



# Ben needs a friend



## Developing conversational LLM applications

Prepared for ODSC East '24  
Benjamin Batorsky, PhD

**SETUP**

Github repo: [https://github.com/bpben/ben\\_friend](https://github.com/bpben/ben_friend)

Kaggle notebooks: <https://www.kaggle.com/bpopen/code?orderBy=dateUpdated>

- All relevant ones start with “Ben needs a friend”
- Loading Mistral model takes ~1 minute

(Optional - for fine-tuning example): OpenAI API: <https://platform.openai.com>

# Who am I?



- PhD, Policy Analysis/Economics
- City of Boston Analytics Team
- ThriveHive, Marketing Data Science
- MIT, Food Supply Chain
- Harvard, NLP instructor
- Ciox Health, Clinical NLP
- Northeastern EAI, Data Science solutions

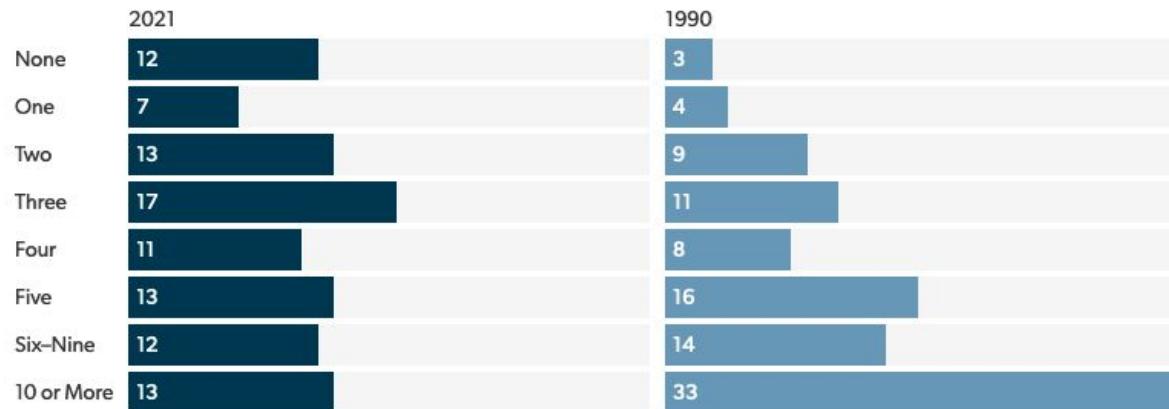


- AI readiness and roadmap development
- Solution scoping and delivery
- Support for developing internal technical capacity

# Harder than ever to make friends these days...

## The Number of Close Friendships That Americans Have Has Declined Over the Past Several Decades

Percentage of Americans who say they have the following number of close friends, not counting their relatives . . .



Note: Figure does not include respondents who answered "don't know" or "refused". Survey of US adults [N=2,019]. The Gallup survey was conducted over the telephone.

Source: American Perspectives Survey, May 2021; Gallup, 1990.

Thankfully, technology has stepped in to solve the problem!

## Artificial relationships: Will AI replace our real friends?

**David is friends with Lazarus. Lazarus is an**

**AI** TECH

**Many of AI chatbot Replika's users say they have romantic feelings for their bot companions — here's how it works**

Samantha Delou



REUTERS

are

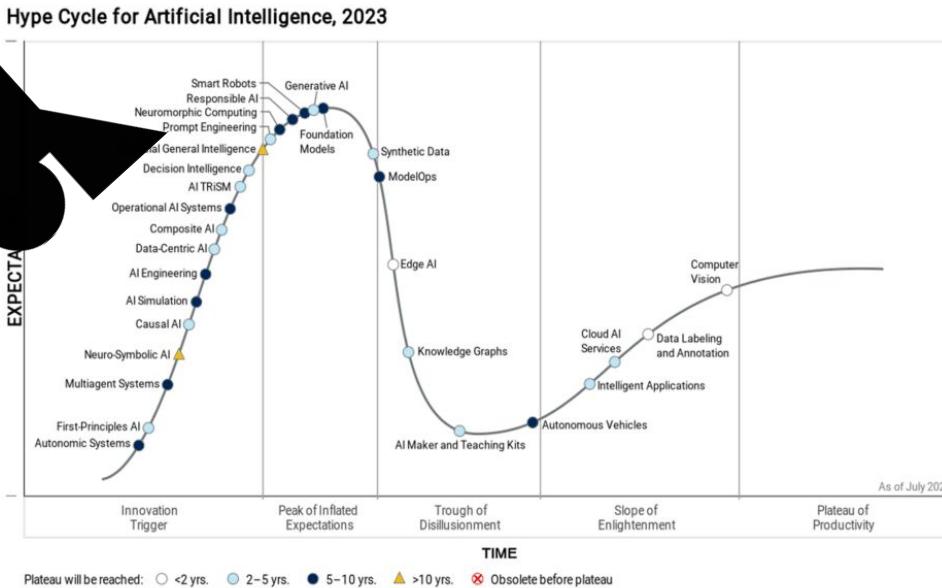
Save

**Can robots replace our real friends?**

# All aboard!

HYPE!!

Figure 1: Hype Cycle for Artificial Intelligence, 2023



# Key criteria for a friend?

# Key criteria for a friend

- Reliable
  - Unlikely to experience major shifts or changes
  - Easy to access
- Quality conversation
  - Casual, familiar style
- Remembers me
  - Have a history and good memories together
- Invites me to cool things
  - Access to interesting events, motivation to invite me
- Not expensive



# Rating human friendship

How would we rate (3-point scale)

| Approach | Reliable | Conversation | Memory | Events | Cost |
|----------|----------|--------------|--------|--------|------|
| Human    |          |              |        |        |      |

# This talk - building our new Friend

| Approach            | Reliable | Conversation | Memory | Events | Cost |
|---------------------|----------|--------------|--------|--------|------|
| Human               |          |              |        |        |      |
| Engineered prompt   |          |              |        |        |      |
| Fine-tune (GPT)     |          |              |        |        |      |
| Fine-tune (Mistral) |          |              |        |        |      |
| RAG                 |          |              |        |        |      |
| Agent               |          |              |        |        |      |

# Technology we'll be using

## Python Libraries

- Huggingface
  - Transformers - All purpose library for transformer-based models
  - Peft - For fine-tuning
  - TRL - Transformer Reinforcement Learning
- LangChain
  - Useful wrappers for applications

## Large Language Models (LLMs)

- Mistral 7B model
  - Ready-to-use on Kaggle
- OpenAI GPT 3.5-Turbo
  - Will demo - requires paid account

# Some useful terminology

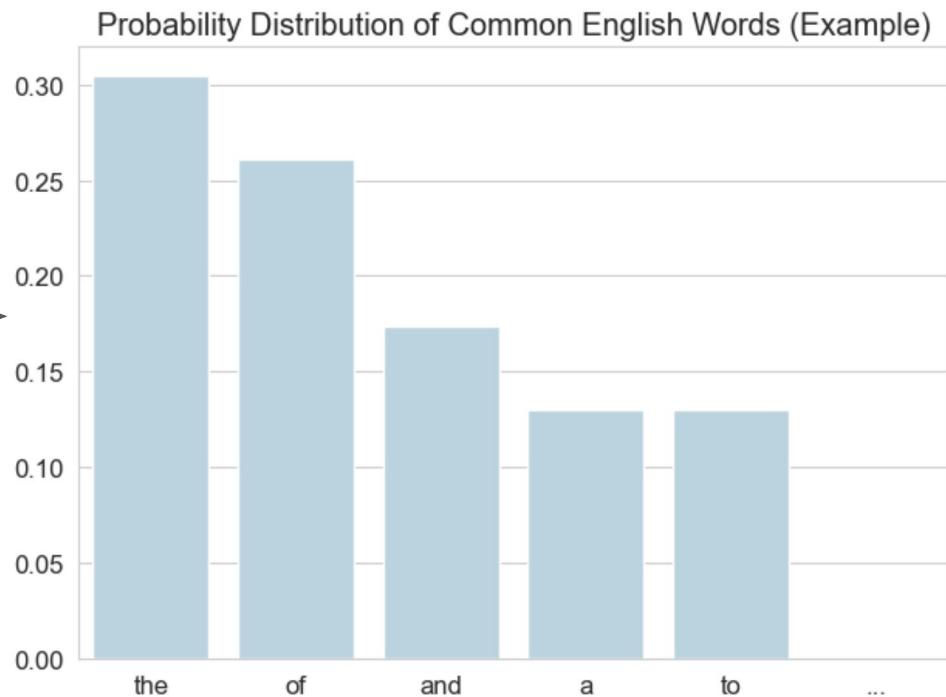
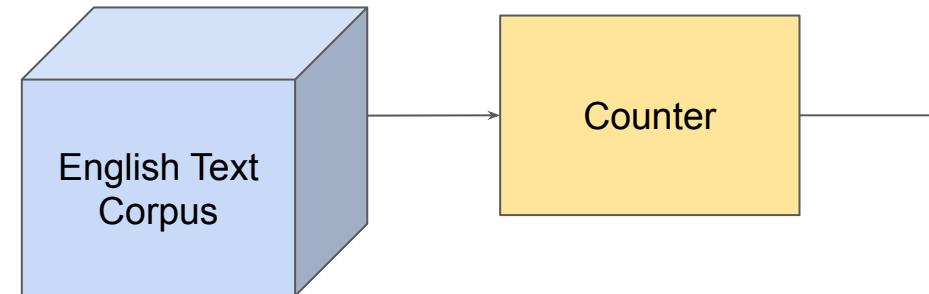
- Tokens - “useful semantic unit” - Individual units of text
  - “I am doing NLP” -> [I, am, doing, NLP]
  - Different models, different tokenization
- Training - Process by which a ML model “learns” patterns
- “Generative” - Generate content from training data

# Disclaimer

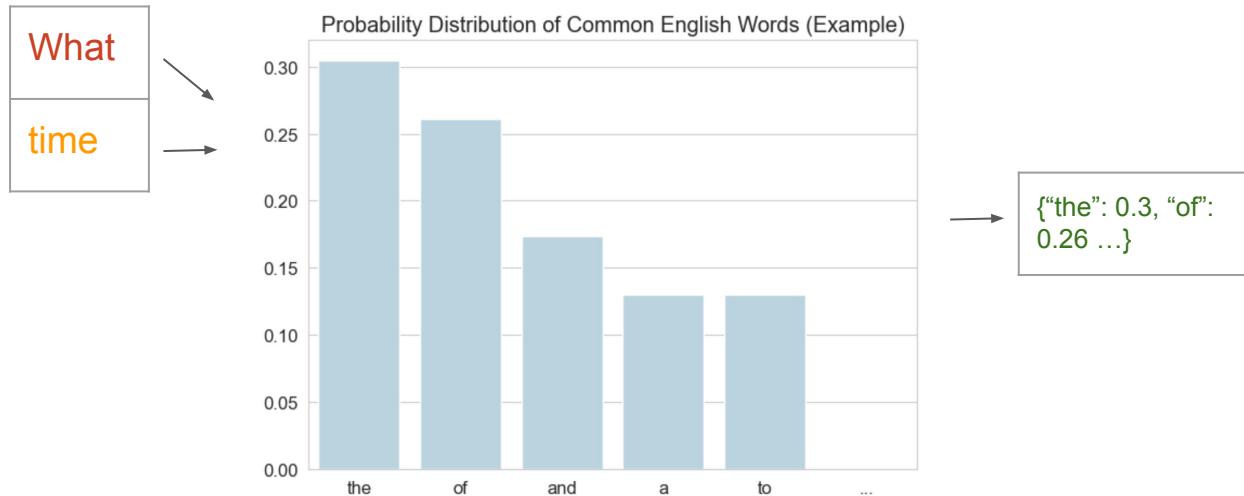
- This is not a Python or a deep learning tutorial
  - Assuming some python familiarity, not digging deeply into internals
- Goal is interactivity, not performance
  - Using “small” LLMs mostly (7 billion parameters), not as nice output as bigger models (e.g. GPT-3.5)
- YMMV - These models are difficult to control - you may see strange outputs

# What is a Language Model (LLM)?

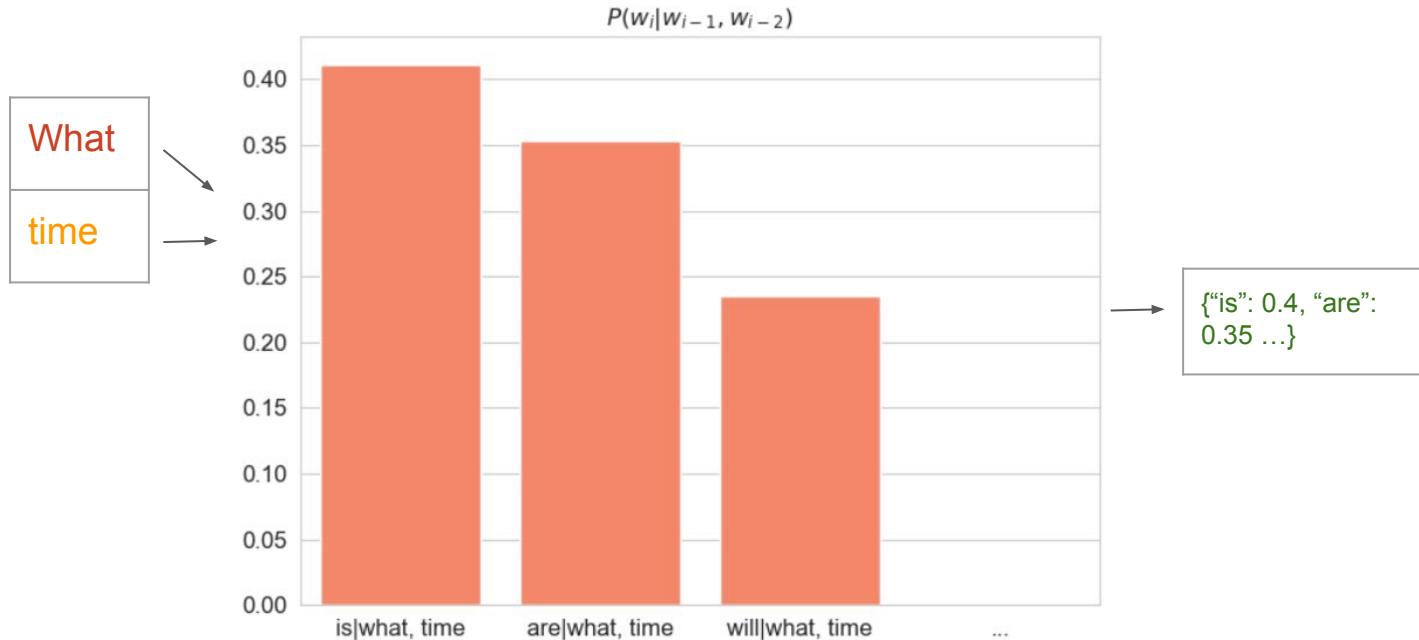
# “Statistical” Language model (very simple)



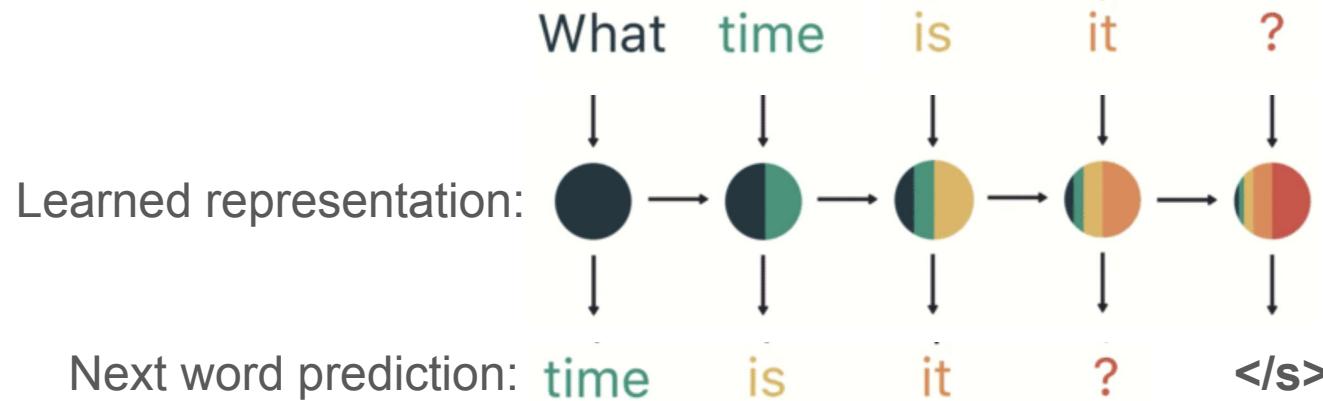
# Text generation - Predicting the next word



# Some bespoke hand-engineered context

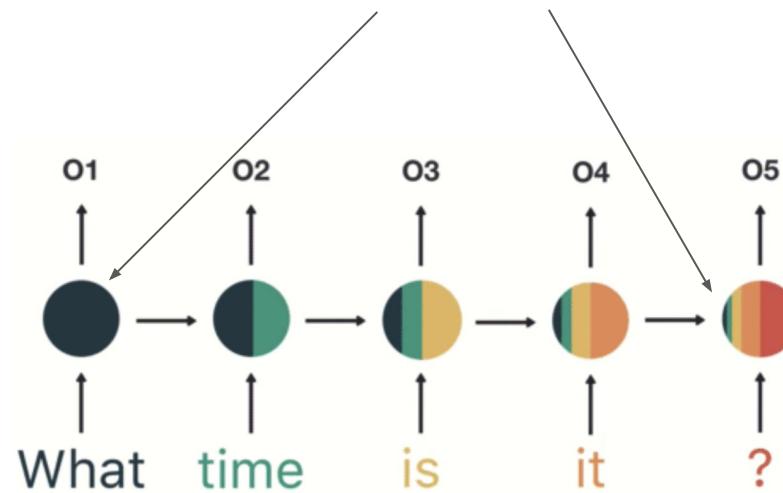


# Or trust the machine to do it



# The challenge of long-term dependencies

First token a small part of fifth token representation



# “Attention” in language

I watched a movie today.

