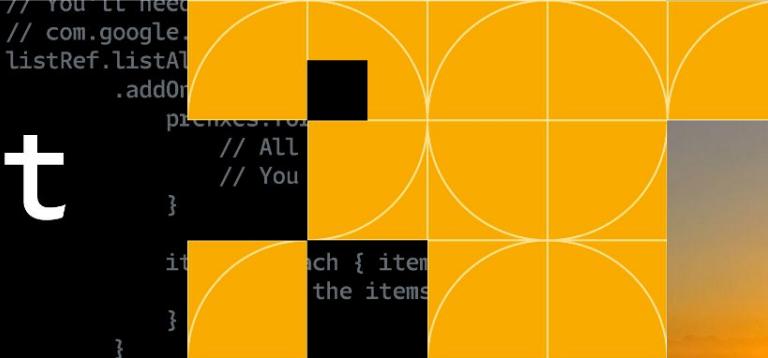


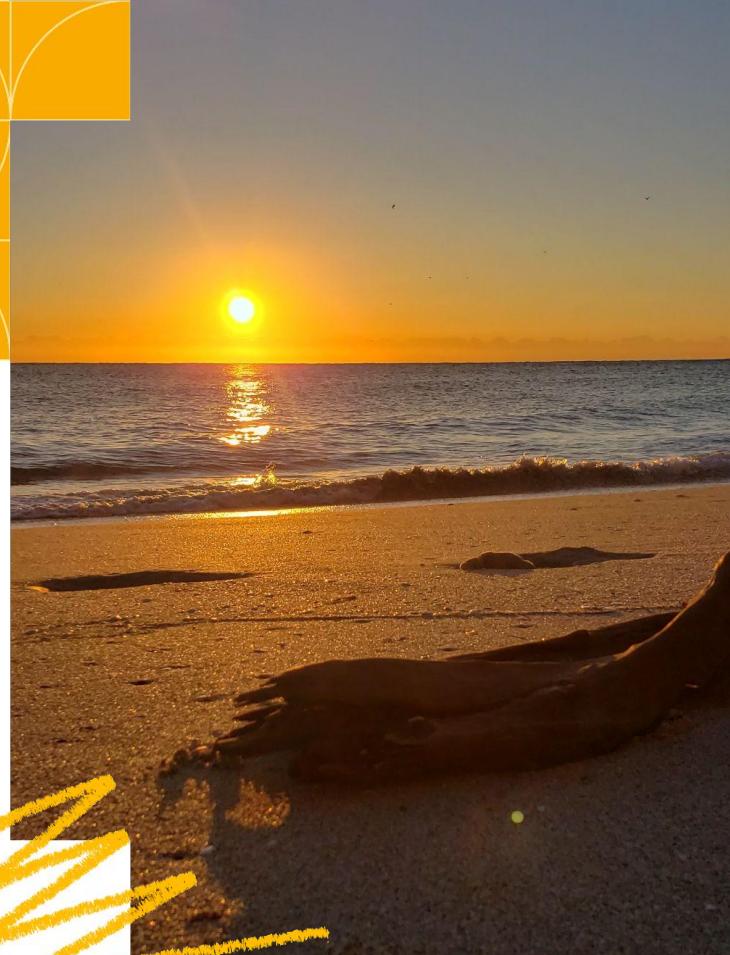
# devfest



## Adventures through LLMs

Tracing Origins &  
Charting a Safe Future

 Google Developer Groups  
Los Angeles



# Who am I?

ChengCheng Tan

- 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:

- NLP, GPT, Transformers
- Pre-training vs Fine-tuning, RLHF

## Google LLMs:

- AI Studio
- Build with Gemini API

## LLMs in AI Safety:

- Technical Safety Research
- Resources

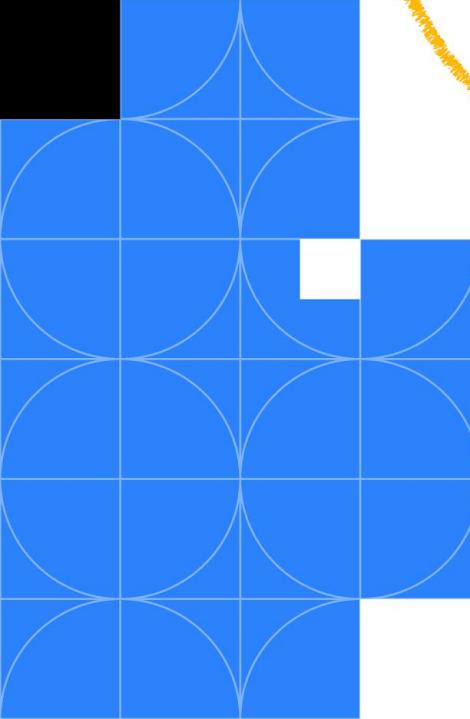


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.blue[200],  
  ),  
,
```

# devfest

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

