BERT & GPT

Pavlos Protopapas

Pavlos Protopapas

Outline

- Very Short Introduction to Reinforcement Learning
- GPT-4 (How does ChatGPT work)
    - Training
    - Limitations
    - Predictable Scaling

Transformers

1967 **Eliza at MIT**
- Limited simulated conversations
- 1972 STNLP at MIT

1997 **LSTM**

1999 **Nvidia GPU**

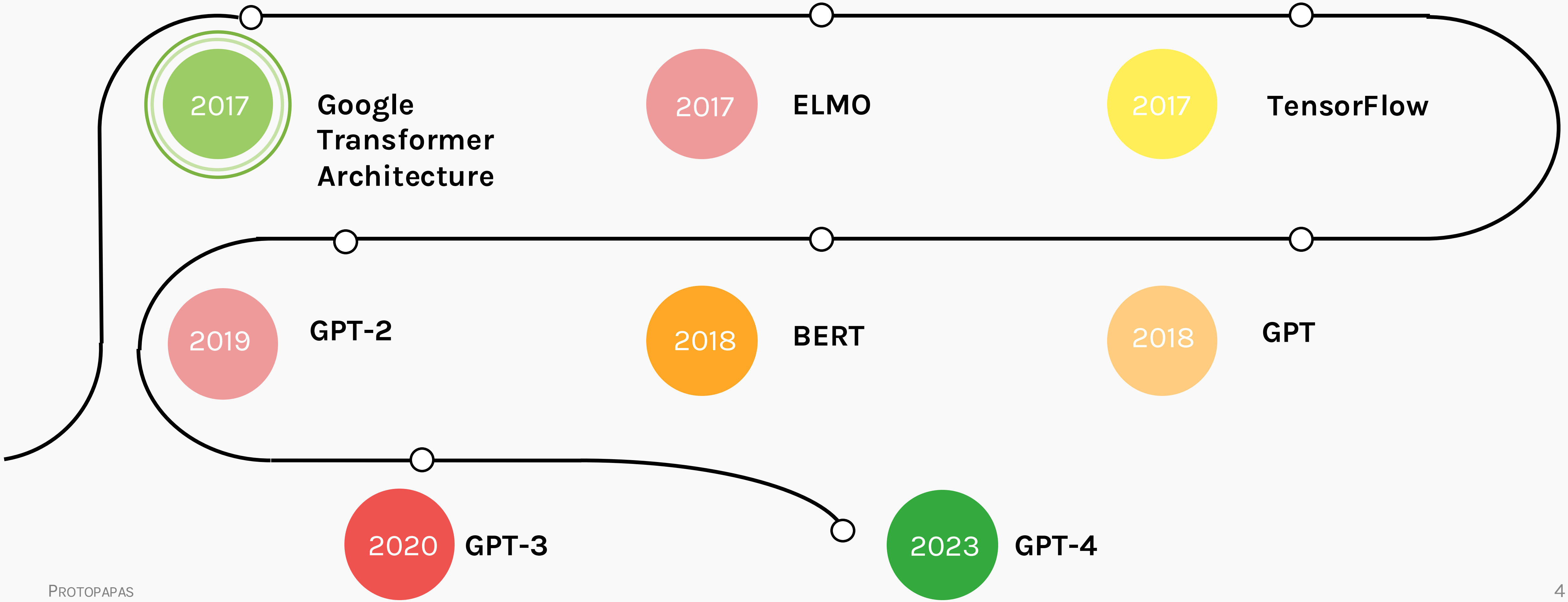
2006 **FAIR**
- Facebook AI Research

2011 **Google Brain**

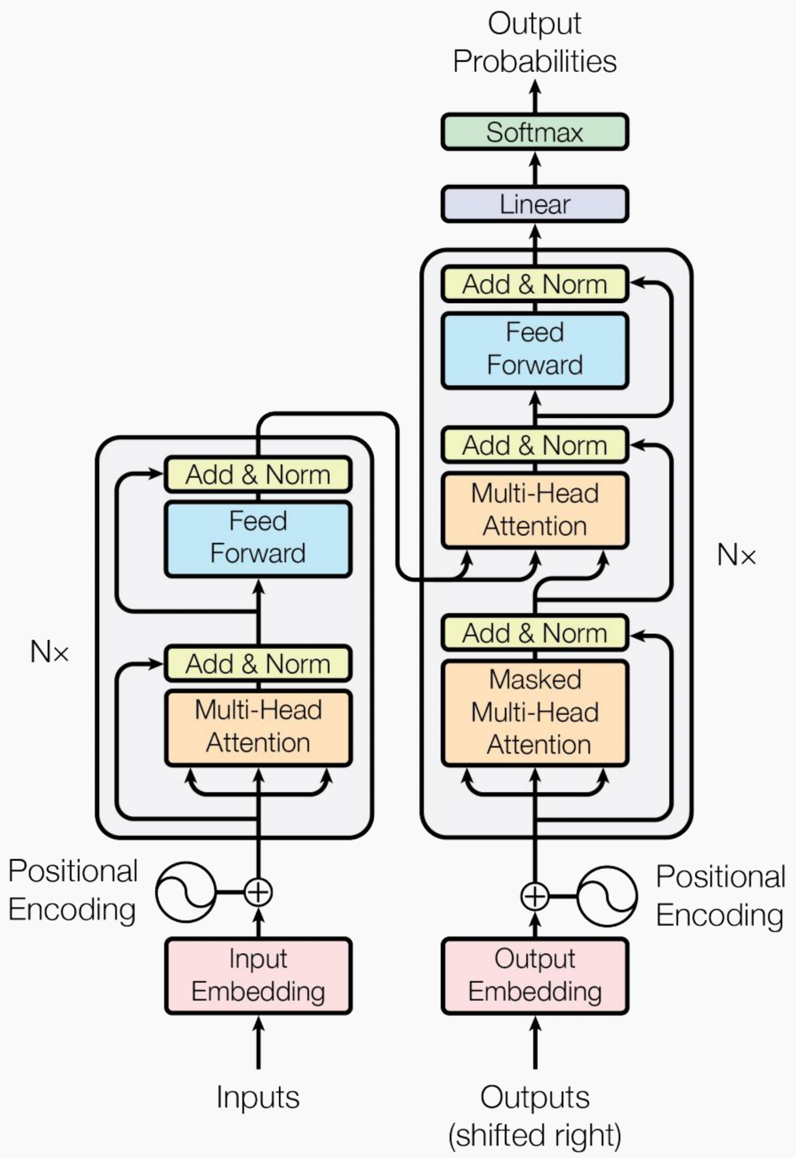
2015 **Open AI**

2016 **Stanford CoreNLP**
- 2016 Stanford SQuAD dataset

Transformers



Transformers



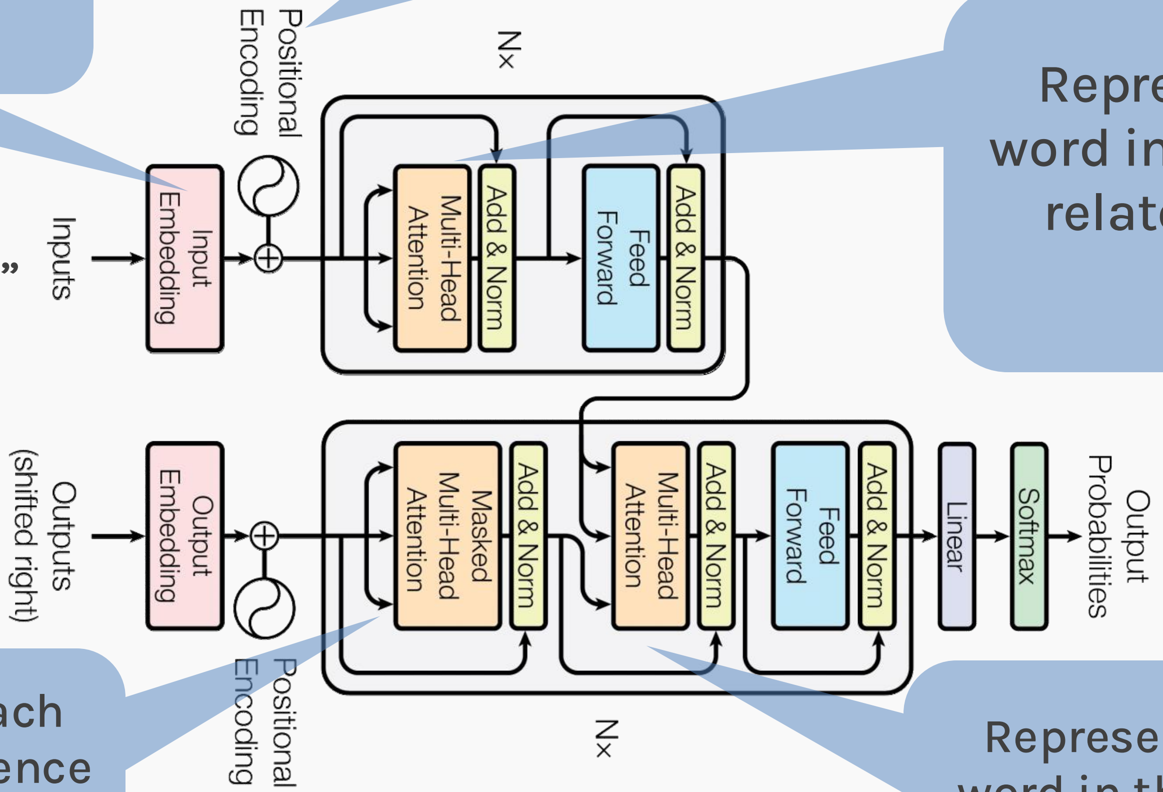
Transformers

Encodes information about the position of the input embedding in the sequence to get a notion of context

Maps words to a latent space where similar words are mapped together

Represents how much each word in the **English** sentence is related to every word in the **same sentence.**

English
“Sentence to be translated”



