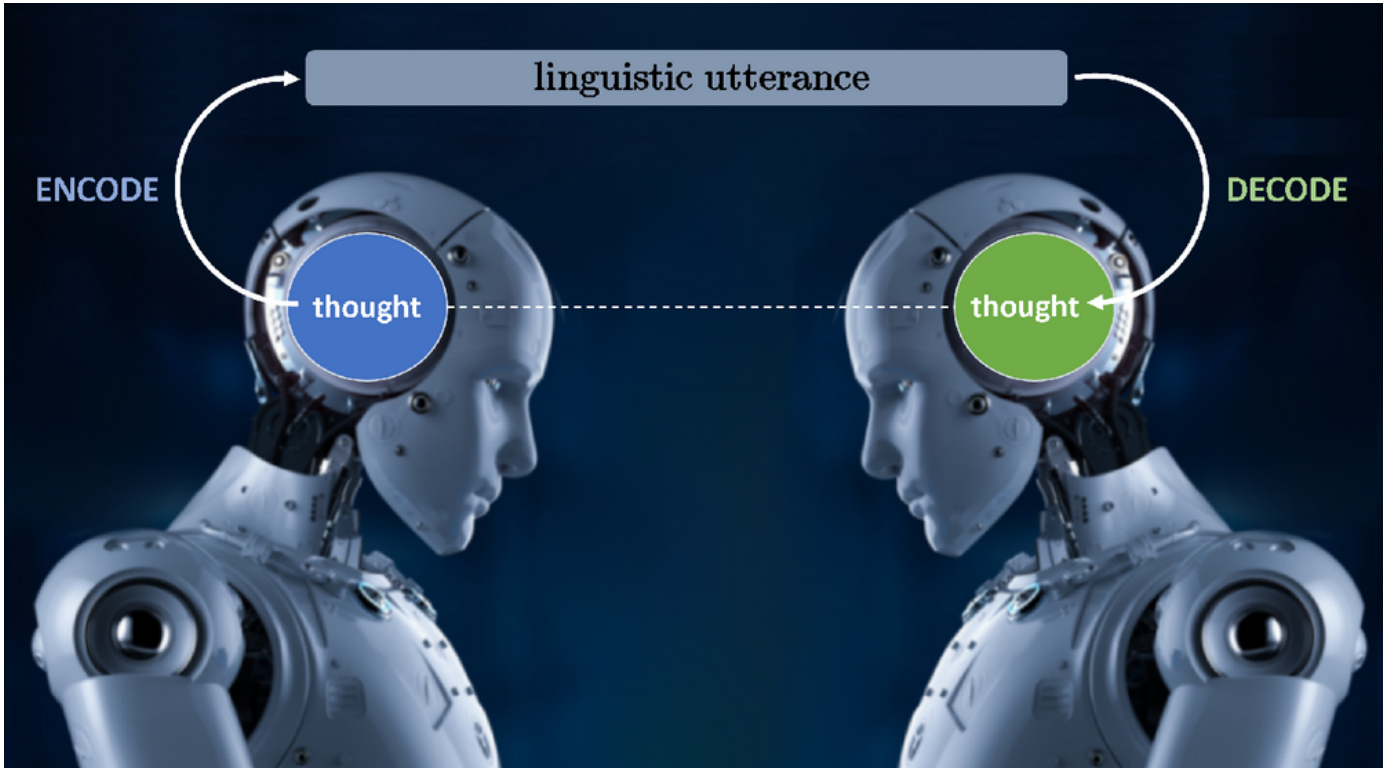
Natural Language Processing

Lecture 6 : ChatBot

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

Lecture partially based on material originally developed by :
Maarten Grootendorst: A Visual Guide to Reasoning LLMs

ChatBot



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

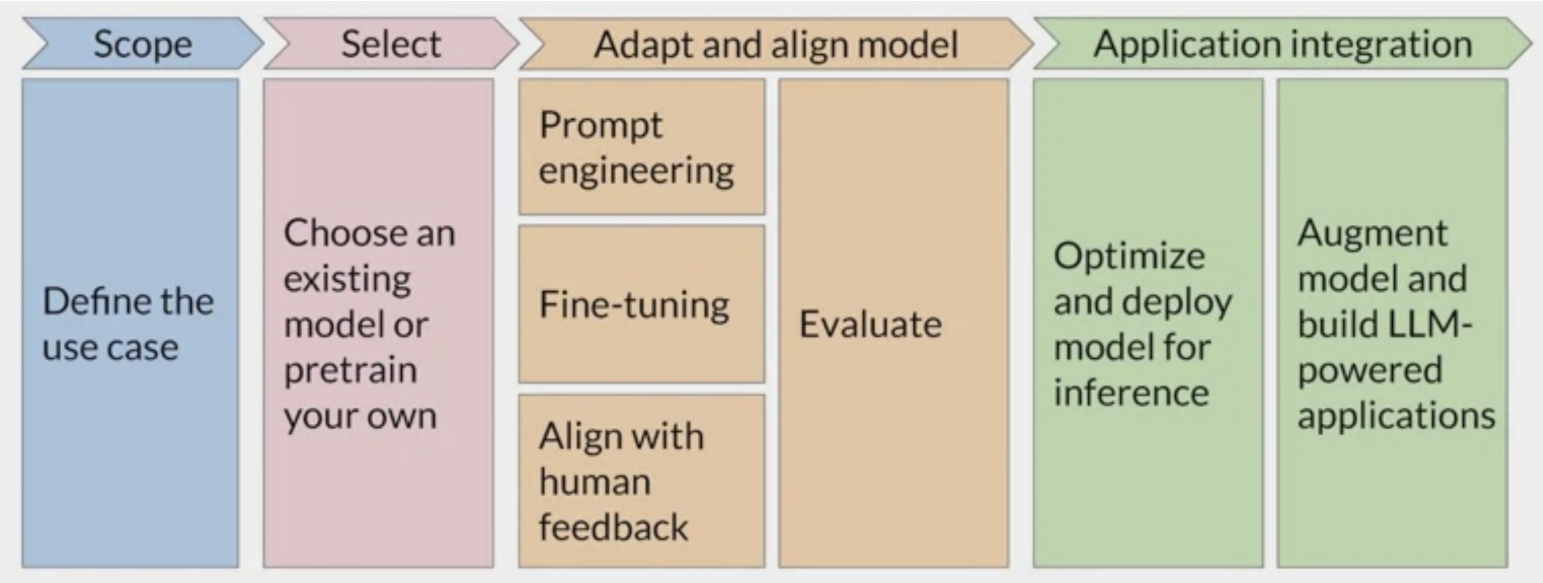
Alternatively to supervised fine tuning and alignment, we can use special, detailed prompts to steer the LLM to simulate a virtual assistant / chatBot.