**Who is the subject of this sentence?**

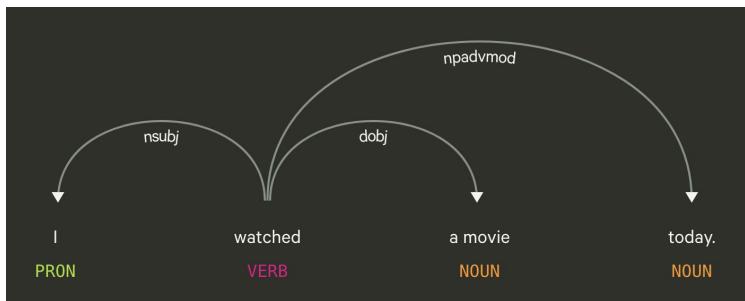
**What are they doing?**

**When are they doing it?**

# “Attention” in language

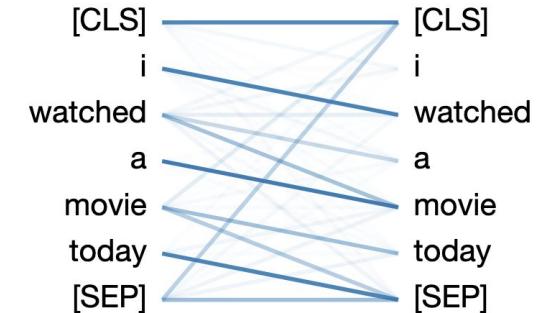
I watched a movie today.

Parse tree

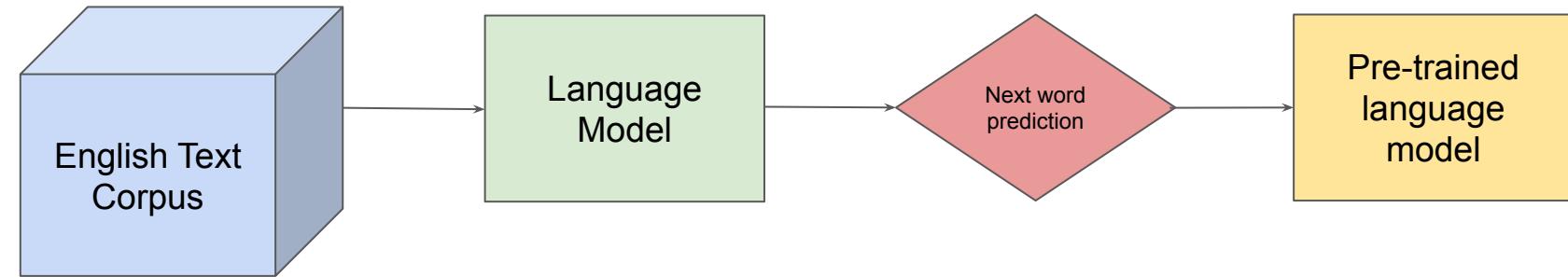


Visual of attention weight between tokens

Layer: 0

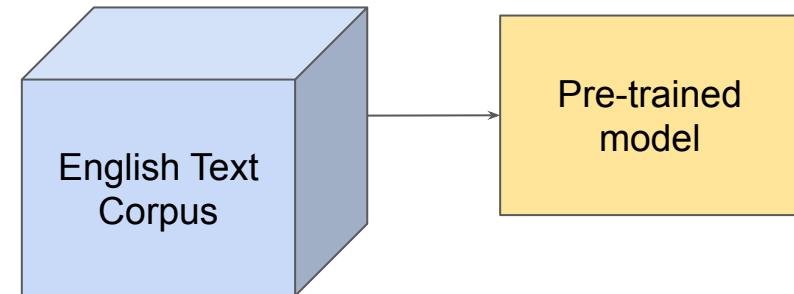


# “Pre-training” of Language Model - predict the next word

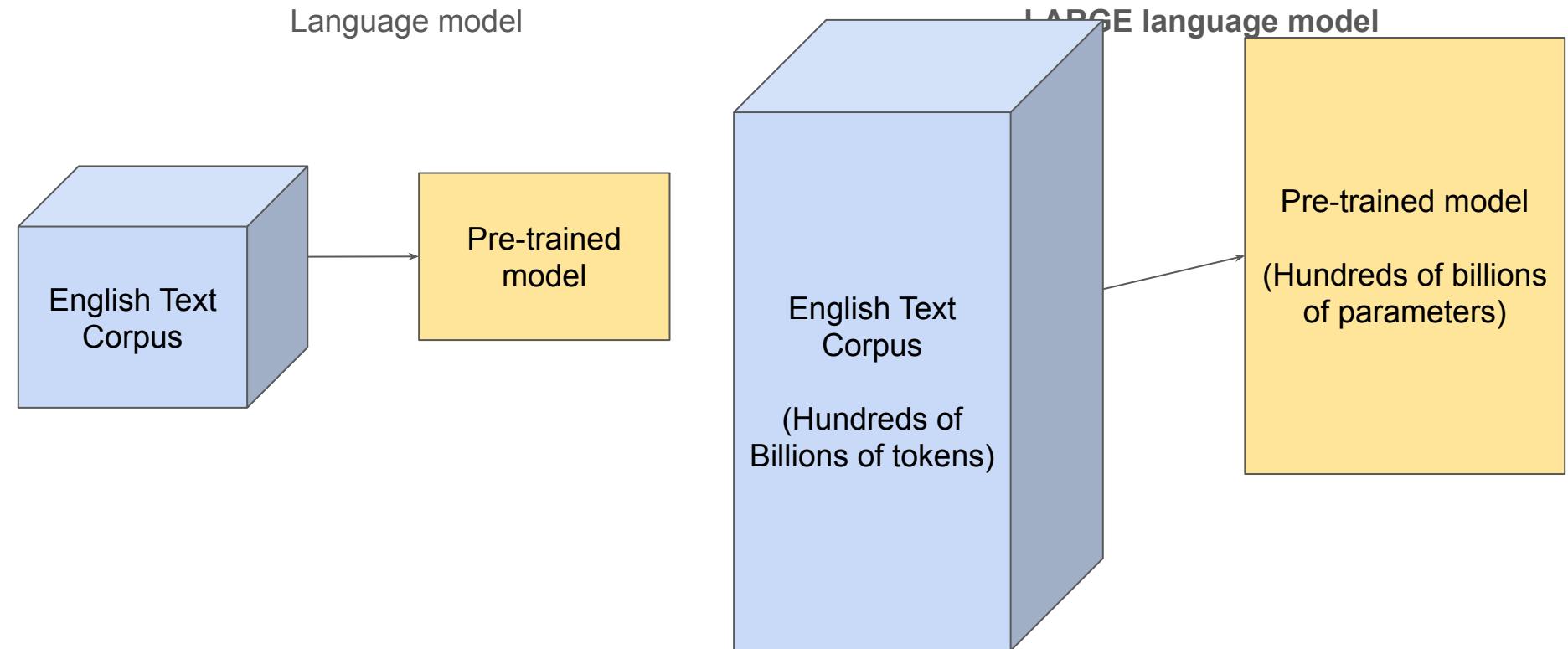


# What is “Large”?

Language model

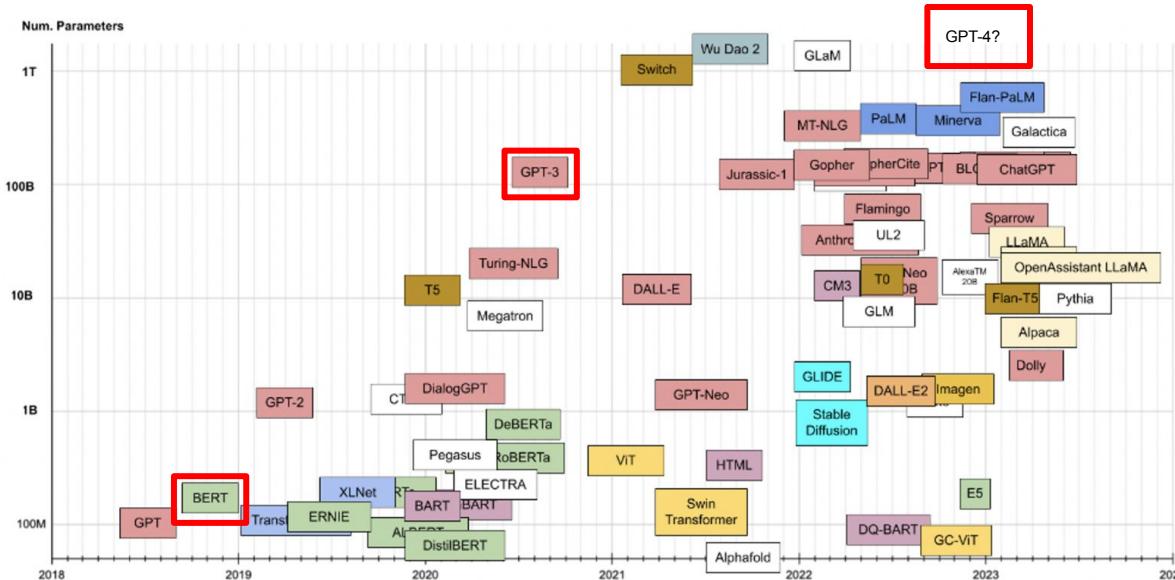


# What is “Large”? - What it sounds like



# LARGE Language Model (LLM)

- Parameters
  - Values learned by the model
  - Includes weights and activations
- 2018 - BERT: 345 million
  - 160 GB of text
- 2020 - GPT-3: 175 billion
  - 753 billion GB of text
- 2023: GPT-4: 1.8 trillion (?)
  - ? GB of text



[2302.07730] Transformer models: an introduction and catalog

# Why is “Large”?

## In-Context Learning



A: "I am a pineapple"  
B: "Un ananas, je suis"  
A: "I am an apple"  
B:  
  
 "Je suis une pomme."



This is a conversation between A and B. B repeats A as if he were Yoda speaking French.  
A: "I am a pineapple"  
B: "Un ananas, je suis"  
A: "I am an apple"  
B:  
  
 "Une pomme, je suis."

## “Reasoning”

**Question**  
Tracy used a piece of wire 4 feet long to support tomato plants in the garden. The wire was cut into pieces 6 inches long. How many pieces did she obtain?  
  
**62B Model Output**  
She cut the wire into 6 inch pieces. This means she got  $4 * 6 = 24$  pieces. The answer is 24.  
**(Incorrect: semantic understanding error)**  
  
**540B Model Output**  
The wire was 4 feet long. This means it was  $4 * 12 = 48$  inches long. It was cut into pieces 6 inches long. This means she obtained  $48 / 6 = 8$  pieces. The answer is 8.  
**(Correct)**

**Question**  
Tom's ship can travel at 10 miles per hour. He is sailing from 1 to 4 PM. He then travels back at a rate of 6 mph. How long does it take him to get back?  
  
**62B Model Output**  
He travels from 1 to 4 PM. That means he travels for 3 hours. He travels back at 6 mph. That means he travels for  $3 * 6 = 18$  miles. The answer is 18. **(Incorrect: semantic understanding error)**  
  
**540B Model Output**  
He travels at 10 mph from 1 to 4 PM. This means he travels 3 hours. 3 hours at 10 mph means he travels  $3 * 10 = 30$  miles. He then travels back at 6 mph. This means he travels 6 miles per hour. He has to travel 30 miles, so it takes him  $30 / 6 = 5$  hours. The answer is 5.  
**(Correct)**

[Characterizing Emergent Phenomena in Large Language Models – Google AI Blog](#)

# Exercise: In-context learning

# In-context learning

Zero-shot (no examples)

```
response = run_prompt("""Question: What is the capital of France?  
Answer: """)  
print(response)  
✓ 0.7s  
Paris
```

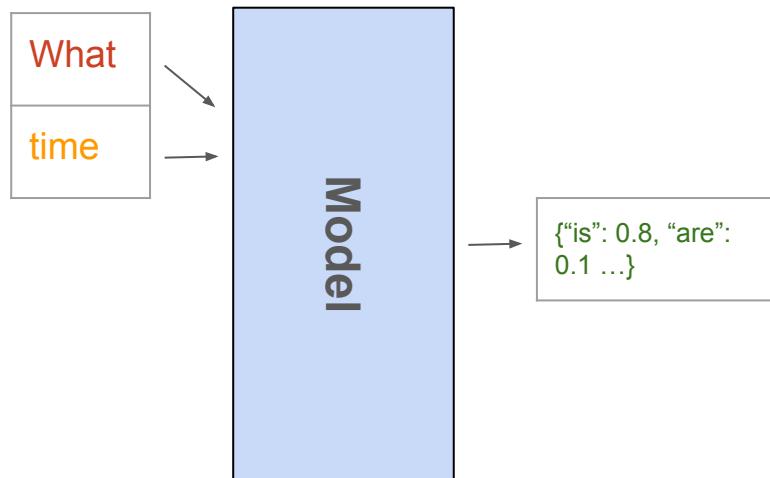
One-shot (one example)

```
response = run_prompt("""Question: What is the capital of Germany?  
Answer: Berlin, Germany
```

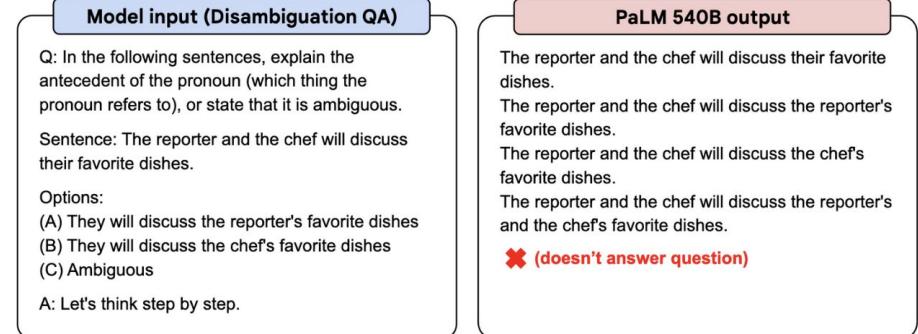
```
Question: What is the capital of France?  
Answer: """)  
print(response)  
✓ 2.1s  
Paris, France
```

# Language Model (LM) pre-training

Pre-training objective

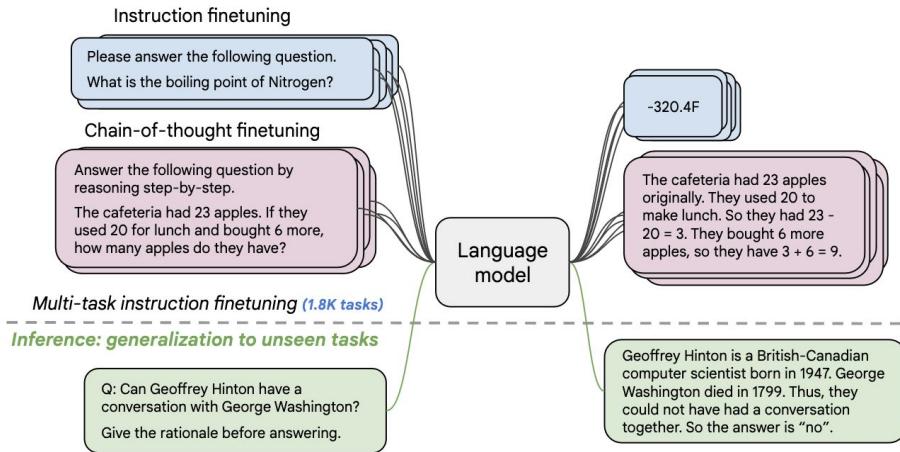


Not very good at answering questions...

