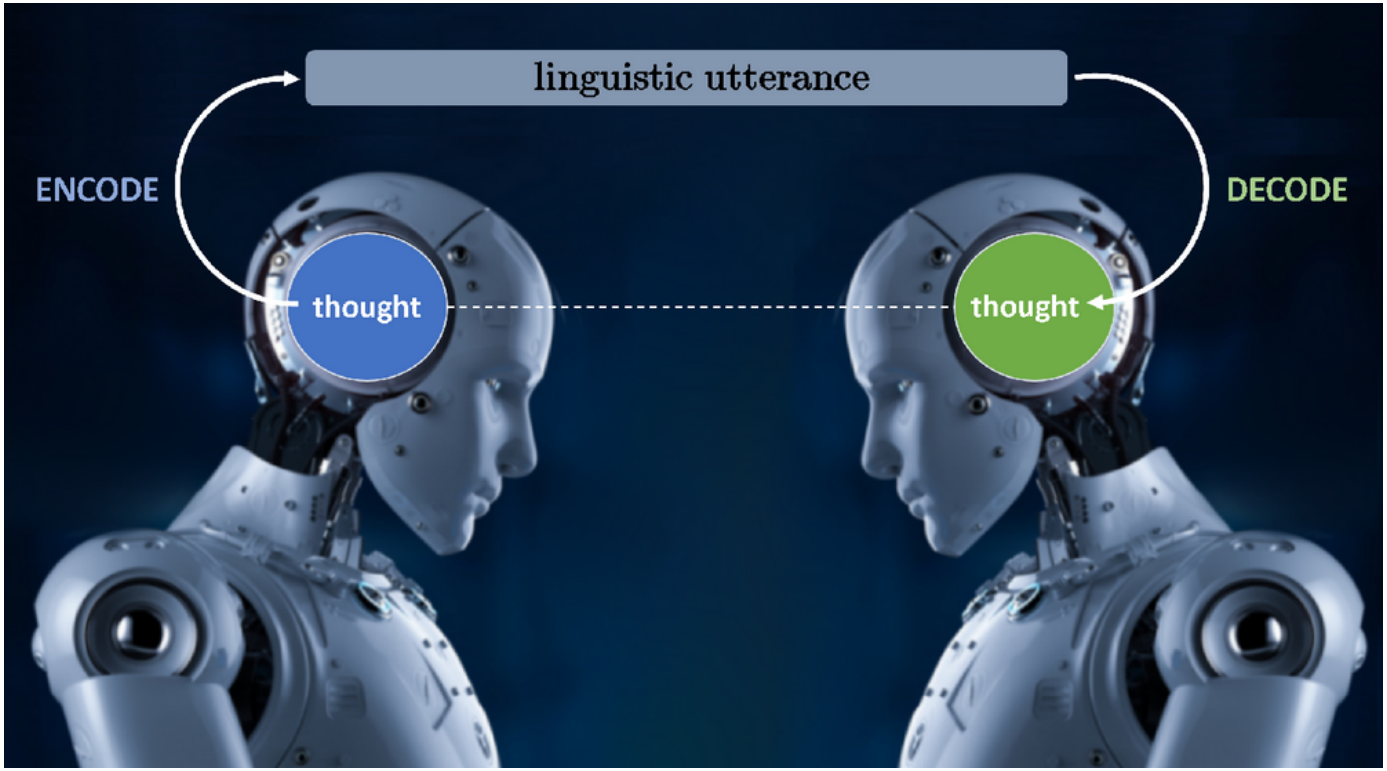
Natural Language Processing

Lecture : ChatBot

Master Degree in Computer Engineering
University of Padua
Lecturer : Giorgio Satta

ChatBots



The gradient, Walid S. Saba

ChatBots

ChatBots are capable of maintaining a conversation with a user in a natural way.

ChatBots can also be used to

- generate creative content
- enhance work productivity
- analyze and extract information from texts

Modern chatBots such as ChatGPT (OpenAI), Gemini (Google), DeepSeek-V3 (DeepSeek), Copilot (Microsoft), Llama (Meta), Claude (Anthropic), etc. are all based on LLM.

ChatBots

LLMs have not been instructed to answer user's questions.

