devfest

Adventures through LLMs

Tracing Origins & Charting a Safe Future

Google Developer Groups
Burnaby

Who am I?

- BA Linguistics & CS, UCLA
- MS CS HCI, Stanford
- LLM + AI Safety
  FAR AI Communications
  AISafety.info + chatbot
- Google WTM Ambassador

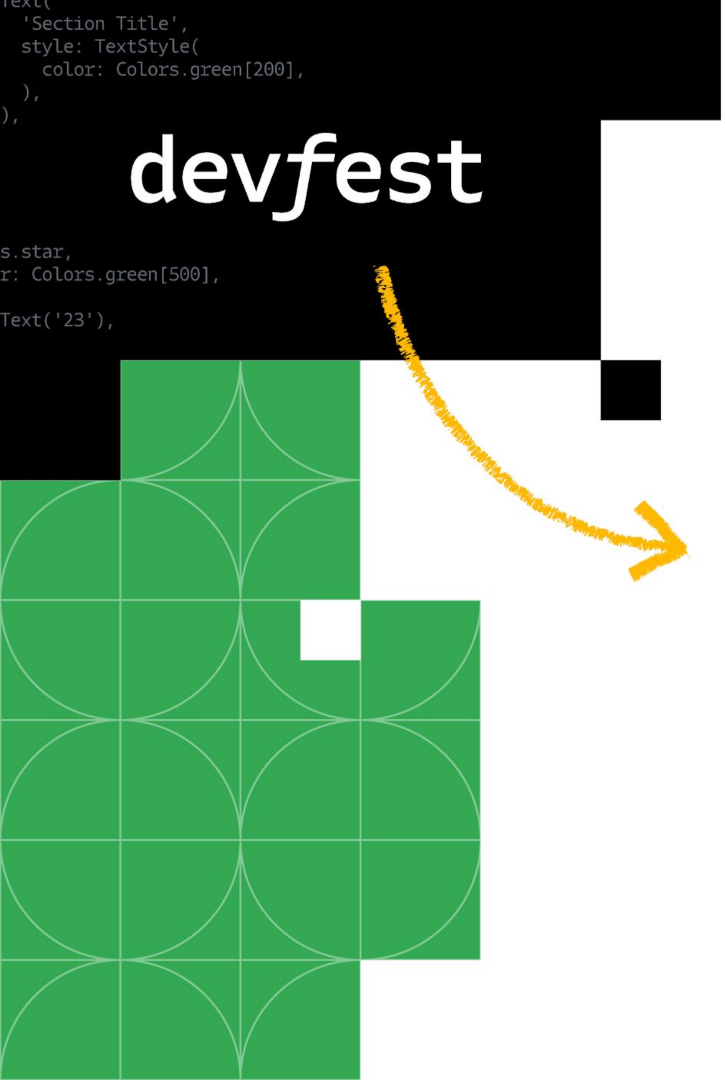
* DISCLAIMER ideas presented here are my own

LLM Overview: Tracing Origins

- Transformers
- Large Language Models
- Pre-training vs Fine-tuning

LLMs in AI Safety: Charting a Safer Future

- What is AI Safety?
- Examples of Technical Research
- Resources



```
Text(
  'Section Title',
  style: TextStyle(
    color: Colors.green[200],
  ),
),
```