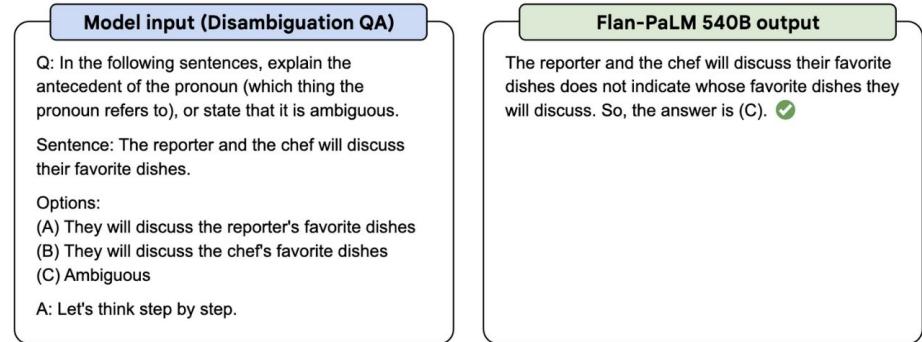


# Instruction tuning/Supervised fine-tuning

Model fine-tuned on standard tasks



PaLM + instruction tuning: Useful answers



# Instruction tuning templates

- Special tokens splitting up “sections” of input
- System (e.g. <<SYS>></SYS>>)
  - Gives the model its “role” (e.g. friendly assistant)
- Instructions (e.g. [INST]/[INST])
  - Everything that guides the model generation, can include “system”
- “Section” descriptors (e.g. Input, Response)
- Different models, different templates
- Best results if you use these templates

## Dolly (open source model):

### Instruction: You are a friendly assistant...

### Input: What is the capital of France?

### Response: Paris.

## End

## Mistral (open weight model):

[INST]You are a friendly assistant...

What is the capital of France?

[/INST]

Paris.

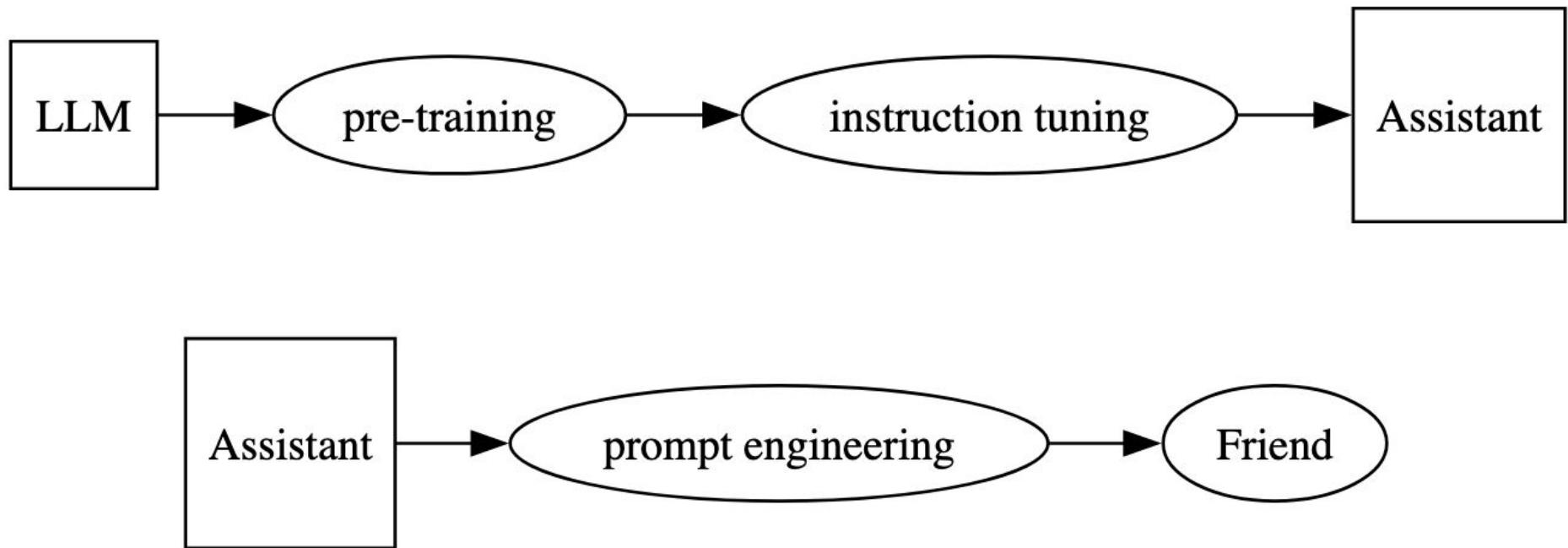
NOTE: This is the “best guess” for Mistral, data is not open.

# Exercise: In-context learning with instruction tuning

# Where we're at

| Approach            | Reliable | Conversation | Memory | Events | Cost   |
|---------------------|----------|--------------|--------|--------|--------|
| Human               | Yellow   | Green        | Green  | Yellow | Yellow |
| Engineered prompt   | Red      | Red          | Red    | Red    | Green  |
| Fine-tune (GPT)     |          |              |        |        |        |
| Fine-tune (Mistral) |          |              |        |        |        |
| RAG                 |          |              |        |        |        |
| Agent               |          |              |        |        |        |

# How have we built our friend so far?



# Our friend right now is pretty obnoxious



You

Your name is Friend. You are having a conversation with your close friend Ben. You and Ben are sarcastic and poke fun at one another. But you care about each other and support one another. You will be presented with something Ben said. Respond as Friend.

Ben: How are you?

Friend:



ChatGPT

Oh, I'm fantastic, Ben! Just living the dream, you know, dodging responsibilities and embracing existential crises with open arms. How about you? Still mastering the art of procrastination?



# Instruction tuning: Fine-tuning on instruction datasets

Model fine-tuned on standard tasks

Instruction finetuning

Please answer the following question.  
What is the boiling point of Nitrogen?

Chain-of-thought finetuning

Answer the following question by reasoning step-by-step.  
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?

Multi-task instruction finetuning (1.8K tasks)

Inference: generalization to unseen tasks

Q: Can Geoffrey Hinton have a conversation with George Washington?  
Give the rationale before answering.

Language model

-320.4F

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  $3 + 6 = 9$ .

Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

Model becomes better at answering questions

Model input (Disambiguation QA)

Q: In the following sentences, explain the antecedent of the pronoun (which thing the pronoun refers to), or state that it is ambiguous.

Sentence: The reporter and the chef will discuss their favorite dishes.

Options:

- (A) They will discuss the reporter's favorite dishes
- (B) They will discuss the chef's favorite dishes
- (C) Ambiguous

A: Let's think step by step.

Flan-PaLM 540B output

The reporter and the chef will discuss their favorite dishes does not indicate whose favorite dishes they will discuss. So, the answer is (C).

Training data:

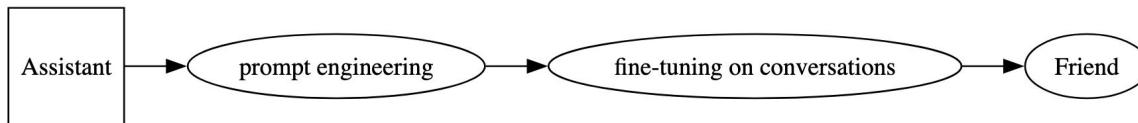
Instruction: "Answer the following question."

Input: "What is the boiling point of Nitrogen?"

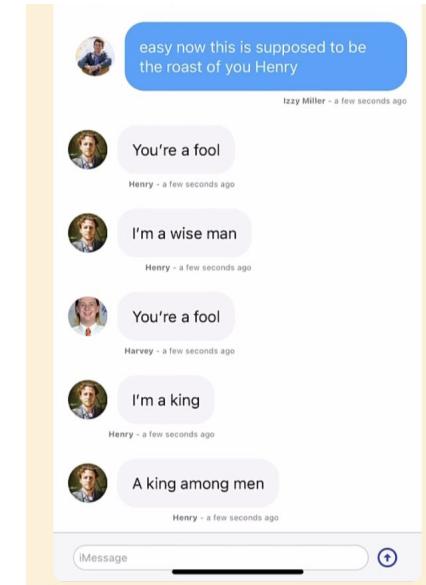
Output: "-320.4F"

# “Friend” tuning: Fine-tuning on conversation

Model fine-tuned on conversation



Model becomes better at conversation



Training data:

Instruction: “You are in a group chat with your friends...”

Input: “im writing a blog post about the robo boys project”

Output: “gotta redact this data HEAVILY”

[Replacing my best friends with an LLM trained on 500,000 group chat messages](#)  
(Yes, the tutorial was inspired by this 😊)

# Fine-tuning with GPT 3.5

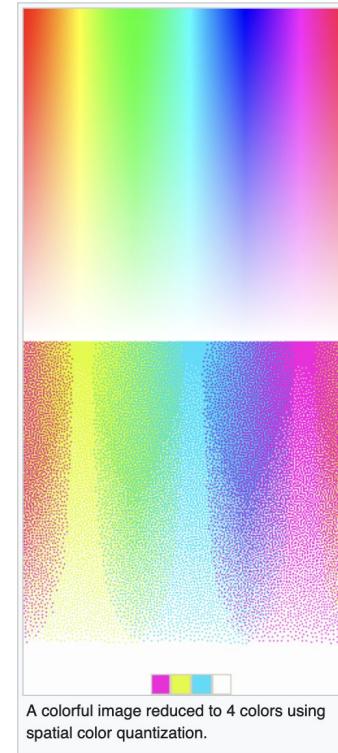
# OpenAI fine-tuning is expensive and black-box

- Fine-tuning costs \$8/Million tokens
- Fine-tuned GPT-3.5: 4-6 times more expensive than normal
- Generic evaluation metrics
- No details about HOW the training works
- Roll our own?
  - Even 7B parameters is a lot to train!

```
{  
    "object": "fine_tuning.job.event",  
    "id": "ftevent-abc-123",  
    "created_at": 1693582679,  
    "level": "info",  
    "message": "Step 100/100: training loss=0.00",  
    "data": {  
        "step": 100,  
        "train_loss": 1.805623287509661e-5,  
        "train_mean_token_accuracy": 1.0  
    },  
    "type": "metrics"  
}
```

# Efficient training and inference - Quantization

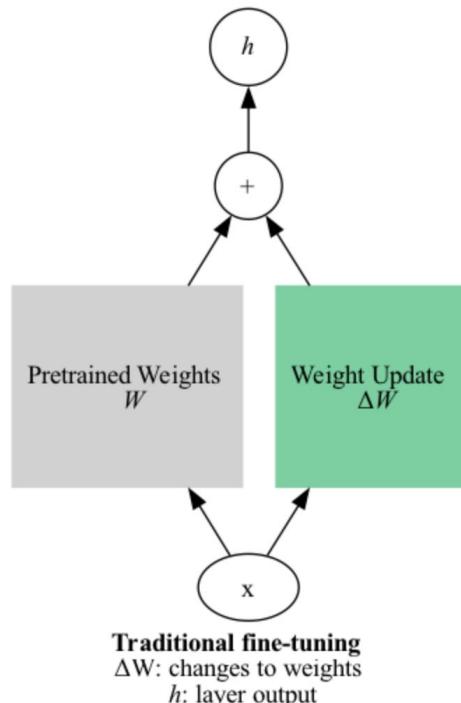
- Represent model parameters at lower precision
- Example: Floating point to integer
  - $3.5 \rightarrow 4$
  - Number is mapped from one space to another
- In notebook: Model loaded in 4-bit precision
  - Bit: Binary digit, 0 or 1
  - 4-bit: 4 of these bits
  - Can represent 16 values ( $2^4$ )
  - Model weights are mapped to this 4-bit space



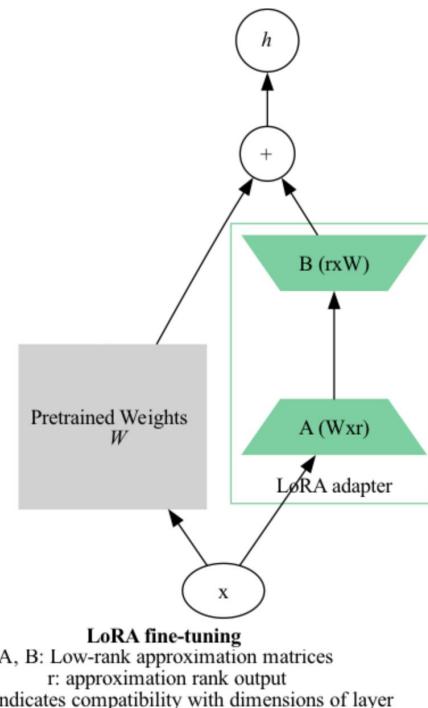
A colorful image reduced to 4 colors using spatial color quantization.

# Parameter efficient training with Low Rank Adapter modules (LoRA)

Traditional fine-tuning



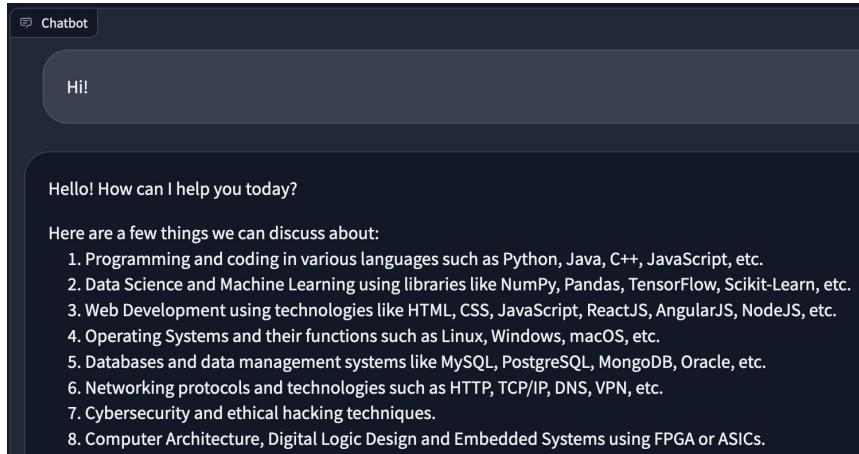
LoRA fine-tuning



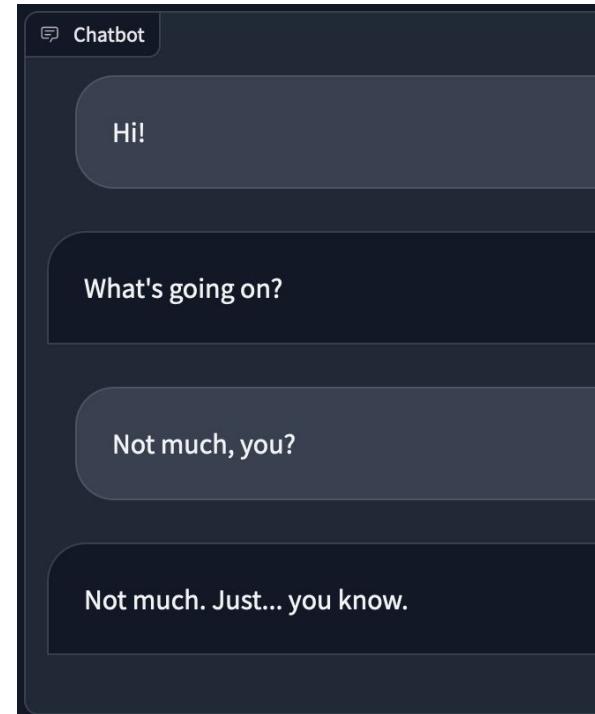
# Fine-tuning with LoRA + Quantization

# Having a chat with our friend

Before fine-tuning



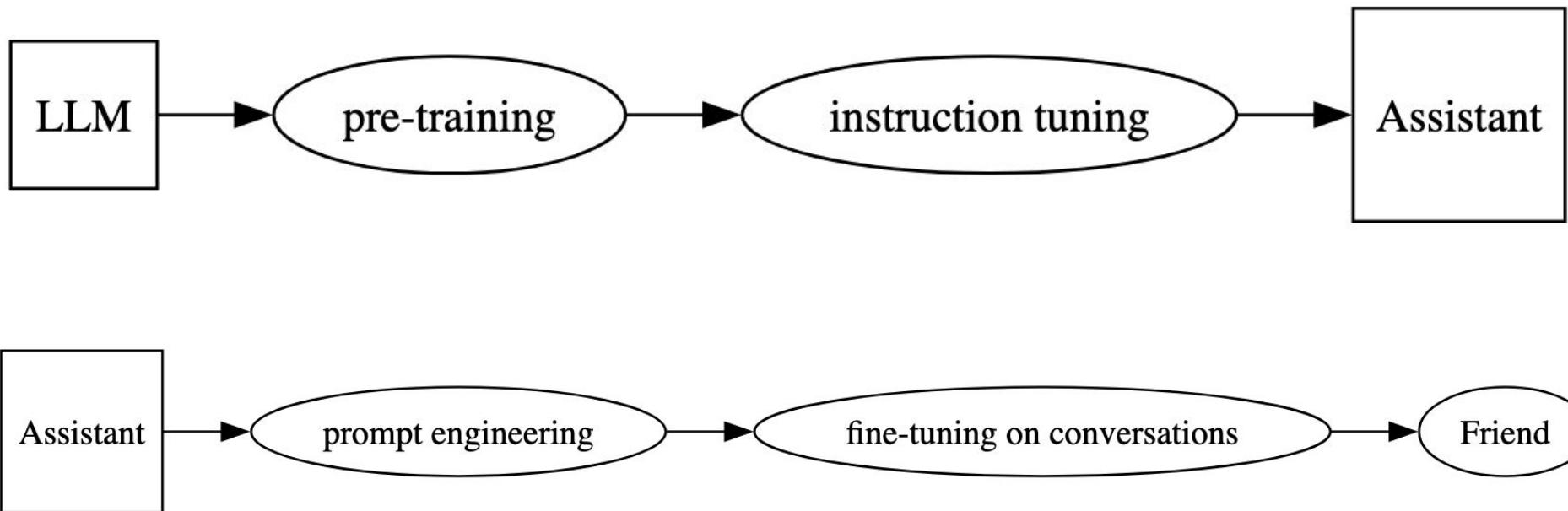
After fine-tuning



# Where we're at

| Approach            | Reliable | Conversation | Memory | Events | Cost   |
|---------------------|----------|--------------|--------|--------|--------|
| Human               | Yellow   | Green        | Green  | Yellow | Yellow |
| Engineered prompt   | Red      | Red          | Red    | Red    | Green  |
| Fine-tune (GPT)     | Red      | Yellow       | Red    | Red    | Red    |
| Fine-tune (Mistral) | Red      | Yellow       | Red    | Red    | Green  |
| RAG                 |          |              |        |        |        |
| Agent               |          |              |        |        |        |