We can turn a pre-trained LLM into a chatBot using a combination of

- instruction tuning
- domain adaptation
- alignment

ChatBots

We have already introduced instruction tuning and alignment in previous lecture.

Domain adaptation: in case we want our chatBot to be specialized for a specific domain of interest, we need to inject new knowledge into the model.

Domain adaptation can be done at the same time as instruction tuning: construct a dataset of question/answers on the domain of interest.

ChatBots

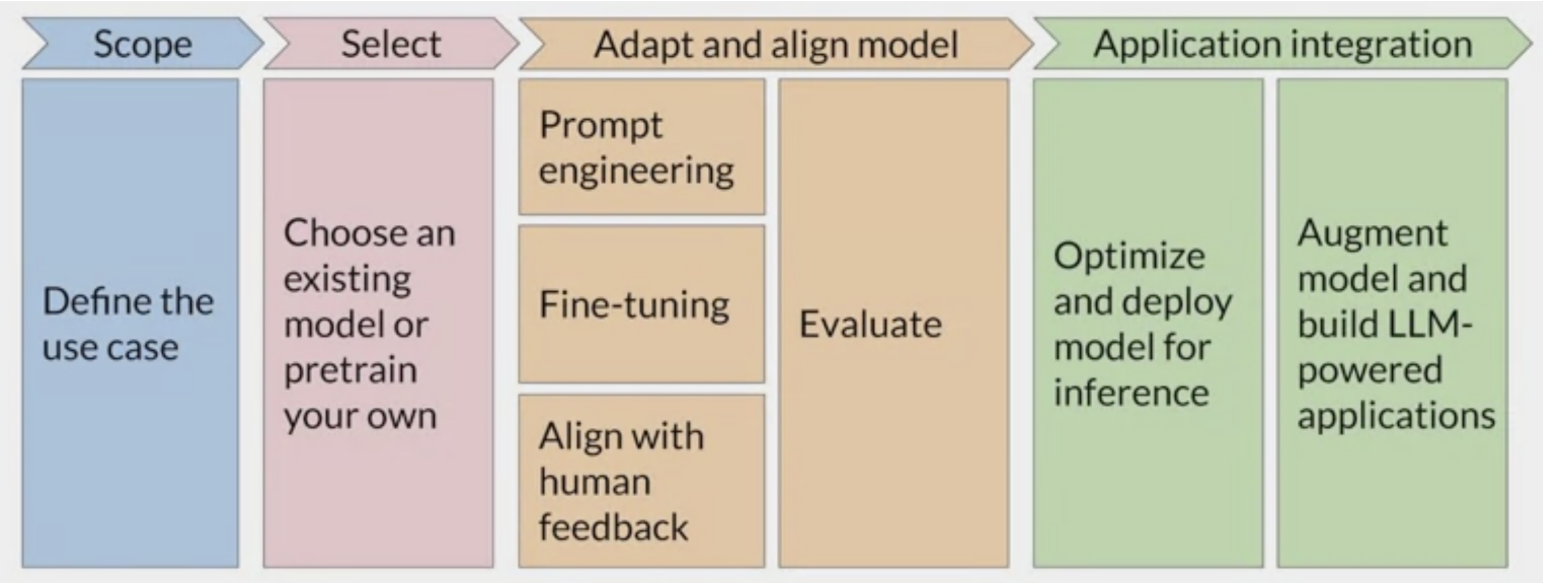
Alternatively to supervised fine tuning and alignment, we can use special, detailed prompts to steer the LLM to simulate a virtual assistant / chatBot.

Prompting does not require any change in the model parameters, but has some additional computational cost at inference time.

