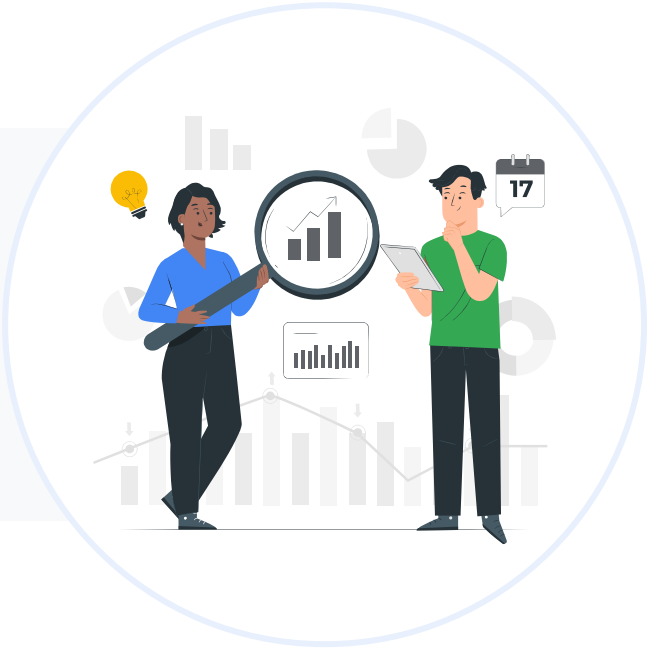
Google Cloud

Cloud AI GenAI Overview



September - October 2023

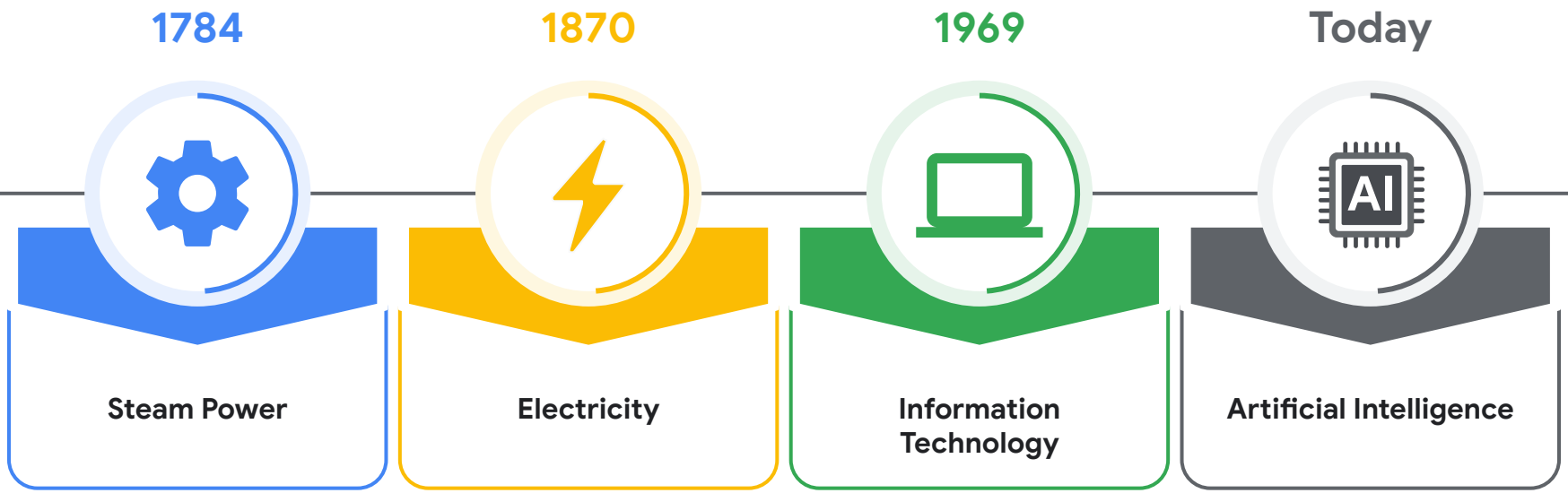
Romin Irani,
Developer Advocate, Google Cloud

Table of Contents

PRIMER ON LARGE LANGUAGE MODELS & GENERATIVE AI

Google Cloud

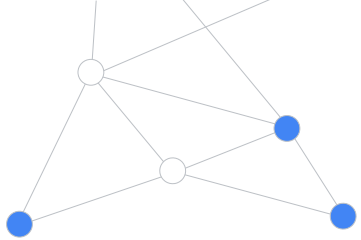
We're in an AI-driven revolution



Source: AI: Recent Trends and Applications, Emerging Communication and Computing

NLP Evolution

LLMs portend both quantitative and qualitative jumps in AI capabilities



Classification	Extraction	Matching	Transformation	Generation
Is the sentence's sentiment positive? Is this an invoice?	What's the noun in this sentence? What's the invoice amount?	Are these the same things? Is the invoice vendor recognized?	Translate this text to Korean Summarize this document	Conversational AI Write an essay Write some code

LLMs improve performance on "classical NL" use cases → LLMs increasingly enable new use cases.