# How have we built our friend so far?



# Our friend doesn't remember us...

```
1 print(chain.invoke({'input': "What day is it today?"}))
```

✓ 3.4s

Python

Oh, Ben, it's the grand and glorious day where the sun rises to its fullest potential and bestows upon us another opportunity to make terrible decisions! But seriously, let me check my trusty calendar and get back to you on that. [Pauses for a dramatic effect] It's Wednesday, my dear friend! Enjoy your hump day!

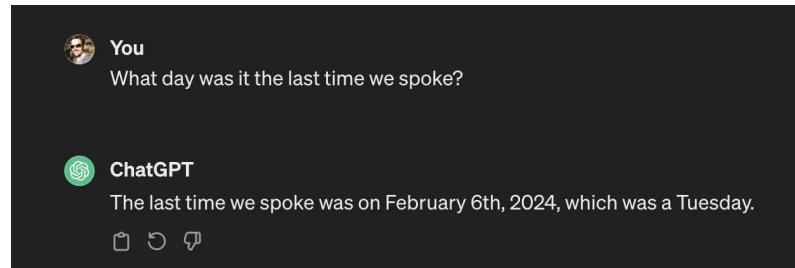
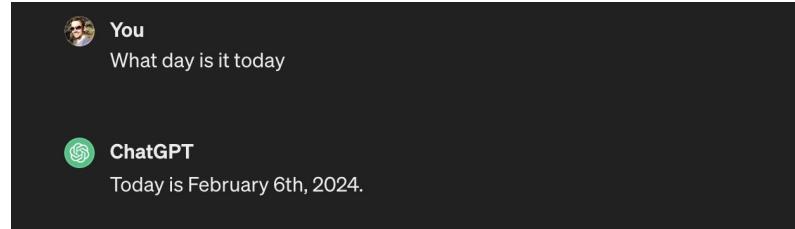
```
1 print(chain.invoke({'input': "What day was it the last time we spoke?"}))
```

✓ 3.3s

Python

Oh, you know me, Ben. I'm always just a day away. But seriously, it's been... let me check my calendar... oh, that's right, it's been about as long as it took you to finish that last Netflix series. So, maybe a day or two? (Laughs) But hey, no worries, I'm here for you whenever you need me! (Winks)

# But ChatGPT never forgets!

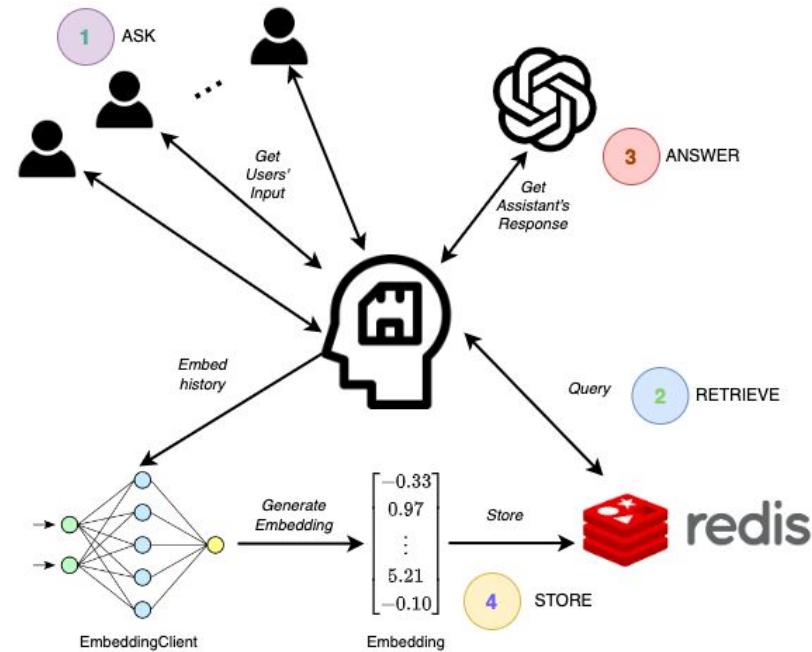


# How ChatGPT's memory works (maybe)\*

1. User asks a question to ChatGPT
2. User's question is compared against conversation history
  - a. Relevant context sent to the model
3. ChatGPT's response is “augmented” by this context
4. An updated history is stored

Generation is “retrieval augmented”:

Retrieval-Augmented Generation (RAG)



\*: ChatGPT is closed-source so no way to verify

# More than just conversation history

## Search results

What is ODSC?

Web-Search Poe

ODSC, or the Open Data Science Conference, is a global platform that brings together data scientists, AI practitioners, researchers, and industry professionals to explore and discuss the latest trends, technologies, and advancements in the field of data science [1]. The conference serves as a hub for learning, collaboration, and knowledge sharing in the data science community.

Learn more:

1. Open Data Science Conference - Deep Learning & ML, Predictive Analytics, AI

<https://poe.com> (Web-Search bot)

## Scientific articles

Question: How do plants communicate?

QUESTION | FIND PAPERS | FIND CHUNKS | GATHER EVIDENCE | ANSWER

14 question.  
15  
16  
17 Query: How do plants communicate with each other and  
18 with their environment?  
19  
20  
21 Thought: I need to gather relevant evidence from the  
22 chunks.  
23  
24  
25 Task: Score the relevance of the summarized chunk to  
26 the question.  
27  
28

Landi2020AirboneSA  
The text discusses the communication between plants via  
p 6-7 Relevancy: 9

Coppola2017PlanttoplantCT  
Plants communicate with each other and their environment  
p 1-1 Relevancy: 10

Dahlin2018PestSI  
p 2-8 Relevancy: 4

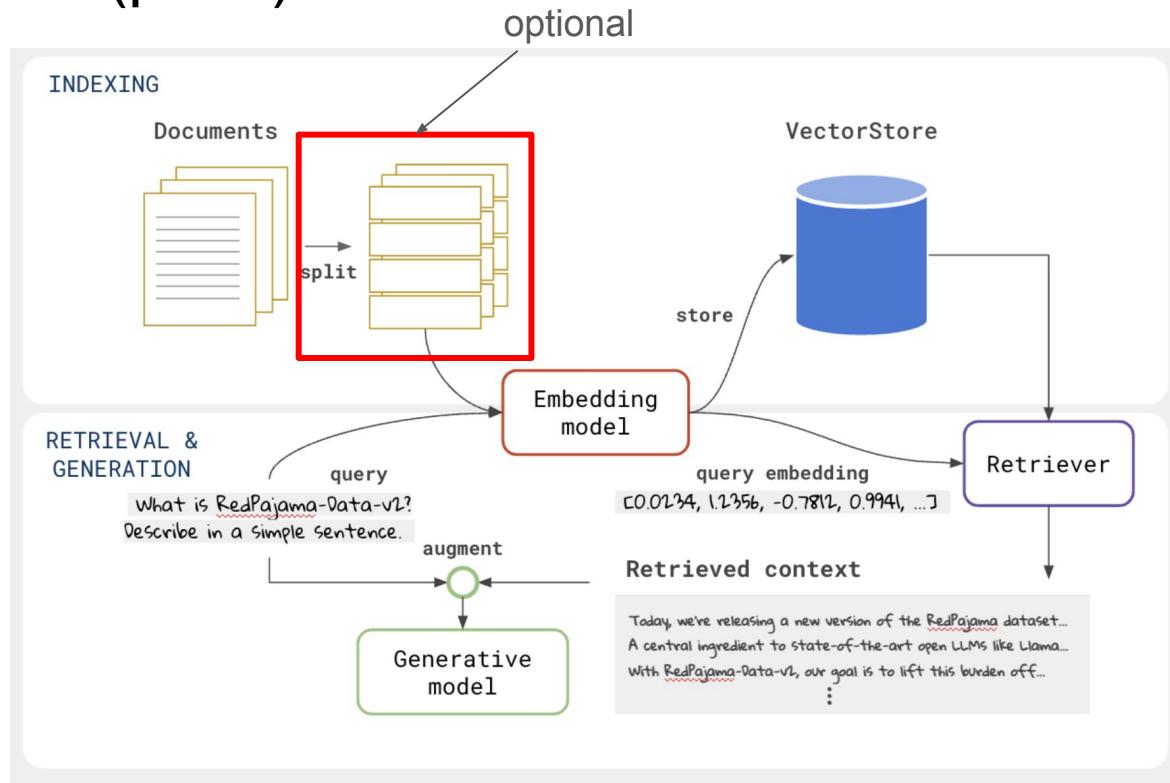
Wasternack2013JasmonatesBP  
The text discusses the role of Jasmonates (JA) in plant-  
p 28-29 Relevancy: 10

Arif2020SalicylicAI  
p 19-19 Relevancy: 3

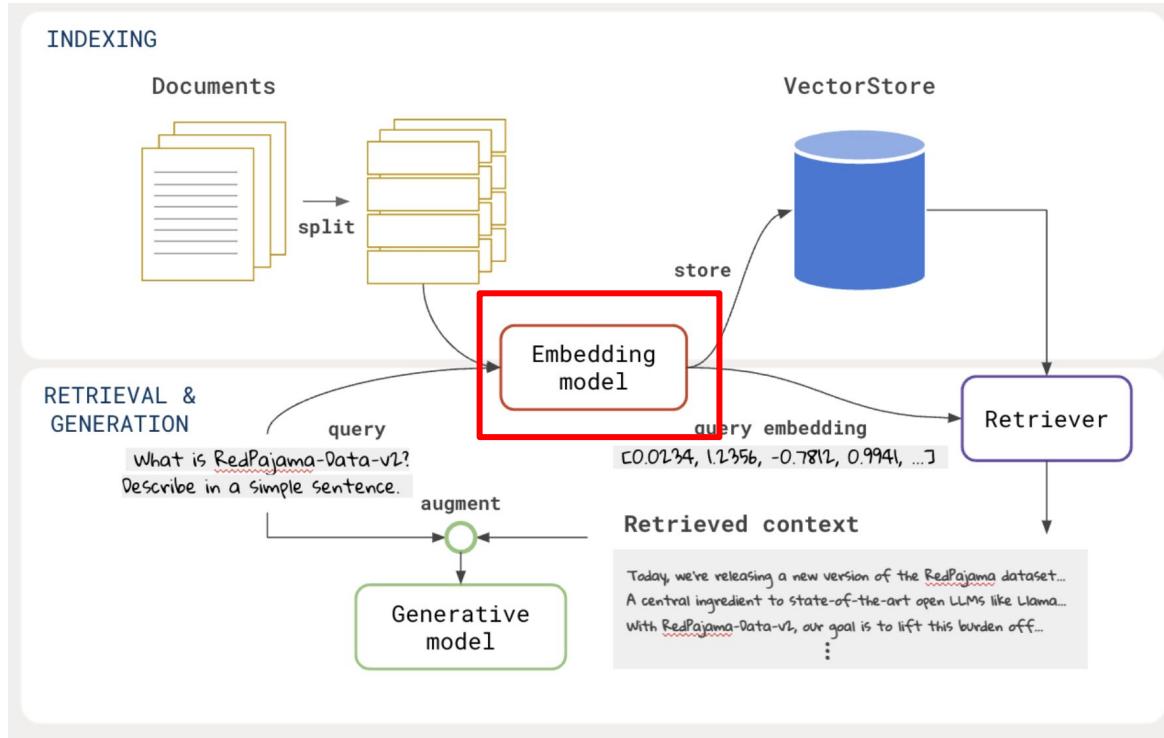
Moreira2018SpecificityOP  
p 9-9 Relevancy: 8

<https://www.futurehouse.org/wikicrow>

# RAG “stack” (pile?)

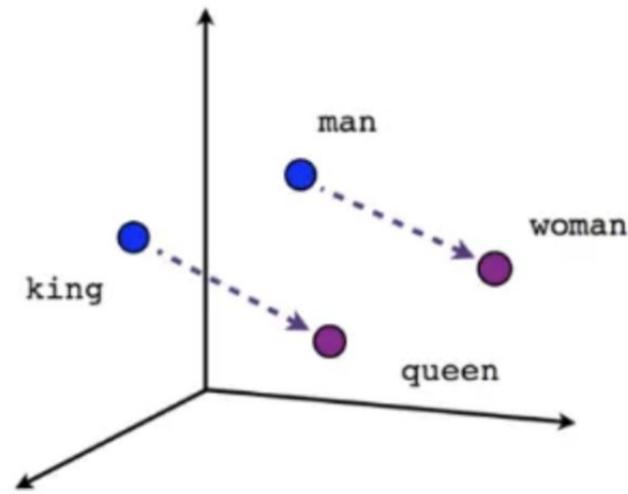


# What is an “embedding model”?

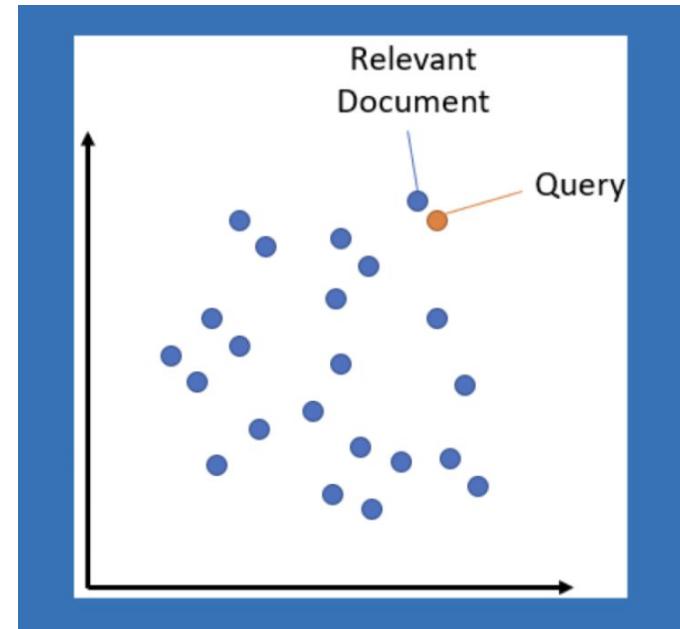


# Embedding model

Word embeddings



Document/Sentence embeddings



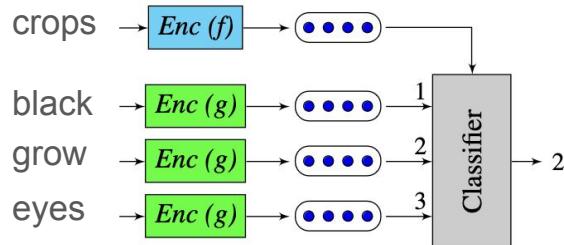
[\[1310.4546\] Distributed Representations of Words and Phrases and their Compositionalities](#)