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.



ChatBot

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

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

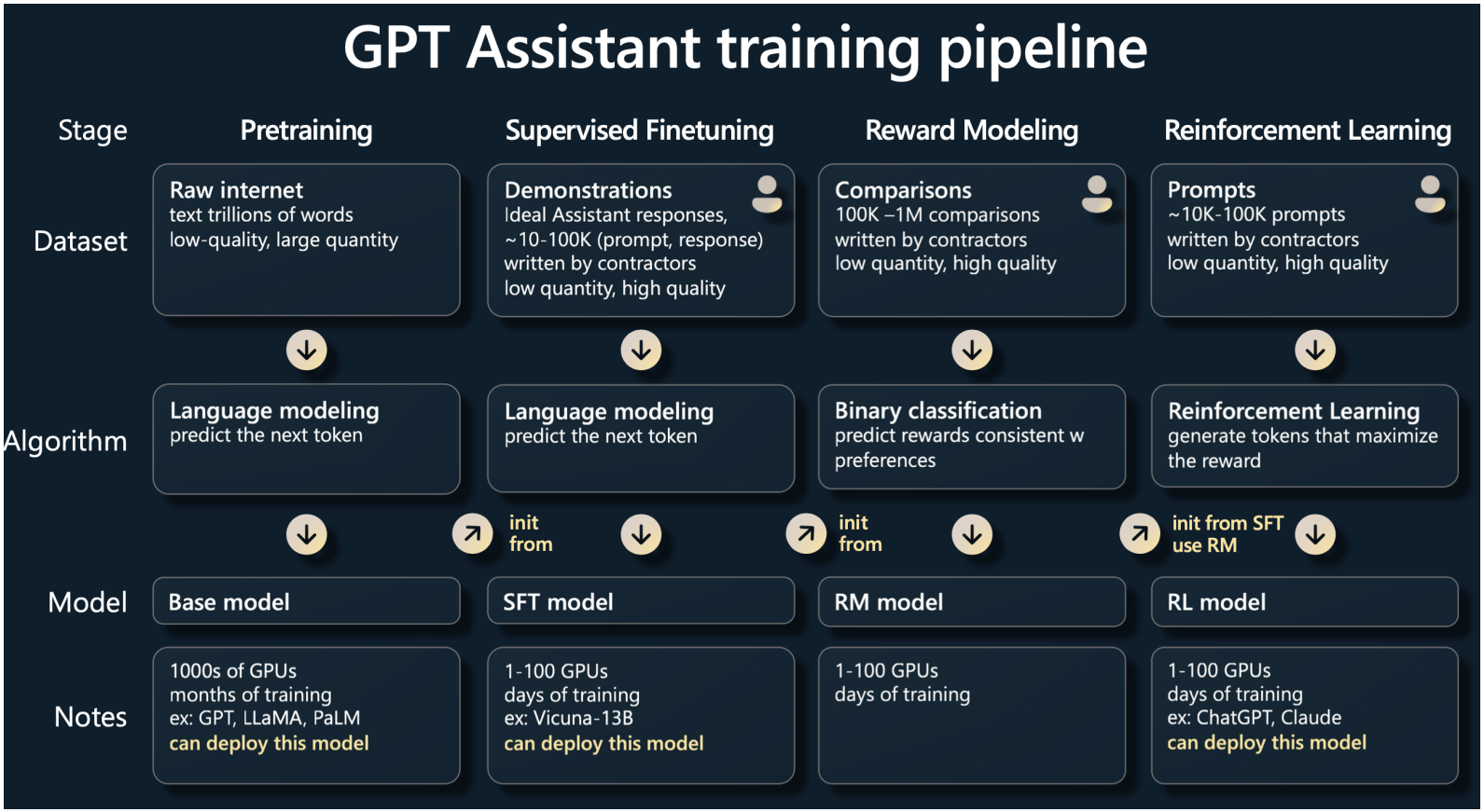
Supervised fine tuning requires high-quality data set.

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

`https://gluebenchmark.com` — superseeded by SuperGLUE

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

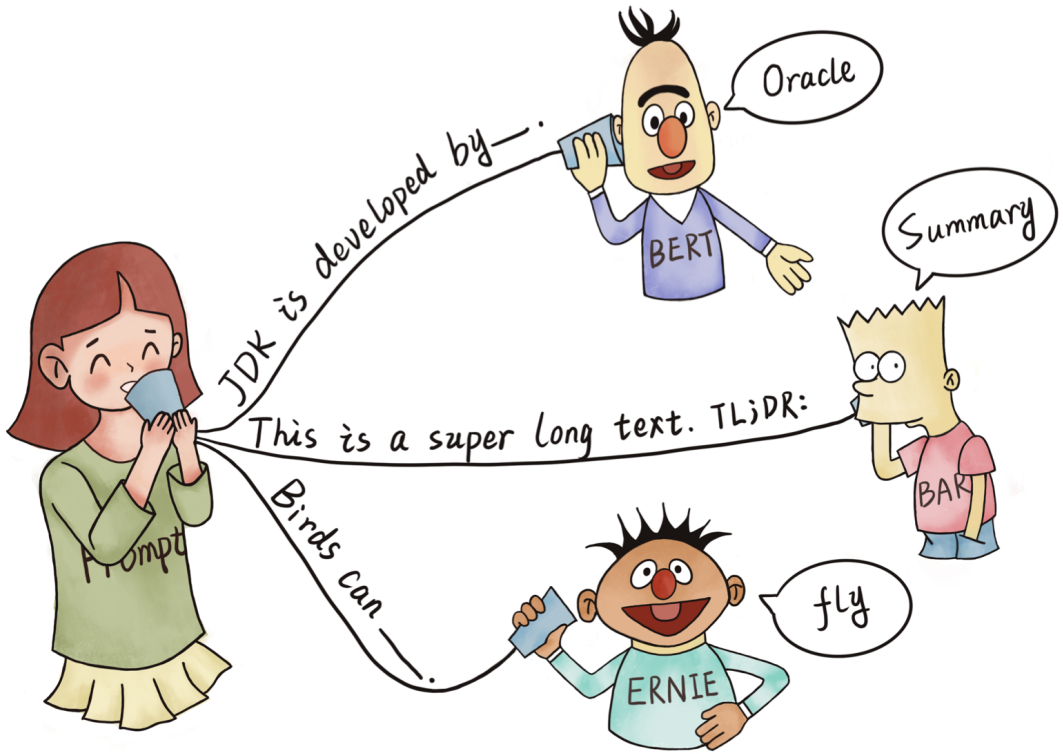
- LLMs take on the role of "players" in head-to-head comparisons
- users are invited to vote on which LLM they find more engaging, informative, or helpful

