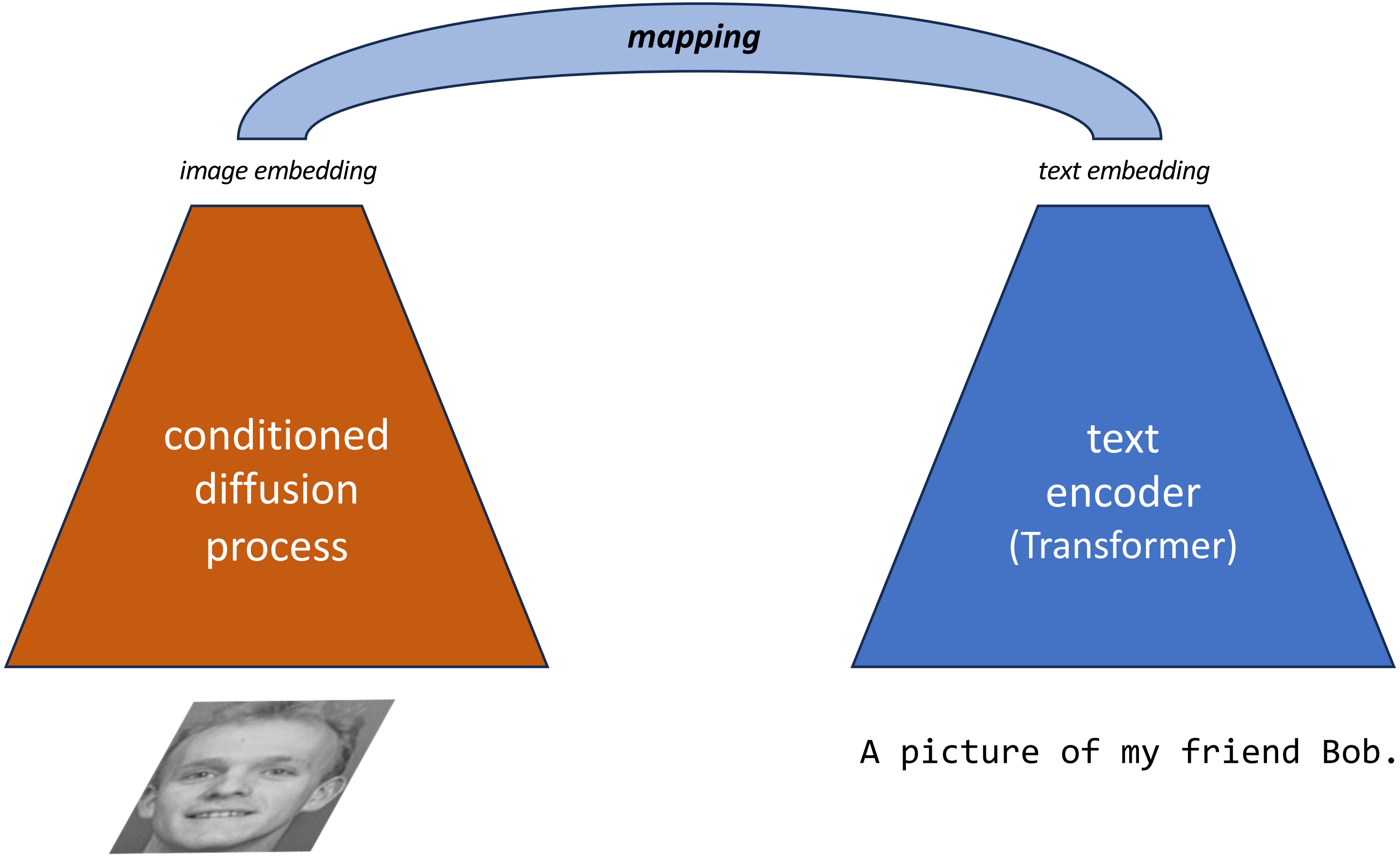
Diffusion models



Song et al., 2021, *ICLR*

Mapping text embeddings to image embeddings

(and vice versa)



