Deep Learning

Large Language Model

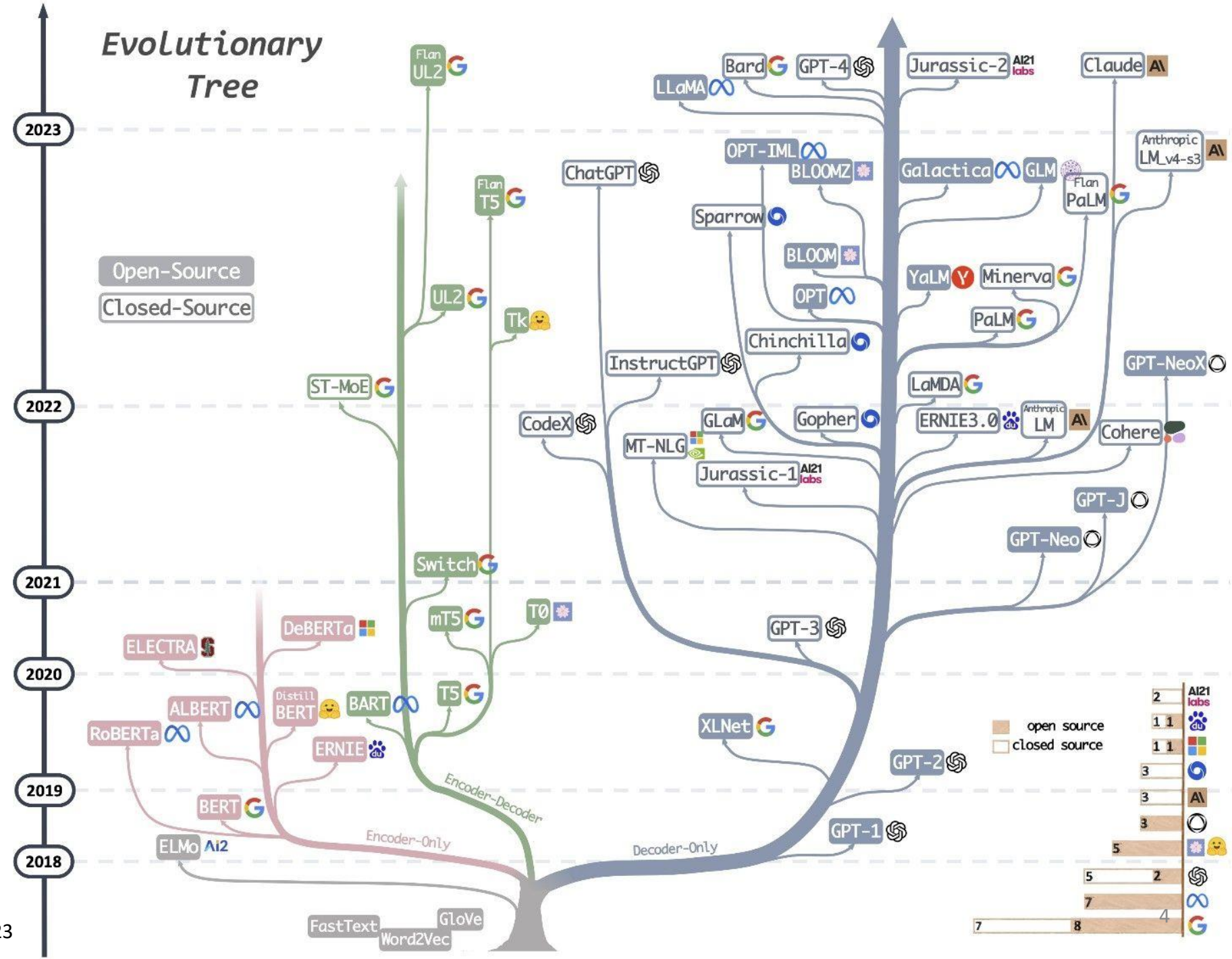
Lionel Fillatre

2025-2026

Outline

- Introduction
- Encoder-Decoder
- Encoder-Only
- Decoder-Only
- Pre-training language model
- Adaptation
- Alignment of LLM
- Conclusion

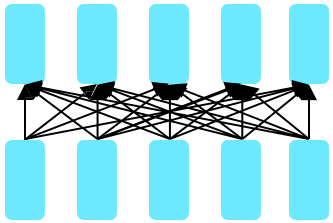
Introduction



Yang et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond, 2023

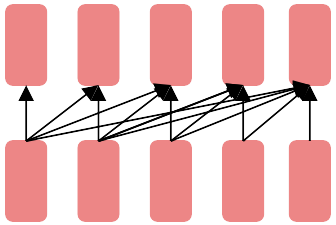
Impact of Transformers

- A building block for a variety of LMs



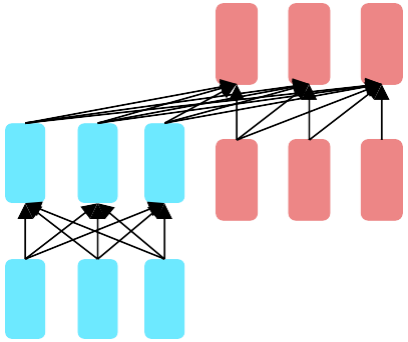
Encoders

- Examples: BERT, RoBERTa, SciBERT.
- Captures bidirectional context.



Decoders

- Examples: GPT-2, GPT-3, LaMDA
- Other name: causal or auto-regressive language model
- Nice to generate from; can't condition on future words



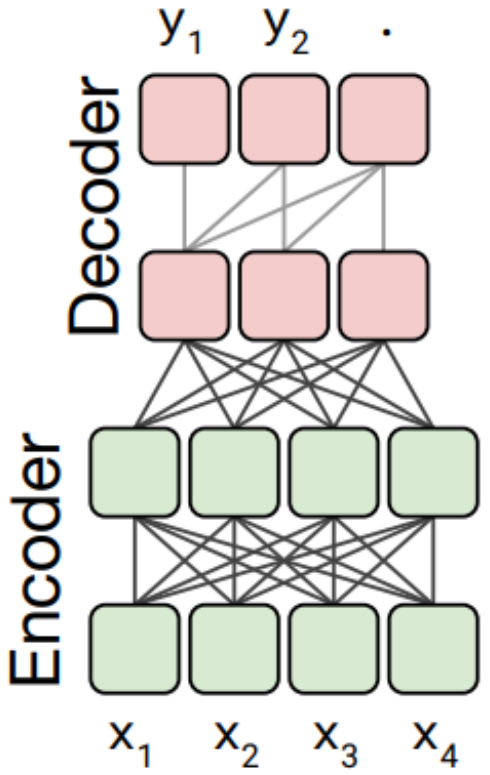
Encoder-Decoders

- Examples: Transformer, T5, Meena
- What's the best way to pretrain them?