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

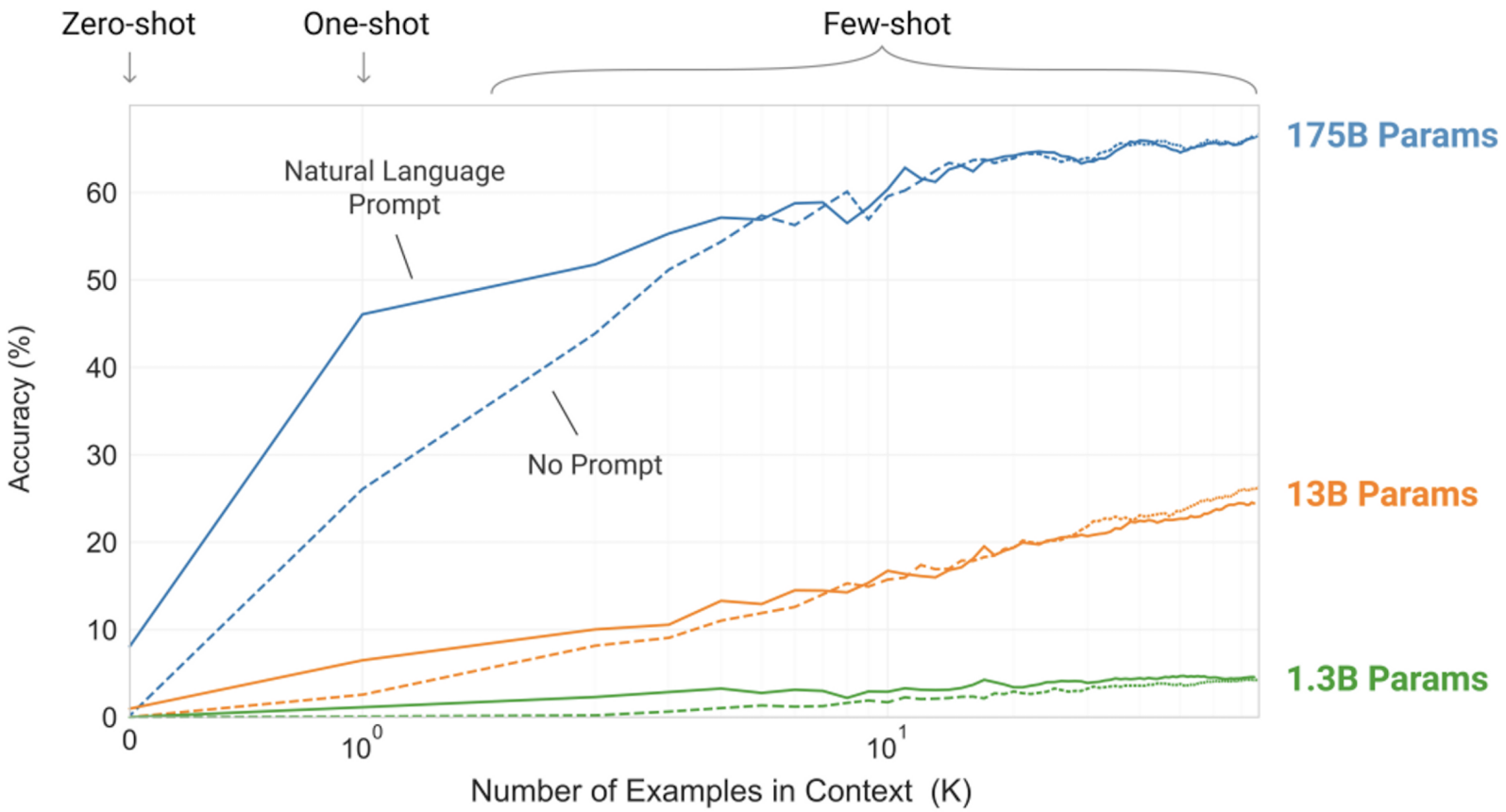
The company anticipated its operating profit to improve. // _____

