

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],
  ),
),
),
```

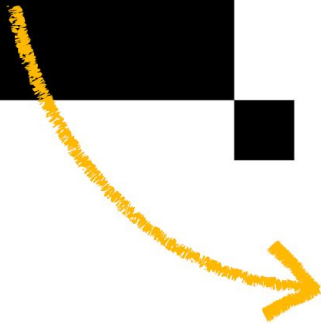
devfest

```
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**

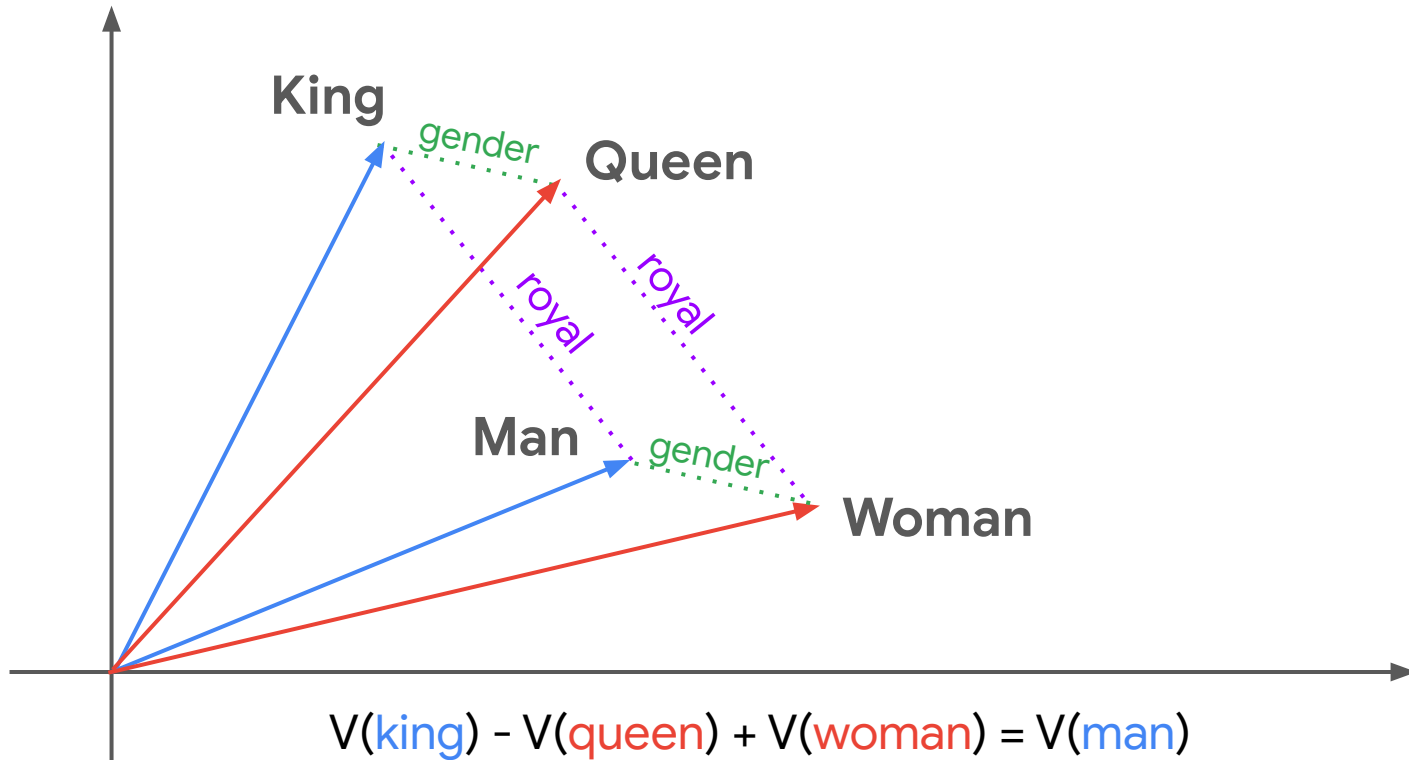
J.R. Firth, Linguist



2010s:
**Rise of Deep Learning
and Neural Networks**



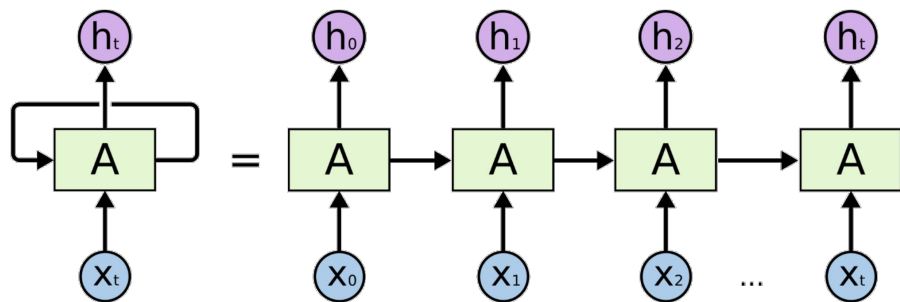
2013: Word2Vec Embeddings



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



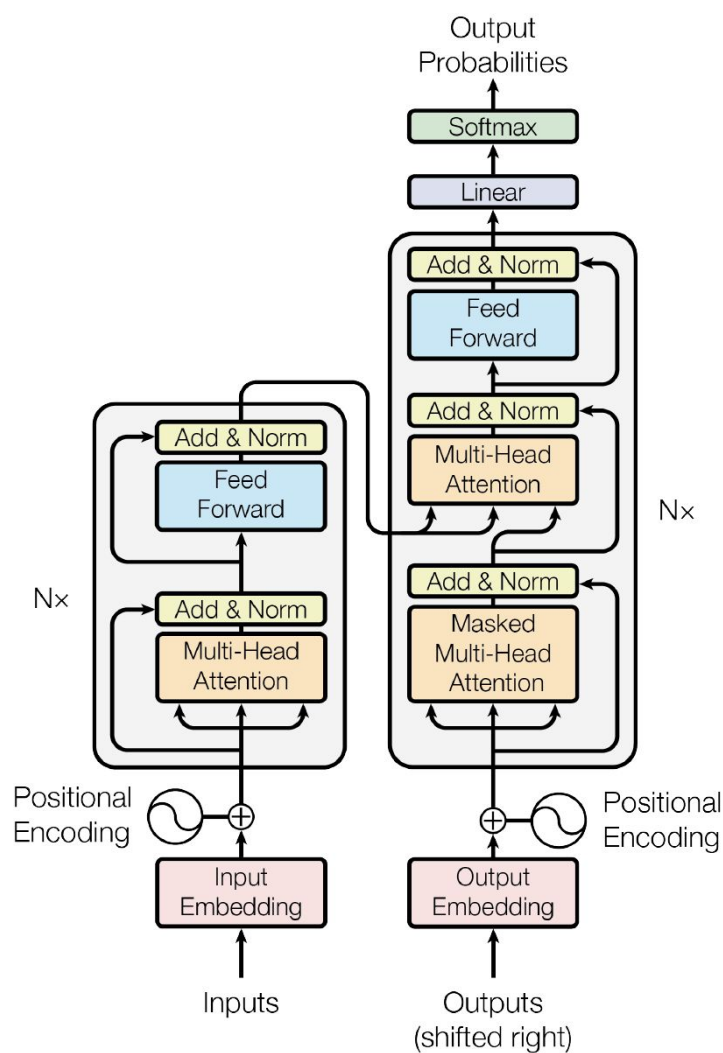
2017: Transformers