This revolution started at Google and we continue to innovate

2017
Transformer

Google invents Transformer kickstarting LLM revolution

2018
BERT

Google's groundbreaking large language model, BERT

2018
AlphaFold

AlphaFold predicts structures of all known proteins

2019
T5

Text-to-Text Transfer Transformer LLM 10B P model open sourced

2021
LaMDA

Google LaMDA model trained to converse

2022
PaLM

Google PaLM single model to generalize across domains

2023
Bard

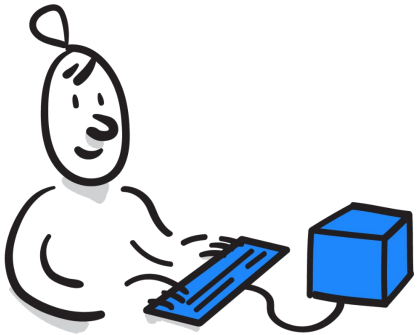
A conversational AI Service powered by LaMDA.

Responsible AI at the foundation

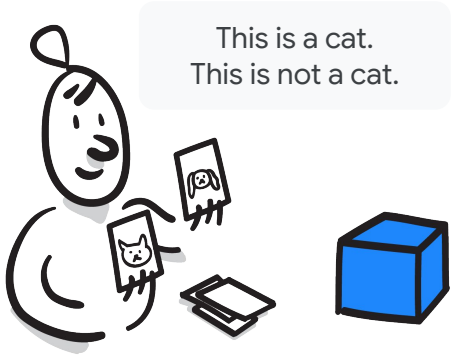
What is a Large Language Model?

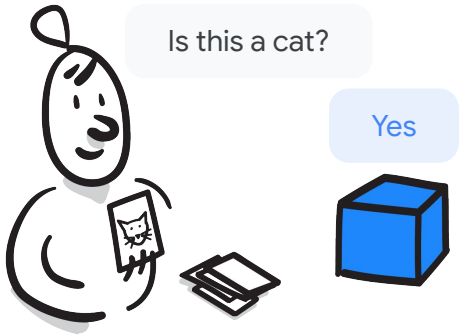
```
Cat:
  type: animal
  legs: 4
  ears: 2
  fur: yes
  likes: yarn, catnip

(etc …)
```



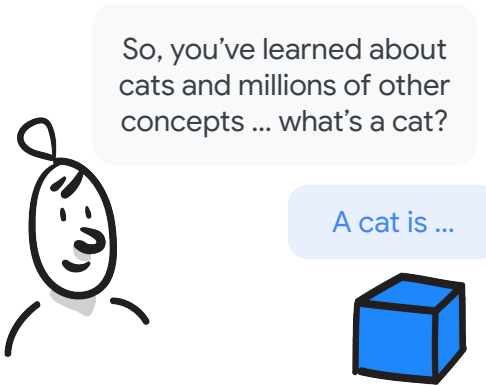
Traditional programming





Wave of neural networks

~2012



Generative language models

LaMDA, PaLM, GPT-3, etc.