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

Text('23'),
```

Google Developer Groups

Burnaby

LLM Overview

Tracing Origins

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

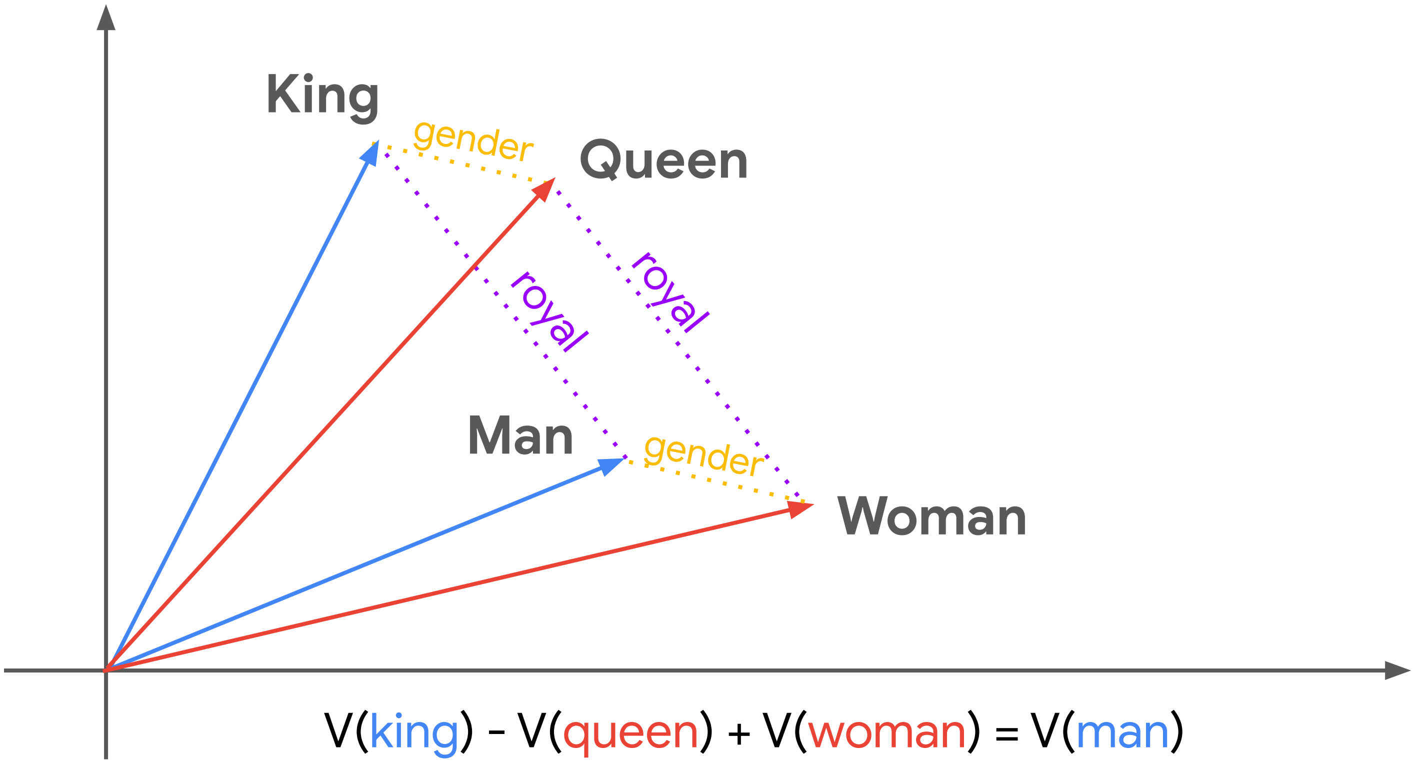
J.R. Firth, Linguist

2010s:

Rise of Deep Learning and Neural Networks

Google Developer Groups

2013: Word2Vec Embeddings



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

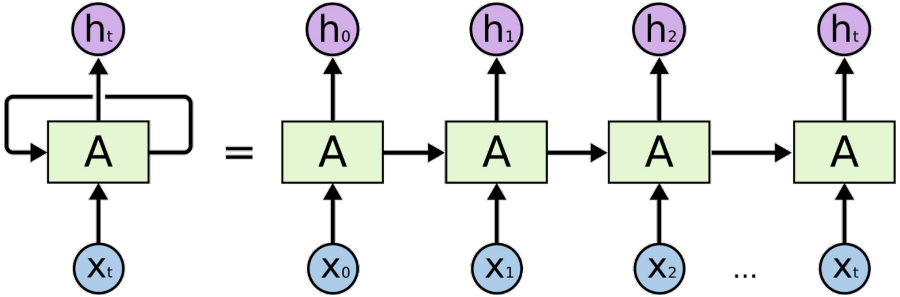
2013: Word2Vec Embeddings

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

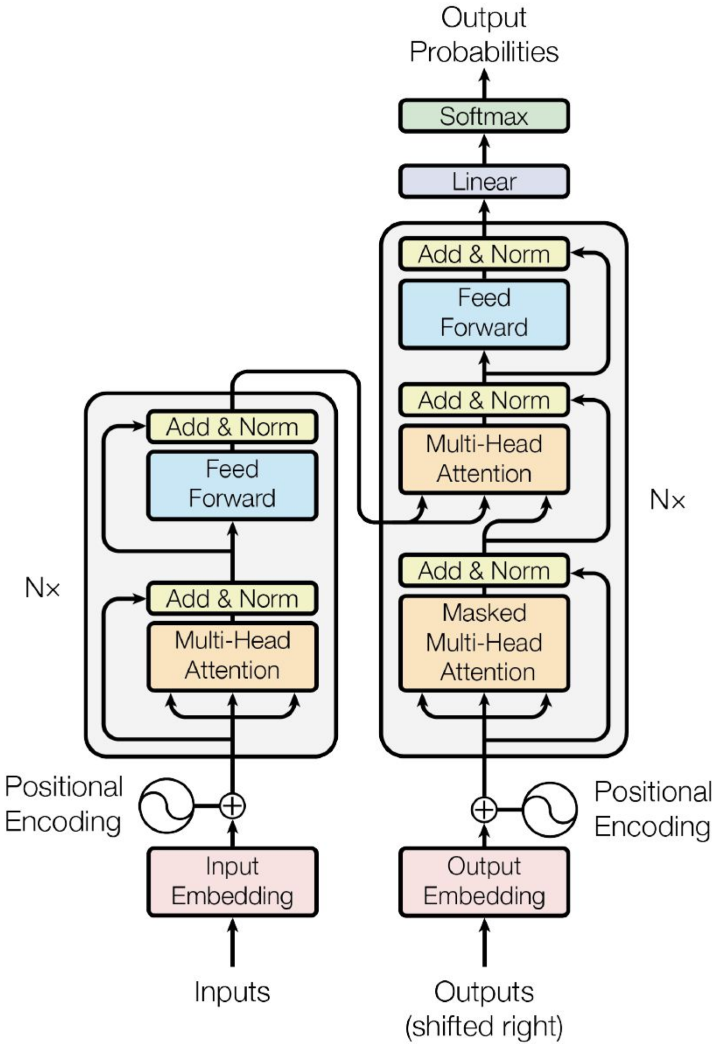
- Slow & Forgetful

2017: Transformers

- Self-attention Mechanism
- Data hungry
- Parallel processing GPU-optimized

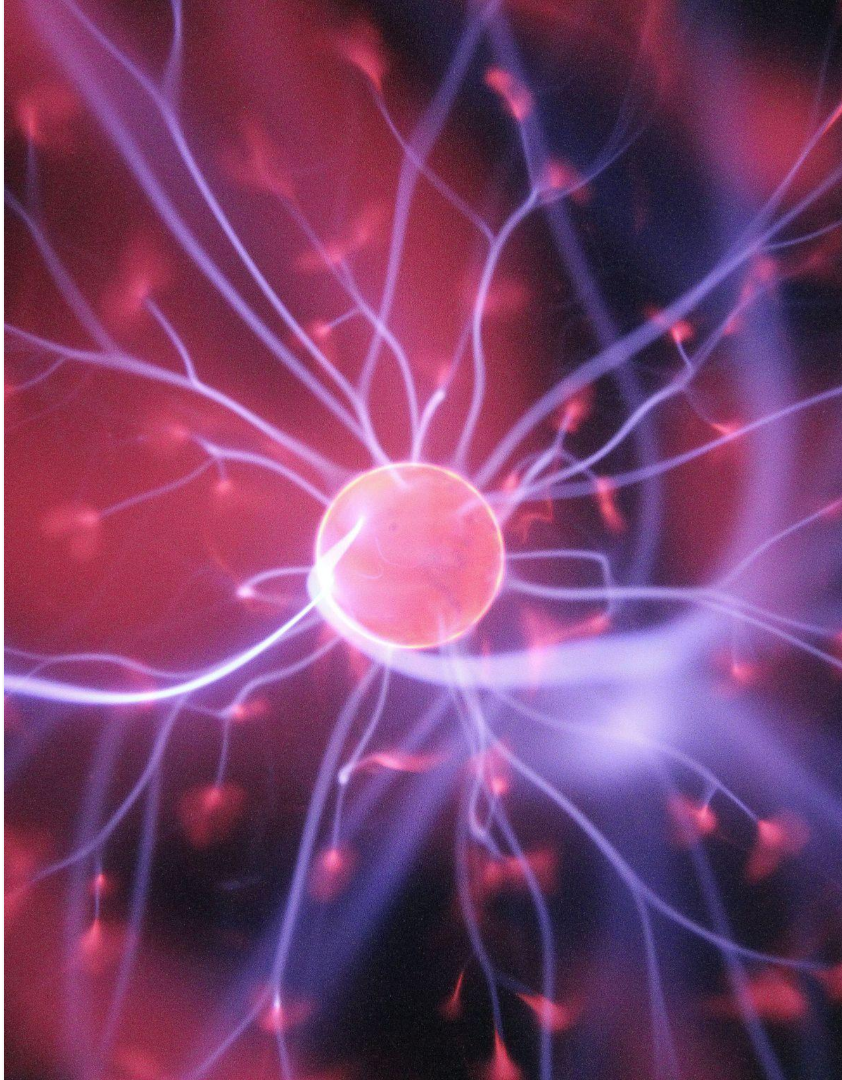
Transformer Architecture

Encoder + Decoder



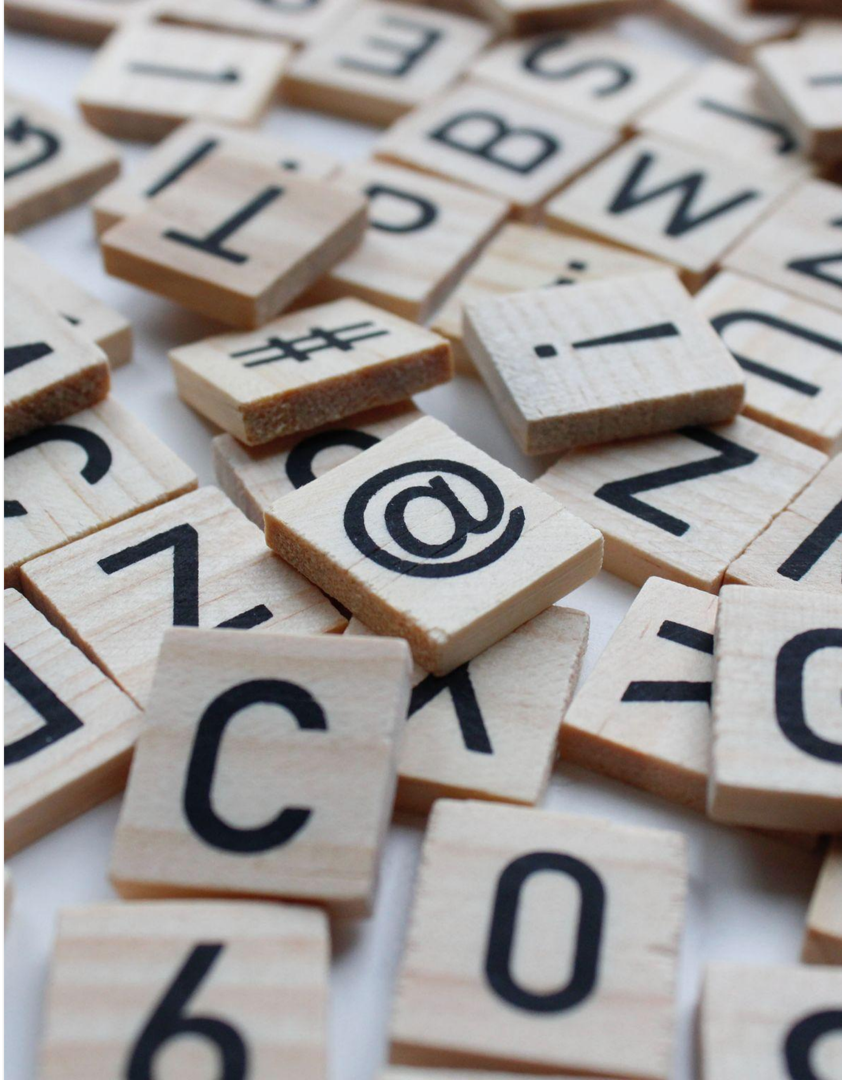