- Self-Attention
- Data Hungry
- Parallel Processing



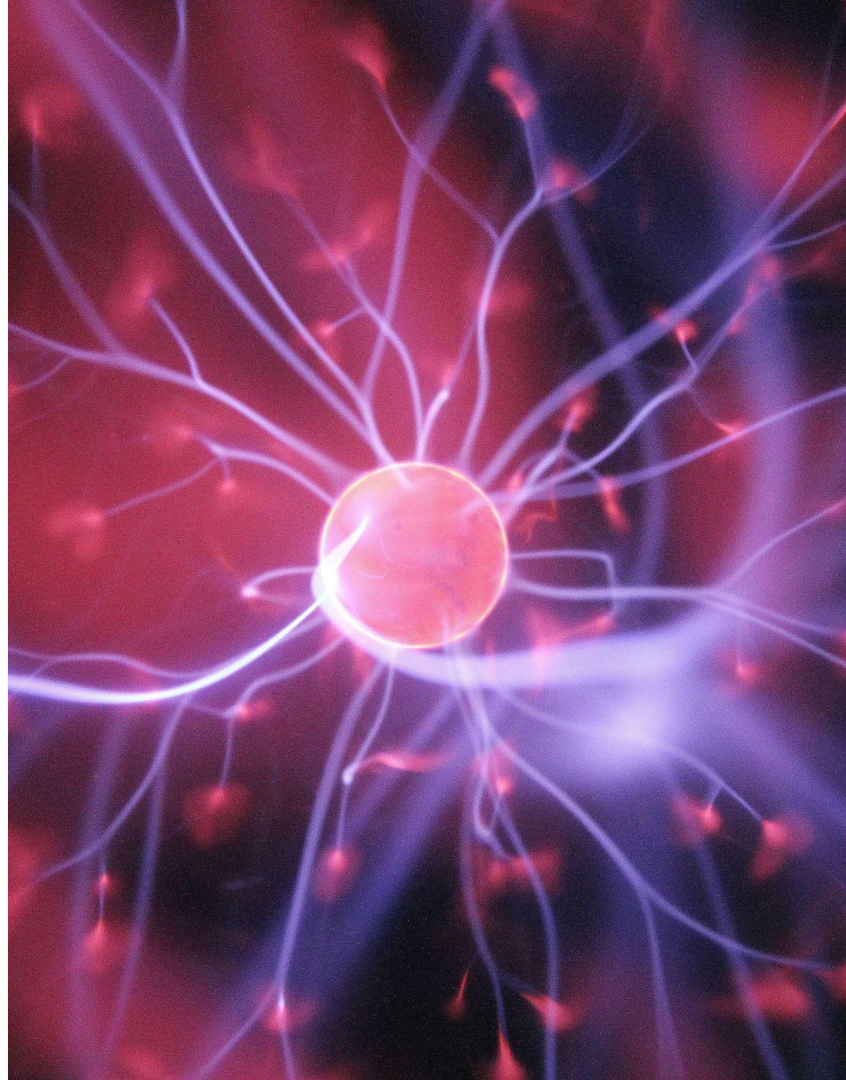
Transformer Architecture

Encoder + Decoder

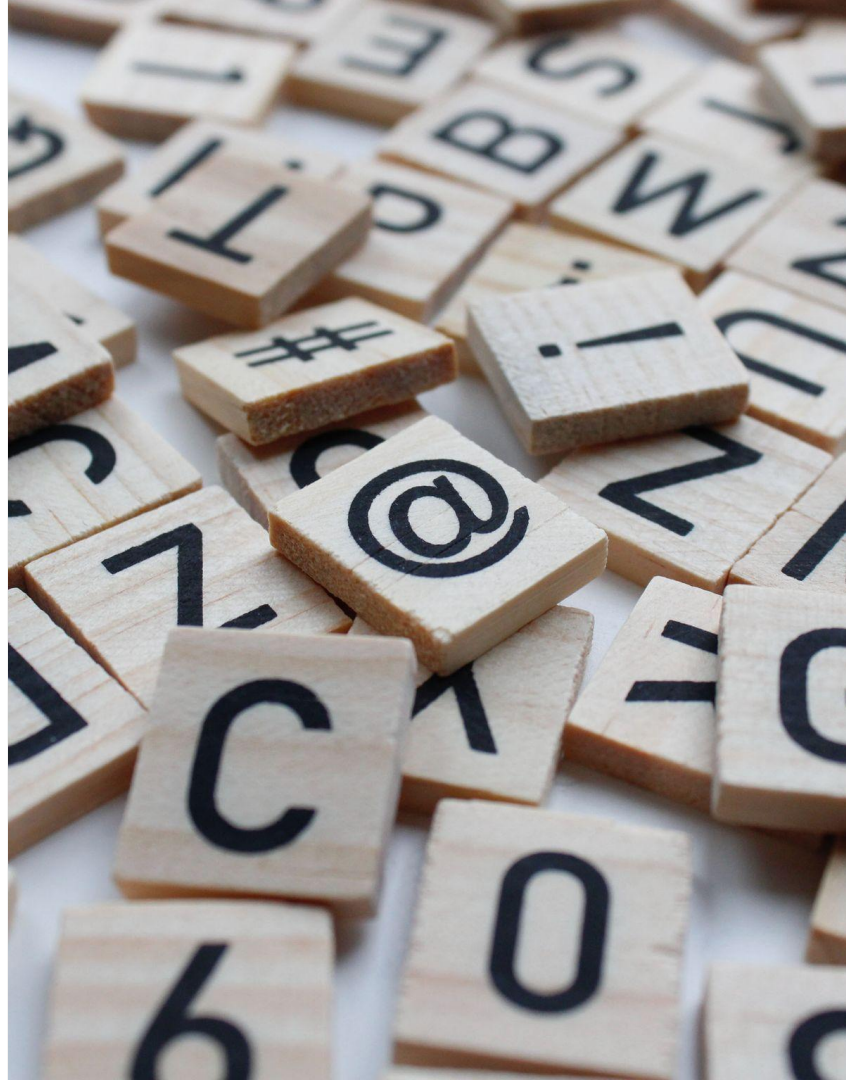


Rise of LLMs

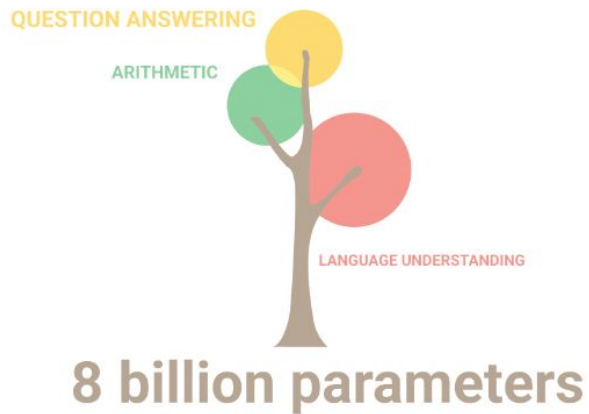
>1 Billion Neurons



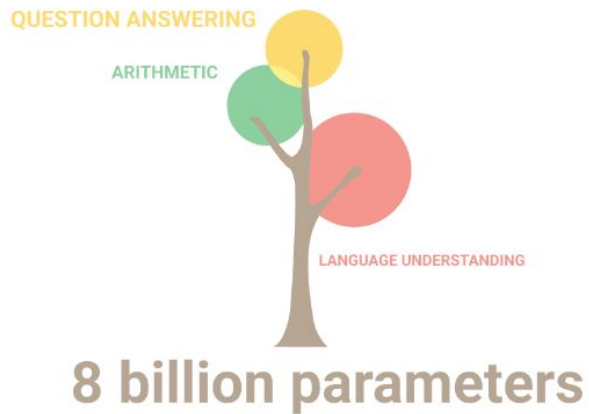
Trained for **Next Word Prediction**



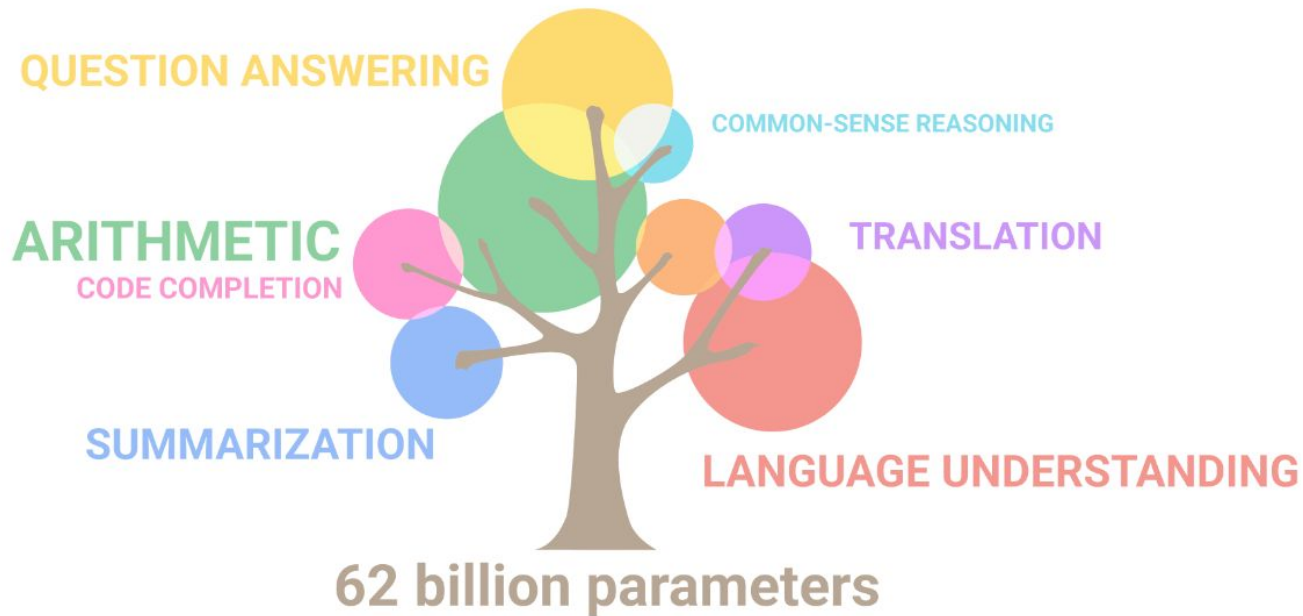
Emergent Abilities



Emergent Abilities



Emergent Abilities



Pre-trained Base

Generalist

VS

Fine-tuned Models