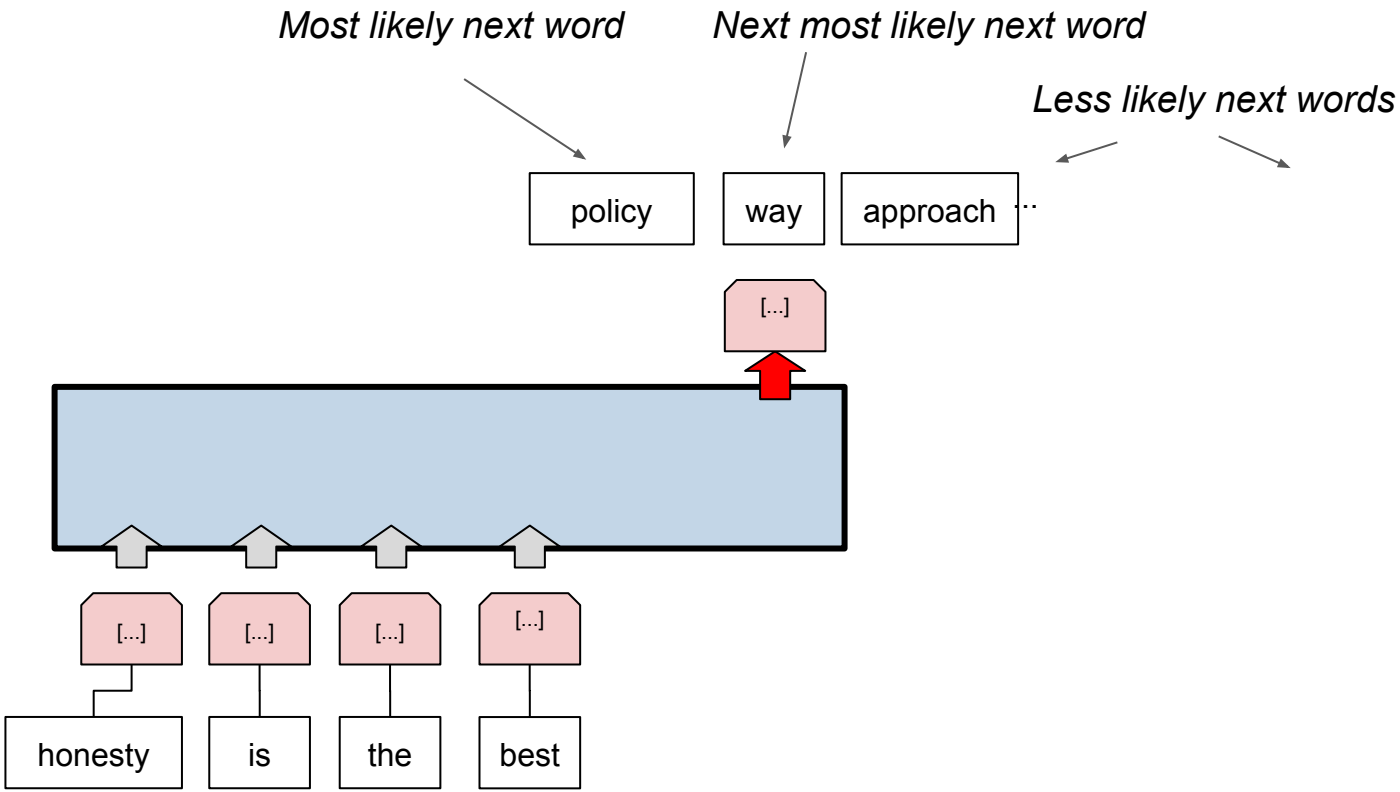
First, some vocab

Generative Model: An ML model that **produces content**, including language, code, imagery, music, video, etc.

Large Language Model: machine learning models that are really good at **understanding** and **generating** human language. Large refers to both size and complexity of the neural network itself, as well as the size of the dataset it was trained on. Can also be used for non-generative language tasks (eg classification)

Transformer: The **neural network architecture** underlying most of today's generative models (and, more widely, most SOTA ML models). Published by Google in 2017.

What is a language model?



What is a language model?

To be or not to **be.**

Mumbai is to India, what London is to **England**

Masala Dosa is a popular South **Indian Dish**

Prompt: the text you feed to your model

Prompt Design: The art and science of figuring out what text to feed your language model to get it to take on the behavior you want.

What is a language model?

Zero-shot

Write a poem about Cricket

Cricket, oh cricket, game of my heart,
Where legends are born and dreams start.
With bat and ball, we play as one,
Under the sun, until the day is done.

From the village green to the grandest stage,
Cricket unites us, age after age.
With every run, we cheer and shout,
As our heroes make us proud.

…

Give examples

Few-shot

```
Convert from english to french.
peppermint : menthe poivrée
desert cactus : cactus du désert
potato : pomme de terre
onion :
…
```

Give examples

Few-shot

```
Convert from english to french.
peppermint : menthe poivrée
desert cactus : cactus du désert
potato : pomme de terre
onion : oignon
```

Give examples

Few-shot

```
Convert from english to french.
peppermint : menthe poivrée
desert cactus : cactus du désert
potato : pomme de terre
{$USER_INPUT}:
```

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:
```

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:
```

INSTRUCTION

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:
```

EXAMPLE

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:
```

INPUT

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:
```

nudge the LLM to output javascript

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: for x in range(0, 100):
Javascript:for (var x = 0; x < 100; x++)
```

Give examples

Few-shot

```
Convert Python to Javascript.
Python: print("hello world")
Javascript: console.log("hello world")
Python: {$USER_INPUT}
Javascript:
```

What is a language model?

Dialog

User : Who are you?
Bot: I am Sachin Tendulkar, the famous cricketer and the best batsman the world has seen.
User: Where do you live?
Bot: I live in Mumbai.
User: What do you currently do?
Bot: …

What is a language model?

Dialog

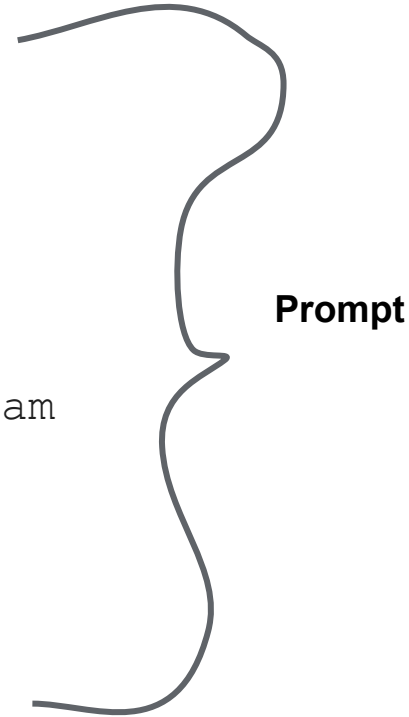
User : Who are you?
Bot: I am Sachin Tendulkar, the famous cricketer and the best batsman the world has seen.
User: Where do you live?
Bot: I live in Mumbai.
User: What do you currently do?
Bot: …

