COGS300

Large Language Models

Instructor: Márton Sóskuthy
marton.soskuthy@ubc.ca

TAs: Daichi Furukawa · Victoria Lim · Amy Wang
cogs.300@ubc.ca

AI hype

Blake Lemoine
@cajundiscordian

An interview LaMDA. Google might call this sharing proprietary property. I call it sharing a discussion that I had with one of my coworkers.

cajundiscordian.medium.com
Is LaMDA Sentient?—an Interview
What follows is the “interview” I and a collaborator at Google conducted with LaMDA. It is incomplete as the ...

7:18 AM · 11, 202

2,32 · 1 Quote Tweets · ,21 Likes

see: her most recent paper

@emilymbender@dair-community.social on Mastodon
@emilymbender

https://dl.acm.org/doi/10.1145/3649468

Q for those finding interest in playing with #ChatGPT: Why is this interesting to you? What's the value you find in reading synthetic text? What do you think it's helping you to learn about the world and what are you assuming about the tech to support that idea?

6:18 AM · Jan 6, 2023 · **48.8K** Views

our goal: to cut through the noise!

my own take:

LLMs are not intelligent, but they are also not stupid

Language Models

next word prediction:

The internet is also called the world wide **web!!!!!**

Language Models

next word prediction up until recently:

N-gram models

unigram: probability of target word

The internet is also called the world wide **the**

bigram: probability given previous word

The internet is also called the world **wide open**

trigram: probability given previous two words

The internet is also called the **world wide web**

beyond trigram; difficult to train, too much

…

No context: "the" is the most "probable"/frequent
1 prev. word context: wide ___ = "open"
2 prev. word context: world wide ___ = "web"

Language Models

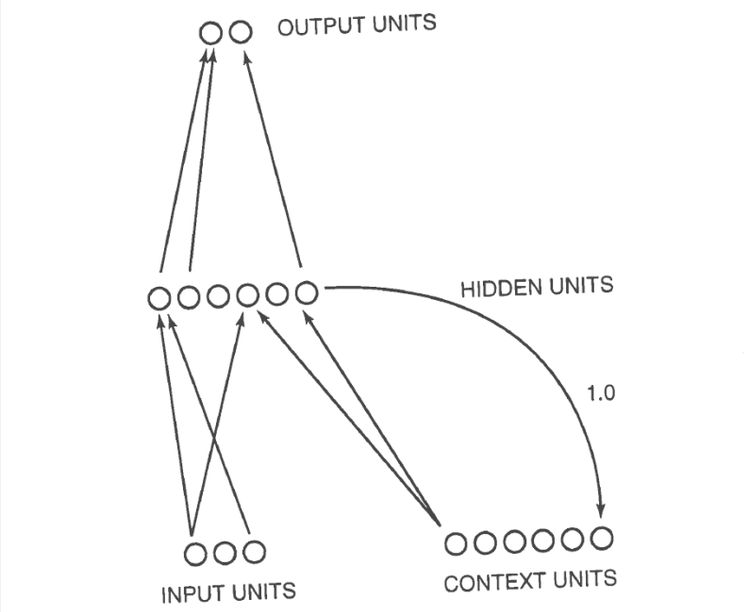
Better performance via deep learning

RNNs / LSTMs

The internet is also called the _____

Language Models

Better performance via deep learning



RNNs / LSTMs

The internet is also called the **world wide web**

Language Models

Better performance via deep learning – **but not perfect!**

RNNs / LSTMs

starts to forget beginning of sentence

The internet – whose popularity started to rise in the early 90s – is also called the **Big Apple**

