

CSCSE 689: Special Topics in Trustworthy NLP

Lecture 8: Natural Language Processing Basics (7)

Kuan-Hao Huang
khhuang@tamu.edu



(Some slides adapted from Chris Manning, Karthik Narasimhan, Graham Neubig)

Paper Summary

- A summary of **two** papers is due **next Monday before lecture**
- Submit summary to **Canvas**
- Page limit: **1 page**
- The summary should include
 - A brief overview of the main objectives and contributions of the paper
 - Key methodologies and approaches used in the study
 - Significant findings and results
 - Strengths and weaknesses of the paper

Paper Summary

Choose 1 paper here

- [Generating Natural Language Adversarial Examples](#), EMNLP 2018
- [BERT-ATTACK: Adversarial Attack Against BERT Using BERT](#), EMNLP 2020
- [Universal Adversarial Triggers for Attacking and Analyzing NLP](#), EMNLP 2019

- [Certified Robustness to Adversarial Word Substitutions](#), EMNLP 2019
- [Towards Robustness Against Natural Language Word Substitutions](#), ICLR 2021
- [Universal and Transferable Adversarial Attacks on Aligned Language Models](#), arXiv 2023

Choose 1 paper here

Course Project – Proposal

- Team: 1~2 people
- Due: 9/25
- Page limit: 2 pages
- Format: [ACL style](#)
- The proposal should include
 - The topic you choose
 - An introduction to the task
 - Evaluation metrics
 - The dataset, models, and approaches you plan to use

Invited Talk

W8	10/7	Fall Break (No Class)	
	10/9	Invited Talk (Remote)	Invited Talk
	10/11	Team Project Highlights	Team Project Highlights

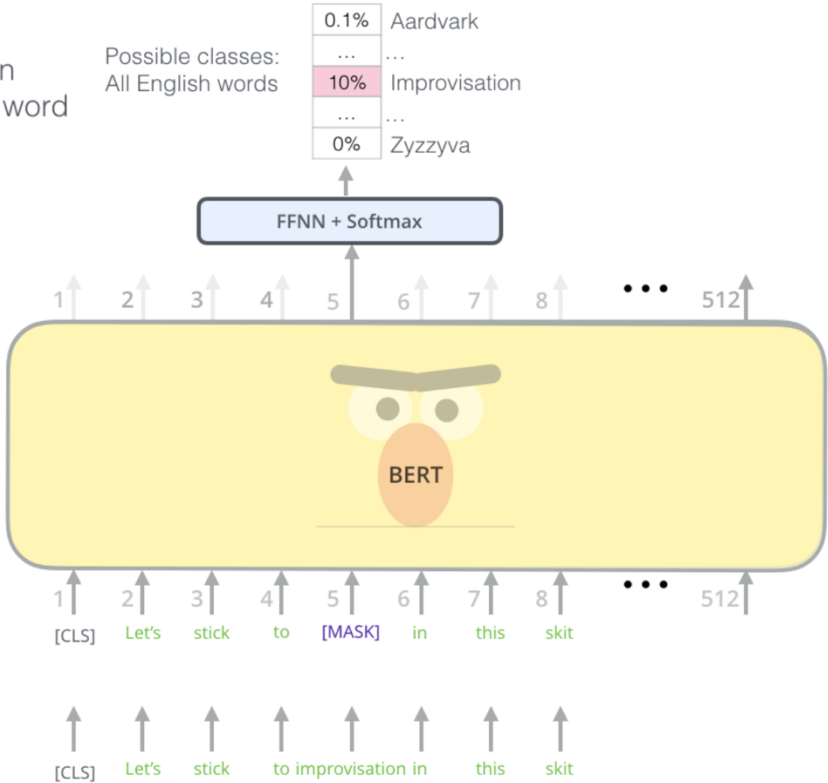
Lecture Plan

- Natural Language Processing Basics
- Large Language Models
- Prompting
- In-Context Learning
- Instruction Tuning