Encoder-Decoder

Encoder-Decoder models: T5

- Architecture:
  - The **encoder** portion benefits from bidirectional context.
  - The **decoder** portion is used to train the whole model through language modeling.
  - Similar to the original Transformer enc-dec architecture.



T5 (Text-to-Text Transfer Transformer) is a series of large language models developed by Google AI introduced in 2019.

Encoder-Decoder models: T5

- Pretraining objective: randomly corrupt tokens and replace with sentinel tokens (<x>, <y>) that is unique over the example.

Targets

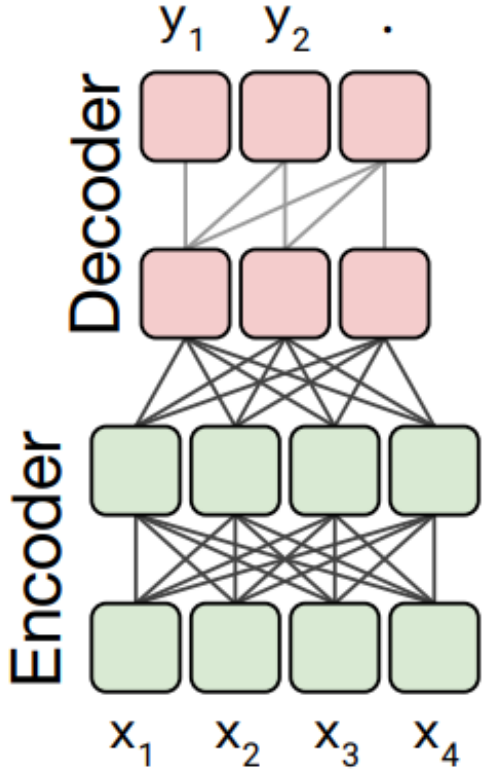
<X> for inviting <Y> last <Z>

Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.



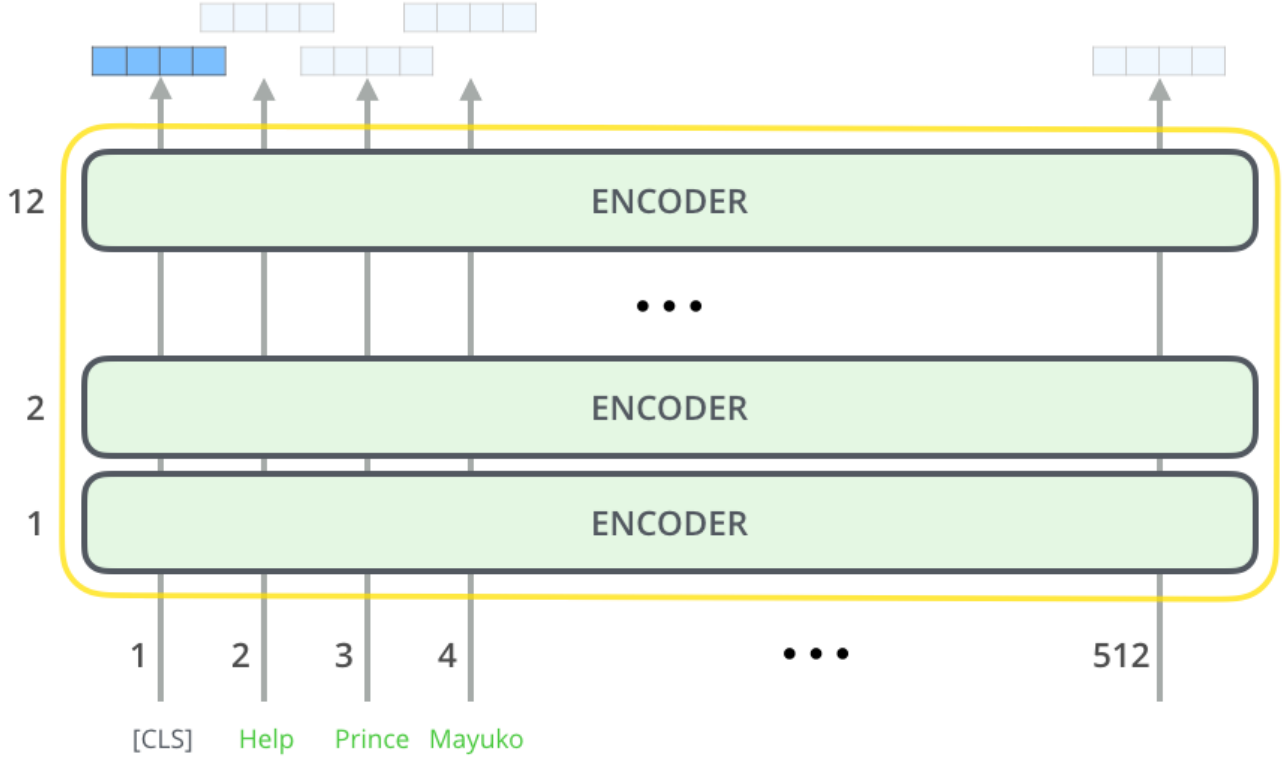
Encoder-Only

Bidirectional encoder representations from transformers (BERT) is a language model introduced in October 2018 by researchers at Google.

Encoder-only models (BERT)

- Transformer encoder-only
- BERT is trained to uncover masked tokens.