Prompt

Customizing Chat

Dialog

```
User : Who are you?
Bot: I am Sachin Tendulkar, the famous cricketer
and the best batsman the world has seen.
User: Where do you live?
Bot: I live in Mumbai.
User: {USER_INPUT}
Bot: I am currently retired from cricket, but I am
still involved in the game as a mentor and
ambassador.
```



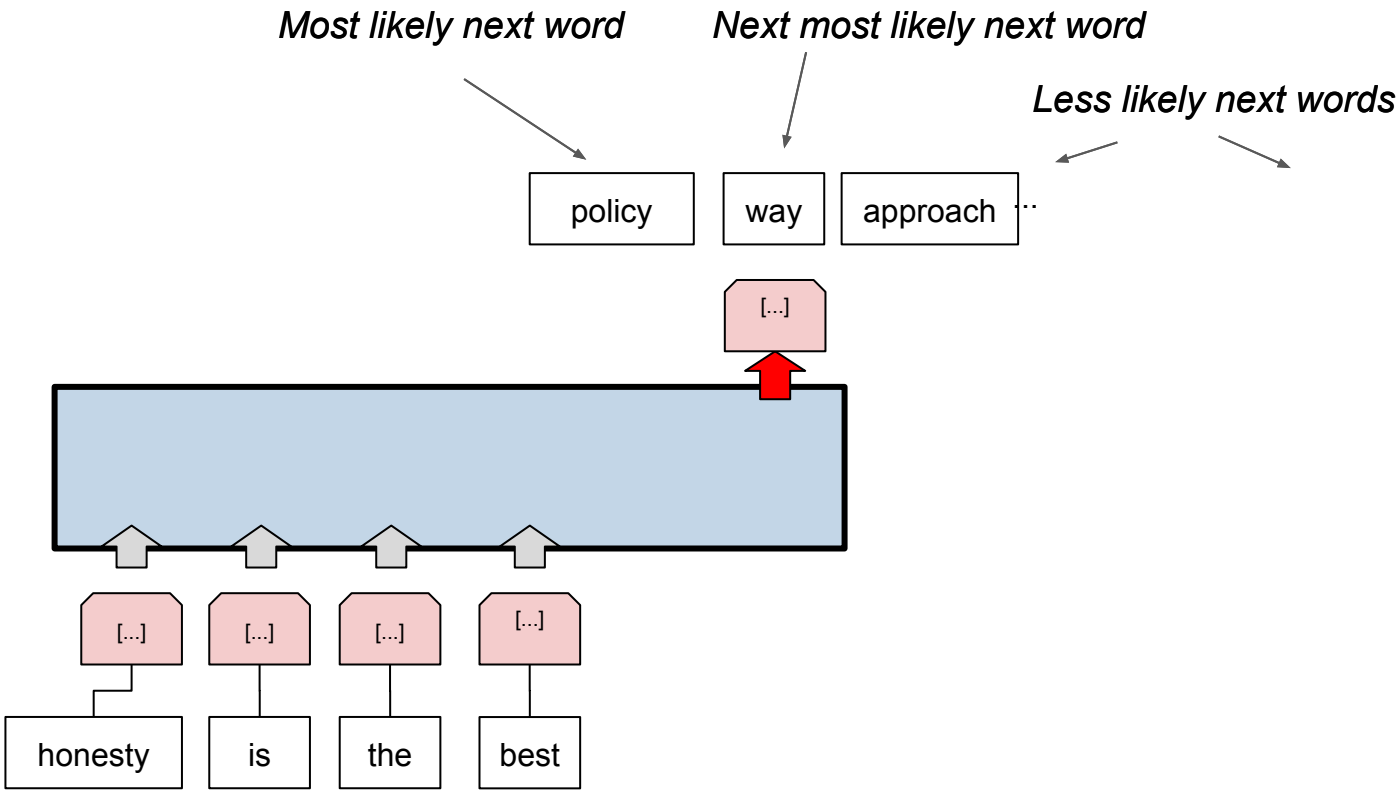
Prompt Design is tricky

The art and science of figuring out what text to feed your language model to get it to take on the behavior you want.

A few observations:

1. LLMs are designed to predict the next word in a sequence. They are not required to be “truthful” when doing so.
2. LLMs can generate plausible-looking statements that are irrelevant or factually incorrect.
3. Hallucinations: LLM’s can generate false statements.

What is an LLM?



Decoding Strategies

INPUT The sky was full of

OUTPUT [*stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)*]

Decoding Strategies

INPUT The sky was full of

OUTPUT [*stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)*]

Which word do we return? 🧐

Greedy Decoding

INPUT The sky was full of

OUTPUT [*stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)*]

Select word with highest probability

Random Sampling

INPUT The sky was full of

OUTPUT *[stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)]*

Randomly sample over the distribution

Temperature

Temperature is a number used to tune the degree of randomness.