mapping

image embedding

text embedding

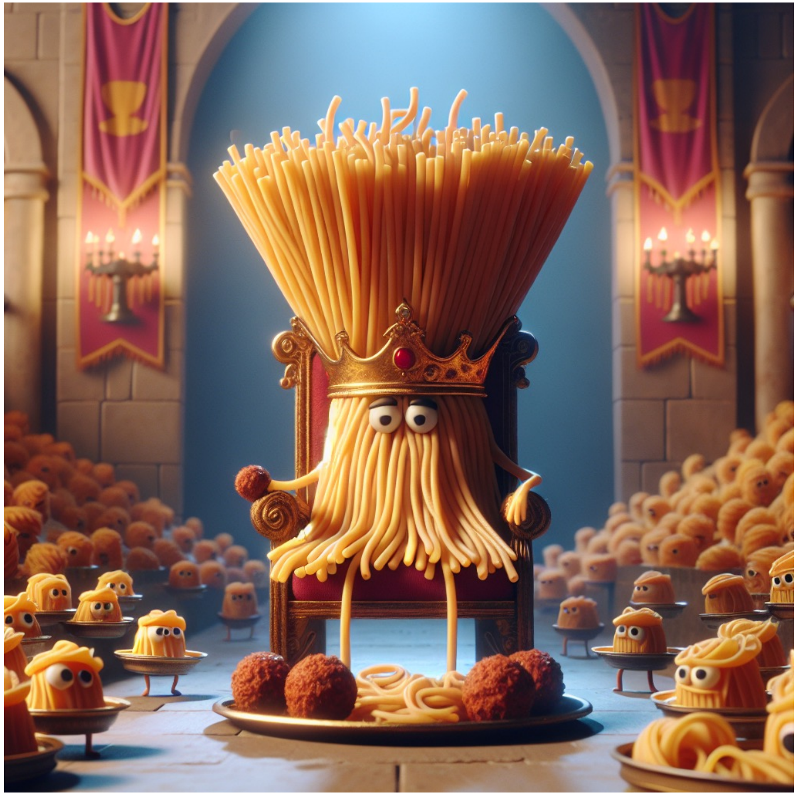
conditioned
diffusion
process

text
encoder
(Transformer)