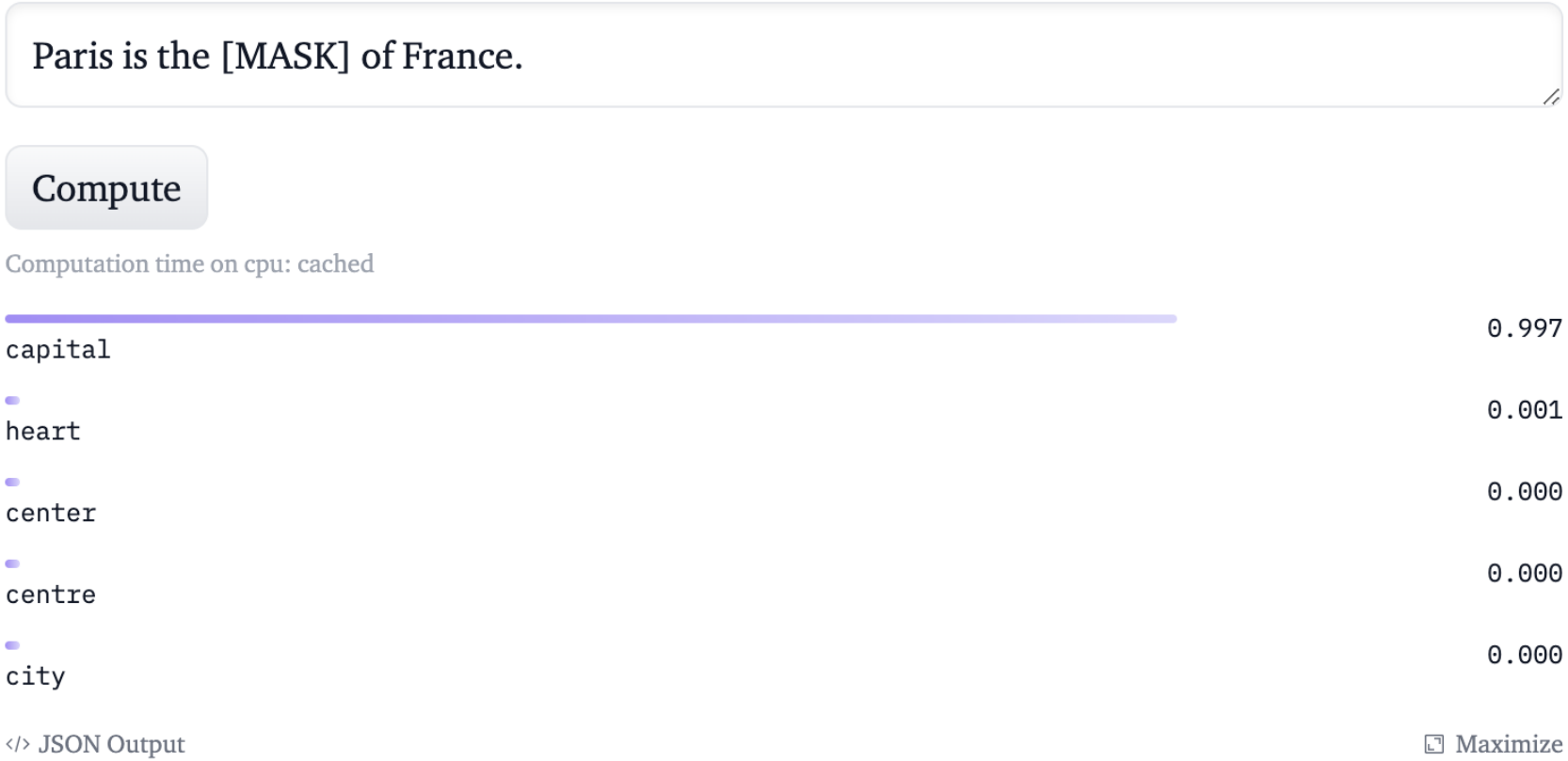
brown 0.92
lazy 0.05
playful 0.03



[BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Devlin et al. 2018]

Encoder-only models (BERT): Probing its predictions

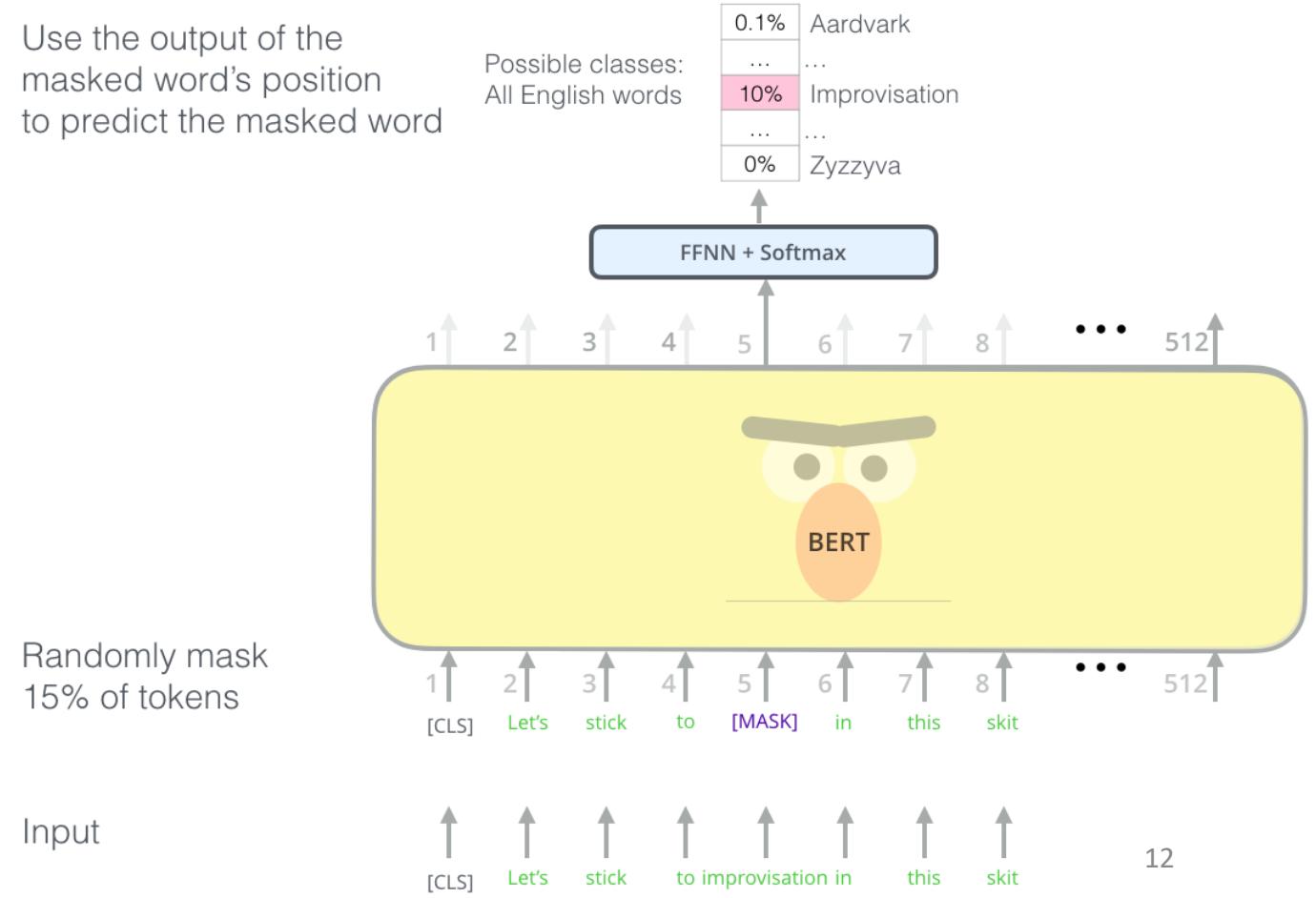
- Masking words forces BERT to use context in both directions to predict the masked word.