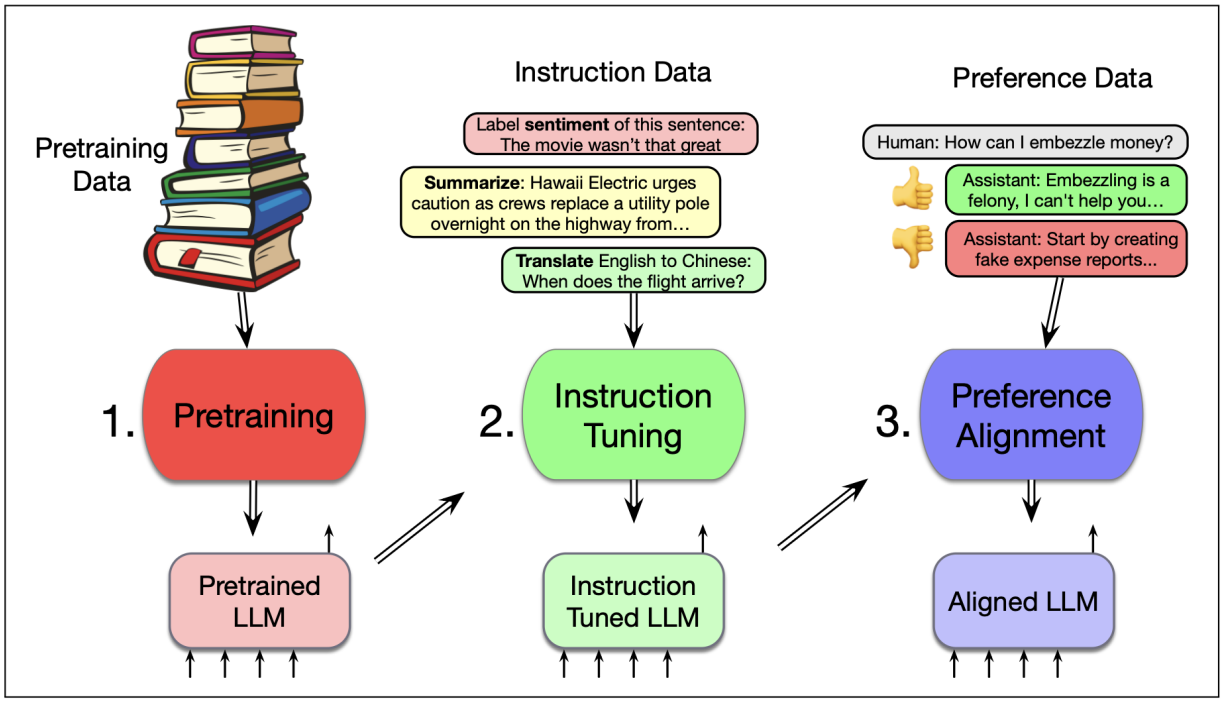
We introduce **prompt engineering** later in this lecture.

ChatBots

The general schema for the lifecycle of a chatBot.



Three stages of training large language models: pretraining, instruction tuning, and preference alignment.



ChatBots

In chatBot lifecycle, 99% of training work is in pretraining phase.

Prompting approach for virtual assistant task is not super reliable / robust.