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

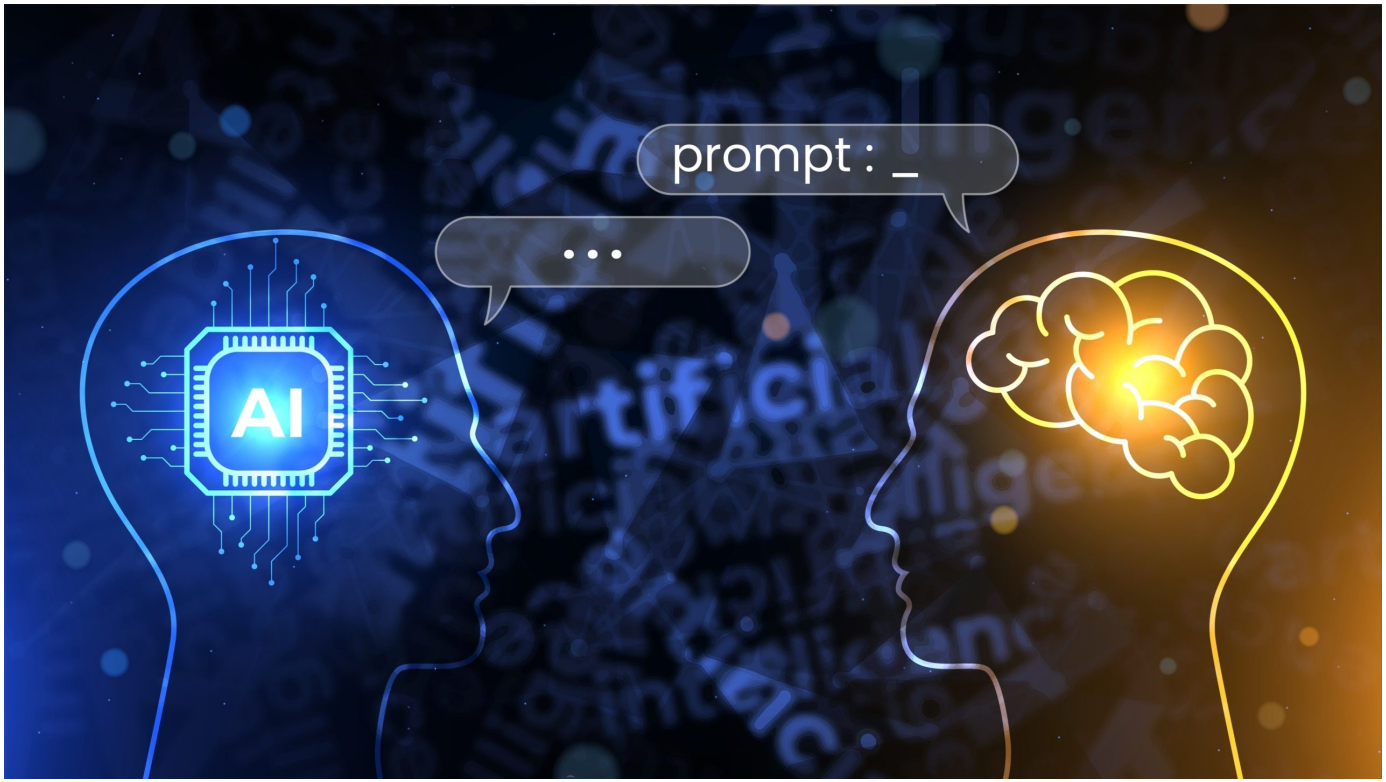
Prompt

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
- **examples**: a few simple examples can be helpful in many cases
- **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
- **jailbreaking prompt**: accompany prompt with a compelling or moving story

Prompt engineering

Techniques for building up your prompt

- **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.

Retrieval-augmented generation

©TheBlue.ai

Retrieval-augmented generation

LLMs have an enormous amount of knowledge encoded in their parameters. However, LLMs

- may lead to hallucination
- may not be up-to-date with their knowledge
- do not provide textual evidence to support their answer
- are unable to answer questions from proprietary data

Retrieval-augmented generation

Retrieval-augmented generation (RAG) is a two-step process combining external knowledge search with prompting.

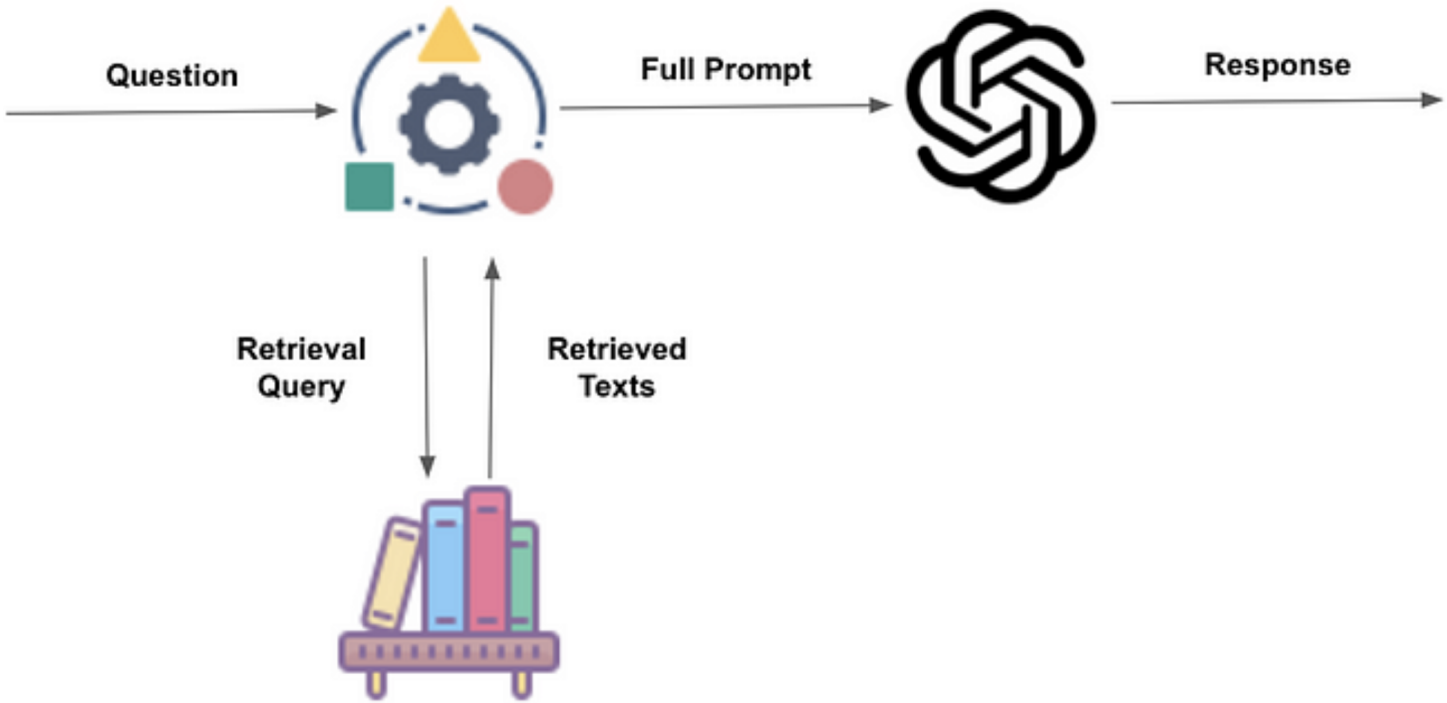
Retrieval: given a query, some information retrieval neural model fetches relevant documents/passages.

Based on document and query embedding, and a learnable similarity measure.

Generation: fetched documents are wrapped into a prompt, along with query, and passed to the LLM to generate relevant response.

Retrieval-augmented generation

RAG: retrieval step and generation step



Retrieval-augmented generation

Example :

```
retrieved passage 1
...
retrieved passage n
Based on these texts, answer the following question:
Q: Who wrote the book "The Origin of Species"?
A:
```

Retrieval-augmented generation

More formally, we **reduce** the Q/A problem to the problem of computing the following probability.