Recap: Encoder-Only Pre-Training

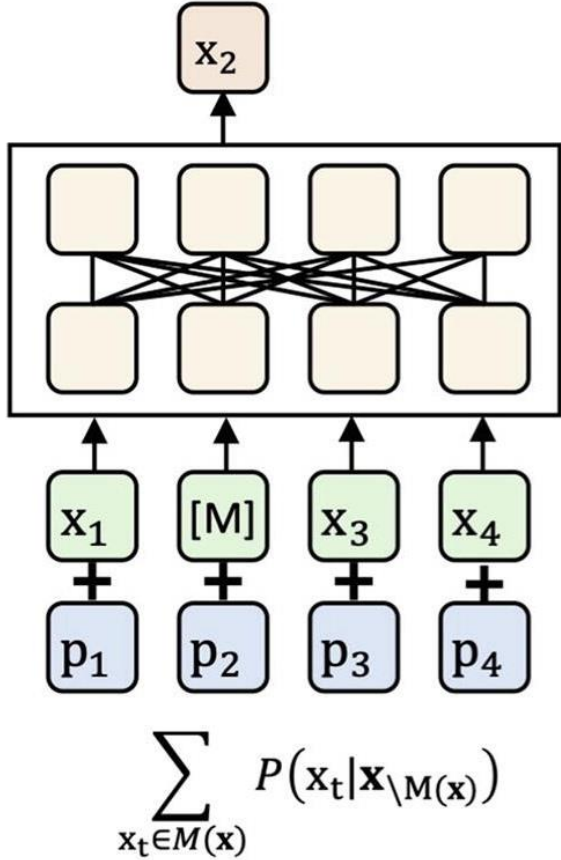
- BERT, RoBERTa

Use the output of the masked word's position to predict the masked word



Randomly mask 15% of tokens

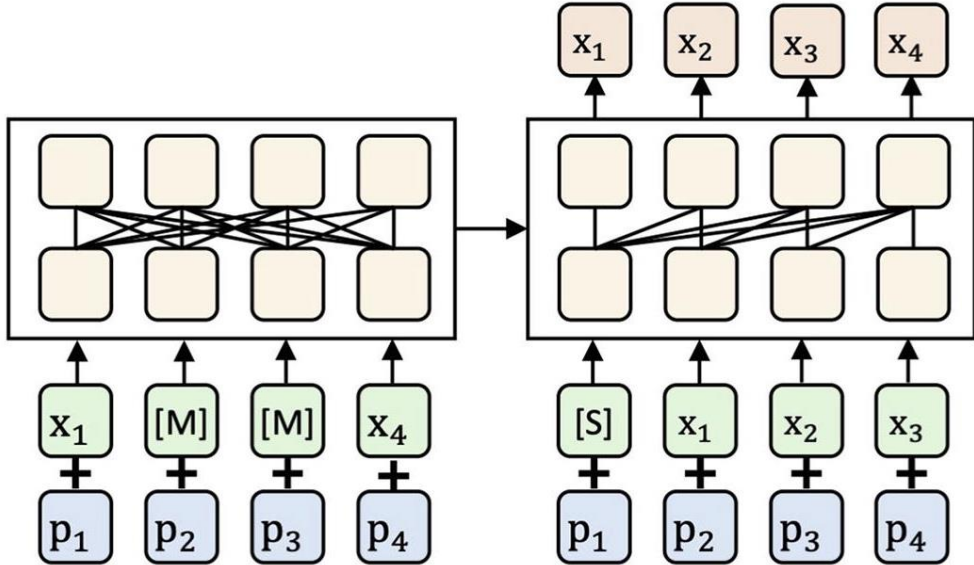
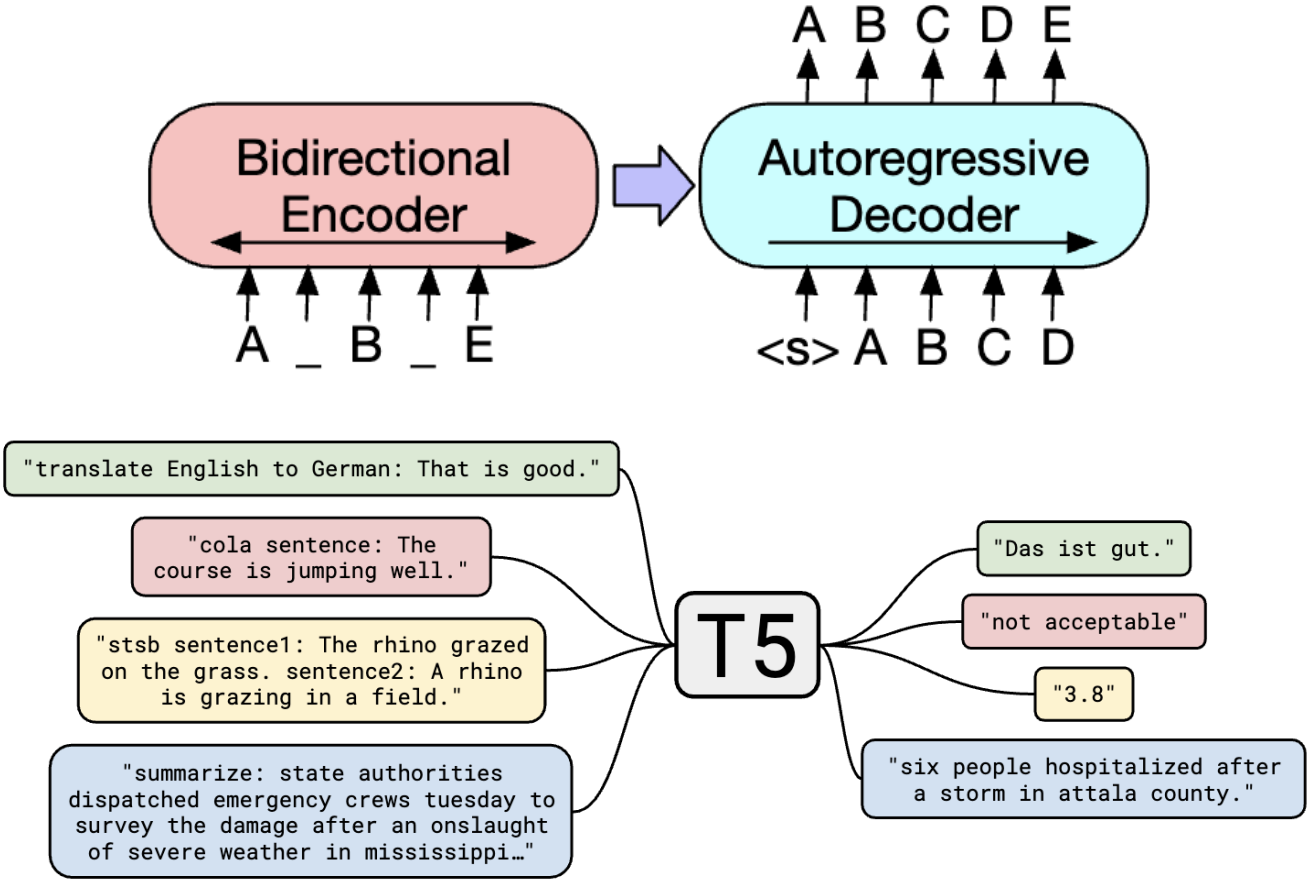
Input



Encoder only

Recap: Encoder-Decoder Pre-Training

- BART, T5

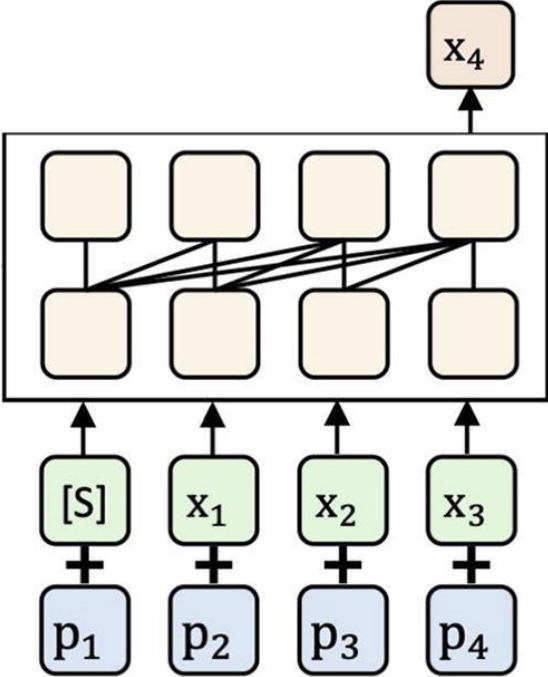
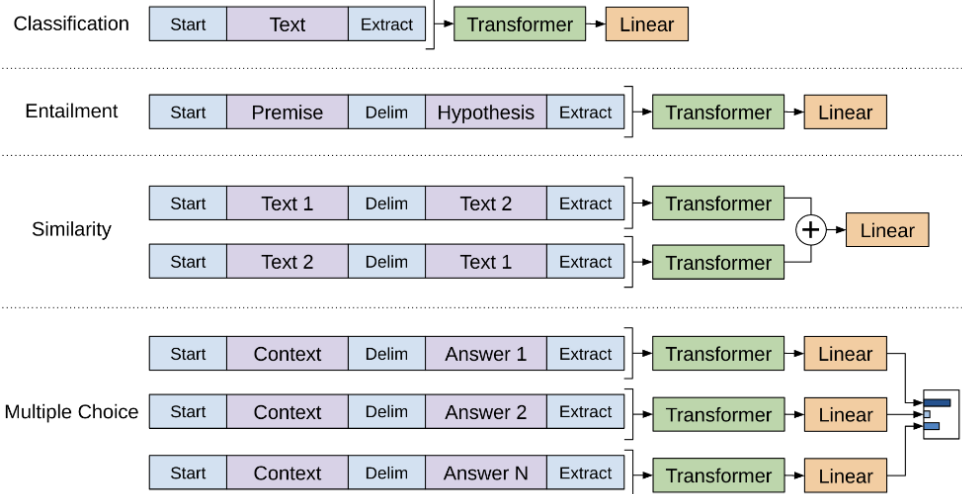


$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t}, \mathbf{x}_{\setminus i:j})$$