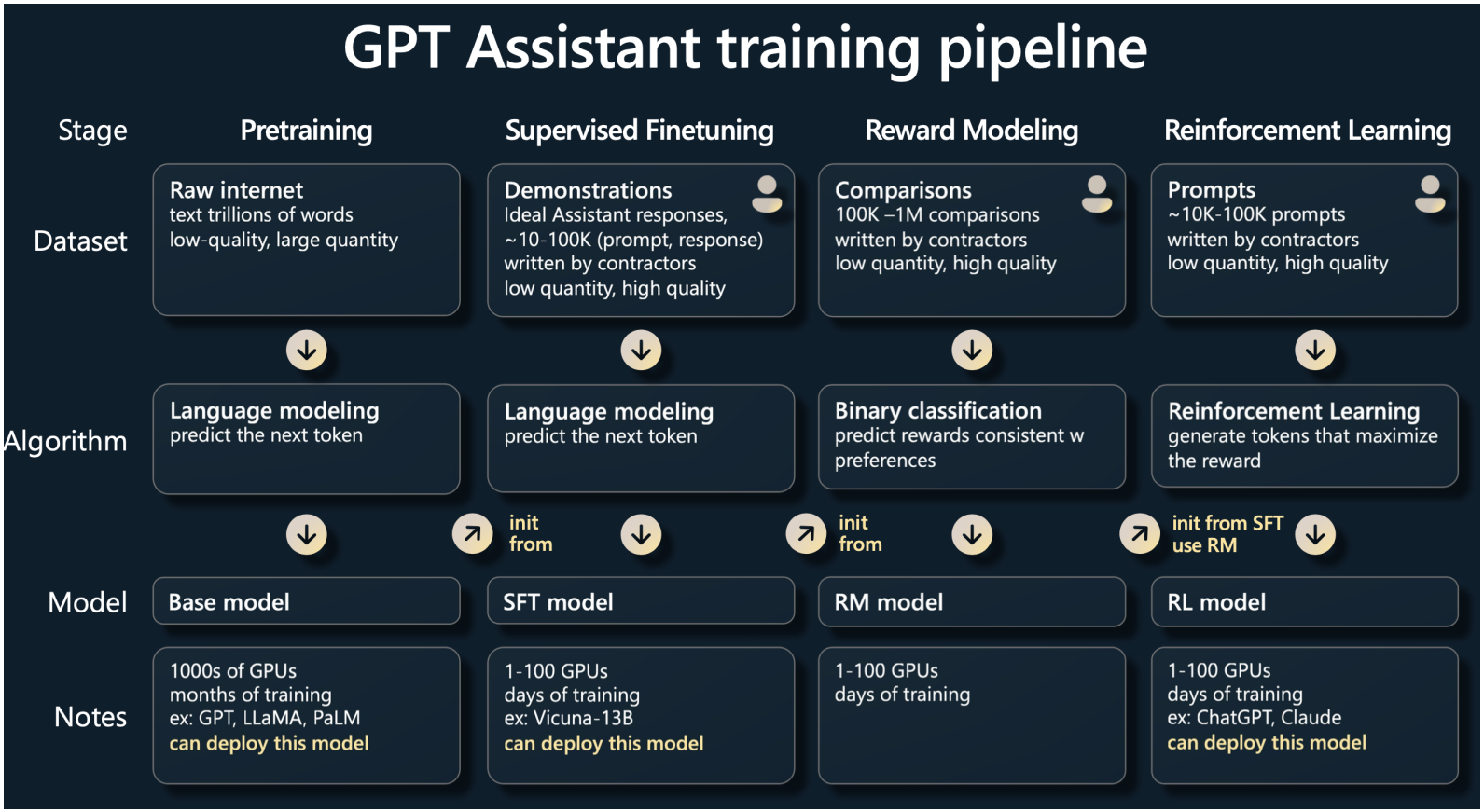
Supervised fine tuning requires high-quality data set.

Video tutorial

Title: State of GPT
Author: Andrej Karpathy
Source: May 23, 2023
Content: This video introduces the basic technologies underlying the development of chat-GPT.

`https://www.youtube.com/watch?v=bZQun8Y4L2A`

Detailed view of the lifecycle of a chat-GPT.



NLU Datasets

General Language Understanding Evaluation (GLUE) benchmark is a collection of 9 datasets for evaluating natural language understanding (NLU) systems:

- Corpus of Linguistic Acceptability (CoLA)
- Stanford Sentiment Treebank (SST)
- Microsoft Research Paragraph Corpus (MRPC)
- Quora Question Pairs (QQP)
- Multi-Genre NLI (MNLI)
- Question NLI (QNLI)
- Recognizing Textual Entailment (RTE)
- Winograd NLI
- Diagnostics Main

NLU Datasets

SuperGLUE has been launched as a harder version of GLUE, after AI models surpassed non-expert human performance on the original set:

- COPA (causal reasoning) and RTE (textual entailment)
- multiRC (multi-sentence QA) and ReCoRD (commonsense reasoning)
- BoolQ (yes/no questions), WiC (word sense disambiguation), and WSC (coreference resolution)
- broad-coverage diagnostic

NLU Datasets

Massive Multitask Language Understanding (MMLU) is a test set to measure a model multitask accuracy.

The test covers 57 tasks, including among others

- science, technology, engineering and mathematics (STEM)
- social science and humanities
- finance, accounting, and marketing
- professional medicine

To attain high accuracy on this test, models must possess extensive world knowledge and problem solving ability.

`https://paperswithcode.com/dataset/mmlu`.

NLU Datasets

Holistic Evaluation of Language Models (HELM) aims to improve the transparency of models, and to offer guidance on which models perform well for specific tasks.