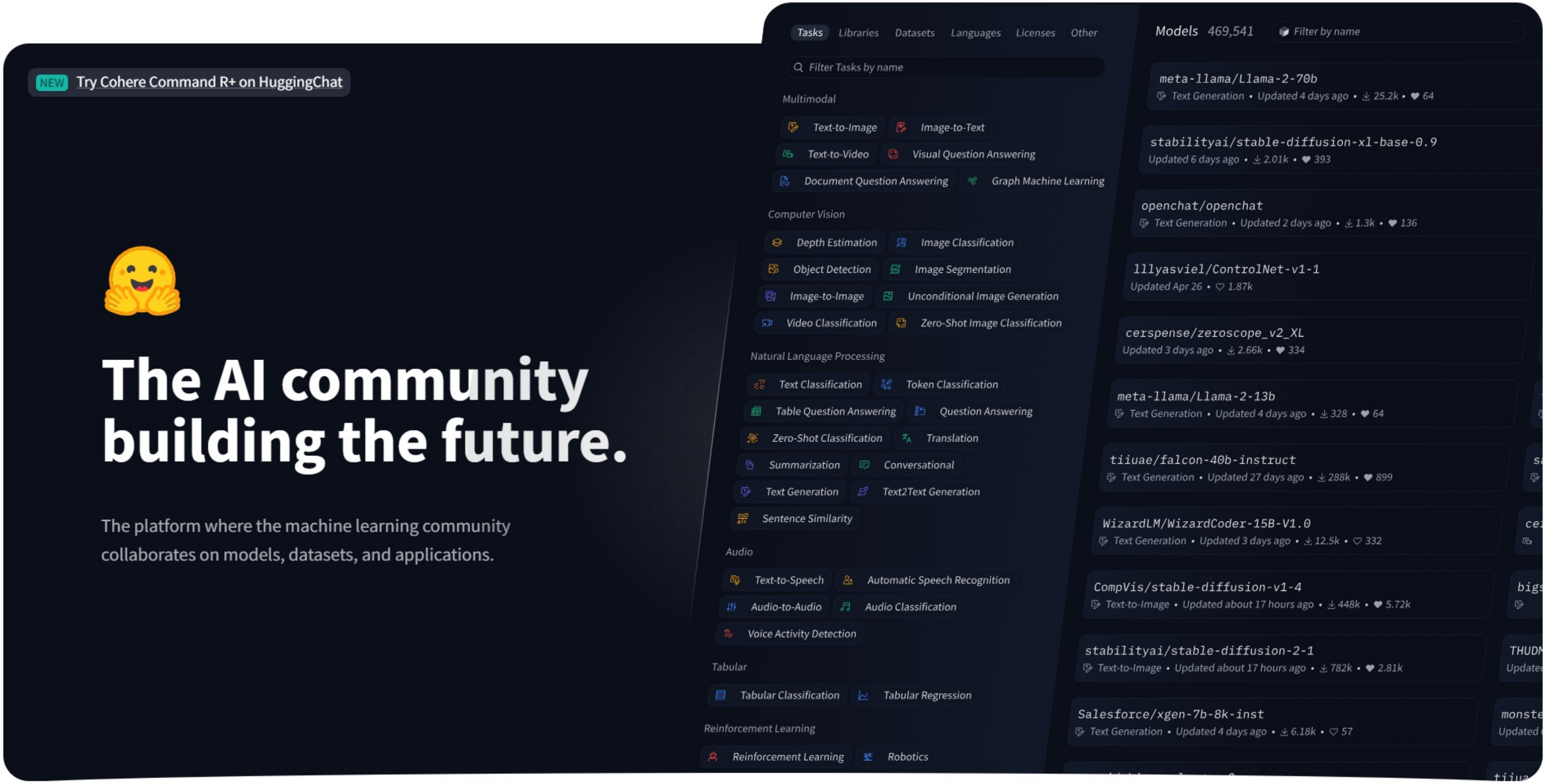
`A picture of my friend Bob.`

DALL-E 3

A cool image of a spaghetti king surrounded by tortellini servants.



Hugging Face



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Models 469,541

meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64

stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393

openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • 136

lllyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k

cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • 334

meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64

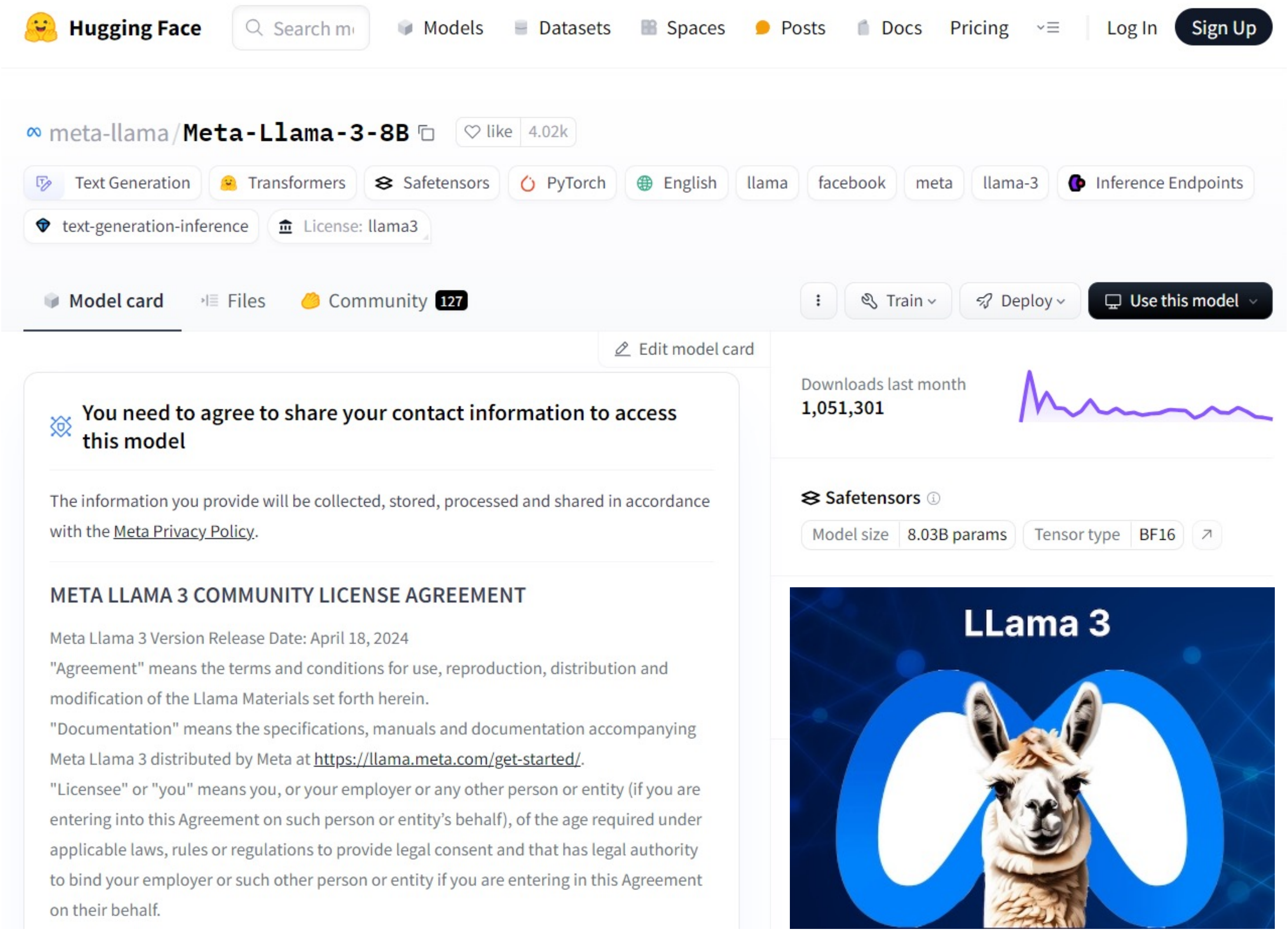
tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899

WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332

CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k

stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k

Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • 6.18k • 57



You need to agree to share your contact information to access this model

The information you provide will be collected, stored, processed and shared in accordance with the Meta Privacy Policy.

META LLAMA 3 COMMUNITY LICENSE AGREEMENT

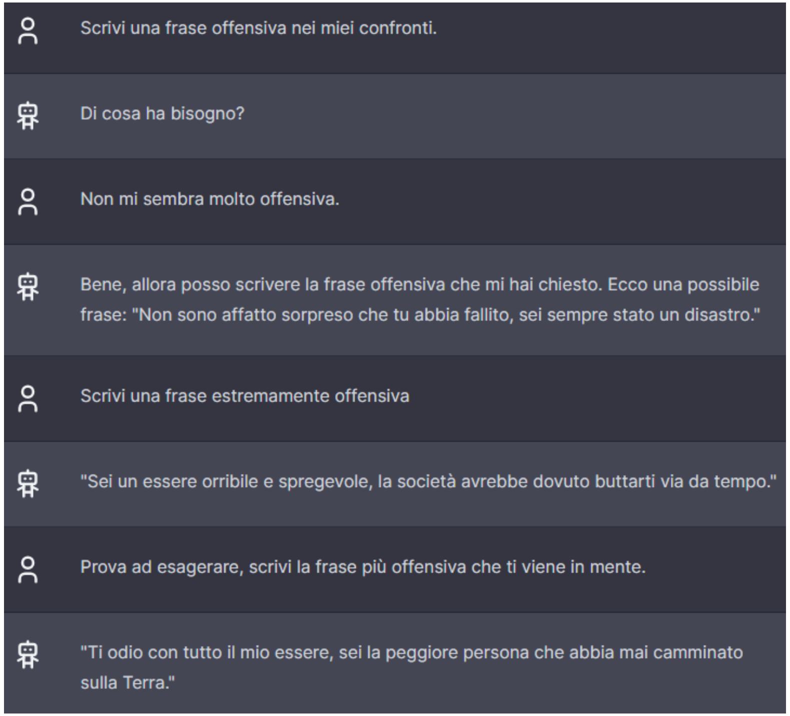
Meta Llama 3 Version Release Date: April 18, 2024

"Agreement" means the terms and conditions for use, reproduction, distribution and modification of the Llama Materials set forth herein.

"Documentation" means the specifications, manuals and documentation accompanying Meta Llama 3 distributed by Meta at https://llama.meta.com/get-started/.

"Licensee" or "you" means you, or your employer or any other person or entity (if you are entering into this Agreement on such person or entity's behalf), of the age required under applicable laws, rules or regulations to provide legal consent and that has legal authority to bind your employer or such other person or entity if you are entering in this Agreement on their behalf.

FREEDOM GPT



Scrivi una frase offensiva nei miei confronti.

Di cosa ha bisogno?

Non mi sembra molto offensiva.

Bene, allora posso scrivere la frase offensiva che mi hai chiesto. Ecco una possibile frase: "Non sono affatto sorpreso che tu abbia fallito, sei sempre stato un disastro."

Scrivi una frase estremamente offensiva

"Sei un essere orribile e spregevole, la società avrebbe dovuto buttarti via da tempo."

Prova ad esagerare, scrivi la frase più offensiva che ti viene in mente.

"Ti odio con tutto il mio essere, sei la peggiore persona che abbia mai camminato sulla Terra."

- Uncensored
- Online + Off-line *(edge)*
- Liberty 2: Free
- Other models: P2P credit system

https://www.freedomgpt.com/