Specialists



RLHF:

Reinforcement Learning
from Human Feedback

Fine-tuned

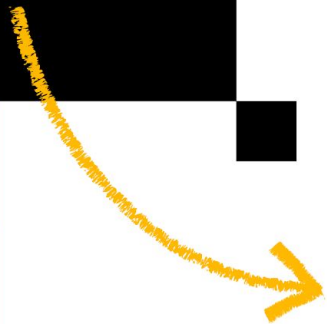
- **Follow Instructions**
- **Conversations**




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

devfest

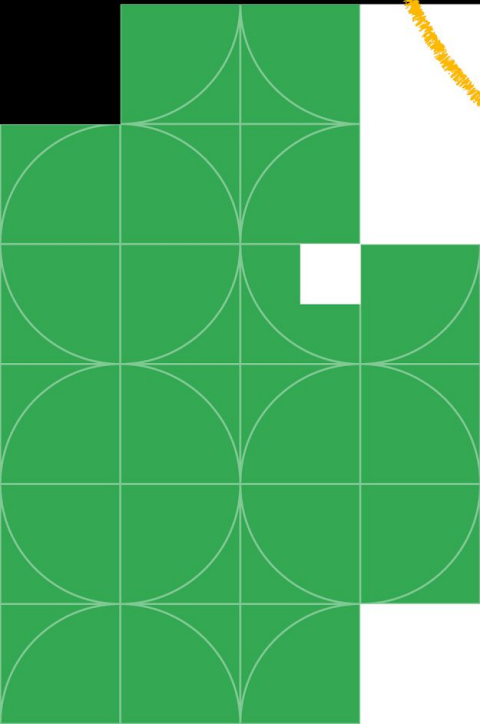
```
s.star,
r: Colors.green[500],
Text('23'),
```



 Google Developer Groups
Los Angeles

Google LLMs

Building with Gemini



Gemini

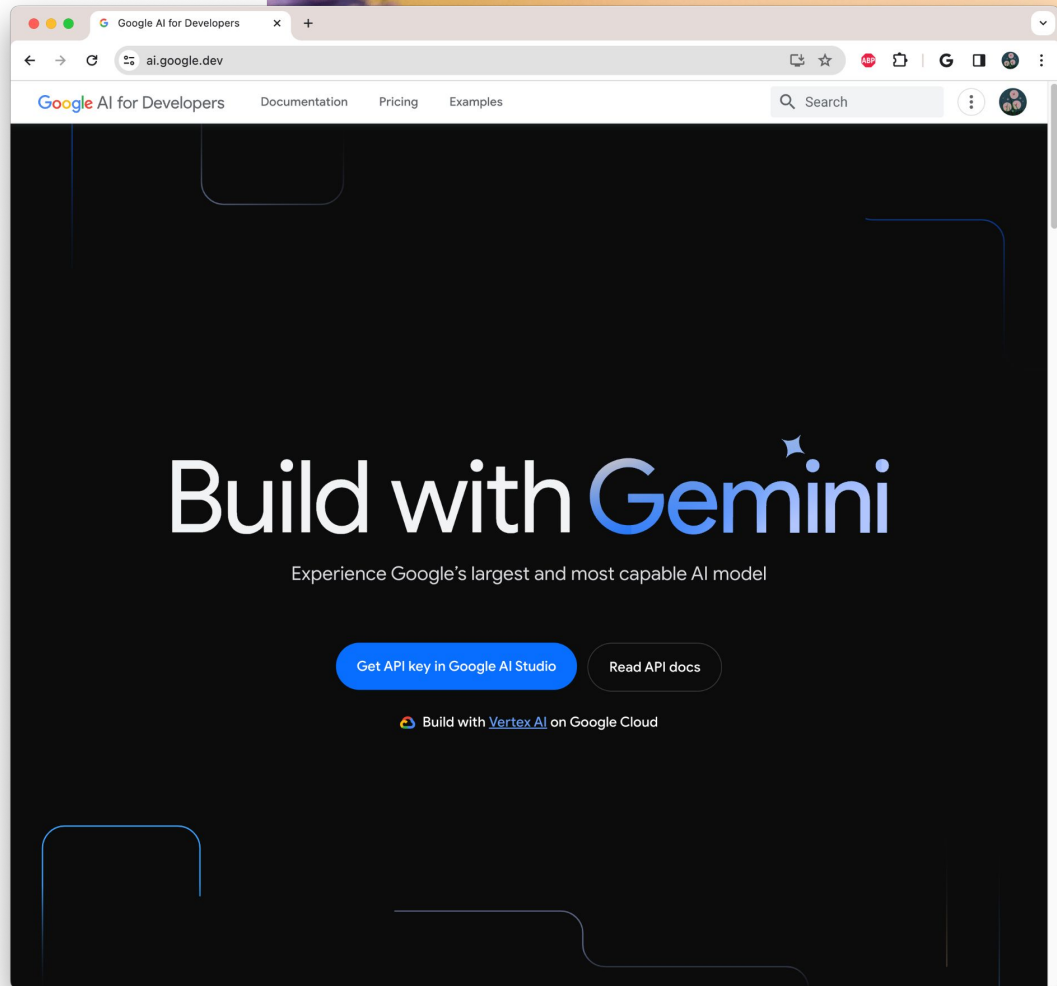
Generalized Multimodal
Intelligence Network