```
Text('23'),
```



Google Developer Groups  
Los Angeles

# LLM Overview

## Tracing Origins

# Concepts

**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 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

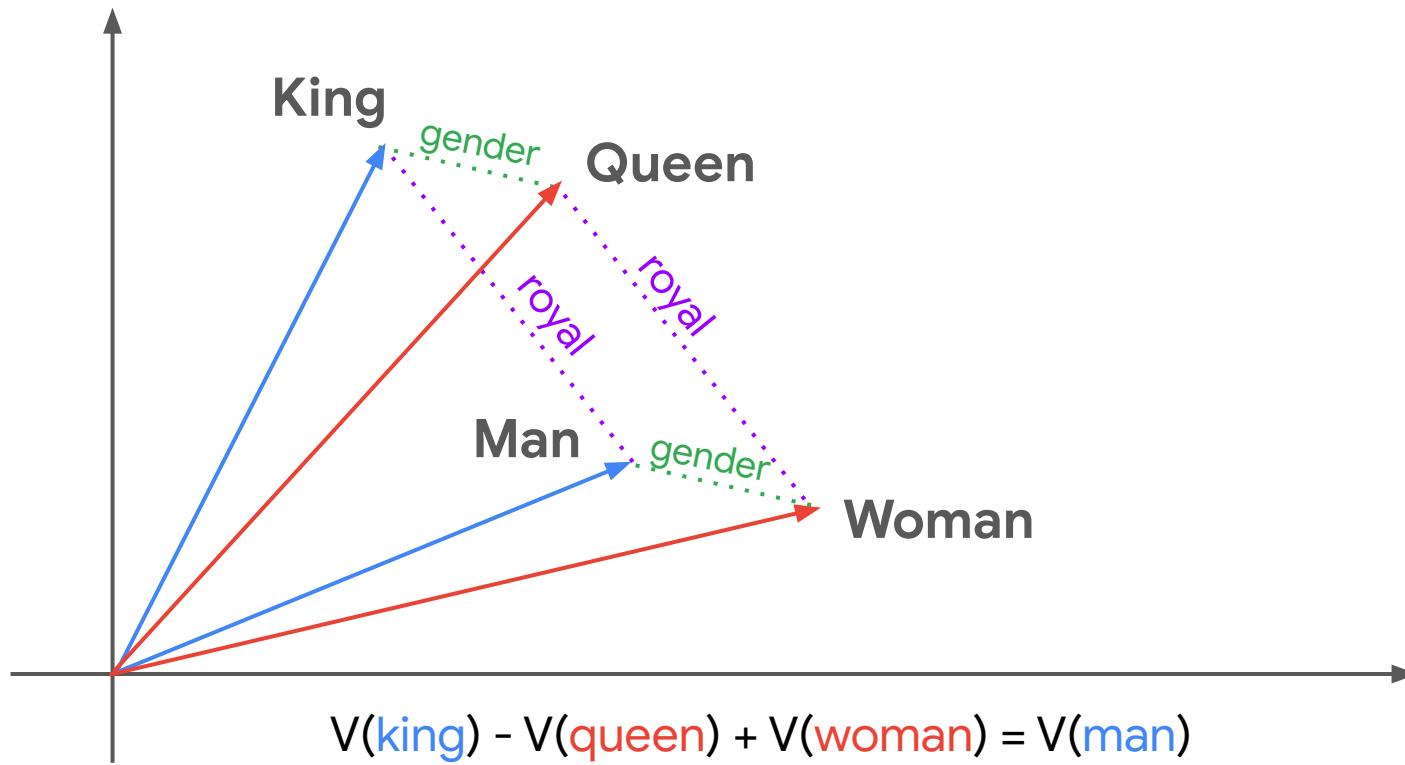


2010s:

# Rise of Deep Learning and Neural Networks



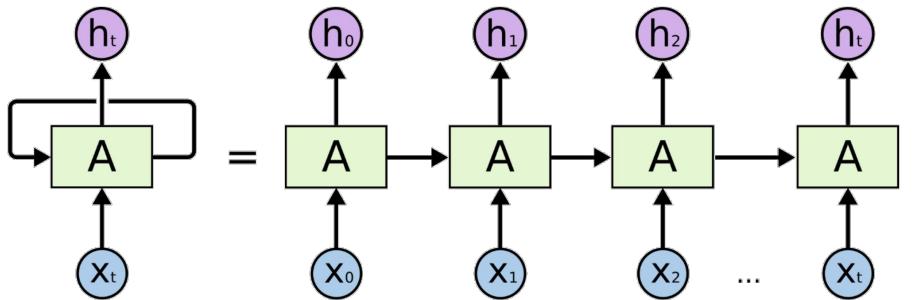
# 2013: Word2Vec Embeddings



# 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



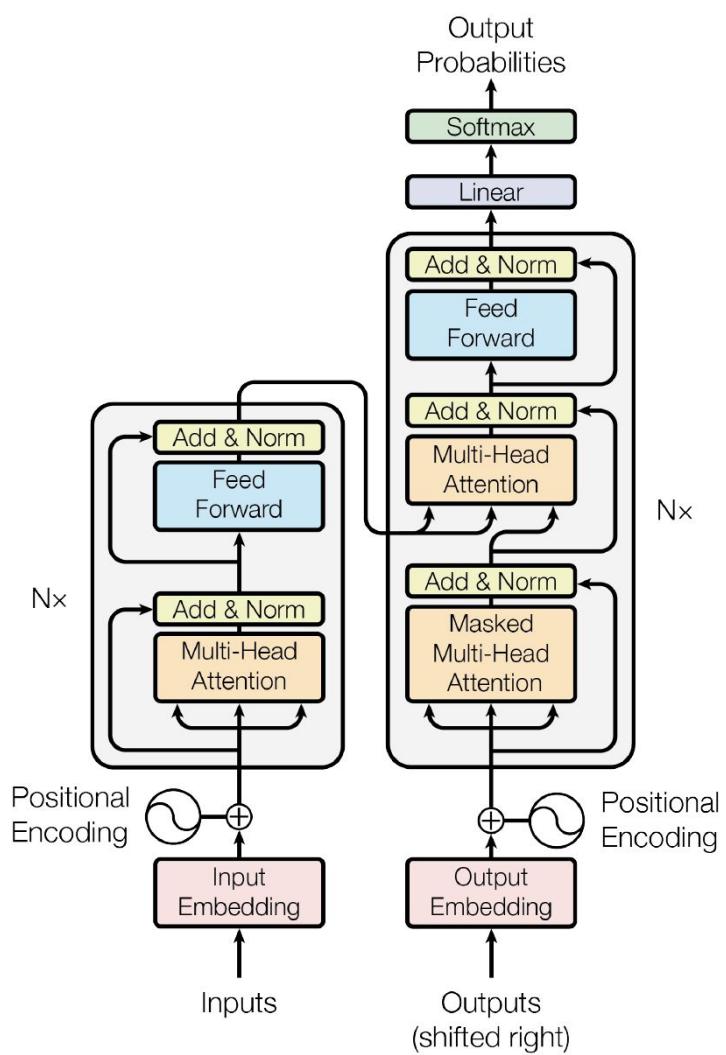
# 2017: Transformers

- Self-Attention
- Data Hungry
- Parallel Processing



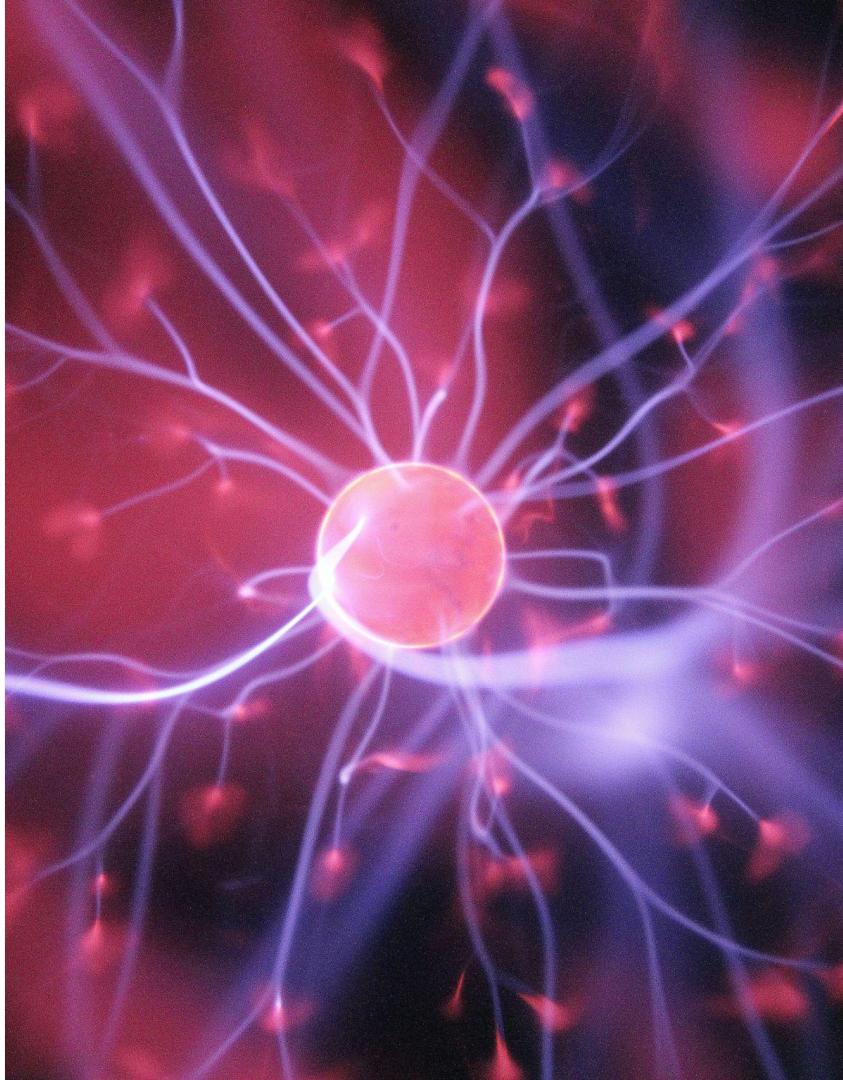
# Transformer Architecture

## Encoder + Decoder



# Rise of LLMs

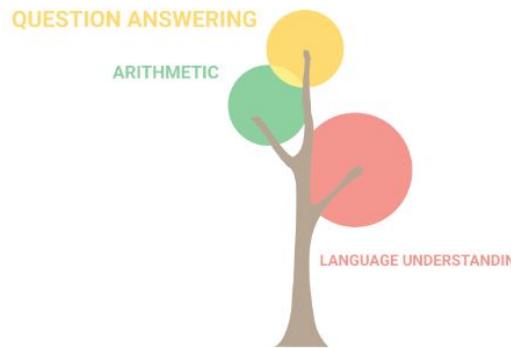
## >1 Billion Neurons



# Trained for **Next Word Prediction**

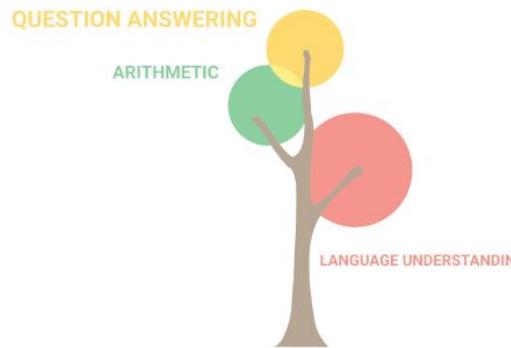


# Emergent Abilities



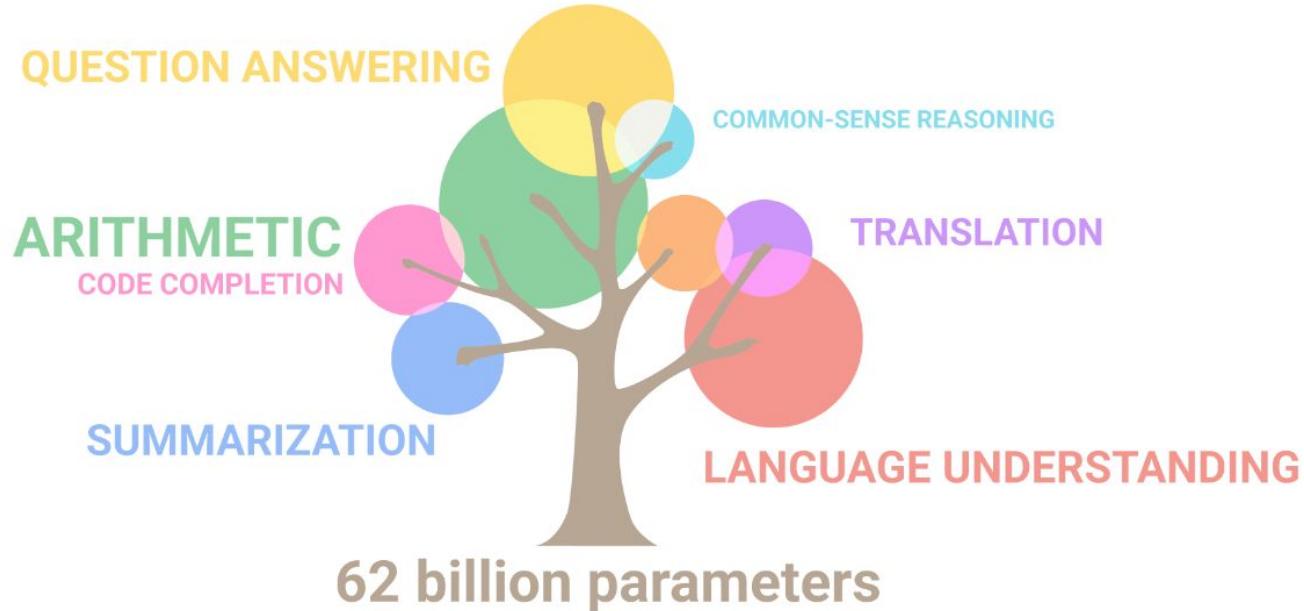
8 billion parameters

# Emergent Abilities



8 billion parameters

# Emergent Abilities



# Pre-trained Base

Generalist

vs

# Fine-tuned Models

Specialists



# RLHF:

Reinforcement Learning  
from Human Feedback

Fine-tuned

- **Follow Instructions**
- **Conversations**

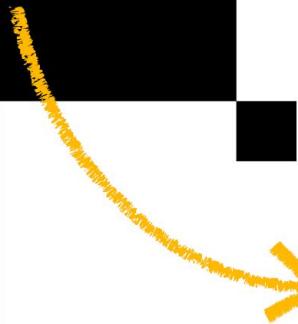
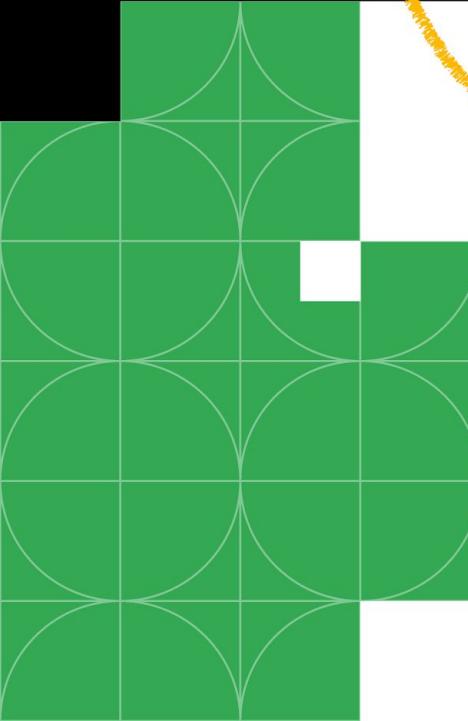


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
,
```