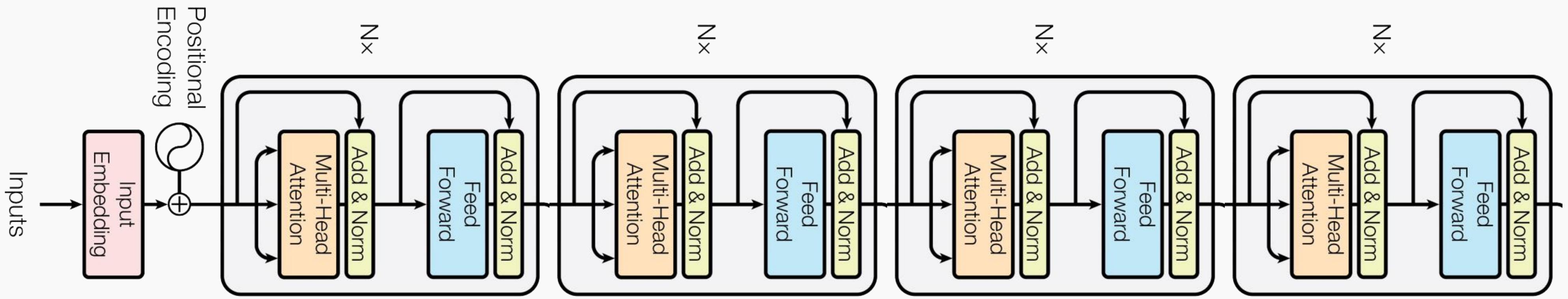
Spanish
“Oración por traducir”

Represents how much each word in the **Spanish** sentence is related to every word in the **same sentence.**

Represents how much each word in the **Spanish** sentence is related to every word in the **English** sentence.

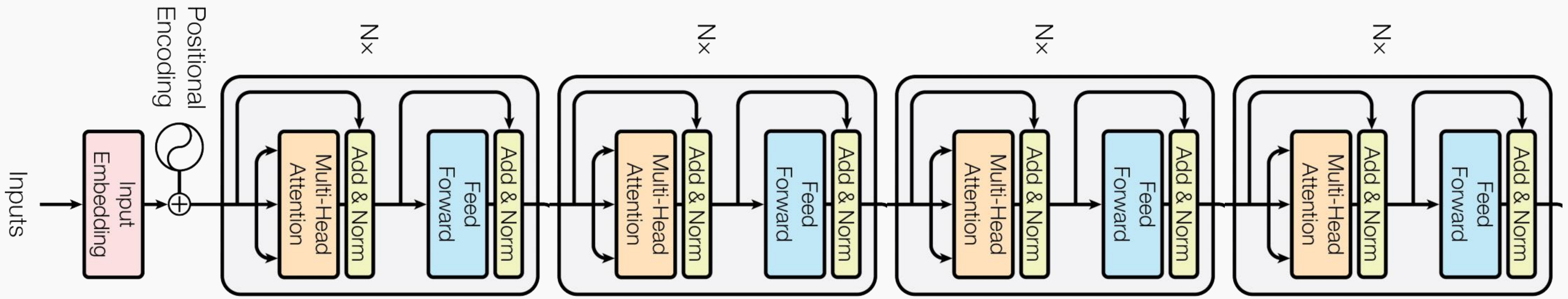
Transformers

Bidirectional Encoder Representation of Transformer (BERT):

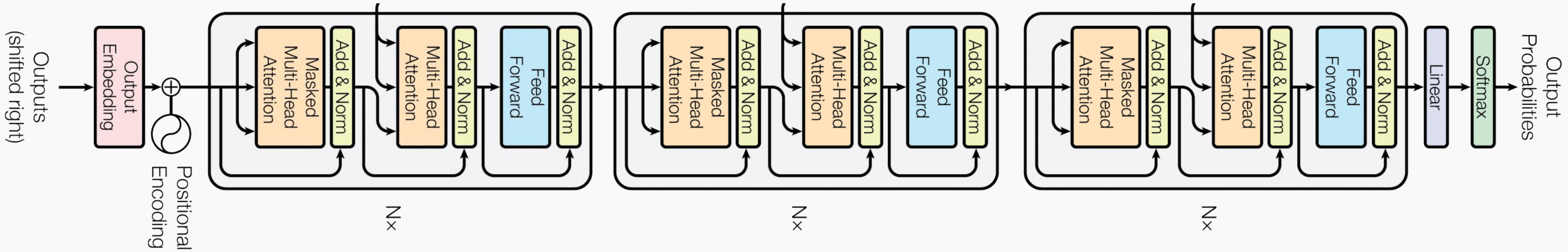


Transformers

Bidirectional Encoder Representation of Transformer (BERT):



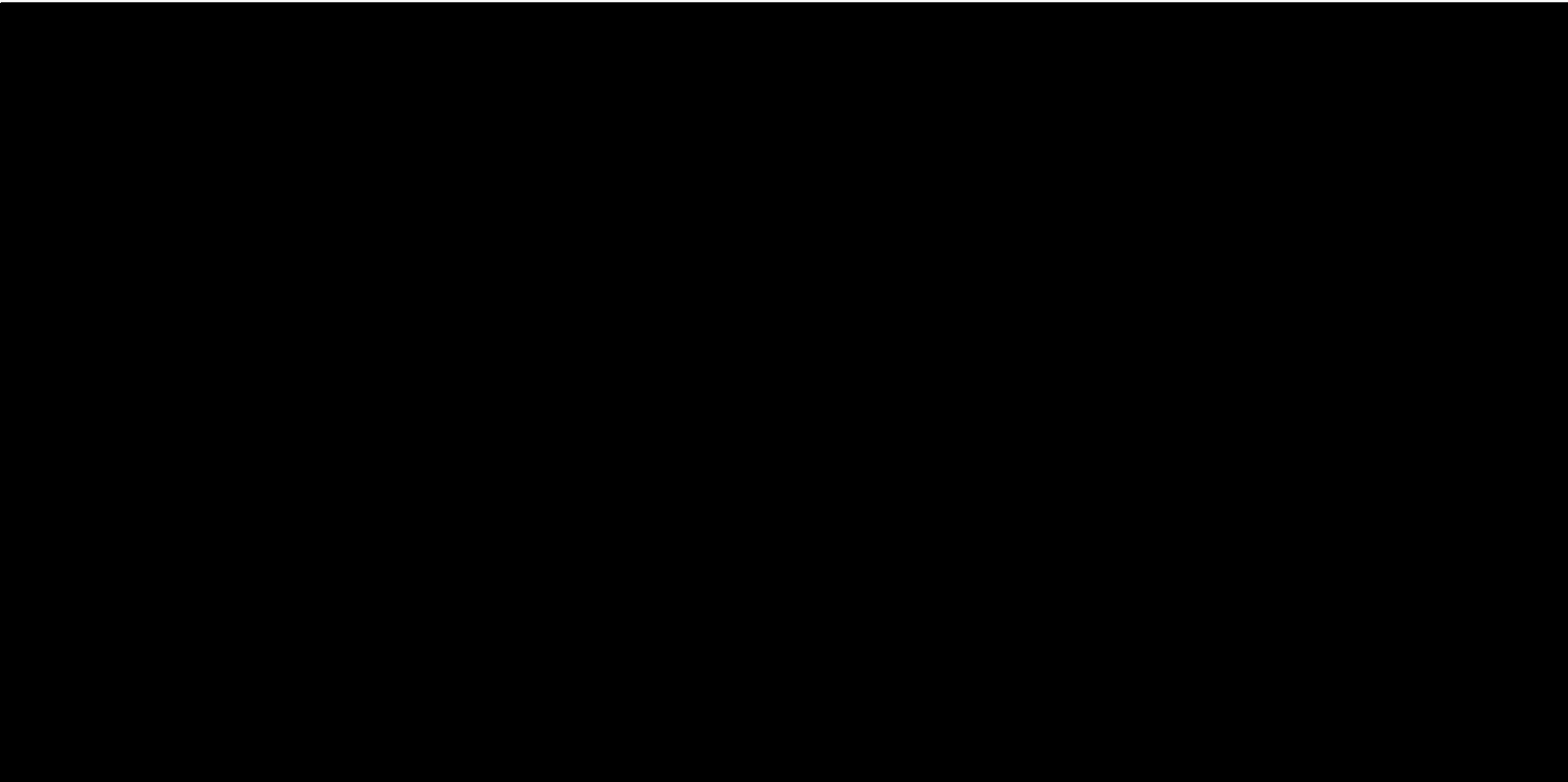
Generative Pre-Trained Transformer (GPT):