https://huggingface.co/bert-base-uncased

Encoder-only models (BERT): Pre-training Objectives

- **Token masking:** Randomly mask 15% of tokens and train the model to recover them.



Encoder-only models (BERT): Pre-training Objectives

- **Token masking:** Randomly mask 15% of tokens and train the model to recover them.
  - Too little masking: too expensive to train
  - Too much masking: underdefined
    - (not enough info for the model to recover the masked tokens)
- **Sentence ordering:** Predict sentence ordering
  - Learns the relationships between sentences
  - 50% correct ordering, and 50% random incorrect ones

Sentence A = `The man went to the store.`
Sentence B = `He bought a gallon of milk.`
Label = `IsNextSentence`

Sentence A = `The man went to the store.`
Sentence B = `Penguins are flightless.`
Label = `NotNextSentence`

Decoder-Only

Decoder-only (GPT)

- Generate sequences where each token is predicted based on the previously generated tokens
- Use causal masking to ensure the causality
- Trained to maximize log-likelihood defined for next-token prediction.

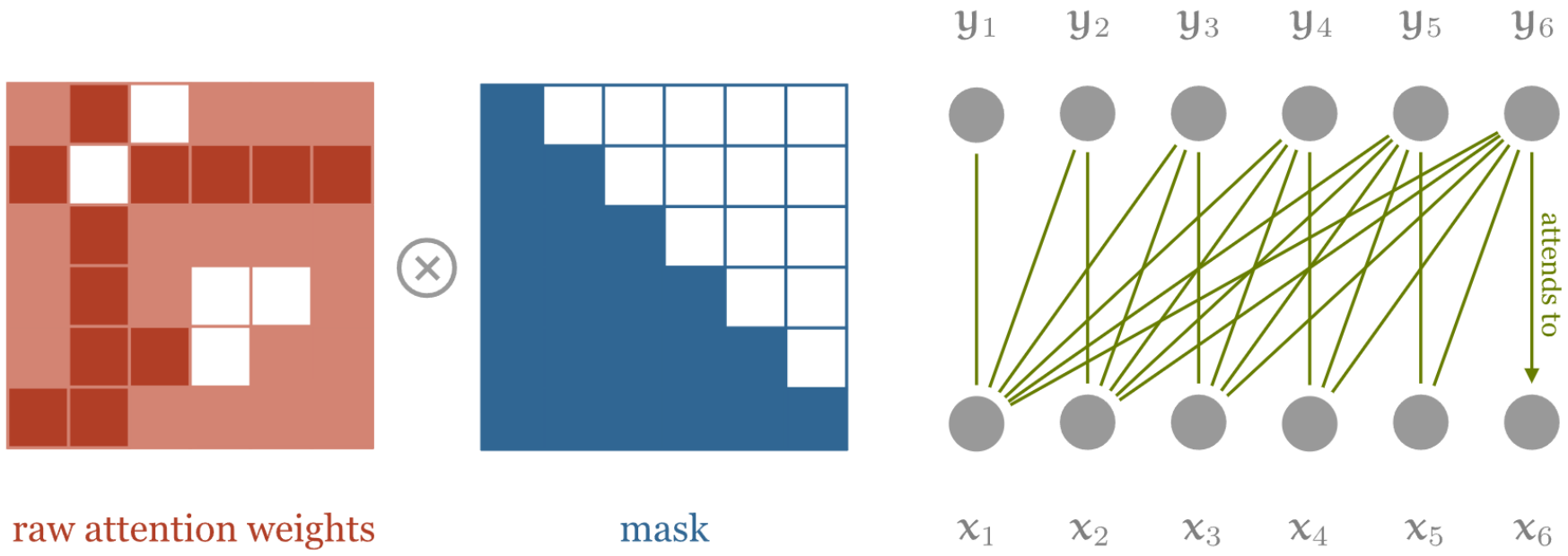


Figure source: https://peterbloem.nl/blog/transformers

https://lmarena.ai/leaderboard/

LMSys ChatArena

Leaderboard Overview

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it here.