Encoder-decoder

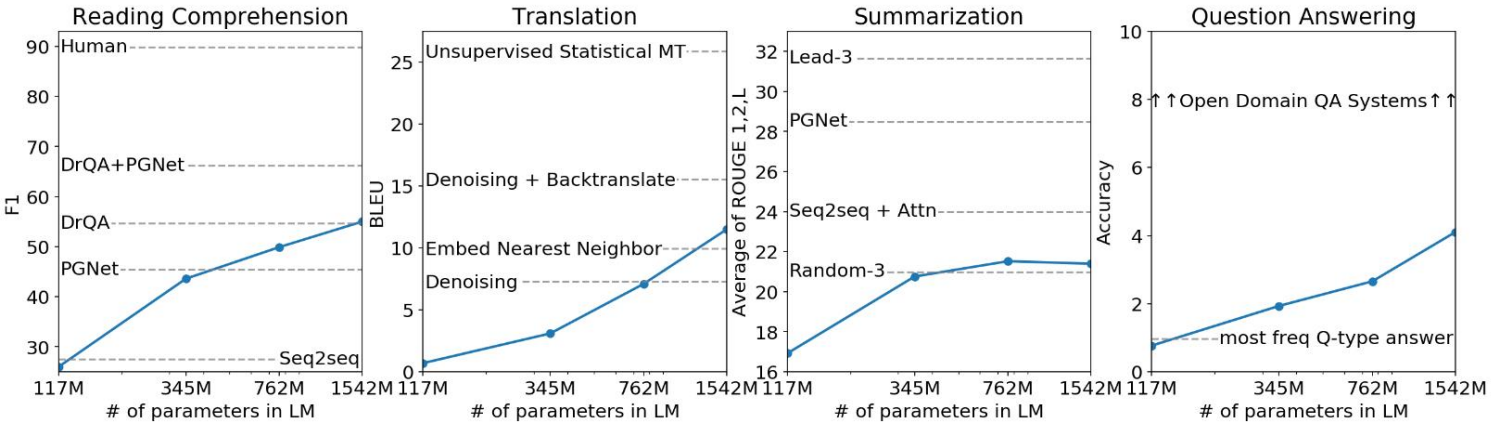
Recap: Decoder-Only Pre-Training

- GPT

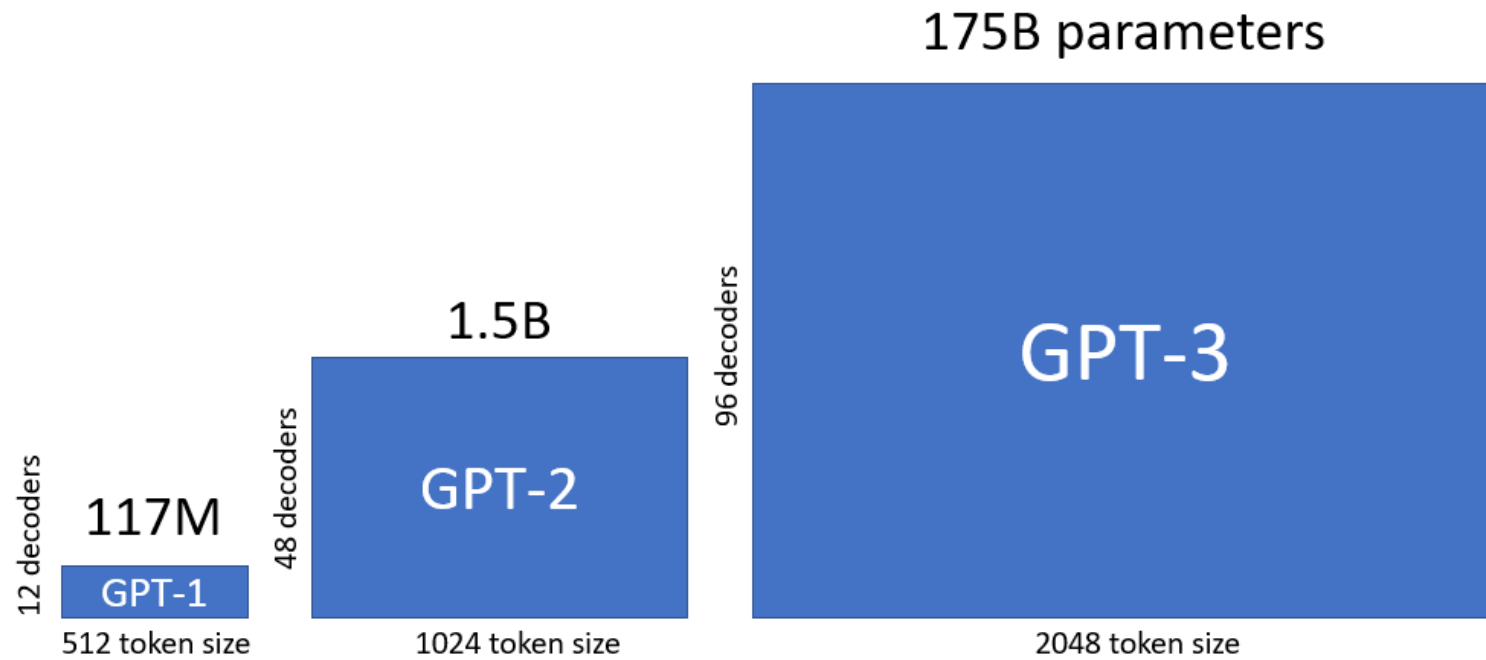


$$\sum_{t=1}^T P(x_t | \mathbf{x}_{<t})$$

Decoder only



Scaling Is The Key



Zero-Shot Prompting

- Prompt → Completion
 - Continue writing

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Zero-Shot Prompting

- Prompt → Completion
 - Continue writing

Prompt

Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

Completion

Language Modeling

Binge ... on | - | and | of | is

Binge **drinking** ... is | and | had | in | was

Binge drinking **may** ... be | also | have | not | increase

Binge drinking may **not** ... be | have | cause | always | help

Binge drinking may not **necessarily** ... be | lead | cause | results | have

Binge drinking may not necessarily **kill** ... you | the | a | people | your

Binge drinking may not necessarily kill **or** ... even | injure | kill | cause | prevent

Binge drinking may not necessarily kill or **even** ... kill | prevent | cause | reduce | injure

Binge drinking may not necessarily kill or even **damage** ... your | the | a | you | someone

Binge drinking may not necessarily kill or even damage **brain** ... cells | functions | tissue | neurons

Binge drinking may not necessarily kill or even damage brain **cells,** ... some | it | the | is | long

Training language models in a large scale → Large Language Models (LLMs)

Zero-Shot Prompting

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

Prompt

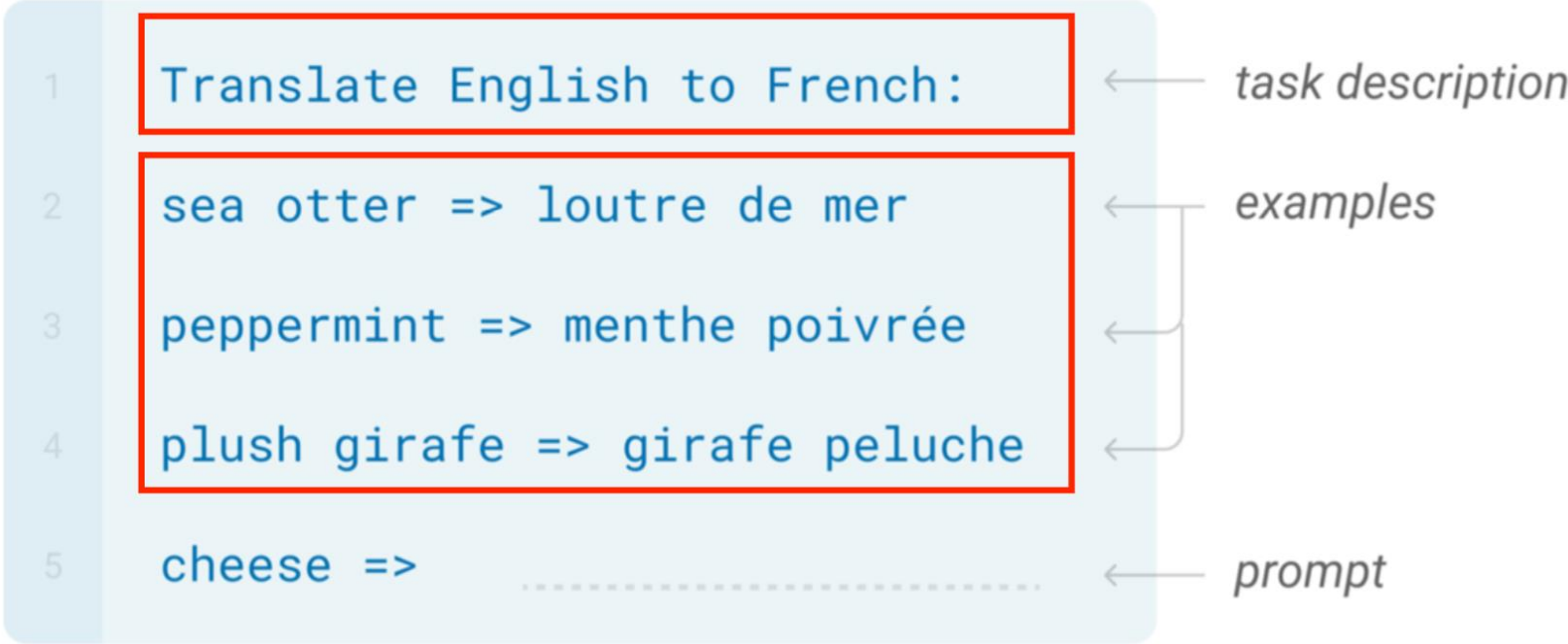
Stephen Curry's clutch barrage seals another Olympic gold for USA. The topic of the above sentence is

sport.

