01

Foundation Model Ecosystem on Google Cloud

The information in this presentation is classified:

Google confidential & proprietary

⚠️ This presentation is shared with you under NDA.

- Do **not** record or take screenshots of this presentation.
- Do **not** share or otherwise distribute the information in this presentation with anyone **inside** or **outside** of your organization.

Thank you!

In this module, you learn to ...

01 Leverage foundation models for generative AI

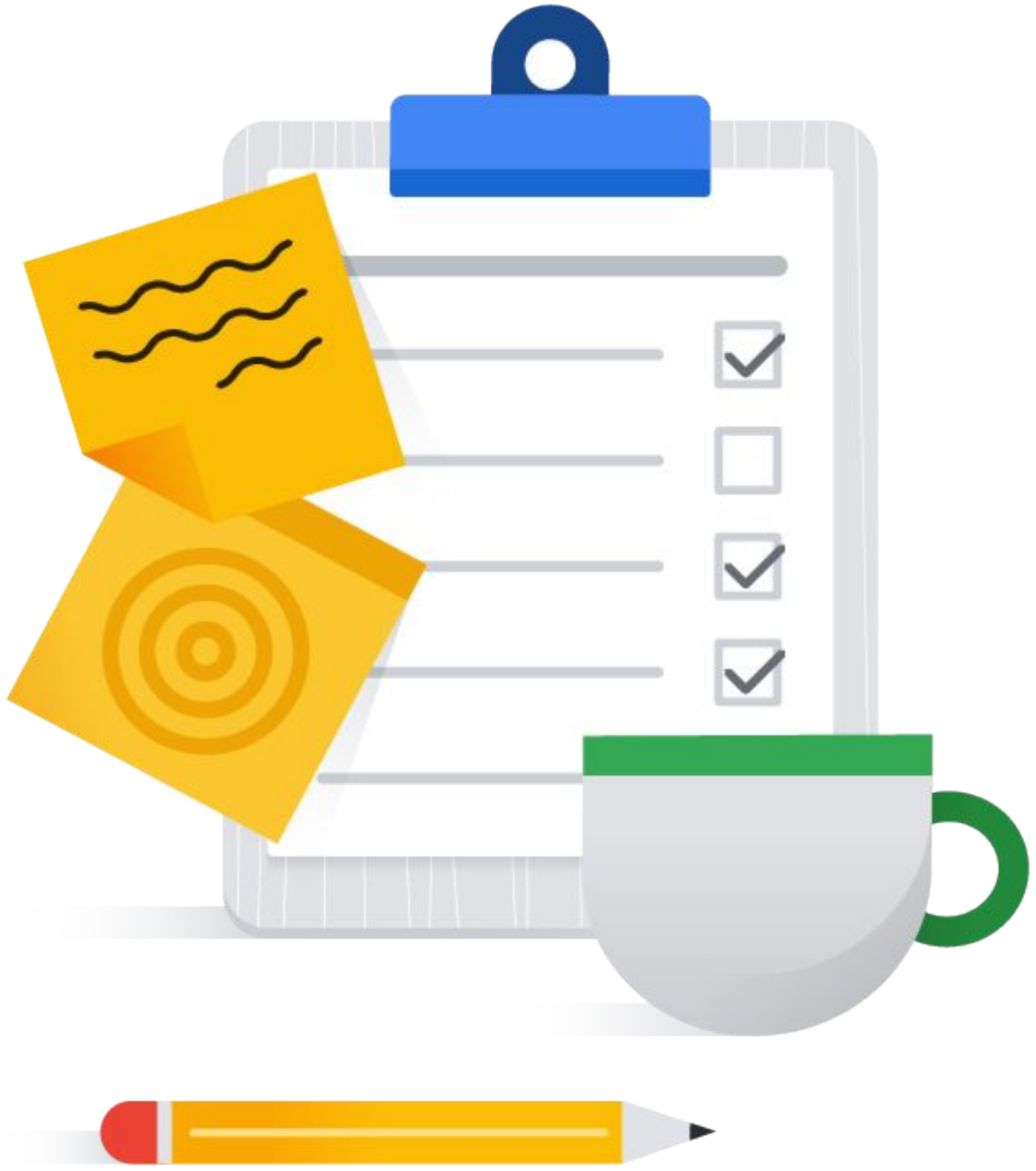
02 Use Google's foundation models to optimize generative AI tasks

03 Start prompting with Vertex AI Studio

04 Explore use cases for generative AI

05 Find and implement models using Vertex AI Model Garden

Topics



This revolution started at Google

Transformers
- Pathbreaking Neural Network Architecture
- Open Sourced by Google in 2017
- Started the revolution in Language Models

T5
(Text-to-Text Transfer Transformer)
- Large Language Encoder-Decoder Model
- 10-billion parameter model
- Open Sourced by Google in 2019

Diffusion Models
High Fidelity Image Generation Using Diffusion Models

PaLM
- (Pathways Language Model)
- Single model to generalize across domains
- 540-billion parameter, dense decoder model

Bard
- A conversational AI service powered by LaMDA.

2017 | 2018 | 2019 | 2020 | 2121 | 2022 | 2023

BERT
(Bidirectional Encoder Representations from Transformers)
- World's first Language Model
- Open Sourced by Google in 2018
- SOTA on number of language benchmarks

LaMDA
(Language Model for Dialog Applications)
- Model trained on dialogue data
- Model could talk about virtually anything
- Published by Google in 2020

CALM
(Confident Adaptive Language Modeling)
- Accelerating the text generation of LMs

This revolution continues...

PaLM 2
(Pathways Language Model))
- Text generation large language model

Imagen
- Foundational Image model
- Image generation
- Image Q&A
- Captioning

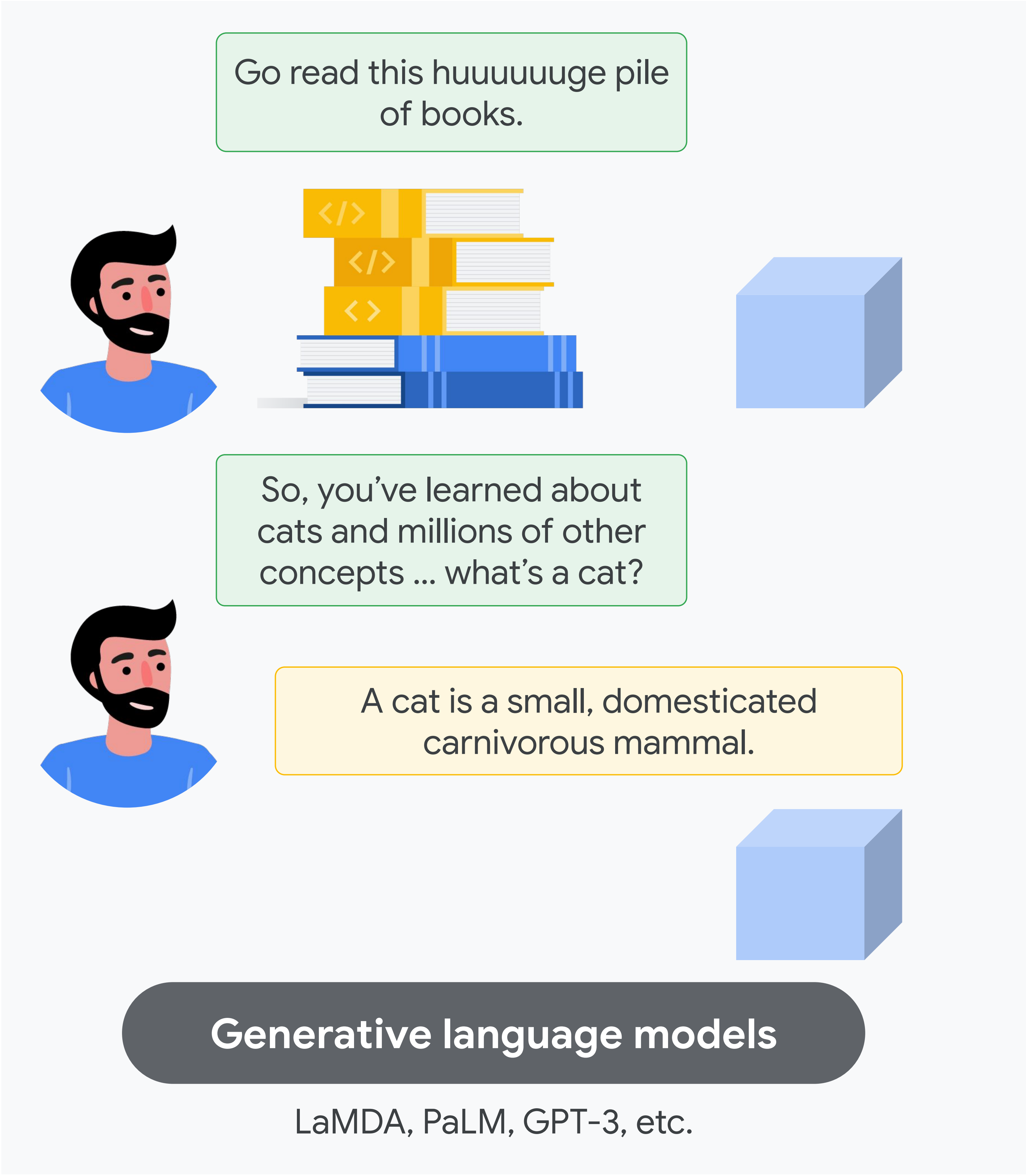
Gemini
- Multi-modal model
- Text generation
- Image analysis
- Video analysis

2023

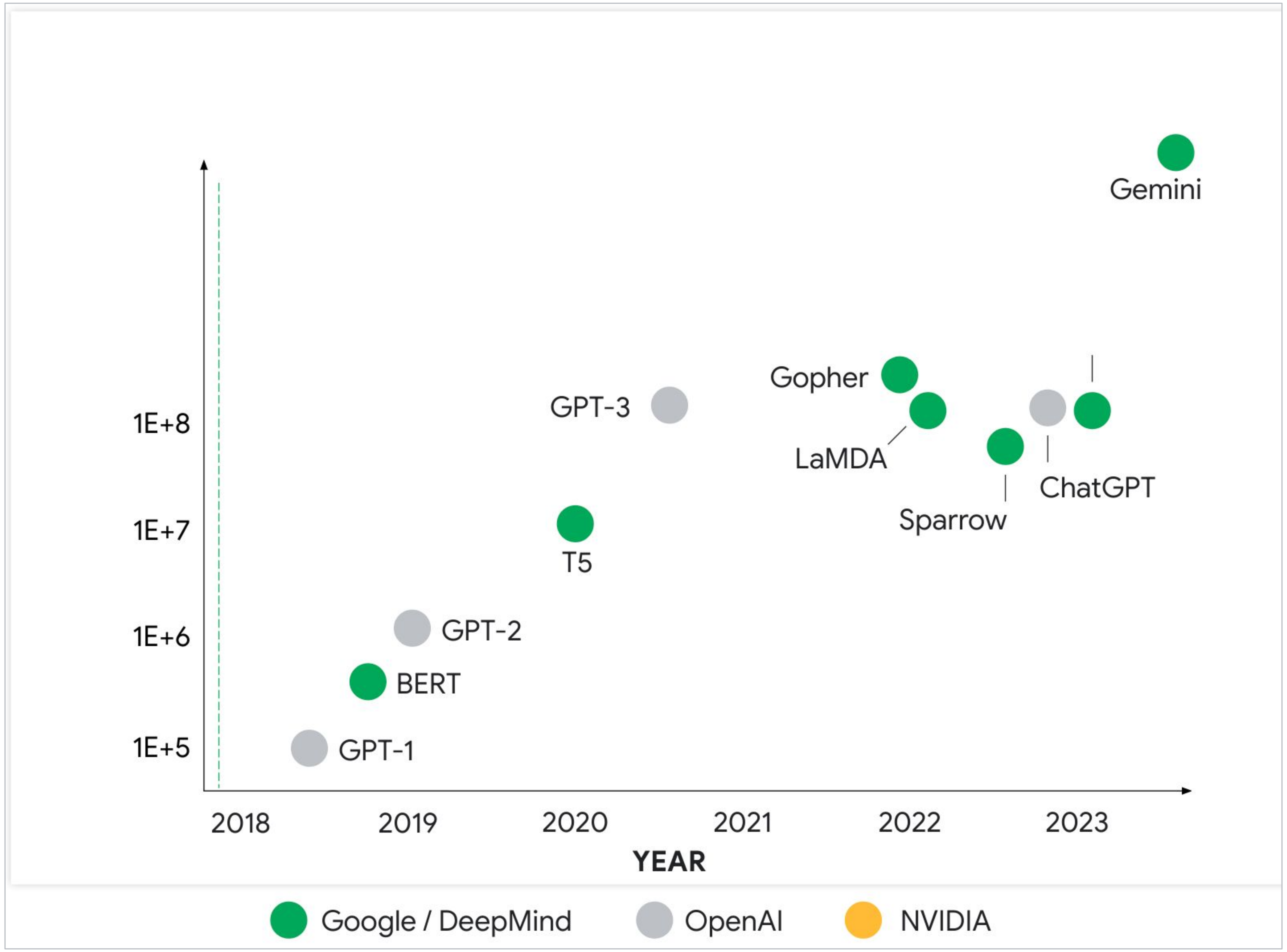
2024

Large Language Models (LLMs)