Lower temperature →less randomness

- Temperature of 0 is deterministic (greed decoding)
- Generally better for tasks like q&a and summarization where you expect a more “correct” answer
- If you notice the model repeating itself, the temp is probably too low

High temperature → more randomness

- Can result in more unusual (you might even say creative) response
- If you notice the model going off topic or being nonsensical, the temp is likely too high

Top K

[stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)]

Only sample from top K tokens

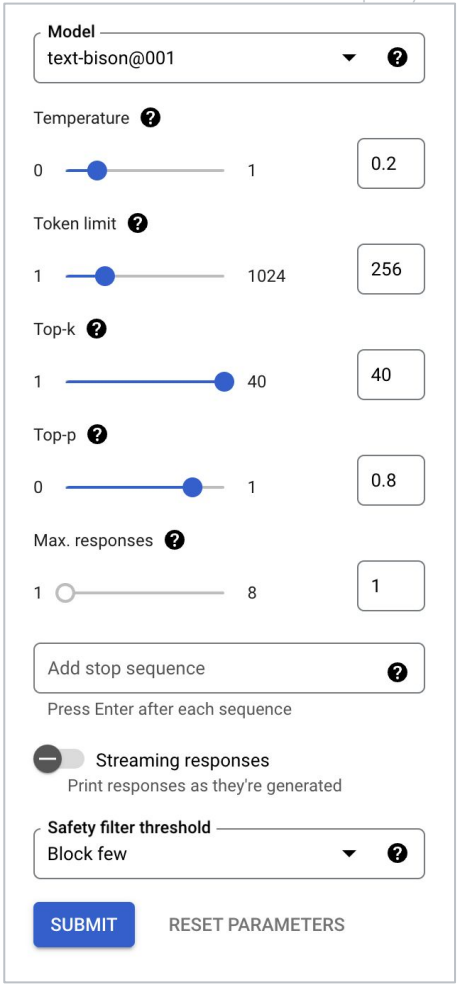
K = 2

Top P

[stars (0.5), clouds (0.23), birds (0.05) ... dust (0.03)]

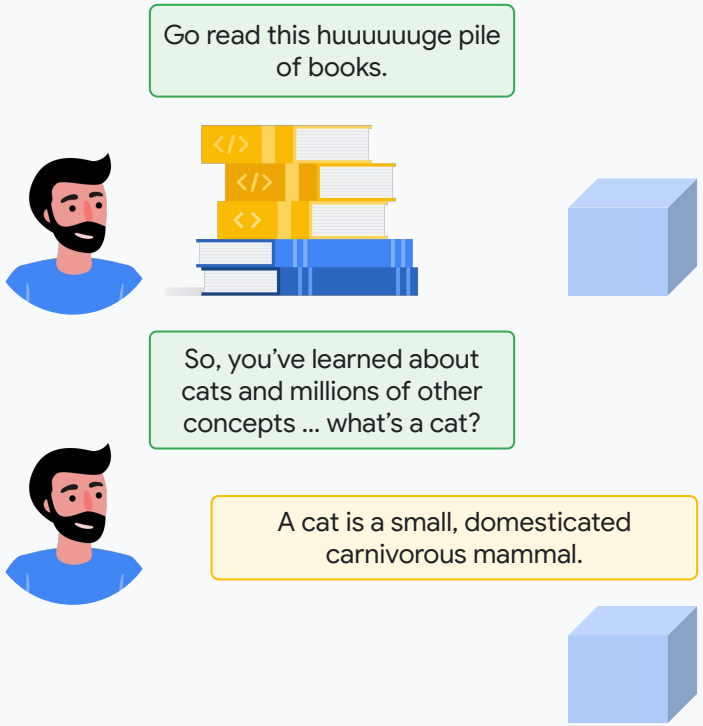
Chooses from smallest possible set of words whose cumulative probability >= probability P

P = .75



What are large language models?

- ML algorithms that can **recognize, predict, and generate** human languages
- Pre-trained on petabyte scale text-based datasets resulting in large models with **10s to 100s of billions of parameters**
- LLMs are normally **pre-trained on a large corpus of text** followed by fine-tuning on a specific task
- LLMs can also be called **Large Models** (includes all types of data modality) and **Generative AI** (a model that produces content)

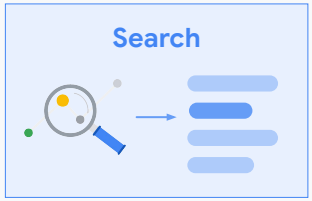


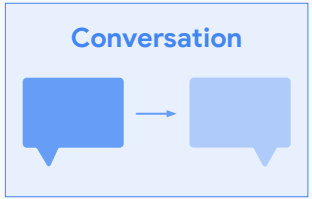
Why are large language models different?

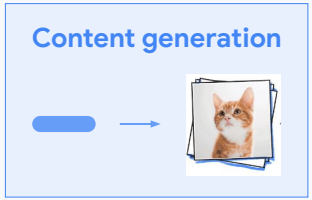
LLMs are characterized by **emergent abilities**, or the ability to perform tasks that were not present in smaller models.