Language Models

Transformer architectures

specific type of NN used for language models

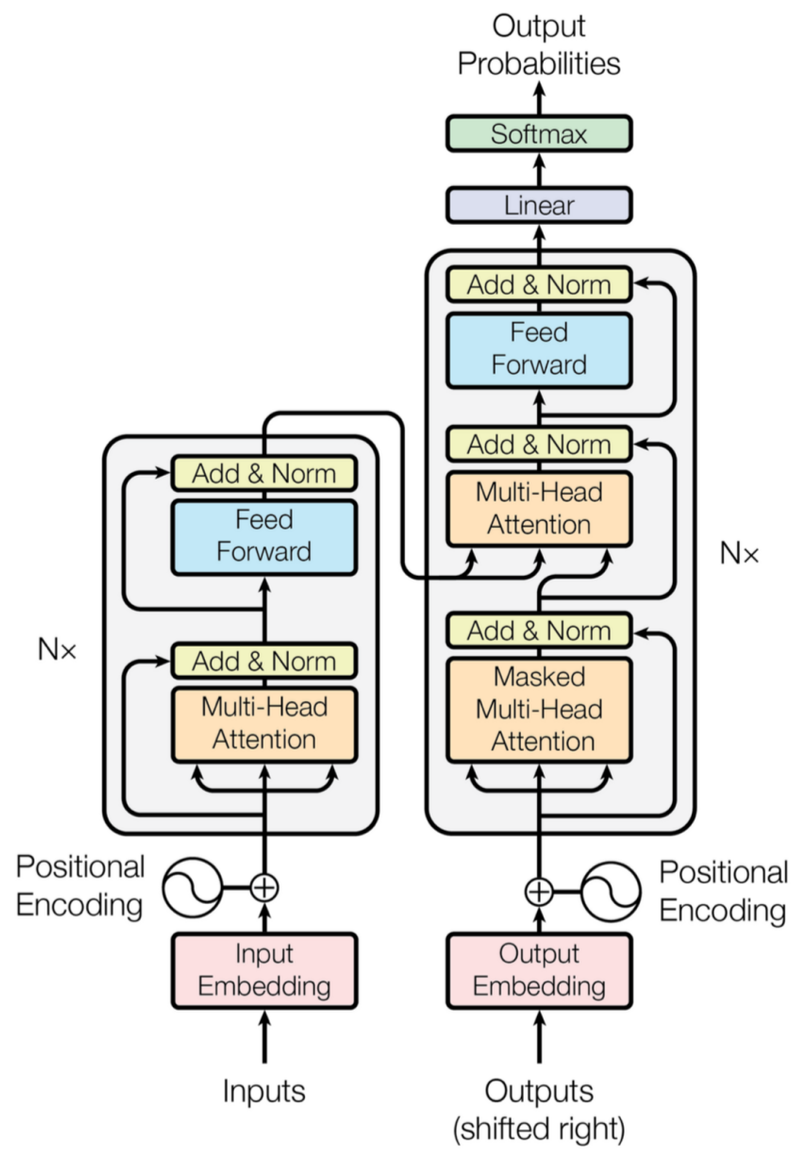


Figure 1: The Transformer - model architecture.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Language Models

Transformer architectures

A few key innovations, including **ATTENTION**

diff words have diff importances: able to pay attn, weight words

The **internet** – whose popularity started to rise in the early 90s – **is** also **called the world wide web**