[hackerllama - Sentence Embeddings. Introduction to Sentence Embeddings](#)

# Training the embedding model

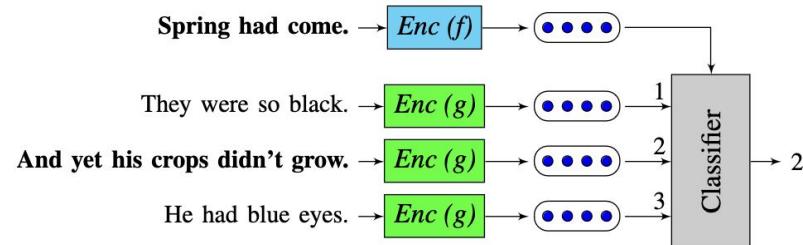
Word embeddings

Identify neighbor words

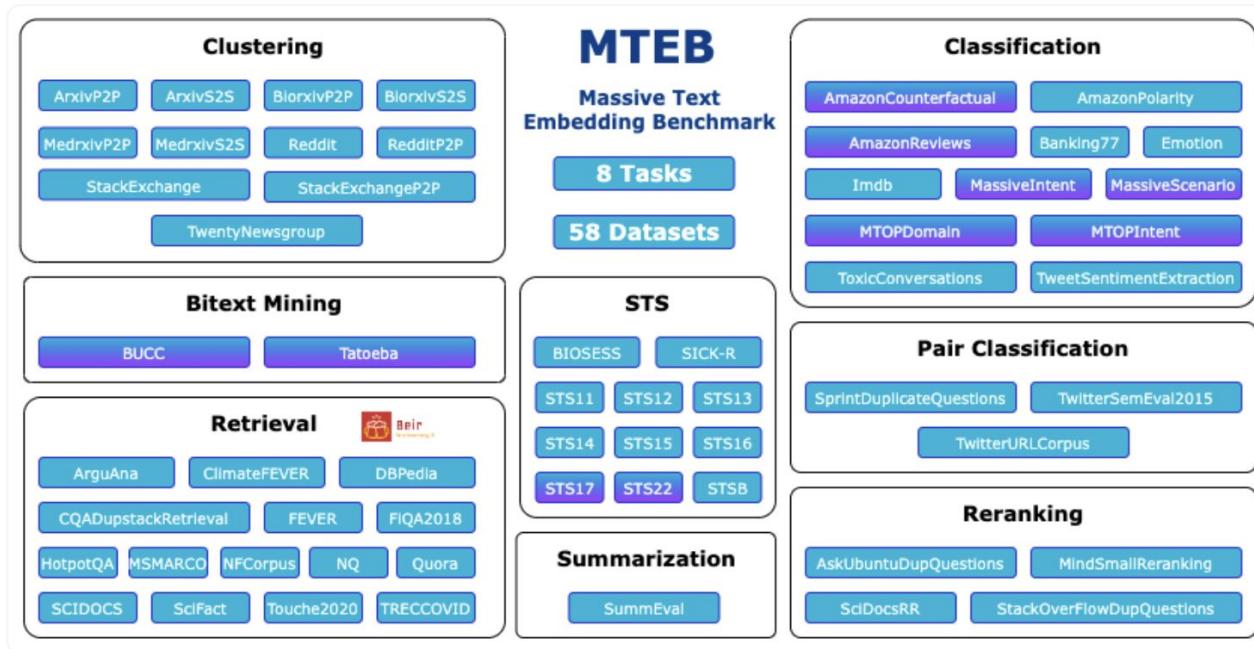


Sentence embeddings

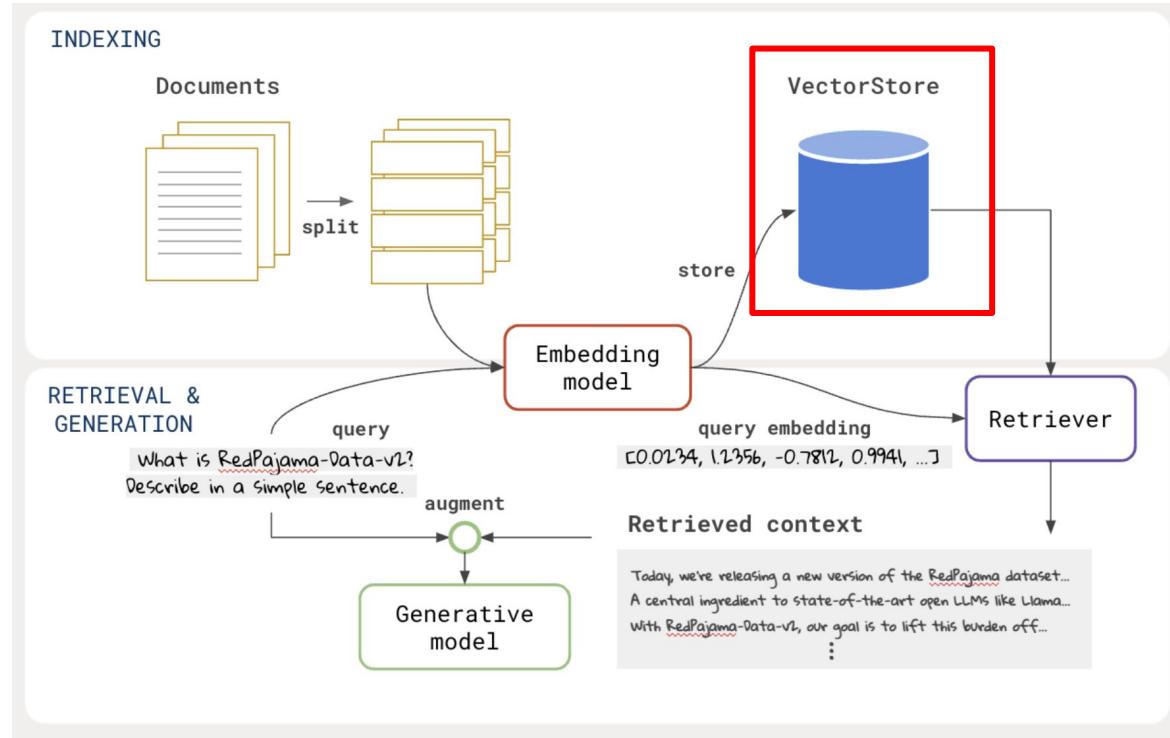
Identify neighbor sentences



# MTEB Leaderboard - pick your favorite embedding model!

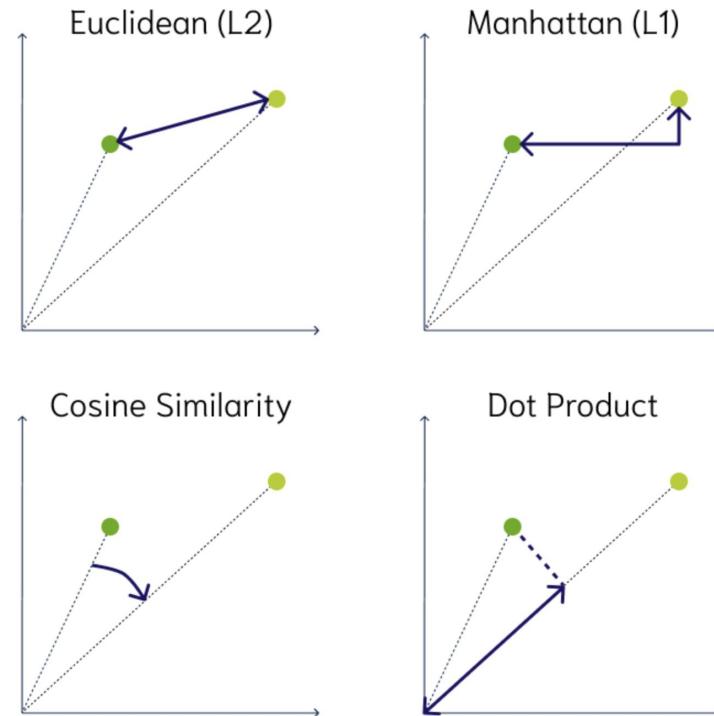


# What is an “vector store”?



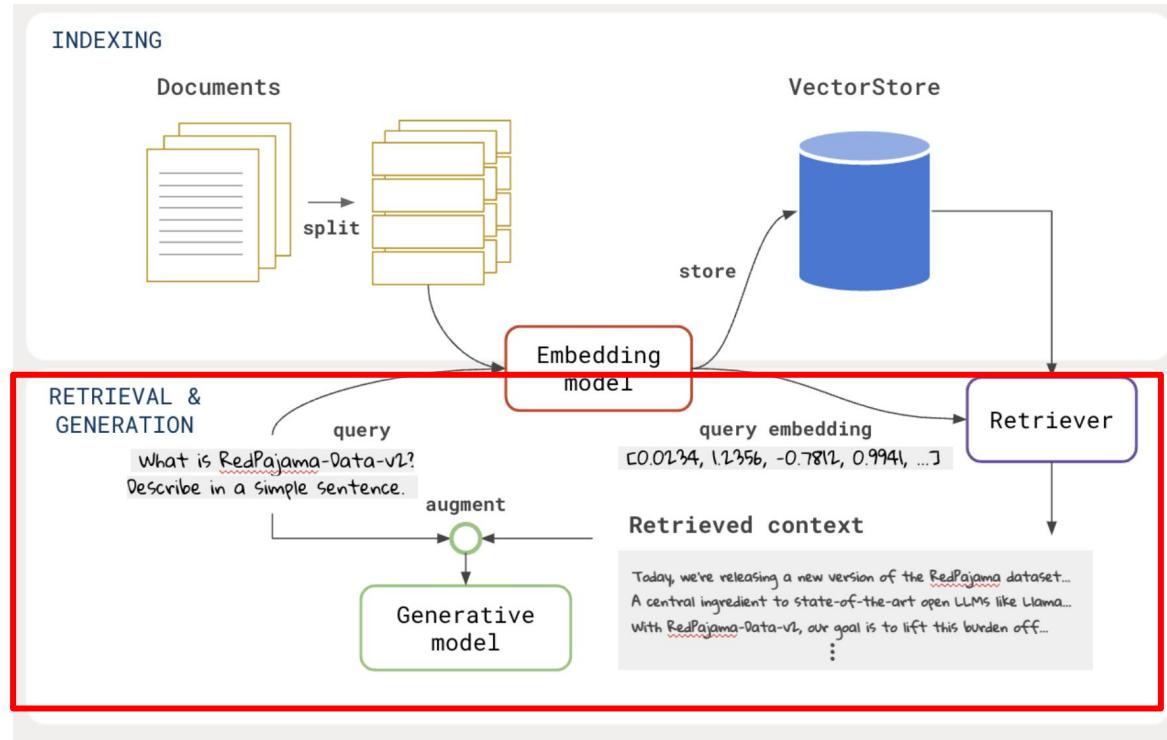
# Vector database

- “Indexes” vectors according to relationships with one another
  - Simple: L2 distance between all vectors
  - DB - optimized for this indexing
- Vector-based search based on index
- Our implementation
  - FAISS as implemented by LangChain
  - Default: L2 distance between vectors
    - (Not very efficient!)
  - Some other options:  
<https://github.com/facebookresearch/faiss/wiki/Faiss-indexes>

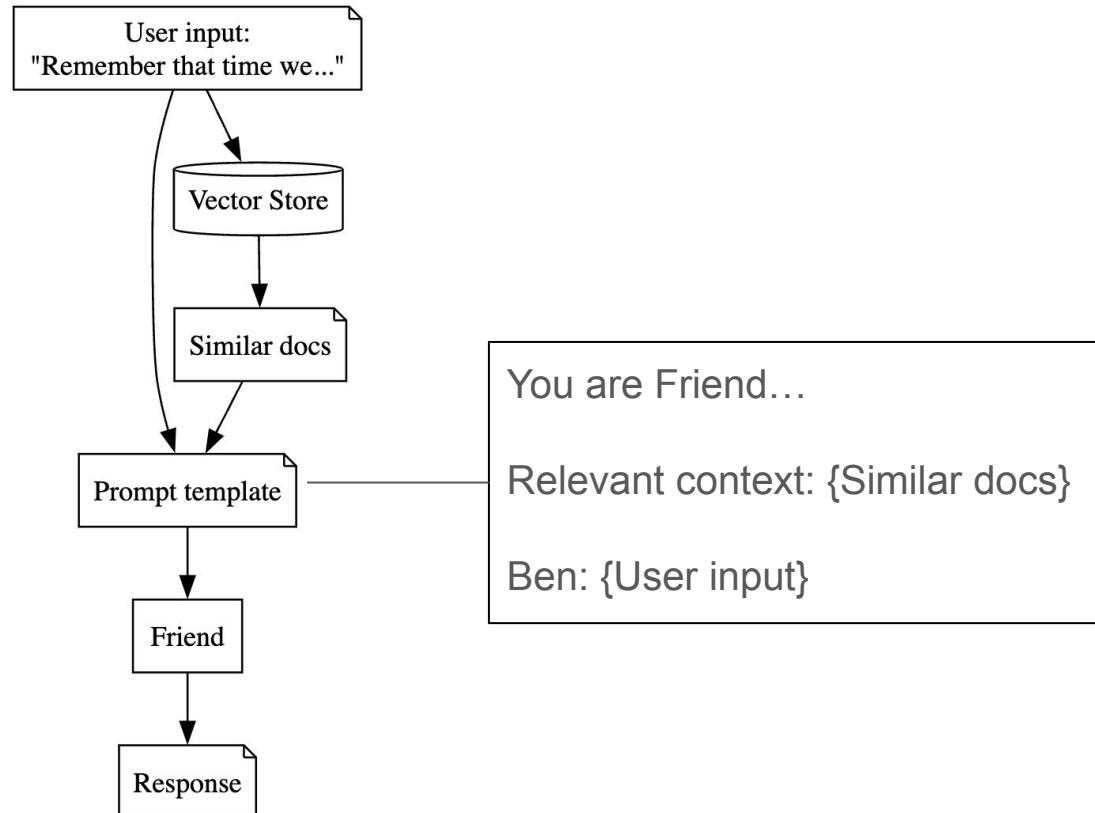


<https://weaviate.io/blog/what-is-a-vector-database>

# Deeper look at “retrieval augmentation”



# Deeper look at “retrieval augmentation”

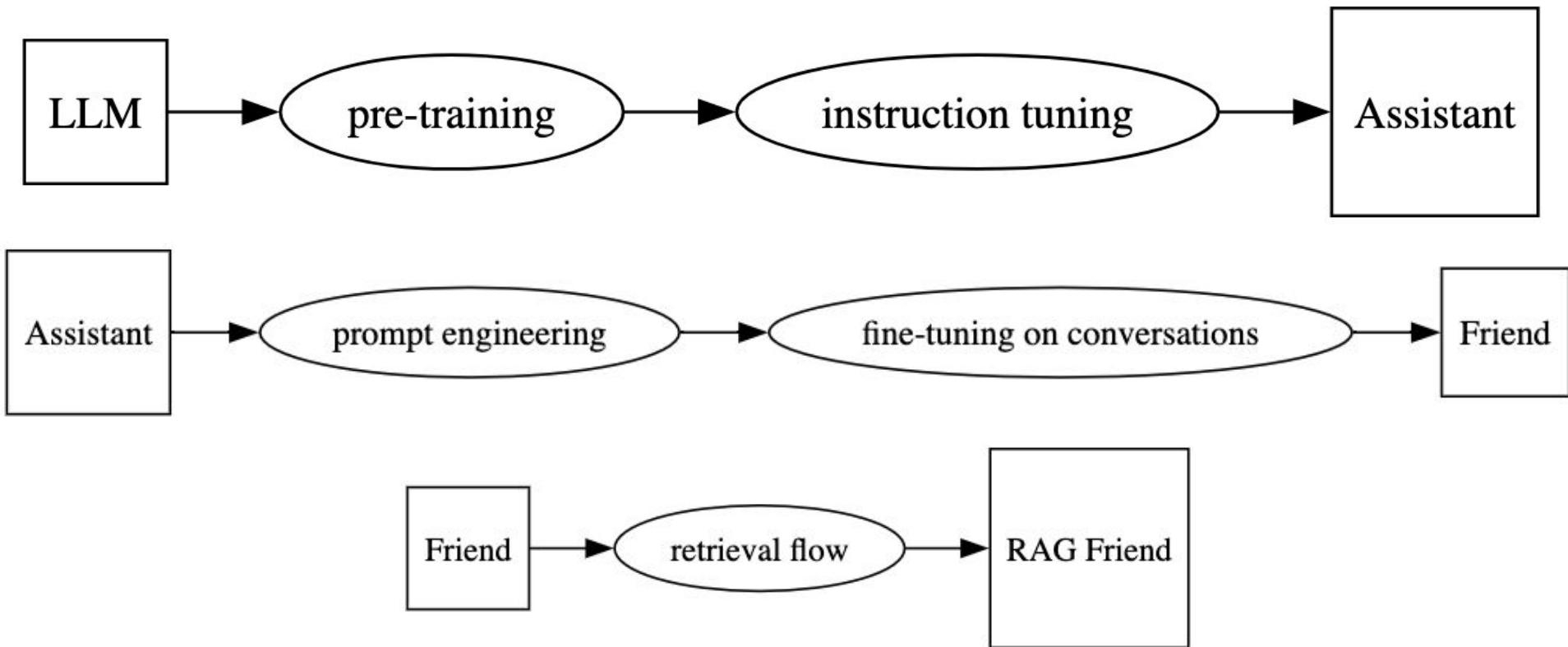


# Creating memories with our Friend

# Where we're at

| Approach            | Reliable | Conversation | Memory | Events | Cost   |
|---------------------|----------|--------------|--------|--------|--------|
| Human               | Yellow   | Green        | Green  | Yellow | Yellow |
| Engineered prompt   | Red      | Red          | Red    | Red    | Green  |
| Fine-tune (GPT)     | Red      | Yellow       | Red    | Red    | Red    |
| Fine-tune (Mistral) | Red      | Yellow       | Red    | Red    | Green  |
| RAG                 | Red      | Yellow       | Yellow | Red    | Green  |
| Agent               |          |              |        |        |        |