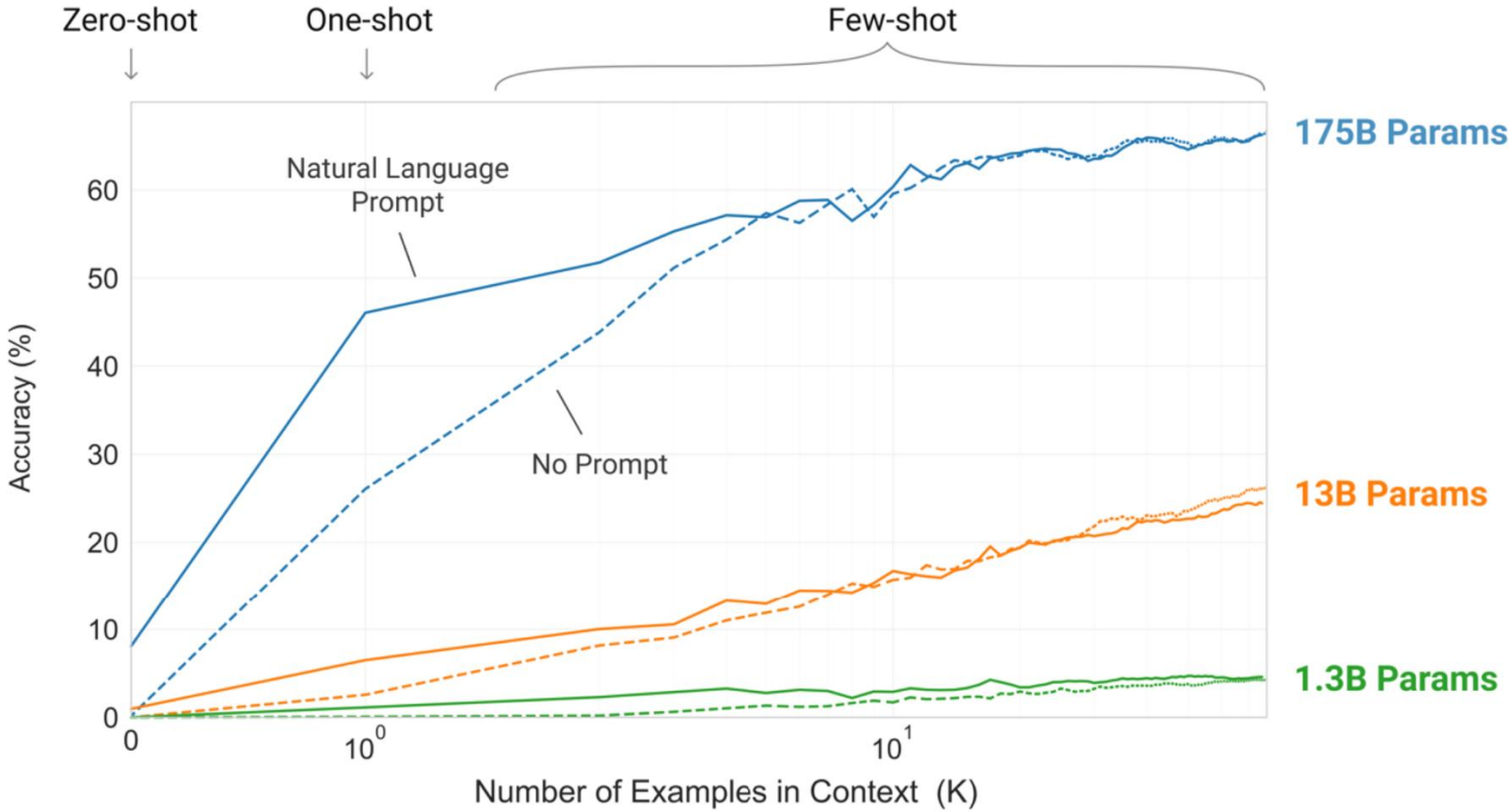
Completion

In-Context Learning

- Few-shot prompting



In-Context Learning

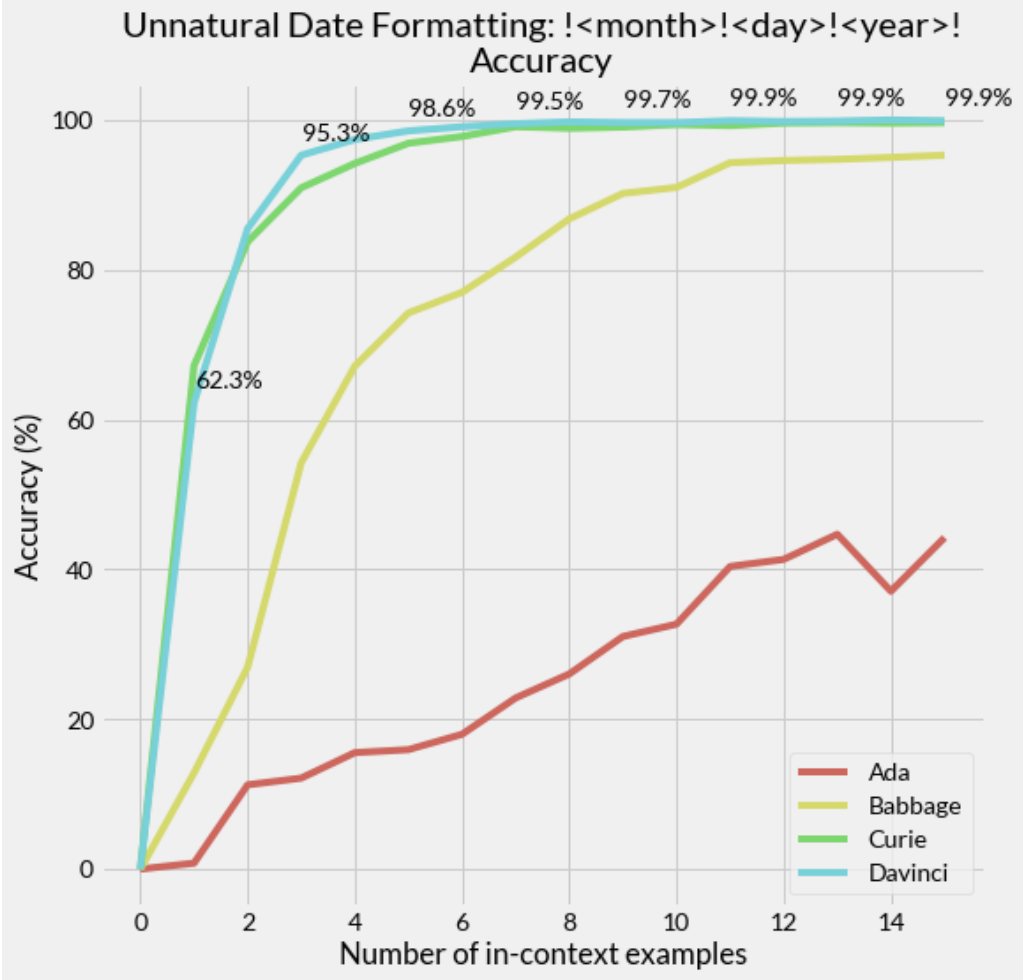


In-Context Learning

Input: 2014-06-01
Output: !06!01!2014!
Input: 2007-12-13
Output: !12!13!2007!
Input: 2010-09-23
Output: !09!23!2010!
Input: **2005-07-23**
Output: !07!23!2005!

in-context examples
test example

!07!23!2005! *model completion*

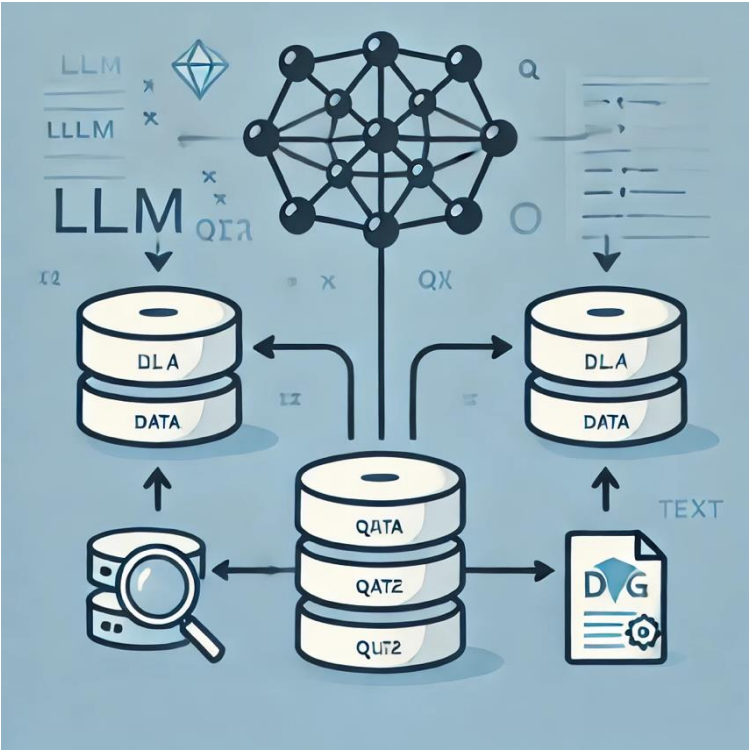


A New Way to Use NLP Models

- Task-specific features + task-specific model
- General embeddings + task-specific model
- General embeddings + general model + task-specific fine-tuning
- General embeddings + general model + **task-specific prompting**