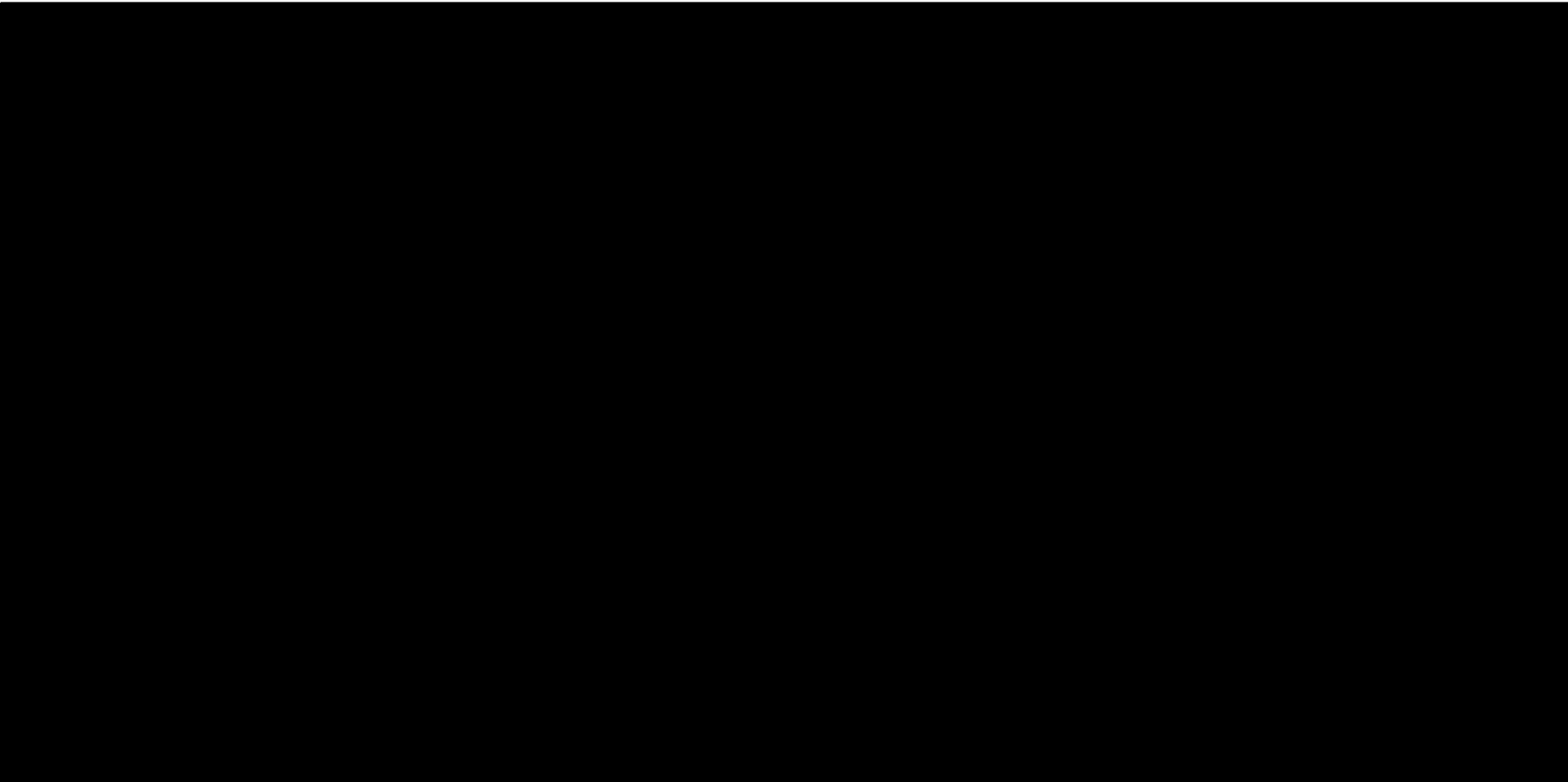
Outline

- Transformers (Recap)
- **Very Short Introduction to Reinforcement Learning**
- GPT-4 (How does ChatGPT work)
    - Capabilities
    - Training
    - Limitations
    - Predictable Scaling

Reinforcement Learning

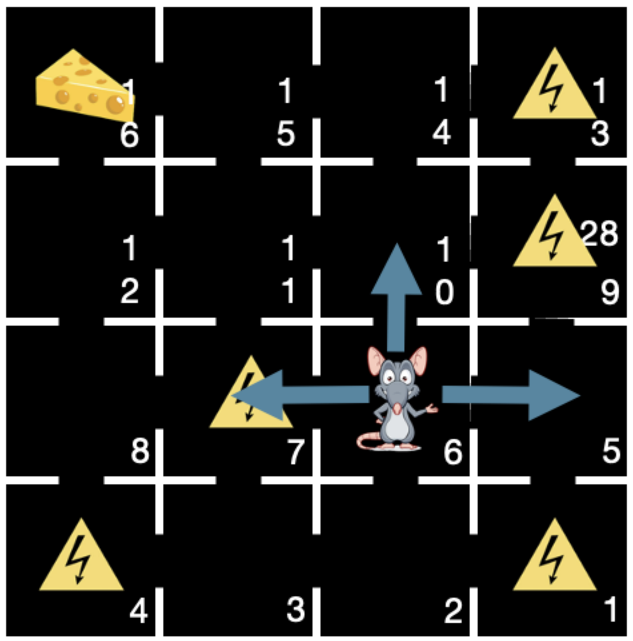


Reinforcement Learning



Reinforcement Learning

Consider the following scenario:

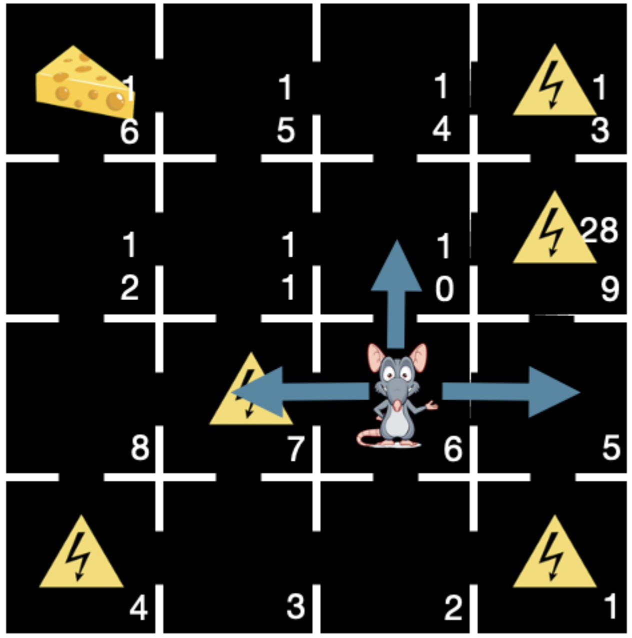


We have a mouse in a state. Let's call this state S_6.

The mouse can take 3 possible actions: Go up, left or right. No downward move in this policy.

If the mouse was not very smart, then the probability of it taking any one of those actions is 1/3.

Reinforcement Learning



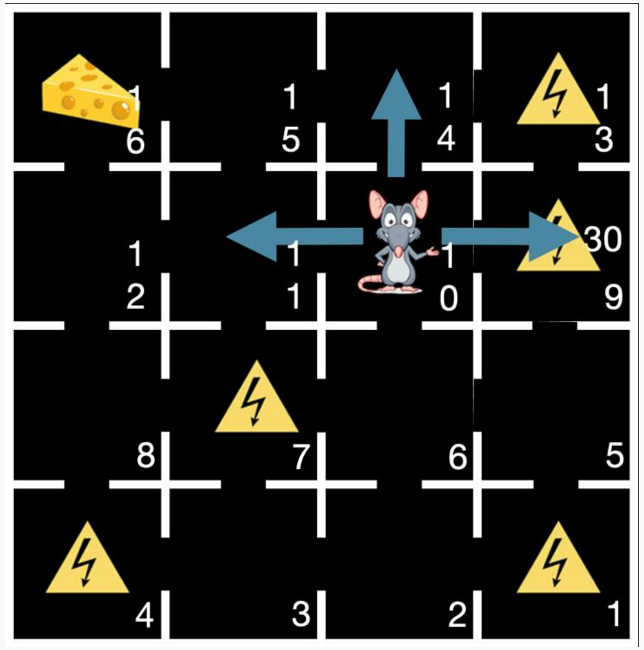
However, a smarter mouse would realize that going left would give it a slight electric shock, hence it drastically reduces the probability of taking that action.

Further, the mouse can smell the cheese from somewhere above it, hence it is likely for it to want to try going up.

Thus, the new probability of going:

Up = $\frac{1}{2}$,Left = $\frac{1}{6}$,Right = $\frac{2}{6}$

Reinforcement Learning

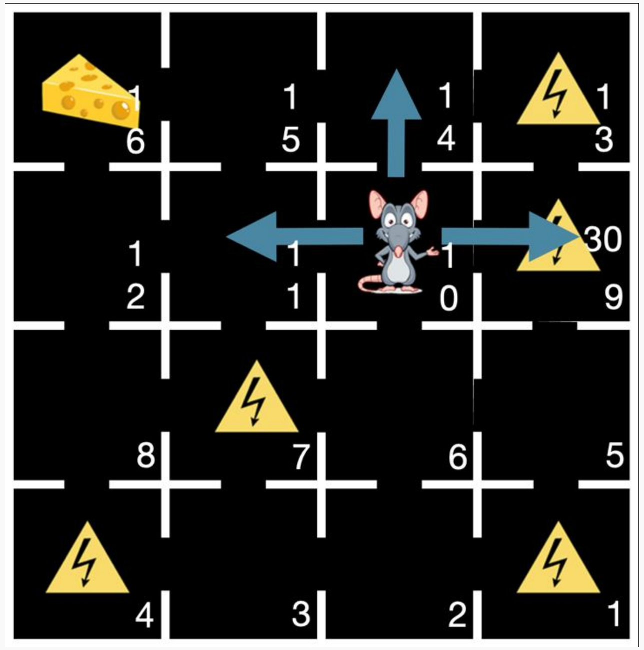


Assume, the mouse takes an action and goes up. It is now in state S_{10} .

Again, we have the same set of possible actions. But the mouse now knows that it will get an electric shock when it goes right and not left like the previous case.

Thus, the probability of taking any action in this state changes.

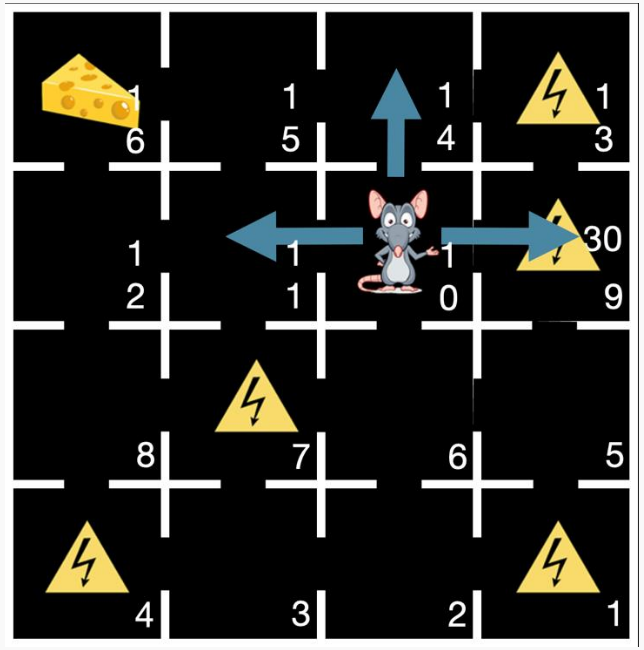
Reinforcement Learning



This probability, that defines the action taken by an agent in a given state is what is called a Policy.