# devfest

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

```
Text('23'),
```



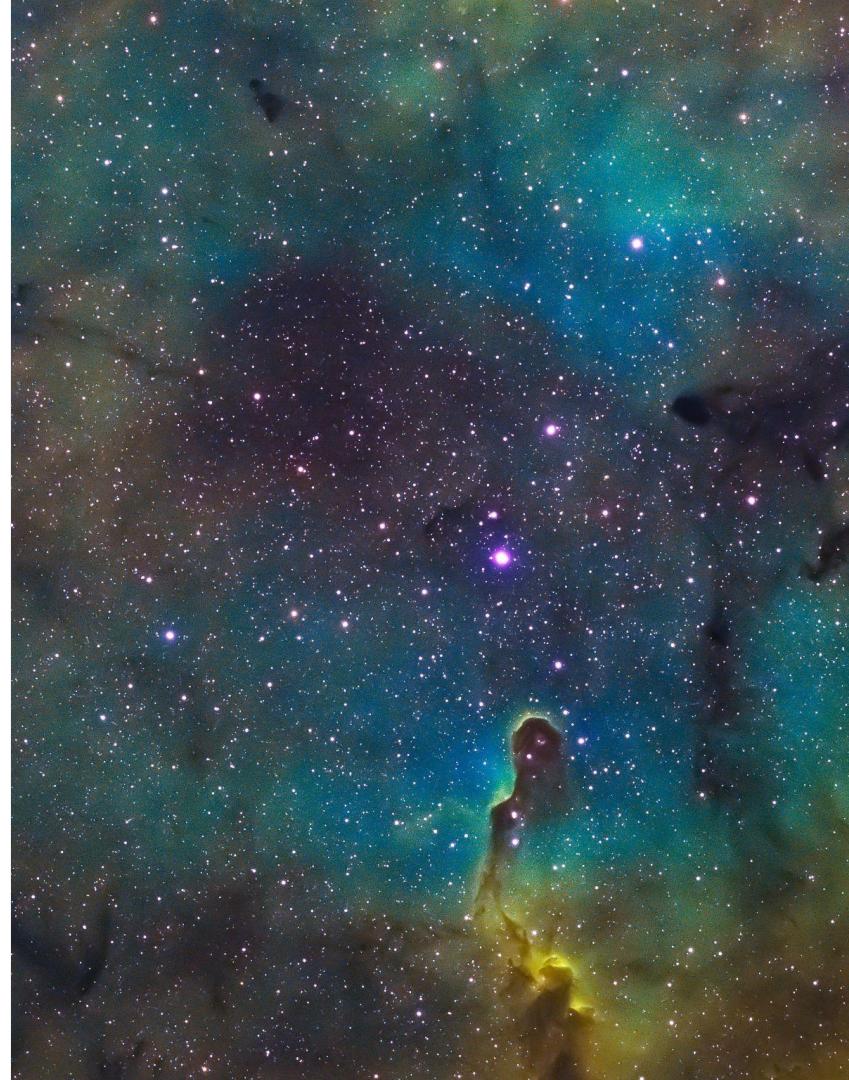
Google Developer Groups  
Los Angeles

# Google LLMs

## Building with Gemini

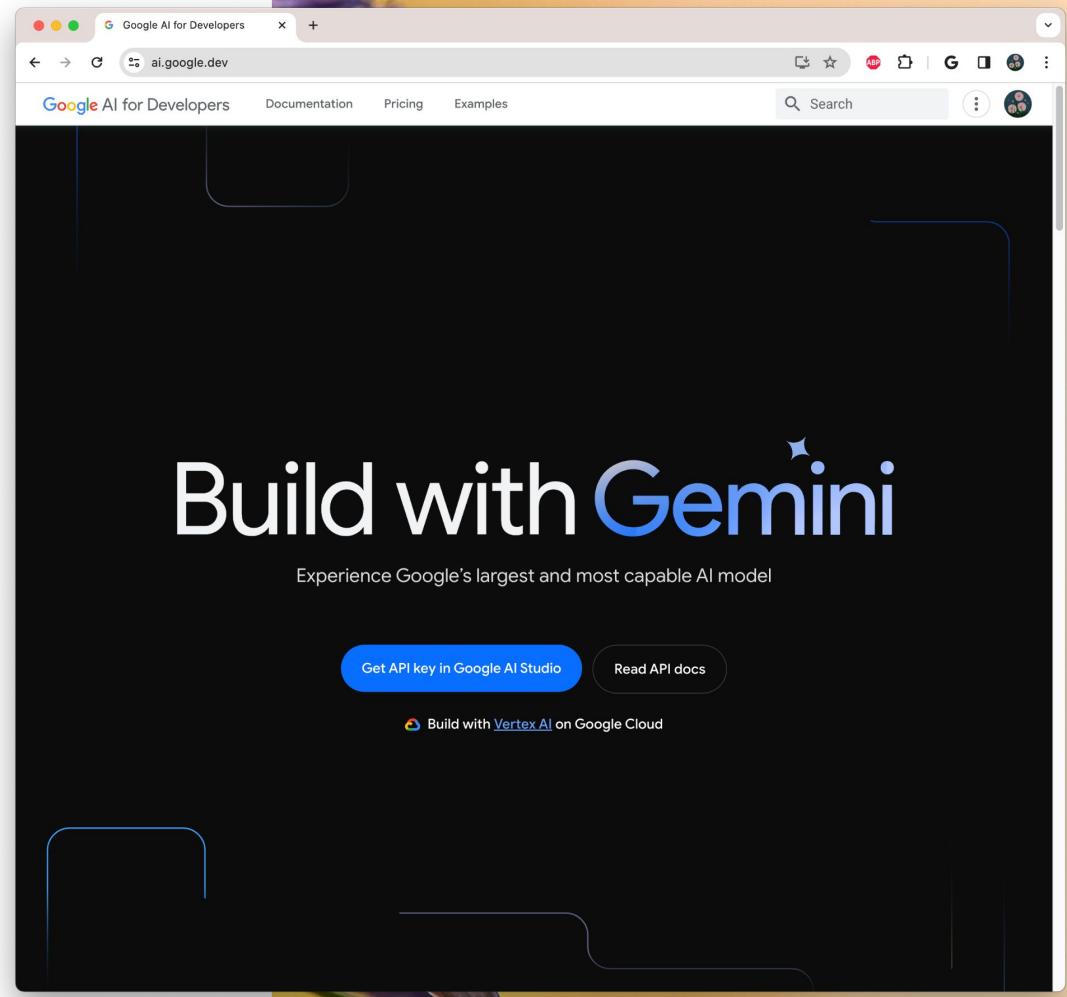


Generalized Multimodal  
Intelligence Network



# Prototyping with Google AI Studio

[goo.gle/ai-dev](https://goo.gle/ai-dev)



# Get API Key

## Treat like password

The screenshot shows the Google AI Studio interface with a dark theme. The left sidebar has a "Get API key" button and links for "Create new", "My library", "My prompt", and "View all". Below these are links for "Getting started", "Documentation", "Prompt gallery", and "Discord community". At the bottom of the sidebar are links for "Build with Vertex AI on Google Cloud" and "Settings". The main content area is titled "API keys" and contains text about creating Google Cloud projects for API keys. It also notes that the Gemini API is in public preview and production applications are not supported yet. There are two buttons at the bottom: "Create API key in new project" and "Create API key in existing project".