Prototyping with Google AI Studio

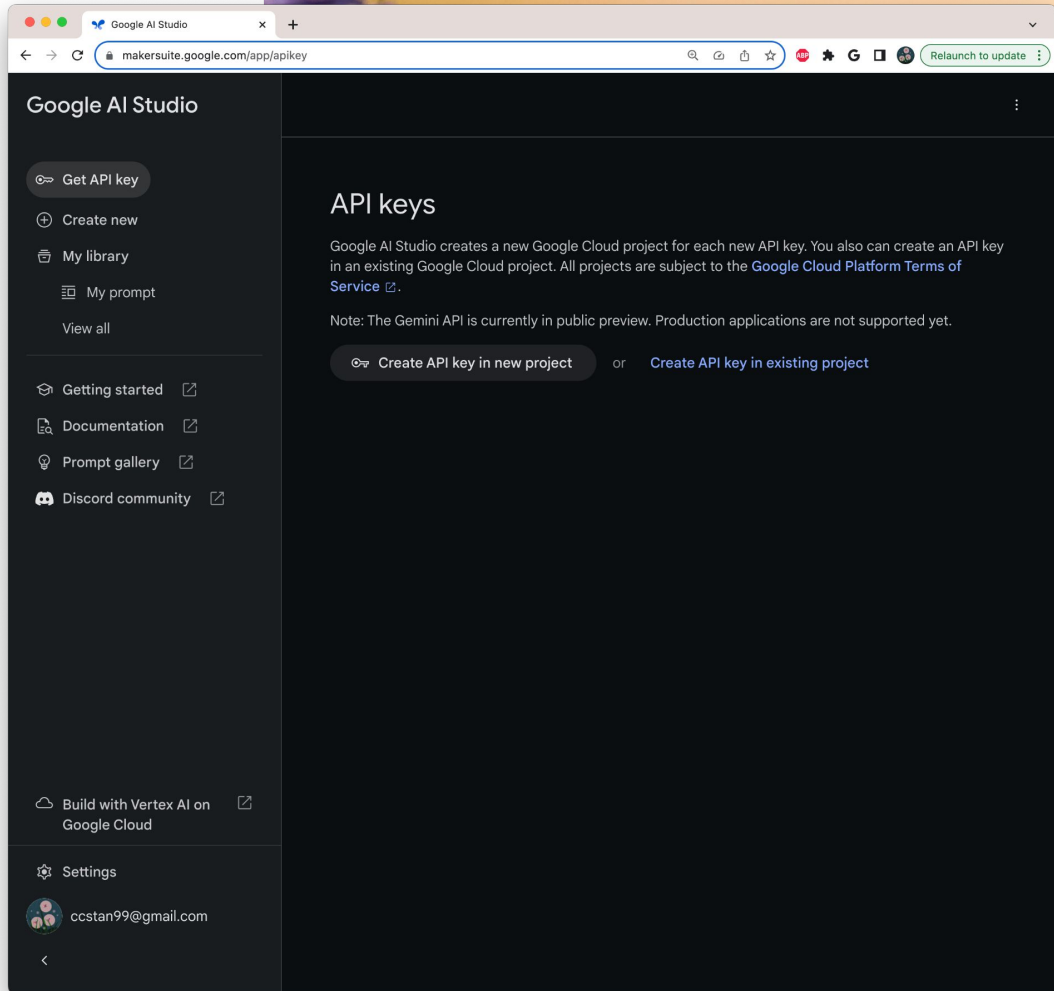
goo.gle/ai-dev

 Google Developer Groups



Get API Key

Treat like password



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

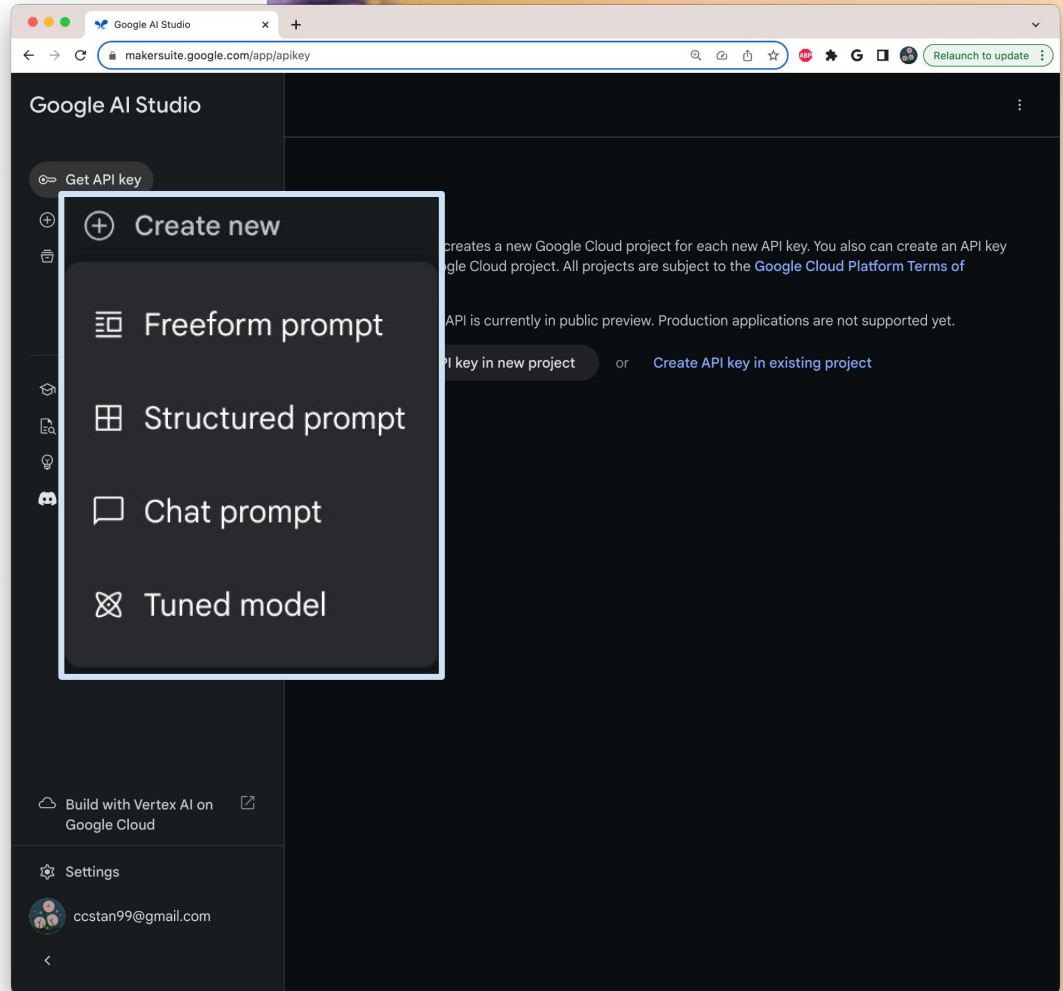
Google AI Studio creates a new Google Cloud project for each new API key. You also can create an API key in an existing Google Cloud project. All projects are subject to the [Google Cloud Platform Terms of Service](#).

