Designing Systems that use AI

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 Break complex processes into discrete tasks

02 Determine if and how AI can be used to help humans perform their tasks

03 Architect AI solutions that leverage Google Cloud tools

Topics

01	Agent Builder
02	Automating Business Processes with AI
03	Product Catalog Management Case Study
04	Architecting the Catalog Management Solution

Agent Builder allows you to build four types of generative AI-powered apps in a few clicks

Select app type

Select the type of application you want to create

Search

Get quality results out-of-box and easily customize the engine

SELECT

Chat

Answer complex questions out-of-the-box

SELECT

Recommendations

Create a content recommendation engine

SELECT

Agent PREVIEW

Built using natural language, agents can answer questions from data, connect with business systems through tools, and more

SELECT

A Search app retrieves the best-matching documents from a data store

- Indexes on embedding similarities + keywords
- Includes the option to enable the system to generate answers to follow-up questions for a full RAG experience
- You choose a type of Generic, Media, or Healthcare for tailored experiences

Search

Get quality results out-of-box and easily customize the engine

A Chat app is an out-of-the-box RAG solution that can answer questions from your data

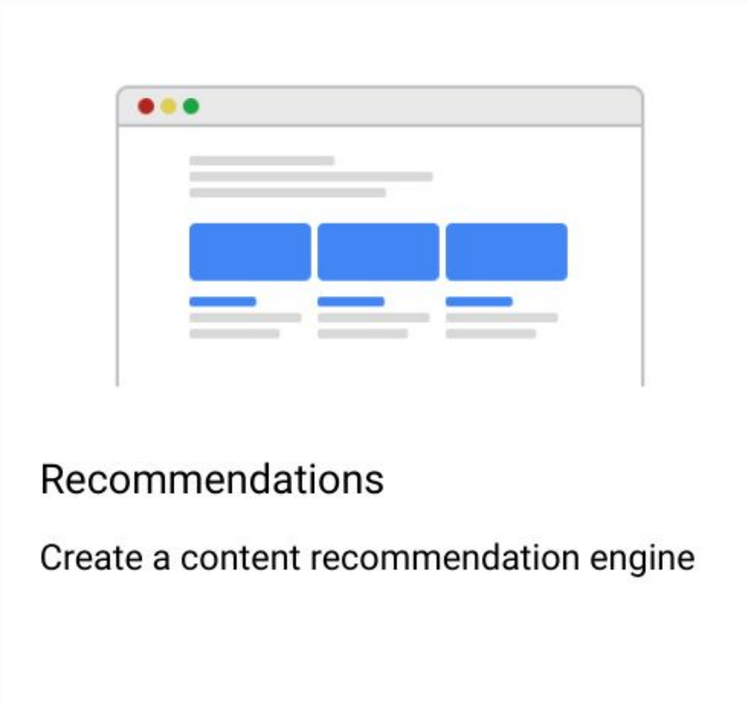
- Searches your data stores to chat or answer questions about your structured and unstructured data
- Creates a Dialogflow CX agent you can further develop
- Offers extensions to interact with it via a widget on your website, messaging on Google Maps, in Facebook Messenger, via phone call, etc.

Chat

Answer complex questions out-of-the-box

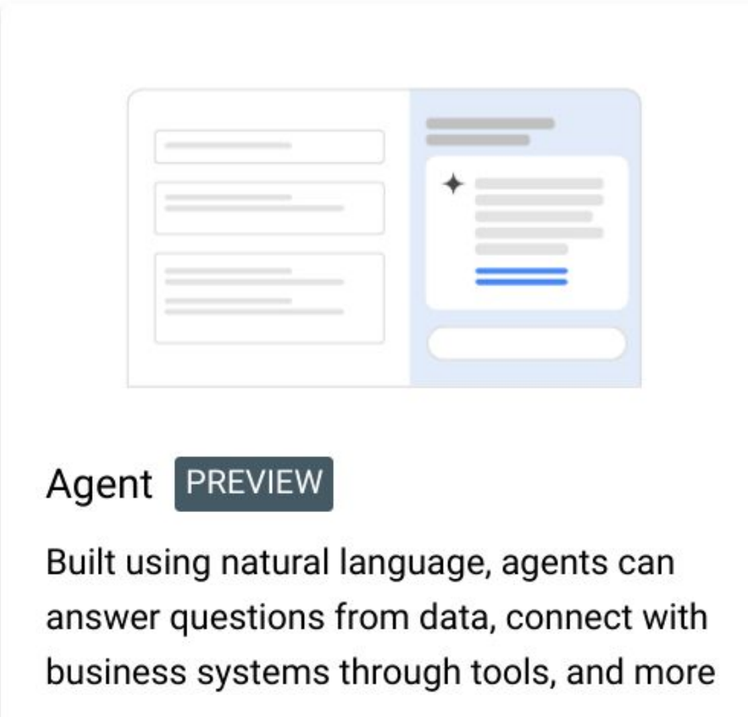
A Recommendations app can provide recommendations for similar content to users