Rise of LLMs

>1 billion neurons

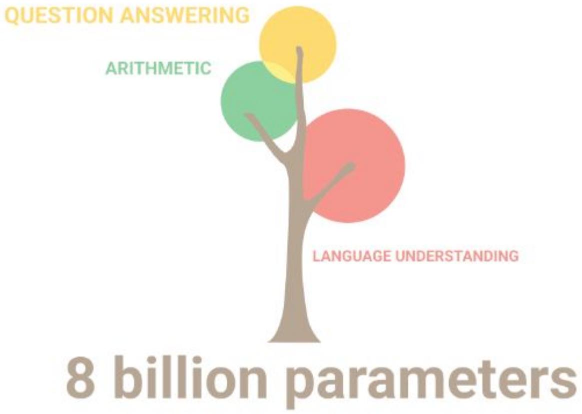


Trained for

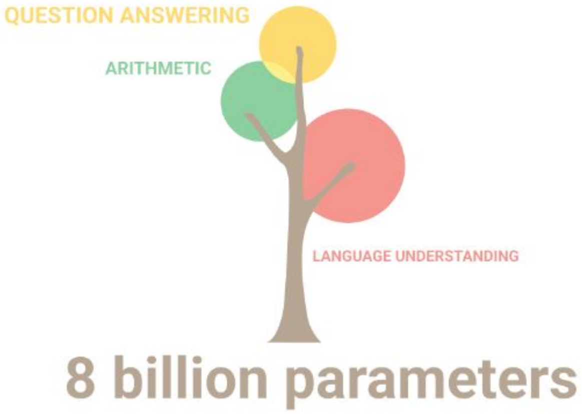
Next Word Prediction



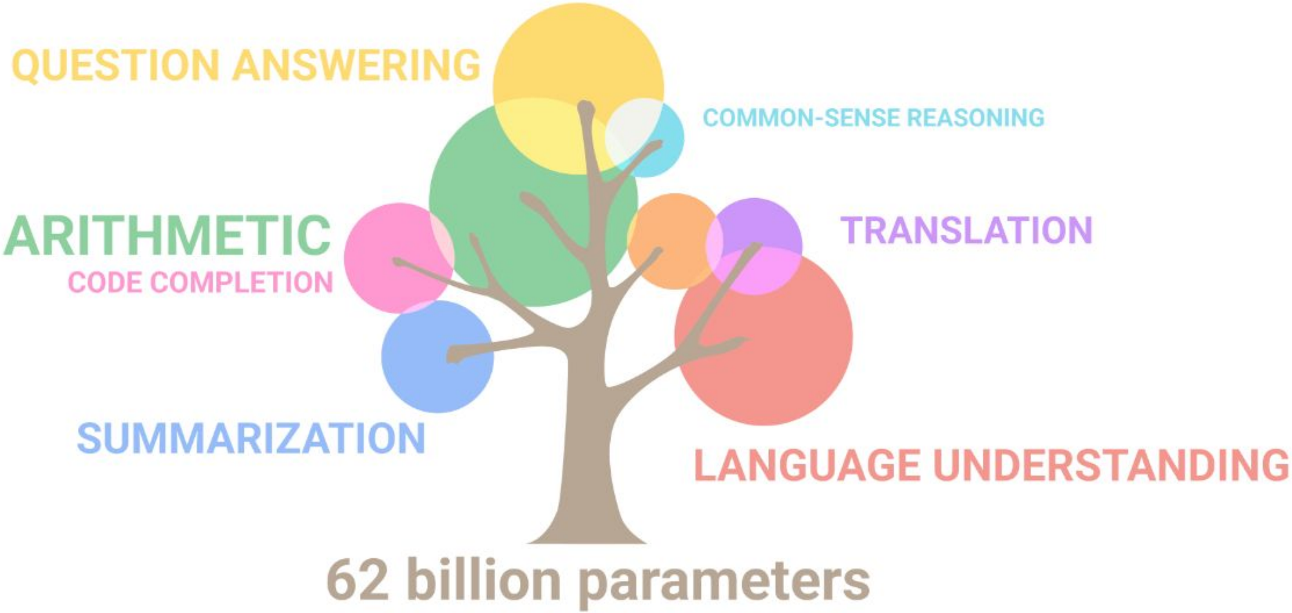
Emergent Abilities



Emergent Abilities



Emergent Abilities