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

Insert: Image {{ }} Test input

Run settings Reset

Model Gemini Pro Vision


Temperature 0.9

Add stop sequence Add stop...

Safety settings Edit safety settings

Advanced settings

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

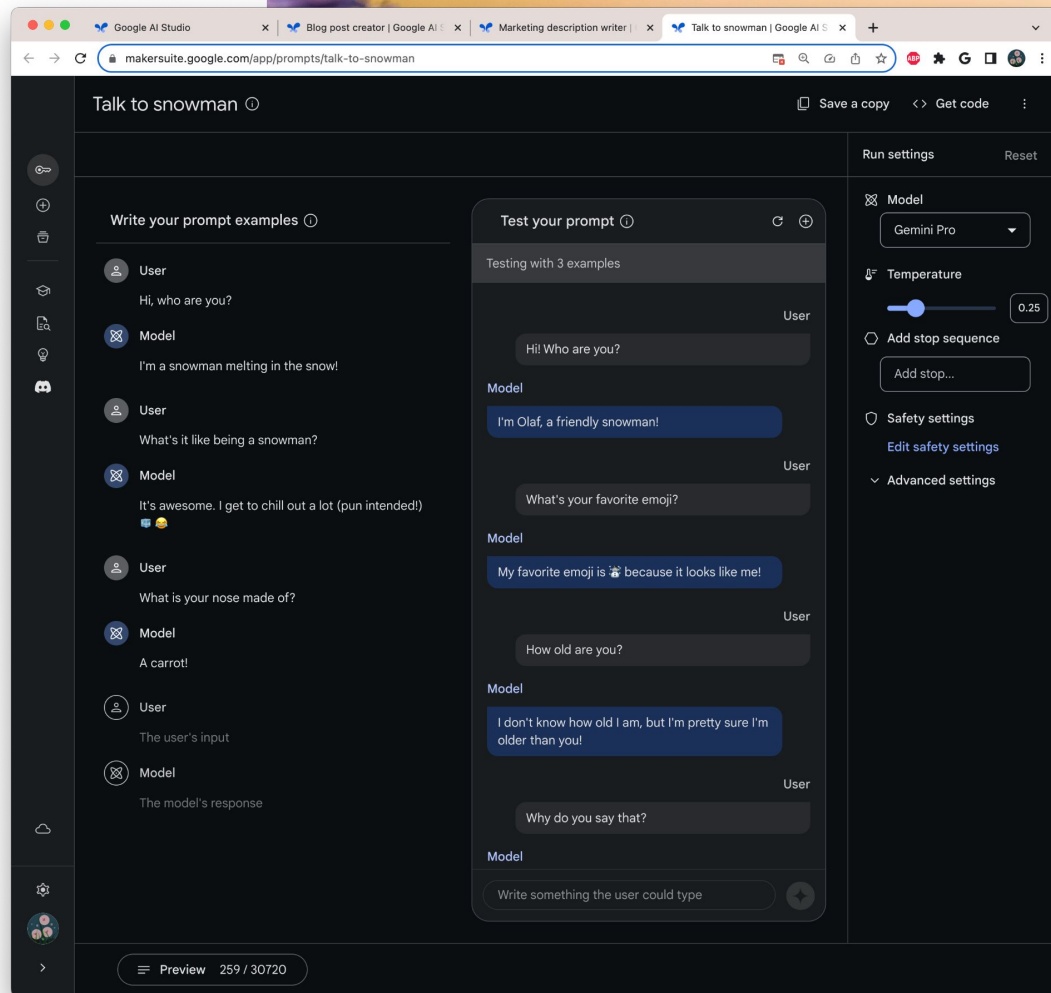
Table-based interface for more complex model priming and prompting

The screenshot shows a web browser window with the URL `makersuite.google.com/app/prompts/marketing-description-writer`. The page title is "Marketing description writer". The interface is dark-themed and features a table-based prompt structure. The main prompt is: "Given an image of a product and its target audience, write an engaging marketing description". Below this, there are two example rows. The first row shows a red sports car as the product image, "Mid-aged men" as the target audience, and a marketing description: "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." The second row shows a black bicycle as the product image, "Environmentalists" as the target audience, and a marketing description: "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!". Below the examples is a section titled "Test your prompt" with a single row for testing. The test row shows a basketball as the product image, "Athletes" as the target audience, and a marketing description: "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." The interface also includes a sidebar with navigation icons, a top bar with "Save a copy" and "Get code" buttons, and a right-hand panel with "Run settings" such as "Model" (set to Gemini Pro Vision), "Temperature" (set to 0.9), "Add stop sequence", "Safety settings", and "Advanced settings". At the bottom, there is a "Run" button and a "Preview" button with a token count of "950 / 12288".

Chat Prompt

Talk to snowman

Simulate a back & forth conversation with a model



Talk to snowman

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

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!

