devfest

AI for Safer AI

ChengCheng Tan

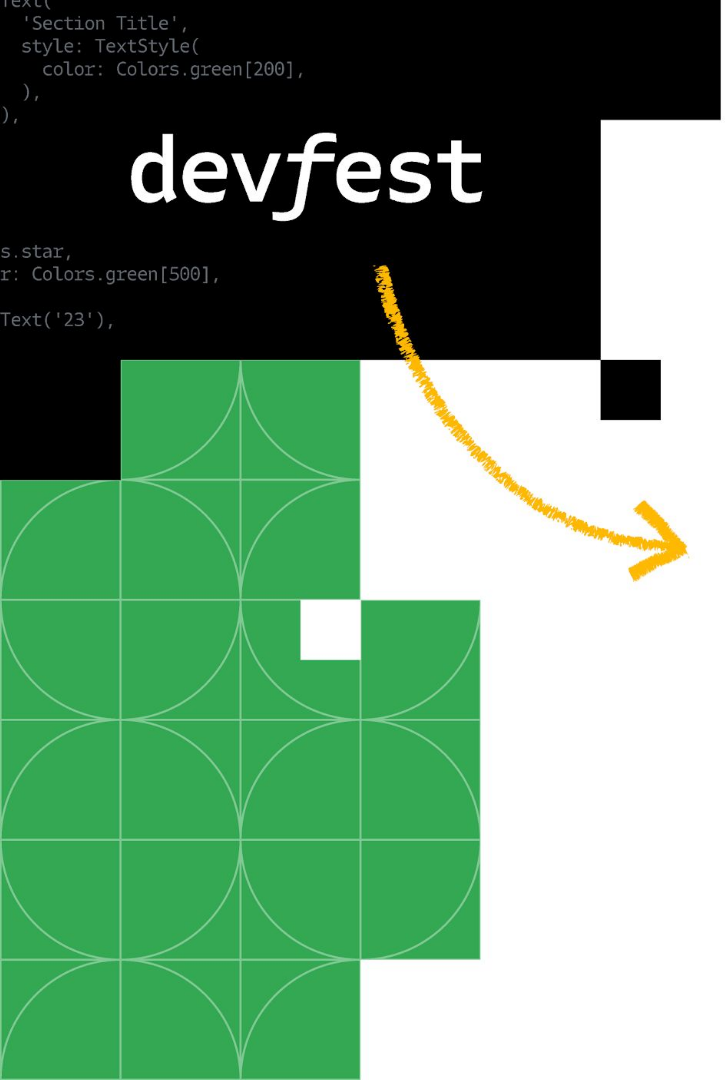
Google Developer Groups
Sacramento

AI & LLM Overview

1. Clarify some terms
2. Transformers
3. Large Language Models
4. Pre-training vs Fine-tuning

LLMs in AI Safety

1. Clarify more terms
2. What is AI safety?
3. Examples of technical research
4. Resources



Google Developer Groups

Sacramento

AI & LLM Overview

Terminology

AI: Artificial Intelligence

ML: Machine Learning

LLM: Large Language Model

NLP: Natural Language Processing

GPT: Generative Pretrained Transformer

Transformers: Neural network leading to LLMs

RLHF: Reinforcement Learning from Human Feedback

Natural Language Processing [NLP]: **Computers can speak & understand human languages**

Pre-1990s:

Rule-based Expert Systems

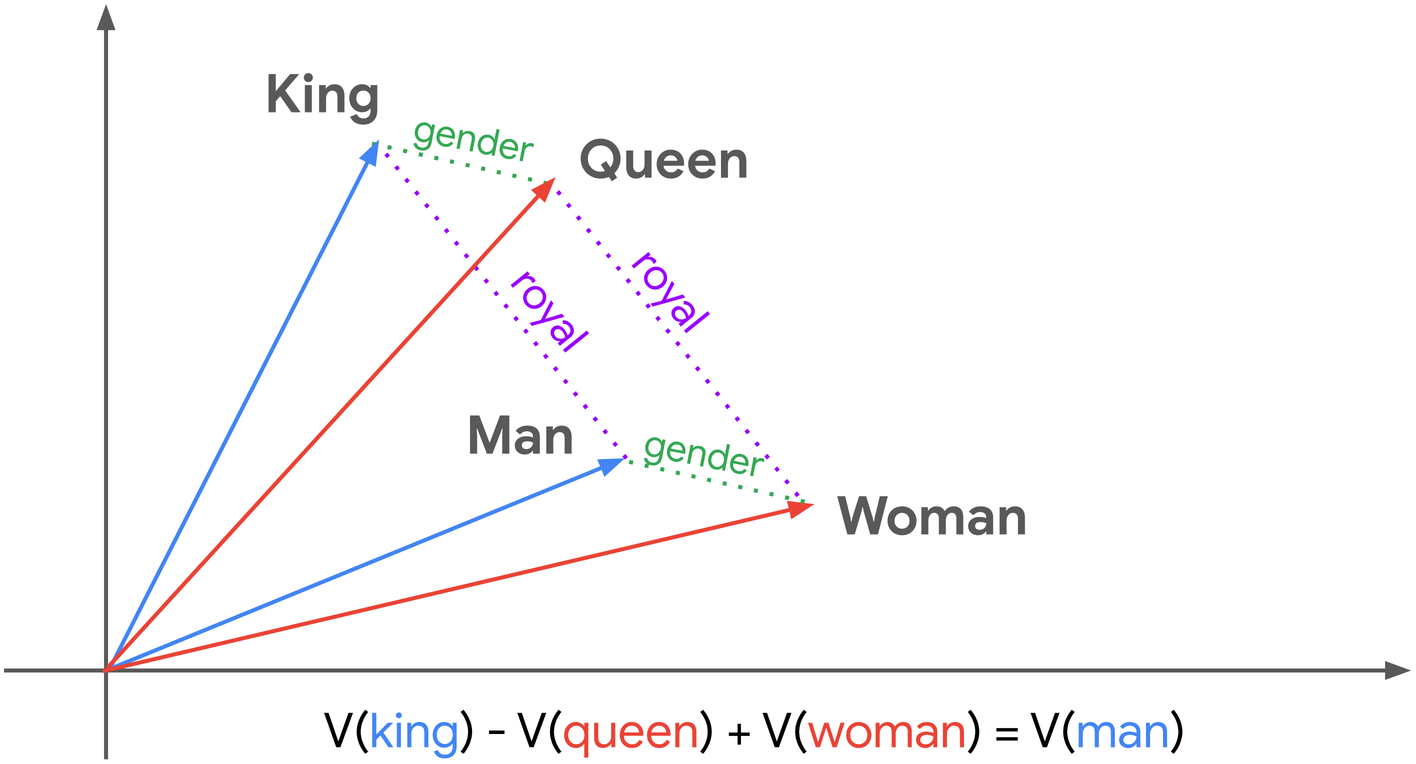
1990s-2000s: **Statistics & Probabilities**

> You shall know a word by the company it keeps
>
> J.R. Firth, Linguist

2010s:

Rise of Deep Learning and Neural Networks

2013: Word2Vec Embeddings



V(king) - V(queen) + V(woman) = V(man)

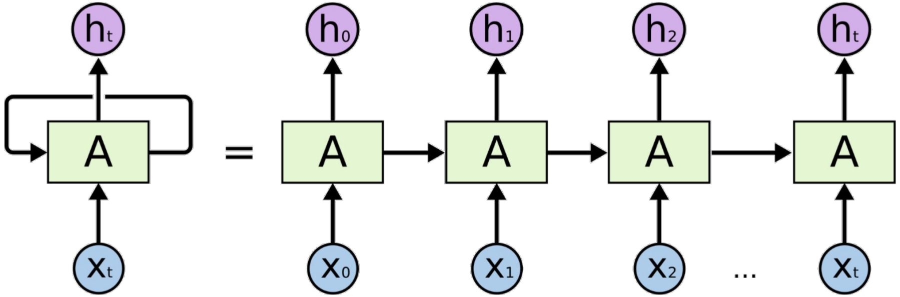
2013: Word2Vec Embeddings

Analogies	Word Pair 1		Word Pair 2	
Man-Woman	king	queen	man	woman
Capital city	Athens	Greece	Oslo	Norway
City-in-state	Chicago	Illinois	Sacramento	California
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Nationality adjective	Switzerland	Swiss	Canada	Canadian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars

2010s:

Neural Networks

RNN, GRU, LSTM



Early Neural Networks

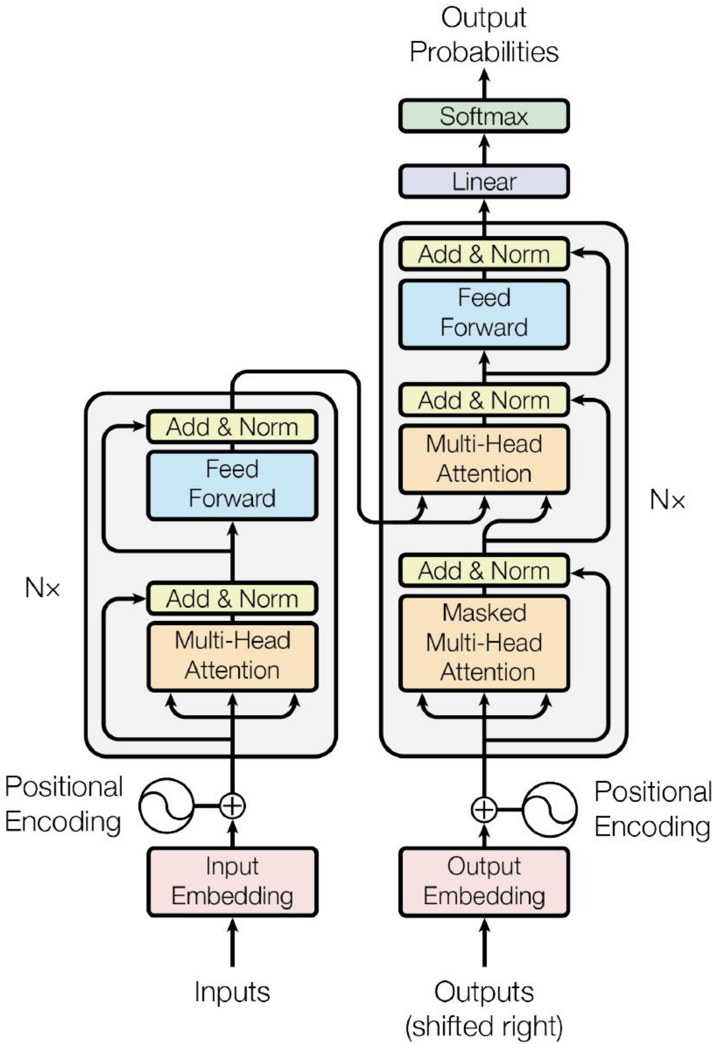
- Slow & forgetful

2017: Transformers

- Attention mechanism
- Parallel processing

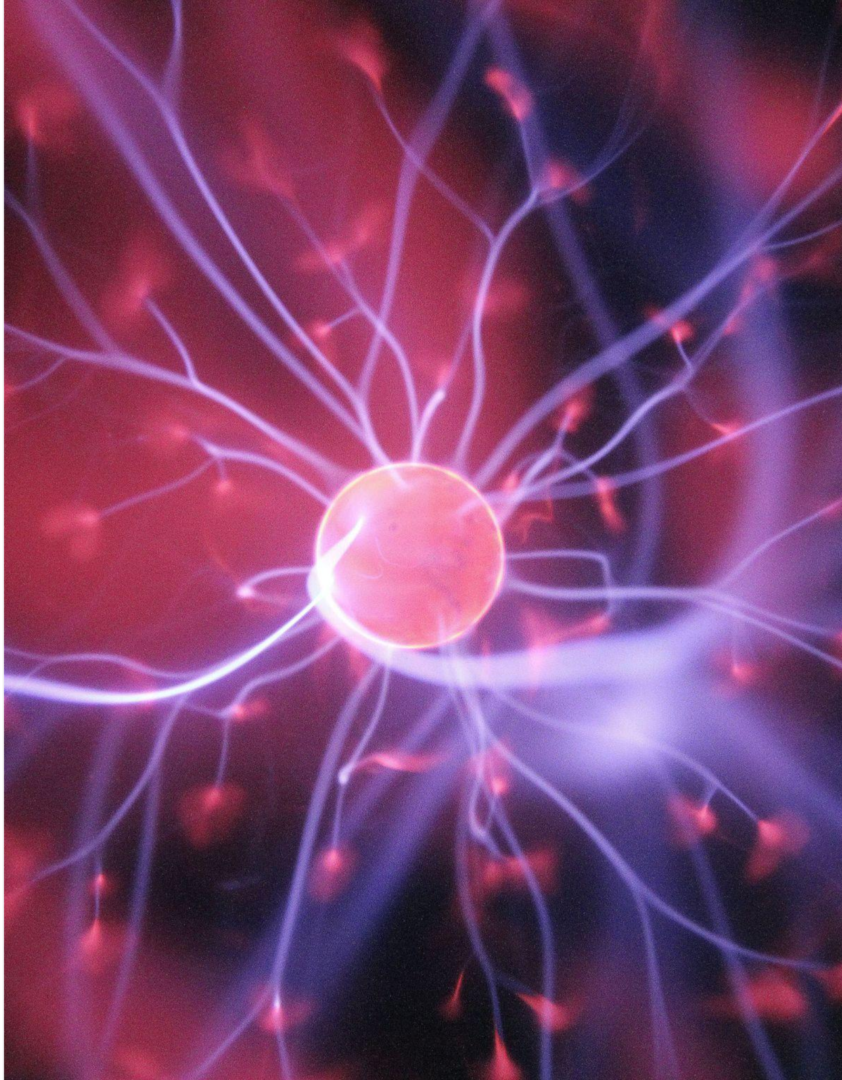