Pre-trained Base

Generalist

vs

Fine-tuned Models

Specialists

RLHF:

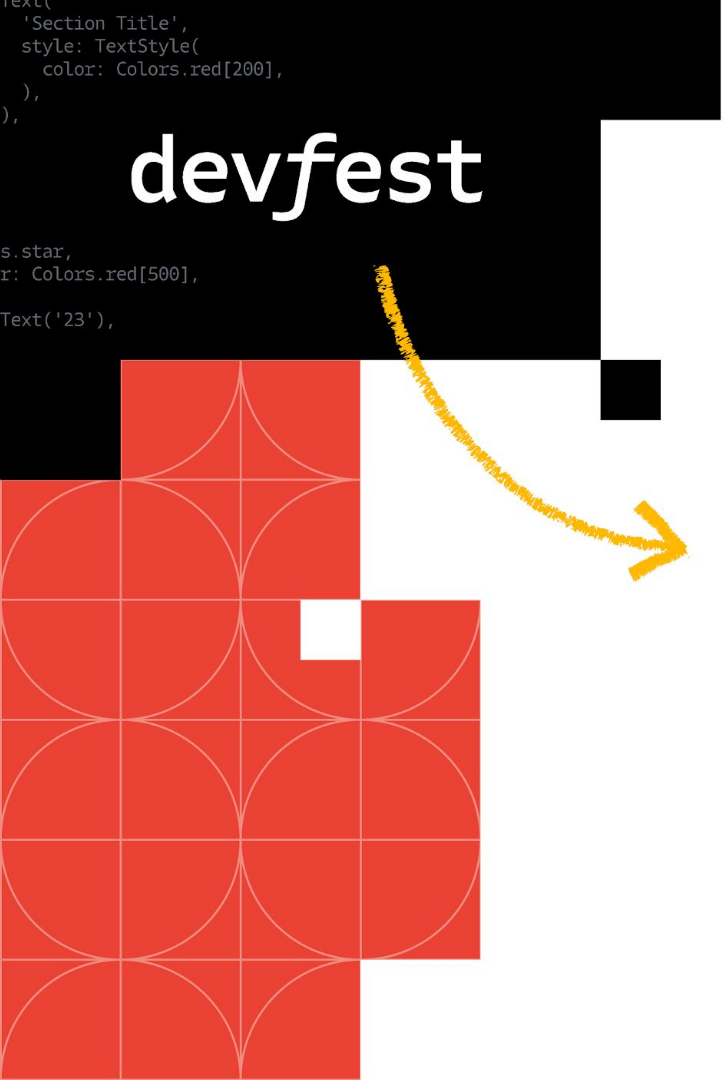
Reinforcement Learning from Human Feedback

Fine-tuned

- Follow Instructions
- Conversations

What’s Next?

- Multimodal
- Open-source
- Agents