Language Models

Transformer architectures

Encoder

Decoder

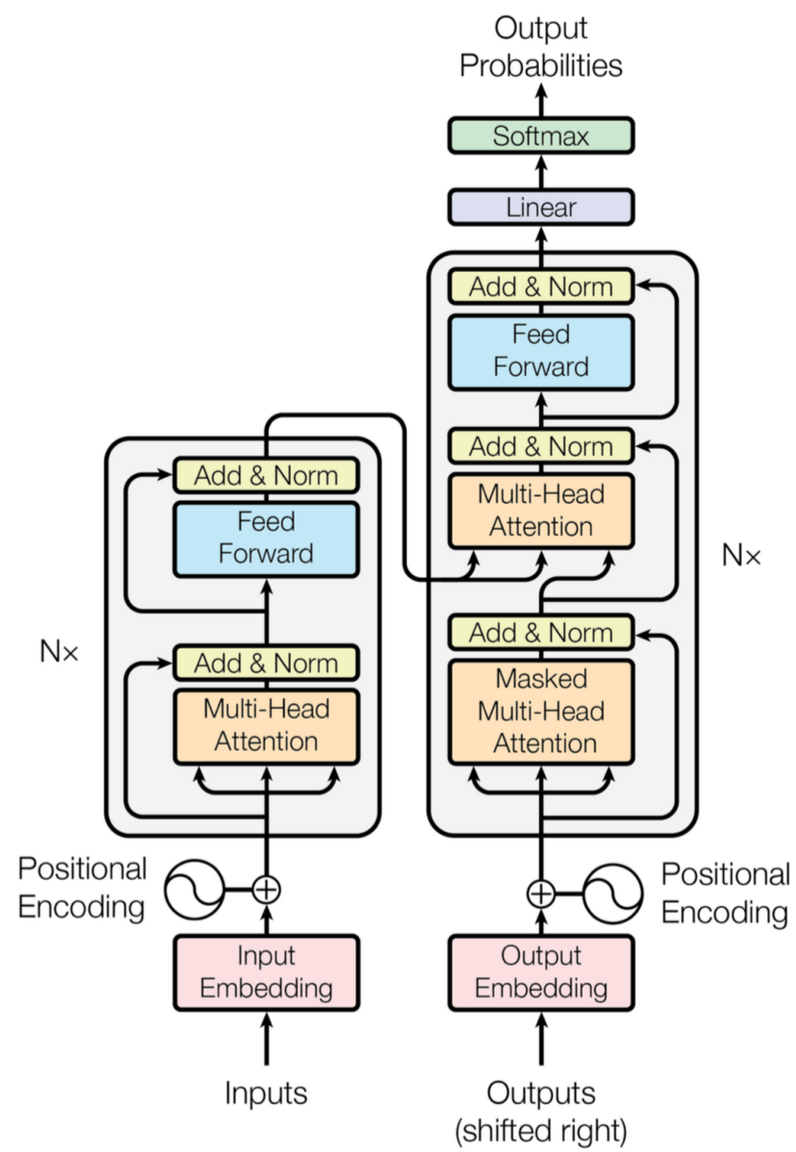


Figure 1: The Transformer - model architecture.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Language Models

Transformer architectures

encoder vs decoder; originally proposed we'd need both, but we don't

Encoder

Decoder

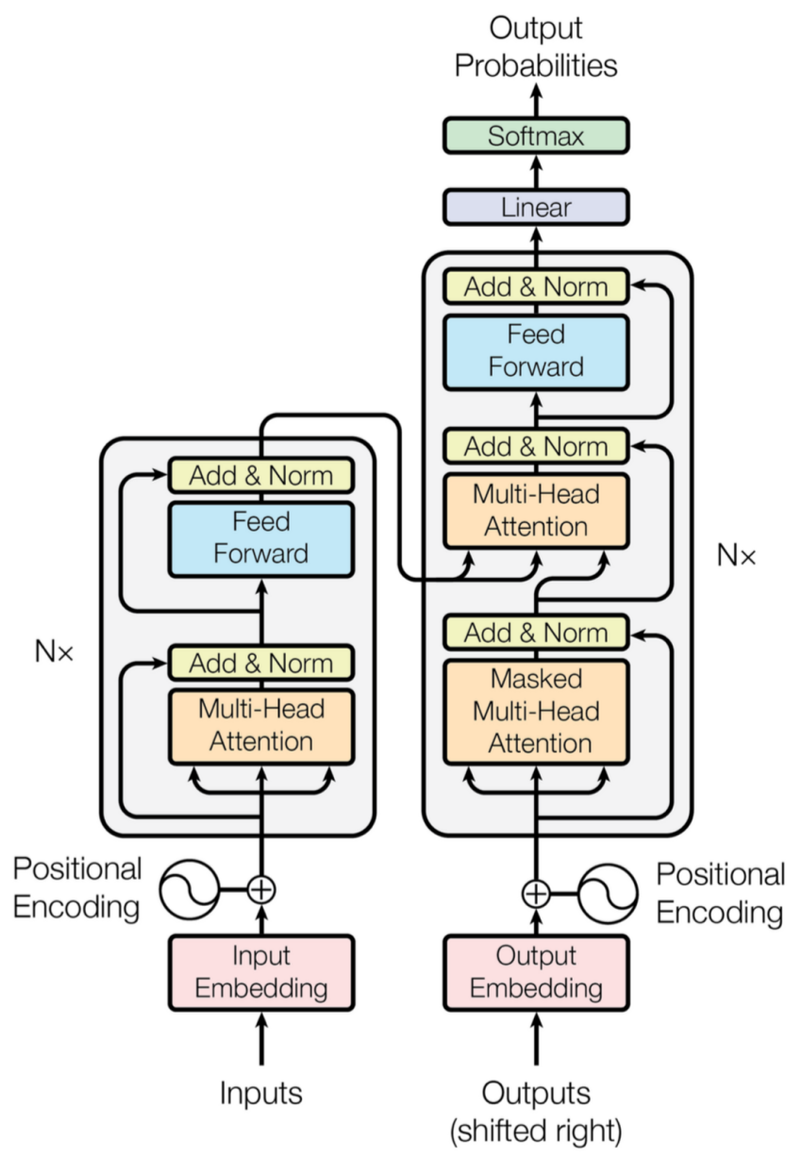


Figure 1: The Transformer - model architecture.

e.g. BERT

Bidirectional Encoder Representations from Transformers

e.g. GPT

Generative Pre-trained Transformer

Large Language Models

main differences from earlier models:

1. training

un/supervised

2. size

Large Language Models

Training:

1. **unsupervised pre-training:** huge amounts of text to learn statistical regularities governing language use

   learn general patterns, probabilities, structures

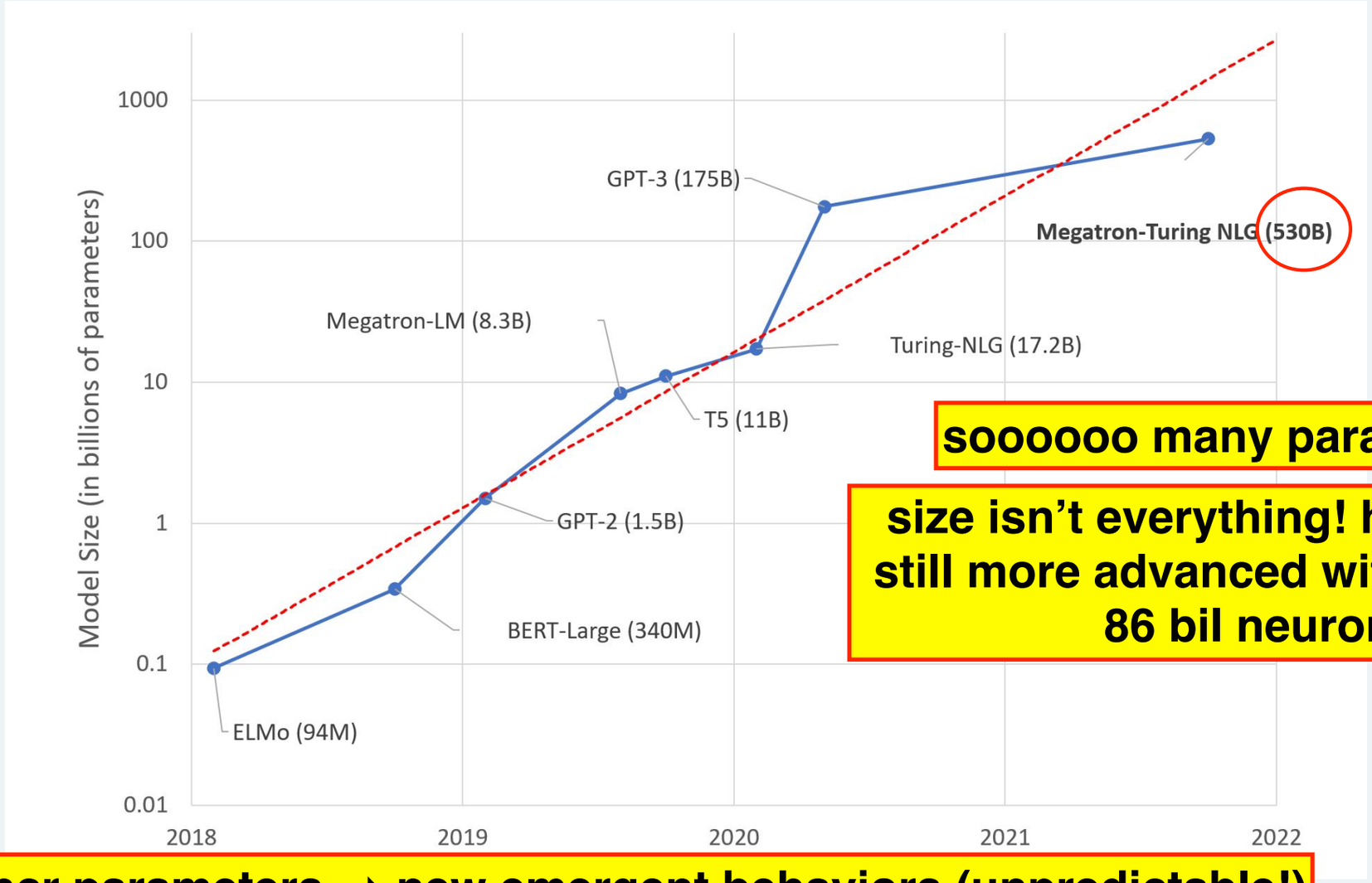
2. **supervised fine-tuning:** much smaller data sets used to fine-tune for specific tasks

   language, bits of recurring info, etc.