Transformer Architecture

Encoder + Decoder



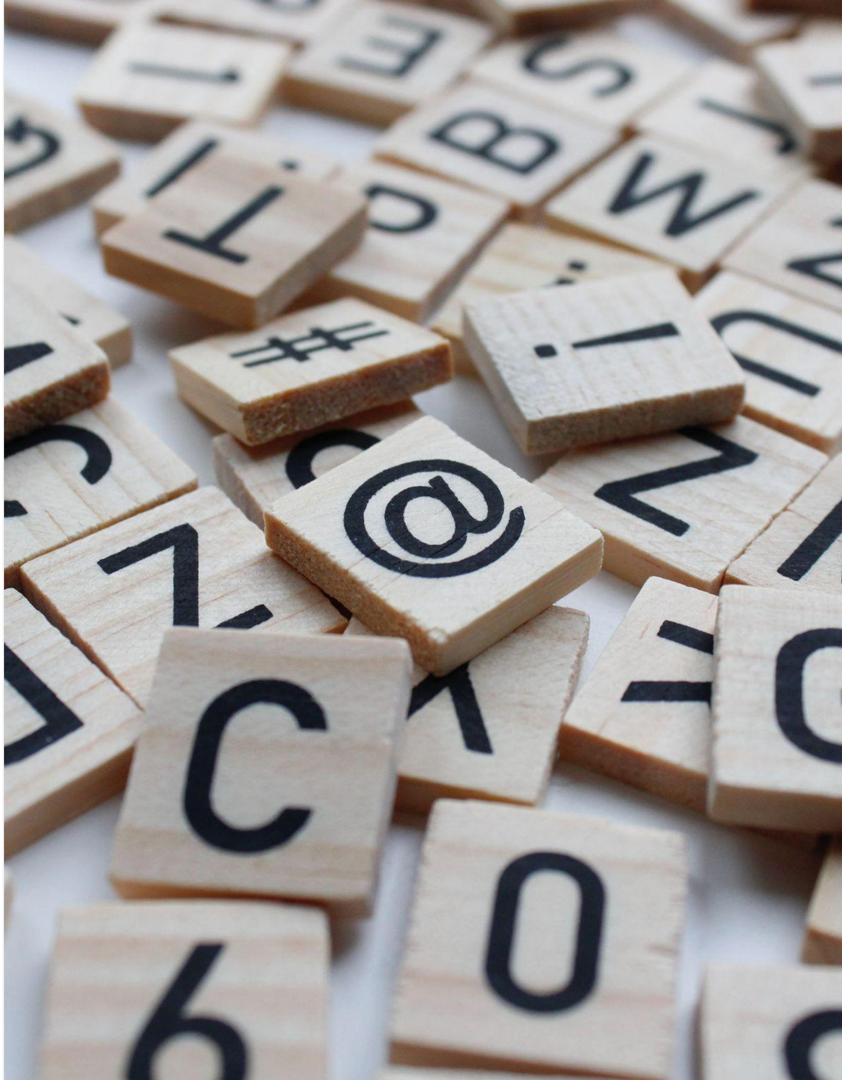
Rise of LLMs

>1 billion neurons



Trained for

next word prediction



Pre-trained base

vs

Fine-tuned models