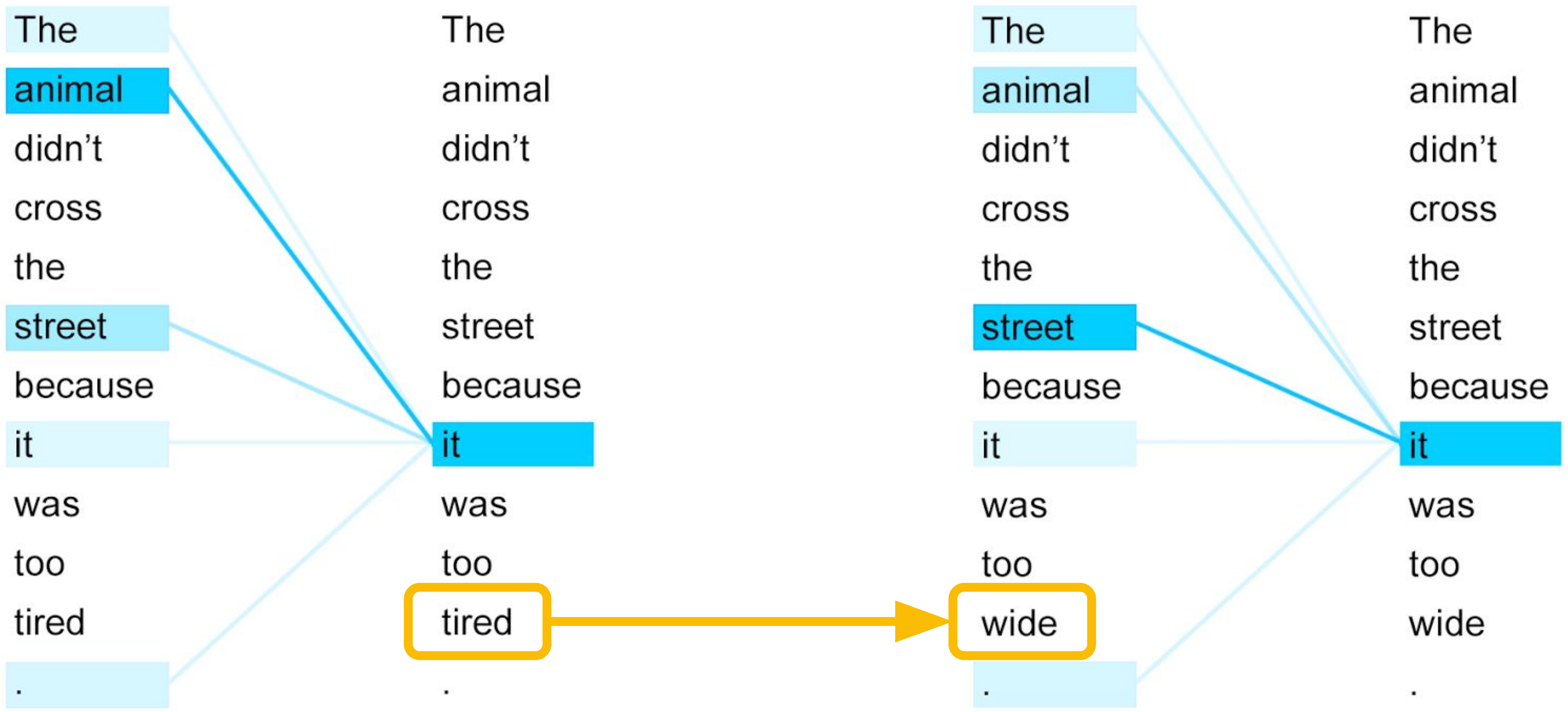
- ML algorithms that can **recognize, predict, and generate** human languages
- Pre-trained on petabyte scale text-based datasets resulting in large models with **10s to 100s of billions of parameters**
- LLMs are normally **pre-trained on a large corpus of text** followed by fine-tuning on a specific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



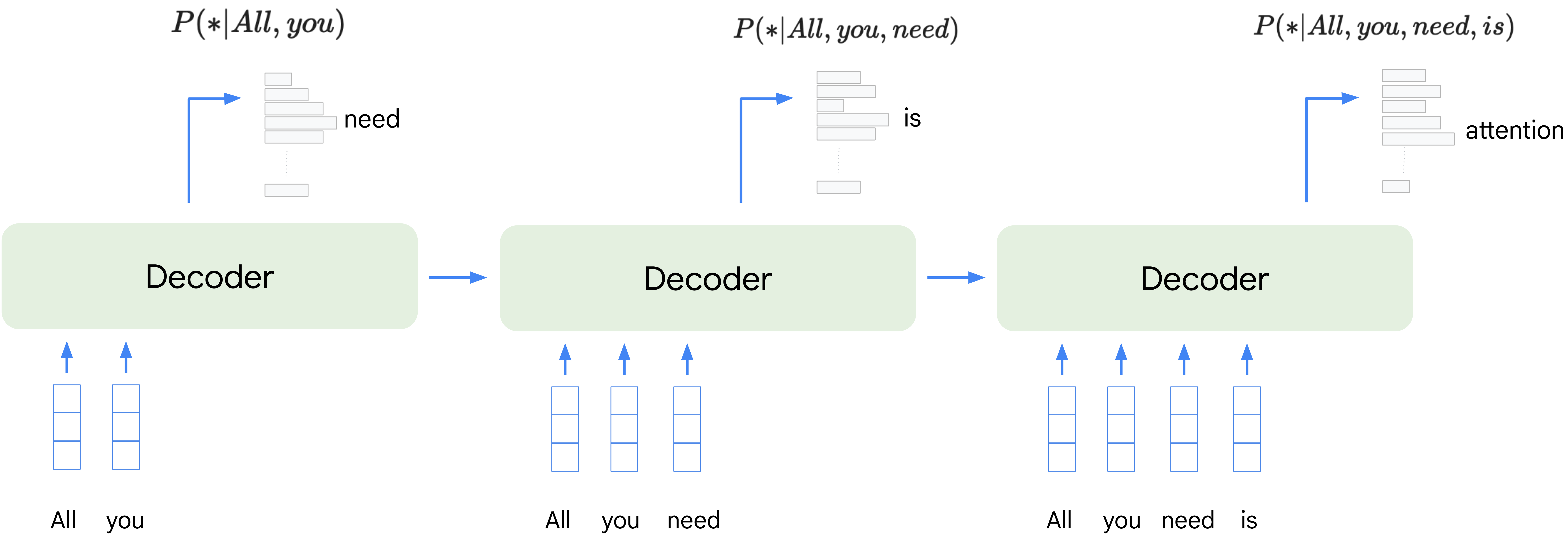
LLMs have driven an explosion in model size



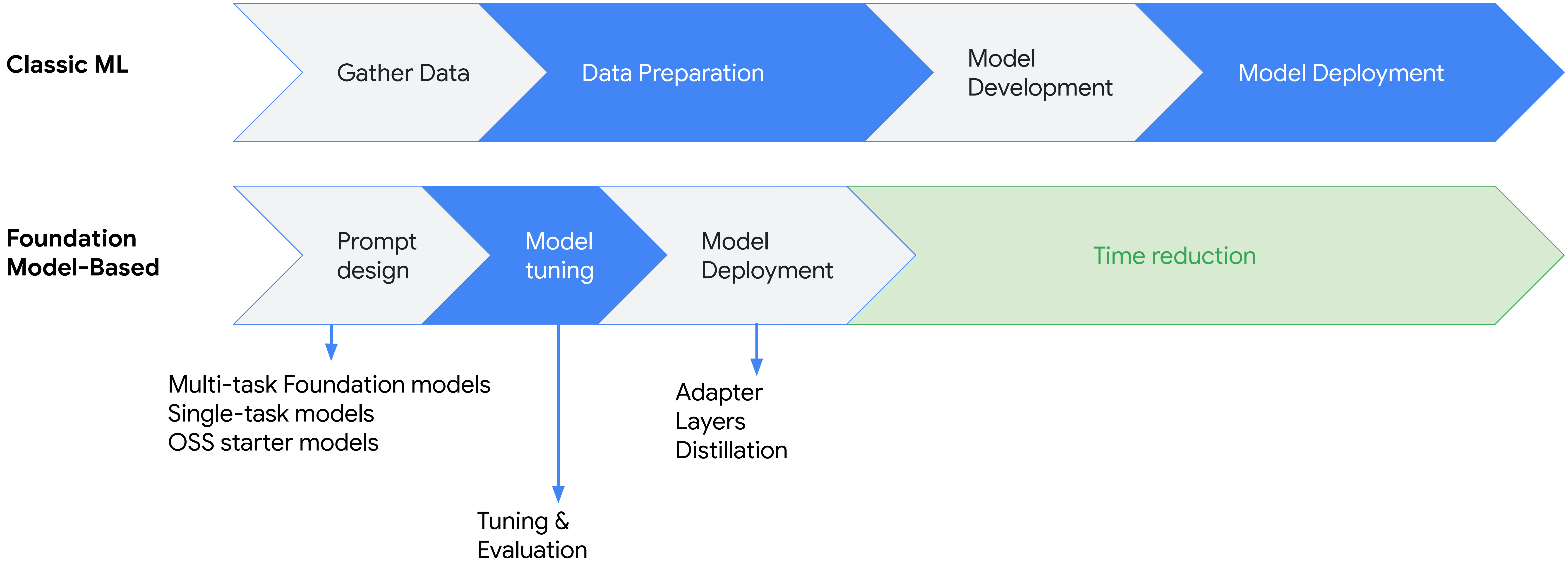
Transformers use a self-attention layer. Enhacing each token's sensitivity to its relationship with other tokens.



The LLM's decoder returns a probability distribution.