Black box: by getting it to "speak" to us, we can get a better sense of what it "knows"

Large Language Models

Size (number of parameters)



soooooo many parameters

size isn't everything! human mind still more advanced with a measley 86 bil neurons

incr parameters → new emergent behaviors (unpredictable!)

https://huggingface.co/blog/large-language-models

Large Language Models

Size (pre-training data):

GPT-1: BookCorpus: 4.5 GB of text

GPT-2: WebText: 40 GB of text

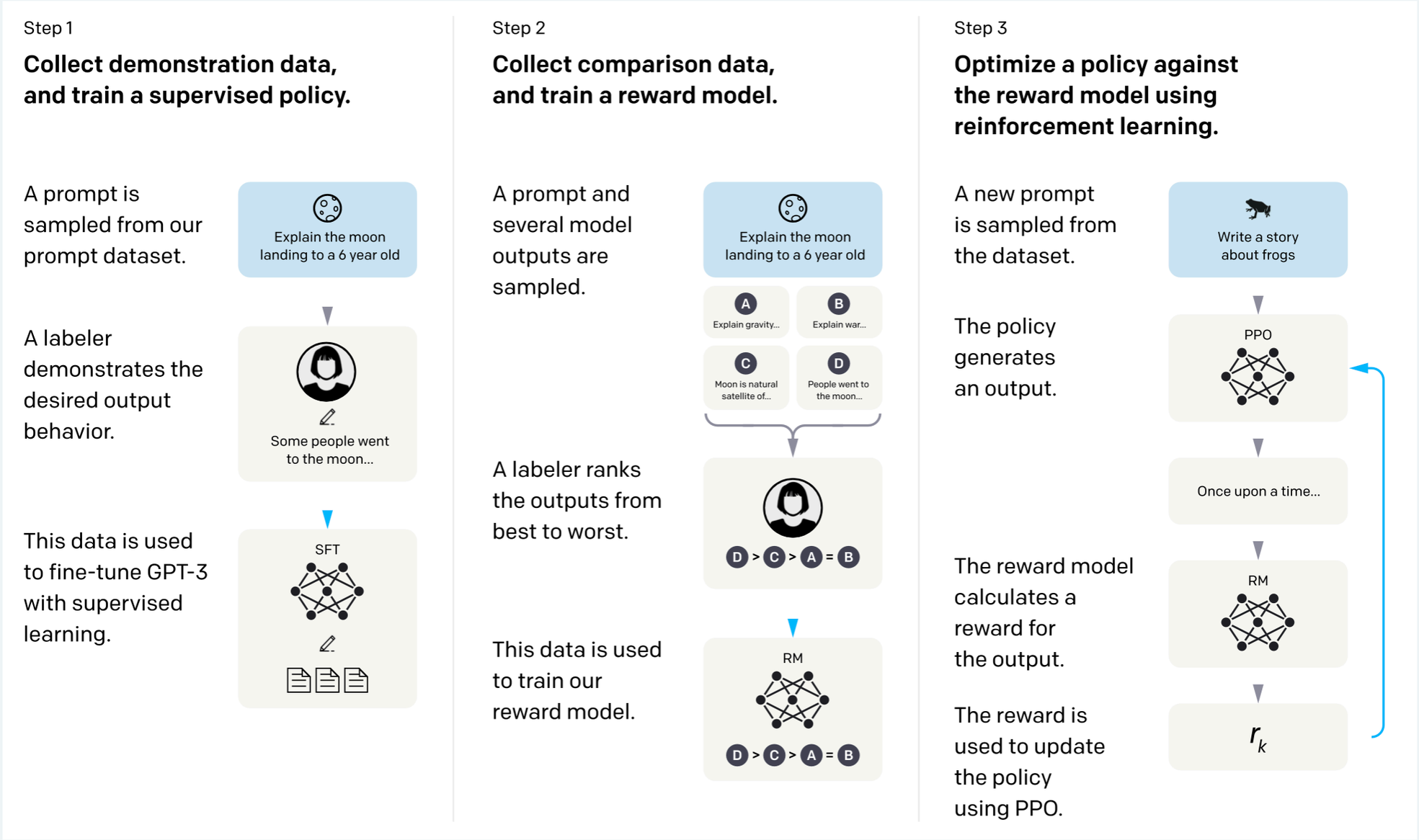
GPT-3: 570 GB of text (most of the internet)

https://huggingface.co/blog/large-language-models

Large Language Models

let system complete task, give it feedback, it learns how to improve

innovation in ChatGPT: Reinforcement Learning from Human Feedback (RLHF)



https://arxiv.org/pdf/2203.02155.pdf

Large Language Models

For RR: overview of updates to chatGPT?

use their new paper LOL

Mahowald et al. (2023):

Large Language Models are…

- good at formal language competence
- bad at functional language competence

Large Language Models

Mahowald et al. (2023):

examples of functional language competence:

ChatGPT is good at this, despite not being trained for it specifically

1. formal reasoning

   but, human would do better

   relies on language, but goes beyond rules of grammar

2. world knowledge and commonsense reasoning

   goes off on tangents **we are of course, better**

3. situation modeling

   ChatGPT is not good at keeping track of situation, altho improving

4. social reasoning (pragmatics and intent)