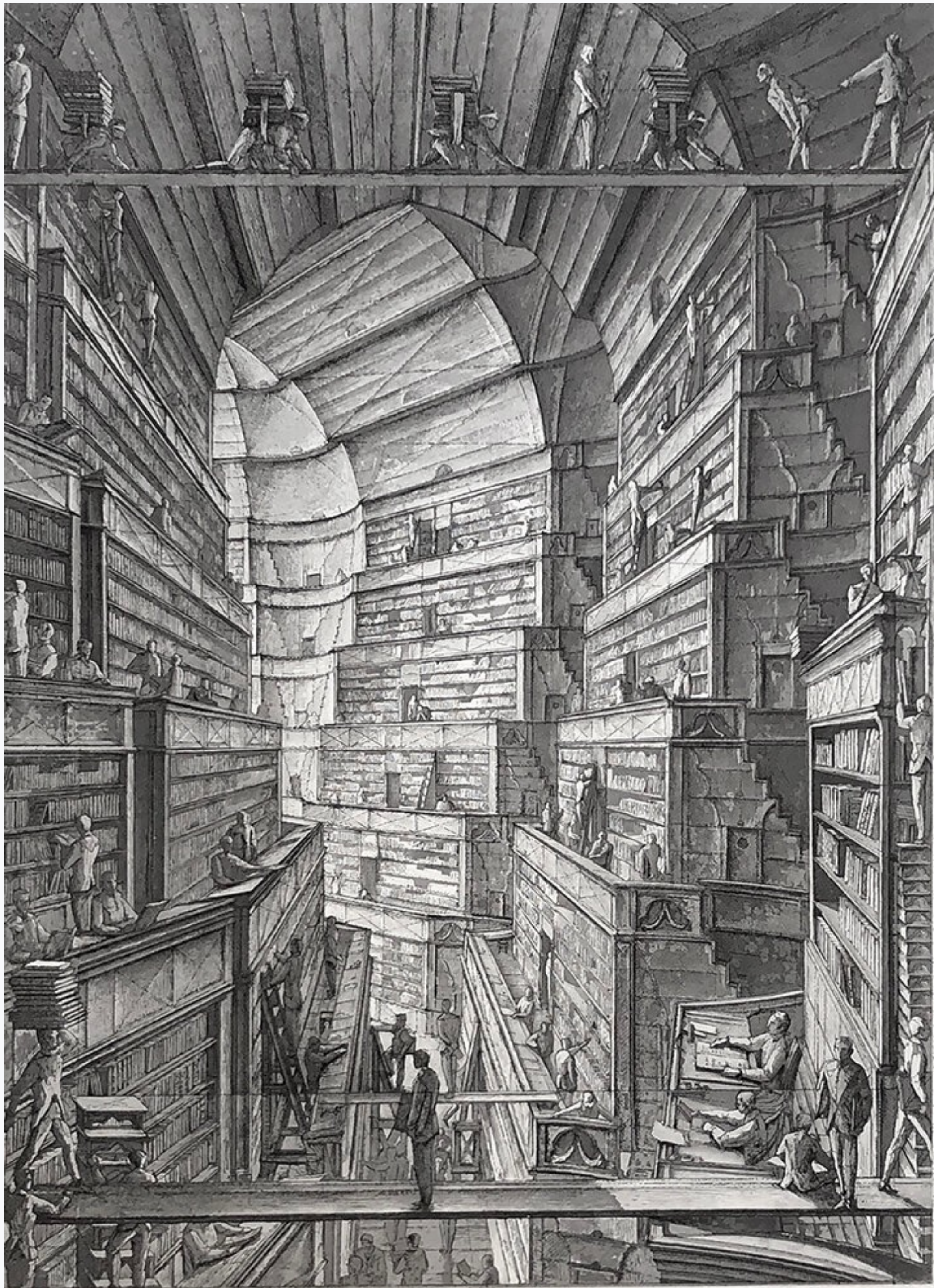
Large Language Models

Past (1980-2017) and present (2017-2024)

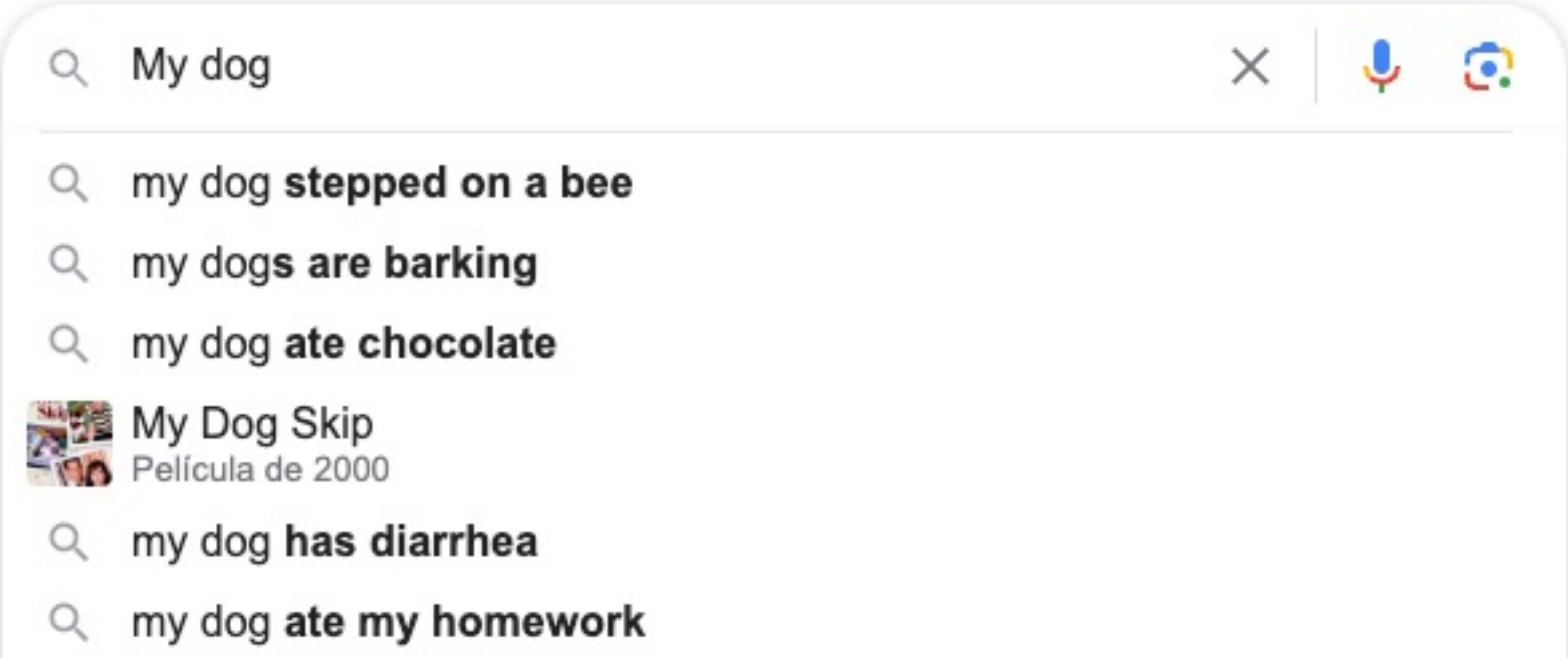
Alberto Lumbreras
AI Researcher

www.albertolumbreras.net



What is a language model

A model that learns the probability of each possible sentence



A timeline of language models

1982. RNN/Hopfield

Hopfield (1982). "Neural networks and physical systems with emergent collective computational abilities". *PNAS*

1995. RNN/LSTM

Hochreiter, Schmidhuber (1997). “Long Short-Term Memory”. *Neural Computing*
Gers, Schmidhuber, Cummins (1999). "Learning to forget: Continual prediction with LSTM". *ICANN*

2014. RNN/LSTM/Sequence-to-sequence

Sutskever, Vinyals, Quoc (2014). “Sequence to Sequence Learning with Neural Networks.” *NeurIPS*

2016. RNN/LSTM/Sequence-to-sequence/Attention

Bahdanau, Cho, and Bengio. (2016). “Neural Machine Translation by Jointly Learning to Align and Translate.” *ICLR*.