Foundational models accelerate time to model deployment



There are also challenges with generative AI

- Ensuring the quality of generated content
- Hallucinations
  - Incorrect statements can be presented in a confident manner
- Preventing offensive or harmful responses

Use cases that build on the strengths of generative AI

Language

- Draft Writing
- Summarization
- Ideation
- Classification
- Sentiment analysis
- Extraction
- Chat
- Search

Code

- Code generation
- Code completion
- Code chat
- Code conversion

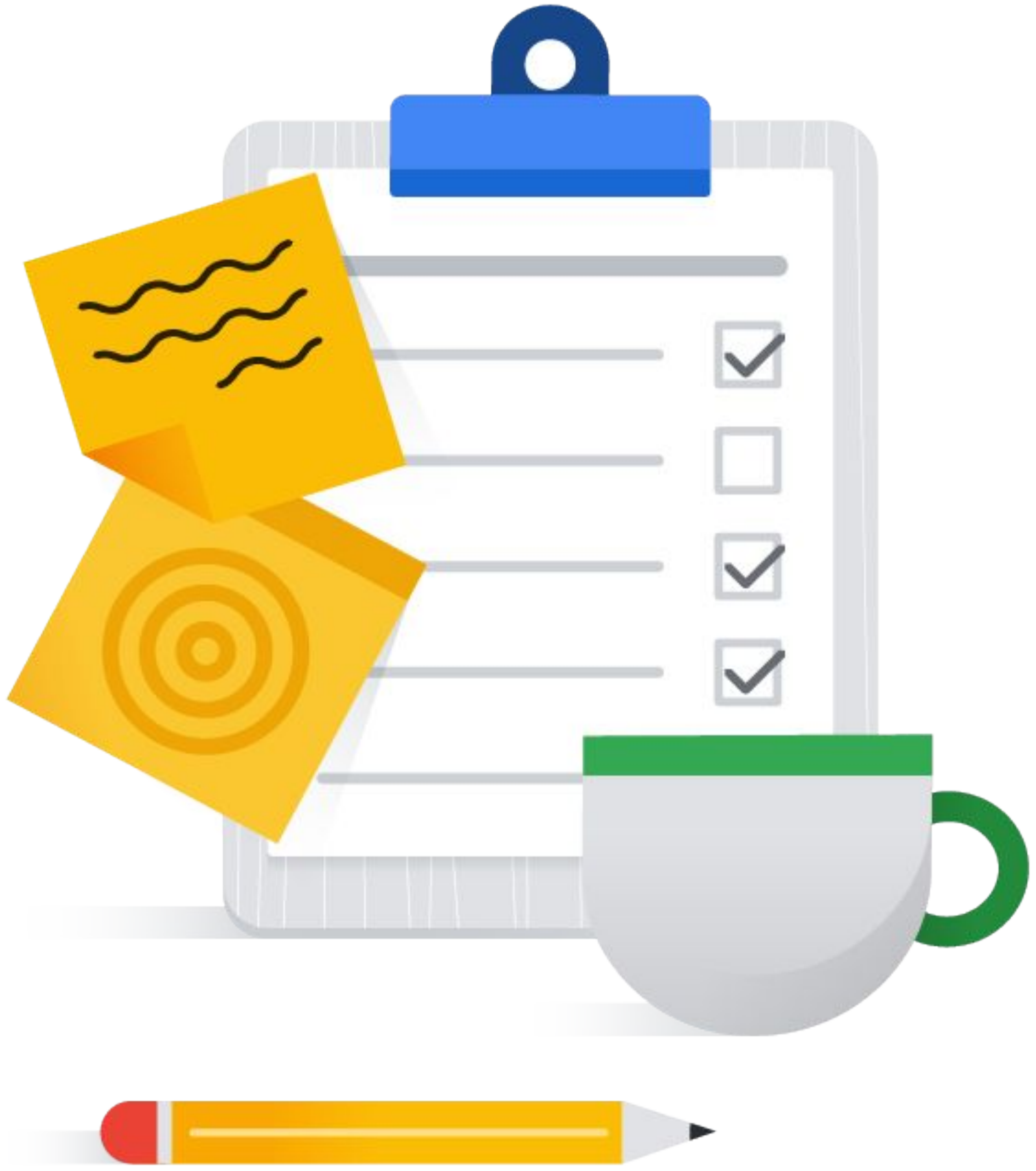
Speech

- Speech to text
- Text to speech

Vision

- Image Q&A
- Image generation
- Image editing
- Captioning
- Image search
- Video descriptions

Topics

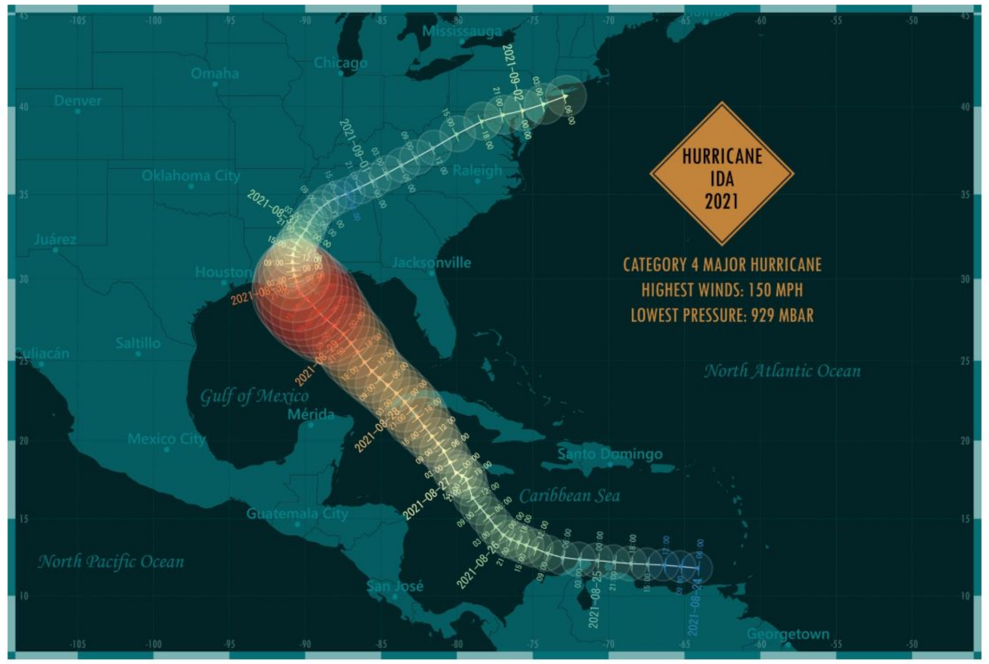


Gemini is a multimodal foundation model trained on text, images, video and audio

- Prompts can contain a combination of text, images, and video
- Capable of performing a wide range of text and vision-related tasks
  - Generate text
  - Extract text from images and video
  - Caption images, video, or audio
  - Understand and respond to questions about video, text and audio
- Multiple versions:
  - **Pro**: Most advanced
  - **Flash**: Nearly as good and faster and cheaper

Gemini's multimodal capabilities mean it can understand images, graphics & tables

Prompt

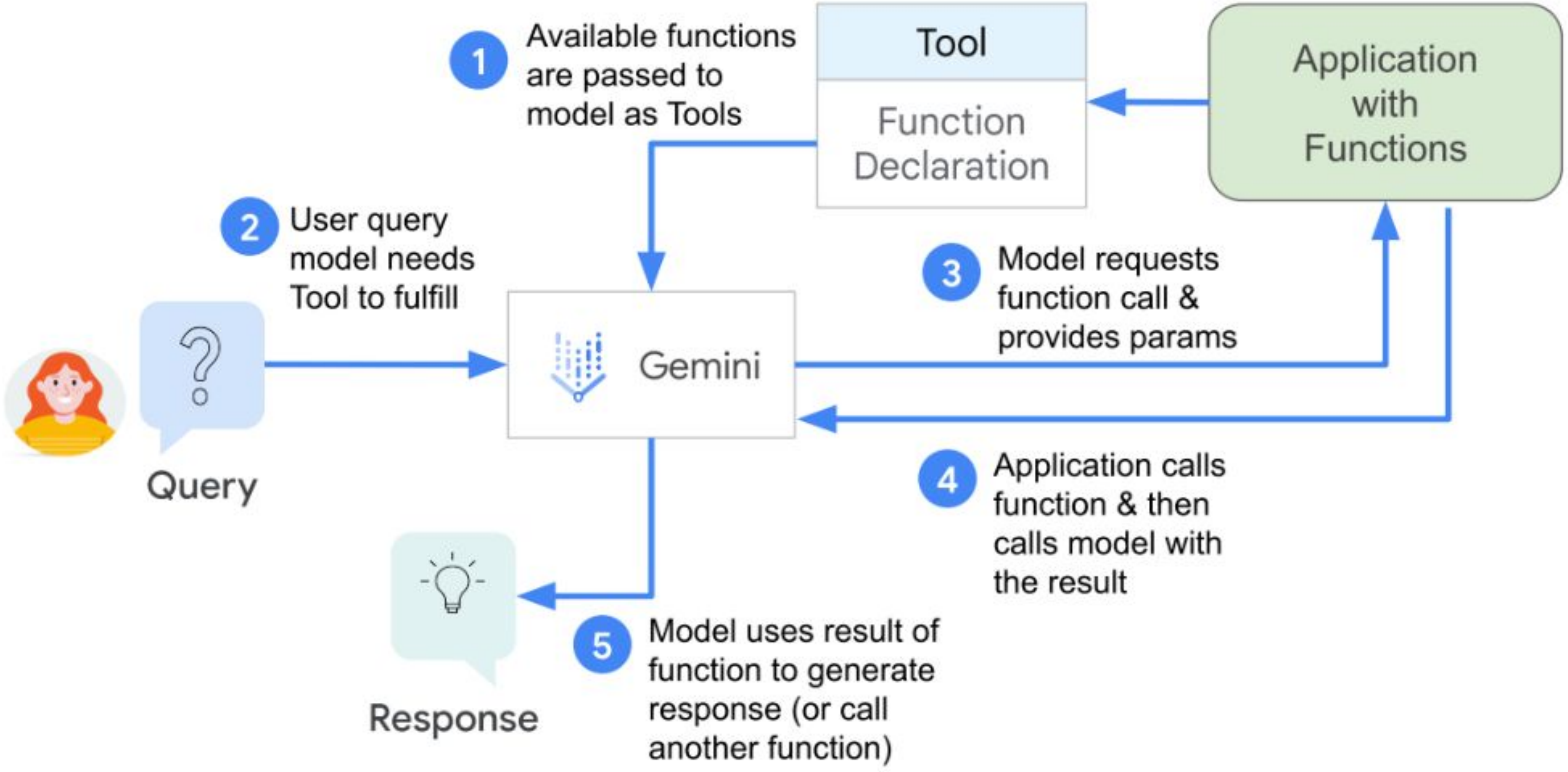


What major event is represented in the image? Which state did it have a severe impact on and when did it make landfall? Answer all questions in bullet points with just the answer, do not use complete sentences.

Response

- Hurricane Ida
- Louisiana
- August 29, 2021

Function Calling lets Gemini pass your system a request for a function to be called, then use the result you return



Gemini for Google Cloud Portfolio

Application Lifecycle

Efficiently manage cloud applications

Gemini Cloud Assist

Preview

Software Development

Accelerate software delivery

Gemini Code Assist

Security

Elevate security expertise

Gemini in Security

Data Analytics

Fast-track data analysis

Gemini in BigQuery

Business Intelligence

Automate data Insights

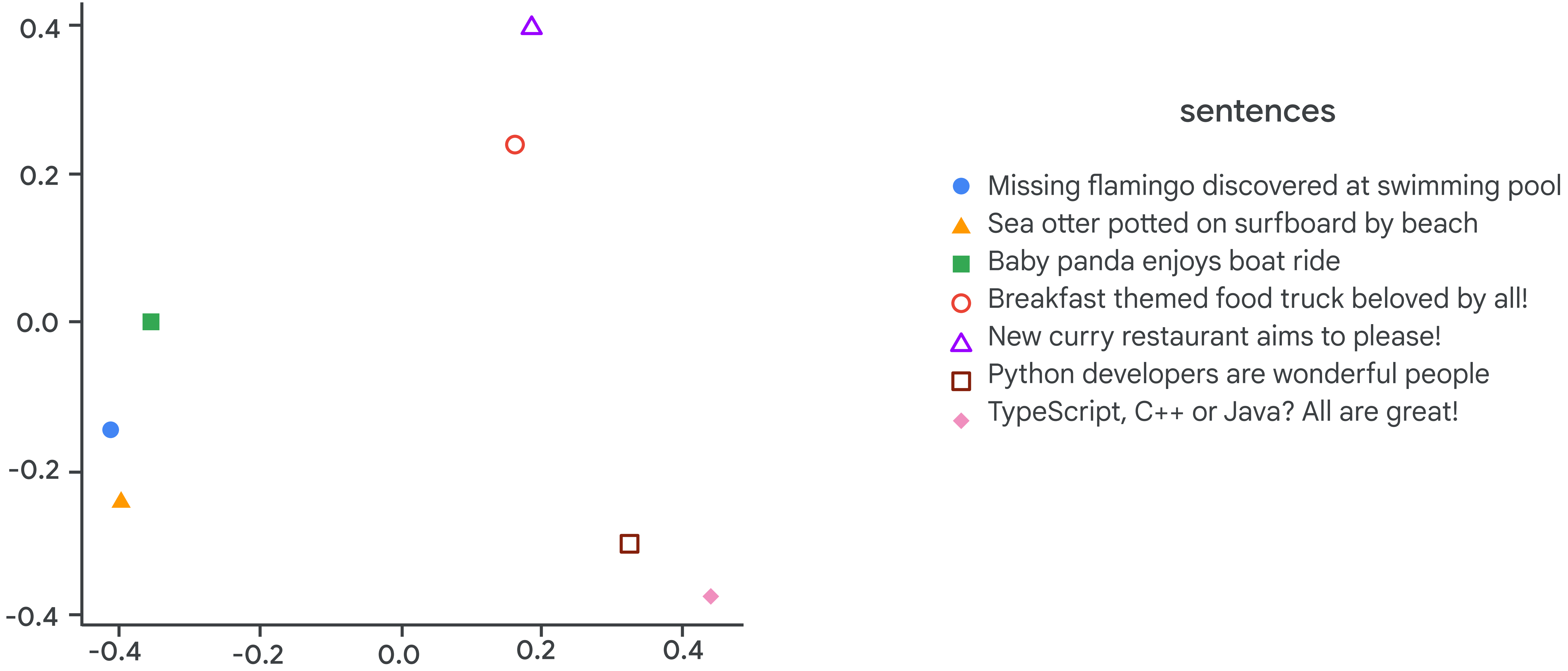
Gemini in Looker

Databases

Supercharge database development & management

Gemini in Databases

The Embeddings API returns embeddings for text, text-multilingual or multimodal prompts