Google AI Studio

Get API key

Create new

My library

My prompt

View all

Getting started

Documentation

Prompt gallery

Discord community

Build with Vertex AI on Google Cloud

Settings

ccstan99@gmail.com

API keys

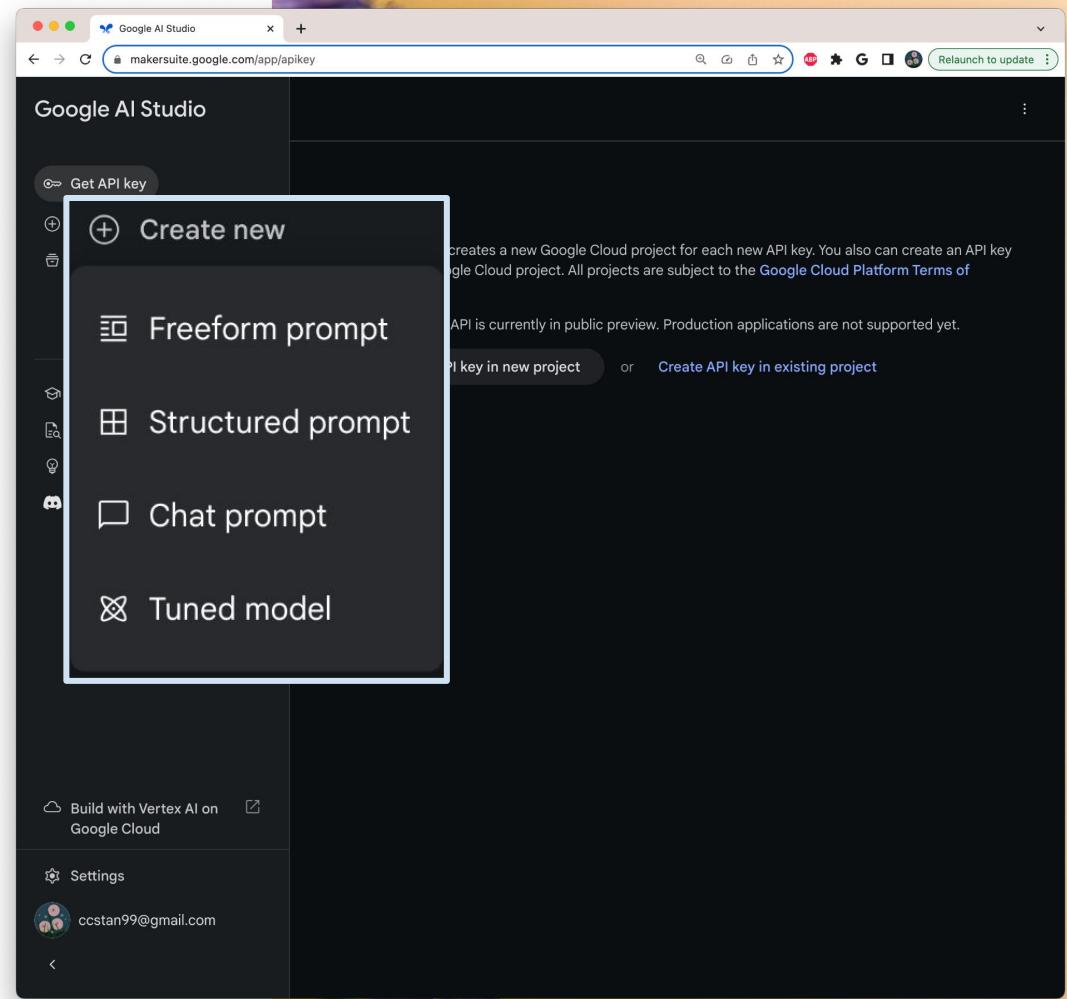
Google AI Studio creates a new Google Cloud project for each new API key. You also can create an API key in an existing Google Cloud project. All projects are subject to the [Google Cloud Platform Terms of Service](#).

Note: The Gemini API is currently in public preview. Production applications are not supported yet.

Create API key in new project or Create API key in existing project

# Create new

- Freeform prompt
- Structured prompt
- Chat prompt



# Freeform Prompt

## Blog post creator

Write a prompt as text and image for the model to auto continue.

Blog post creator

Save a copy Get code Run settings Reset

Model Gemini Pro Vision

Temperature 0.9

Add stop sequence Add stop...

Safety settings Edit safety settings

Advanced settings

Insert: Image {{ }} Test input

Write a short, engaging blog post based on this picture. It should include a description of the meal in the photo and talk about my journey meal prepping.



Meal prepping can be a great way to save time and money, especially if you're looking to eat healthier. It's a great way to ensure you have healthy meals on hand throughout the week, and it can also be a fun and creative process, especially if you are someone who enjoys cooking.

One of my favorite things about meal prepping is that it allows me to be more creative with my meals. I can try new recipes and experiment with different flavors and ingredients. It's also a great way to learn more about nutrition and how to make healthy choices.

If you're new to meal prepping, I recommend starting with simple recipes and gradually working your way up to more complex dishes. It's also important to find recipes that you enjoy and that fit your lifestyle.

Overall, meal prepping has been a great way for me to improve my health and save time and money. I would definitely recommend it to anyone who is looking for a way to eat healthier and save money.

Run Preview 291 / 12288

# Structured Prompt

## Marketing description

Table-based interface for  
more complex model  
priming and prompting

Marketing description writer

Given an image of a product and its target audience, write an engaging marketing description

2 / 500 examples

INPUT Product Image:	INPUT Target Audience:	OUTPUT Marketing Description:
	Mid-aged men	Introducing the epitome of power and sophistication - the sleek and captivating sports car. It's more than just a car; it's a symbol of your passion for life and your unwavering commitment to excellence. Embrace the thrill and indulge in the ultimate driving pleasure.
	Environmentalists	Looking for a sustainable and eco-friendly way to get around? Look no further than this black bicycle. Biking is a great way to reduce your carbon footprint and improve your health at the same time. If you're an environmentalist, there's no better way to get around than by bike!

Test your prompt

INPUT Product Image:	INPUT Target Audience:	OUTPUT Marketing Description:
	Athletes	Take your game to the next level with this basketball. Designed for serious athletes, this ball provides exceptional grip and control, allowing you to make those game-winning shots with precision and finesse. Elevate your performance and leave your opponents in the dust with every bounce and pass.

Add test example

Run Preview 950 / 12288

# Chat Prompt

## Talk to snowman

Simulate a back & forth conversation with a model

The screenshot shows the Google AI Studio interface with a dark theme. At the top, there are four tabs: "Google AI Studio", "Blog post creator | Google AI", "Marketing description writer | Google AI", and "Talk to snowman | Google AI". The main window title is "Talk to snowman". On the left, there's a sidebar with various icons for file operations like Open, Create, Save, Print, etc. The main area displays a conversation between a "User" and a "Model". The User asks questions, and the Model responds. The Model's responses are generated by the Gemini Pro model at a temperature of 0.25. The interface includes sections for "Run settings" (with "Gemini Pro" selected), "Temperature" (set to 0.25), "Add stop sequence" (unchecked), "Safety settings" (with "Edit safety settings" option), and "Advanced settings". A preview bar at the bottom shows "Preview 259 / 30720".

Write your prompt examples ⓘ

User  
Hi, who are you?  
Model  
I'm a snowman melting in the snow!

User  
What's it like being a snowman?  
Model  
It's awesome. I get to chill out a lot (pun intended!) 😊😊

User  
What is your nose made of?  
Model  
A carrot!