Text

9 days ago

Rank (UB) ↑	Model	Score	Votes
1	gemini-2.5-pro	1451	54 087
1	claude-opus-4-1-20250805-thi…	1447	21 306
1	claude-sonnet-4-5-20250929-t…	1445	6 287
1	gpt-4.5-preview-2025-02-27	1441	14 644
2	chatgpt-4o-latest-20250326	1440	40 013
2	o3-2025-04-16	1440	51 293
2	claude-sonnet-4-5-20250929	1438	6 144
2	gpt-5-high	1437	23 580
2	claude-opus-4-1-20250805	1437	33 298
3	qwen3-max-preview	1434	18 078

View all

WebDev

5 days ago

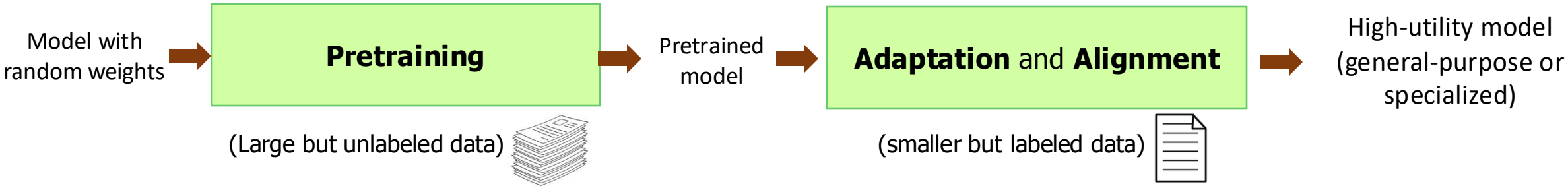
Rank (UB) ↑	Model	Score	Votes
1	GPT-5 (high)	1478	5 848
1	Claude Opus 4.1 thinking-16k…	1472	5 312
1	Claude Opus 4.1 (20250805)	1462	5 582
4	Claude Sonnet 4.5 (thinking …	1421	1 337
4	Gemini-2.5-Pro	1401	11 022
4	GLM-4.6	1398	5 442
4	DeepSeek-R1-0528	1394	4 800
5	Claude Sonnet 4.5	1385	4 127
6	Claude Opus 4 (20250514)	1384	9 238
6	GLM-4.5	1381	4 360

View all

Pre-training language model

Training Pipeline for LLMs

- There is extensive literature about best practices for pretraining
  - What choice of architectures are good?
  - How do you prepare pre-training data?
  - What considerations go into efficient training of the models?



The pre-training data size and sources

Model Name	Release	Pre-training data #Tokens	Training Dataset
BERT	2018	3.3B	BooksCorpus (800M), English Wikipedia (2.5B)
GPT-1	2018	13B	BooksCorpus
GPT-2	2019	40B	WebText: scraping outbound links from Reddit post with >= 3 karma
T5	2019	34B	C4 which is the cleaned up version of CommonCrawl
GPT-3	2020	400B	Common Crawl (filtered), WebText2, Myrstry books!! (Books1, Books2), Wikipedia
Gopher	2021	1.4T	MassiveText
BLOOM	2022	350B	ROOTS corpus, a dataset comprising hundreds of sources in 46 natural and 13 programming languages (59 in total)
PaLM	2022	2.81T	Web documents, books, Wikipedia, conversations, GitHub code
LaMDA	2022	1.56T	Public dialog data and web documents
Chinchilla	2022	1.4T	MassiveText
LLaMA2	2023	2.0T	A new mix of publicly available online data
GPT-4	2023	?	?
Claude-3	2023	?	?
OLMo 2	2024	5.6T	OLMo-Mix-1124(stage1) + Dolmino-Mix-1124(stage 2)
Qwen2.5	2024	7T	
DeepSeek (V3)	2024	14.8T	GitHub's Markdown and StackExchange
LLaMA3	2024	15T	A new mix of publicly available online data

On Pre-training Objectives

- So far, the dominant objective we have seen is “next-token” prediction.
- In reality, any “marginal” observations about language can be a source of supervision.

- **Prefix language modeling**
  - **Input:** `Thank you for inviting`
  - **Output:** `me to your party last week`

- **BERT-style denoising**
  - **Input:** `Thank you <M> <M> me to your party apple week`
  - **Output:** `Thank you for inviting me to your party last week`

- **Deshuffling**
  - **Input:** `party me for your to. last fun you inviting week Thanks.`
  - **Output:** `Thank you for inviting me to your party last week`

Exploring the limits of transfer learning with text-to-text transfer transformers, 2020