2017. Self-Attention/Transformers

Vaswani et al. 2017. “Attention Is All You Need.” *NeurIPS*

Parallelizable training

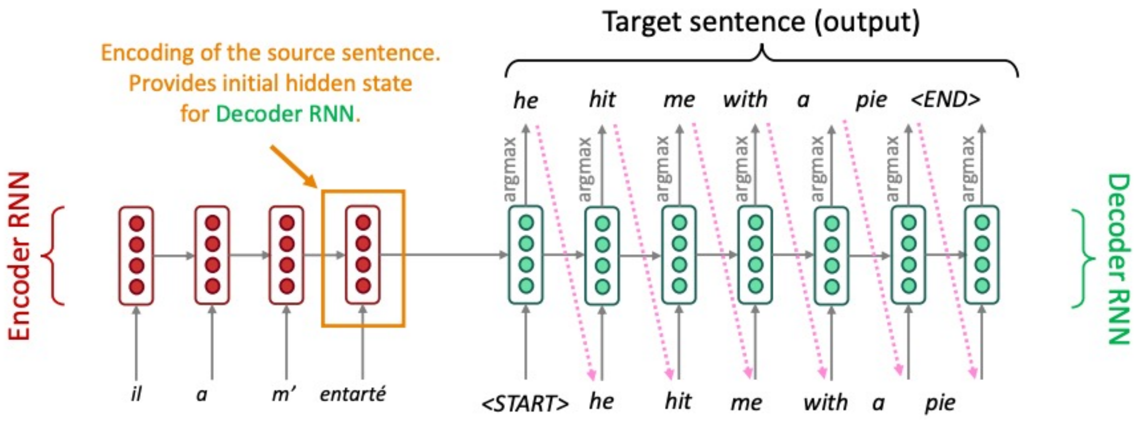
GPU

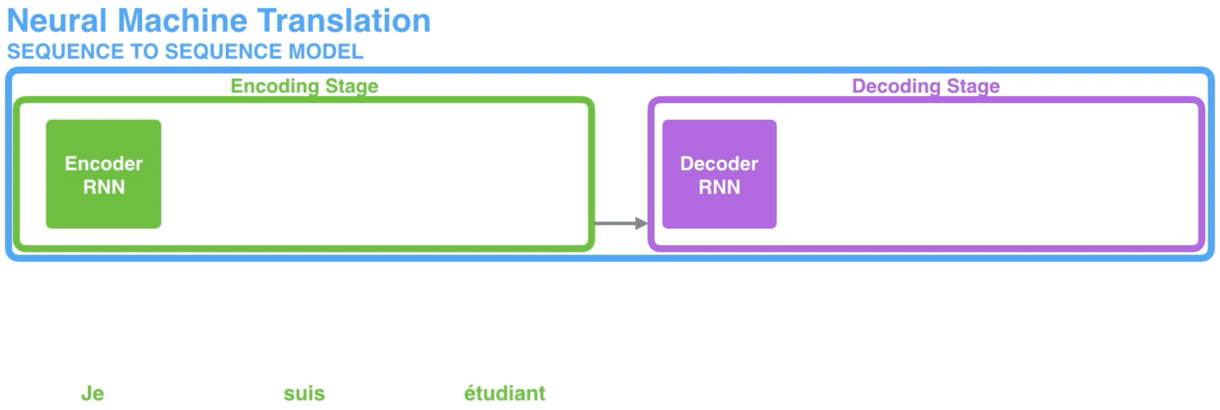
Data

Recurrent Neural Networks / Sequence-to-Sequence (2014)

And encoder RNN and a decoder RNN

- First, encode the input sentence (encoder)
- Then, decode looking only at the full encoded sentence.
- Encoder and decoder are RNNs





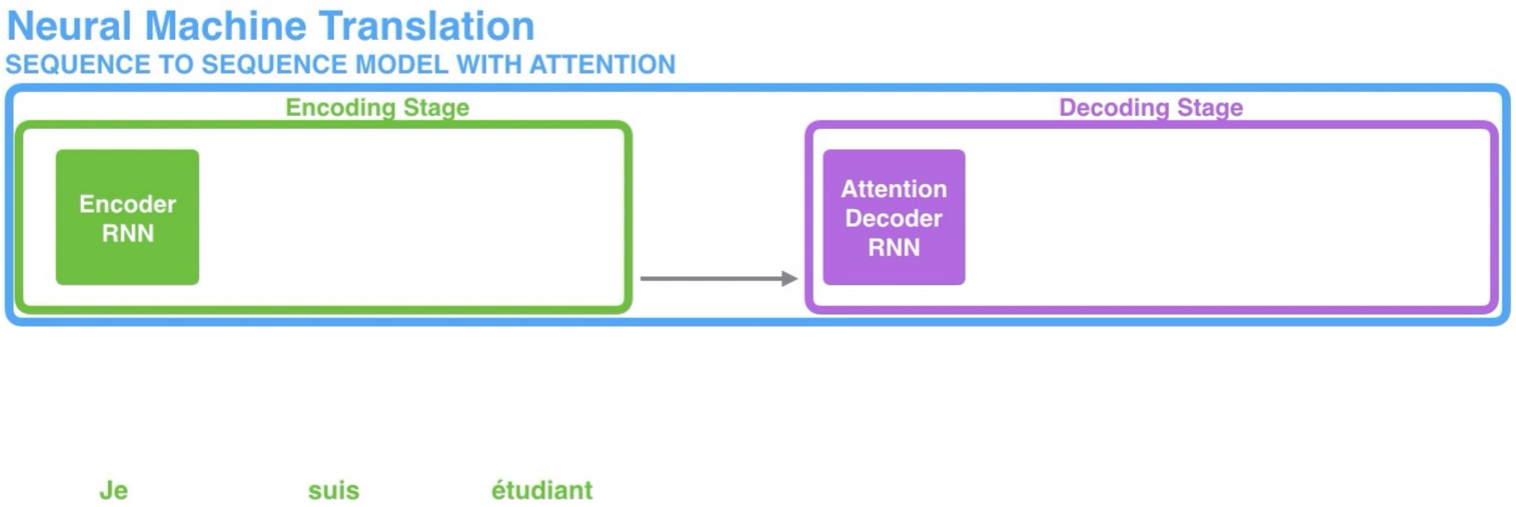
- Machine Translation trained **end-to-end**

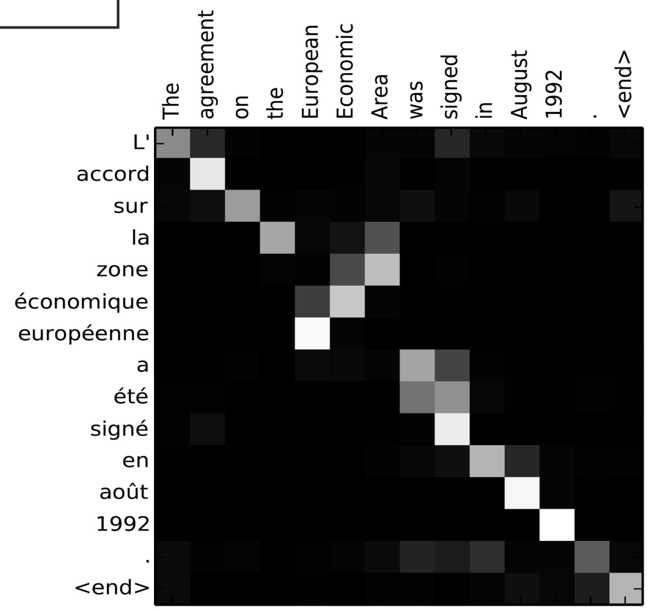
- Decoder only sees **last context vector**

Recurrent Neural Networks / Seq2Seq / Attention (2016)

An improvement to seq2seq models for machine translation

- Potentially use all encoder hidden states when decoding
- Learn to decide which hidden states are more relevant at each timestep (attention)





- Attention is not affected by distance

- Naïve implementation needs Lx x Ly comparisons

A timeline of language models

Everything changed in 2017

1982. RNN/Hopfield

Hopfield (1982). "Neural networks and physical systems with emergent collective computational abilities". *PNAS*

1995. RNN/LSTM

Hochreiter, Schmidhuber (1997). “Long Short-Term Memory”. *Neural Computing*
Gers, Schmidhuber, Cummins (1999). "Learning to forget: Continual prediction with LSTM". *ICANN*

2014. RNN/LSTM/Sequence-to-sequence

Sutskever, Vinyals, Quoc (2014). “Sequence to Sequence Learning with Neural Networks.” *NeurIPS*

2016. RNN/LSTM/Sequence-to-sequence/Attention

Bahdanau, Cho, and Bengio. (2016). “Neural Machine Translation by Jointly Learning to Align and Translate.” *ICLR*.

2017. Self-Attention/Transformers

Vaswani et al. 2017. “Attention Is All You Need.” *NeurIPS*

Parallelizable training

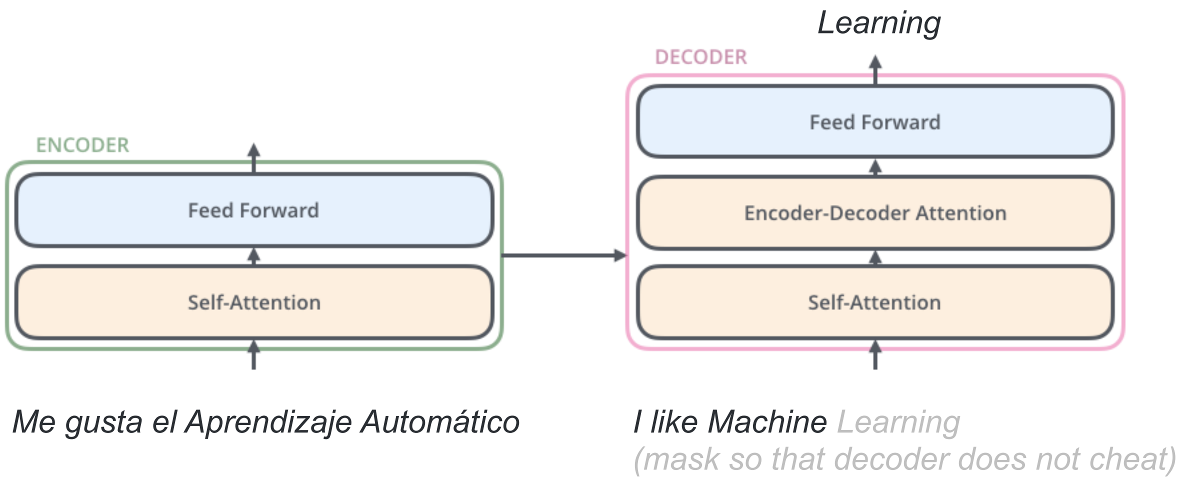
GPU

Data

Transformer (2017)