- You choose a type of Generic, Media (like next news article or YouTube video), or Retail (for ecommerce similar products)
- For Media recommendations, you can additionally select recommendation types like “More like this” based on a given piece of content or “Recommended for you” for overall recommendations
- For Media recommendations you can also optimize recommendations for click-through rate, conversion rate, or watch duration per session

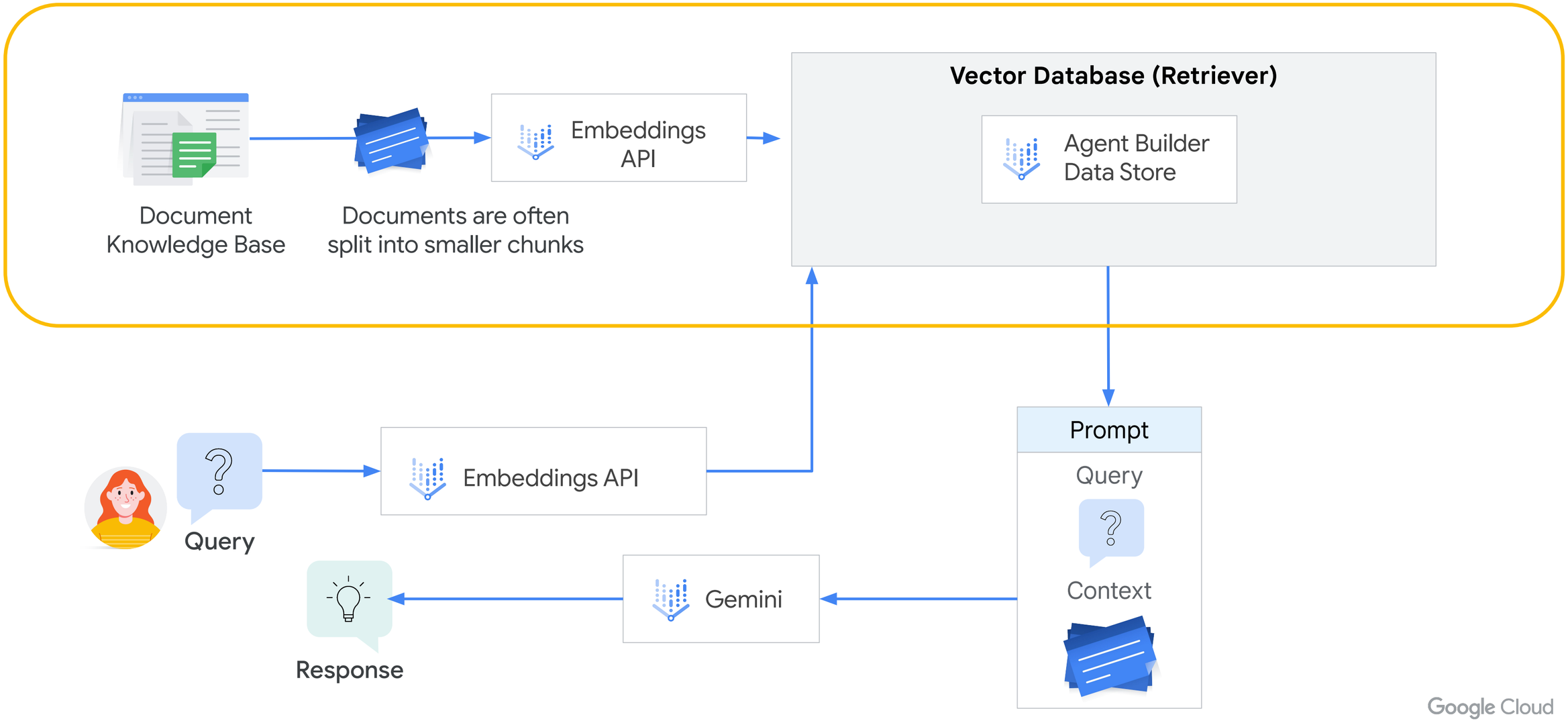


An Agent is similar to a Chat app, but empowered with function-calling tool use and a target goal

- Like Chat, creating an Agent app creates a Dialogflow agent
- Multi-step instructions for the agent can be provided to get it to follow certain steps towards a goal
- It can be provided Data stores, functions, and APIs to call to take actions

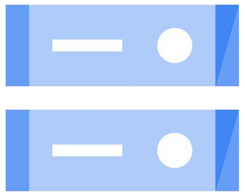


Agent Builder Data Stores



Data stores allow your apps to respond based on content from many sources

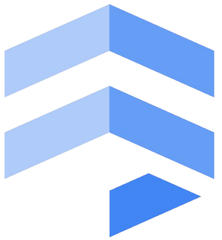
Cloud Storage



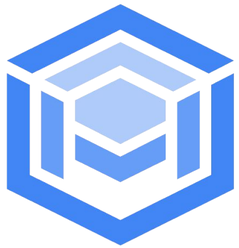
BigQuery

Website Content

Firestore

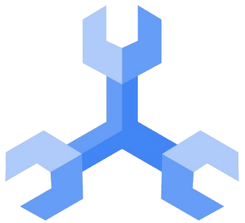


AlloyDB and AlloyDB AI



Cloud SQL

Cloud Spanner

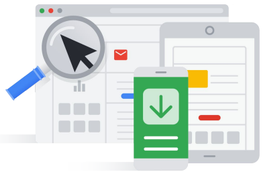


Google Drive, APIs & more

Document AI Layout Parsers empower smart loading & chunking

Understand and Extract

Smart chunking



Enhance Generative AI Apps

Built for developers

Parsing complex documents

Text

```
{“text”: {
   “content”:
      “Albert Einstein (1879-1955) was a renowned
      German-born theoretical physicist, celebrated for his