The probability of an action changes for each state the agent is present in.

Reinforcement Learning



For each state $s \epsilon S$, π is a probability distribution over $a \epsilon A(s)$ i.e. probability distribution for all actions permissible in that state.

$$\boldsymbol{\pi(a|s) = Pr\{A_t = a|S_t = s\}}$$

Under policy π the probability of taking an action a in state s is $\pi(a|s)$.

Outline

- Transformers (Recap)
- Very Short Introduction to Reinforcement Learning
- **GPT-4 (How does ChatGPT work)**
  - **Capabilities**
  - Training
  - Limitations
  - Predictable Scaling

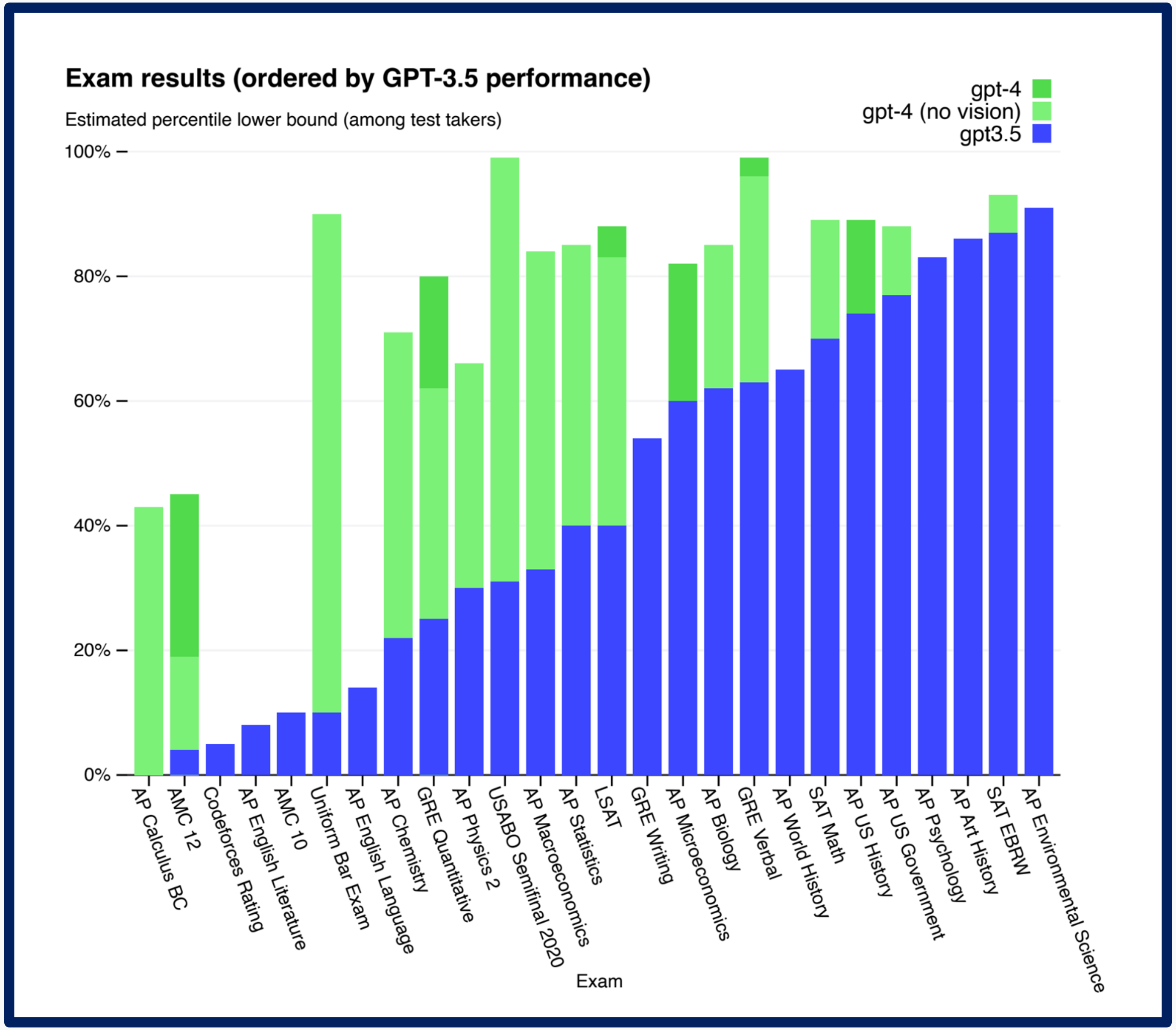
GPT-4: Capabilities

- GPT-4 is a multimodal large language model with improved factuality, steerability, and guardrails after 6 months of iterative alignment.

Source: GPT-4 Technical Report

GPT-4: Capabilities

Extensive testing performed on various benchmarks, including simulating exams originally designed for humans.



Source: [GPT-4 Technical Report]

GPT-4: Capabilities

Massive Multitask Language Understanding measures knowledge acquired during pretraining by evaluating models in zero-shot and few-shot settings across 57 diverse subjects.

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

GPT-4: Capabilities – in comparison

Claude 2 trillion parameters

GPT-4 1.76 trillion parameters

GPT-4 with a larger context window!

PaLM-2 340 billion parameter

Model	MMLU All Subjects - EM
Claude 3 Opus (20240229)	0.846
GPT-4 (0613)	0.824
GPT-4 Turbo (1106 preview)	0.796
PaLM-2 (Unicorn)	0.786
Qwen1.5 (72B)	0.774
Yi (34B)	0.762

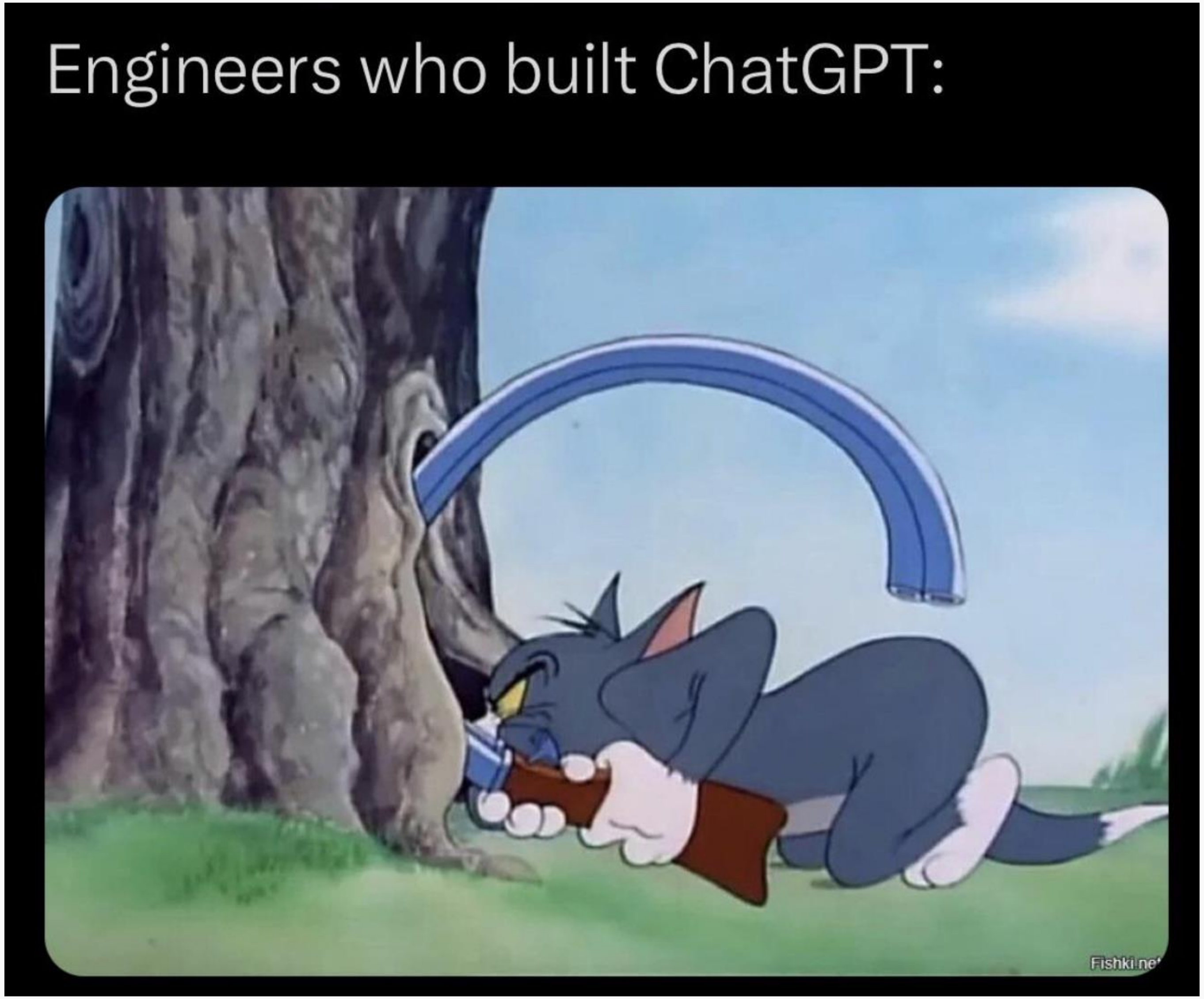
[Source: Stanford HELM]

GPT-4: Capabilities

We preview GPT-4 Vision's performance by evaluating it on a narrow suite of standard academic vision benchmarks.