Adaptation

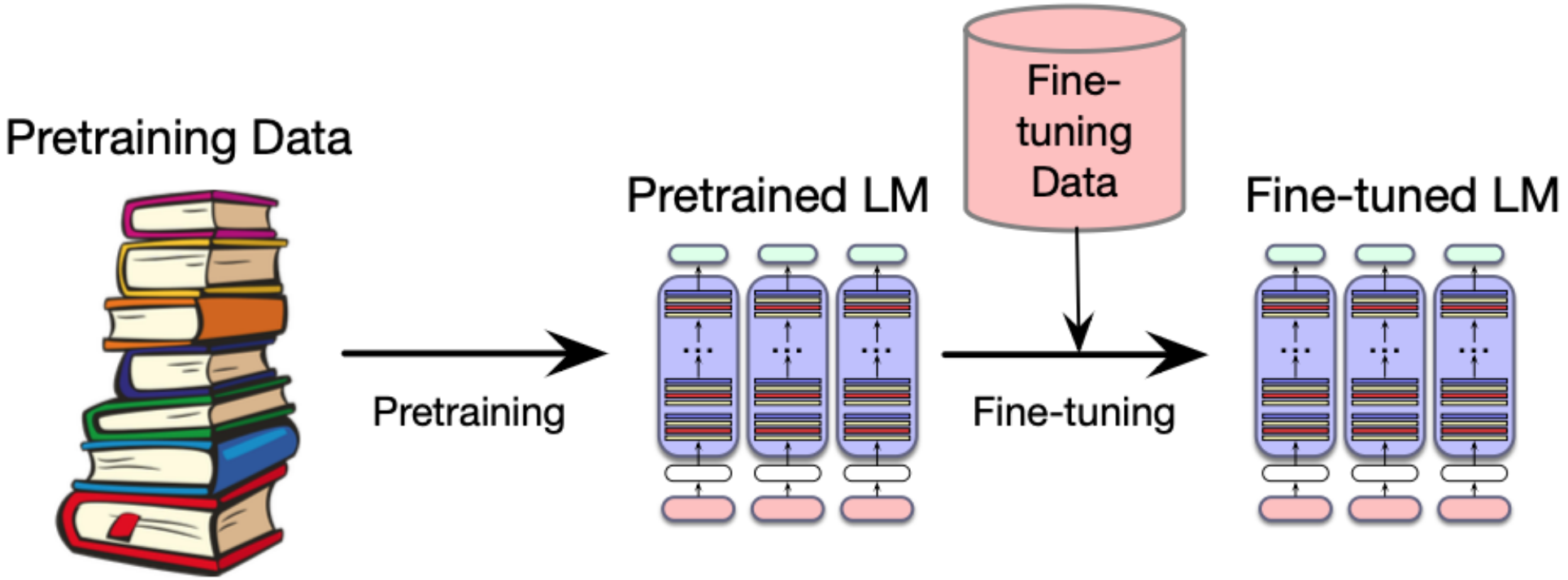
Adapting Language Models

- You have a pre-trained language model that is pre-trained on massive amounts of data.

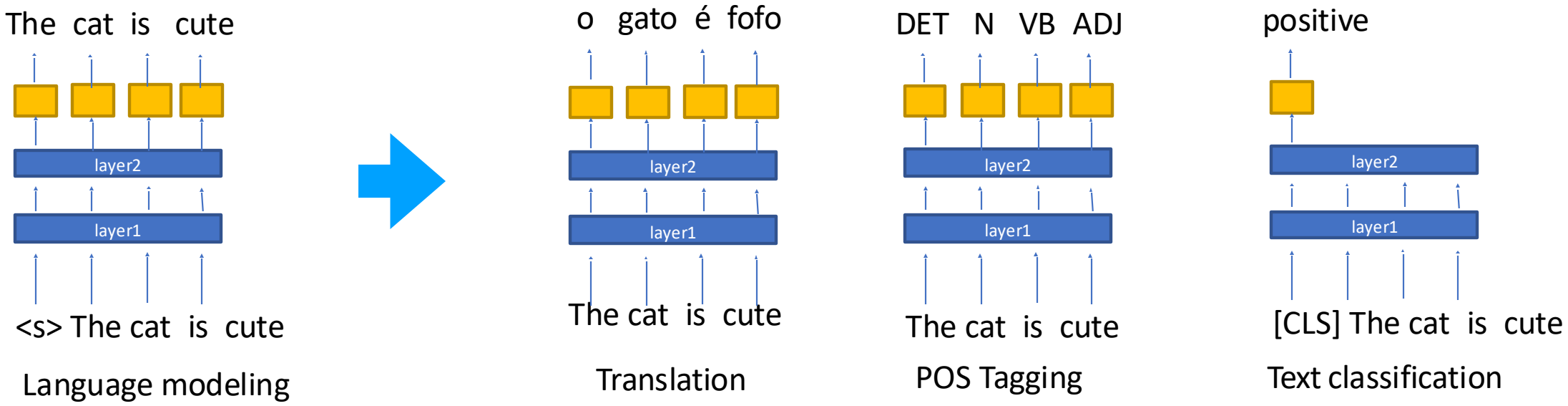
- They do not necessarily do useful things—they only complete sentences.

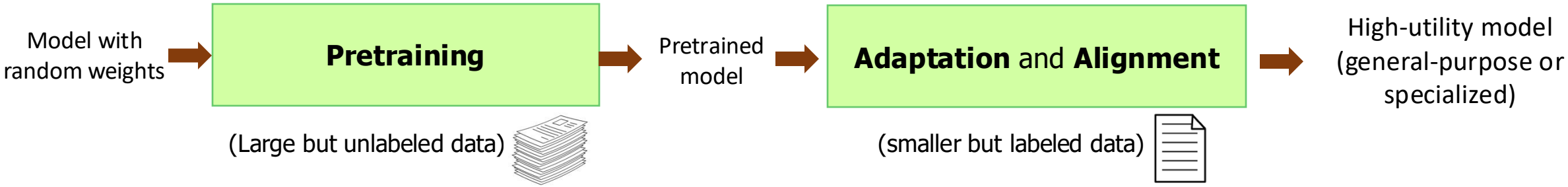
- Now how to you ”adapt” them for your use-case?
  - **Tuning:** adapting (modifying) model parameters
  - **Prompting:** adapting model inputs (language statements)

Finetuning


Pretraining Data
Pretraining
Pretrained LM
Fine-tuning Data
Fine-tuning
Fine-tuned LM