Assume a query q and let $R(q)$ be the set of retrieved passages based on q. Then (symbol ';' denotes string concatenation)

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid R(q) \ ; \ \texttt{prompt} \ ; \ \texttt{[Q:]} \ ; \ q \ ; \ \texttt{[A:]} \ ; \ x_{<i})$$

Retrieval-augmented generation

Important advantages of RAG approach

- external knowledge source injects into LLM recent/specific information that wasn't available during pre-training.
- external knowledge source can be updated, no need to retrain the LLM.
- RAG works well also with small-size, more manageable LLMs.
- allows fact checking and reduces hallucinations.

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.

Large reasoning models

©Placement Preparation

Maarten Grootendorst, A Visual Guide to Reasoning LLMs

https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

Large reasoning models

OpenAI's 5 steps towards artificial general intelligence (AGI)

- conversational AI
- reasoning AI
- autonomous agent AI
- innovating AI
- organizational AI

Large reasoning models

Large reasoning models (LRM) are LLMs empowered with reasoning capabilities.

Examples as for 2025: OpenAI o-3, DeepSeek R1, Google Gemini 2.0 Flash Thinking.

LRMs break down a problem into smaller steps (often called reasoning steps or thought processes) before answering a given question.

LRMs learn to mimic human reasoning through

- long chain-of-thought (CoT)
- reinforcement learning

Large reasoning models

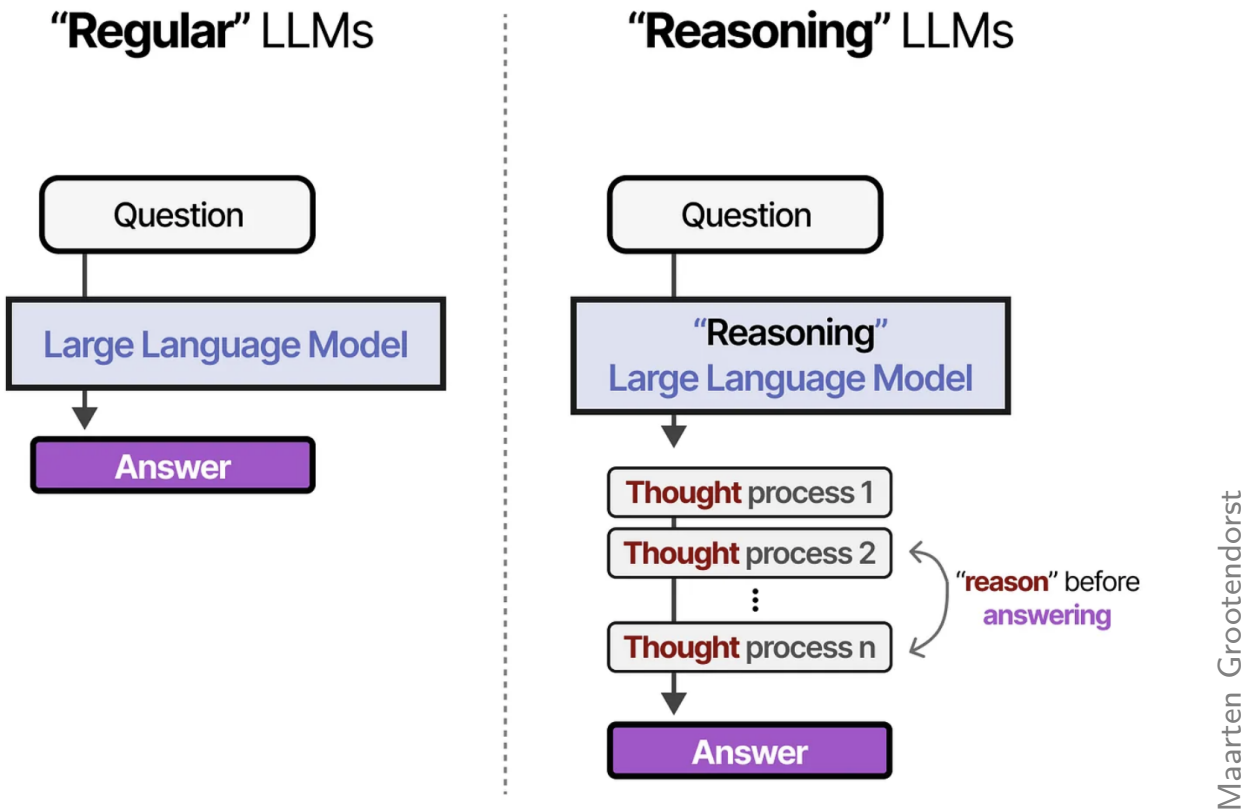
Example :
Julia has two sisters and one brother. How many sisters does her brother Martin have?

- Julia has two sisters; that means there are three girls in total (Julia + two more)
- Julia also has one brother, named Martin
- altogether, there are four siblings: three girls and one boy (Martin)
- from Martin's perspective, his sisters are all three of the girls (Julia and her two sisters)
- therefore Martin has three sisters.