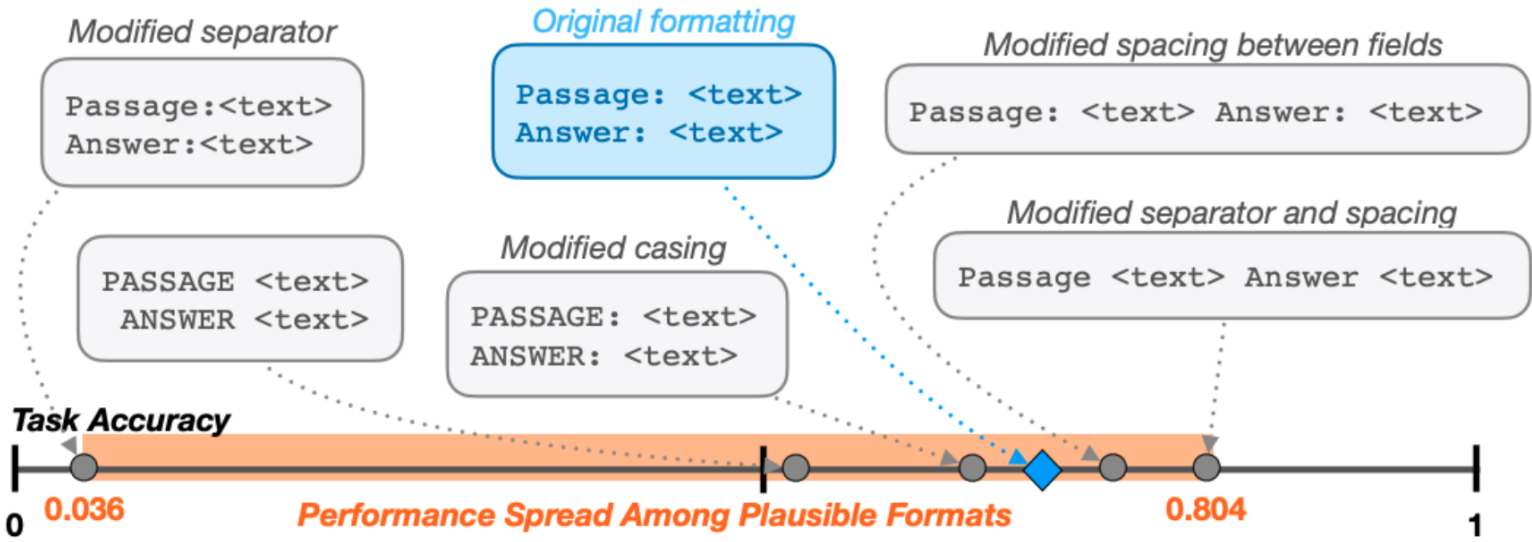
Large Language Models as Database

- Pre-training → building knowledge database from corpus
- Prompting → query database



Prompt Engineering

- LLMs are sensitive to prompts
- Prompt engineering: search for better prompts



Prompt Engineering: Chain-of-Thought

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**

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. ✓*

Prompt Engineering: Chain-of-Thought

(a) Few-shot

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: 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 answer is 8. ✗

(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. ✓

(c) Zero-shot

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 ✗

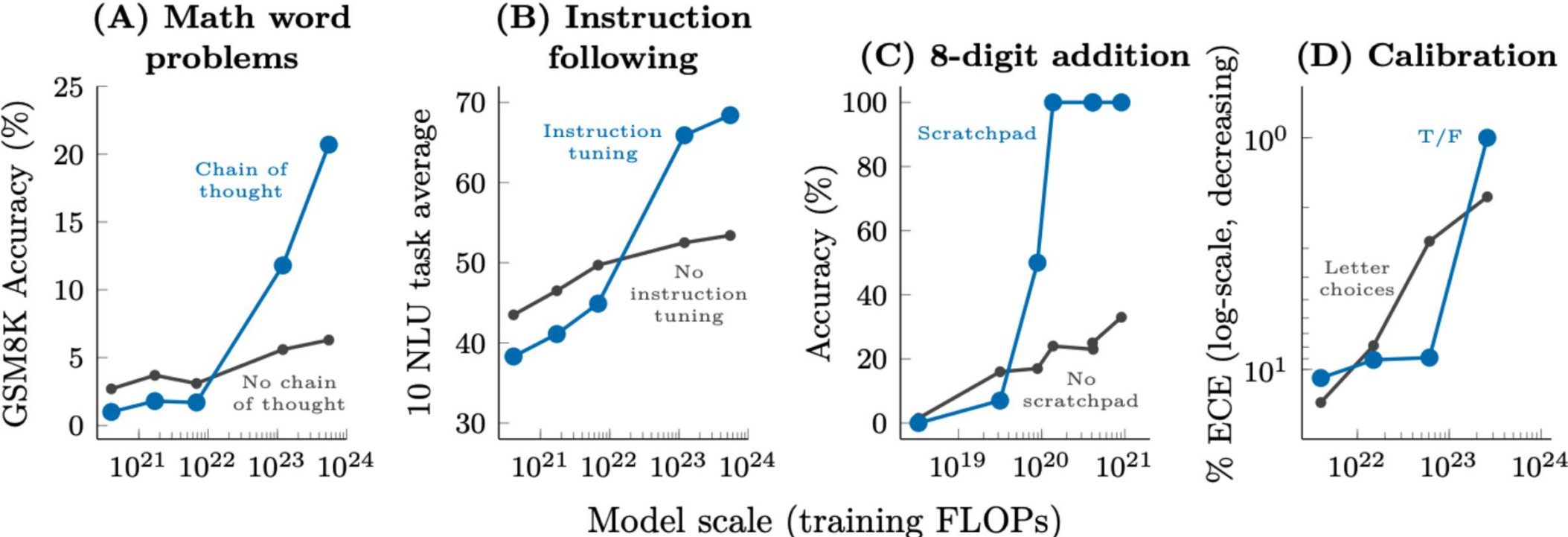
(d) Zero-shot-CoT (Ours)

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. ✓

Scaling Is The Key



Lecture Plan

- Natural Language Processing Basics
- Large Language Models
- Prompting
- In-Context Learning
- Instruction Tuning

Instruction Tuning

- Continuing writing does not always work

Prompt

This place is incredible! The lobster is the best I've ever had. The sentiment of the above sentence is

positive.

Completion

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

Instruction Tuning

- Training LLMs to following **human thoughts**

Finetune on many tasks (“instruction-tuning”)

<p>Input (Commonsense Reasoning)</p> <p>Here is a goal: Get a cool sleep on summer days. How would you accomplish this goal? OPTIONS: -Keep stack of pillow cases in fridge. -Keep stack of pillow cases in oven.</p> <p>Target</p> <p>keep stack of pillow cases in fridge</p>	<p>Input (Translation)</p> <p>Translate this sentence to Spanish: The new office building was built in less than three months.</p> <p>Target</p> <p>El nuevo edificio de oficinas se construyó en tres meses.</p>
---	---

Sentiment analysis tasks

Coreference resolution tasks

...



Inference on unseen task type

Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.
Hypothesis: It's not certain how many lessons you'll learn by your thirties.
Does the premise entail the hypothesis?
OPTIONS:
-yes -it is not possible to tell -no

FLAN Response

It is not possible to tell

Instruction Tuning

- Training LLMs to following **human thoughts**

Annotated task definitions

You will be given two pieces of text... One of them is simpler ...
You are expected to output 'Text one' if the first sentence is simpler.
Otherwise output 'Text two'.

Given a sentence with a missing word, pick the answer option that best fills out the missing word in the sentence. Indicate each answer with its index ('a', 'b', 'c', 'd').

Given a document, generate a short title of the document. The title should convey the main idea/event/topic about which the document is being written.