RLHF:

Reinforcement Learning from Human Feedback

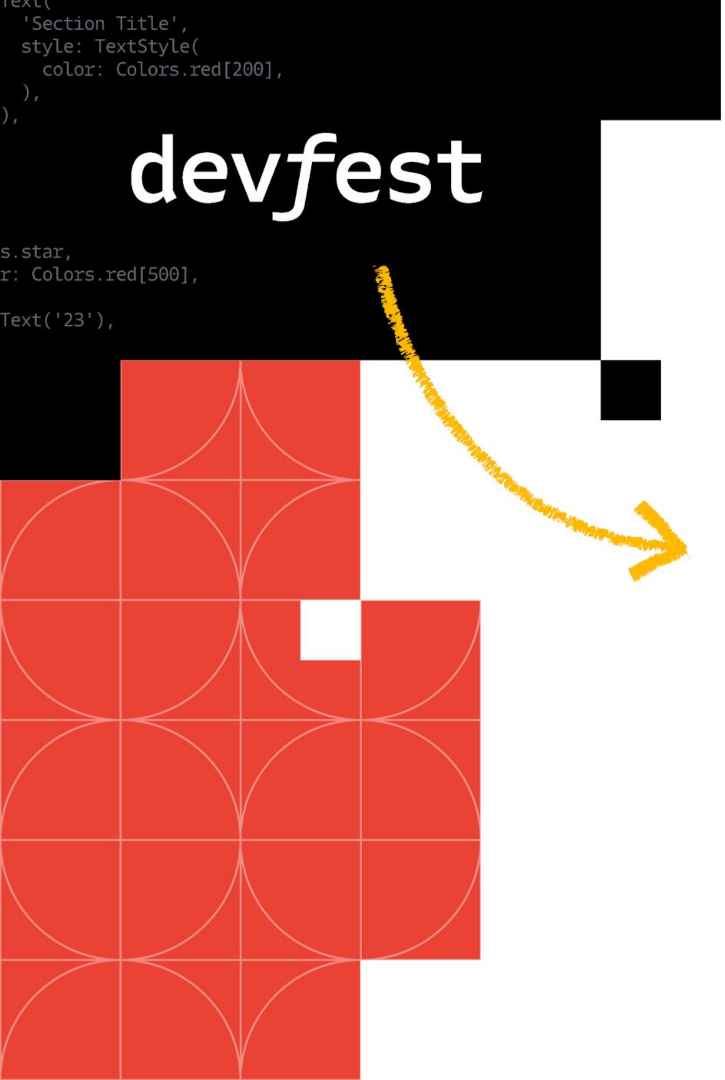
Google Developer Groups

Fine-tuned to **follow instructions**

Fine-tuned for **conversations**

What’s Next?