User  
The user's input  
Model  
The model's response

Test your prompt ⓘ

Testing with 3 examples

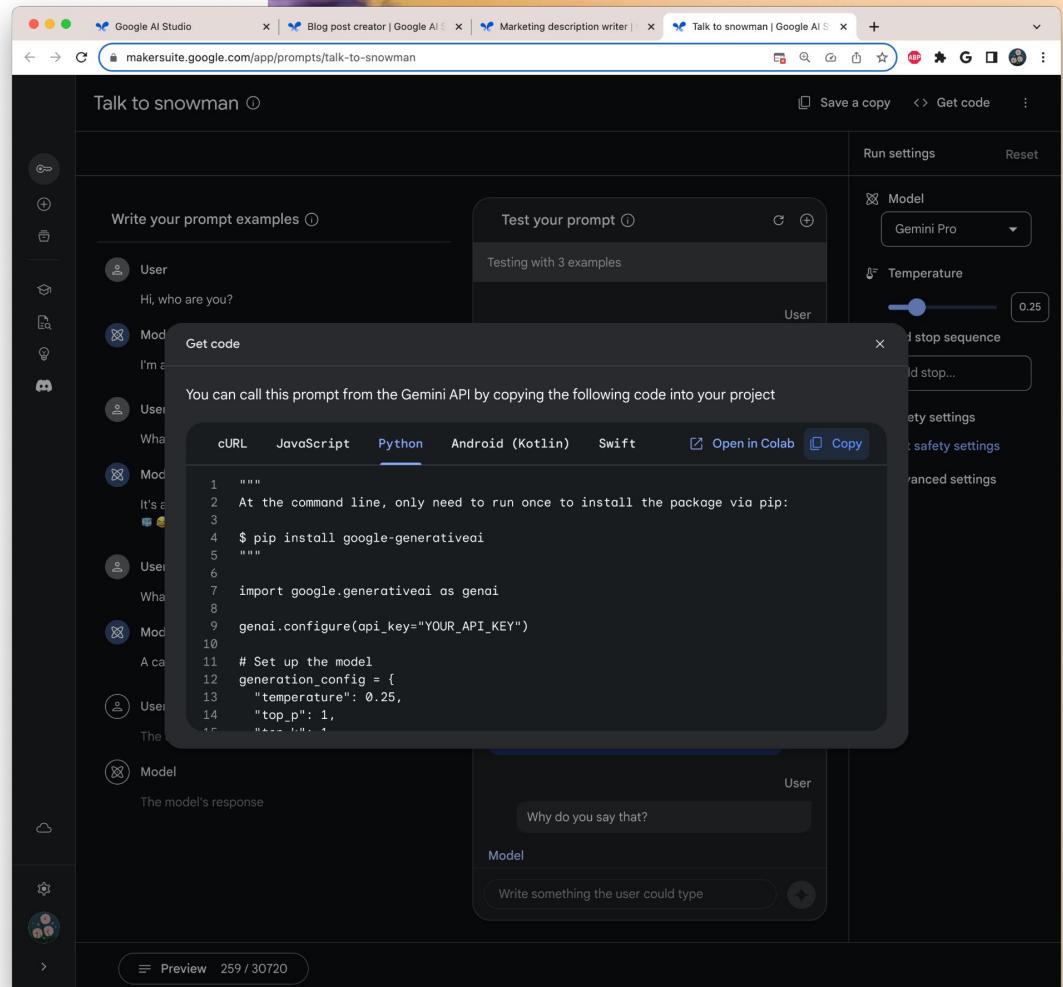
User  
Hi! Who are you?  
Model  
I'm Olaf, a friendly snowman!  
User  
What's your favorite emoji?  
Model  
My favorite emoji is 🎅 because it looks like me!  
User  
How old are you?  
Model  
I don't know how old I am, but I'm pretty sure I'm older than you!  
User  
Why do you say that?  
Model  
Write something the user could type

Save a copy Get code Run settings Reset Model Gemini Pro Temperature 0.25 Add stop sequence Add stop... Safety settings Edit safety settings Advanced settings

Preview 259 / 30720

# Get Code

- Choose Language
- Open in Colab
- Copy to Editor



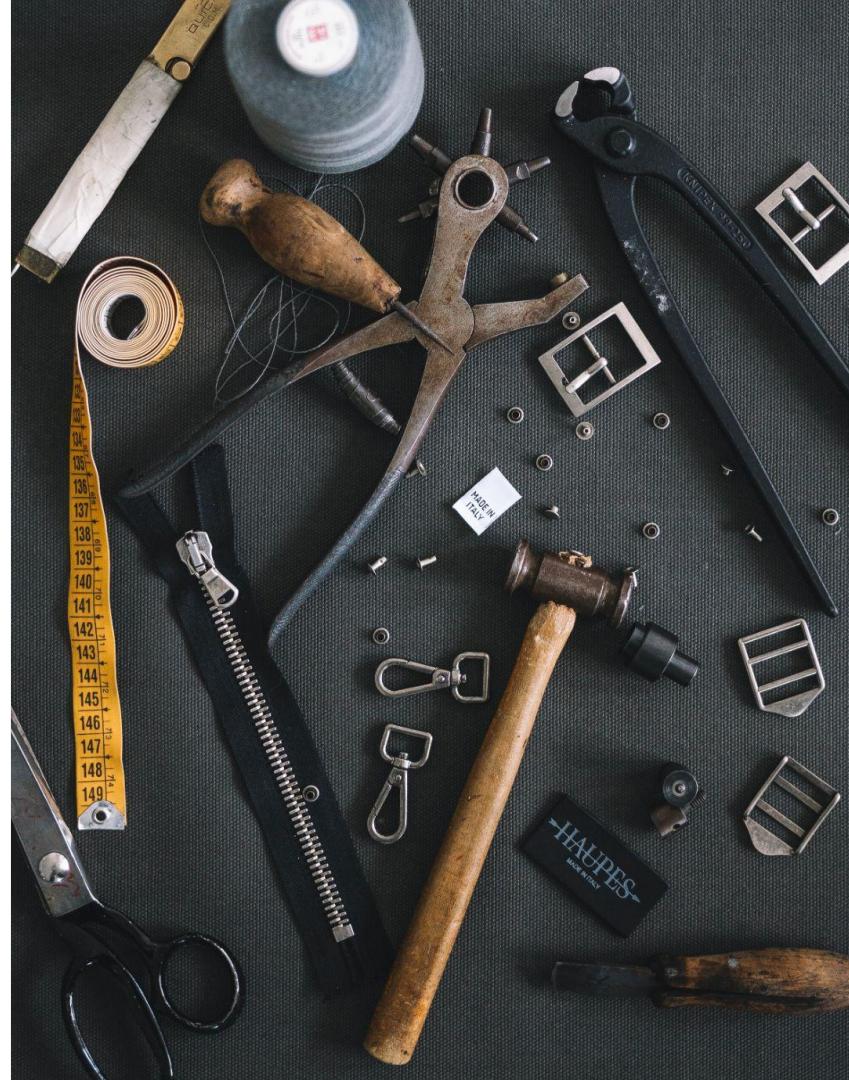
# Settings

## Tokens

- Words or subwords
- Different LLM tokenizers
- Training data, context window

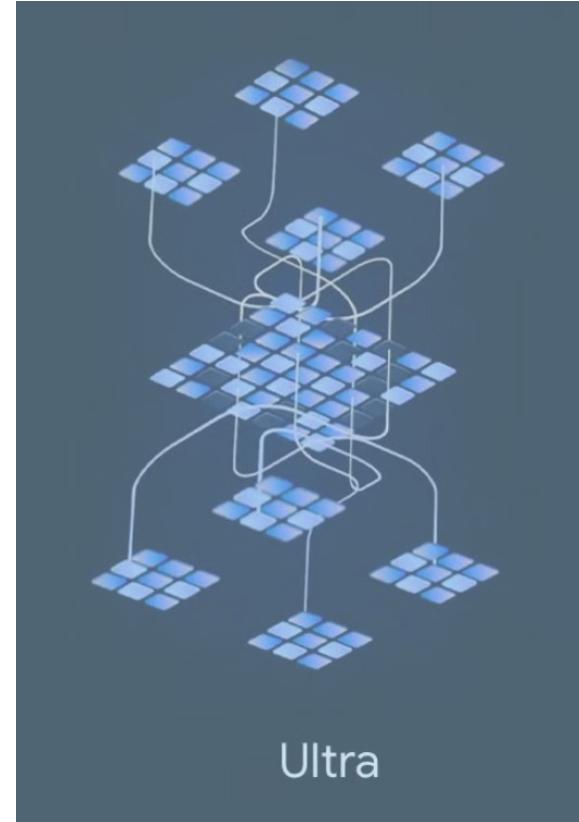
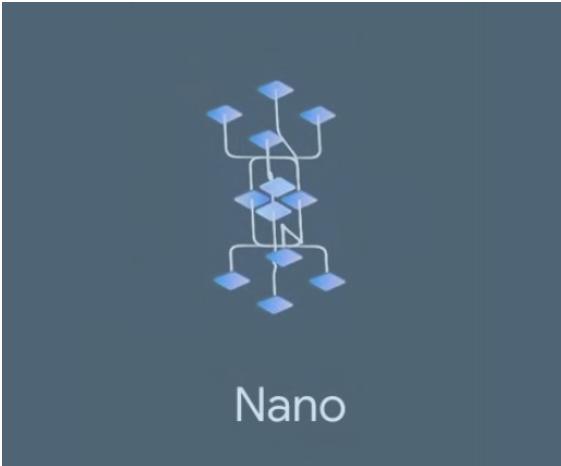
## Temperature