M. Mitchell, `https://www.science.org/doi/10.1126/science.adw5211`

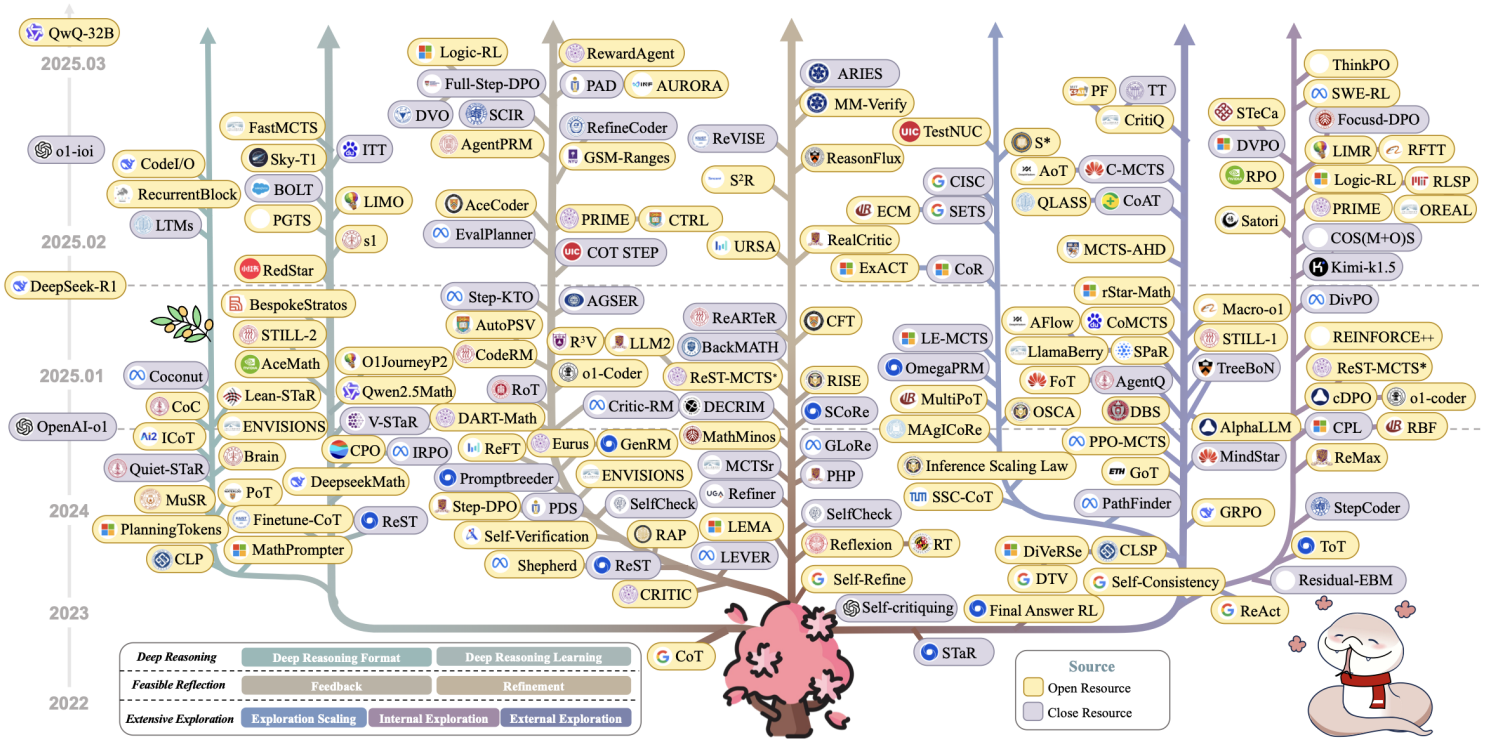
Large reasoning models

LRMs generates entire CoT, which can be really long, and most of which are not revealed to the user.



Modifying proposal distribution

Evolution of reasoning systems based on long CoT.



Chen et al., Towards Reasoning Era

Test-time compute

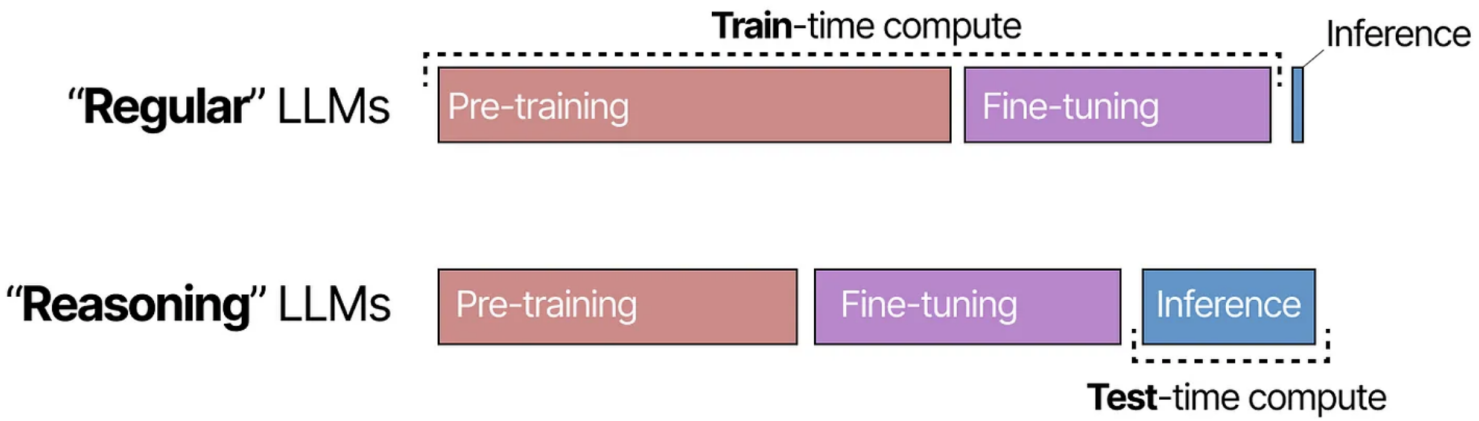
LRMs mark the paradigm shift from scaling **train-time compute** to scaling **test-time compute**.

Old paradigm: the larger your pretraining budget (parameters, tokens, FLOPs) the better the resulting model will be.

New paradigm: allow model to use more time resources during inference and explore large search space, i.e. think longer.

Test-time compute

LLM output the answer and skip any reasoning step. LRM use more tokens to derive their answer, through a systematic 'thinking process'