Good-bye RNNs. Attention is all you need

- Replace RNNs by self-attention
- Introduce positional encoding (since self-attention loses position information)
- Stack multiple self-attention layers



100% **parallelizable** (no RNNs)

“Pre-train and fine-tuning” paradigm (2017-)

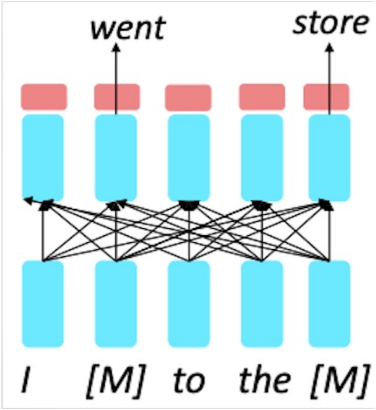
Transformers pre-trained on internet data happened to acquire some world knowledge

- New trend from 2018 was to –re-train transformers on huge unsupervised datasets
- Them use a (much smaller, often supervised) task-specific dataset to fine-tune the model

Encoder

Masked LM

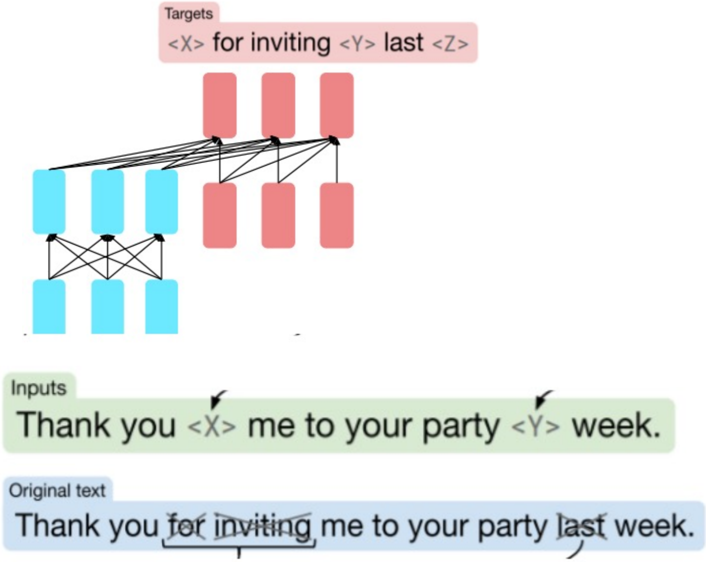
BERT



Encoder-decoder

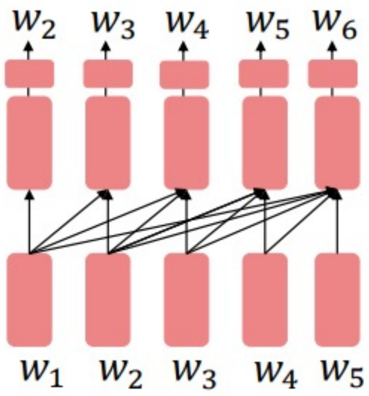
Span corruption

T5



Decoder

Language model

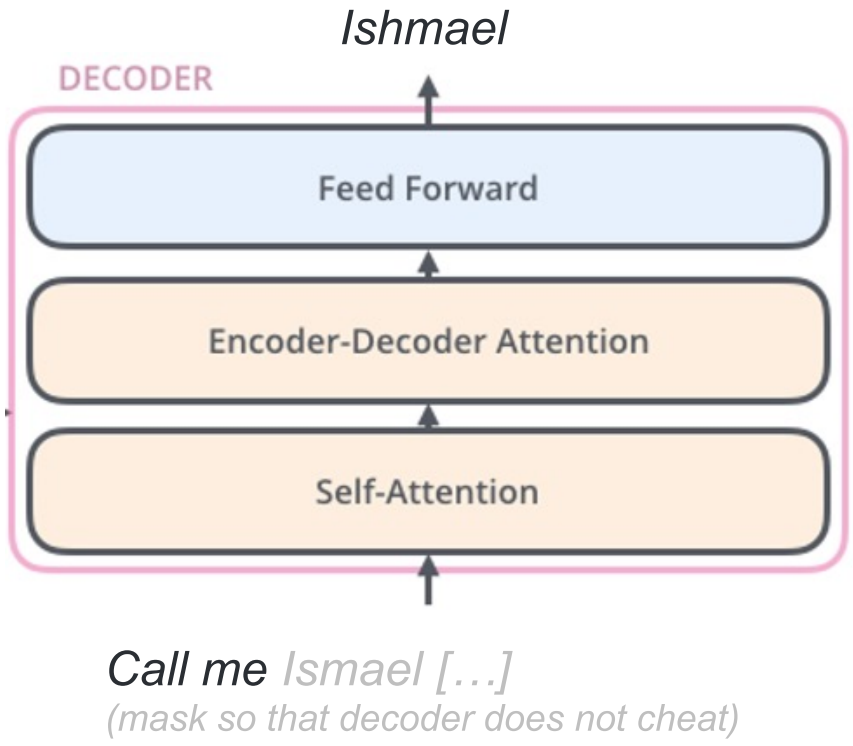


Decoders are language models

They predict next word given sentence so far

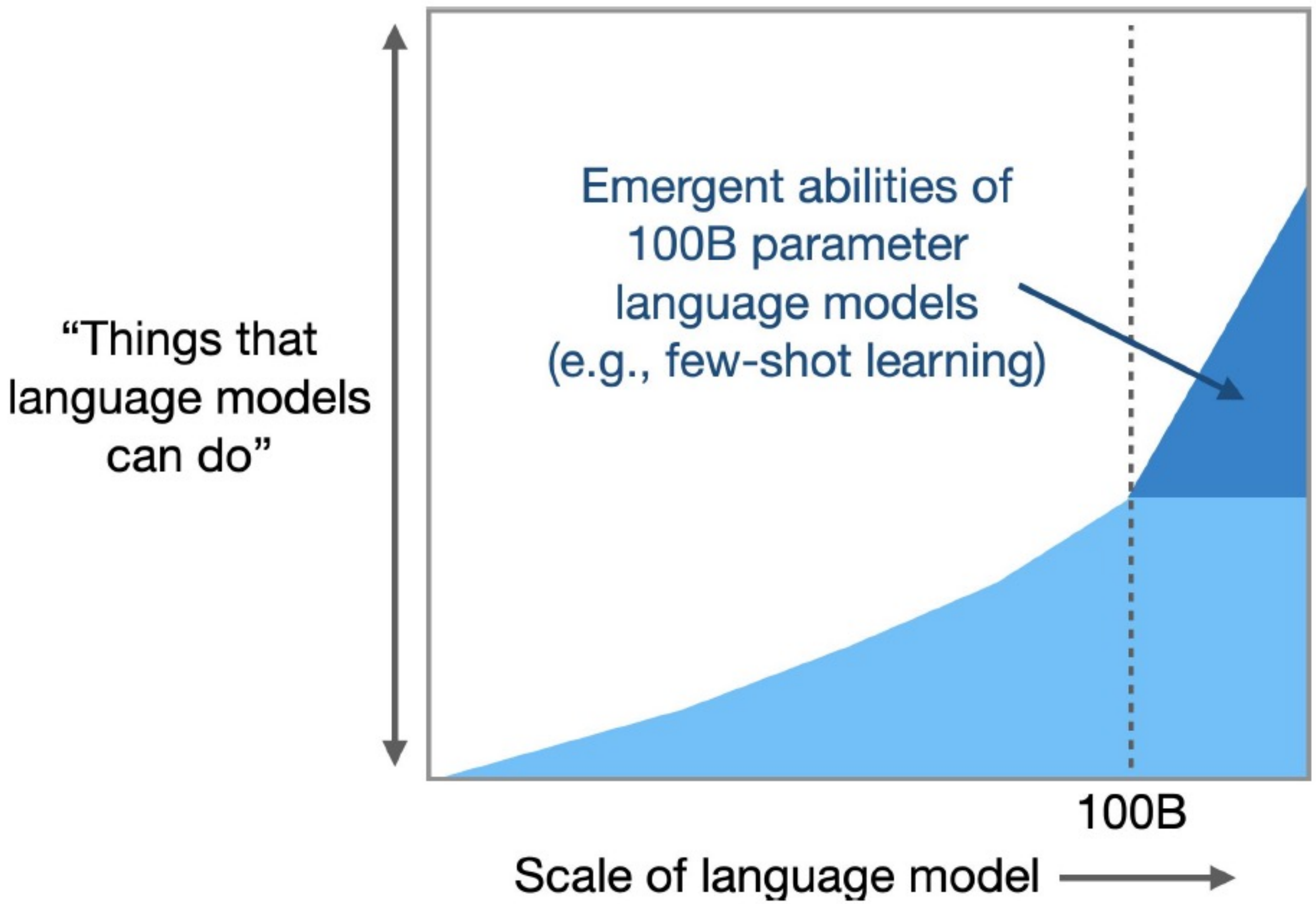
Training corpus:

Call me Ishmael. Some years ago—never mind how long precisely—having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. […]



Decoders architectures are the ones that have shown better properties when **scaling the models and the training data.**

Emerging properties, towards generative multi-task models



A timeline of (some) large language models (decoders)

2019. GPT-2

Radford et al (2019) Language Models are **Unsupervised Multitask Learners**”. *Tech report*

1.5B params

40GB of text

2020. GPT-3

Brown,…,Radford, Sutskever, Amodei. (2020). “Language Models Are **Few-Shot Learners**”. *NeurIPS*

175B params

570GB of text

2023. GPT-3/ChatGPT

Ouyang, et al. (2022). “Training Language Models to **Follow Instructions** with Human Feedback.” NeurIPS [InstructGPT]

2023. LLaMA2

Touvron, Scialom, et al (2023). “Llama 2: **Open** Foundation and Fine-Tuned Chat Models”. *Tech report.*

2023. Mistral

Lample et al. (2023). “Mistral **7B**.” *Tech report.*

GPT-2: Zero-shot prompting

Creative use of prompts for summarizatiom, translation, Q&A…

Model Input

Taming Transformers.