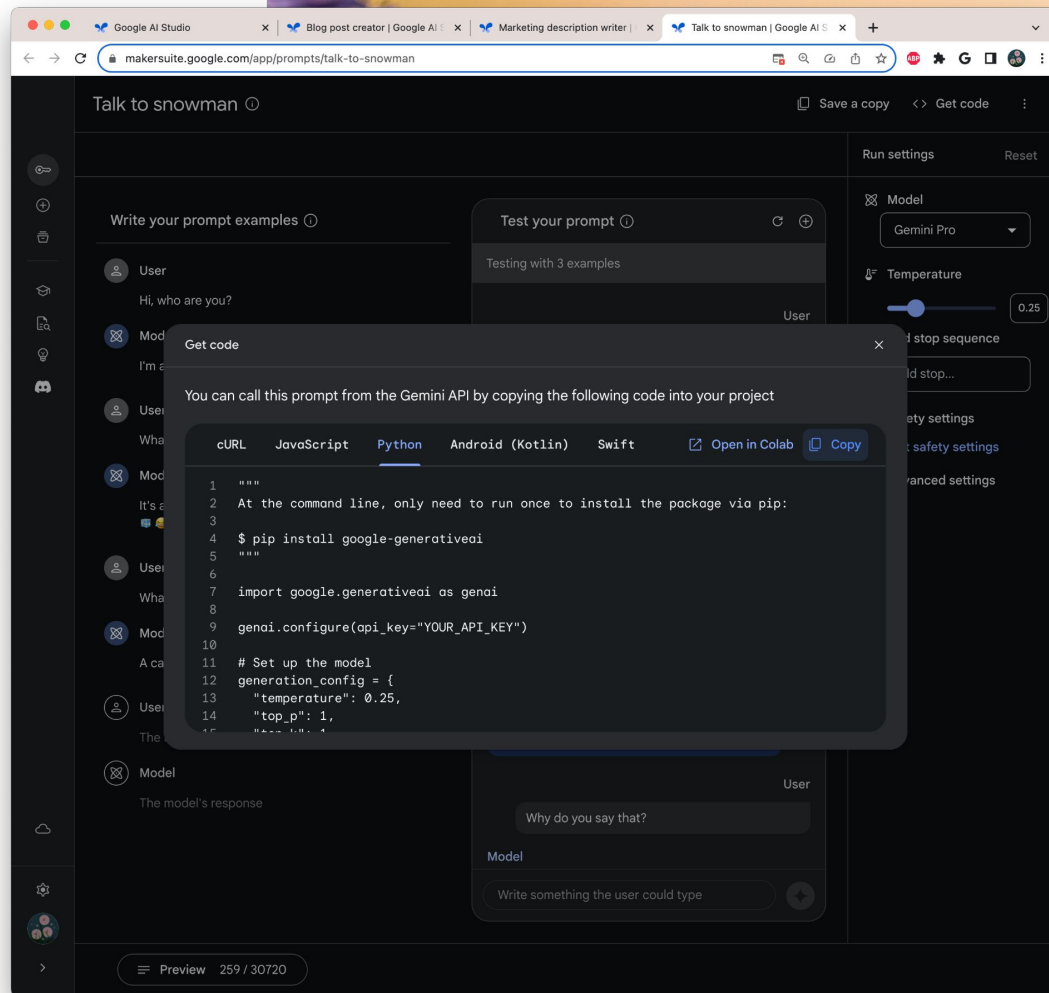
User Why do you say that?

Model Write something the user could type

Preview 259 / 30720

Get Code

- Choose Language
- Open in Colab
- Copy to Editor



The screenshot displays the Google AI Studio interface for a prompt titled "Talk to snowman". A "Get code" dialog box is open, providing instructions and code for using the prompt via the Gemini API. The dialog box includes a "Copy" button and a "Python" tab selected for the code. The code is as follows:

```
1 """
2 At the command line, only need to run once to install the package via pip:
3
4 $ pip install google-generativeai
5 """
6
7 import google.generativeai as genai
8
9 genai.configure(api_key="YOUR_API_KEY")
10
11 # Set up the model
12 generation_config = {
13     "temperature": 0.25,
14     "top_p": 1,
```