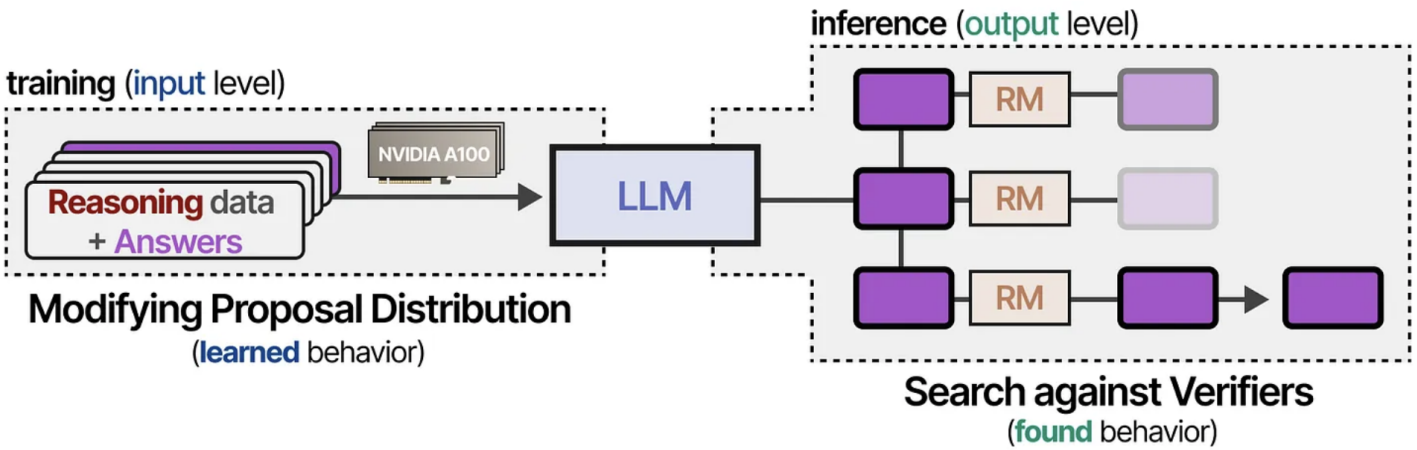


Maarten Grootendorst

Test-time compute

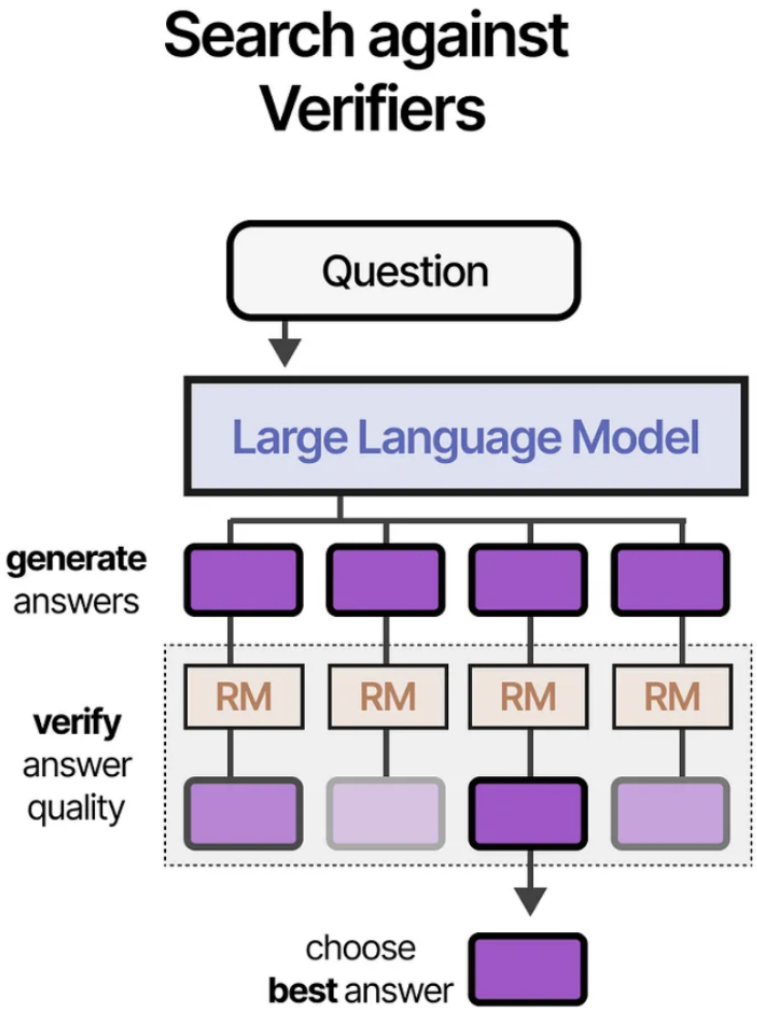
Test-time compute can be roughly put into two categories

- search against verifiers: sampling CoT generations and selecting the best answer (output-focused)
- modifying proposal distribution: the model is trained to create improved reasoning steps (input-focused)



Maarten Grootendorst

Search against verifiers



Search against verifiers

Search against verifiers involves two steps

- multiple samples of reasoning processes and answers are created
- a verifier scores the generated output

The **verifier** is based on a LLM, fine-tuned for judging the quality of the process and acting as a reword model (RM).

Two methodologies for the verifier

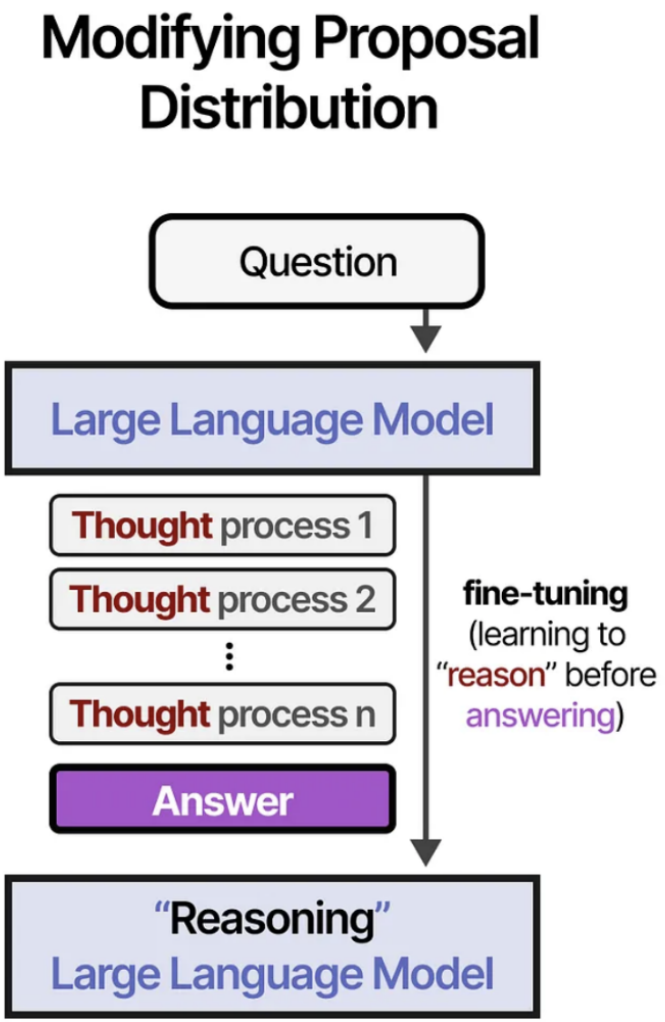
- outcome verifier: only judges the answer
- process verifier: judges the reasoning that leads to the answer