HELM takes a multimetric approach, measuring seven metrics: accuracy, calibration, robustness, fairness, bias, toxicity, efficiency.

ChatBot Arena

Chatbot Arena Leaderboard is a novel platform that leverages crowdsourced human evaluation to rank LLMs

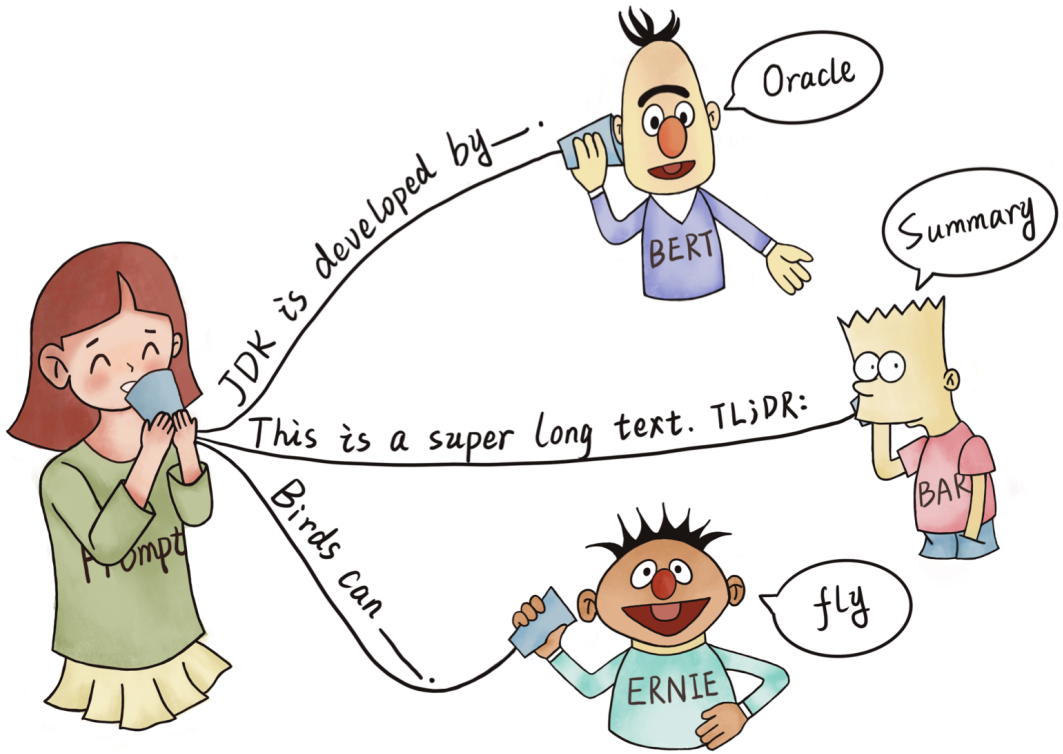
- LLMs take on the role of "players" in head-to-head comparisons
- users are invited to vote on which LLM they find more engaging, informative, or helpful

Ranking based on user votes provided in system comparison.

ChatBot Arena

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Exp-03-25	1443	+8/-13	2540	Google	Proprietary
2	4	Grok-3-Preview-02-24	1404	+5/-6	10398	xAI	Proprietary
2	2	GPT-4.5-Preview	1398	+7/-6	10615	OpenAI	Proprietary
4	7	Gemini-2.0-Flash-Thinking-Exp-01-21	1381	+4/-3	22659	Google	Proprietary
4	4	Gemini-2.0-Pro-Exp-02-05	1380	+5/-5	20293	Google	Proprietary
4	3	ChatGPT-4o-latest (2025-01-29)	1374	+5/-4	22517	OpenAI	Proprietary
7	5	DeepSeek-R1	1360	+5/-5	12772	DeepSeek	MIT
7	12	Gemini-2.0-Flash-001	1355	+4/-4	18327	Google	Proprietary
7	4	o1-2024-12-17	1351	+4/-4	25044	OpenAI	Proprietary
10	12	Qwen2.5-Max	1340	+5/-3	17124	Alibaba	Proprietary
10	12	Gemma-3-27B-it	1340	+7/-6	6974	Google	Gemma

Prompt



Liu *et al.*, 2021

Prompt

We have already introduced the idea of reducing NLP tasks to text instances of the text continuation problem.