      contributions to theory of relativity and quantum
      mechanics. His famous equation E=mc² and Nobel
      Prize-winning work on the photoelectric effect have
      left an indelible mark on science.”,
   “role”: “paragraph”,
   “bounding_region”: ...}
},
```

Image

```
{“image”: {
   “uri”: “gs://vertex_ai_search/output/image1.jpeg”,
   “content”: “Figure 1. Handsome Albert.”,
   “bounding_region”: ...}
},
```

Albert Einstein

1 Introduction

1.1 Early Life and Background

Albert Einstein (1879-1955) was a renowned German-born theoretical physicist, celebrated for his contributions to theory of relativity and quantum mechanics. His famous equation $E=mc^2$ and Nobel Prize-winning work on the photoelectric effect have left an indelible mark on science.

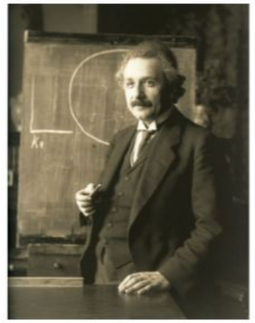


Figure 1. Handsome Albert.

2 Scientific Achievements

2.1 Theory of Relativity and Quantum Mechanics

In 1905, Einstein published groundbreaking papers on the photoelectric effect, Brownian motion, and his special theory of relativity.

$$E = mc^2$$

His work showed the equivalence of mass and energy and proposed a general theory of relativity. He made notable contributions to statistical mechanics and quantum theory.

2.2 Later Endeavors

In his late career, Einstein grappled with the randomness in quantum theory and aimed to create a unified field theory encompassing electromagnetism and gravitation. These pursuits led to his isolation from mainstream physics.

Table 1. Albert Einstein: A Profile of His Life, Scientific Legacy, and Contributions.