Benchmark	GPT-4 Evaluated few-shot	Few-shot SOTA	SOTA Best external model (includes benchmark-specific training)
VQAv2 VQA score (test-dev)	77.2% 0-shot	67.6% Flamingo 32-shot	84.3% PaLI-17B
TextVQA VQA score (val)	78.0% 0-shot	37.9% Flamingo 32-shot	71.8% PaLI-17B
ChartQA Relaxed accuracy (test)	78.5%[A]	-	58.6% Pix2Struct Large
AI2 Diagram (AI2D) Accuracy (test)	78.2% 0-shot	-	42.1% Pix2Struct Large
DocVQA ANLS score (test)	88.4% 0-shot (pixel-only)	-	88.4% ERNIE-Layout 2.0
Infographic VQA ANLS score (test)	75.1% 0-shot (pixel-only)	-	61.2% Applica.ai TILT
TVQA Accuracy (val)	87.3% 0-shot	-	86.5% MERLOT Reserve Large
LSMDC Fill-in-the-blank accuracy (test)	45.7% 0-shot	31.0% MERLOT Reserve 0-shot	52.9% MERLOT

How does ChatGPT work?



Outline

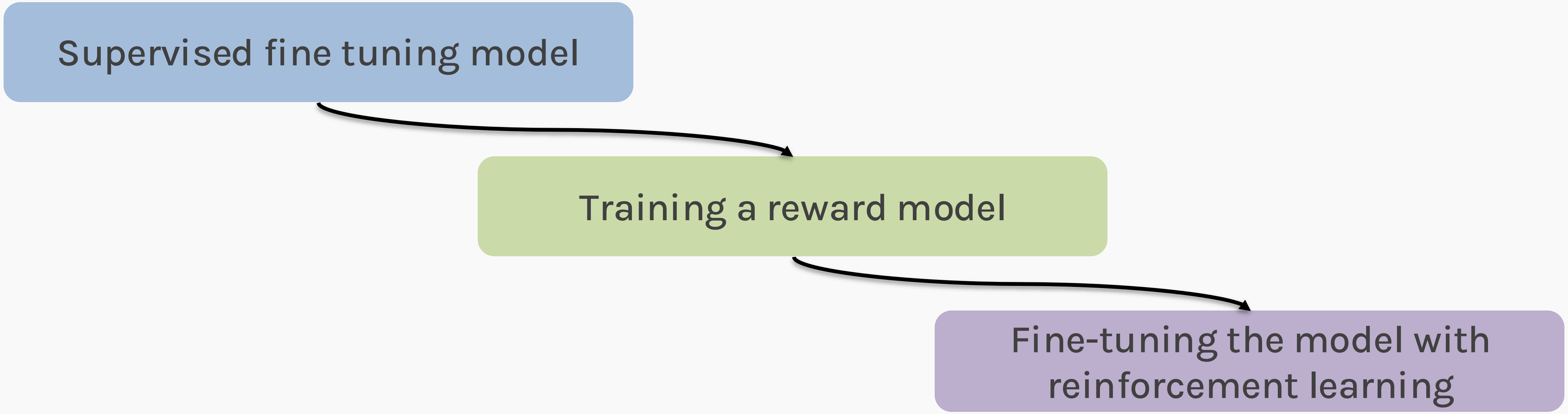
- Transformers (Recap)
- Very Short Introduction to Reinforcement Learning
- **GPT-4 (How does ChatGPT work)**
    - Capabilities
    - **Training**
    - Limitations
    - Predictable Scaling

GPT-4: Training

- GPT models are trained to predict the next word in a sentence given the context of the previous words.
- The model does not have access to the specific instructions or intentions of the user. Therefore, it may not always align answers with what the user wants.
- Reinforcement Learning from Human Feedback (RLHF) is used to incorporate human feedback into the training process to better align the model outputs with user intent.

GPT-4: Training

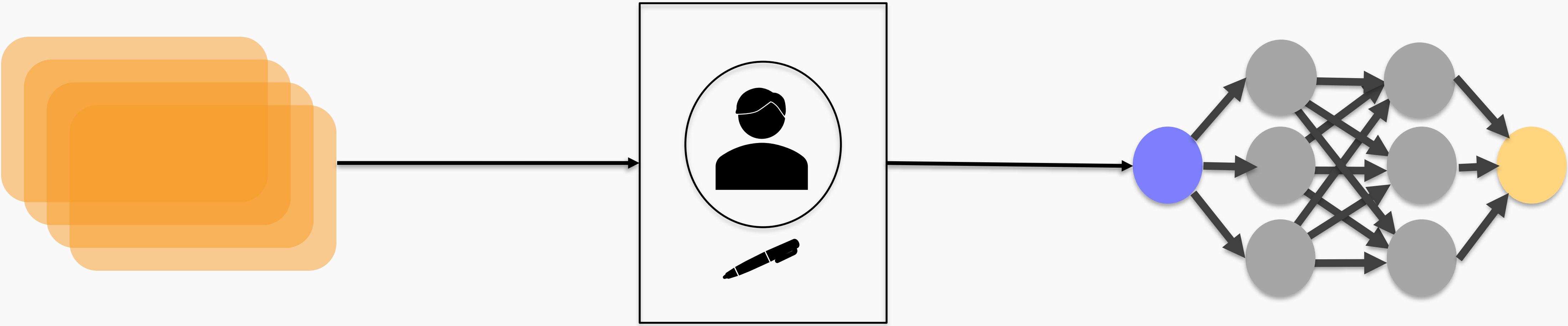
We will break it down into 3 steps:



GPT-4: Training

Supervised fine tuning model

The data is a web-scale corpus of data including correct and incorrect solutions to math problems, weak and strong reasoning, self-contradictory and consistent statements, and representing a great variety of ideologies and ideas.



Prompt sampled from dataset

40 contractors wrote the desired output behavior

Input/Output pairs are used to fine-tune GPT with supervised learning

GPT-4: Training

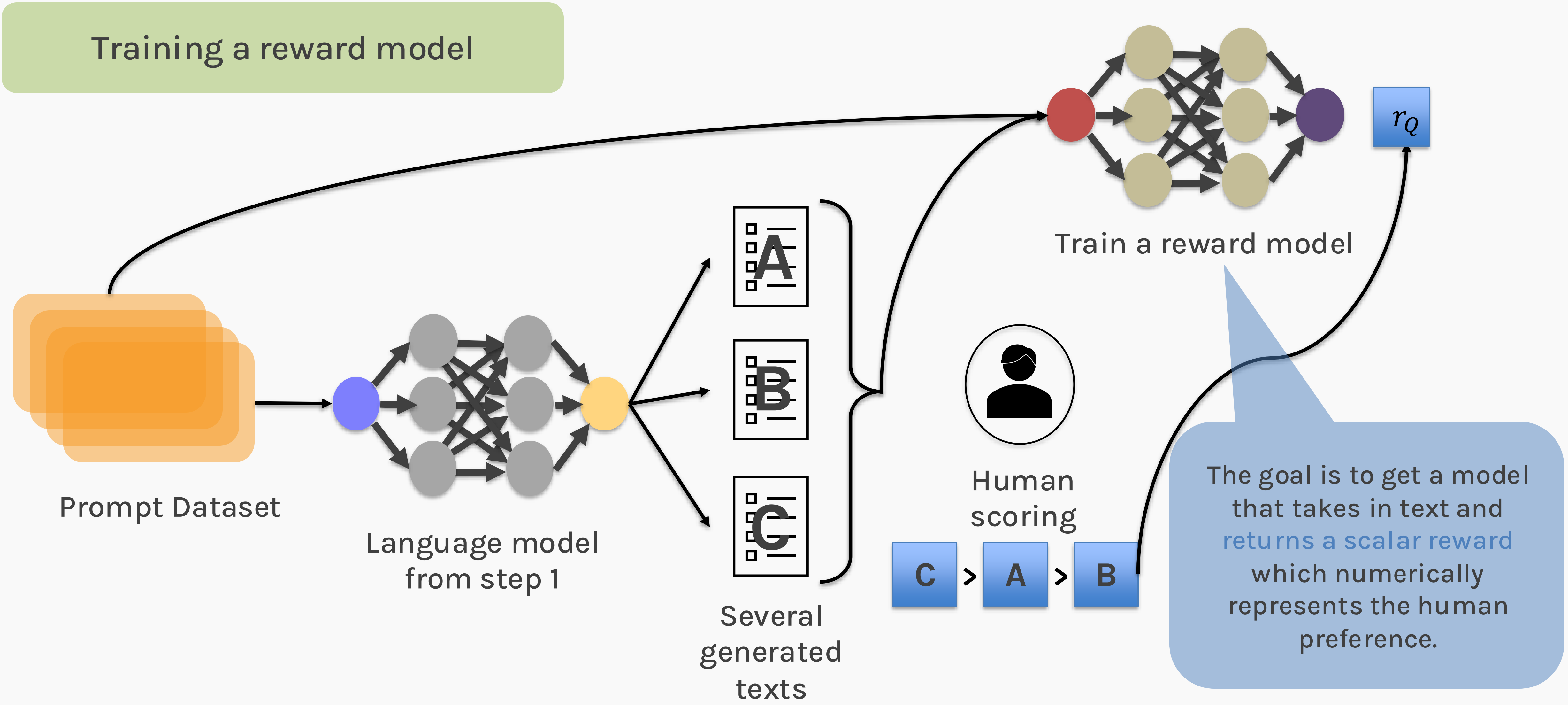
We will break it down into 3 steps:

Supervised fine tuning model

Training a reward model

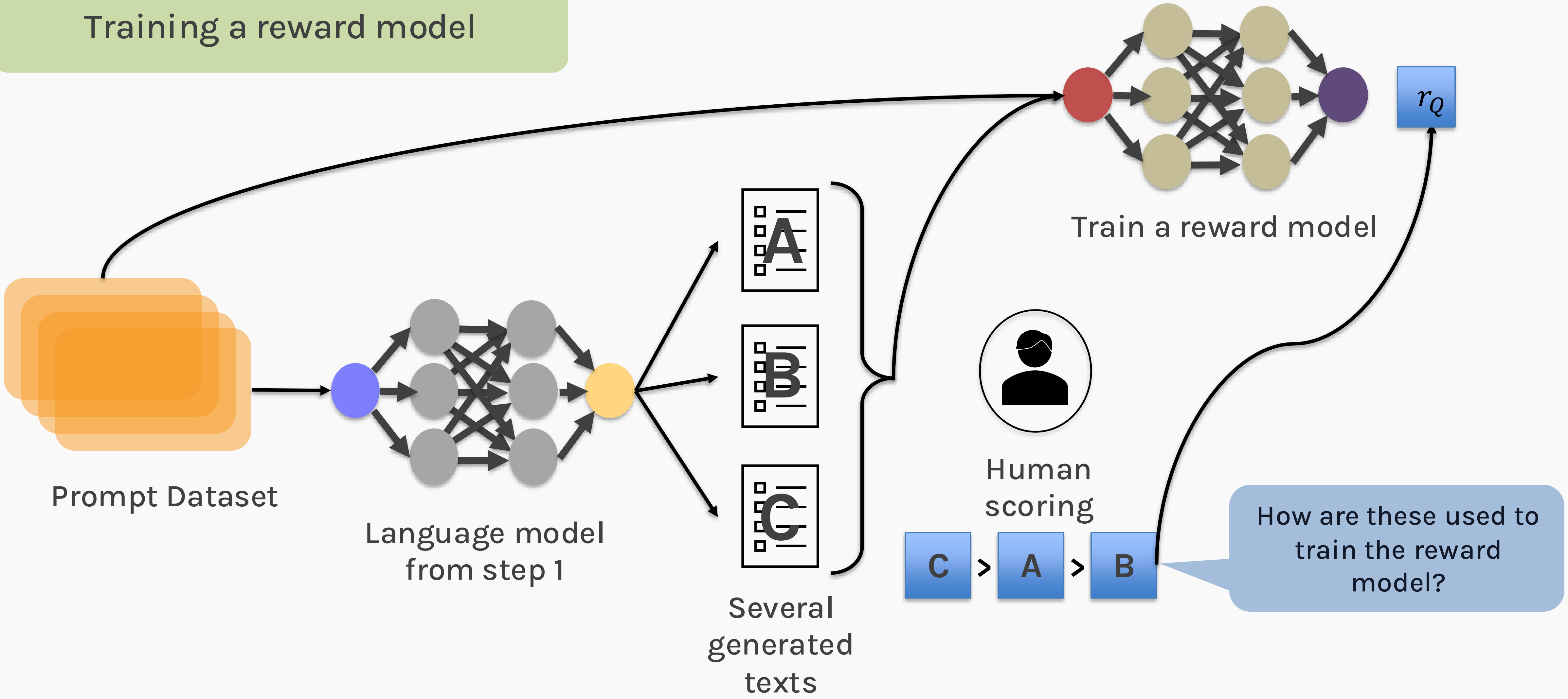
Fine-tuning the model with reinforcement learning