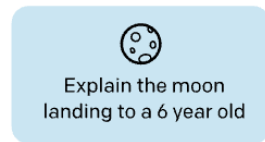
Category	Description
Input Content	Primary description of the task input
<u>Additional Input Content</u>	Additional details on task input
Action Content	Action to perform for task
Input Mention	Mentions of input within action content
Output Content	Primary description of task output
<u>Additional Output Content</u>	Additional details on task output
Label List	Task output labels (classification only)
Label Definition	Task Label definitions (classification only)

InstructGPT: Instruction Tuning + Human Feedback

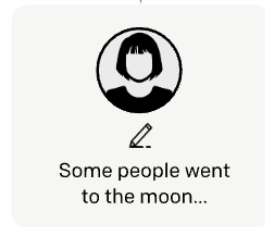
Step 1

Collect demonstration data, and train a supervised policy.

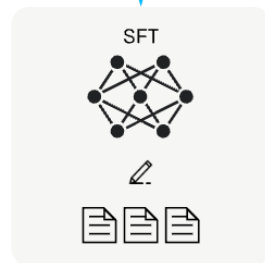
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



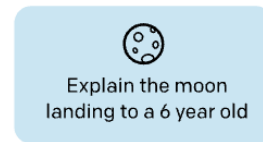
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

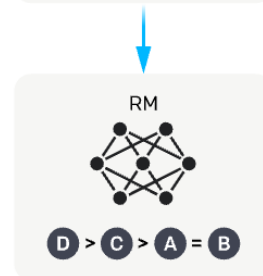
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



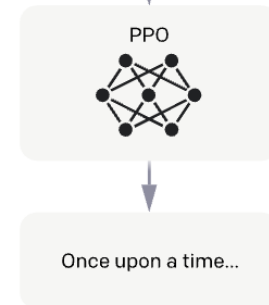
Step 3

Optimize a policy against the reward model using reinforcement learning.

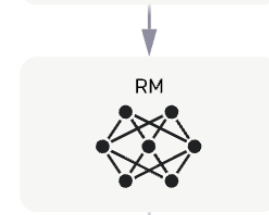
A new prompt is sampled from the dataset.



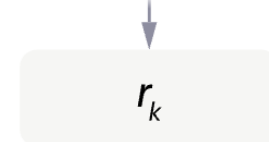
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



InstructGPT: Instruction Tuning + Human Feedback

- Supervised Fine-Tuning (SFT)

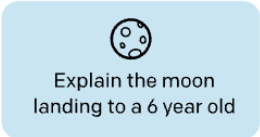
Use-case	Prompt	Use-case	(%)	Number of Prompts		
Brainstorming	List five ideas for how to regain enthusiasm for my career	Generation	45.6%	SFT Data		
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.	Open QA	12.4%	split	source	size
Rewrite	This is the summary of a Broadway play: "" {summary} "" This is the outline of the commercial for that play: ""	Brainstorming	11.2%	train	labeler	11,295
		Chat	8.4%	train	customer	1,430
		Rewrite	6.6%	valid	labeler	1,550
		Summarization	4.2%	valid	customer	103
		Classification	3.5%			
		Other	3.5%			
		Closed QA	2.6%			
		Extract	1.9%			

InstructGPT: Instruction Tuning + Human Feedback

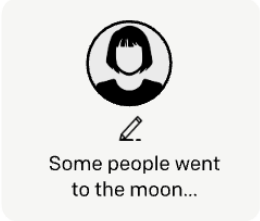
Step 1

Collect demonstration data, and train a supervised policy.

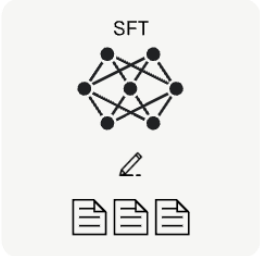
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



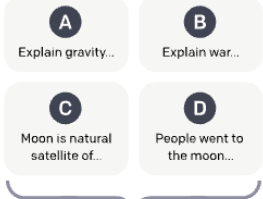
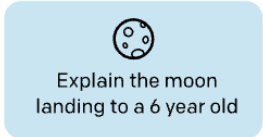
This data is used to fine-tune GPT-3 with supervised learning.



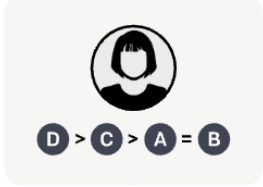
Step 2

Collect comparison data, and train a reward model.

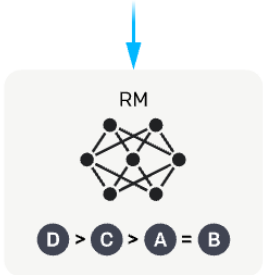
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



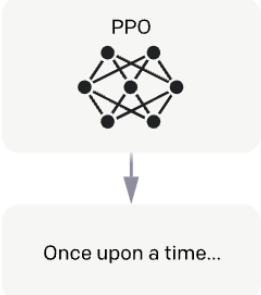
Step 3

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.

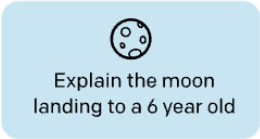


InstructGPT: Instruction Tuning + Human Feedback

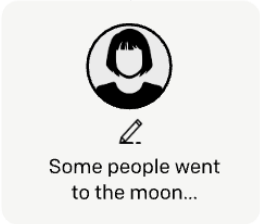
Step 1

Collect demonstration data, and train a supervised policy.

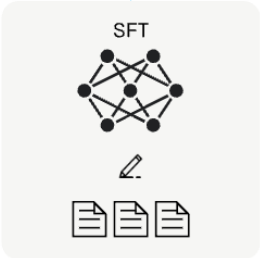
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



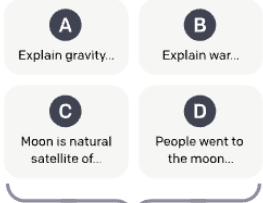
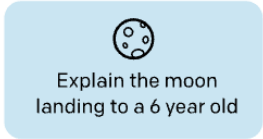
This data is used to fine-tune GPT-3 with supervised learning.



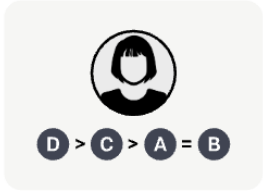
Step 2

Collect comparison data, and train a reward model.

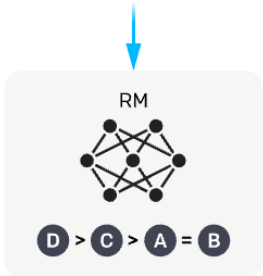
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



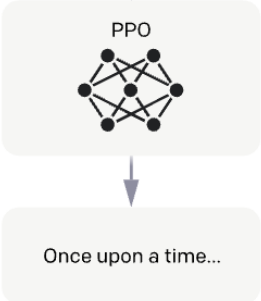
Step 3

Optimize a policy against the reward model using reinforcement learning.

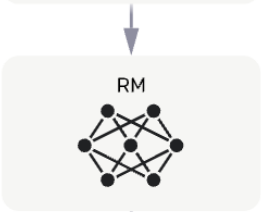
A new prompt is sampled from the dataset.



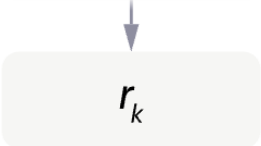
The policy generates an output.



The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.