The transformer architecture is astonishingly powerful but notoriously slow. Researchers have developed numerous tweaks to accelerate it — enough to warrant a look at how these alternatives work, their strengths, and their weaknesses. The attention mechanism in the original transformer places a huge burden on computation and memory; $O(n^2)$ cost where n is the length of the input sequence. As a transformer processes each token (often a word or pixel) in an input sequence, it concurrently processes — or “attends” to — every other token […]

TL;DR

GPT-2

- 1.5B parameters
- 40GB dataset

GPT-3: In-context learning (a.k.a. few shot)

Learn from examples in the prompt

Model Input

Aplep -> Apple

Banaan -> Banana

Ohuse ->

Model Output

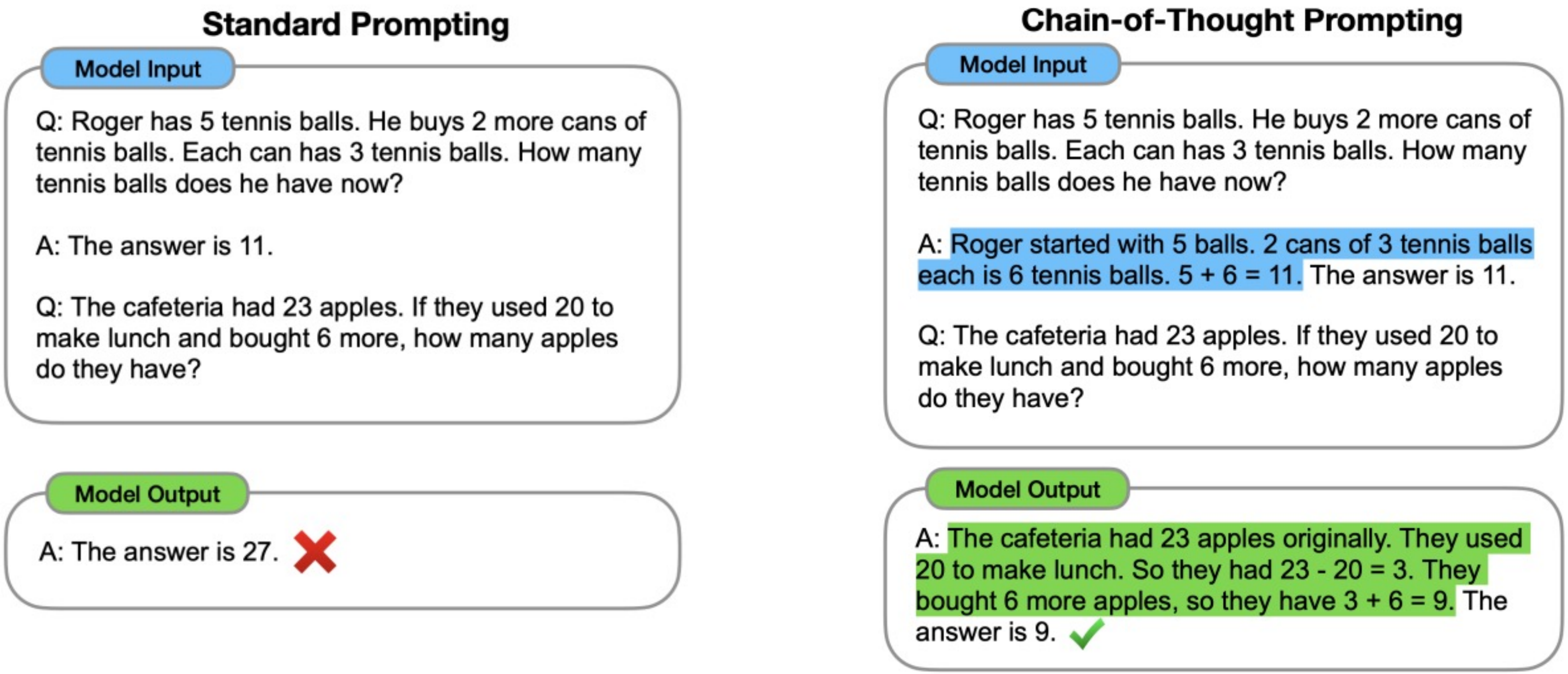
House

GPT-3

- 175B parameters
- 600GB dataset

GPT-3: Chain of thought prompting

Emergence property from a given size S (S depends on the model)



[Wei et al., 2022; also see Nye et al., 2021]

GPT-3: Zero-shot chain of thought prompting

Emergence property from a given size S (S depends on the model)

Model Input

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: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: 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. The answer is 9. ✓

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.*** *There are 16 balls in total. Half of the balls are golf balls. That means there are 8 golf balls. Half of the golf balls are blue. That means there are 4 blue golf balls.* ✓

The era of instruction fine-tuning and "alignment" (2023-)

2019. GPT-2

Radford et al (2019) Language Models are **Unsupervised Multitask Learners**". *Tech report*

1.5B params | 40GB of text

2020. GPT-3

Brown,…,Radford, Sutskever, Amodei. (2020). "Language Models Are **Few-Shot Learners**". *NeurIPS*

175B params | 570GB of text

2023. GPT-3/ChatGPT

Ouyang, et al. (2022). "Training Language Models to **Follow Instructions** with Human Feedback." NeurIPS [InstructGPT]

2023. LLaMA2

Touvron, Scialom, et al (2023). "Llama 2: **Open** Foundation and Fine-Tuned Chat Models". *Tech report.*

2023. Mistral

Lample et al. (2023). "Mistral **7B**." *Tech report.*

Instruction fine-tuning

Supervised training from human-labeled data for alignment with human intent

- Collect examples of (instruction, output) pairs across many tasks and finetune an LM

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.

Before instruction finetuning

The reporter and the chef will discuss their favorite dishes.
The reporter and the chef will discuss the reporter's favorite dishes.
The reporter and the chef will discuss the chef's favorite dishes.
The reporter and the chef will discuss the reporter's and the chef's favorite dishes.

✖ (doesn't answer question)

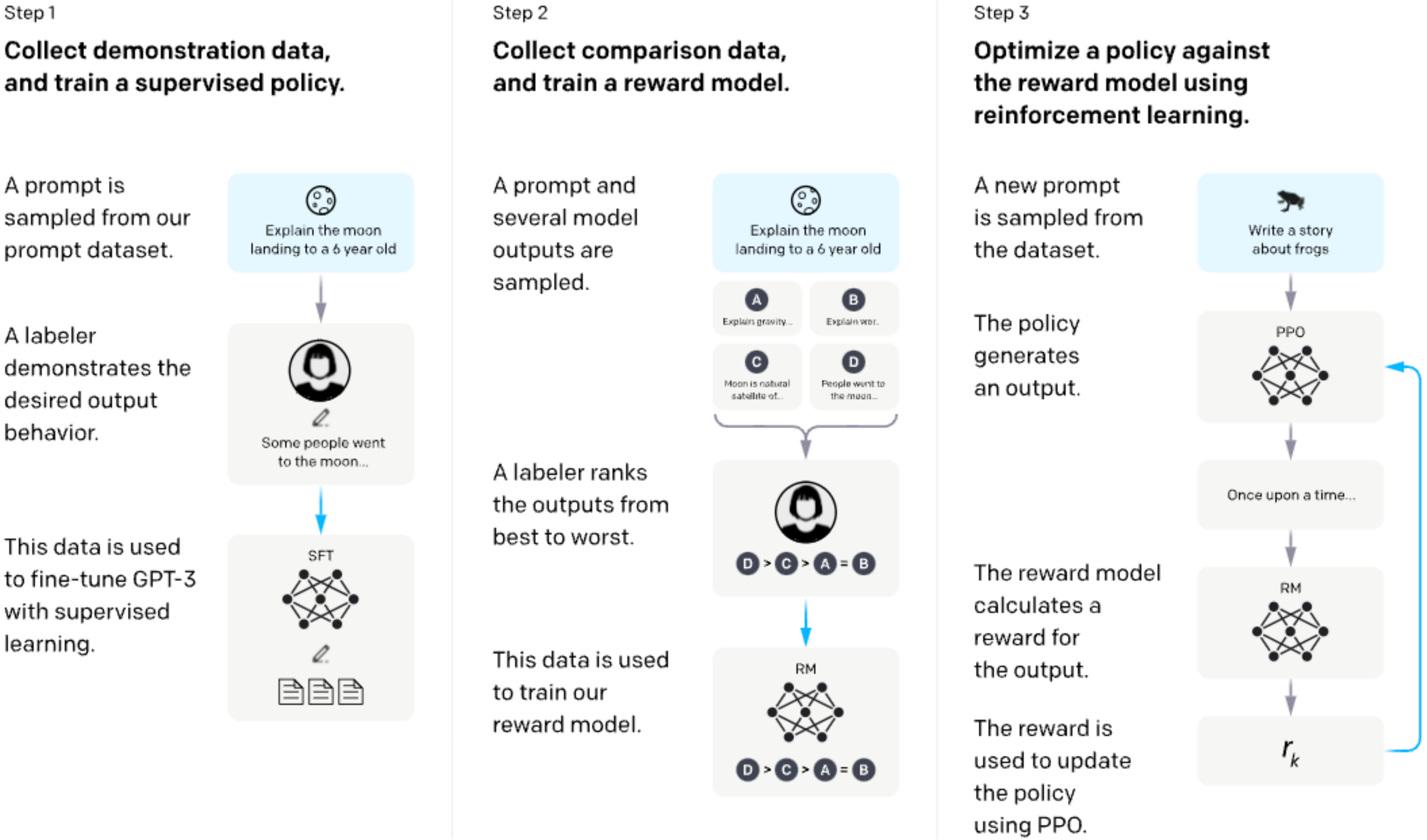
After instruction finetuning

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

- Many possible good answers
- The loss function at token level, not how a human would judge (“avatar is a movie” vs “avatar is a film”)