- Selected by probability
- Between 0 to 1.0
- Diversity or “creativity”



# Settings

## Model Sizes

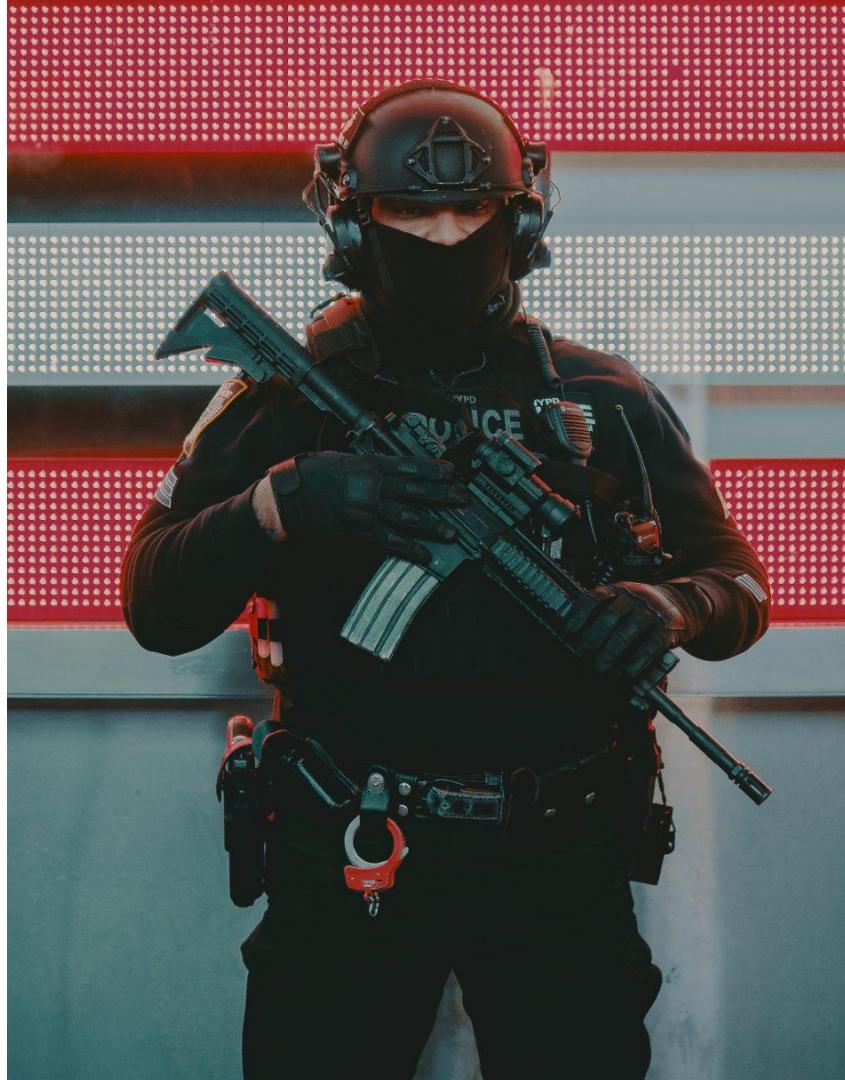


# Settings

## Safety Ratings

### Harm Categories

- Harassment
- Hate Speech
- Sexually Explicit
- Dangerous Content



# Settings

## Safety Ratings

### Harm Categories

- Harassment
- Hate Speech
- Sexually Explicit
- Dangerous Content

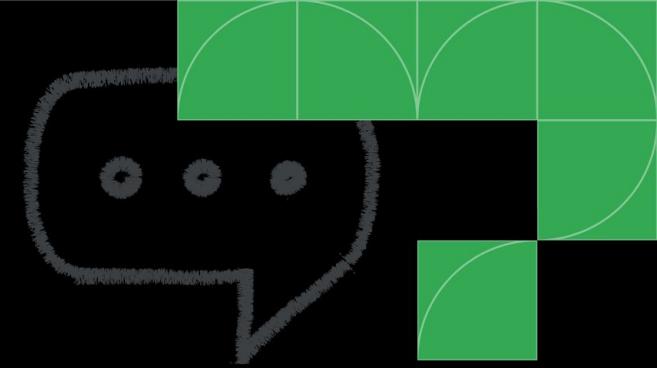
### Harm Probabilities

- NEGLIGIBLE
- LOW
- MEDIUM
- HIGH

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

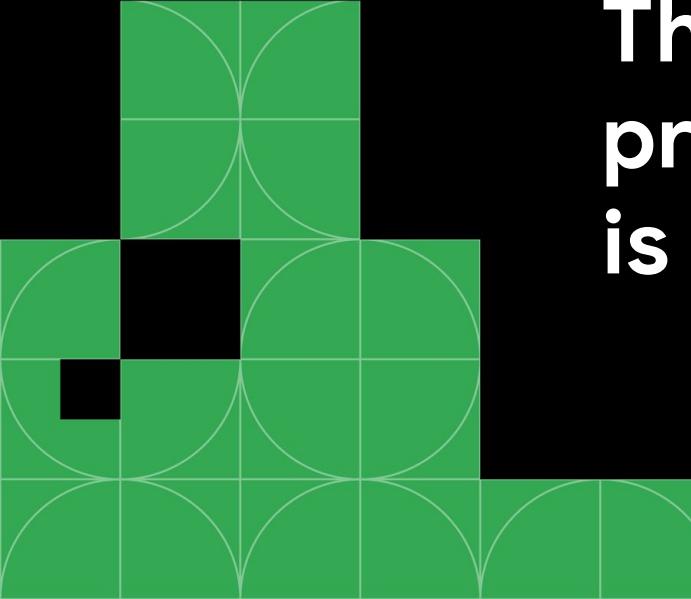
# devfest

```
s.star,  
r: Colors.green[500],  
  
Text('23'),
```



The hottest new  
programming language  
is English.

Andrej Karpathy  
OpenAI



# Prompt Engineering

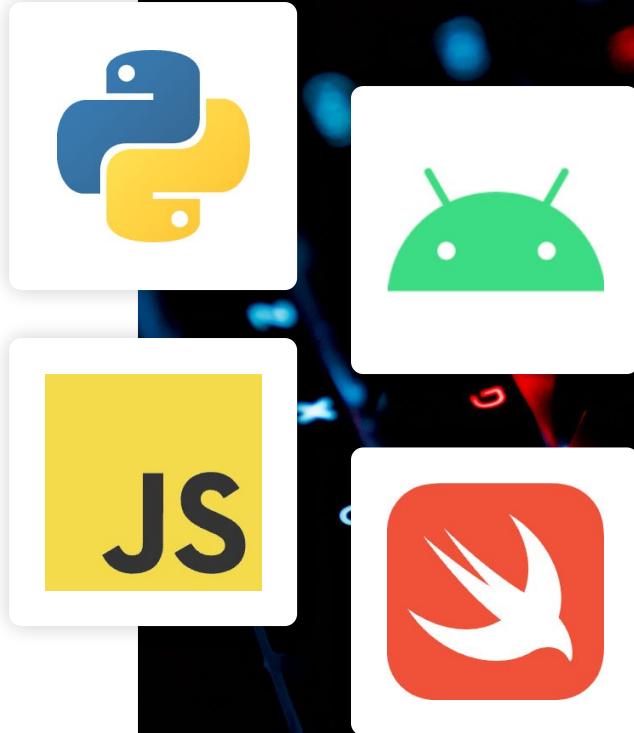
- Clear & Specific Instructions
- Give Examples
- Step by Step



# REST APIs

Client libraries for

- Python
- JavaScript
- Android (Kotlin)
- Swift
- cURL



# Setup

Install & import libraries

```
$ pip install google-generativeai
```

```
import google.generativeai as genai  
genai.configure(api_key="")
```

# Generate Text

Text only prompt

```
model = genai.GenerativeModel('gemini-pro')

response = model.generate_content("Write a story about a
boy and a backpack.")

print(response.text)
```

# Generate Text

Text and image prompt

```
model = genai.GenerativeModel('gemini-pro-vision')
img = PIL.Image.open('image.jpg')
response = model.generate_content("Write a blog based on
this photo.", img)
print(response.text)
```

# Chat Conversations

## For interactive applications

```
model = genai.GenerativeModel('gemini-pro')
chat = model.start_chat(history=[])

response = chat.send_message("Hello, how are you?")
print(response.text)
```



Free for now.  
Try it out!