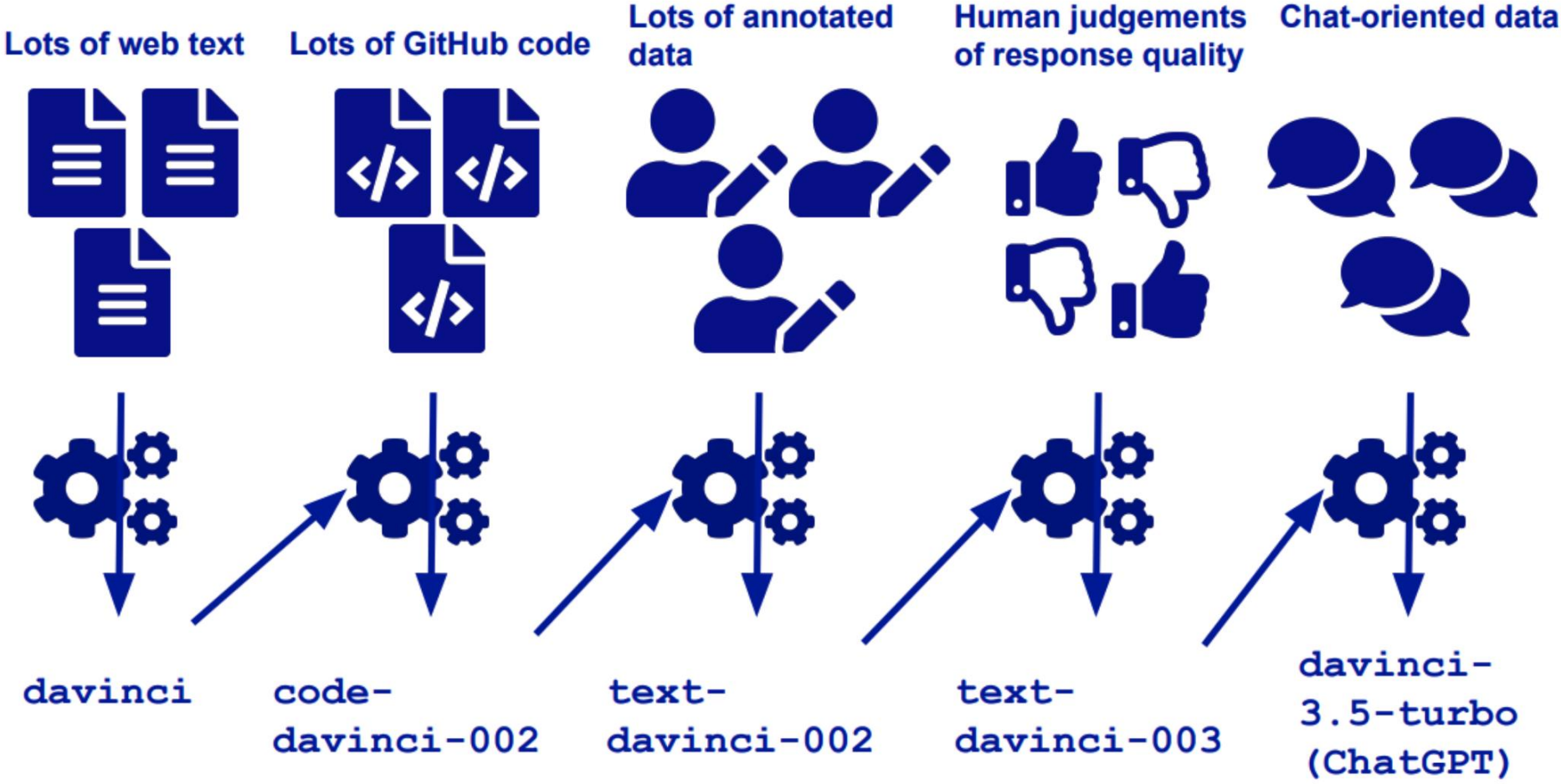
ChatGPT: InstructGPT + Dialogue Data

Introducing ChatGPT

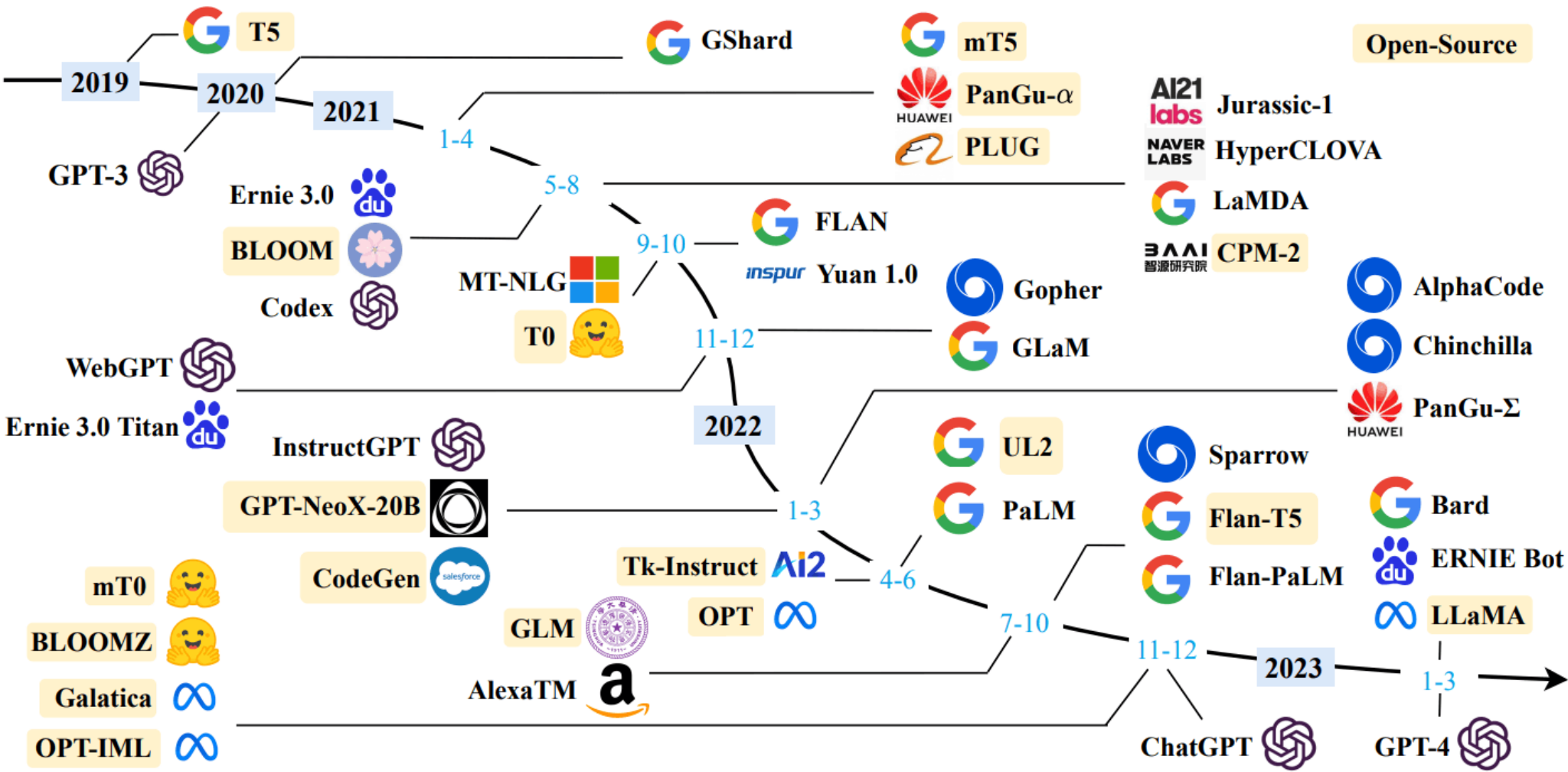
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.

“We trained this model using Reinforcement Learning from Human Feedback (RLHF), **using the same methods as InstructGPT**, but with slight differences in the data collection setup. We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant. We gave the trainers access to model-written suggestions to help them compose their responses. **We mixed this new dialogue dataset with the InstructGPT dataset**, which we transformed into a dialogue format.”

ChatGPT: InstructGPT + Dialogue Data

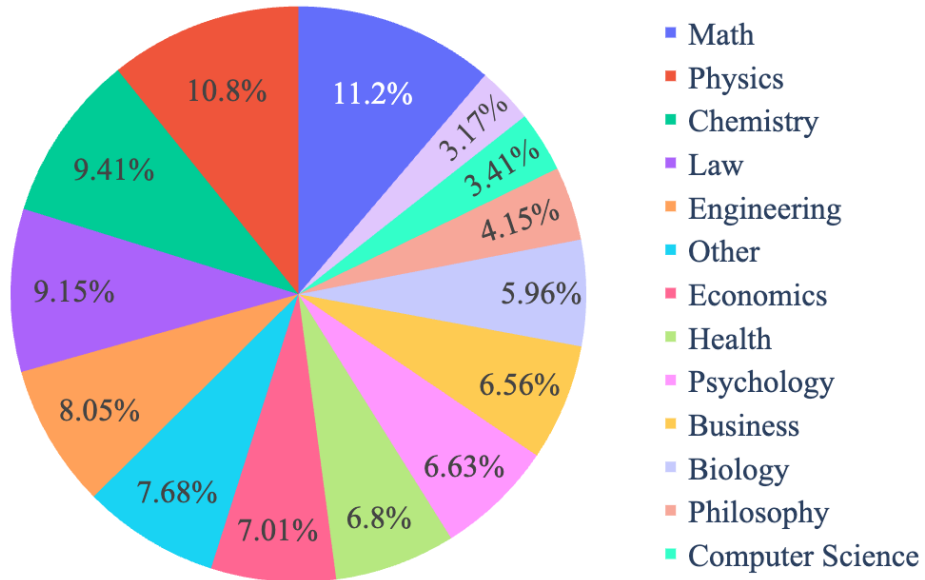


Large Language Models



Benchmark

- MMLU and MMLU-pro



Find all c in \mathbb{Z}_3 such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.
 (A) 0 (B) 1 (C) 2 (D) 3

Figure 14: An Abstract Algebra example.