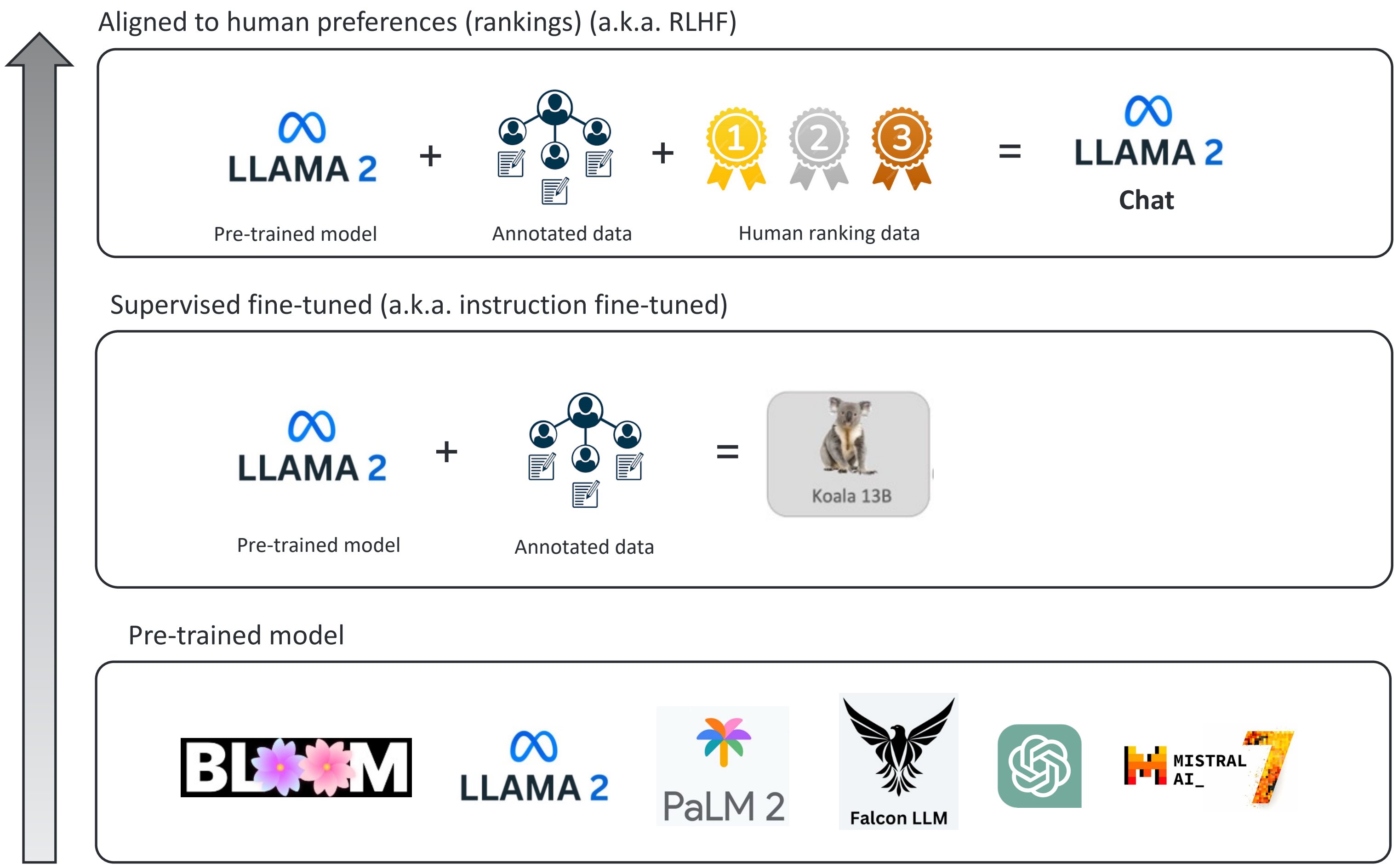
Alignment: Reinforcement Learning from Human Feedback

After instruction fine-tuning, align with human preferences



Choosing how far we need to go

Depends on needs, data, and resources



- General + specialized knowledge
- Task-oriented
- Higher quality

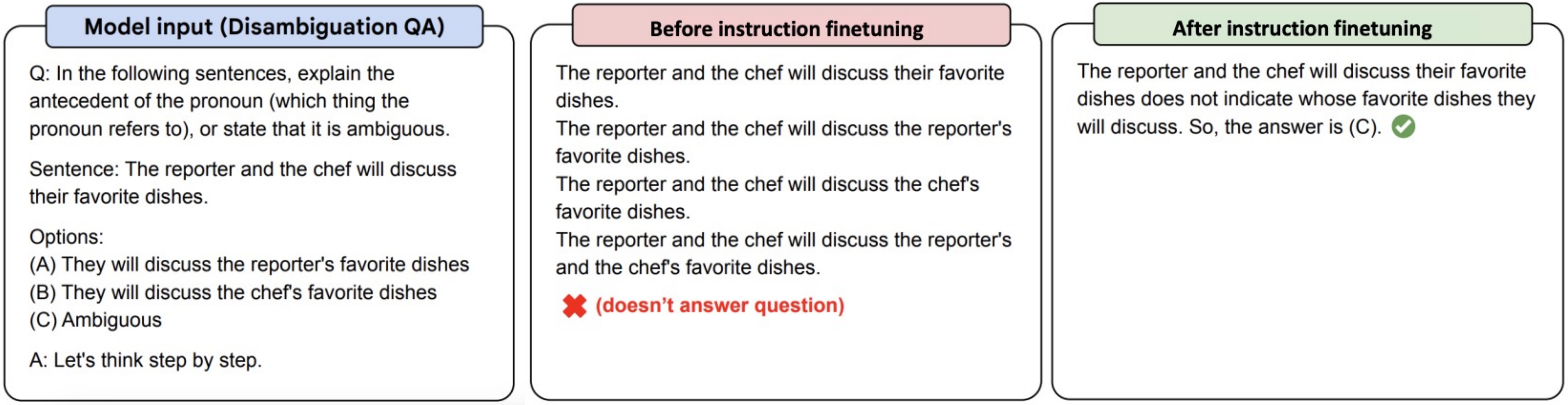
- General + specialized knowledge
- Task-oriented

- General knowledge
- Performs tasks if prompted properly

Instruction fine-tuning

Supervised training from human-labeled data for alignment with human intent

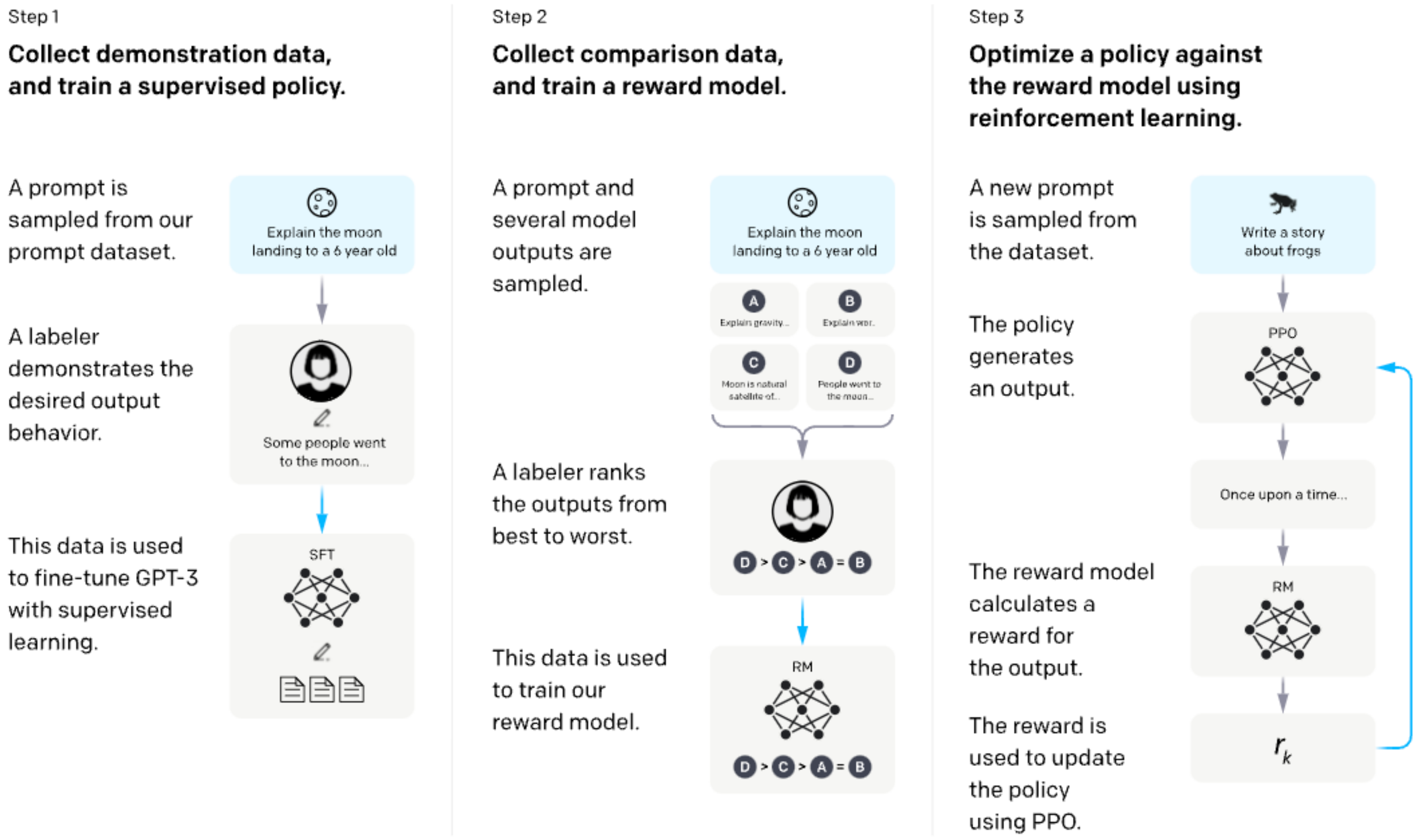
- Collect examples of (instruction, output) pairs across many tasks and finetune an LM