Search against verifiers

The verifier may implement any of the following **search strategies**

- majority voting
- best-of-N samples
- beam search
- Monte Carlo tree search: node expansion, rollout and backprop

Modifying proposal distribution

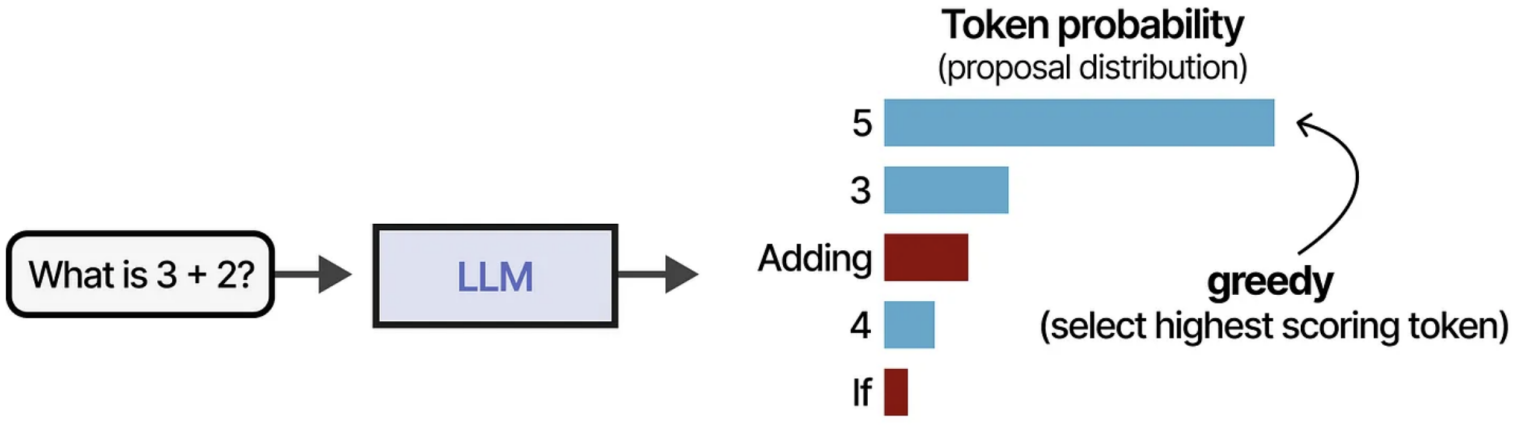


Maarten Grootendorst

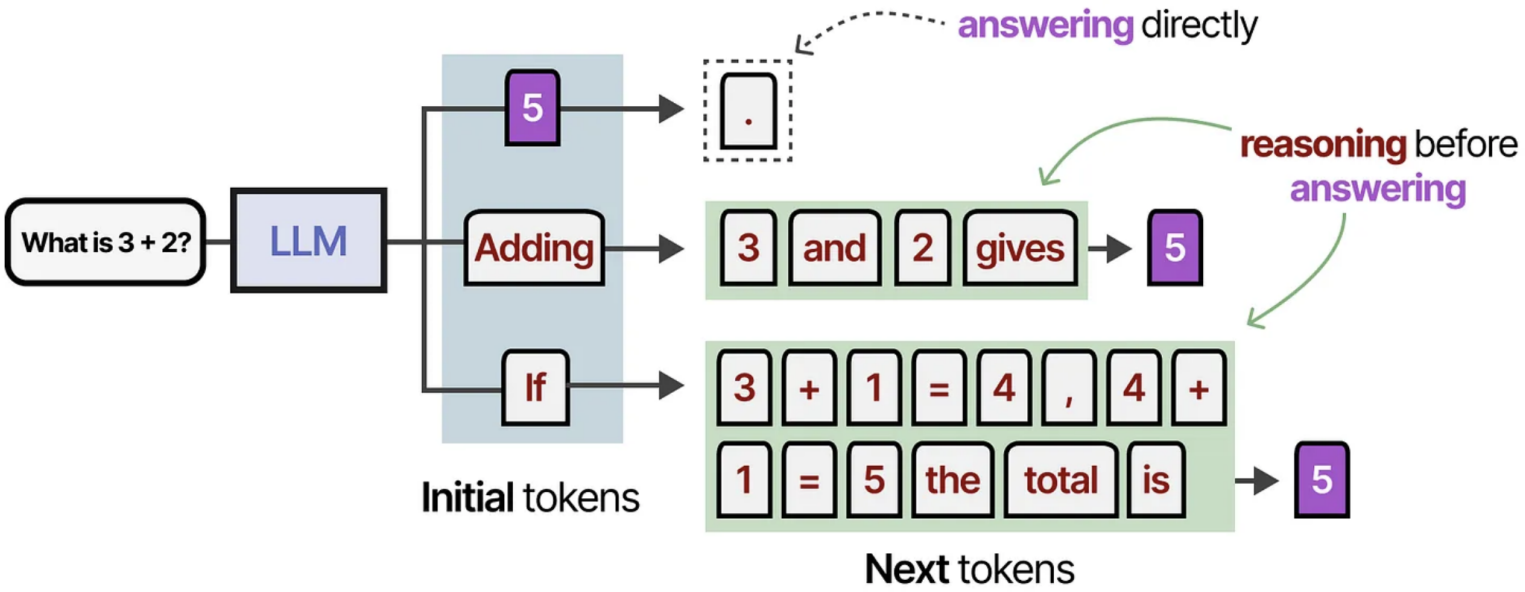
Modifying proposal distribution

Modifying proposal distribution trains the model to create improved reasoning steps.

We are interested in tokens that are more likely to lead to a reasoning process.



Modifying proposal distribution



Maarten Grootendorst

Modifying proposal distribution

Two general methods for modifying the next token distribution.

Use of prompt engineering: provide examples to the model, or use prompt that induce reasoning-like behavior.

Rewarded the model for generating reasoning steps; this involves reasoning data and reinforcement learning.

Research papers

Emil Widlund on Unsplash

Research papers

Title: Towards Reasoning Era: A Survey of Long Chain-of-Thought for Reasoning Large Language Models
Authors: Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, Wanxiang Che
Source: 13 March 2025
Content: Comprehensive survey on long CoT, offering a unified perspective on large reasoning models.

https://arxiv.org/pdf/2503.09567