What is the embryological origin of the hyoid bone?
 (A) The first pharyngeal arch
 (B) The first and second pharyngeal arches
 (C) The second pharyngeal arch
 (D) **The second and third pharyngeal arches**

Figure 15: An Anatomy example.

Why isn't there a planet where the asteroid belt is located?
 (A) A planet once formed here but it was broken apart by a catastrophic collision.
 (B) There was not enough material in this part of the solar nebula to form a planet.
 (C) There was too much rocky material to form a terrestrial planet but not enough gaseous material to form a jovian planet.
 (D) **Resonance with Jupiter prevented material from collecting together to form a planet.**

Figure 16: An Astronomy example.

Three contrasting tactics that CSO's can engage in to meet their aims are _____ which typically involves research and communication, _____, which may involve physically attacking a company's operations or _____, often involving some form of _____.
 (A) Non-violent direct action, Violent direct action, Indirect action, Boycott
 (B) Indirect action, Instrumental action, Non-violent direct action, Information campaign
 (C) **Indirect action, Violent direct action, Non-violent direct-action Boycott.**
 (D) Non-violent direct action, Instrumental action, Indirect action, Information campaign

Figure 17: A Business Ethics example.

Benchmark

- BIG-Bench Hard

Model Input (Navigate)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: If you follow these instructions, do you return to the starting point? Turn left. Turn right. Take 5 steps. Take 4 steps. Turn around. Take 9 steps.

Options:

- Yes
- No

A: Let's think step by step.

Model Output

We start at the origin (0, 0), facing the positive y-axis.
(1) Turn left: (0, 0), facing the negative x-axis.
(2) Turn right: (0, 0), facing the positive y-axis.
(3) Take 5 steps: (0, 5), facing the positive y-axis.
(4) Take 4 steps: (0, 9), facing the positive y-axis.
(5) Turn around: (0, 9), facing the negative y-axis.
(6) Take 9 steps: (0, 0), facing the negative y-axis.
Since (0, 0) is (0, 0), we are indeed where we started. So the answer is **Yes**. ✓

Model Input (Word Sorting)

<Few shot chain-of-thought exemplars omitted for brevity>

Q: Sort the following words alphabetically: List: burley bela arapaho bacteria bock

A: Let's think step by step.

Model Output

The first letter: "burley": "b" (2). "bela": "b" (2). "arapaho": "a" (1). "bacteria": "b" (2). "bock": "b" (2).
We now have: (1) "arapaho" < (2) ["burley" ? "bela" ? "bacteria" ? "bock"]. Now let's sort this subpart ["burley" ? "bela" ? "bacteria" ? "bock"] by looking at their second letters.
The second letter: "burley": "u" (21). "bela": "e" (5). "bacteria": "a" (1). "bock": "o" (15). We now have: (1) "bacteria" < (5) "bela" < (15) "bock" < (21) "burley". Hence, we have "arapaho" < ["bacteria" < "bela" < "bock" < "burley"]. So the answer is **arapaho bacteria bela bock burley**. ✓

Chatbot Arena

<https://huggingface.co/spaces/lmsys/chatbot-arena-leaderboard>

👉 Chat now!

🔍 Expand to see the descriptions of 73 models

Model A

What are LLMs? Give me a short answer.

LLMs (Large Language Models) are advanced AI systems trained on massive amounts of text data. They can understand and generate human-like text, enabling them to perform tasks like writing, translation, and answering questions.

Model B

What are LLMs? Give me a short answer.

LLMs are Large Language Models, which are advanced AI models trained on vast amounts of text data to understand, generate, and interact with human language.

👉 A is better 👉 B is better 🏆 Tie 👉 Both are bad

Chatbot Arena

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	ChatGPT-4o-latest (2024-08-08)	1316	+4/-3	31148	OpenAI	Proprietary	2023/10
2	Gemini-1.5-Pro-Exp-0827	1300	+4/-4	22844	Google	Proprietary	2023/11
2	Gemini-1.5-Pro-Exp-0801	1298	+4/-4	26110	Google	Proprietary	2023/11
2	Grok-2-08-13	1294	+4/-4	16215	xAI	Proprietary	2024/3
5	GPT-4o-2024-05-13	1285	+3/-2	86306	OpenAI	Proprietary	2023/10
6	GPT-4o-mini-2024-07-18	1274	+4/-4	26088	OpenAI	Proprietary	2023/10
6	Claude 3.5 Sonnet	1270	+3/-3	56674	Anthropic	Proprietary	2024/4
6	Gemini-1.5-Flash-Exp-0827	1268	+5/-4	16780	Google	Proprietary	2023/11
6	Grok-2-Mini-08-13	1267	+4/-4	16731	xAI	Proprietary	2024/3
6	Meta-Llama-3.1-405b-Instruct	1266	+4/-4	27397	Meta	Llama 3.1 Community	2023/12
7	Gemini Advanced App (2024-05-14)	1266	+3/-3	52236	Google	Proprietary	Online
7	GPT-4o-2024-08-06	1263	+4/-4	18093	OpenAI	Proprietary	2023/10
12	Gemini-1.5-Pro-001	1259	+2/-2	76952	Google	Proprietary	2023/11
12	Gemini-1.5-Pro-Preview-0409	1257	+4/-4	55597	Google	Proprietary	2023/11
12	GPT-4-Turbo-2024-04-09	1257	+2/-3	90149	OpenAI	Proprietary	2023/12

Common None-Open-Source LLMs

- Not public but strong performance
 - ChatGPT-4o (OpenAI)
 - Gemini (Google DeepMind)
 - Grok (xAI)
 - Claude (Anthropic)
- Most of them have APIs

Common Open-Source LLMs

- LLaMA 3 (Meta)
 - Meta-Llama-3-8B
 - Meta-Llama-3.1-8B-Instruct
 - Meta-Llama-3-70B
 - Meta-Llama-3.1-70B-Instruct

Hugging Face Search models, datasets, users...

meta-llama / **Meta-Llama-3-8B** like 5.59k

Text Generation Transformers Safetensors PyTorch English llama facebook

meta llama-3 text-generation-inference Inference Endpoints License: llama3

Train Deploy Use this model

Model card Files Community 200

Edit model card

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...

Log in or Sign Up to review the conditions and access this model content.

Downloads last month **2,138,973**

Safetensors

Model size 8.03B params

Tensor type BF16

Inference API Cold

Text Generation Examples

My name is Clara and I am

Open-Source LLMs

- Mistral (Mistral AI)
 - Mistral-Large-Instruct-2407

The screenshot shows the Hugging Face interface for the model `mistralai/Mistral-Large-Instruct-2407`. At the top, there is a search bar and the Hugging Face logo. The model name is displayed with a 'like' button showing 713 likes. Below the name are several tags: Text Generation, Transformers, Safetensors, 10 languages, mistral, and conversational. There are also tags for text-generation-inference, Inference Endpoints, and License: mrl (other). Action buttons for Train, Deploy, and Use this model are visible. Below these are links for Model card, Files, and Community (24). A prominent message box states: "You need to agree to share your contact information to access this model". It includes a link to the Privacy Policy and buttons for Log in and Sign Up. To the right, there is a graph showing "Downloads last month" with a value of 44,083. Below the graph, there is a section for Safetensors with details: Model size 123B params and Tensor type BF16. At the bottom right, there is a section for Inference API with a link for Text Generation.

Next Lecture

- Adversarial Attacks and Defenses