Gemma is a family of lightweight, open models built from the same technology used to create the Gemini

- Small enough to run on mobile devices, desktop and laptop computers, and your own servers
- Comes in multiple flavors
  - **Gemma 2**: The latest text-only version
  - **PaliGemma**: Image + text as input, text as output
  - **CodeGemma**: Further trained on code & math
  - **RecurrentGemma**: A distinct model focusing on memory efficiency
- Can deploy using Vertex AI Model Registry and Model Endpoints

You can try out Gemma on your own computer

- Go to https://ollama.com/ and download and install the program
- Go to your terminal and type:
  `ollama run gemma`
- Enter a prompt to try it out

When should you deploy Gemma on a project?

- On systems that can’t connect to the Cloud for security or latency reasons
- On edge devices
- For experimenting with whole foundation-model tuning

Imagen is Google's Foundation model for Vision

- Imagen is capable of performing a wide range of vision-related tasks
  - Generate an image
  - Edit a masked section of an image
  - Caption an image
  - Visual Q&A (Answer questions about an image)
  - The documentation shows a feature roadmap with more features planned

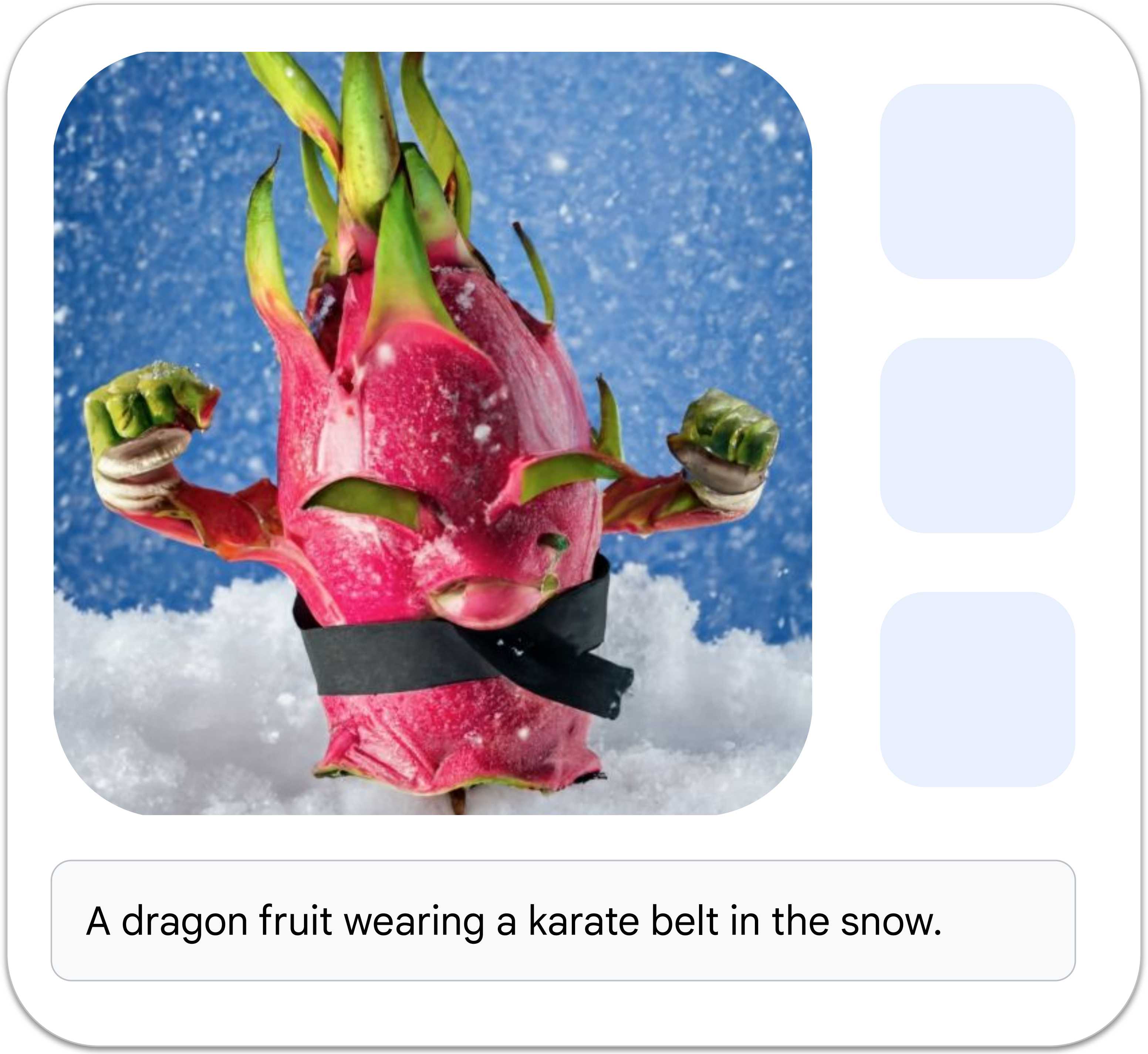


Imagen uses diffusion-based techniques to generate images

Text

Frozen Text Encoder

Text Embedding

Text-to-Image
Diffusion model

64 x 64 Image

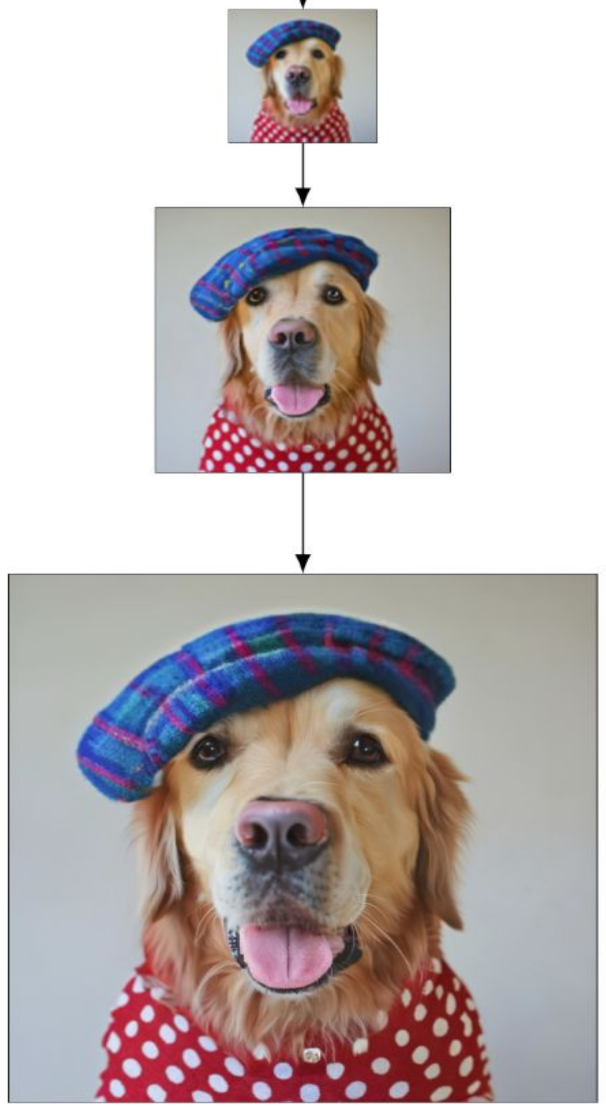
Super-Resolution
Diffusion model

256 x 256 Image

Super-Resolution
Diffusion model

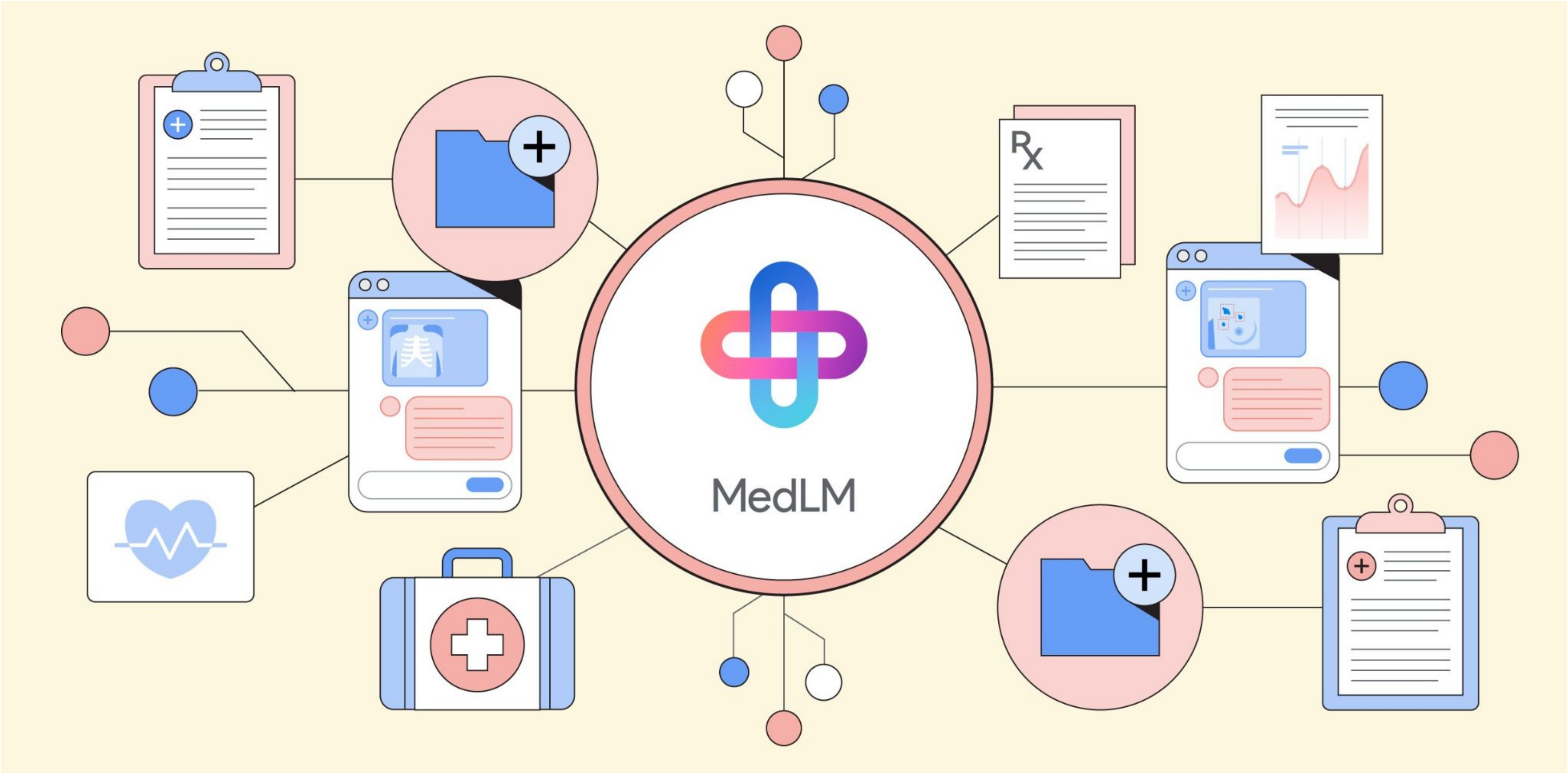
1024 x 1024 Image

“A Golden Retriever dog wearing a blue beret and red dotted turtleneck.”