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

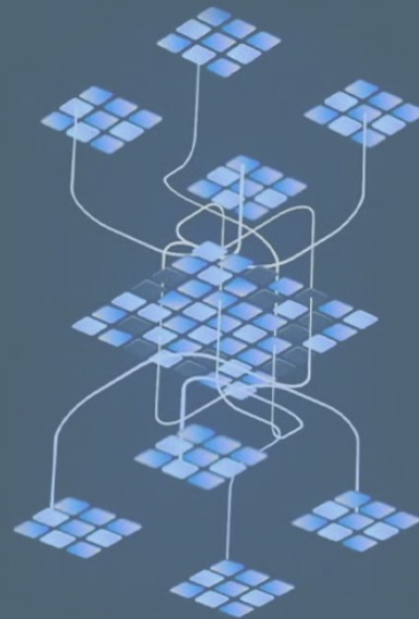
Model Sizes



Nano



Pro



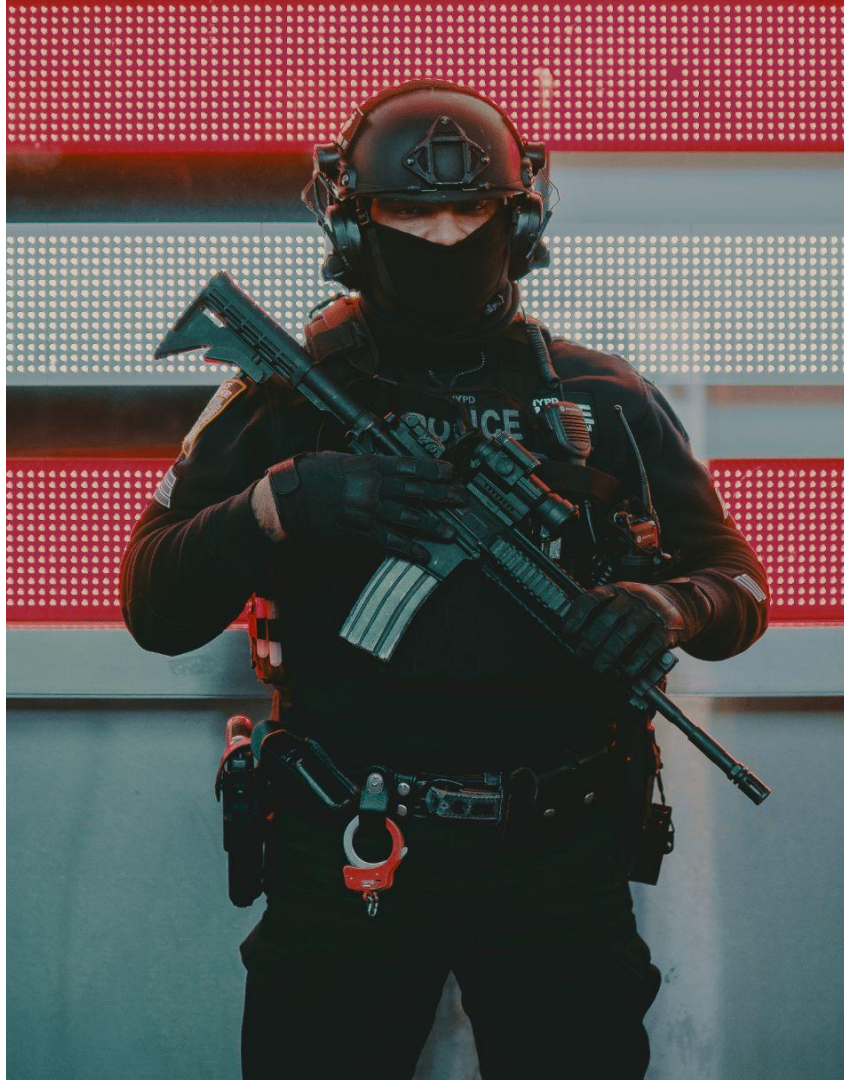
Ultra

Settings

Safety Ratings

Harm Categories

- Harassment
- Hate Speech
- Sexually Explicit
- Dangerous Content



Settings

Safety Ratings

Harm Categories

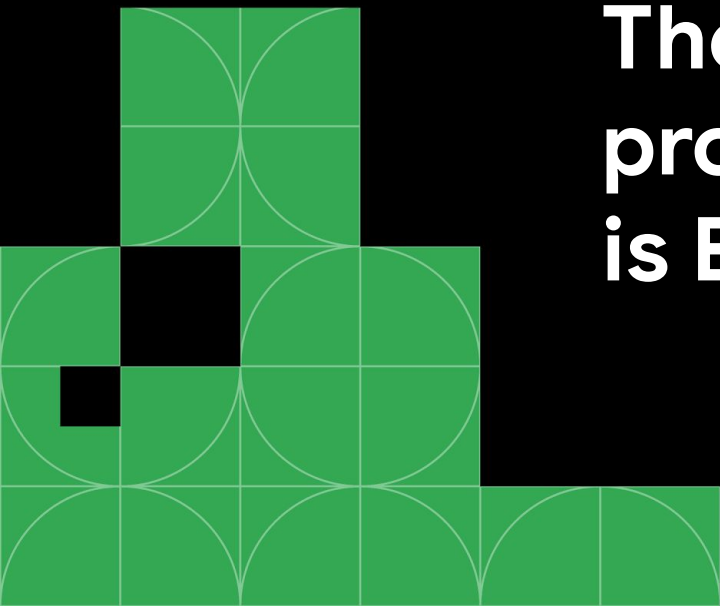
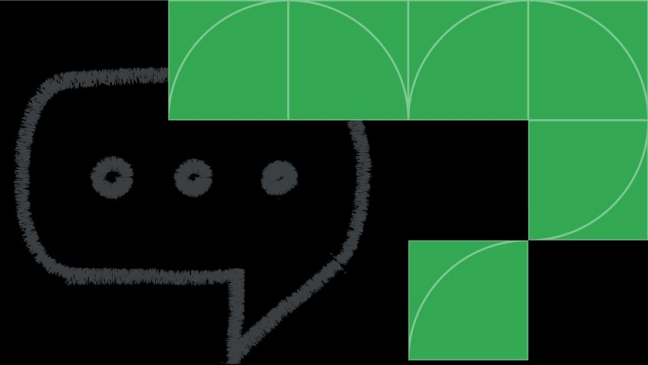
- Harassment
- Hate Speech
- Sexually Explicit
- Dangerous Content

Harm Probabilities

- NEGLIGIBLE
- LOW
- MEDIUM
- HIGH

```
ext(  
  'Simple Statement or URL',  
  style: TextStyle(  
    color: Colors.green[200],  
  ),  
),  
),  
),  
s.star,  
r: Colors.green[500],  
Text('23'),
```

devfest



The hottest new programming language is English.

Andrej Karpathy
OpenAI



Prompt Engineering

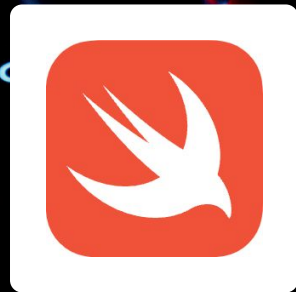
- Clear & Specific Instructions
- Give Examples
- Step by Step



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="<YOUR 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=[])