Category	Description
Birth	14 March 1879
Place of Birth	Ulm, Kingdom of Württemberg, German Empire
Citizenship	Kingdom of Württemberg (until 1896)
	Stateless (1896-1901)
	Switzerland (1901-1955)
	Austria (1911-1912)
	Kingdom of Prussia (1914-1918)
Education	Federal polytechnic school (Dipl., 1900)
	University of Zurich (PhD, 1905)

Parsing complex documents

Table


```
{“table”: {
      “caption”:
         “Table 1. Albert Einstein: Profile of His Life,
         Scientific Legacy, and Contributions.”,
   “row_count”: 10, “col_count”: 2,
   “cells”: [
      {“content”: “Category”, “row_index”: 0,
      “col_index”: 0, “row_span”: 1, “col_span”: 1},
      {“content”: “Description”, “row_index”: 0,
      “col_index”: 1, “row_span”: 1, “col_span”: 1},
      {“content”: “Birth”, “row_index”: 1,
      “col_index”: 0, “row_span”: 1, “col_span”: 1},
      {“content”: “14 March 1879”, “row_index”: 1,
      “col_index”: 1, “row_span”: 1, “col_span”: 1},
      ...
   ],
   “bounding_region”: ...}
},
```


Albert Einstein

1 Introduction

1.1 Early Life and Background

Albert Einstein (1879-1955) was a renowned German-born theoretical physicist, celebrated for his contributions to theory of relativity and quantum mechanics. His famous equation $E=mc^2$ and Nobel Prize-winning work on the photoelectric effect have left an indelible mark on science.

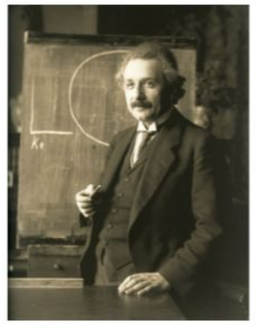


Figure 1. Handsome Albert.

2 Scientific Achievements

2.1 Theory of Relativity and Quantum Mechanics

In 1905, Einstein published groundbreaking papers on the photoelectric effect, Brownian motion, and his special theory of relativity.

$$E = mc^2$$

His work showed the equivalence of mass and energy and proposed a general theory of relativity. He made notable contributions to statistical mechanics and quantum theory.

2.2 Later Endeavors

In his late career, Einstein grappled with the randomness in quantum theory and aimed to create a unified field theory encompassing electromagnetism and gravitation. These pursuits led to his isolation from mainstream physics.

Table 1. Albert Einstein: A Profile of His Life, Scientific Legacy, and Contributions.

Category	Description
Birth	14 March 1879
Place of Birth	Ulm, Kingdom of Württemberg, German Empire
Citizenship	Kingdom of Württemberg (until 1896)
	Stateless (1896-1901)
	Switzerland (1901-1955)
	Austria (1911-1912)
	Kingdom of Prussia (1914-1918)
Education	Federal polytechnic school (Dipl., 1900)
	University of Zurich (PhD, 1905)