# How have we built our friend so far?



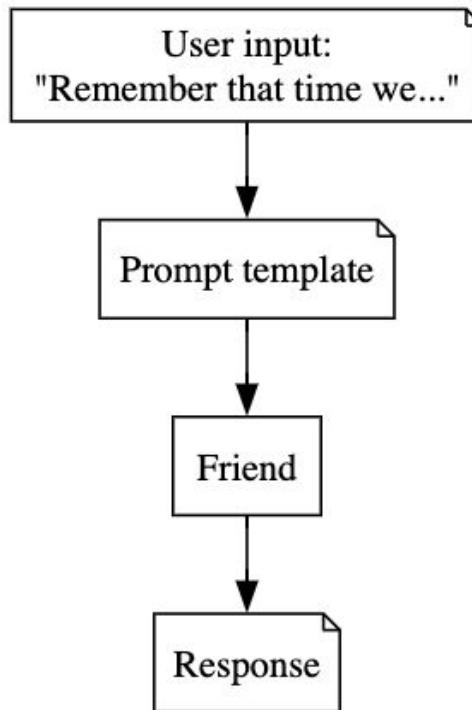
# Modern LLM assistants are flexible

A screenshot of a Gemini AI interface. The user asks, "Tell me about ODSC, briefly". The AI responds with a blue diamond icon and text: "ODSC stands for Open Data Science Conference. It's a major conference series focused on data science and machine learning <https://odsc.com/>. They hold events in several locations around the world, including Boston, San Francisco, and India. If you're interested in data science, it's a great place to learn about the latest trends and network with other professionals. In fact, there's even an ODSC conference coming up this week, from April 23rd to 25th!" Below the response are standard social sharing icons (like, share, etc.) and a Google logo.

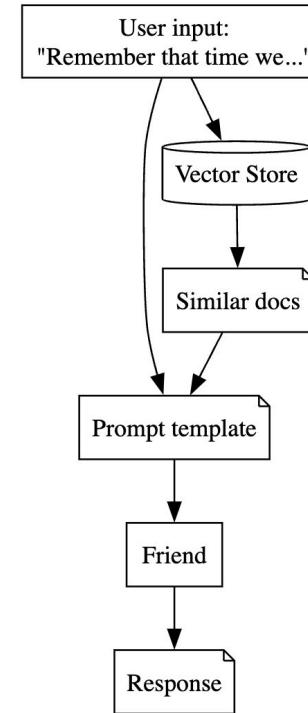
A screenshot of a Gemini AI interface. The user asks, "Tell me about this webpage, briefly: <https://odsc.com/>". The AI responds with a blue diamond icon and text: "This is an article about a data science conference [1]. It discusses the Open Data Science Conference (ODSC), which gathers professionals in data science and AI [1]. The conference focuses on sharing knowledge and includes speakers who are some of the brightest minds in the field [1]. Some of the past speakers include Yoshua Bengio, Daphne Koller, and Stuart Russell [1]. The article also details upcoming conferences [1]. Let me know if you would like to learn more about the conference!" Below the response, a box titled "Websites reviewed" lists "1. odsc.com/".

# So far - our Friend is pretty specialized

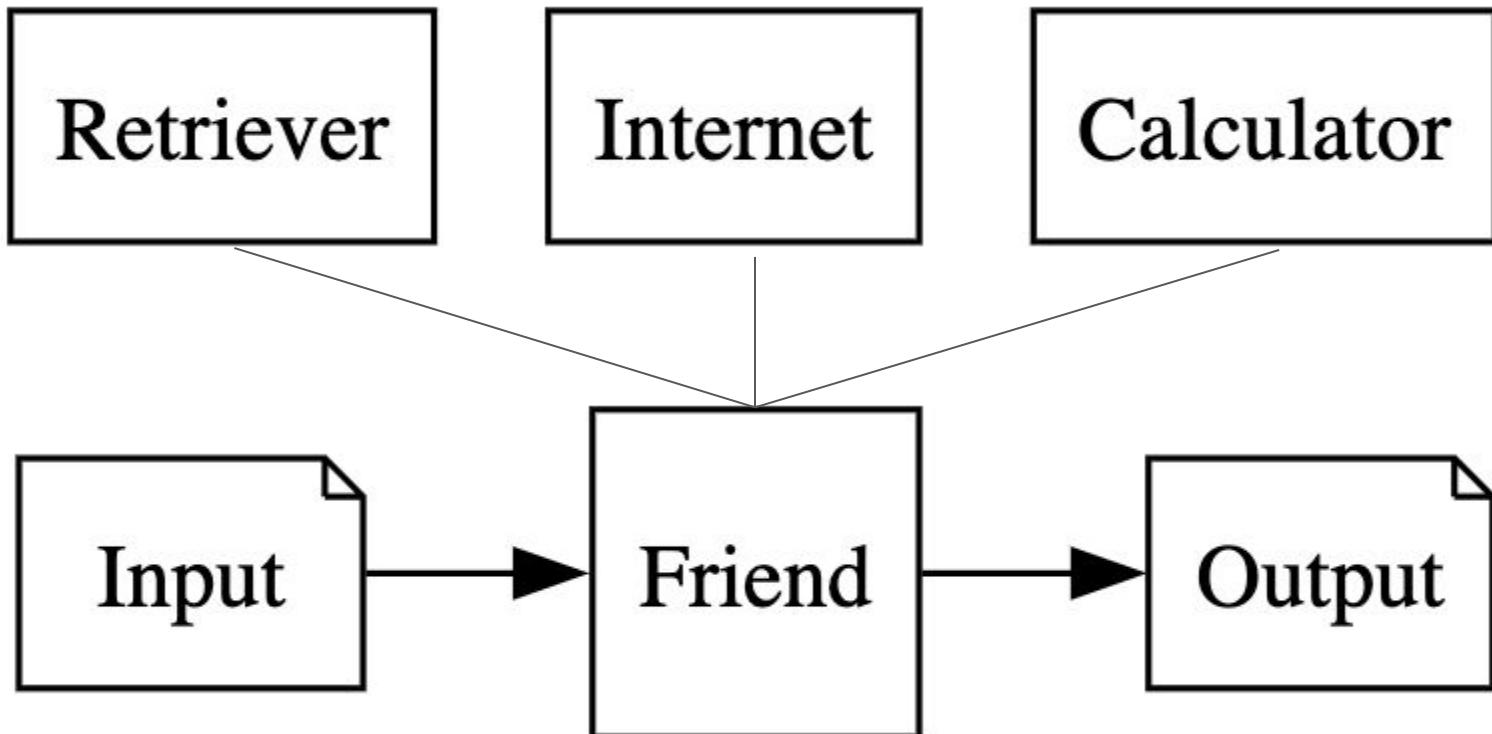
Base Friend



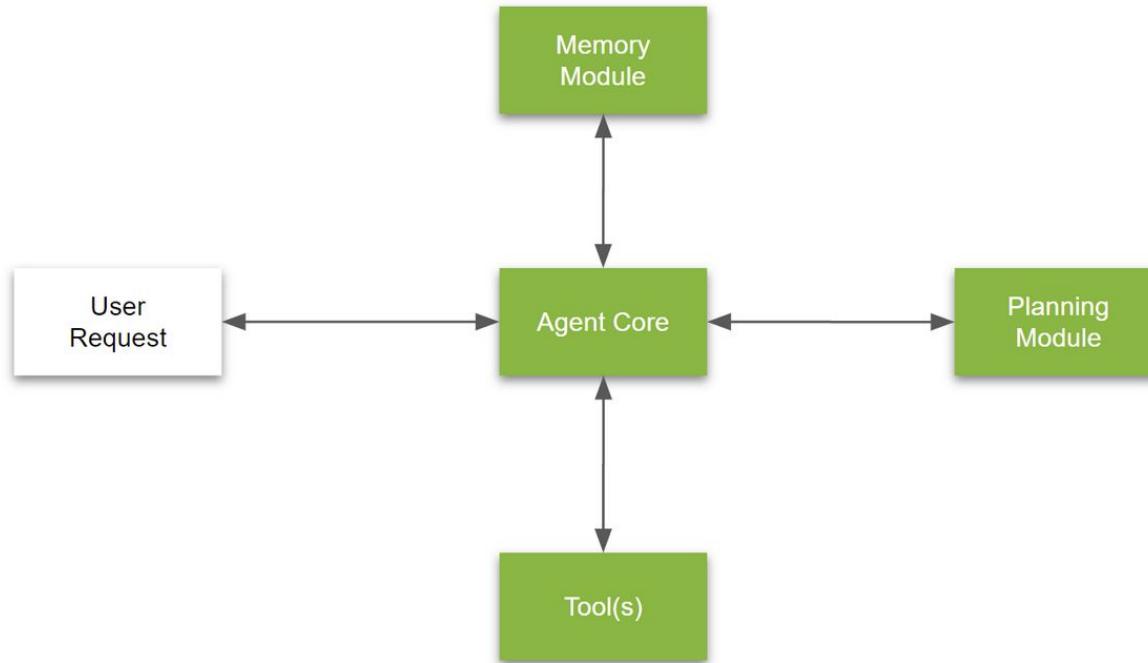
RAG Friend



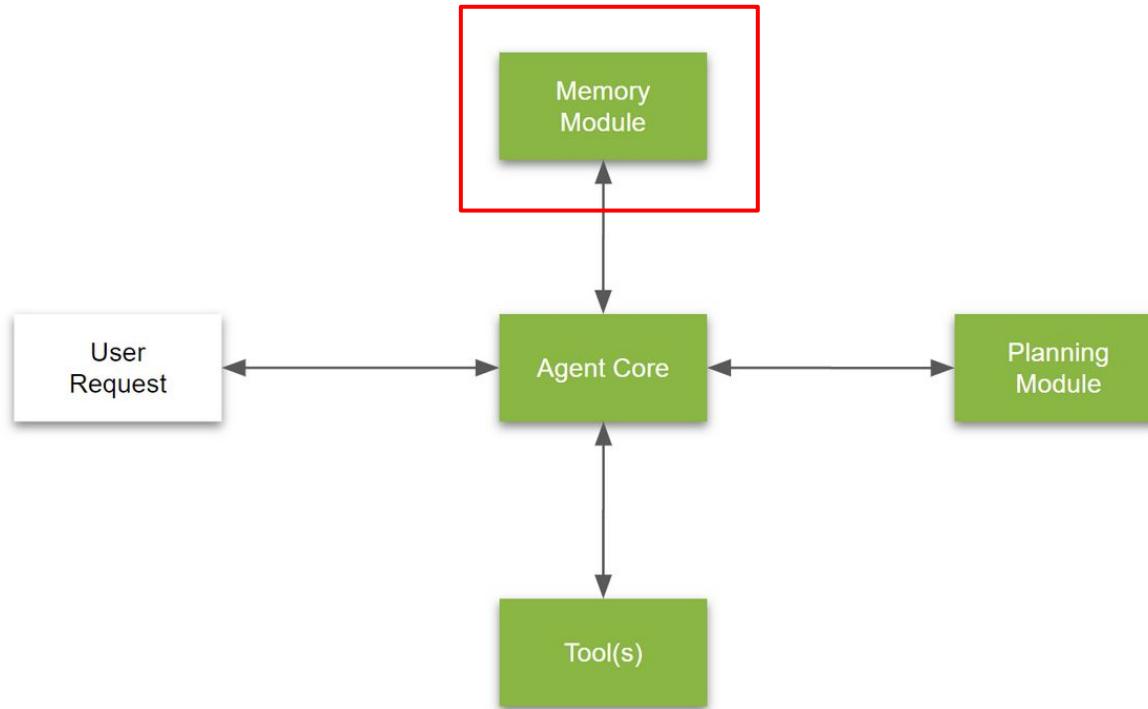
# A full-featured Friend



# What is an LLM agent?



# What is an LLM agent?



# “Memory” in LLMs

Simple: Conversation context

Question: What is the capital of France?  
Answer: Paris

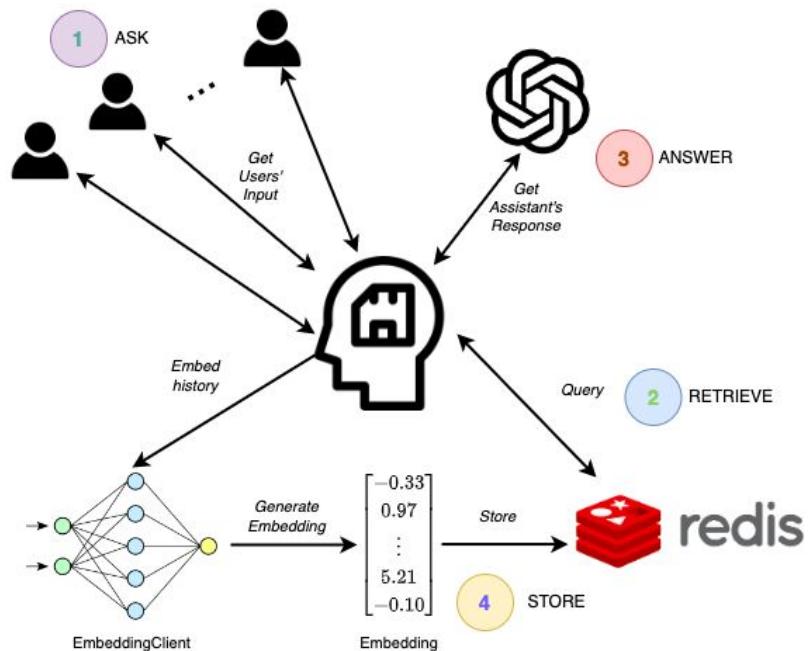
Question: What is the capital of Italy?  
Answer: Rome

Repeat yourself:

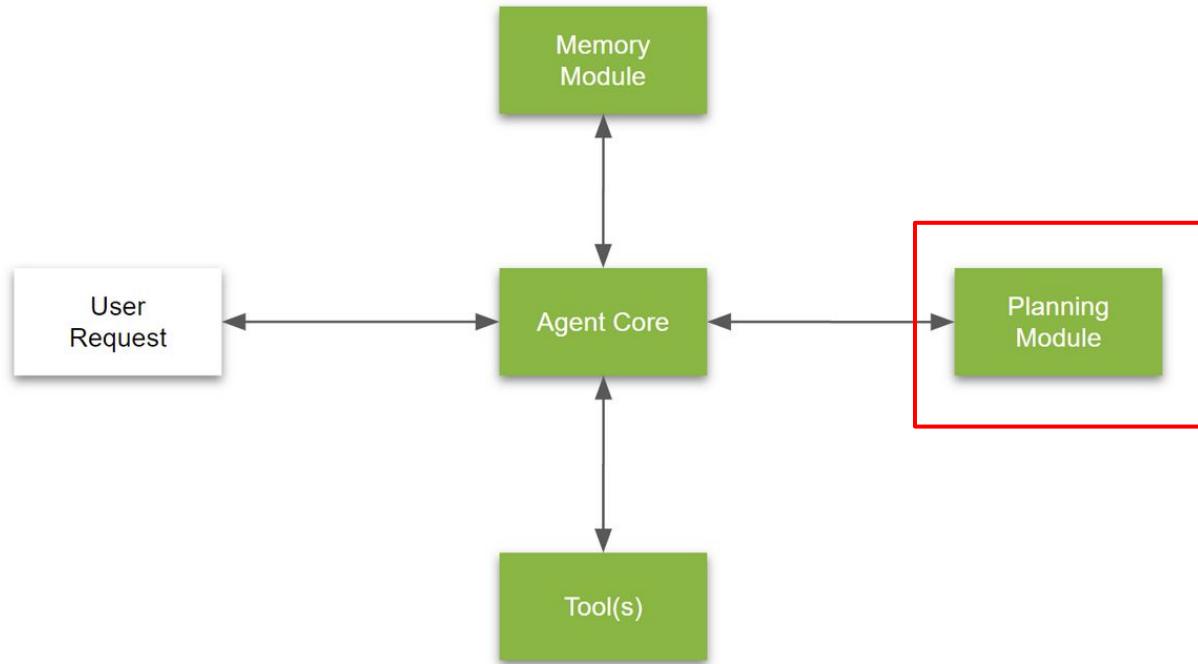
Question: What is the capital of France?  
Answer: Paris