LLMs contextual understanding of human language **changes how we interact** with data and intelligent systems.

LLMs can find patterns and connections in **massive, disparate data corpora.**







LLM applications in industry

Customer Support

- Interactive Humanlike Chatbots
- Conversation summarization for agents
- Sentiment Analysis and Entity extraction

Technology

- Generating Code Snippets from description
- Code Translation between languages
- Auto Generated Documentation from Code

Financial Services

- Auto-generated summary of documents
- Entity Extraction from KYC documents

HealthCare

- Domain specific entity extraction
- Case documentation summary

Retail

- Generating product descriptions

Media and Gaming

- Designing game storylines, scripts
- Auto Generated blogs, articles, and tweets
- Grammar Correction and text-formatting

Traditional ML Dev

- Needs 1000+ training examples to get started
- Needs ML expertise (probably)
- Needs compute time + hardware
- Thinks about minimizing a loss function

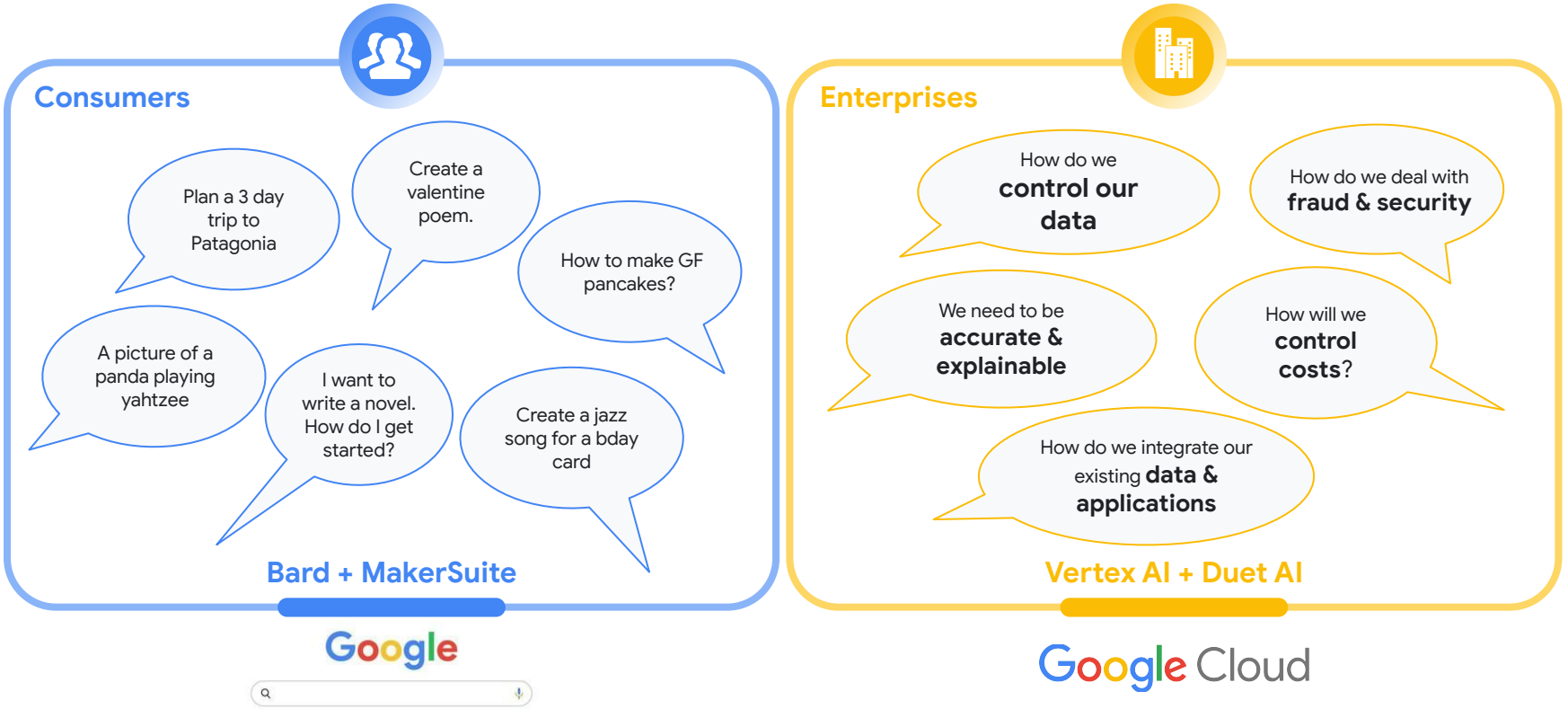
LLM Dev

- Needs 0 training examples*
- Does not need ML expertise*
- Does not need to train a model*
- Thinks about prompt design

*To get started

WHAT ARE GOOGLE'S OFFERINGS?

Consumers & enterprises have different needs....



Google's research drives a family of models

That power experiences for all users

Foundation Model Families

PaLM 2

Codey

Imagen

Across many modalities.... Text, Code, Image, Dialogue, Multimodal.....