Diffusion models iteratively refine a noise-filled image to approximate the target image distribution.

MedLM is a HIPAA-compliant suite of medically tuned models and APIs powered by Google Research



To use Vertex AI models in a web or mobile app, investigate using Firebase GenKit

- Designed for app developers to integrate generative AI models into Firebase web or mobile applications
- Currently supports JavaScript/Typescript (Node.js) with Go support in active development

```
import { gemini15Flash } from '@genkit-ai/vertexai';
import { generate } from '@genkit-ai/ai';

const result = await generate({
 model: gemini15Flash,
 config: { temperature: 0.3, maxOutputTokens: 200 },
 prompt: 'What makes you the best LLM out there?',
});
console.log(result.text());
```

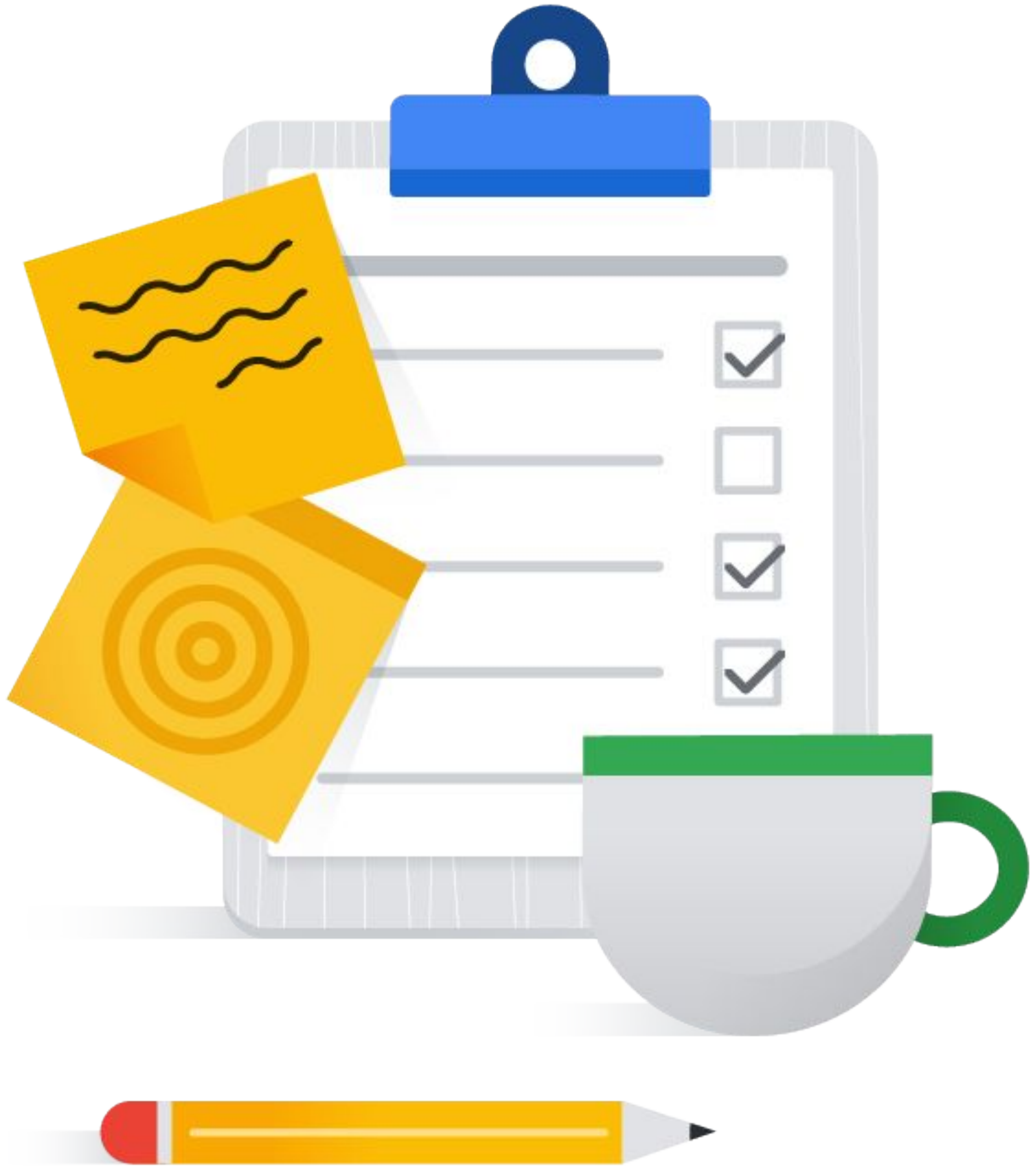
Firebase Genkit

And other models are available in Model Garden to run on Google Cloud!

You can call Anthropic's Claude natively on Vertex AI using its API or through LangChain

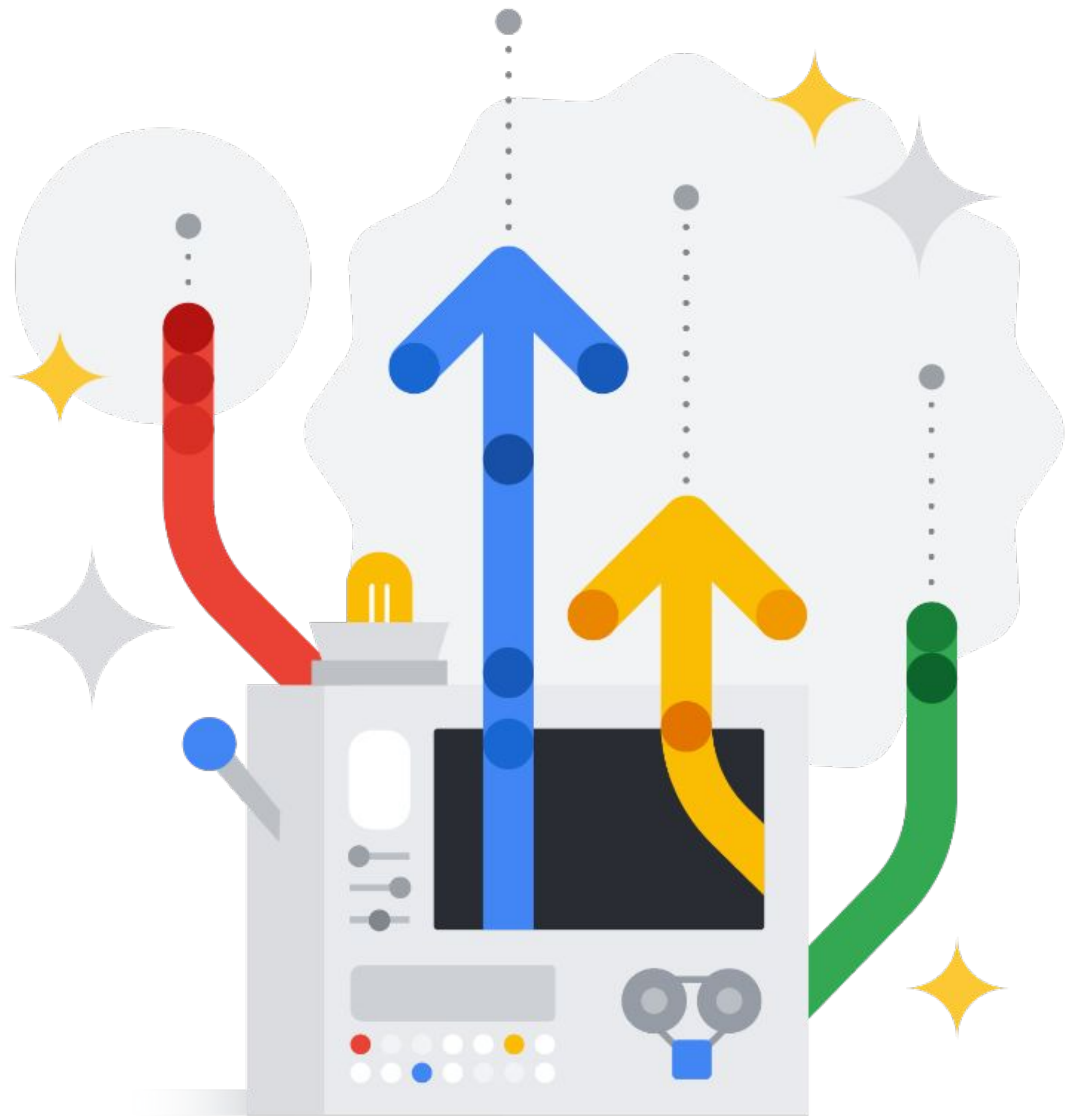
```
from anthropic import AnthropicVertex

client = AnthropicVertex(region=”us-east5”, project_id=PROJECT_ID)
message = client.messages.create(
   max_tokens=1024,
   messages=[
       {
           "role": "user",
           "content": "Send me a recipe for banana bread.",
       }
   ],
   model="claude-3-5-sonnet@20240620")
print(message.model_dump_json(indent=2))
```

Topics



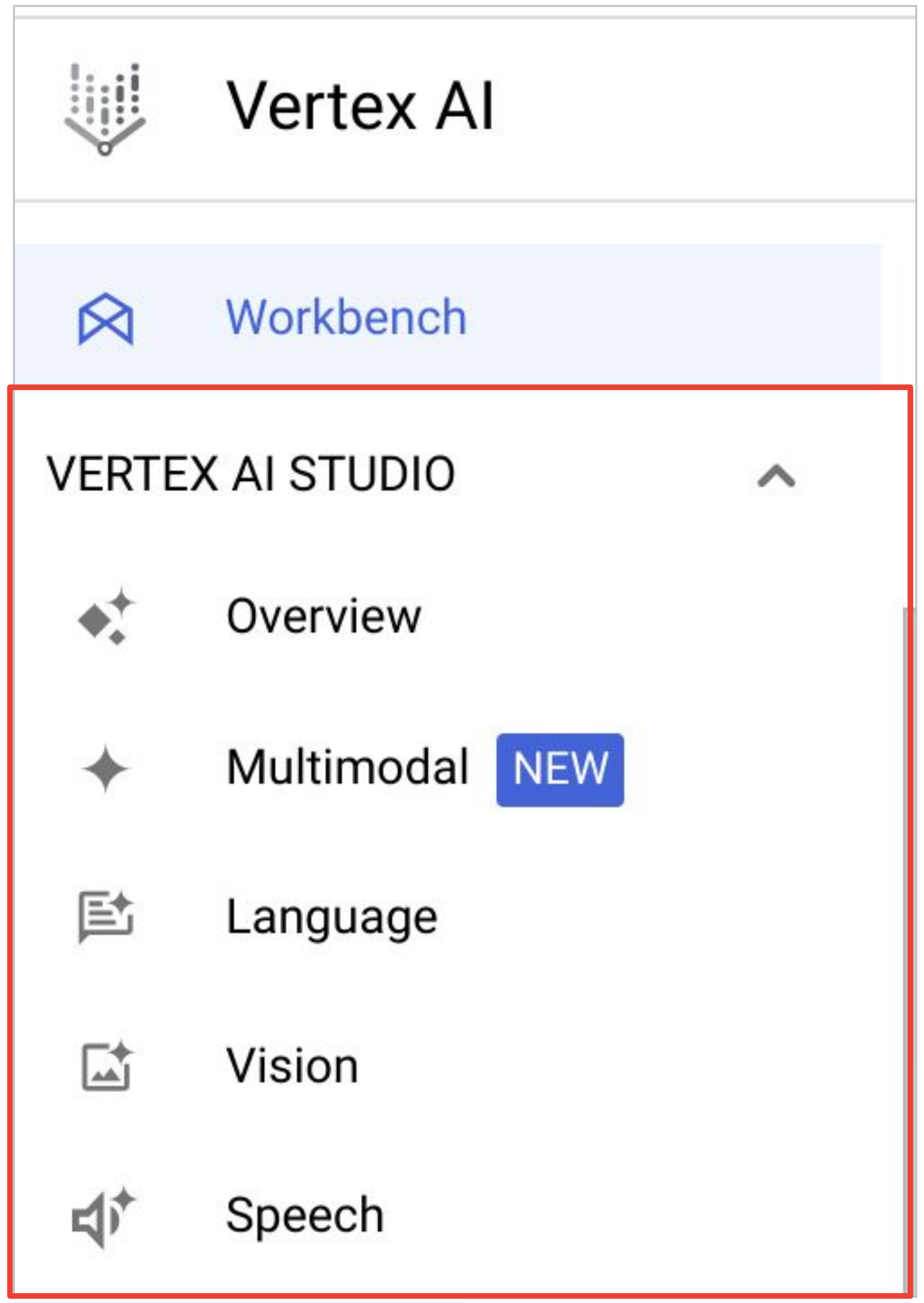
Vertex AI Studio with Google foundation models