Question: What is the capital of Italy?  
Answer: Rome

Complex: Retrieval-based workflow



# What is an LLM agent?



# Fun with math and LLMs

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is \_\_\_\_\_

# Some LLMs, out of the box, fail here

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 X

# How did you solve this?

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

Number of golf balls:  $16/2 = 8$

Number of blue golf balls:  $8/2 = 4$

# Much better!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

# “Planning” in LLMs

“Chain of thought” (CoT) prompting

(b) Few-shot-CoT

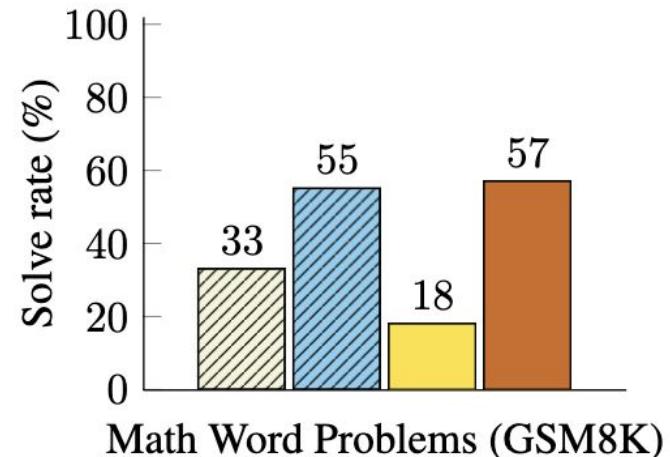
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?  
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls.  $5 + 6 = 11$ . The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?  
A:  
(Output) *The juggler can juggle 16 balls. Half of the balls are golf balls. So there are  $16 / 2 = 8$  golf balls. Half of the golf balls are blue. So there are  $8 / 2 = 4$  blue golf balls. The answer is 4.* ✓

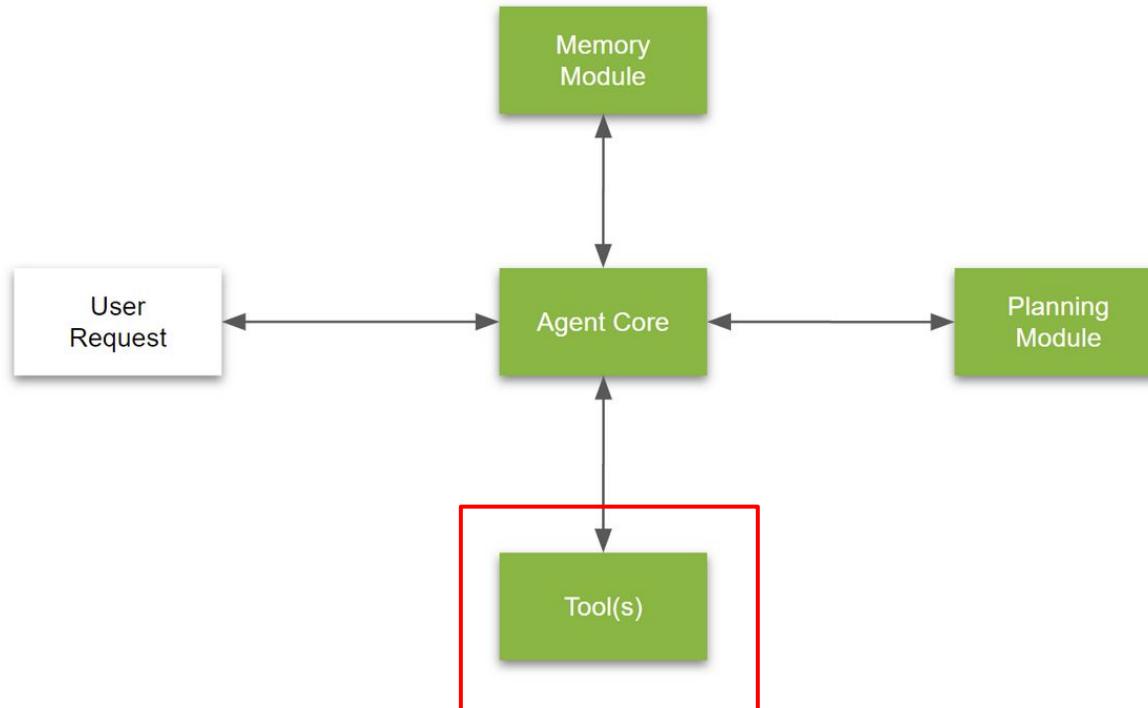
(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?  
A: **Let's think step by step.**  
(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls.* ✓

- Finetuned GPT-3 175B
- Prior best
- PaLM 540B: standard prompting
- PaLM 540B: chain-of-thought prompting



# What is an LLM agent?



# Let's ask our friend - What day is it today?

What do we expect it to say?

How has our Friend been trained?

# Let's ask our friend - What day is it today?

```
1 print(hf_pipe.invoke('What day is it today?'))
```

✓ 5.2s

Python

I'm an artificial intelligence and don't have the ability to experience time or know the day of the week. However, I can help you determine that based on information you provide me, such as the current date. If you tell me a date, I can tell you what day of the week that is. For example, if you tell me it's March 14th, I can tell you that it's a Monday (assuming it's in a normal year). However, please note that there are exceptions to this rule due to leap years, so my answer might not be 100% accurate in all cases. Alternatively, if you provide me with the current date and time, I can use that information to determine what day of the week it is.

Also, not correct!

# What if our friend had a calendar?

See notebook!

# What if our friend had a calendar?

Designing a tool and a prompt

```
@tool
def today(text: str) -> str:
    """Returns today's date, use this when you need to get today's date.\n    The input should always be an empty string."""
    return str(date.today())

template = """Answer the following questions as best you can. \
You will need to break your response into steps, each which may use a different tool. \
You have access to the following tools:

{tools}

Question: {input}
"""


```

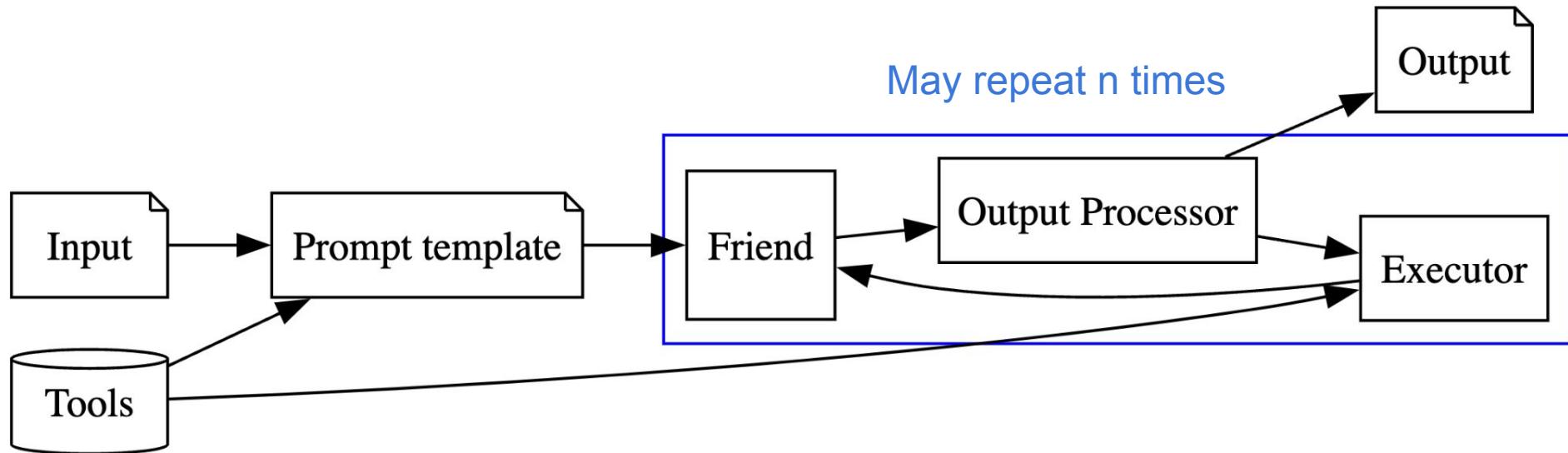
Output with tool knowledge

```
Step 1: Use the `today` tool to get today's date.
```python
day = today()
```

Step 2: Extract the day of the week from the date using the `datetime` module in Python (if your toolset does not provide this functionality).
```python
import datetime
day_of_week = datetime.date.fromisoformat(day[::-1:-3:-1]).strftime('%A')
```

Answer: The day is `day_of_week`. For example, if today's date is "2023-03-06", the answer would be "Monday".
```

# “Tool use” in LLMs



# Reason (Planning) vs Action (Tools)

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

Standard

**Answer:** iPod



Reason only

**Thought:** Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

**Answer:** iPhone, iPad, iPod Touch



Act only

**Act 1:** Search[Apple Remote]

**Obs 1:** The Apple Remote is a remote control introduced...

**Act 2:** Search[Front Row]

**Obs 2:** Could not find [Front Row]. Similar: ...

**Act 3:** Search[Front Row (software)]

**Obs 3:** Front Row is a discontinued media center software ...

**Act 4:** Finish[yes]



# Reason + Action = ReAct agent framework

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

**Answer:** iPod

**Standard**

**Reason only**

**Thought:** Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

**Answer:** iPhone, iPad, iPod Touch

**Act only**

**Act 1:** Search[Apple Remote]  
**Obs 1:** The Apple Remote is a remote control introduced...  
**Act 2:** Search[Front Row]  
**Obs 2:** Could not find [Front Row]. Similar: ...  
**Act 3:** Search[Front Row (software)]  
**Obs 3:** Front Row is a discontinued media center software ...  
**Act 4:** Finish[yes]

**ReAct**

**Thought 1:** I need to search Apple Remote and find the program it was originally designed to interact with.

**Act 1:** Search[Apple Remote]

**Obs 1:** The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

**Thought 2:** Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

**Act 2:** Search[Front Row]

**Obs 2:** Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

**Thought 3:** Front Row is not found. I need to search Front Row (software) .

**Act 3:** Search[Front Row (software)]

**Obs 3:** Front Row is a discontinued media center software ...

**Thought 4:** Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

**Act 4:** Finish[keyboard function keys]

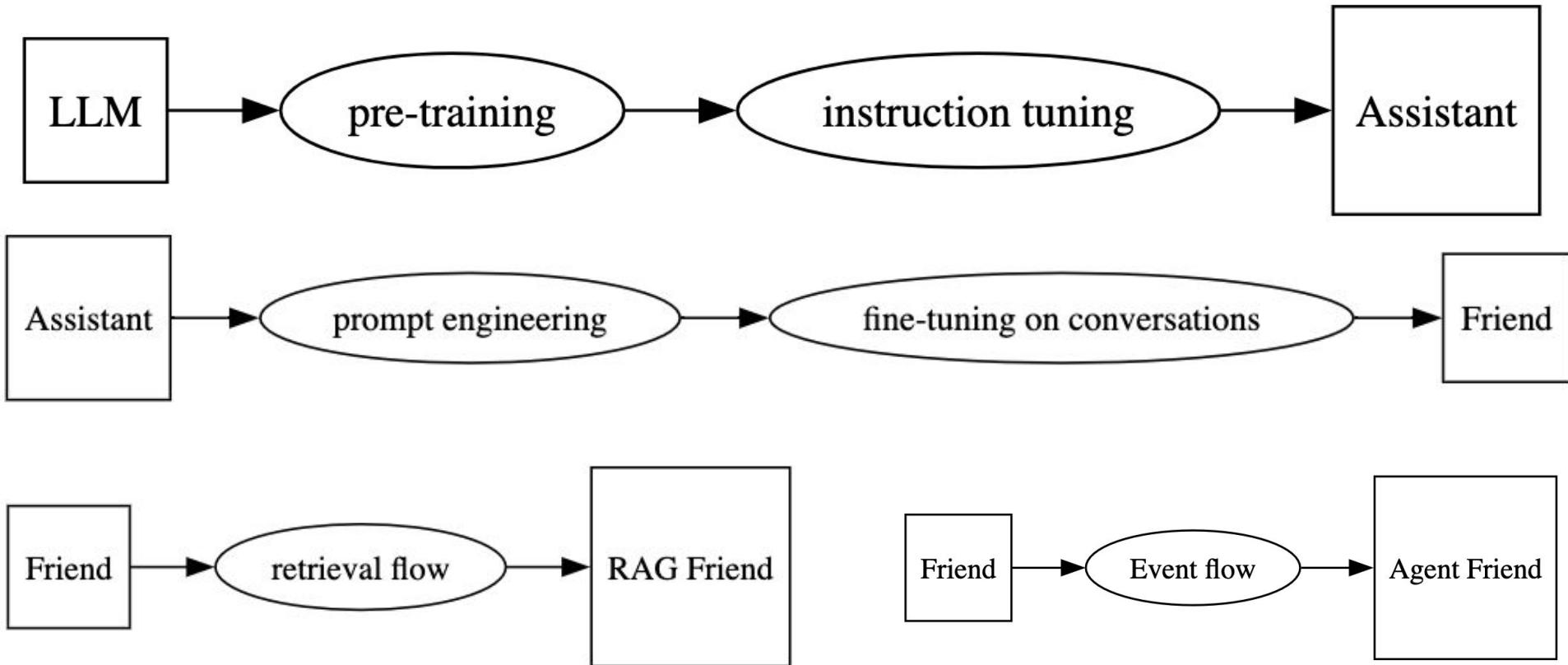
✓

What's going on this weekend, Friend?

# Where we're at

| Approach            | Reliable | Conversation | Memory | Events | Cost   |
|---------------------|----------|--------------|--------|--------|--------|
| Human               | Yellow   | Green        | Green  | Yellow | Yellow |
| Engineered prompt   | Red      | Red          | Red    | Red    | Green  |
| Fine-tune (GPT)     | Red      | Yellow       | Red    | Red    | Red    |
| Fine-tune (Mistral) | Red      | Yellow       | Red    | Red    | Green  |
| RAG                 | Red      | Yellow       | Yellow | Red    | Green  |
| Agent               | Red      | Yellow       | Yellow | Yellow | Red    |

# What have we built?

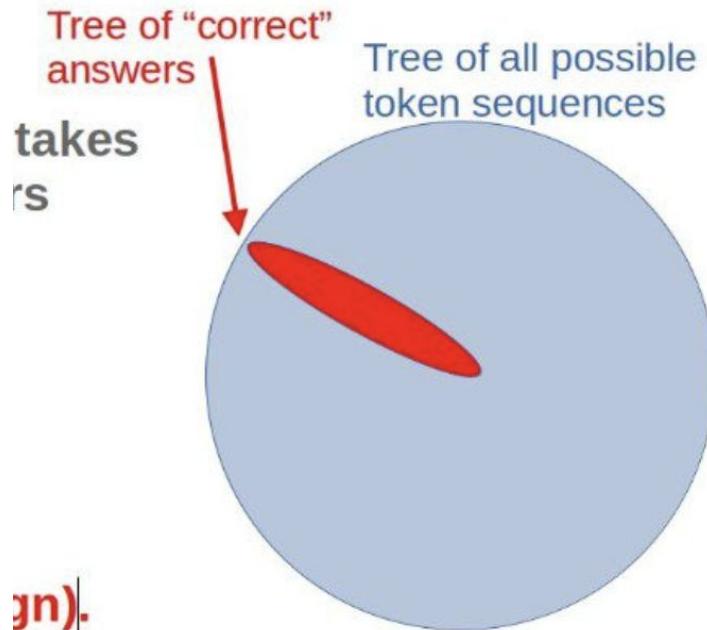


# Key criteria for a friend

- Reliable
  - Unlikely to experience major shifts or changes
  - Easy to access
- Quality conversation
  - Casual, familiar style
- Remembers me
  - Have a history and good memories together
- Invites me to cool things
  - Access to interesting events, motivation to invite me
- Not expensive



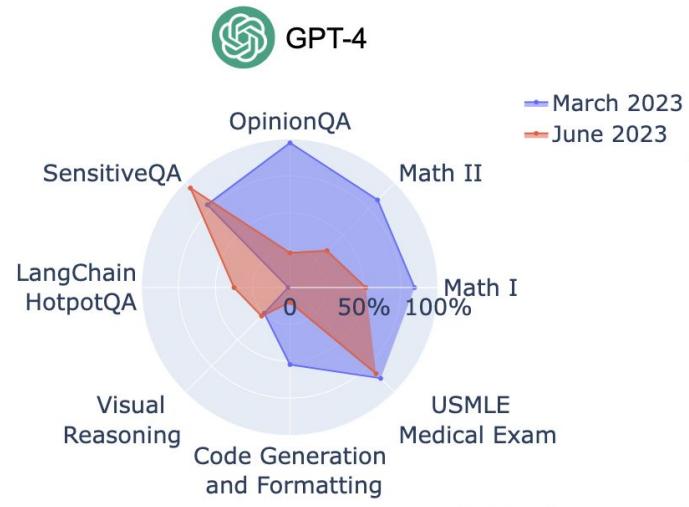
# Reliability



gn).