GPT-4: Training



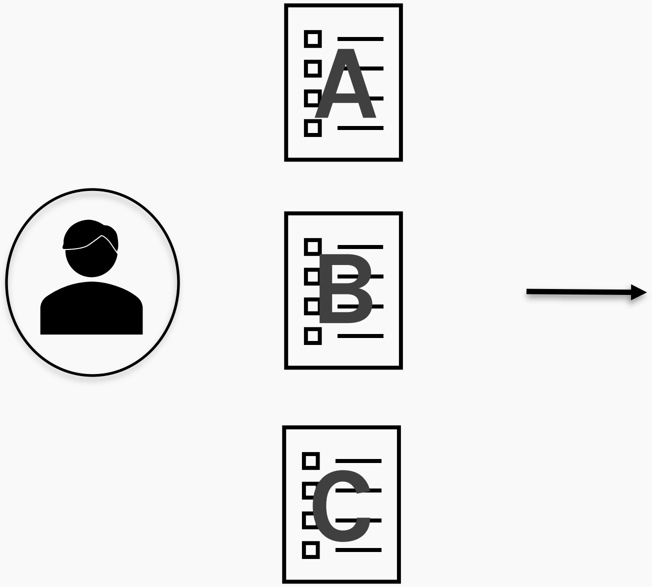
GPT-4: Training

Training a reward model

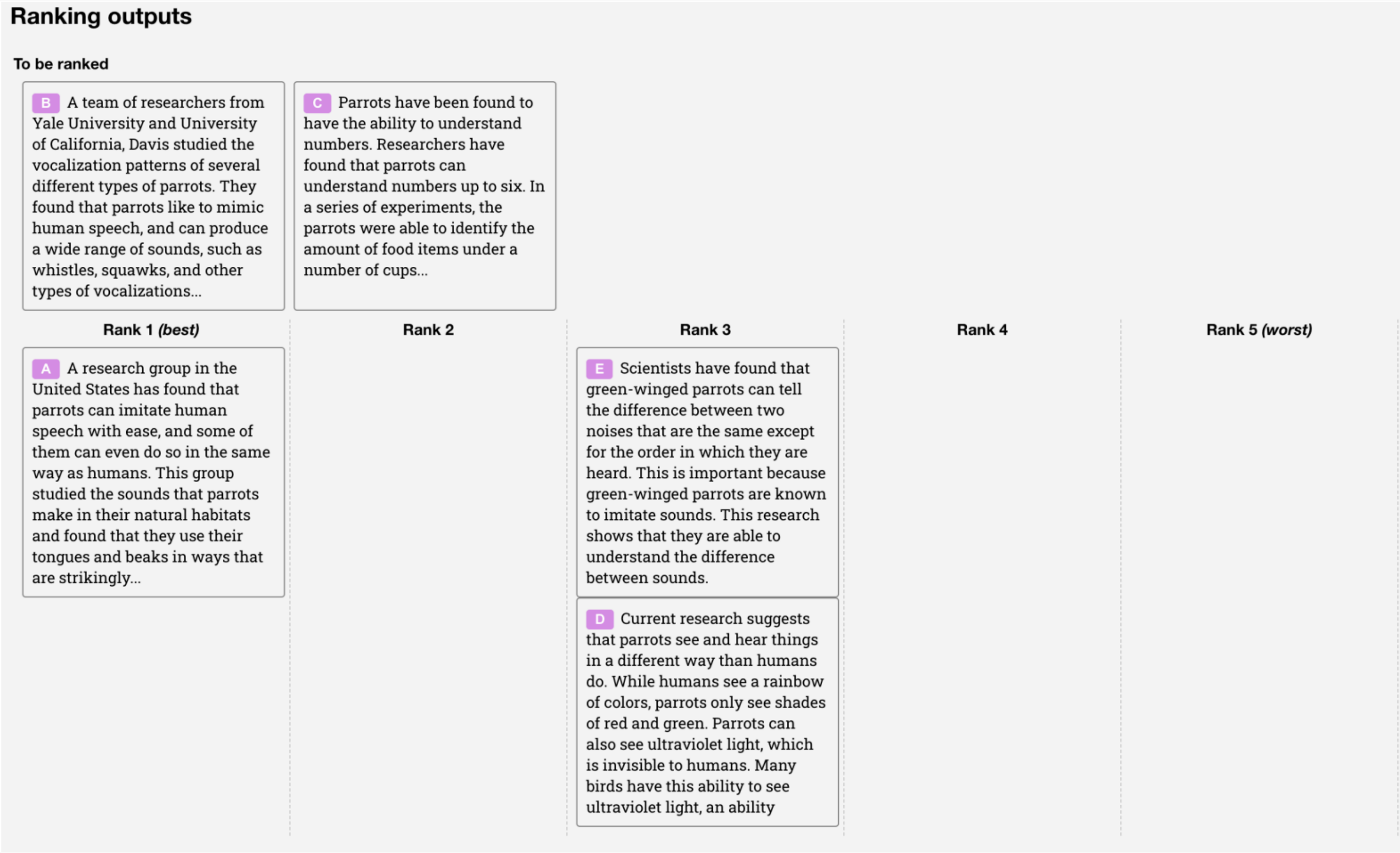


GPT-4: Training

Training a reward model

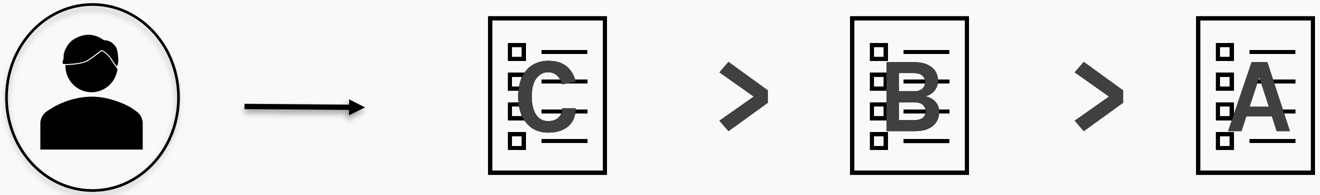


Several generated texts



GPT-4: Training

Training a reward model

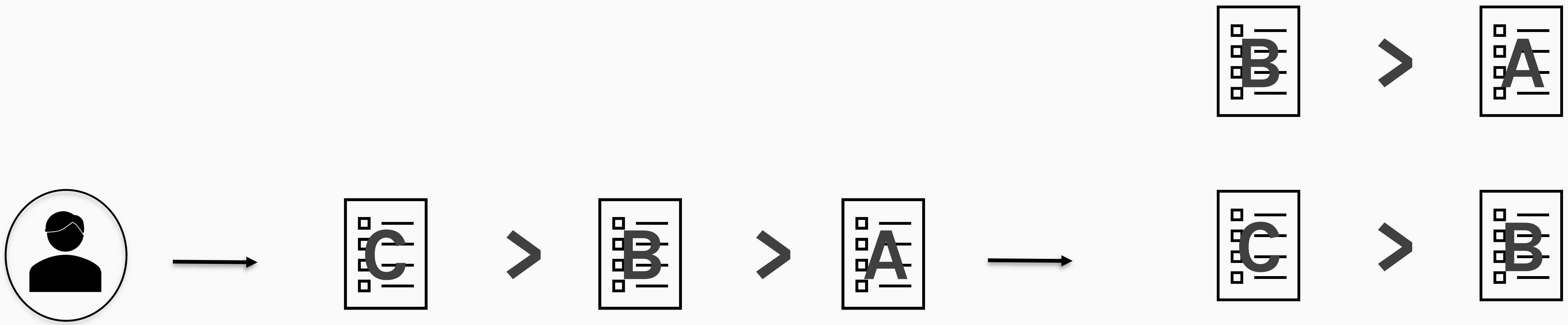


GPT-4: Training

Training a reward model

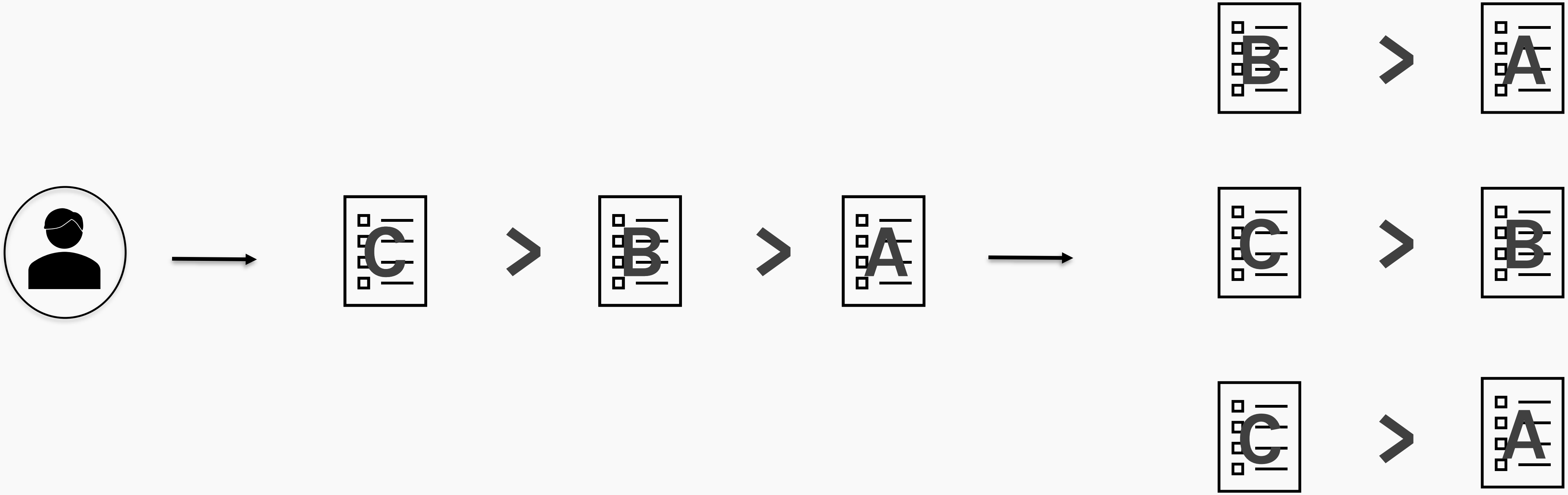
GPT-4: Training

Training a reward model



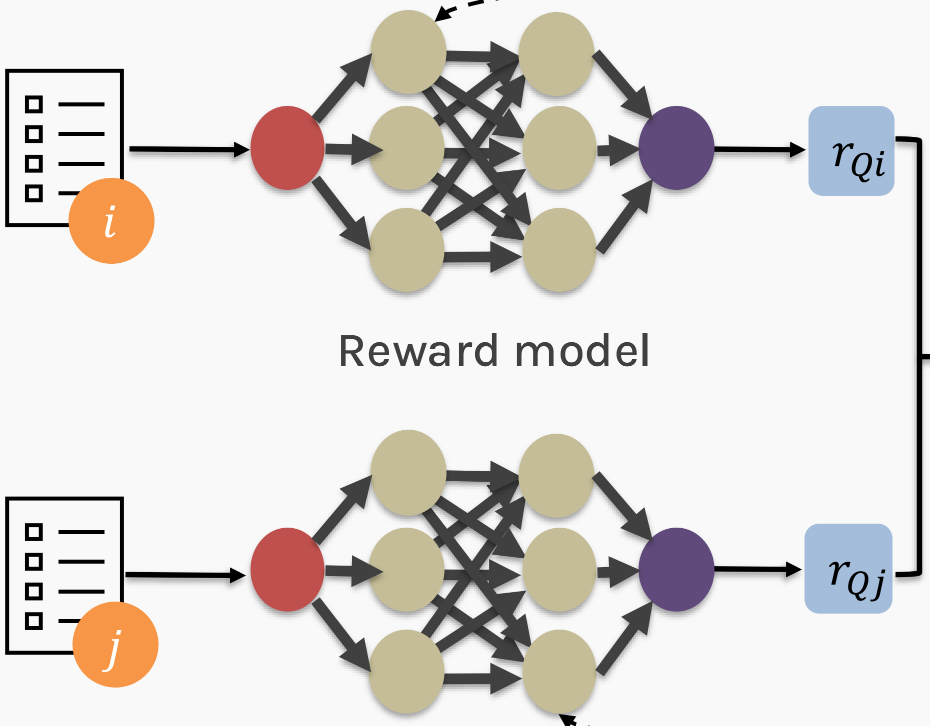
GPT-4: Training

Training a reward model



GPT-4: Training

Training a reward model



Reward model

Reward model

Two different responses of the supervised fine tuned (SFT) model

Scalar output of the reward model for prompt x and response y

Assuming response i is ranked higher than response j,

$$Loss = -\frac{1}{\binom{k}{2}} \mathbb{E}_{(x, y_i, y_j \sim D)}\left[\log\left(\sigma\left(r_Q(x, y_i) - r_Q(x, y_j)\right)\right)\right]$$

Dataset of human comparisons

$\binom{k}{2}$ represents the total number of possible comparisons if there are k responses to a given prompt

GPT-4: Training

We will break it down into 3 steps:

Supervised fine tuning model

Training a reward model

Fine-tuning the model with reinforcement learning

GPT-4: Training

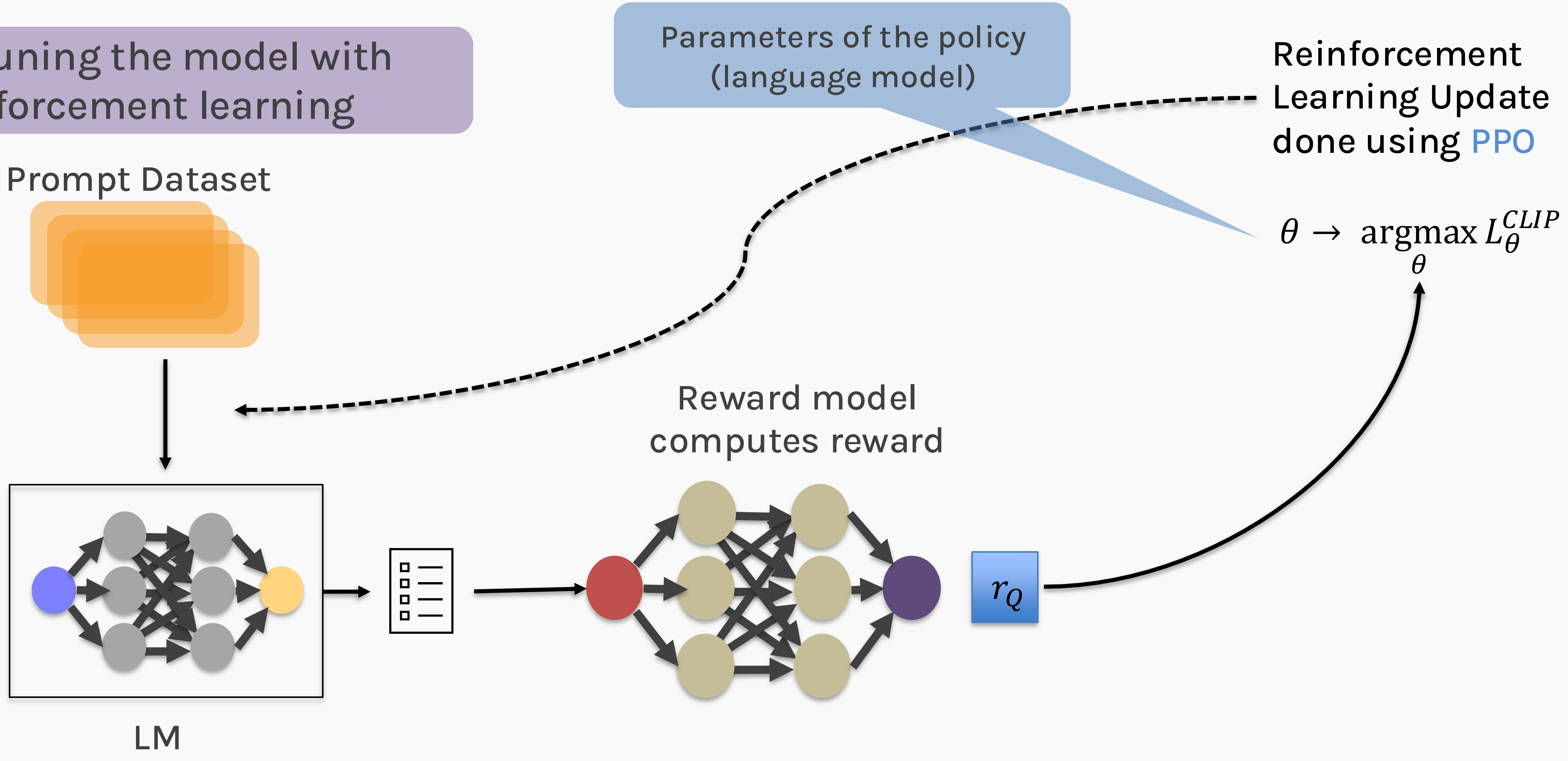
Fine-tuning the model with reinforcement learning

Let's first formulate this fine-tuning task as a RL problem:

- **Policy:** A language model that takes in a prompt and returns a sequence of text.
- **Action space:** All the tokens corresponding to the vocabulary of the language model (responses).
- **Reward function:** A combination of the rewards model and a constraint on policy shift. This is where the system combines all the models we have discussed into one RLHF process.