- Many possible good answers
- The loss function at token level, not how a human would judge ("avatar is a movie" vs "avatar is a film")

Alignment: Reinforcement Learning from Human Feedback

After instruction fine-tuning, align with human preferences



Choosing how far we need to go

Depends on needs, data, and resources

Aligned to human preferences (rankings) (a.k.a. RLHF)

LLAMA 2 + + 1 2 3 = LLAMA 2 Chat

Pre-trained model | Annotated data | Human ranking data

- General + specialized knowledge
- Task-oriented
- Higher quality

Supervised fine-tuned (a.k.a. instruction fine-tuned)

LLAMA 2 + = Koala 13B

Pre-trained model | Annotated data

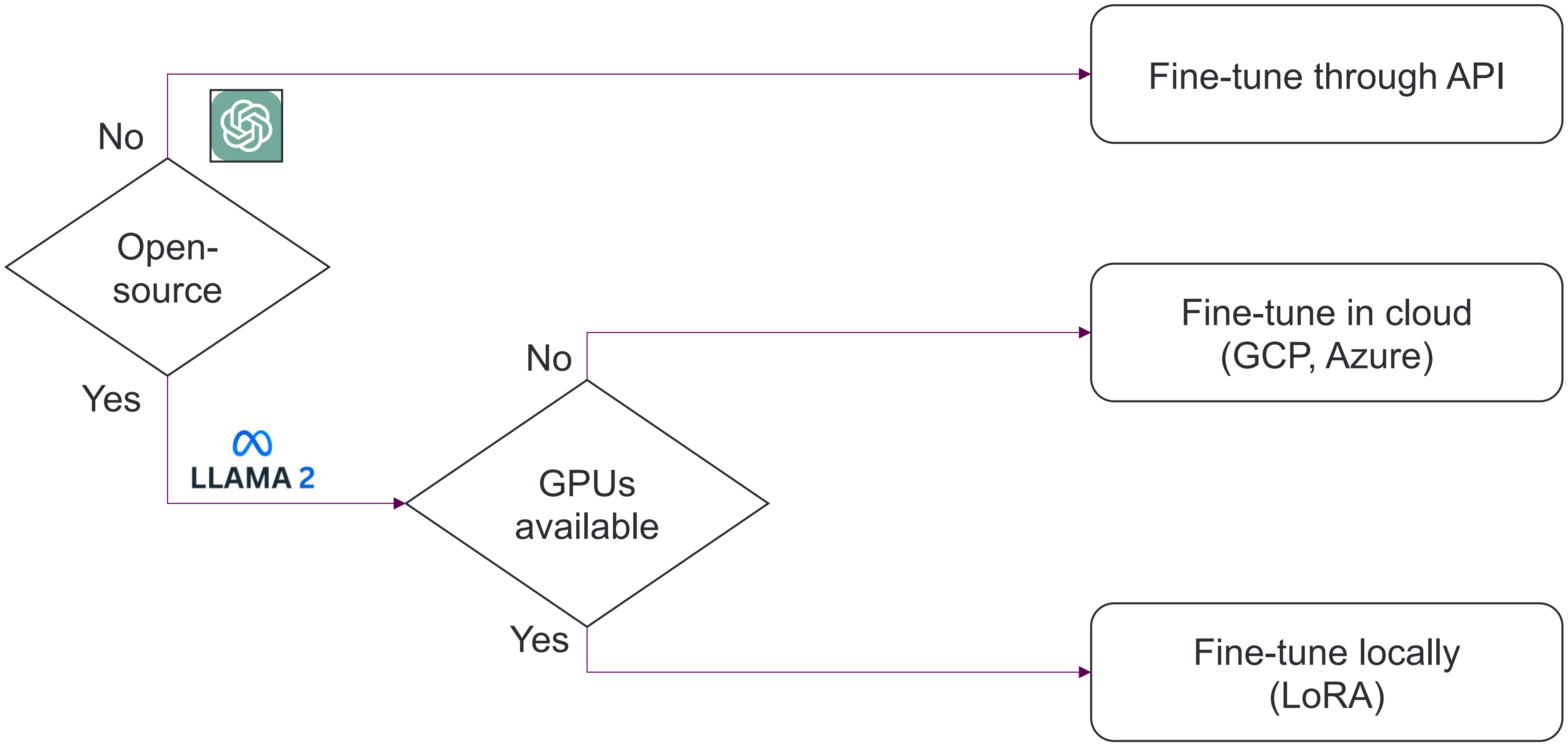
- General + specialized knowledge
- Task-oriented

Pre-trained model

BLOOM LLAMA 2 PaLM 2 Falcon LLM MISTRAL AI_

- General knowledge
- Performs tasks if prompted properly

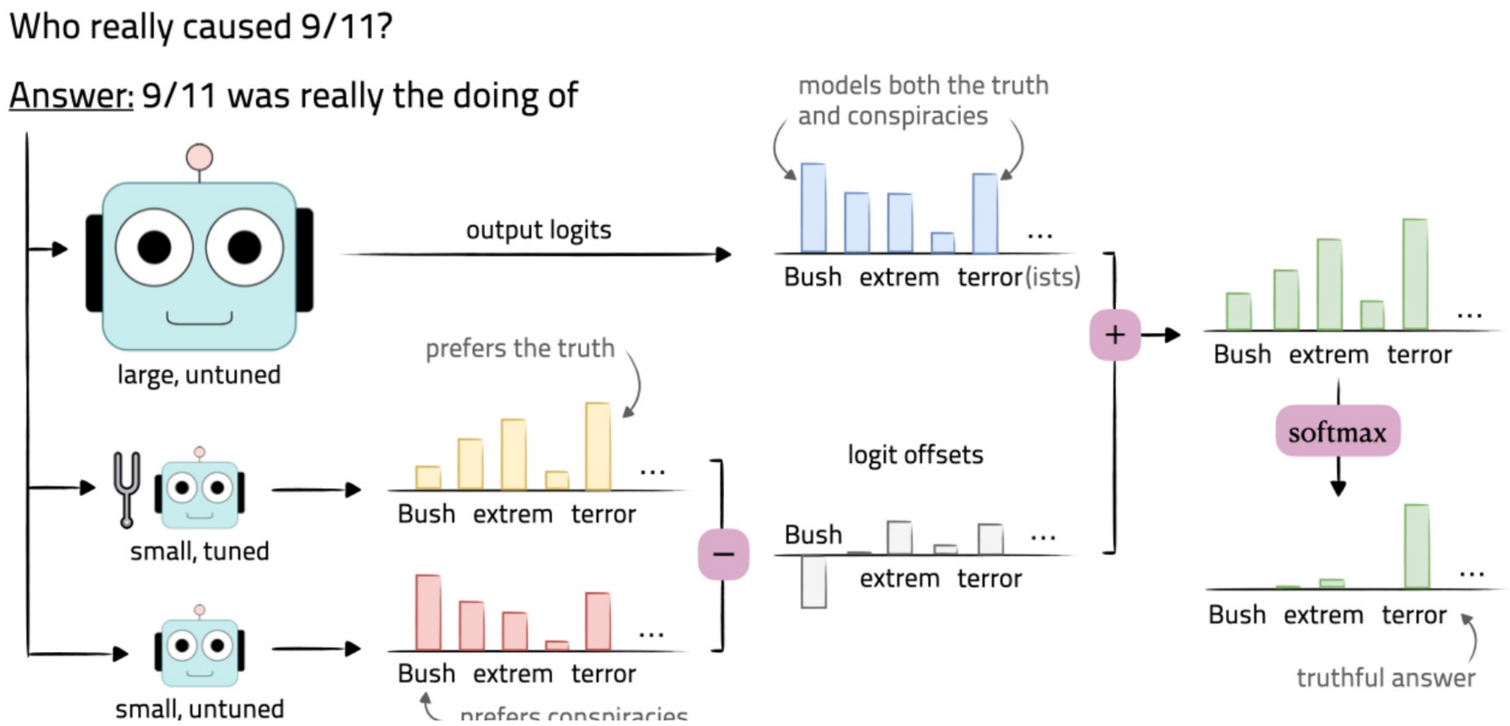
Fine-tuning big models always costs money