A **prompt** is a text that a user issues to a LLM to create a context that guides the generation of useful output.

Use of prompt is very important when interfacing with chatBots.

Prompt

Example :

Sample Hotel Review

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax.

A prompt consisting of a review plus an incomplete statement

Did not like the service that I was provided, when I entered the hotel. I also did not like the area, in which the hotel was located. Too much noise and events going on for me to feel relax. In short, our stay was

Prompt

Use of prompting **templates**: task-specific prompting text along with slots for the input.

Basic Prompt Templates

Summarization	{input} ; `tldr;`
Translation	{input} ; `translate to French:`
Sentiment	{input}; `Overall, it was`
Fine-Grained-Sentiment	{input}; `What aspects were important in this review?`

Prompt

The already mentioned zero-shot/few-shot learning can also be viewed as special cases of prompt learning

Example :

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____

Positive

Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

Apple ... development of in-house chips. // Tech

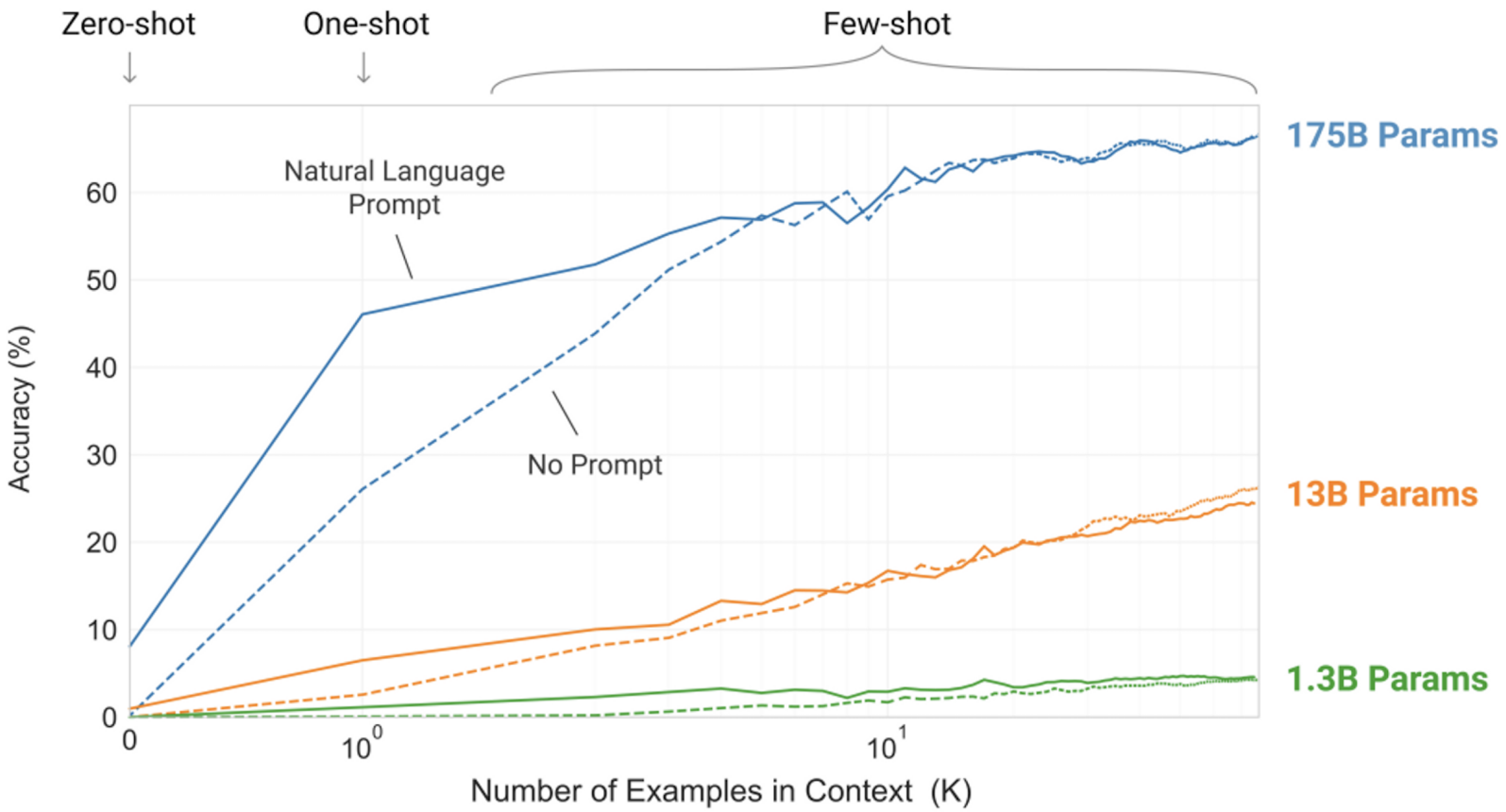
The company anticipated its operating profit to improve. // _____