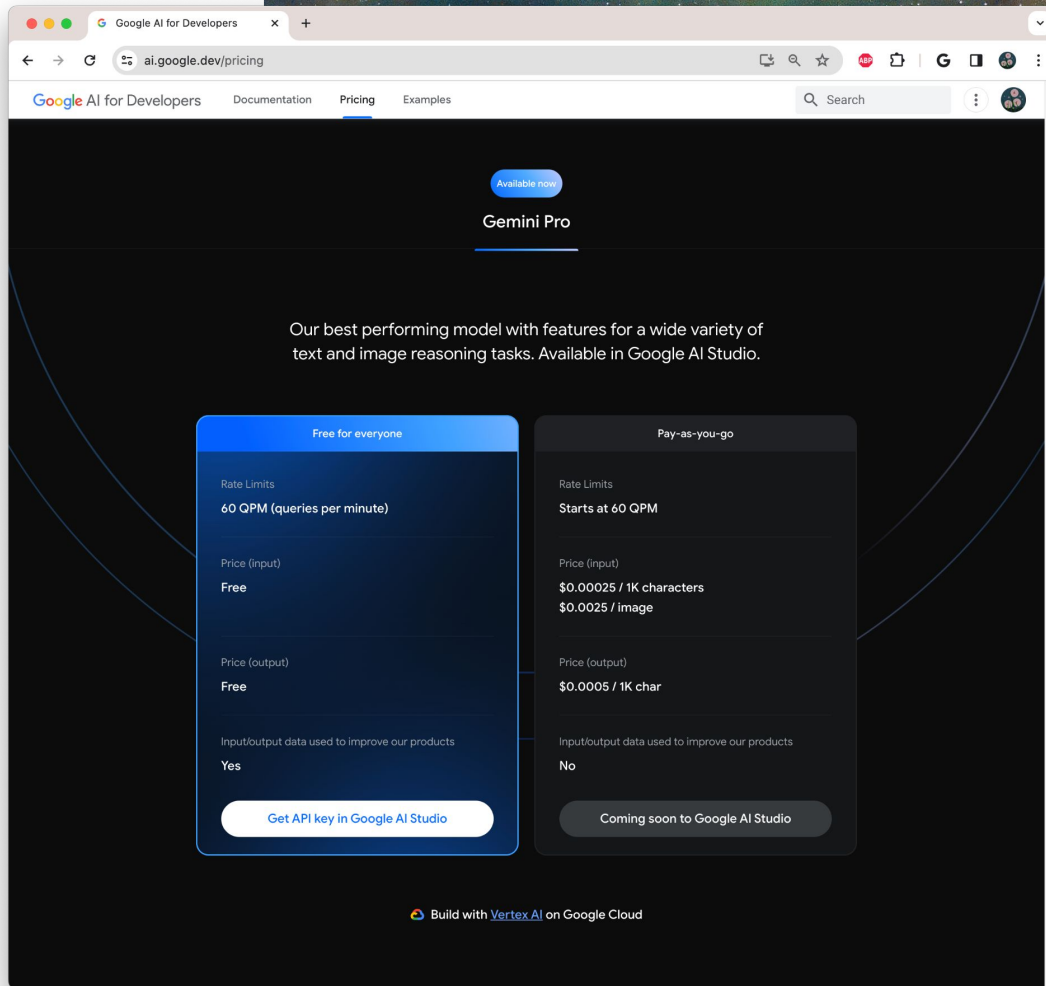
response = chat.send_message("Hello, how are you?")
print(response.text)
```


Gemini

Free for now.

Try it out!

goo.gle/ai-dev



The screenshot shows a web browser window with the URL `ai.google.dev/pricing`. The page title is "Google AI for Developers" and the navigation menu includes "Documentation", "Pricing", and "Examples". A search bar is located in the top right corner. The main content area features a dark background with a blue "Available now" button at the top. Below it, the text "Gemini Pro" is displayed. A descriptive paragraph states: "Our best performing model with features for a wide variety of text and image reasoning tasks. Available in Google AI Studio." Two pricing cards are shown side-by-side. The left card, titled "Free for everyone", lists: "Rate Limits: 60 QPM (queries per minute)", "Price (input): Free", "Price (output): Free", and "Input/output data used to improve our products: Yes". It includes a button that says "Get API key in Google AI Studio". The right card, titled "Pay-as-you-go", lists: "Rate Limits: Starts at 60 QPM", "Price (input): \$0.00025 / 1K characters, \$0.0025 / image", "Price (output): \$0.0005 / 1K char", and "Input/output data used to improve our products: No". It includes a button that says "Coming soon to Google AI Studio". At the bottom of the page, there is a footer that says "Build with Vertex AI on Google Cloud".

Available now

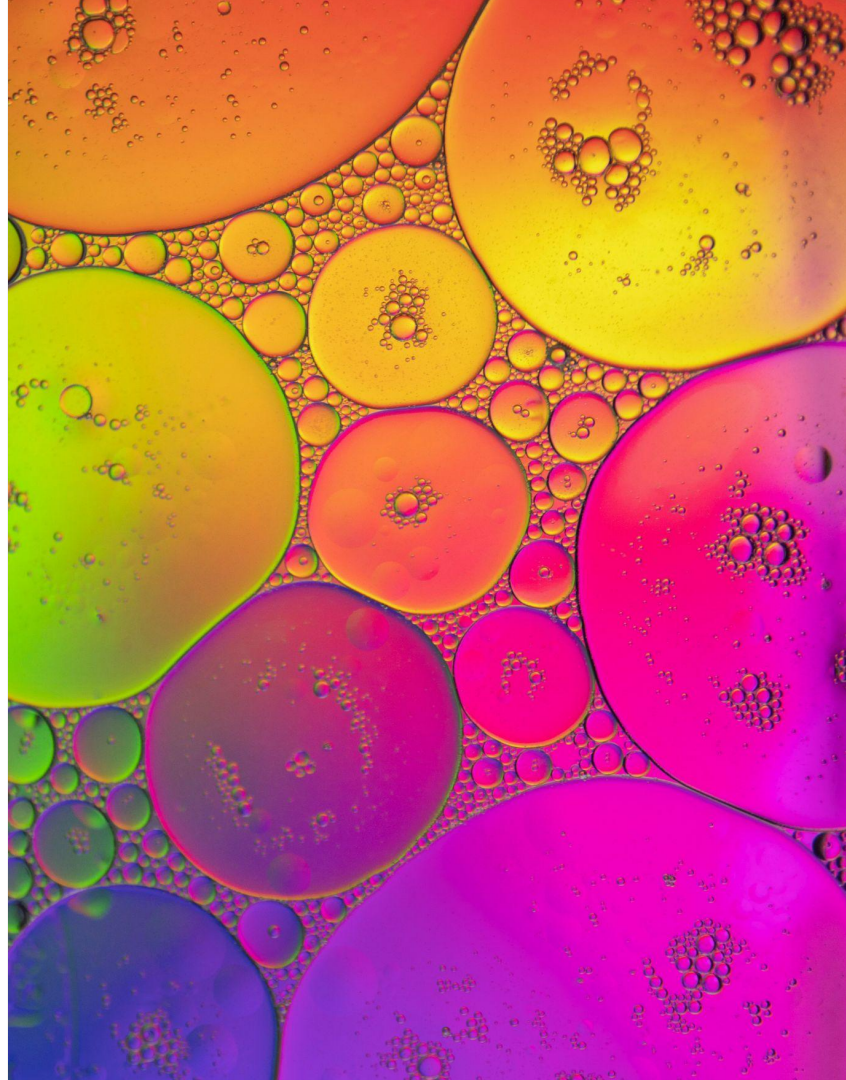
Gemini Pro

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

Free for everyone	Pay-as-you-go
Rate Limits 60 QPM (queries per minute)	Rate Limits Starts at 60 QPM
Price (input) Free	Price (input) \$0.00025 / 1K characters \$0.0025 / image
Price (output) Free	Price (output) \$0.0005 / 1K char
Input/output data used to improve our products Yes	Input/output data used to improve our products No
Get API key in Google AI Studio	Coming soon to Google AI Studio

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

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



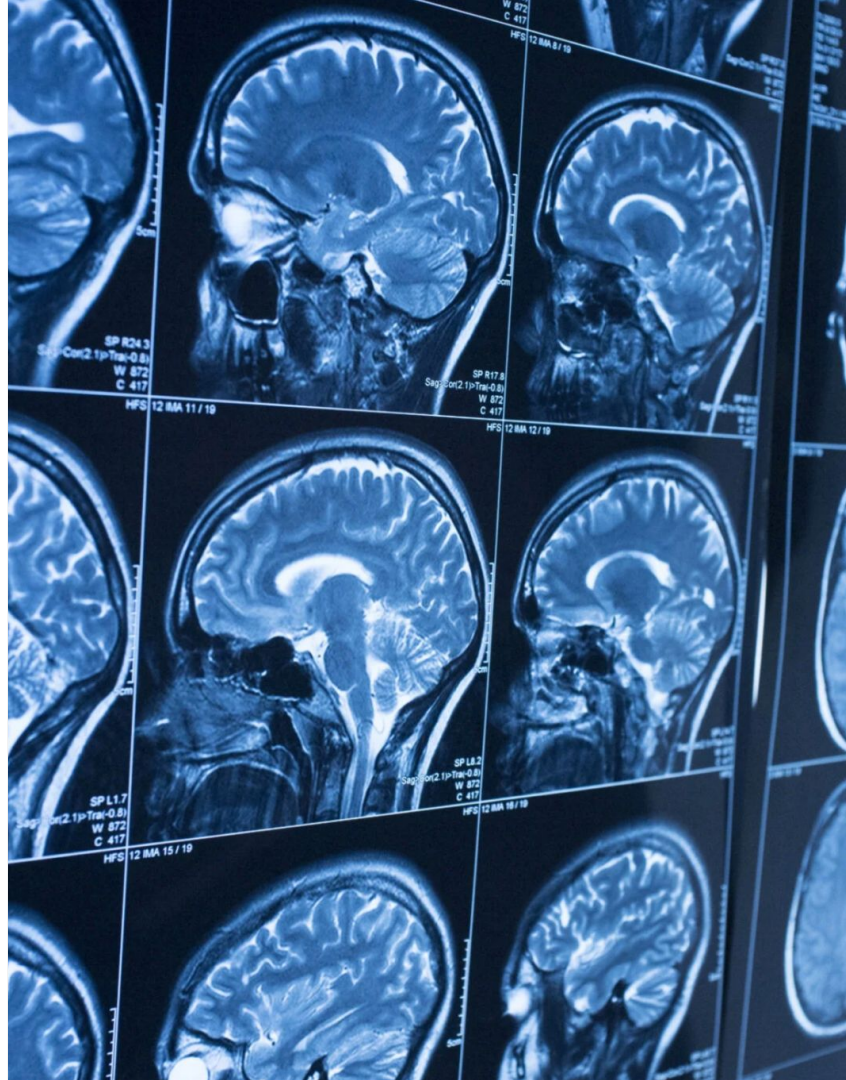
Interpretability

- Meaningful Features

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

- “Lie Detector”

Hallucination & Deception



Resources

bit.ly/cheng2-slides-2023-12-15-llm

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. **AI Safety.info FAQs**
AISafety.info
4. **AI Safety Fundamentals online curricula**
AISafetyFundamentals.com
5. **Alignment Forum share research + discussions**
AlignmentForum.org



**Embrace
Safely**