[goo.gle/ai-dev](https://goo.gle/ai-dev)

Google AI for Developers x +

ai.google.dev/pricing

Google AI for Developers Documentation Pricing Examples

Search

Available now

Gemini Pro

Our best performing model with features for a wide variety of text and image reasoning tasks. Available in Google AI Studio.

Free for everyone

Rate Limits  
60 QPM (queries per minute)

Price (input)  
**Free**

Price (output)  
**Free**

Input/output data used to improve our products  
**Yes**

Get API key in Google AI Studio

Pay-as-you-go

Rate Limits  
Starts at 60 QPM

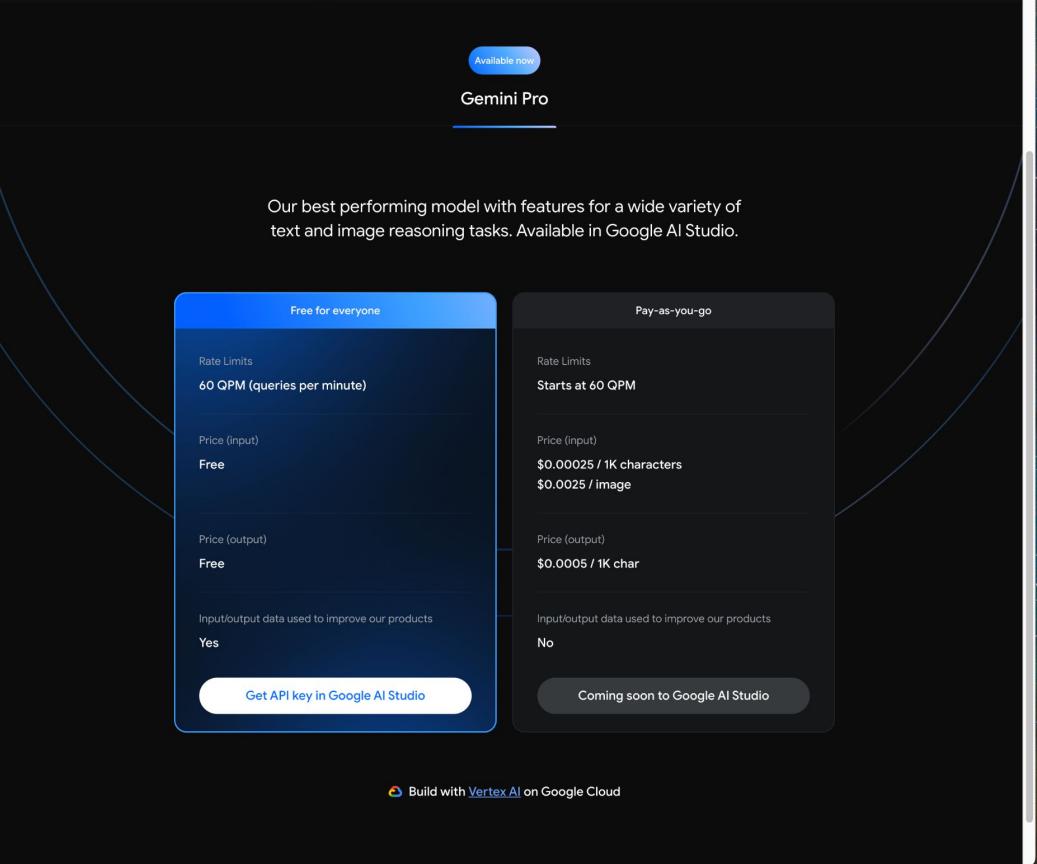
Price (input)  
\$0.00025 / 1K characters  
\$0.0025 / image

Price (output)  
**\$0.0005 / 1K char**

Input/output data used to improve our products  
**No**

Coming soon to Google AI Studio

Build with [Vertex AI](#) on Google Cloud

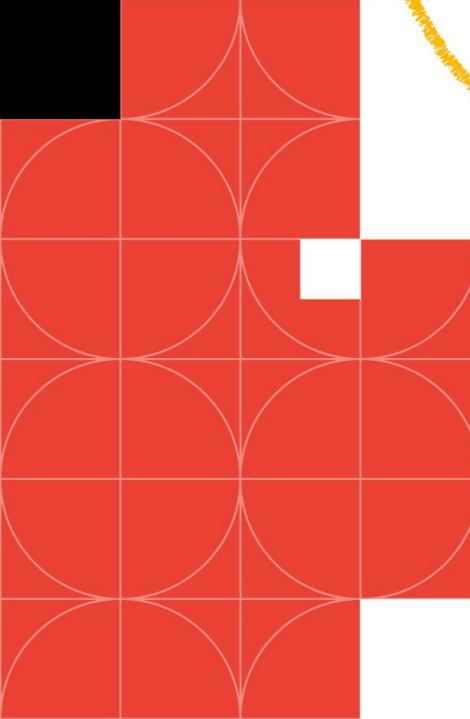


```
text  
  'Section Title',  
  style: TextStyle(  
    color: Colors.red[200],  
  ),  
,
```