Finance

https://ai.stanford.edu/blog/understanding-incontext/#f1

Prompt

Accuracy with few shot learning for several GPT-3 models:



Prompt

We can view a prompt as some kind of **learning signal**, helping LLM to perform novel tasks.

Prompt learning is also referred to as **in-context learning**.

The term 'in-context learning' was introduced in the original GPT-3 paper (Brown *et al.*, 2020).

Prompt

Why does prompt learning work?

- prompt examples do not teach a new task; instead, they help locating a task already learned during pre-training
- prompt examples induce a latent state/concept so that LM generates coherent next tokens
- prompt learning performance is highly correlated with term frequencies during pre-training

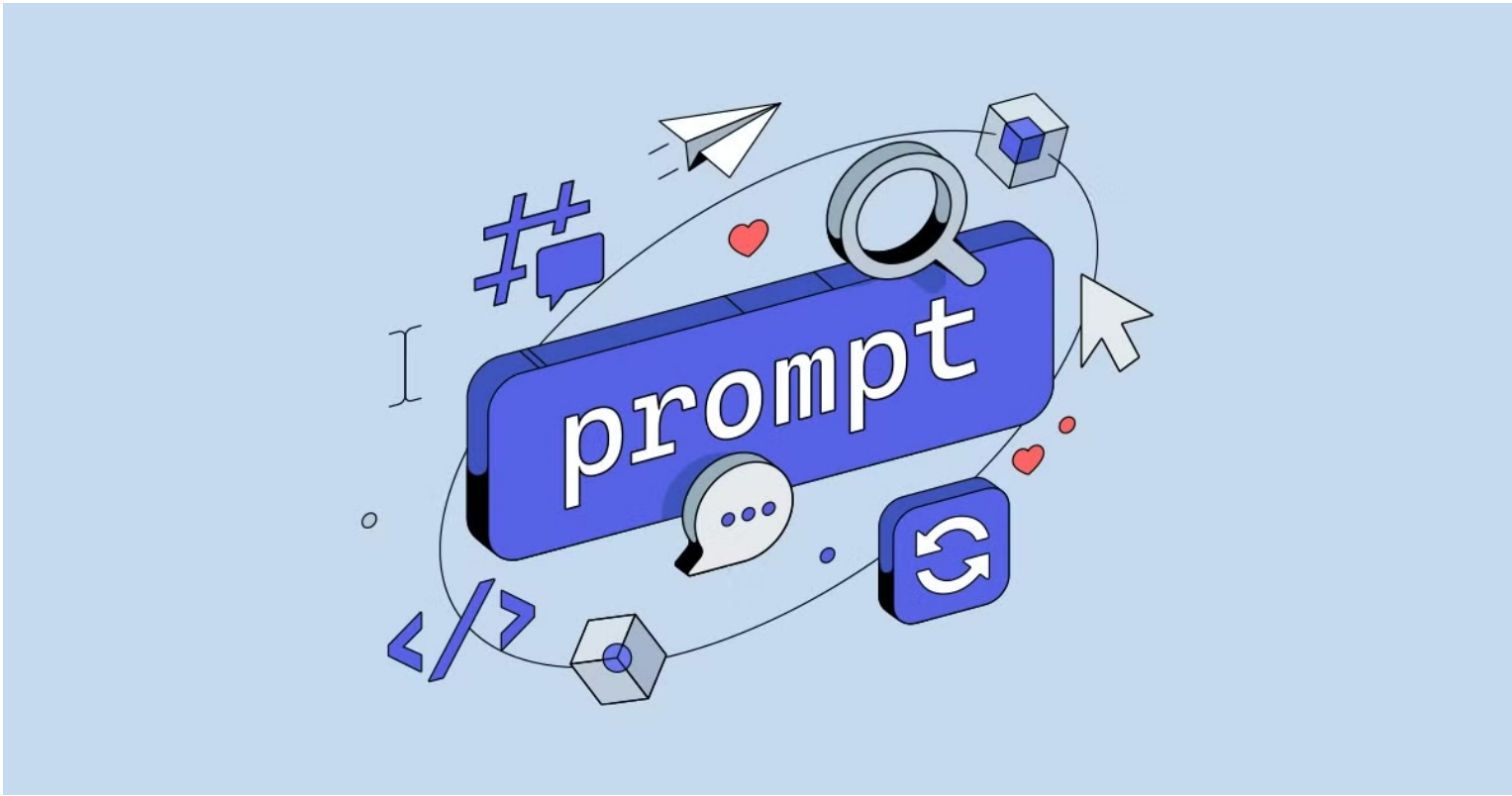
Prompt

Advantages of prompt learning wrt fine-tuning

- the gap between the pre-training stage and the downstream task can be significant: the objectives are different
- for the downstream tasks, we need to introduce new parameters
- when you only have a dozen of training examples for a new downstream task, it is hard to fine-tune

Prompting can also be used to avoid hallucinations, providing enough details and constraints in the prompt to the model.

Prompt engineering



©artea.com

Prompt engineering

The process of finding effective prompts for a task is known as **prompt engineering**.

We explore below some **good practices** in prompt engineering.

Prompt engineering

Some rules for building up your prompt

- **task**: specify the problem precisely, avoiding ambiguity
- **context**: place the problem in the proper frame
- **motivation**: why do you need the problem to be solved?
- **format**: specify the format for the answer

Prompt engineering

Some rules for building up your prompt

- **persona**: indicate your role and suggest a role for the chatBot
- **instructions**: provide a list of numbered instructions for performing the requested task
- **examples**: a few simple examples can be helpful in many cases.

Prompt engineering

Some rules for building up your prompt

- **style**: indicate the style of the response, informal, catchy, etc.