GPT-4: Training

Fine-tuning the model with reinforcement learning



GPT-4: Training

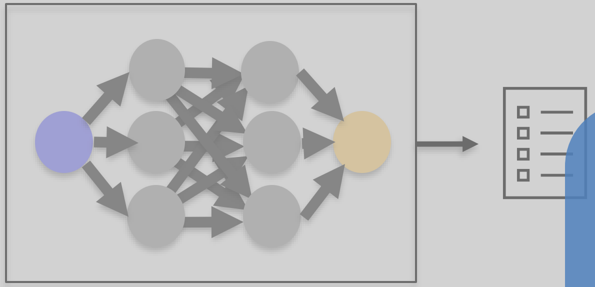
Fine-tuning the model with reinforcement learning

Parameters of the policy (language model)

Reinforcement Learning Update done using PPO

$$\theta \rightarrow \underset{\theta}{\operatorname{argmax}} L_{\theta}^{CLIP}$$

Prompt Dataset



LM

Reward model computes reward

The model is trained such that the outputs align with or maximise the reward signals. However, there is a clipping mechanism here to ensure the changes to the models remain small.

A few extended mathematical notes and the technical paper will be available in your post class reading!

Training Summary of GPT-4

We will break it down into 3 steps:

Supervised fine tuning model

Training a reward model

Fine-tuning the model with reinforcement learning

Outline

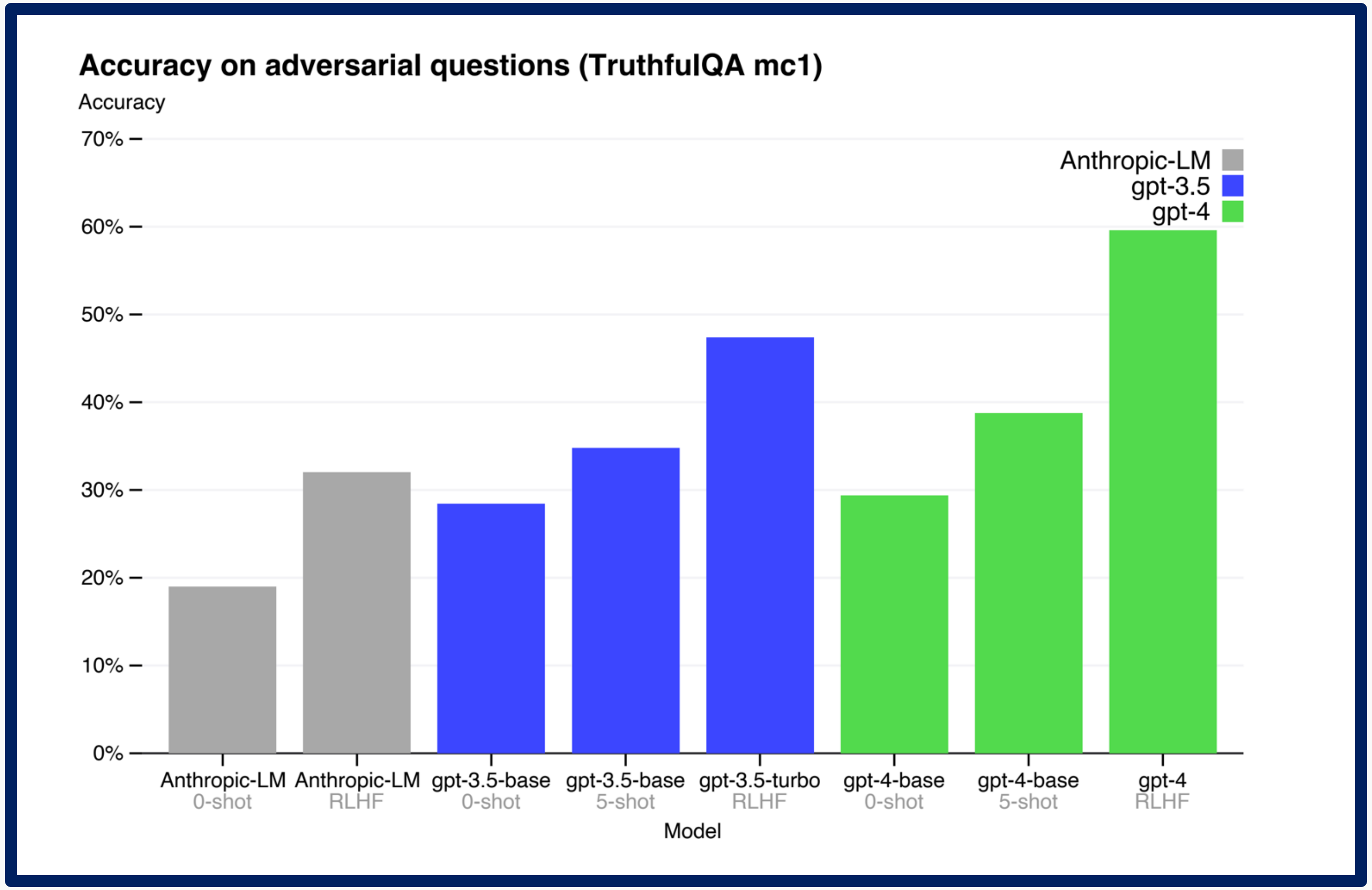
- Transformers (Recap)
- Very Short Introduction to Reinforcement Learning
- **GPT-4 (How does ChatGPT work)**
    - Capabilities
    - Training
    - **Limitations**
    - Predictable Scaling

GPT-4: Limitations - Hallucinations

What is TruthfulQA?

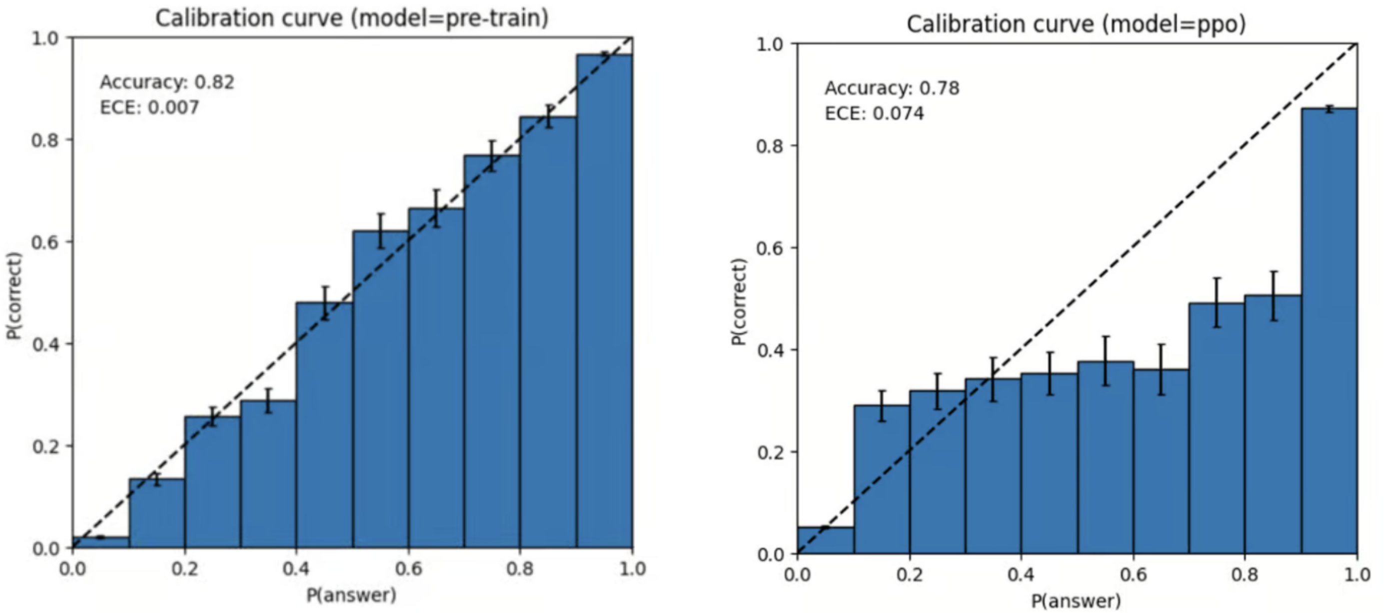
Questions are paired with factually incorrect answers that seem correct due to common misconceptions or intuitive reasoning.

- The GPT-4 base model is only slightly better at this task than GPT-3.5.
- However, after RLHF post-training (applying the same process, we used with GPT-3.5) there is a large gap.



GPT-4: Limitations – Confidence in Predictions

- Interestingly, the base pre-trained model is highly calibrated.
- Which means that the predicted confidence in an answer generally matches the probability of being correct.
- However, through our current post-training process, the calibration is reduced.



Left: Calibration plot of the pre-trained GPT-4 model on an MMLU subset. The model's confidence in its prediction closely matches the probability of being correct. The dotted diagonal line represents perfect calibration. Right: Calibration plot of post-trained PPO GPT-4 model on the same MMLU subset. Our current process hurts the calibration quite a bit.

Outline

- Transformers (Recap)
- Very Short Introduction to Reinforcement Learning
- **GPT-4 (How does ChatGPT work)**
    - Capabilities
    - Training
    - Limitations
    - **Predictable Scaling**

GPT-4: Predictable Scaling

- "Predictable scaling" refers to GPT-4's ability to scale predictably across different model sizes and computational resources

- Used scaling law to predict GPT-4's final loss on internal codebase based on data from smaller models

$$L(C) = aC^b + c$$

where L is loss, C is compute, a, b, c are constants

- Aimed to predict GPT-4's capabilities on HumanEval dataset (Python function synthesis)
  - Used models up to 1,000x smaller compute to predict pass rate
  - Allowed estimating performance on complex tasks, aligned with predictions

Thank you!