```
Text(
  'Section Title',
  style: TextStyle(
    color: Colors.red[200],
  ),
),
```

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

Text('23'),
```

Google Developer Groups

Burnaby

LLMs in AI Safety

Charting a Safer Future

```
Text(
  '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

Some concrete examples of technical safety research...

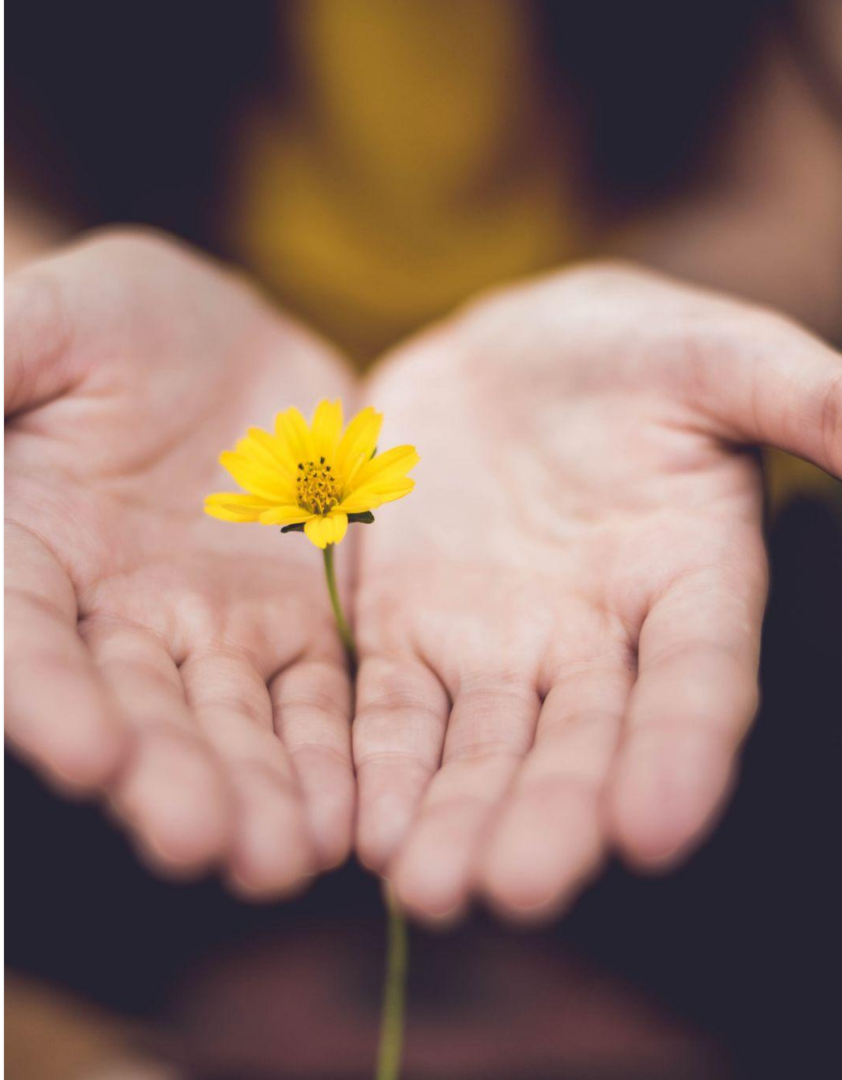
Value Alignment

- RLHF

  OpenAI's Alignment

- Sycophancy

  Agreeable but untrue



Evaluations

- Inverse Scaling

  Bigger isn't always better

- Jailbreaking

  Guardrails vs Attack suffix

Interpretability

- Meaningful Features

  Topics, i.e. legal text, space, time

- “Lie Detector”

  Hallucination & Deception

```
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

bit.ly/cheng2-slides-2023-11-25

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

Embrace Safely

ChengCheng Tan

ccstan99@gmail.com

cheng2-tan

@cheng2_tan

Google Developer Groups