Detect document layout for smart chunking

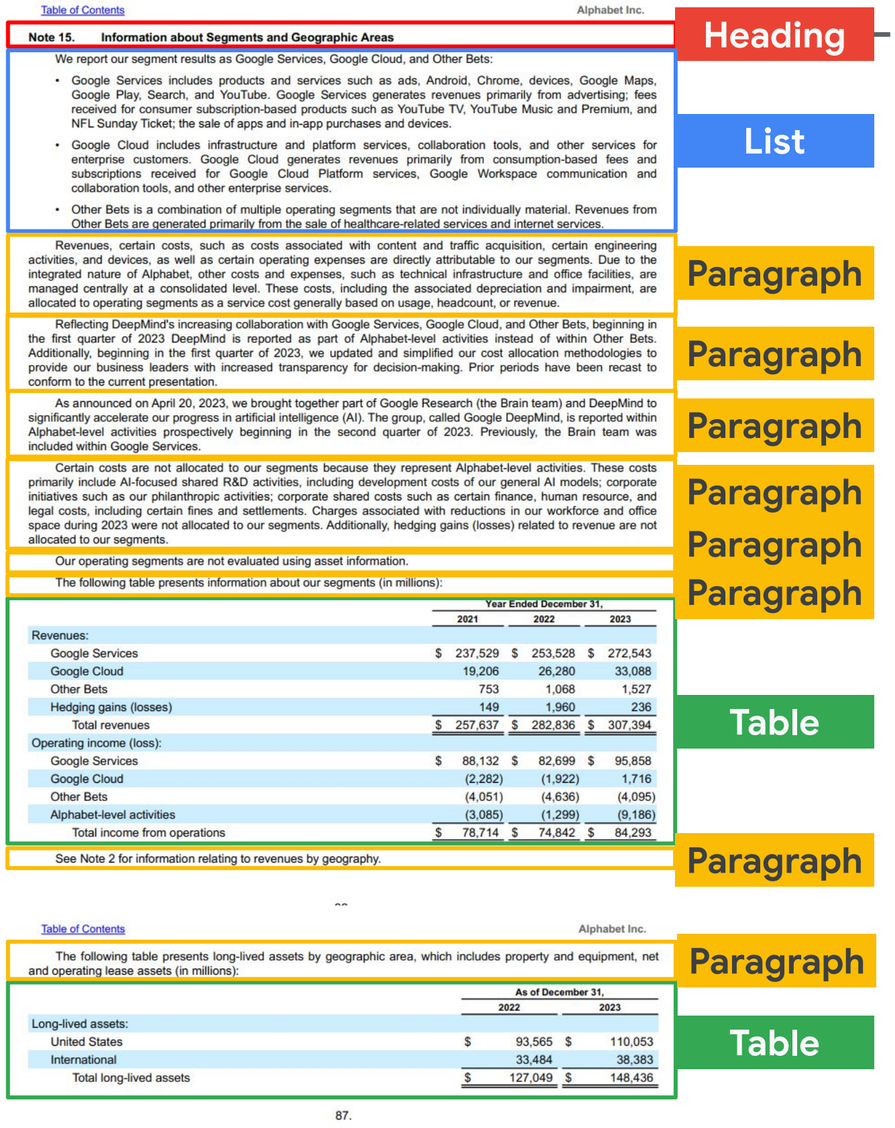


Table of Contents Alphabet Inc.

Note 15. Information about Segments and Geographic Areas

We report our segment results as Google Services, Google Cloud, and Other Bets:

- Google Services includes products and services such as ads, Android, Chrome, devices, Google Maps, Google Play, Search, and YouTube. Google Services generates revenues primarily from advertising; fees received for consumer subscription-based products such as YouTube TV, YouTube Music and Premium, and NFL Sunday Ticket; the sale of apps and in-app purchases and devices.
- Google Cloud includes infrastructure and platform services, collaboration tools, and other services for enterprise customers. Google Cloud generates revenues primarily from consumption-based fees and subscriptions received for Google Cloud Platform services, Google Workspace communication and collaboration tools, and other enterprise services.
- Other Bets is a combination of multiple operating segments that are not individually material. Revenues from Other Bets are generated primarily from the sale of healthcare-related services and internet services.

Revenues, certain costs, such as costs associated with content and traffic acquisition, certain engineering activities, and devices, as well as certain operating expenses are directly attributable to our segments. Due to the integrated nature of Alphabet, other costs and expenses, such as technical infrastructure and office facilities, are managed centrally at a consolidated level. These costs, including the associated depreciation and impairment, are allocated to operating segments as a service cost generally based on usage, headcount, or revenue.

Reflecting DeepMind's increasing collaboration with Google Services, Google Cloud, and Other Bets, beginning in the first quarter of 2023 DeepMind is reported as part of Alphabet-level activities instead of within Other Bets. Additionally, beginning in the first quarter of 2023, we updated and simplified our cost allocation methodologies to provide our business leaders with increased transparency for decision-making. Prior periods have been recast to conform to the current presentation.

As announced on April 20, 2023, we brought together part of Google Research (the Brain team) and DeepMind to significantly accelerate our progress in artificial intelligence (AI). The group, called Google DeepMind, is reported within Alphabet-level activities prospectively beginning in the second quarter of 2023. Previously, the Brain team was included within Google Services.