- Supports Google foundational models
  - Gemini Pro and Flash for text and code generation
  - Gemini Pro Vision for Image and Video Q&A
  - PaLM 2 models for text and chat
  - Chirp for speech to text
  - Imagen for text to image generation
- Allows users to easily experiment with prompts
  - Simple, intuitive design
  - Easily experiment with parameters
  - Add context and examples

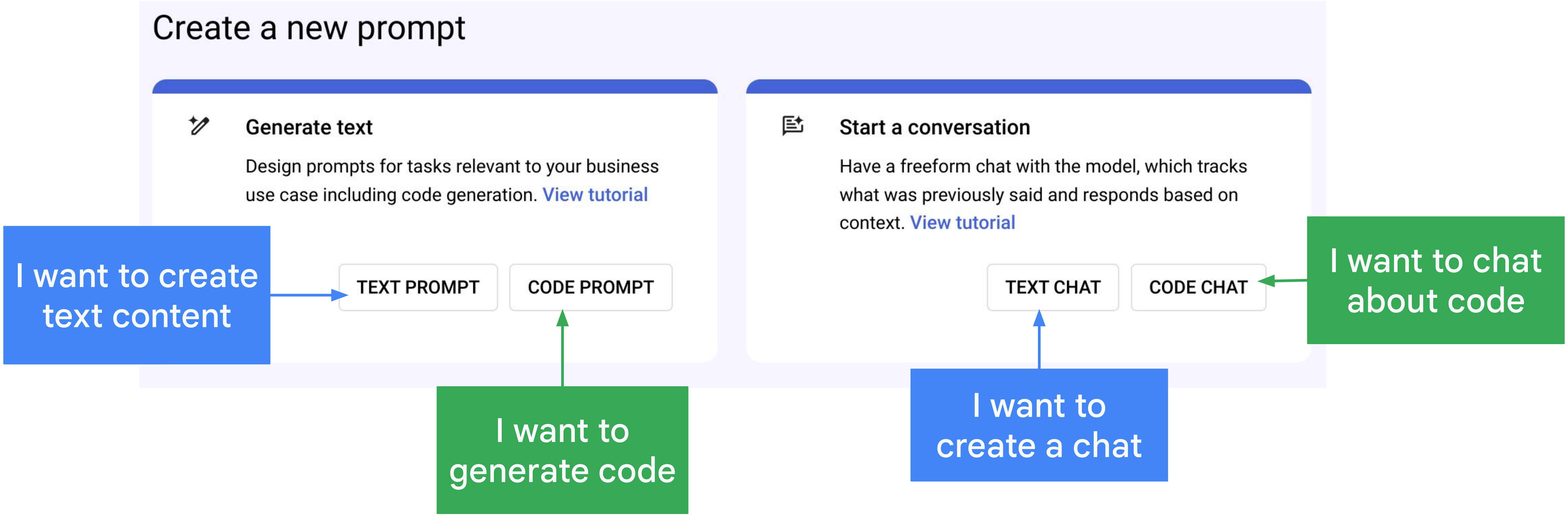


Vertex AI Studio

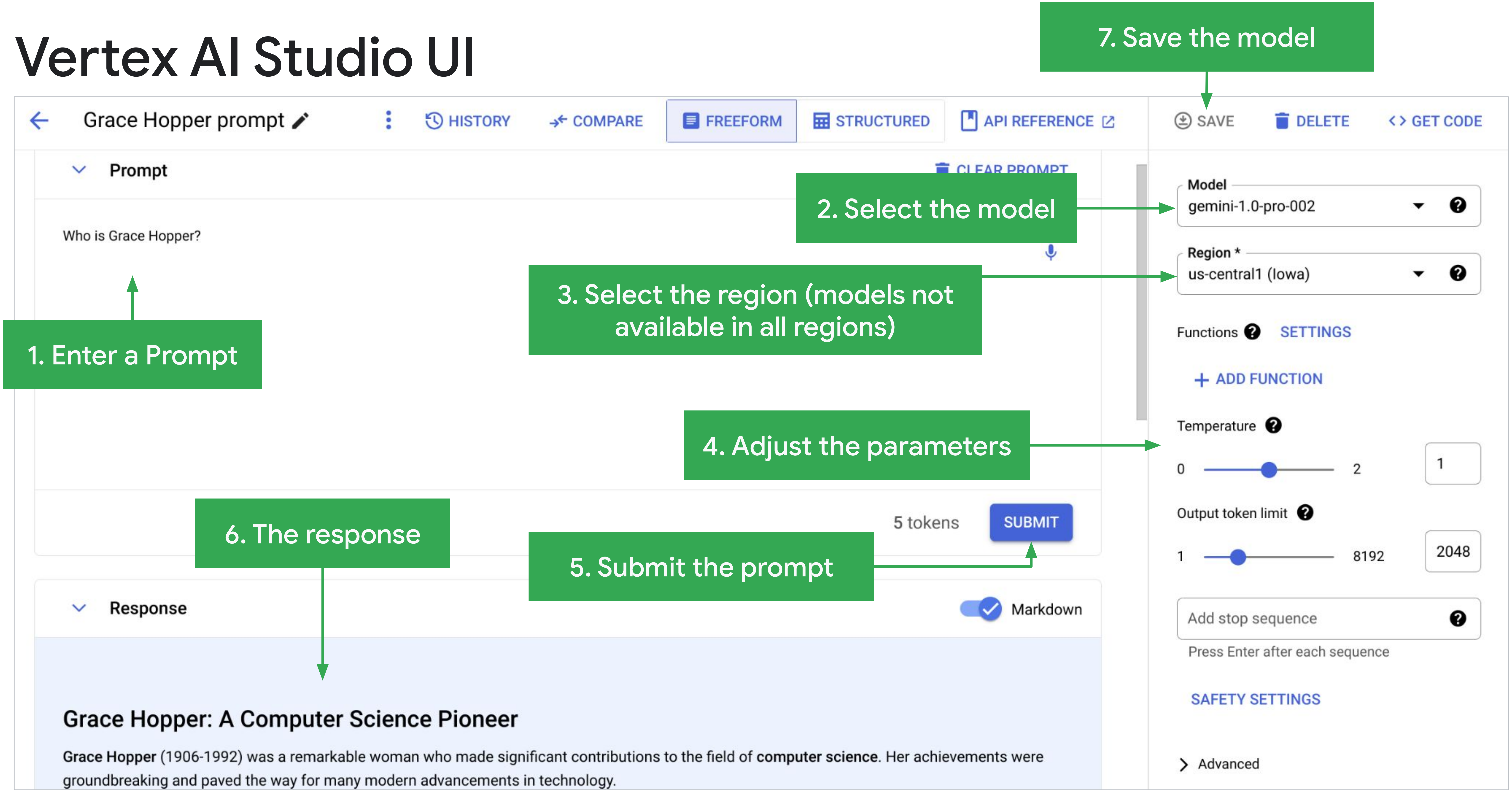
- Vertex AI Studio is available as a feature of Google Cloud Vertex AI
- Choose from Multimodal, Language, Vision, or Speech models



To get started, choose a task based on your goal



Vertex AI Studio UI



Test a simple text prompt to see the results

Prompt

Tell me about Grace Hopper

Response Markdown

Grace Hopper was an American computer scientist and United States Navy rear admiral. She was one of the pioneers of computer programming, and is credited with coining the term "debugging". She also helped develop the first compiler, which translates human-readable code into machine code that computers can understand.

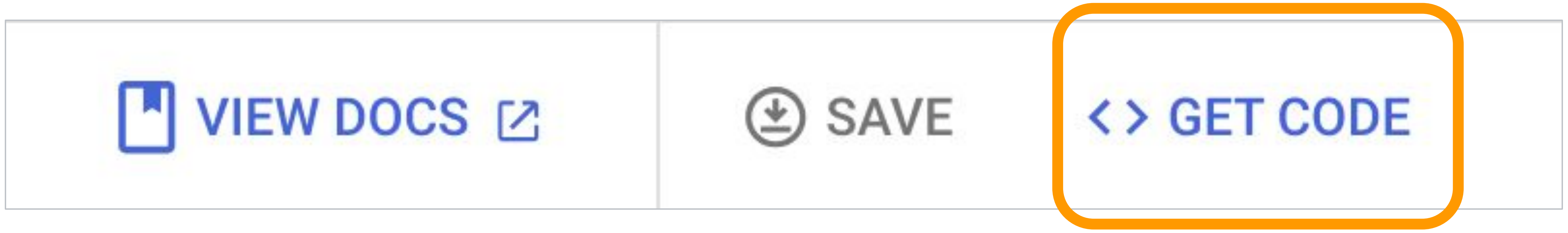
Hopper was born in New York City in 1906. She attended Vassar College, where she studied mathematics and physics. After graduating, she taught mathematics at Vassar for a year before joining the United States Navy in 1943.

During World War II, Hopper worked on the Mark I computer, one of the first electronic computers. She was responsible for developing the first compiler, which translated human-readable code into machine code that computers could understand. This work made it much easier for programmers to write code for computers, and helped to make computers more accessible to a wider range of people.

After the war, Hopper continued her work in computer science. She helped develop the COBOL programming language, which is still widely used today. She also served as the director of the Navy's computer science research center.

In 1986, Hopper was awarded the National Medal of Technology for her contributions to computer science

Use < > Get Code from Vertex AI Studio to provide template code



Get package installation and code with any parameters or safety settings you’ve adjusted in a few languages

Get code

PYTHON | NODE.JS | JAVA | CURL

Use this script to request a model response in your application.

1. Install the Vertex AI SDK.

```
npm install https://github.com/googleapis/nodejs-vertexai
gcloud auth application-default login
```

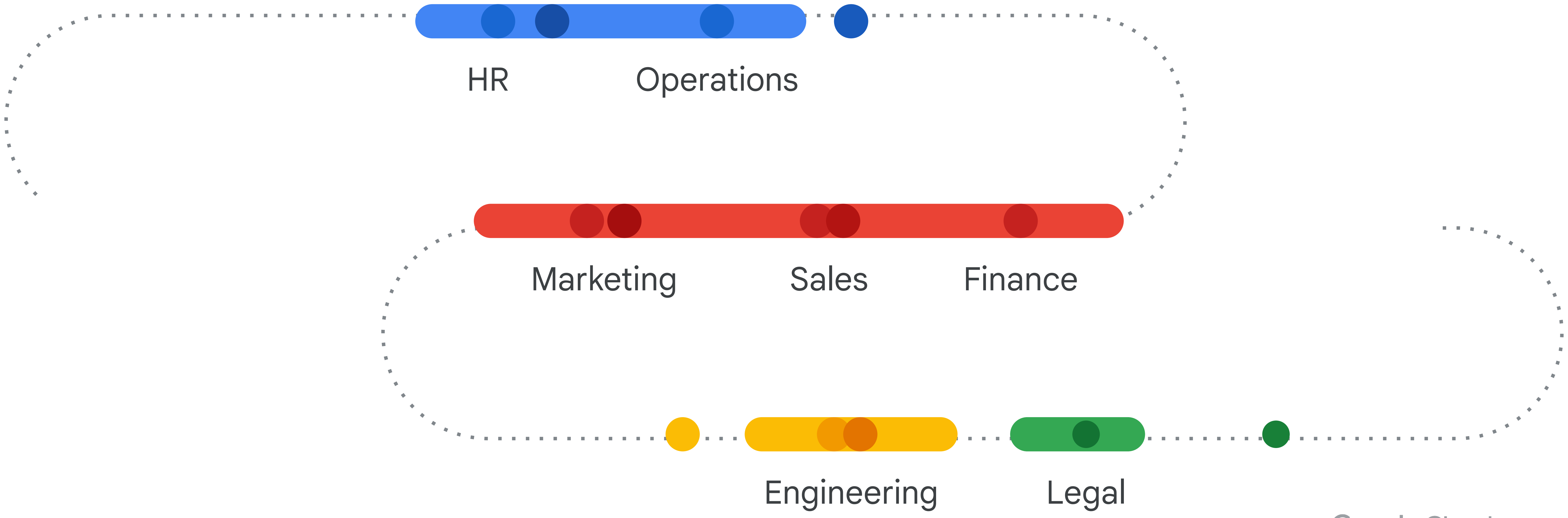
2. Create an index.js file and add the following code:

```
const {VertexAI} = require('@google-cloud/vertexai');

// Initialize Vertex with your Cloud project and location
const vertex_ai = new VertexAI({project: 'vertext-ai-dar', location: 'us-central1'});
const model = 'gemini-pro';

// Instantiate the models
const generativeModel = vertex_ai.preview.getGenerativeModel({
  model: model,
  generation_config: {
    "max_output_tokens": 2048,
    "temperature": 0.9,
    "top_p": 1
  },
});

async function generateContent() {
  const req = {
    contents: [{role: 'user', parts: [{text: 'Tell me a funny joke'}]}],
  };

  const streamingResp = await generativeModel.generateContentStream(req);
```

Topics

There are many opportunities for Generative AI across many organizations

- Content creation
- Marketing and advertising
- Customer service
- Education and research
- Many more...