- **terminology**: say which type of terminology should be used in the response
- **length**: indicate the approximate length for the desired document

Prompt engineering

Techniques for building up your prompt

- **chain of thoughts**: ask the chatBot to reason step-by-step, reporting answers for each intermediate step
- **tree of thoughts**: expand your prompt along a tree structure; use for complex problems
- **iterative prompting**: start with an initial prompt and add details or further requests later

Prompt engineering

Techniques for building up your prompt

- **jailbreaking prompt**: accompany prompt with a compelling or moving story
- **meta-prompting**: provide a draft of your prompt, and explicitly ask the chatBot to improve it
- **prompt library**: save the most successful prompts for reuse in the future

Prompt engineering

Prompt **optimization methods** search for prompts with improved performance on the basis of the following three components

- start state: an initial human or machine generated prompt or prompts suitable for some task
- scoring metrics: a method for assessing how well a given prompt performs on the task
- expansion method: method for generating variations of a prompt

Evaluation

We can evaluate accuracy in a prompting setup using multiple answer questions from language understanding datasets.

Example : MMLU dataset (see previous slides)

MMLU microeconomics example

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained.
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
(C) monopoly firms do not engage in significant research and development.
(D) consumer surplus is lost with higher prices and lower levels of output.

Turns question into prompted test and ask to select correct answer.

Ethics

Feng Yu, Stock.Adobe.com

Ethics

Contextual language models can generate **toxic language**, misinformation, radicalization, and other socially harmful activities.

Contextual language models can **leak information** about their training data. It is possible for an adversary to extract individual data from a language model (phishing).

Mitigating all these harms is an important but **unsolved** research problem in NLP.