Certain costs are not allocated to our segments because they represent Alphabet-level activities. These costs primarily include AI-focused shared R&D activities, including development costs of our general AI models; corporate initiatives such as our philanthropic activities; corporate shared costs such as certain finance, human resource, and legal costs, including certain fines and settlements. Charges associated with reductions in our workforce and office space during 2023 were not allocated to our segments. Additionally, hedging gains (losses) related to revenue are not allocated to our segments.

Our operating segments are not evaluated using asset information.

The following table presents information about our segments (in millions):

	Year Ended December 31,		
	2021	2022	2023
Revenues:			
Google Services	$ 237,529	$ 253,528	$ 272,543
Google Cloud	19,206	26,280	33,088
Other Bets	753	1,068	1,527
Hedging gains (losses)	149	1,960	236
Total revenues	$ 257,637	$ 282,836	$ 307,394
Operating income (loss):			
Google Services	$ 88,132	$ 82,699	$ 95,858
Google Cloud	(2,282)	(1,922)	1,716
Other Bets	(4,051)	(4,636)	(4,095)
Alphabet-level activities	(3,085)	(1,299)	(9,186)
Total income from operations	$ 78,714	$ 74,842	$ 84,293

See Note 2 for information relating to revenues by geography.

Table of Contents Alphabet Inc.

The following table presents long-lived assets by geographic area, which includes property and equipment, net and operating lease assets (in millions):

	As of December 31,	
	2022	2023
Long-lived assets:		
United States	$ 93,565	$ 110,053
International	33,484	38,383
Total long-lived assets	$ 127,049	$ 148,436

87.

Heading

List

Paragraph

Paragraph

Paragraph

Paragraph

Paragraph

Paragraph

Table

Paragraph

Paragraph

Table

Chunk 1

List

Chunk 2

Paragraph

Paragraph

Paragraph

Paragraph

Paragraph

Chunk 3

Paragraph

Table

Paragraph

Chunk 4

Paragraph

Table

Chunk 1

Note 15. Information about Segments and Geographic Areas

We report our segment results as Google Services, Google Cloud, and Other Bets:
* Google Services includes products and services such as ads, Android, Chrome, devices, Google Maps, Google Play, Search, and YouTube ...
* Google Cloud includes infrastructure and platform services, collaboration tools, and other services for enterprise customers ...
* Other Bets is a combination of multiple operating segments that are not individually material ...

Chunk 4

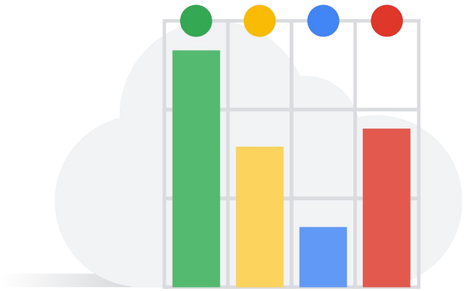
Note 15. Information about Segments and Geographic Areas

The following table presents long-lived assets by geographic area, which includes property and equipment, net and operating lease assets (in millions):

-	As of December 31,
-	2022	2023
Long-lived assets:		
United States	$ 93,565	$ 110,053
International	33,484	38,383
Total long-lived assets	$ 127,049	$ 148,436