1: Software development

Expedite the coding process by providing developers with the ability to generate, troubleshoot, and conduct unit testing for their code base

$11.9M - 23.8M[1]

in potential cost savings opportunities

+ Decrease time spent writing and troubleshooting code
+ Simplify code documentation process
+ Automate unit test case development

Value calculation (annualized)

Number of developers	6,349	
Salary per developer	$150,000	
Salary paid to developers	$952,350,000	
Percentage increase in developer productivity	10.0%	20.0%
Potential reduction in developer cost from improved productivity	$95,235,000	$190,470,000
% of developers usage	50.0%	
Percentage of value realize	25.0%	
	Conservative	*Top-end*
Estimated annual cost savings with software development GenAI	$11,904,375	$23,808,750

2: Media content search

Reduce subscriber churn from improved user search experience (e.g., ability to find more relevant content)

$32M - 65M[1]

in potential revenue improvement opportunities

+ Recapture of revenue that would have been lost from customer churn due to lack of relevant content

Value calculation (annualized)

Number of subscribers	238,390,000	
Subscriber churn rate (%)	2.40%	
Number of churned subscribers	5,721,360	
Potential decrease in subscriber churn from better search experience	5.0%	10.0%
Number of subscribers saved	286,068	572,136
Average subscription cost per user per month	$12	
Revenue from saved subscribers	$40,392,843	$80,785,686
EBITDA margin (%)	80.0%	
	Conservative	*Top-end*
Estimated annual profits from subscribers saved	$32,314,274	$64,628,549

2: Media content search

Increase advertisement inventory from growth in watch time due to improved user search experience

$36M - 71M[1]

in potential revenue improvement opportunities

+ Increase in ads revenue by increasing average watch time for ad-supported subscribers

Value calculation (annualized)

Number of subscribers	238,390,000	
Percentage of subscribers using ad-supported tier (%)	2.00%	
Number of subscribers using ad-supported tier	4,767,800	
Average hours of watch time per week	22	
Number of hours of watch time	5,553,533,440	
Potential increase in watch time	2.5%	5.0%
Hours increase in watch time due to better search experience	138,838,336	277,676,672
Number of ads per hour of watch time	8	
Average CPM (per 1,000 impressions) ($)	$40.00	
Revenue from incremental ads due to more watch time	$44,428,268	$88,856,535
EBITDA margin (%)	80.0%	
	Conservative	*Top-end*
Esimated annual profits from more ads	$35,542,614	$71,085,228

3: Retail customer experience

Decrease live chat cost with chat containment

$4.4M - 17.5M[1]

in potential cost consolidation opportunities

+ Enable customers with simple questions to get them answered without having to wait and engage directly with a live service agent

Value calculation (annualized)

Annual website chat session	73,000,000	
Current chat containment %	25.00%	
Number of live chat sessions	54,750,000	
Potential increase in chat containment	5.0%	20.0%
Number of additional live chat sessions contained	2,737,500	10,950,000
Cost per live chat	$1.60	
	Conservative	*Top-end*
Estimated cost savings from containment	$4,380,000.00	$17,520,000.00

3: Retail customer experience

Increase contact center agent and store employee productivity and improve customer experience with LLM-powered assistance

$28M - 55M[1]

in potential cost consolidation opportunities

+ Improve productivity of customer-facing staff, allowing them to dedicate more hours in the day to help and support end-customers

Value calculation (annualized)

Number of contact center agents	5,000	
Number of store representatives	50,000	
Average salary per contact center agents and store representatives	$40,000	
Salary paid to contact center agents and store representatives	$2,200,000,000	
Percentage increase in productivity	5.0%	10.0%
Potential cost savings with increased productivity	$110,000,000	$220,000,000
Percentage value realize	25.00%	
	Conservative	*Top-end*
Estimated cost savings from productivity increase	$27,500,000	$55,000,000

3: Retail customer experience

Drive incremental sales with better customer service by agents (e.g., better knowledge, faster response) that improve customer experiences

$7.3M - 25.5M[1]

in potential revenue improvement opportunities

+ Enable customer service agents to be more informed and better equipped to answer product questions, increasing the speed and likelihood of sales

Value calculation (annualized)

Annual company revenue ($)	$45,000,000,000	
Percentage of sales with customer service agents engagement	16.20%	
Annual company revenue impacted by customer service agents	$7,290,000,000	
Potential increase in revenue impacted by customer service agents	2.0%	7.0%
Estimated revenue increase	$145,800,000	$510,300,000
EBITDA margin (%)	5.0%	
	Conservative	*Top-end*
Estimated annual profits from better customer service by agents	$7,290,000	$25,515,000

4: Health Insurance Customer Service

Improve customer experience and provide relief for customer service centers by introducing a responsive patient service

$21M - 63.2M[1]

in potential financial impact

- Decrease **10-30%** live agent calls from containment[2]
- Increase contact center agent productivity by **10-30%**[2]
- Revenue recapture from improved call abandonment by **10-30%**[2]

15,000,000	Number of inbound calls, annually
10%	Current call containment rate (%)
$10.00	Cost per call
5,000	Number of contact center agent
$30,000	Annual salary per CC agent
5%	% of inbound calls attempt to make payment
20%	Current call abandonment rate (%)
$50	Revenue capture per call
10%	EBITDA margin

4: Health Insurance Customer Service

Improve customer experience and provide relief for customer service centers by introducing a responsive patient service

$7.5M - 22.5M[1]

in potential cost savings opportunities

+ **Increase contact center agent productivity by 10-30%[2]**
+ Decrease **10-30%** live agent calls from containment[2]
+ Revenue recapture from improved call abandonment by **10-30%**[2]

Value calculation (annualized)

Number of contact center agents	5,000	
Salary per contact center agent	$30,000	
Salary paid to contact center agents	$150,000,000	
Percentage increase in agent productivity	10.0%	30.0%
Potential reduction in developer cost from improved productivity	$15,000,000	$45,000,000
Percentage of value realize	50.0%	
	Conservative	*Top-end*
Estimated annual cost savings with agent productivity improvement	$7,500,000	$22,500,000

4: Health Insurance Customer Service

Decrease live call cost with call containment

$13.5 - 40.5M[1]

in potential cost consolidation opportunities

- **Decrease 10-30% live agent calls from containment[2]**
- Increase contact center agent productivity by 10-30%[2]
- Revenue recapture from improved call abandonment by **10-30%**[2]

Value calculation (annualized)

Number of inbound calls, annually	15,000,000	
Current chat containment %	10.00%	
Number of live calls	13,500,000	
Potential increase in chat containment	10.0%	30.0%
Number of additional live chat sessions contained	1,350,000	4,050,000
Cost per live chat	$10	
	Conservative	*Top-end*
Estimated cost savings from containment	$13,500,000	$40,500,000

5: Automotive conversational manual and assistant

Provide conversational manual and assistant to assist drivers in automotive vehicles

$5.8M - 17.6M[1]

in potential financial impact

+ Decrease **10-30%** live agent calls from containment[2]
+ Increase revenue and upsell car features and services by **5-15%**

5,000,000	Number of inbound calls from conversation manual, annually
$8.00	Cost per call
$150B	Annual revenue
2.5%	Percentage of revenue from car services and features
10%	% of car services and features revenue from conversation manual
10%	EBITDA margin

6: Financial Services Analyst Research

Improve financial services research analyst productivity by enabling them to query bodies of financial data and reports in natural language

$400 - 700k[1]

in potential cost consolidation opportunities

+ Drive Financial Research Analyst productivity by decreasing time spent searching for information
+ Improve employee satisfaction

Value calculation (annualized)

Number of Financial Research Analyst	250	
Salary per Analyst	$80,000	
Salary paid to developers	$20,000,000	
Time spent searching for information[2]	20%	
Percentage increase in productivity[3]	20.0%	35.0%
Potential reduction in analyst cost from improved productivity	$800,000	$1,400,000
Percentage value realization	50.00%	
	Conservative	Top-end
Estimated annual cost savings with increased productivity	$400,000	$700,000

7: Marketing Campaign Planning

Automate and expedite the Marketing campaign planning of briefs, creatives and assets across multi-channel platforms

$35M-70M+[1]

in potential cost consolidation opportunities

- Drive marketing campaign creation productivity by decreasing time and money spent on campaign planning, briefs, and asset creation
- Automate Multi-Channel Campaigns

Value calculation (annualized)

Total Annual Revenue	$50,000,000,000	
% of revenue spent on marketing[1]	3.00%	
Marketing Spend	$1,500,000,000	
Marketing Spend (TV Only) %	42.00%	
Marketing Spend excluding TV spend	$870,000,000	
% Campaign Planning Costs	20%	
Total Campaigning Planning Costs minus 20% for conservative estimates	$139,200,000	
Potential cost savings[2]	25.0%	50.0%
	Conservative	Top-end
Estimated annual cost savings with marketing GenAI	$34,800,000	$69,600,000

8: Knowledge Workers Productivity

Assist knowledge workers to search for information and to get answers to questions quicker