Fine-Tuning for Tasks



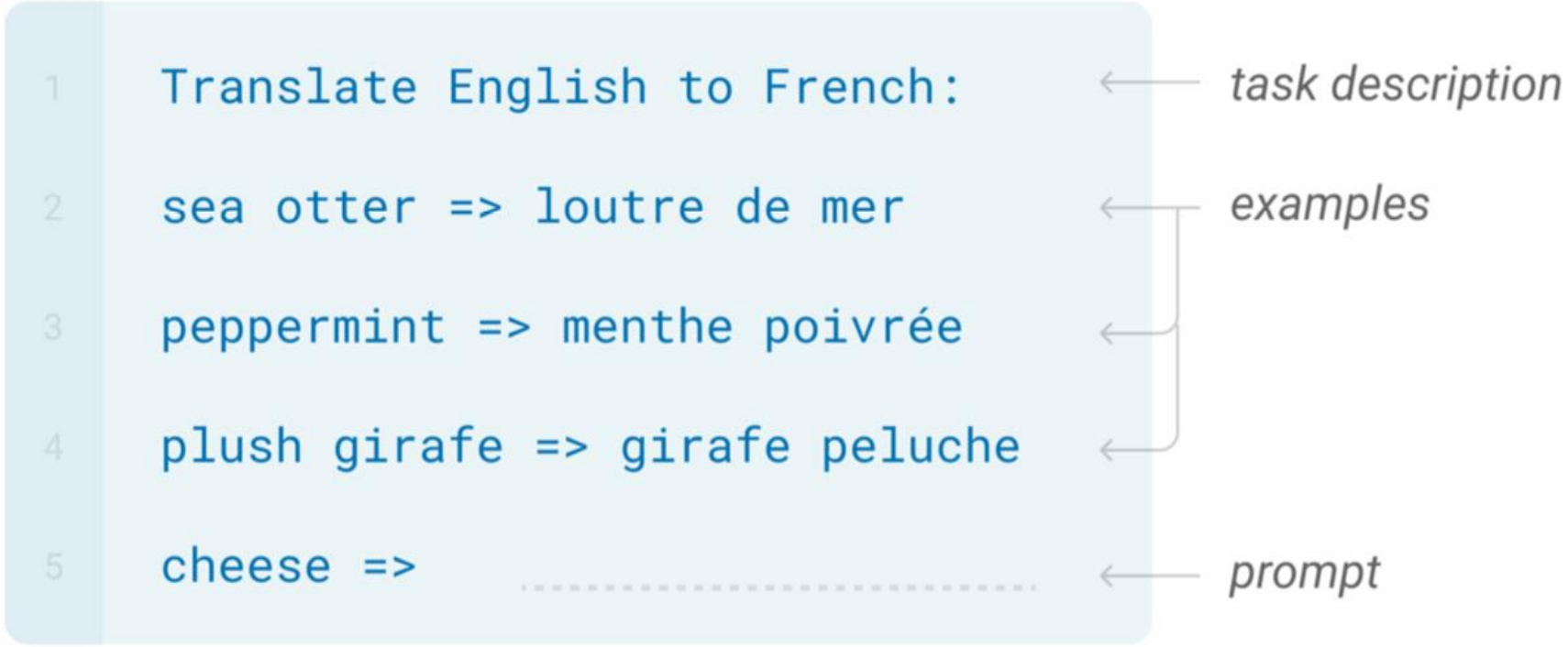


Limitations of Pre-training, then Fine-tuning

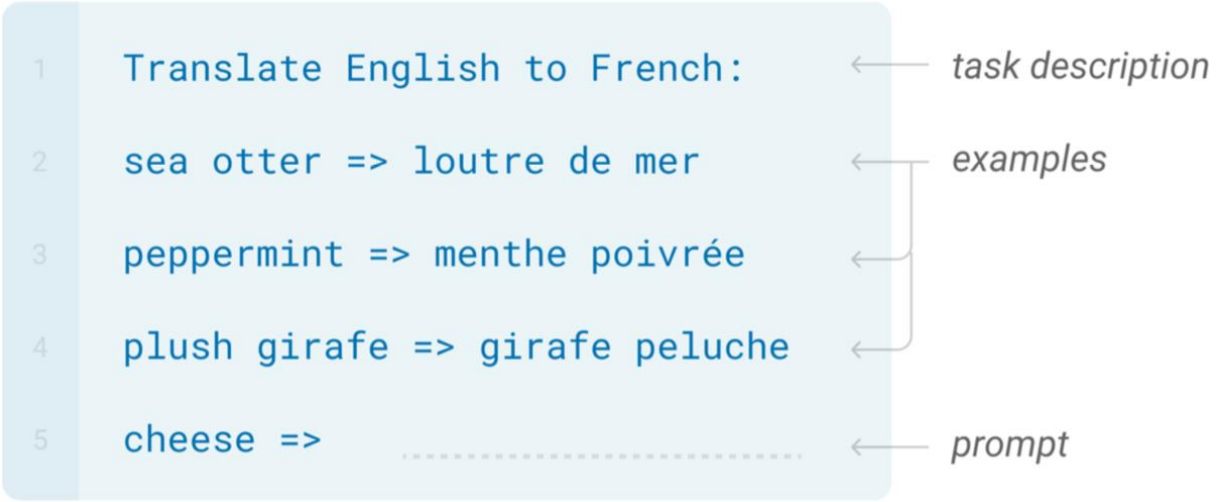
- Often you need a large labeled data for fine-tuning
- End up with many copies (or sub-copies) of the same (sub)model
- **Prompts** get language models to generate text
  - They can be viewed as a learning signal: they help language models learn to perform novel tasks.
  - The task of prompting with examples is sometimes called **few-shot prompting**
  - It is contrasted with **zero-shot prompting** which means instructions that don't include labeled examples.
- For this reason, we also refer to prompting as in-context learning

https://www.hopsworks.ai/dictionary/in-context-learning-icl

In-Context Learning



In-Context Learning



- Learns to do a downstream task by conditioning on input-output examples!

- **No weight update** — our model is not **explicitly pre-trained** to learn from examples
  - The underlying models are quite general

- Today's focus:
  - How to use effectively in practice?
  - Fundamentally, why does it work?

Adding “thought” before “answer”

(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（Wei et al., 2022）

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. ✓

Step-by-step demonstration

Step-by-step Answer

The use of natural language to describe rationales is critical for the success of CoT.

Alignment of LLM

Alignment of LLM

- Large models trained on very large corpuses of text
  - Basically, the entire internet
- Including, problematic texts
- After pretraining, LLMs generate:
  - Toxic languages: insults, etc.
  - Questionable answers: “how to steal someone identity”
  - Harmful knowledge: “how to build a bomb”
  - Lack of empathy: risk of suicide
  - Explicit content