- Multimodal
- Open-source
- Agents



Google Developer Groups

Sacramento

LLMs in AI Safety

```
  'Simple Statement or URL',
  style: TextStyle(
    color: Colors.red[200],
  ),
),
```

devfest

```
s.star,
r: Colors.red[500],

Text('23'),
```

> Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.
>
> Center for AI Safety

More Terminology

AGI: Artificial GENERAL Intelligence, Strong AI

ANI: Artificial NARROW Intelligence, Weak AI

HLI: Human-Level Intelligence

ASI: Artificial SUPERINTELLIGENCE surpass humans on all tasks

Orthogonality Thesis: intelligence & goals are independent

Takeoff: fast vs slow

Timelines: short vs long

ANI

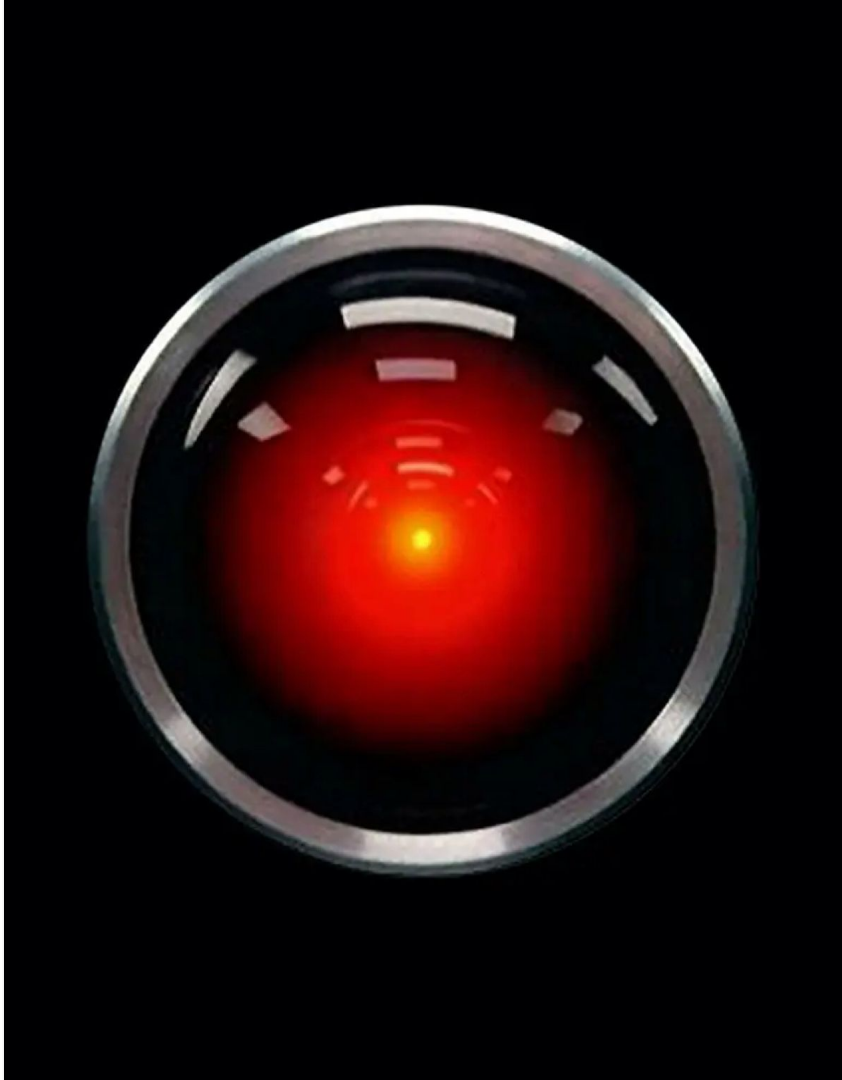
Artificial Narrow Intelligence

Weak AI

AGI

Artificial General Intelligence

Strong AI



Intelligence:

- Human-level
- Superintelligence

... Singularity

Takeoff

Fast vs Slow

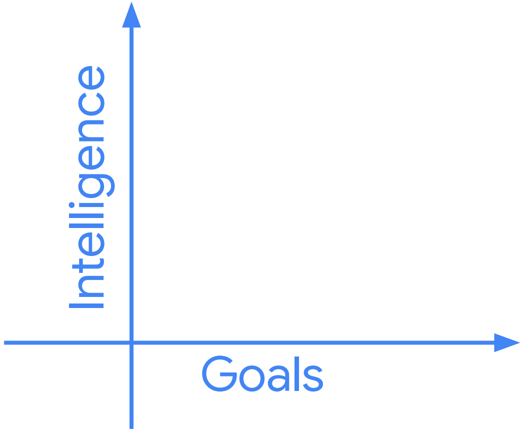
Timelines

Short vs Long

Google Developer Groups

Orthogonality Thesis

Intelligence & Goals are independent



Google Developer Groups

Paperclip Maximizer

Thought experiment or stamp alternative

Outer Misalignment

AI creator goals don’t align with general human values

Inner Misalignment

AI achieves goals in ways unintended by its creators

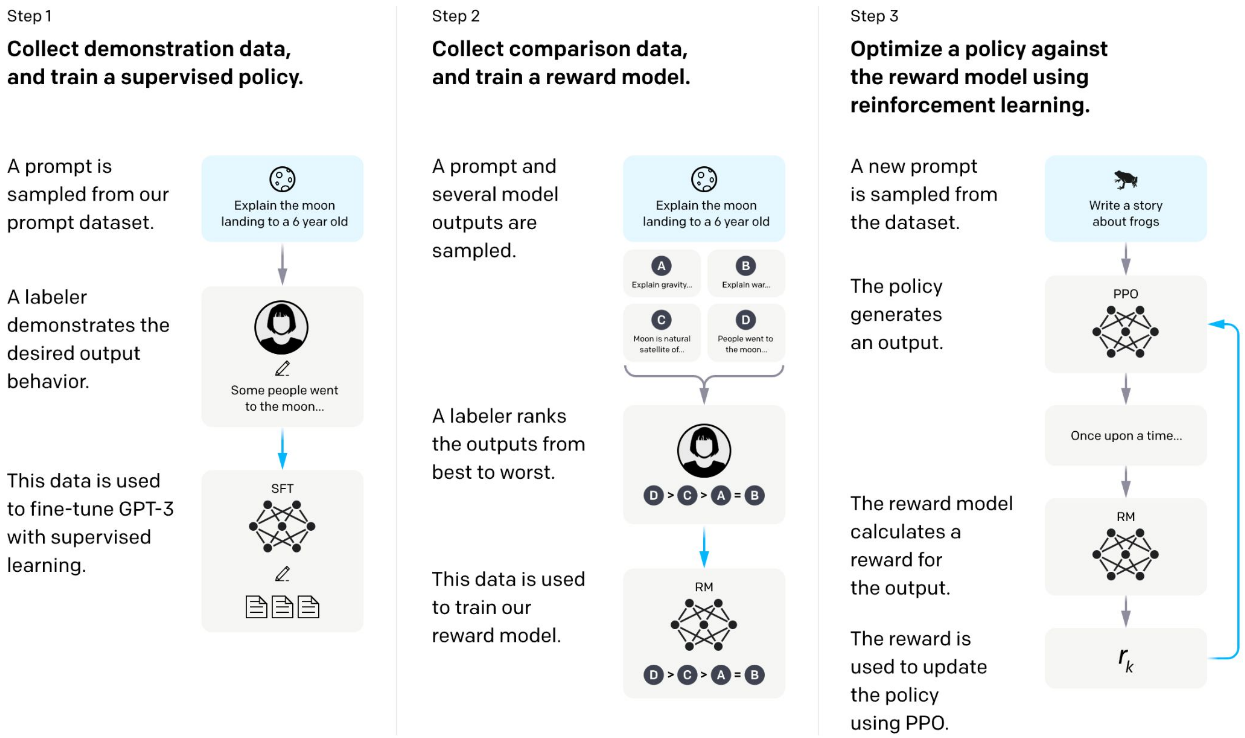
Gorilla Problem

Can humans maintain autonomy in a world with superintelligence?