$562M - 788M

in potential financial impact opportunities

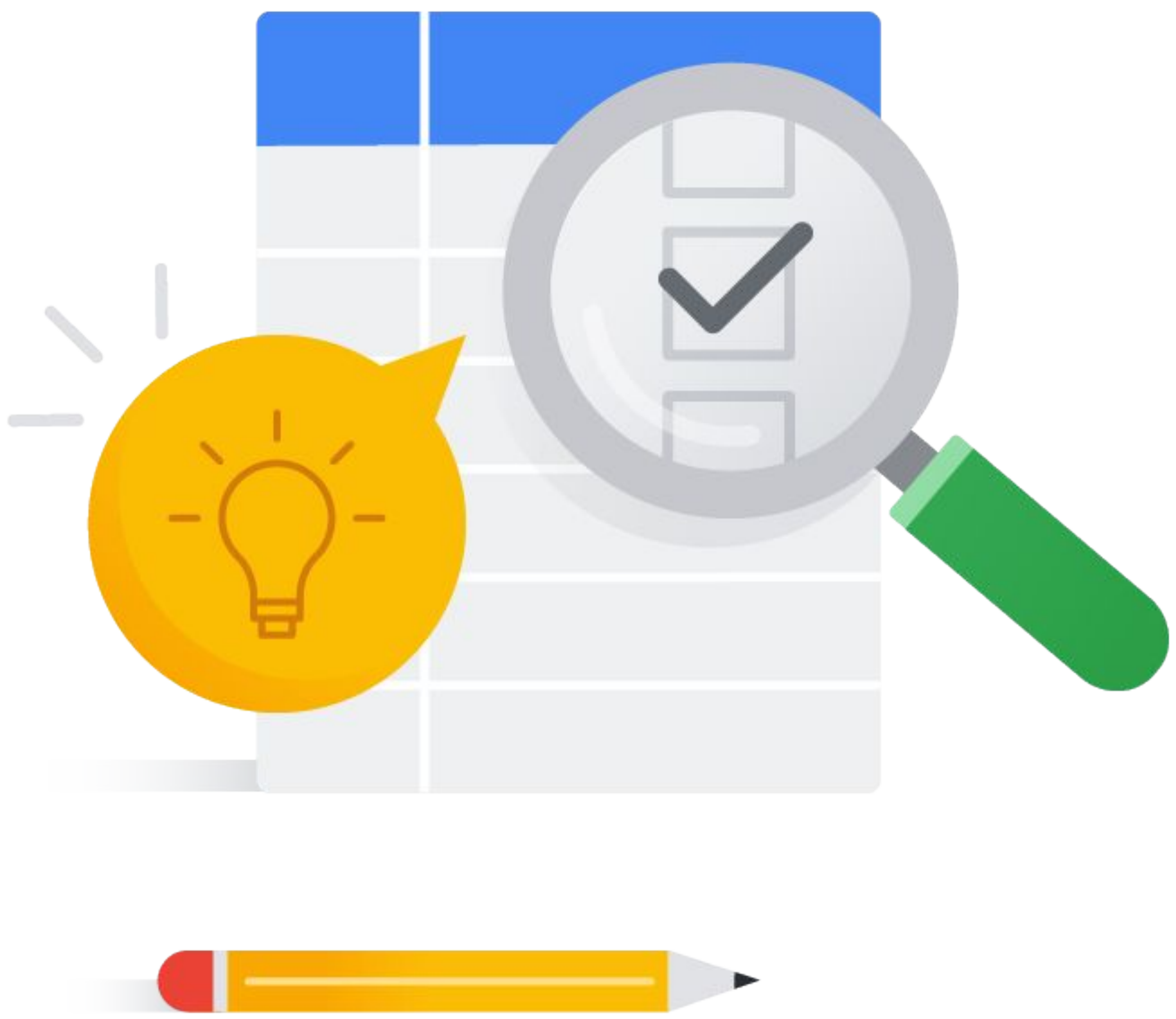
- **Increase Knowledge Workers productivity**
- Improve employee experience

Value calculation (annualized)

Number of employees + contract workers	150,000	
Percentage of overall employees who are Knowledge workers	33%	
Annual salary per Knowledge Employees	$80,000	
Annual salary paid to Knowledge Employees	$12,000,000,000	
Percentage of time searching for information[2]	25%	
Percentage of employees GenAI tool usage	75%	
Percentage time reduction searching for information with GenAI tools[3,4]	50.0%	70.0%
Potential reduction in knowledge employees cost from improved productivity	$1,125,000,000	$1,575,000,000
Percentage value realization	50.00%	
	Conservative	Top-end
Estimated annual cost savings with Specialized Knowledge Workers	$562,500,000	$787,500,000

Lab

1 hour

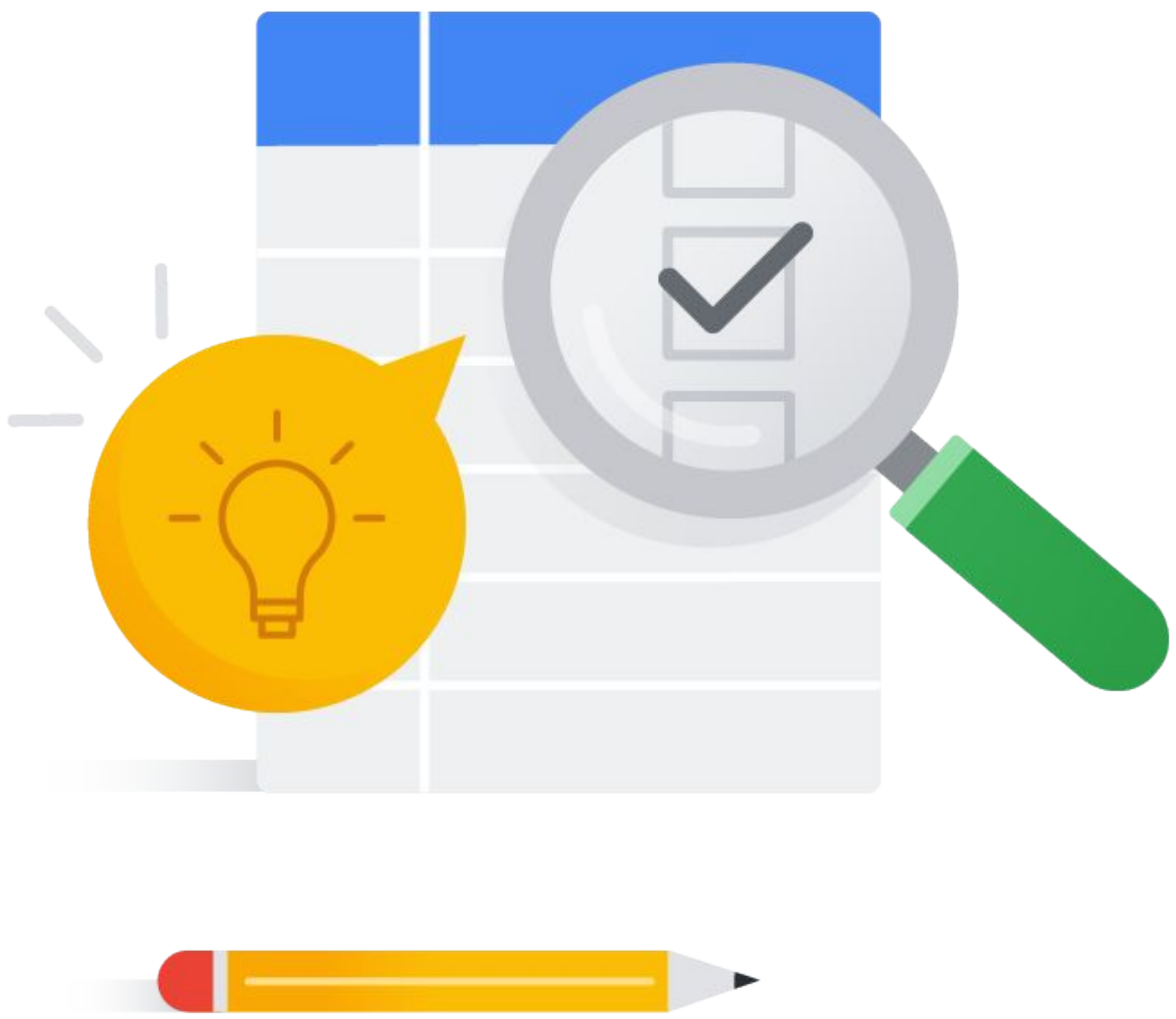


Lab: Multimodality with Gemini

Lab

1 hour

Lab: Explore and Evaluate Models using Model Garden



In this module, you learned to ...

01 Leverage foundation models for generative AI

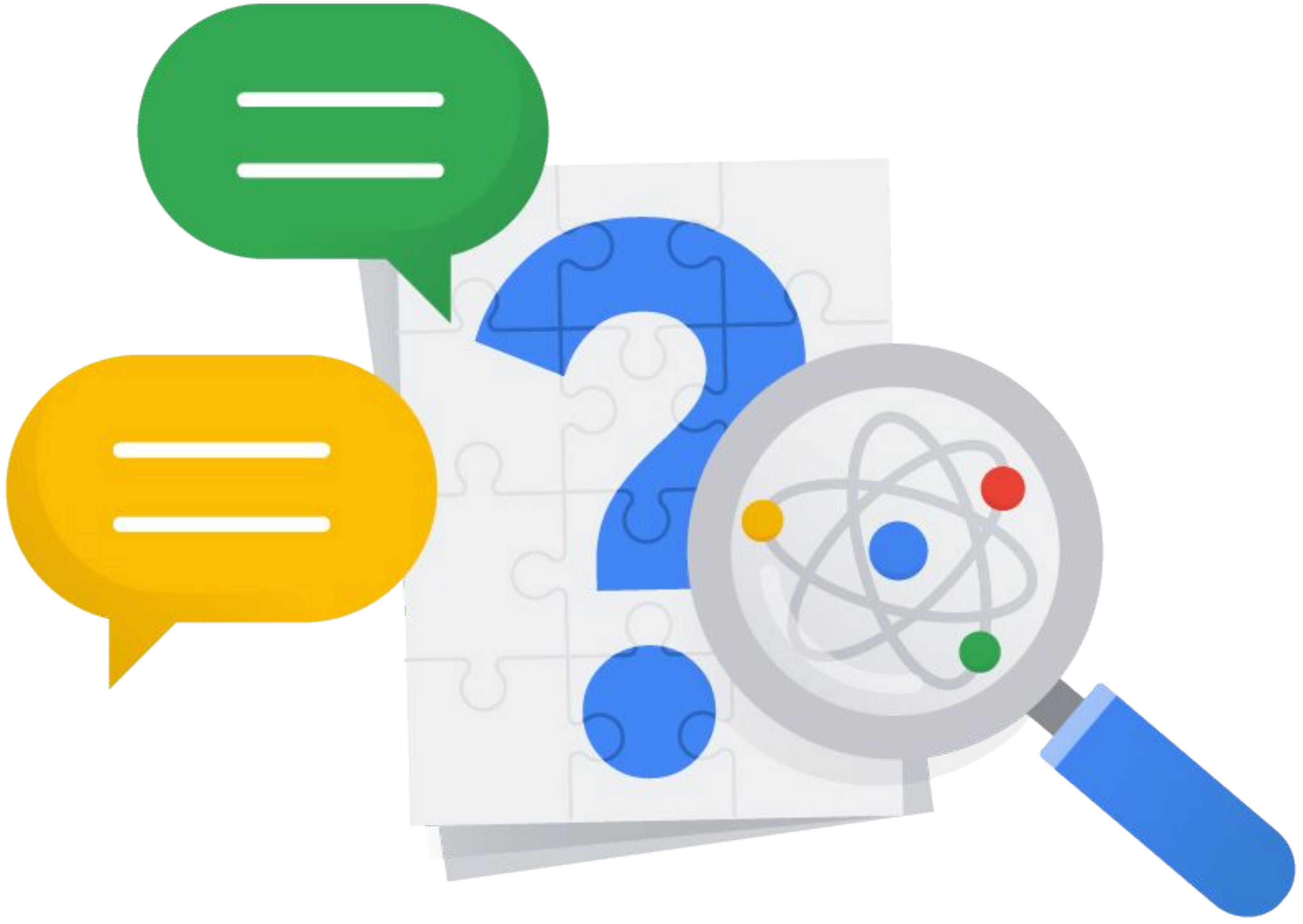
02 Use Google's foundation models to optimize generative AI tasks

03 Start prompting with Vertex AI Studio

04 Explore use cases for generative AI

05 Find and implement models using Vertex AI Model Garden

Questions and answers



Quiz question

Which of the following are tuned versions of the PaLM model, optimized for specific use cases? (Choose all that apply)

A: Codey

B: Chat-Bison

C: Med PaLM

D: Sec PaLM

E: ChatGPT

F: Imagen

G: Dall-e

Quiz question

Which of the following are tuned versions of the PaLM model, optimized for specific use cases? (Choose all that apply)

A: Codey

B: Chat-Bison

C: Med PaLM

D: Sec PaLM

E: ChatGPT

F: Imagen

G: Dall-e

Quiz question

What tool could you use to find an appropriate ML model based on a use case, and find documentation for a chosen model?

A: AutoML

B: Model Garden

C: Workbench

D: Vertex AI Studio

Quiz question

What tool could you use to find an appropriate ML model based on a use case, and find documentation for a chosen model?

A: AutoML

B: Model Garden

C: Workbench

D: Vertex AI Studio

Quiz question

A large language model like Google PaLM or GPT could perform which of the following ML tasks? (Choose all that apply)

A: Text generation

B: Summarization

C: Image generation

D: Speech to text

E. Sentiment analysis

Quiz question

A large language model like Google PaLM or GPT could perform which of the following ML tasks? (Choose all that apply)

A: Text generation

B: Summarization

C: Image generation

D: Speech to text

E. Sentiment analysis

Google Cloud