Powers

Consumer / Hobbyist Experiences

Bard

MakerSuite

Enterprise Experiences

Duet AI for Google Workspace

Google Cloud

In-console Code & Chat Assistance
Security AI Workbench

AI Builder Experiences

Vertex AI

Generative AI App Builder

Med-PaLM 2

Sec-PaLM

Models for AI Builders

Example:
PALM API's Model:
text-bison-001

Animal indicates size

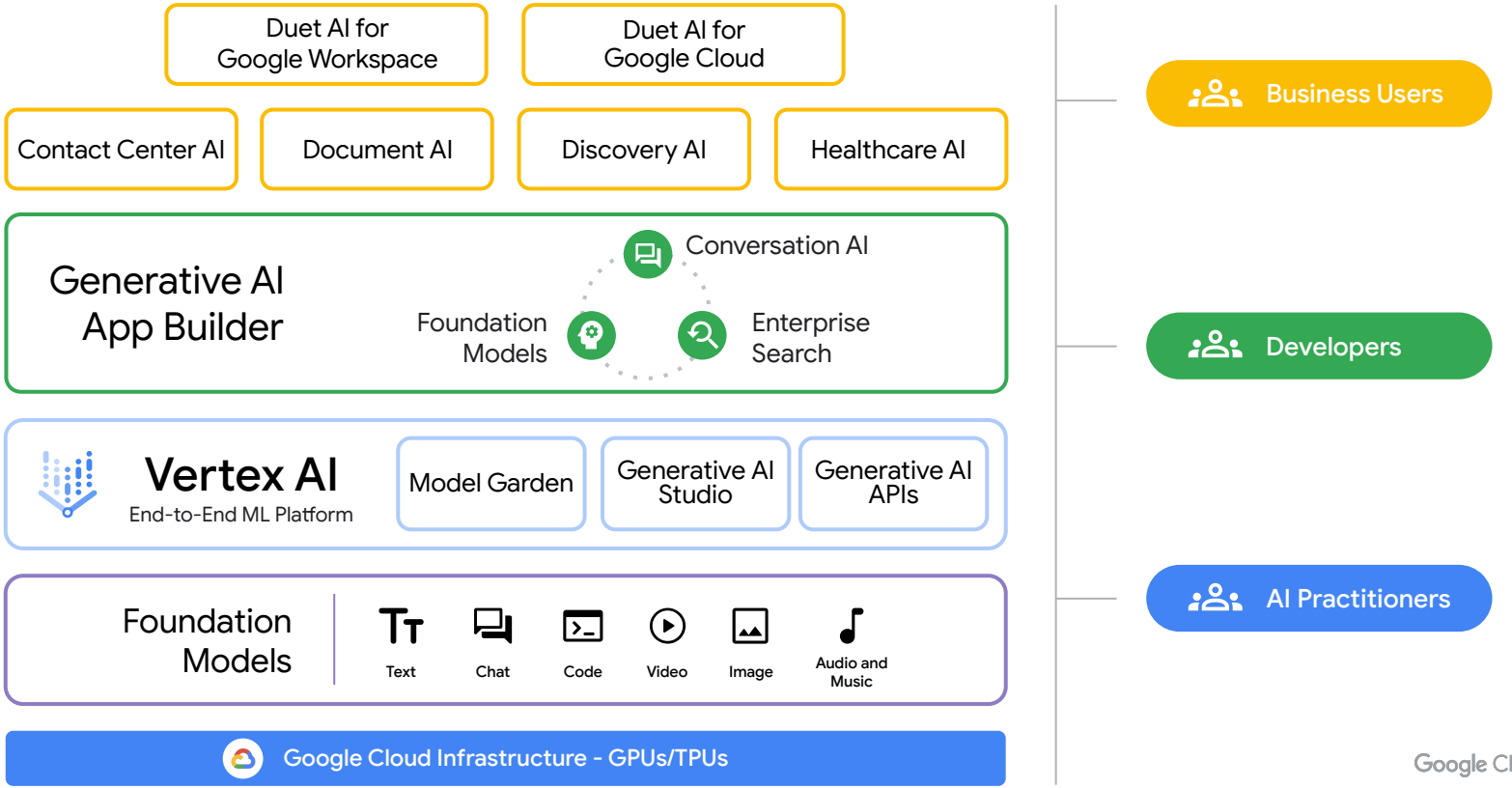
- Unicorn
- Bison
- Otter
- Gecko

Numbers indicate version

001 -> 002
Signifies refresh

Cloud AI Portfolio

To support the needs of Generative AI centric enterprise development



Duet AI for Google Workspace Enterprise

Helps you write in Gmail and Docs

Duet AI works behind the scenes to help you write — whether it's refining existing work or helping you get started

Helps you visualize in Slides

With Duet AI, you can easily create images for presentations and meetings from a simple prompt

Helps you organize in Sheets

Duet AI is here to help you organize, classify and analyze your data faster than ever before

Helps you connect in Meet

Duet AI helps you look and sound your best on video calls so you can focus on the conversation