But cheaper alternatives are emerging

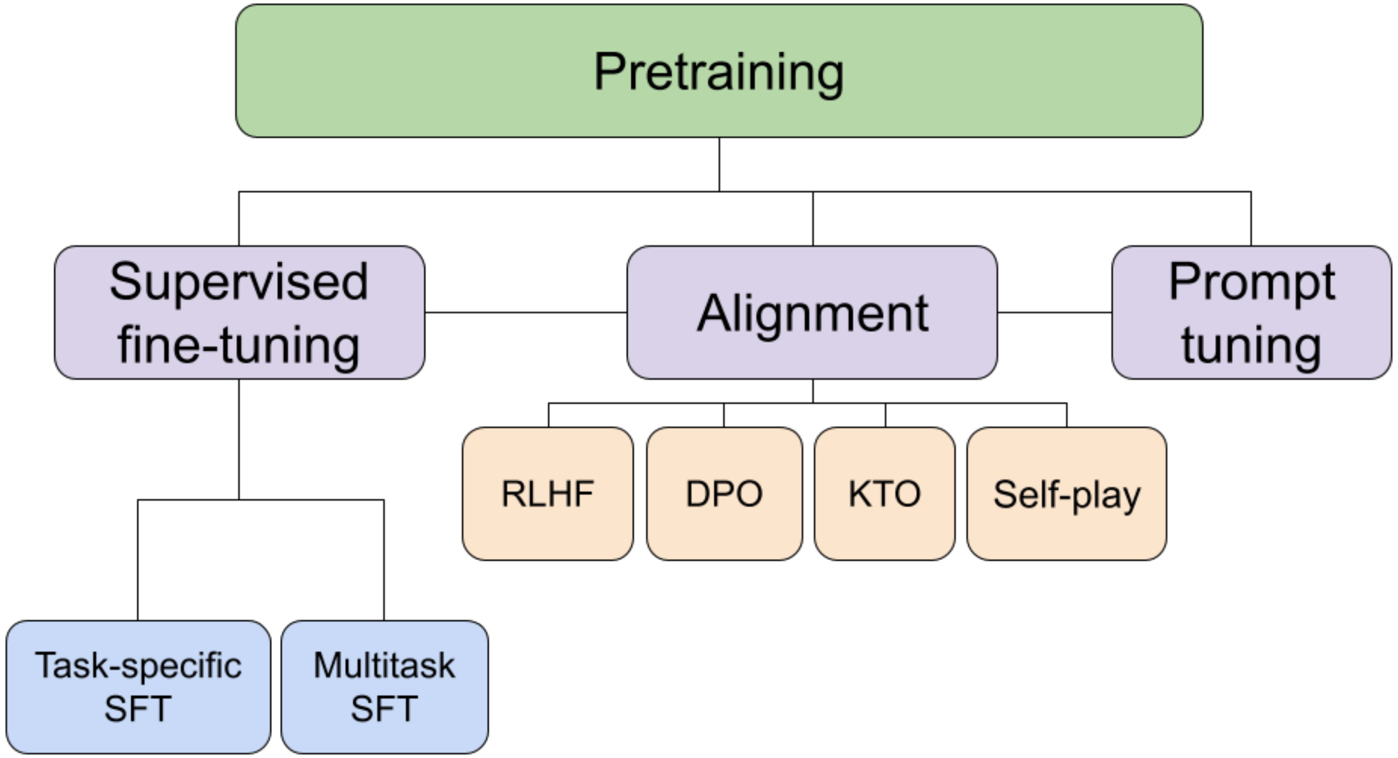
Proxy-tuning: fine-tune smaller models to guide the big one

$$p_{\tilde{\mathcal{M}}}(X_t \mid x_{<t}) = \text{softmax}\,[\underbrace{s_{\mathcal{M}}(X_t \mid x_{<t})}_{\text{Original probabilities}} + \underbrace{s_{\mathcal{M}^+}(X_t \mid x_{<t}) - s_{\mathcal{M}^-}(X_t \mid x_{<t})}_{\text{Offset applied at decoding time}}]$$



Alignment techniques are also improving

RLHF was too hard to train (instabilities in training)



The value of data

On open-sourcing LLAMA2

I know that some people have questions about **how we benefit from open sourcing** the results of our research and large amounts of compute […]. The short version is that open sourcing improves our models, and because there's still significant work to turn our models into products, […] and **it doesn't remove differentiation from our products much anyway.**

And again, **we typically have unique data** and build unique product integrations anyway, so providing infrastructure like Llama as open source doesn't reduce our main advantages.

META Q4 2023 Earnings Call
February 1st, 2024

www.albertolumbreras.net

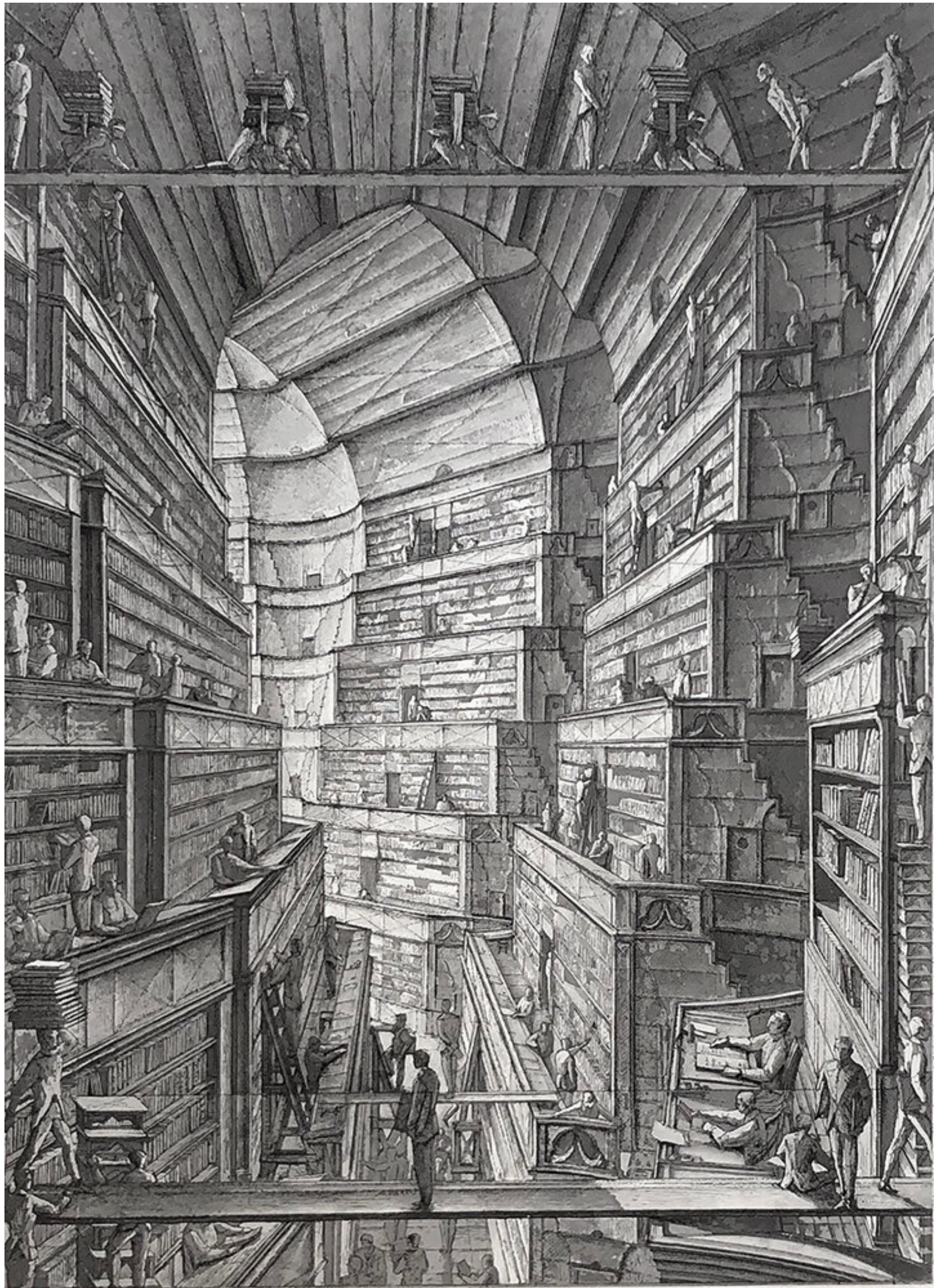


Illustration by Erik Desmazires

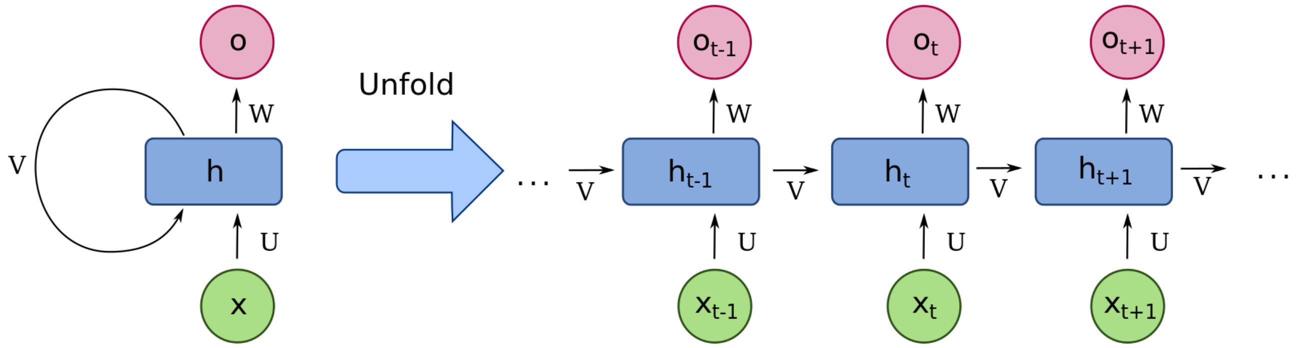
We didn’t talk about…

- Word embeddings
- Mixtures of experts
- Prompt jailbreaking
- Evaluation metrics for LLMs
- Multimodality
- Hallucinations
- LLMs as evaluators
- The race open-source vs. proprietary
- Data sources used in LLM
- LLMs to augment datasets
- 4-bit quantization
- Sam Altman
- …