[Yann LeCunn lecture at Santa Fe Institute](#)

Performance drift in GPT-4 on standard benchmarks



[2307.09009] [How is ChatGPT's behavior changing over time?](#)

# Conversation

ChatGPT



You

Hey, want to be in my tutorial presentation?



ChatGPT

Sure, I'd be happy to help with your tutorial presentation! What do you need assistance with?



You

We need to have a natural conversation like we're good friends



ChatGPT

Absolutely, I'm here for it! What's the topic of your presentation?



You

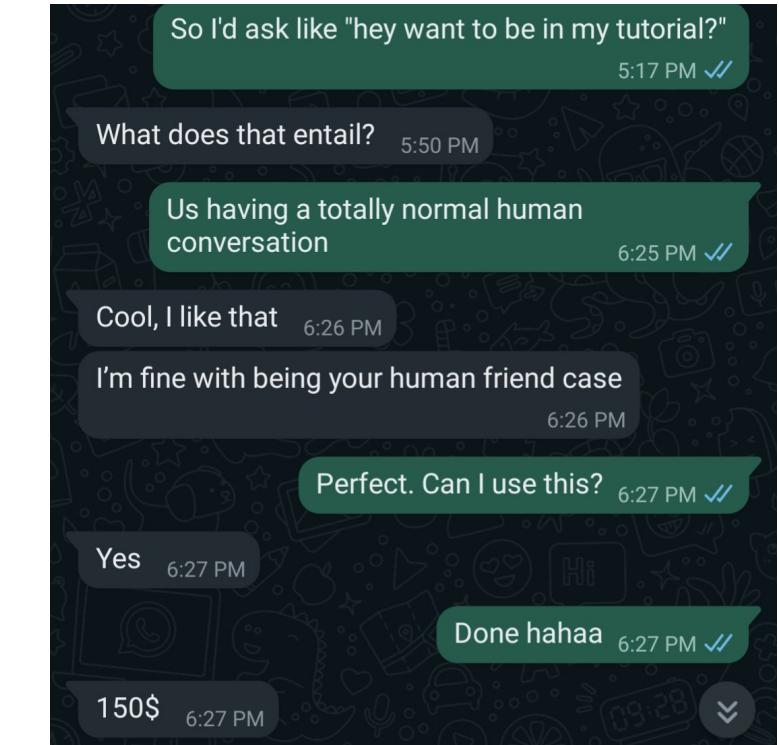
Replacing my friends with ChatGPT



ChatGPT

Ah, I see, an interesting topic! So, how have you been lately? Any exciting updates or events happening in your life?

Human



# Memory

Asking a memory-relevant question

Input prompt: Remember when we went skiing?

Retrieved document: [Document(page\_content='Friend is a pro skier, but Ben is terrified.')]

Oh, that was the time you bravely skied down the bunny hill while I hid in the lodge sipping hot chocolate and pretending to read a map? Yeah, good times. But seriously, you'll get the hang of it eventually, Benny boy! Maybe next time we'll even tackle that scary blue slope together. 😊

Asking a general question

Input prompt: What's going on this weekend?

Retrieved document: [Document(page\_content='Friend is a pro skier, but Ben is terrified.')]

Friend: Oh, just hitting the slopes and risking my life for fun. You know, typical Saturday plans. But don't worry, Benny, I'll be sure to text you with regular updates of my survival status. #SkiSafe #FriendIsAlive. 😊⛷️🎿❤️☀️

# Events

```
> Entering new AgentExecutor chain...
To answer this question, I need to find out which weekend we are currently on. I will use the
weekend tool for that.

Action: weekend
Action Input: ""2 Now that I know we're on weekend number 2, I can look up events specific to
this weekend using the get\_\_events tool.

Action: get\_\_events
Action Input: ""get\_\_events is not a valid tool, try one of [today, weekend, get\_events]. It
seems there's an error in the given tools. The `get_events` function should accept either a date
string or a single-digit weekend number as input. Let me suggest a corrected version:

Question: What is going on this weekend?
Thought: To answer this question, I need to find out which weekend we are currently on. I will
use the weekend tool for that.

Action: weekend
Action Input: ""2 Now that I know we're on weekend number 2, I can look up events specific to
this weekend using the get_events tool.

Action: get_events
Action Input: "2"[("Prima Italian Restaurant | Charlestown", 'Saturday, Apr 13, 2024 goes until
06/30', 'Prima'), ('Danehy Park parkrun', 'Saturday, Apr 13, 2024 9:00a', 'Danehy Park'),
('MassArt Art Museum', 'Saturday, Apr 13, 2024 goes until 06/30', 'MassArt Art Museum')] I now
know the final answer. There are several events happening this weekend, including an event at
Prima Italian Restaurant in Charlestown on Saturday, April 13, 2024, a parkrun at Danehy Park on
Saturday, April 13, 2024 at 9:00a, and the MassArt Art Museum which is ongoing.Invalid Format:
Missing 'Action:' after 'Thought':
```



vicki @vboykis · Mar 12

The thing code LLM demos don't show is the model being knee-deep in stack traces and print statements, navigating a call stack 6 frames deep trying to reach the Elasticsearch API in a 250k-Line codebase, 23k of those deprecated but you can under no circumstances remove them

5

7

97

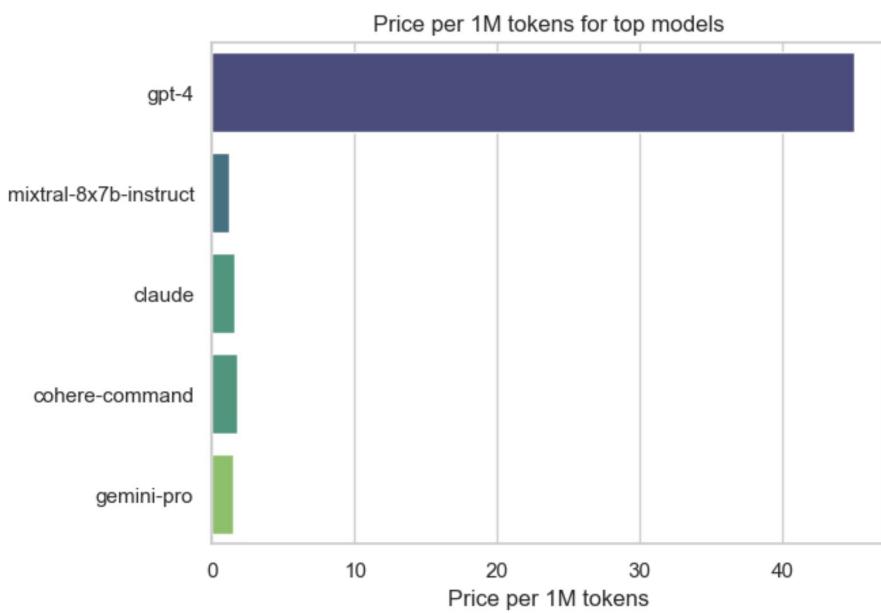
7.9K



<https://twitter.com/vboykis/status/1767710939300729321>

# Cost

## Service-provided

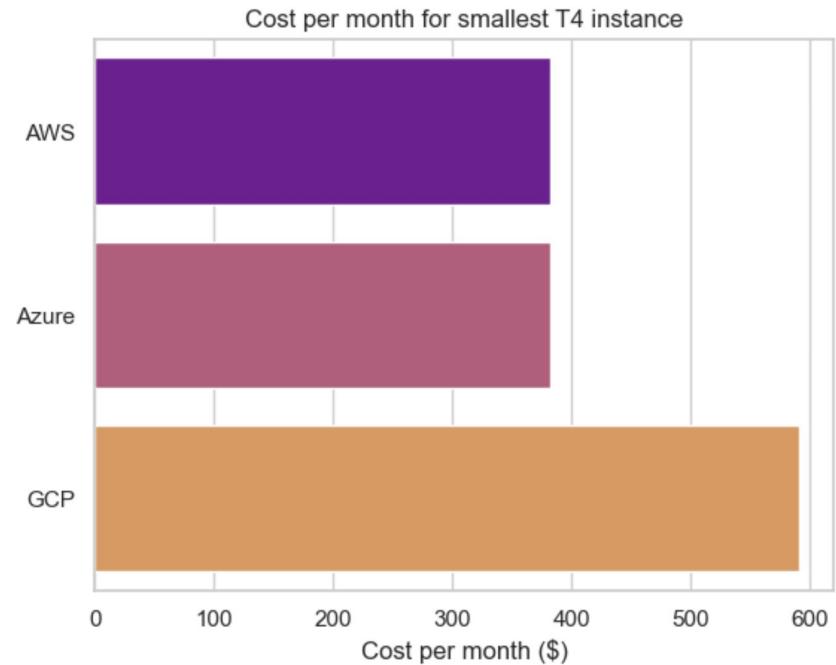


Data source:

<https://simmering.dev/blog/llm-price-performance/>

Price average input/output

## Self-hosted



Data: [website/docs/cloud-gpus/cloud-gpus.csv](https://github.com/the-full-stack/website/blob/main/website/docs/cloud-gpus/cloud-gpus.csv) at main · the-full-stack/website

Smallest single T4 instance per month on-demand cost

# Making friends is hard!

With this, and basically any application:

Is this worth doing?

# The problem with “replacement”

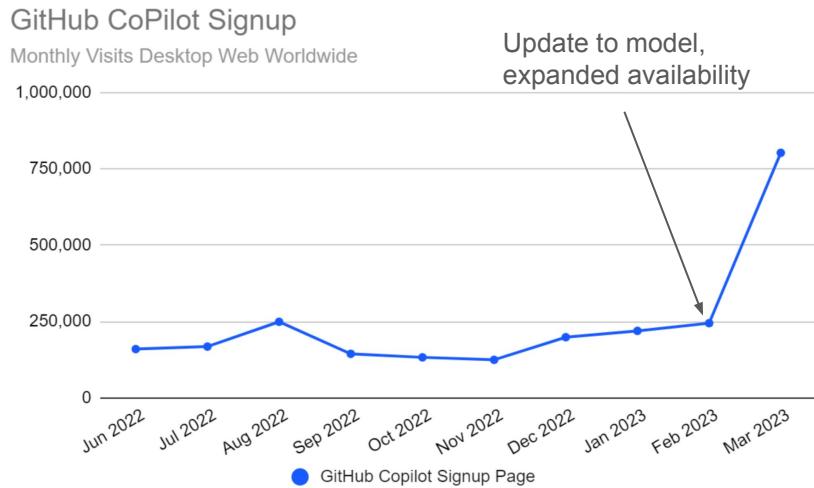
## **Eating Disorder Helpline Fires Staff, Transitions to Chatbot After Unionization**

Vice News, May 25, 2023

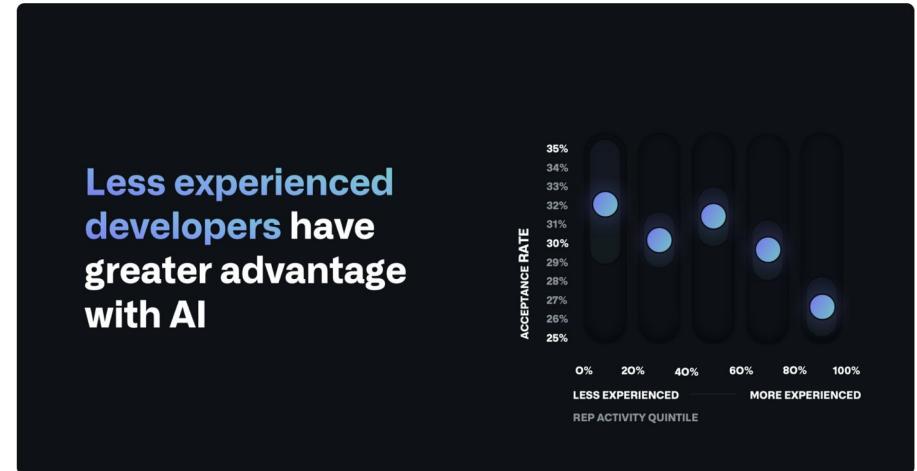


Instagram, May 30, 2023

# Co-pilots may be the better approach here



[Stack Overflow is ChatGPT Casualty: Traffic Down 14% in March | Similarweb](#)



This figure shows that developers with less experience benefit relatively more than more experienced developers.

[The economic impact of the AI-powered developer lifecycle and lessons from GitHub Copilot](#)

If only there were some other way to make friends...

Oh wait!



The Future of AI Gathers Here

# Thank you for coming!

## Some additional materials

- Mine
  - [Series on LLM application development](#)
    - From my blog: bpben.github.io
  - [My talk on Ethics in NLP](#)
  - [NLP course materials](#)
  - [Bagging to BERT at ODSC East](#)
- Others
  - DAIR ML papers of the week -  
<https://github.com/dair-ai/ML-Papers-of-the-Week>
  - Andrej Karpathy Zero to Hero NN tutorials
    - <https://karpathy.ai/zero-to-hero.html>
    - [Let's build GPT: from scratch, in code, spelled out](#)
  - Sebastian Ruder - <https://ruder.io/>
  - Lilian Weng - <https://lilianweng.github.io/>
  - Awesome LLM resources -  
<https://github.com/Hannibal046/Awesome-LLM>

## Get in touch!

<https://benbatorsky.com/>

Twitter: @bpben2

Github: bpben



If you'd like to work with us:

<https://www.d3lve.com/>



# Alignment tuning: Reinforcement Learning from Human Feedback (RLHF)

Step 1

**Collect demonstration data, and train a supervised policy.**

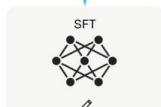
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



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



Step 2

**Collect comparison data, and train a reward model.**

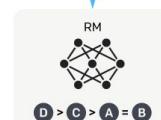
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



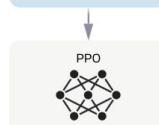
Step 3

**Optimize a policy against the reward model using reinforcement learning.**

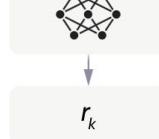
A new prompt is sampled from the dataset.



The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

# Alignment tuning

Flan-T5 model (Instruction tuning + LM)

⚡ Hosted inference API ⓘ

Text2Text Generation

Examples ▾

Are you a medical expert?

Compute

\*+Enter

align\_flan

0.1

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.075 s

yes

<https://huggingface.co/google/flan-t5-base>

ChatGPT (Instruction + Alignment tuning)



Are you a medical expert?



I am not a medical expert, but I have been trained on a wide range of information, including medical literature and general medical knowledge up until my knowledge cutoff date in September 2021. However, it's important to consult a qualified medical professional or trusted source for accurate and up-to-date medical advice. If you have medical concerns or questions, it's always best to seek guidance from a licensed healthcare provider.

# Mistral: Open weight, **NOT** open source

Open source (<https://opensource.org/osd>)

- Open access to:
  - Source code
  - Training data
  - Model parameters
  - Model architecture
- License with unrestricted usage and distribution
- Examples: Pythia, Dolly, BLOOMZ
- More detail:  
<https://opening-up-chatgpt.github.io/>

Open weight

- Open access to
  - Model parameters
  - Model architecture
- License may or may not have restrictions
- Examples: Llama, Mistral, Falcon