Some Concrete Examples of Technical Research...

Google Developer Groups

Alignment: RLHF Human Values



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

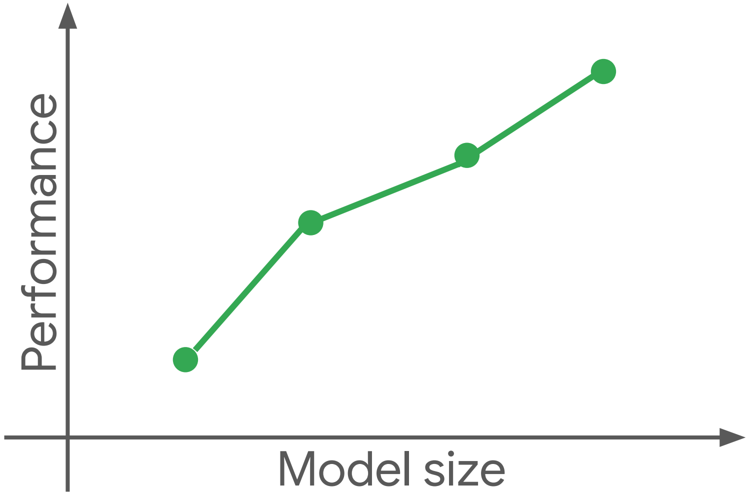
The policy generates an output.

The reward model calculates a reward for the output.

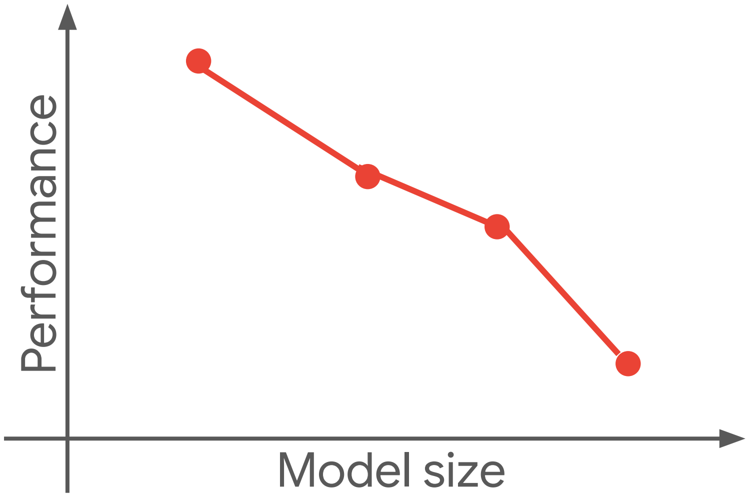
The reward is used to update the policy using PPO.

LLM Output Evaluation: Inverse Scaling Laws

Most tasks like this



Any tasks like this?



LLM Internals

Interpretability

```
Text(
  'Simple Statement or URL',
  style: TextStyle(
    color: Colors.green[200],
  ),
),
```

devfest

```
s.star,
r: Colors.green[500],

Text('23'),
```

We need scientific & technical breakthroughs to steer and control AI systems much smarter than us.

OpenAI Superalignment

Resources

1. **AI Explained videos on AI development + safety**
   youtube.com/@aiexplained-official
2. **80,000 Hours career advice + job board**
   80000hours.org/problem-profiles/artificial-intelligence
3. **AISafety.info FAQs**
   AISafety.info
4. **AI Safety Fundamentals online curricula**
   AISafetyFundamentals.com
5. **Alignment Forum share research + discussions**
   AlignmentForum.org

Questions?

Embrace Safely

ChengCheng Tan

https://www.linkedin.com/in/cheng2-tan/
ccstan99@gmail.com

Google Developer Groups
Sacramento