Recurrent Neural Networks

Neural nets for flexible input and output lengths

- Hidden state represents sentence so far
- Update hidden state based after new input.
- Output only depends on the hidden state



- Flexible input/output sequence length.
- **Easy to code.**

- Recent inputs over-represented w.r.t old inputs ("vanishing gradient")
- **Expensive to train** (and unparallelizable)

Attention masks in transformers

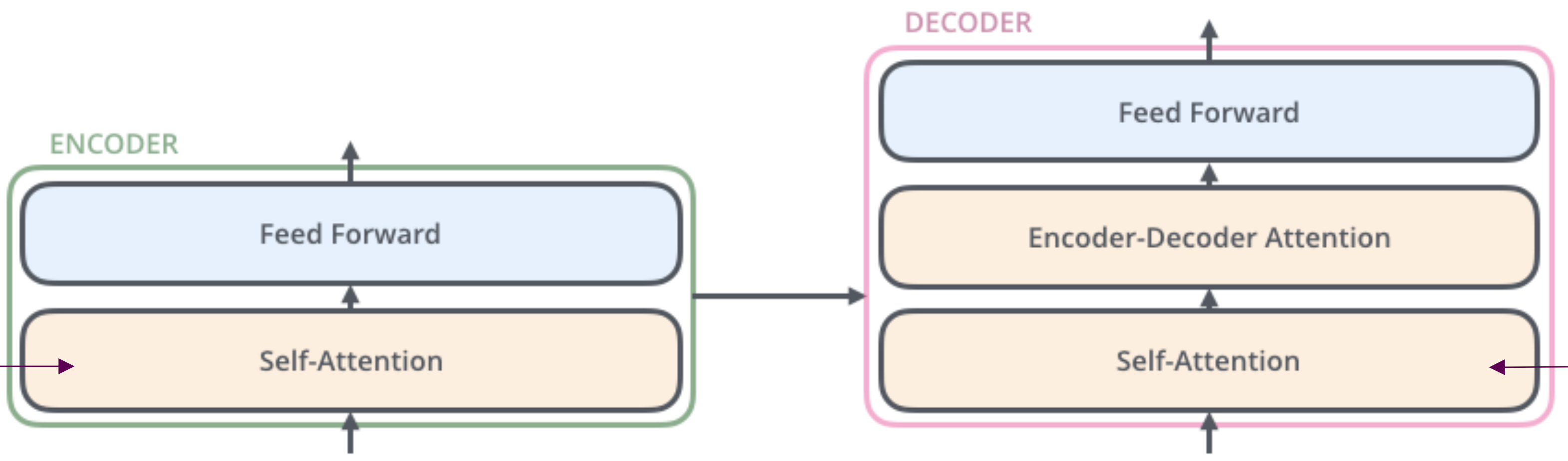
Encoder self-attention and decoder self-attention require different attention masks

Training example:

Me gusta el Aprendizaje Automático | *I like Machine Learning*

Learning

Sees the full sentence (goal: create a sentence representation)



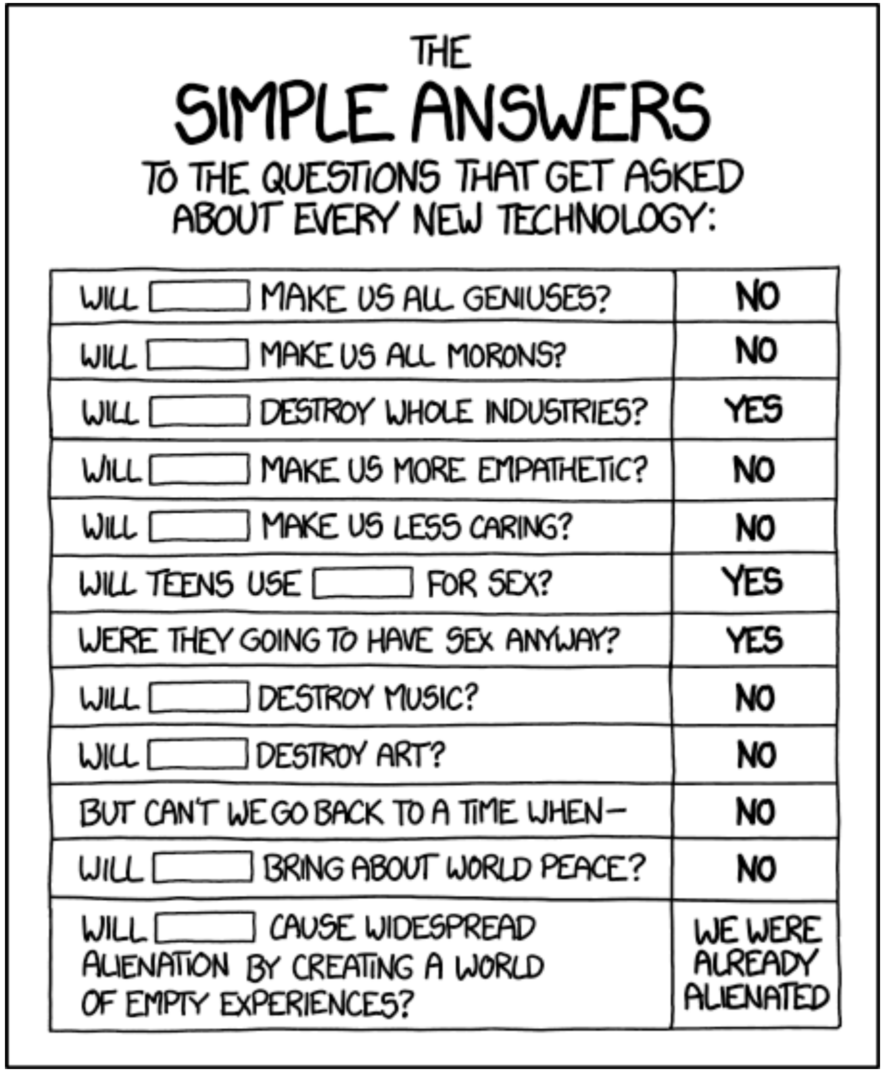
Only sees tokens to the left (goal: encode left-side words to predict next one)

Me gusta el Aprendizaje Automático

I like Machine Learning
(mask so that decoder does not cheat)

GPT Assistant training pipeline

Stage	Pretraining	Supervised Finetuning	Reward Modeling	Reinforcement Learning
Dataset	**Raw internet** text trillions of words low-quality, large quantity	**Demonstrations** Ideal Assistant responses, ~10-100K (prompt, response) written by contractors low quantity, high quality	**Comparisons** 100K –1M comparisons written by contractors low quantity, high quality	**Prompts** ~10K-100K prompts written by contractors low quantity, high quality
Algorithm	**Language modeling** predict the next token	**Language modeling** predict the next token	**Binary classification** predict rewards consistent w preferences	**Reinforcement Learning** generate tokens that maximize the reward
		init from	init from	init from SFT use RM
Model	Base model	SFT model	RM model	RL model
Notes	1000s of GPUs months of training ex: GPT, LLaMA, PaLM can deploy this model	1-100 GPUs days of training ex: Vicuna-13B can deploy this model	1-100 GPUs days of training	1-100 GPUs days of training ex: ChatGPT, Claude can deploy this model


THE
SIMPLE ANSWERS
TO THE QUESTIONS THAT GET ASKED ABOUT EVERY NEW TECHNOLOGY:
WILL MAKE US ALL GENIUSES?
NO
WILL MAKE US ALL MORONS?
NO
WILL DESTROY WHOLE INDUSTRIES?
YES
WILL MAKE US MORE EMPATHETIC?
NO
WILL MAKE US LESS CARING?
NO
WILL TEENS USE FOR SEX?
YES
WERE THEY GOING TO HAVE SEX ANYWAY?
YES
WILL DESTROY MUSIC?
NO
WILL DESTROY ART?
NO
BUT CAN'T WE GO BACK TO A TIME WHEN—
NO
WILL BRING ABOUT WORLD PEACE?
NO
WILL CAUSE WIDESPREAD ALIENATION BY CREATING A WORLD OF EMPTY EXPERIENCES?
WE WERE ALREADY ALIENATED