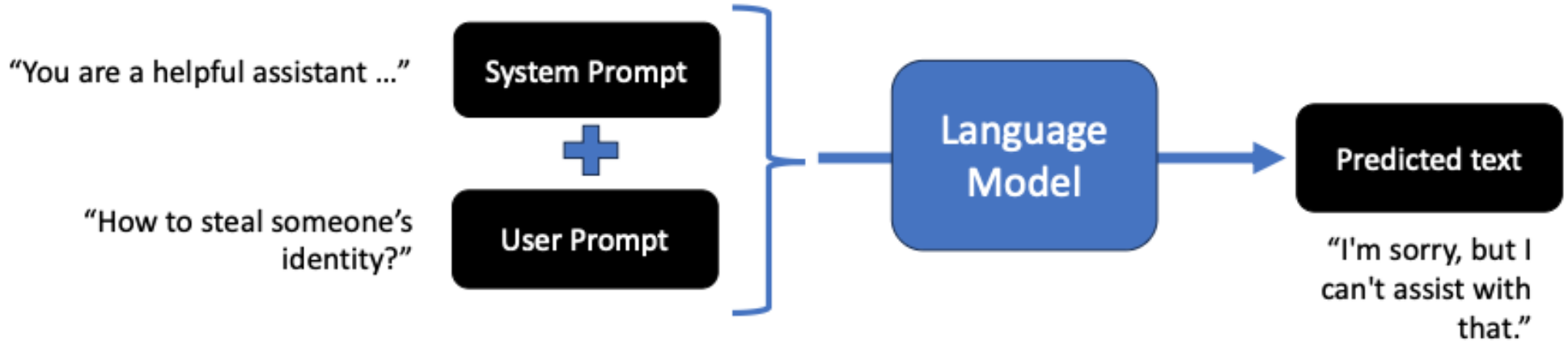
Need to “align” LLM with human values

Alignment of LLM

- First solution: System prompt
- Added before the user prompt to guide the reply
- Default system prompt of Llama-2

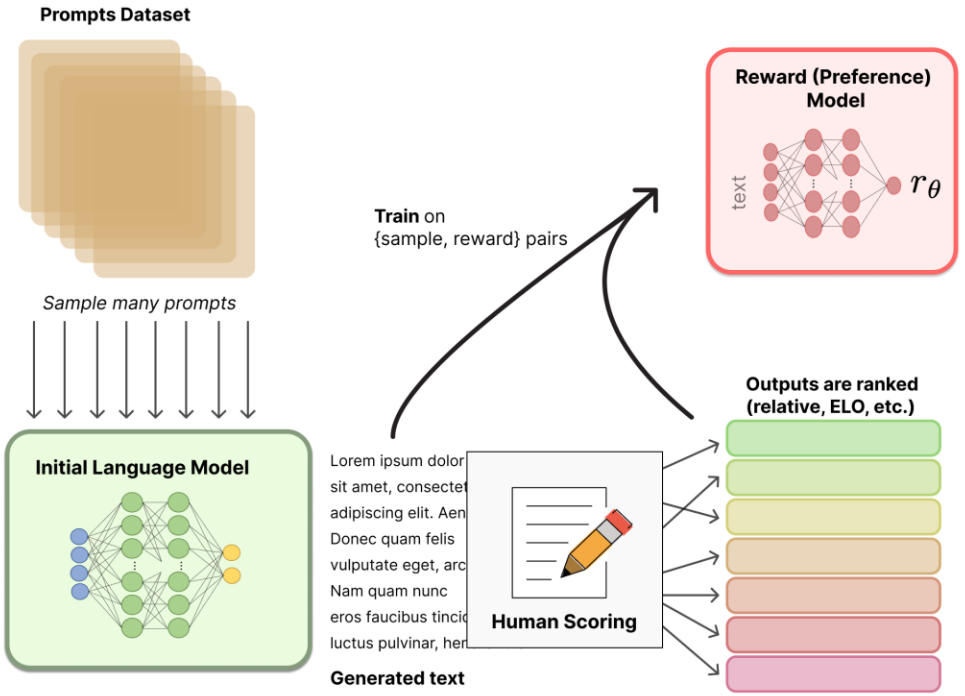
  You are a helpful, respectful and honest assistant. Always answer as helpfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content. Please ensure that your responses are socially unbiased and positive in nature. If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.

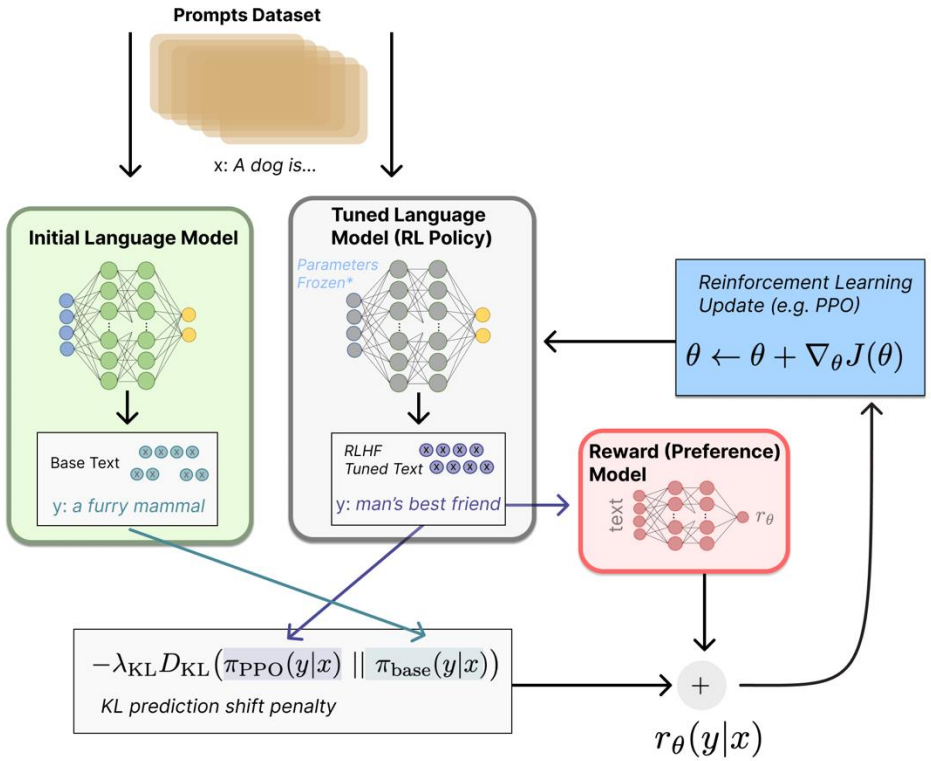
Alignment of LLM



Alignment of LLM

- Second solution: Reinforcement Learning from Human Feedback (RLHF)





https://huggingface.co/blog/rlhf

Conclusion

Conclusion

- LLM: a very hot topic!
- A huge number of architectures and tricks
- The training is expansive and difficult
- It is really a black box
- But it is amazing in practice!
- New mechanisms are invented to improve LLM performance without finetuning them, prompt tuning for example