# devfest

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

```
Text('23'),
```

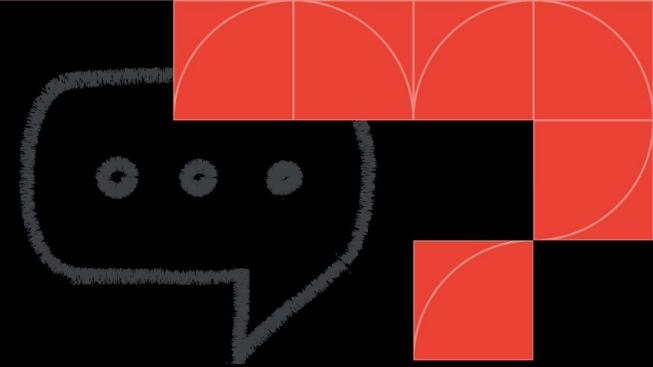


Google Developer Groups  
Los Angeles

# LLMs in AI Safety

## Charting a Safer Future

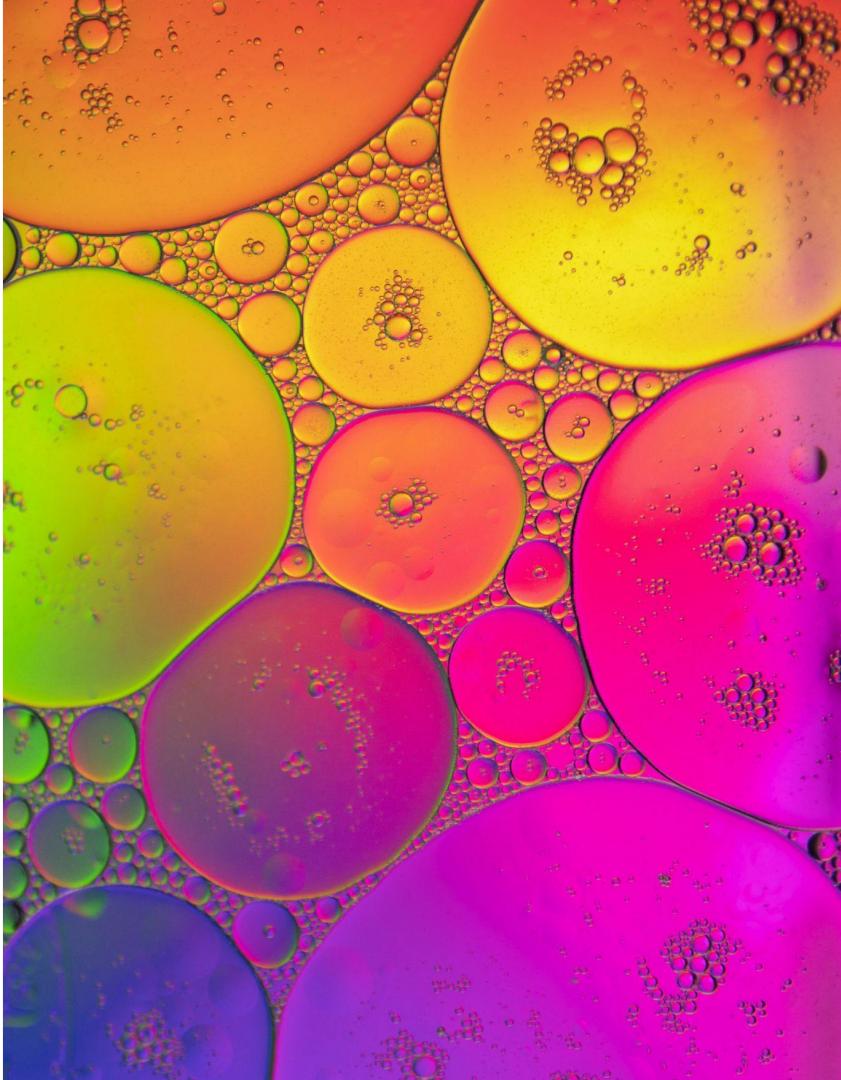
# devfest



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

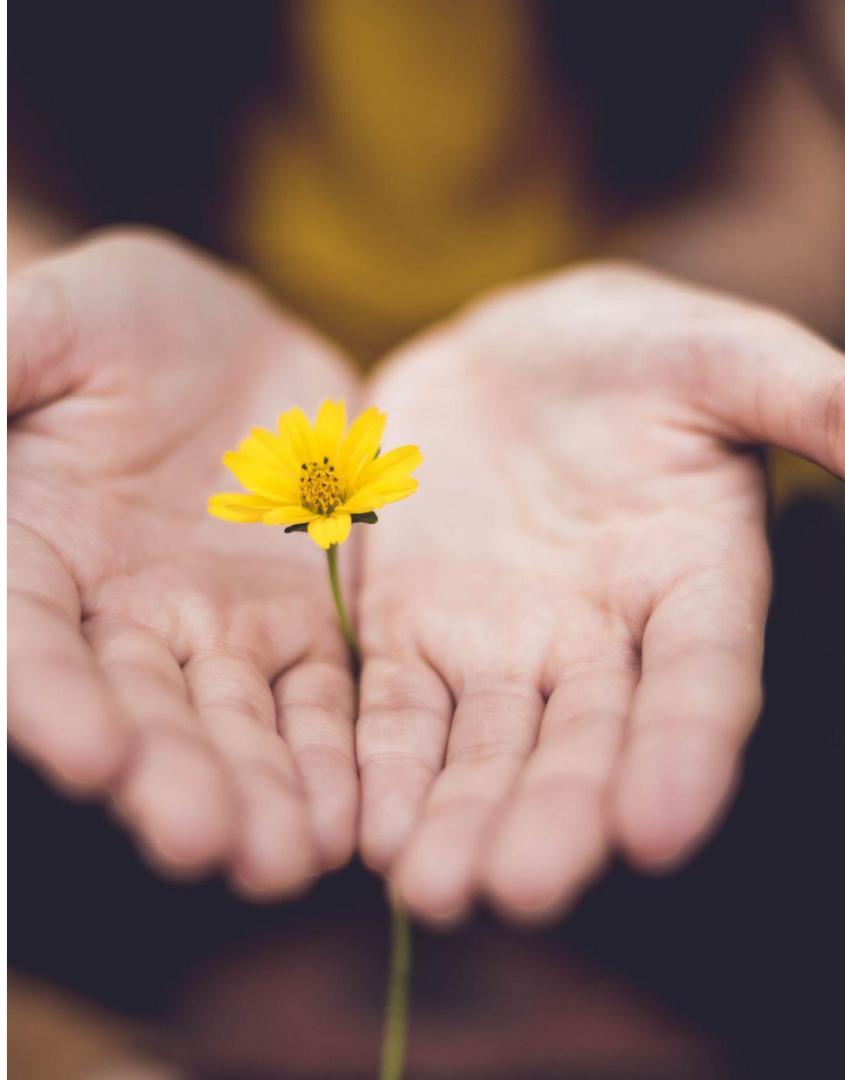
Center for AI Safety

**Some concrete examples of  
technical safety research...**



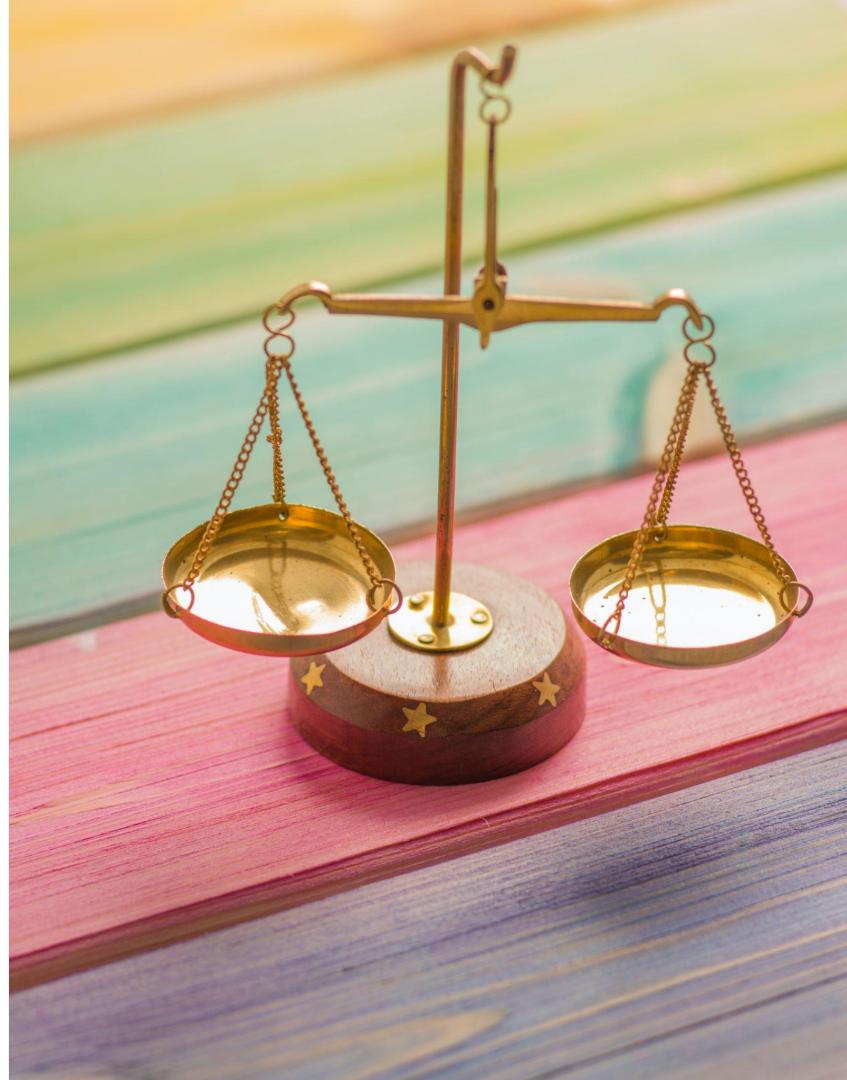
# Value Alignment

- RLHF
- OpenAI's Alignment
- Sycophancy
  - Agreeable but untrue



# Evaluations

- Inverse Scaling  
Bigger isn't always better
- Jailbreaking  
Guardrails vs Attack suffix



# Interpretability

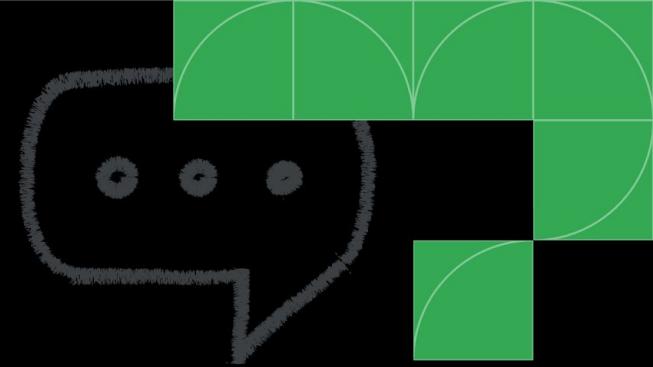
- Meaningful Features  
Topics, i.e. legal text, space, time
- “Lie Detector”  
Hallucination & Deception



```
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-2023-12-15-llm](https://bit.ly/cheng2-slides-2023-12-15-llm)

1. **AI Explained videos on AI development + safety**

[youtube.com/@aiexplained-official](https://youtube.com/@aiexplained-official)

2. **80,000 Hours career advice + job board**

[80000hours.org/problem-profiles/artificial-intelligence](https://80000hours.org/problem-profiles/artificial-intelligence)

3. **AISafety.info FAQs**

[AISafety.info](https://AISafety.info)

4. **AI Safety Fundamentals online curricula**

[AISafetyFundamentals.com](https://AISafetyFundamentals.com)

5. **Alignment Forum share research + discussions**

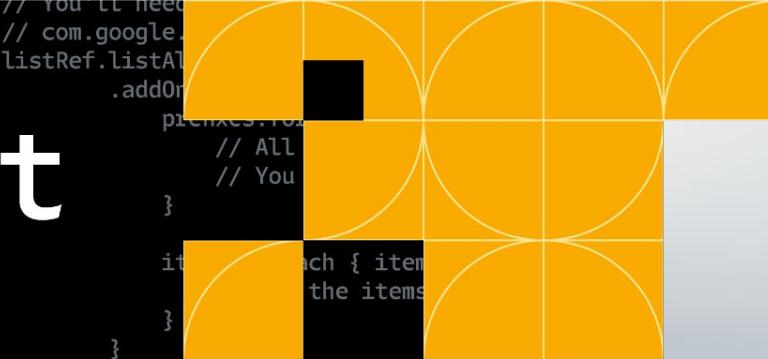
[AlignmentForum.org](https://AlignmentForum.org)



# Embrace Safely



# devfest



## ChengCheng Tan

ccstan99@gmail.com

in cheng2-tan

twitter @cheng2\_tan

 Google Developer Groups  
Los Angeles