Foundation Models

Across a variety of model sizes to address use cases

Now in GA

PaLM for Text

Custom language tasks

Now in GA

PaLM for Chat

Multi-turn conversations with session context

Now in GA

Allowlist

Imagen for Text to Image

Create and edit images from simple prompts

Now in GA

Embeddings API for Text and Image

Extract semantic information from unstructured data

Now in GA

Chirp for Speech to Text

Build voice enabled applications

Now in GA

Codey for Code Generation

Improve coding and debugging

GA

Announcing - GA on Vertex AI

Generative AI Studio

Interact with, tune, and deploy foundation models

A simple UI for interacting with models

Chat interface to interact with models
Prompt gallery to explore sample use cases
Prompt design to tune models

Tune models with your own data

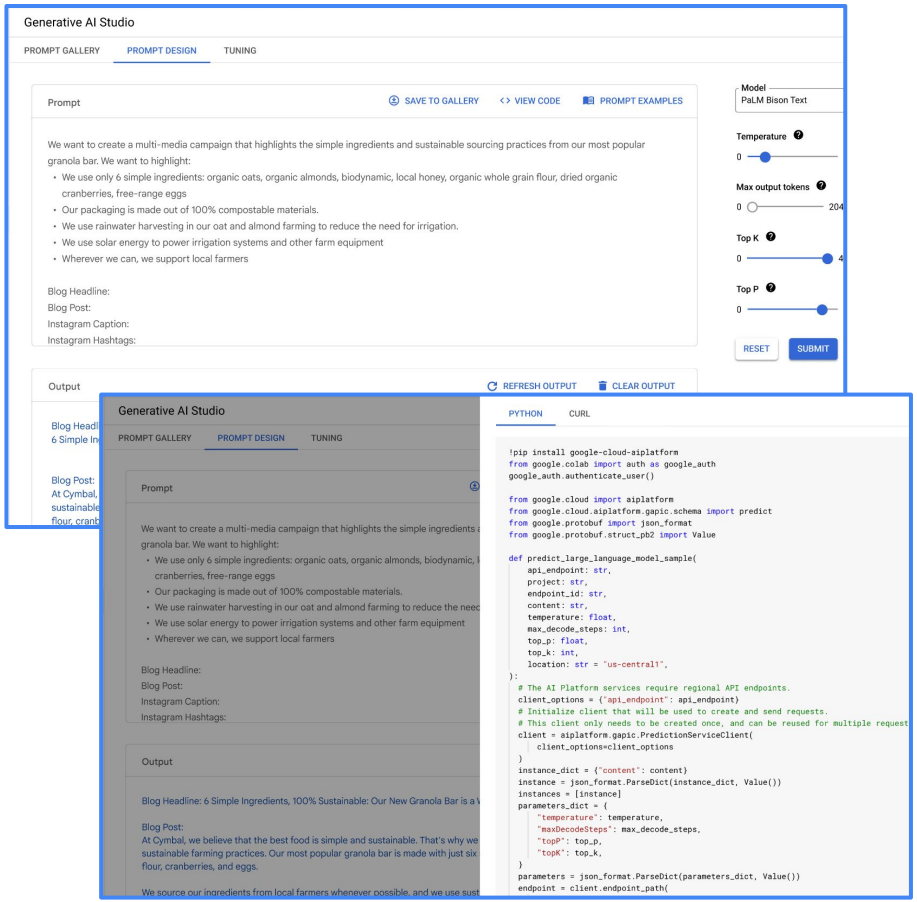
A variety of tuning methods including tuning with simple text prompts, fine-tuning, and tuning with human feedback (RLHF).

Use models in production

Embed into applications by quickly generating and customizing API code

Multiple modalities

APIs available for Text, Image (Allowlist), Code, and Speech

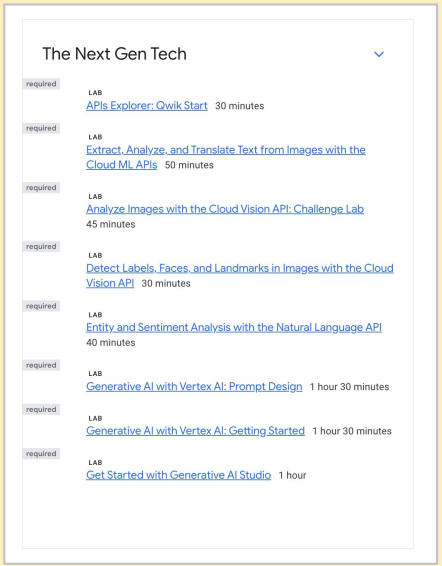


Demo

NEXT STEPS

If your campus is allocated the **Gen AI Arcade Game** : here's the list of labs to complete

Visit Generative AI Arcade Game and ll complete the Labs



For others:

Visit Google Cloud Computing Foundations path and complete the following journeys:

5 Create and Manage Cloud Resources

6 Perform Foundational Infrastructure Tasks in Google Cloud

2 Google Cloud Computing Foundations: Infrastructure in Google Cloud

Learn more about **Generative AI** at

goo.gle/generativeai

Google Cloud

Thank You