Limitations

Maximum PDF file size:
20MB



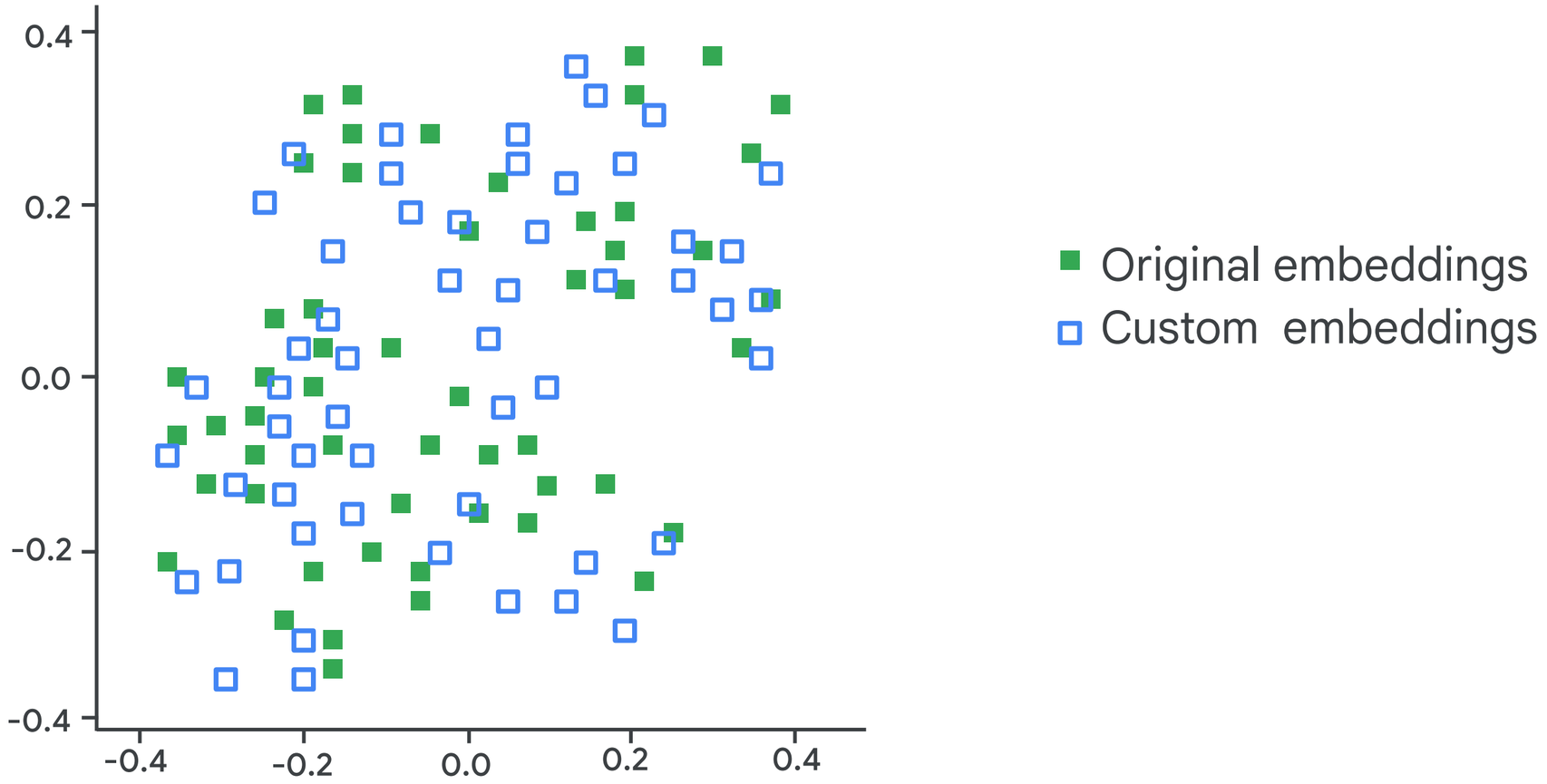
Usage limit:
100 document files per project per day

Batch processing:
Not supported

Supported file types

File type	Detected elements						
HTML	Paragraph	Table	Title	List	Heading	⊗	⊗
PDF	Paragraph	Table	Title	⊗	Heading	⊗	⊗
DOCX	Paragraph	Table	Title	List	Heading	Header	Footer

You can use custom embeddings in Agent Builder

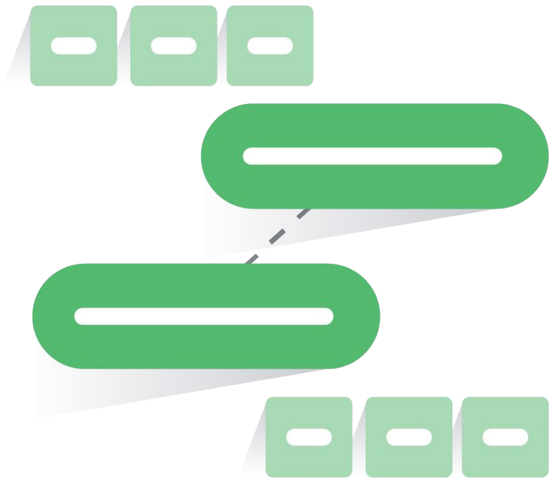


Use custom embeddings with care

Google recommends letting Vertex AI Search create and use embeddings for you

Use if:

- Embeddings trained on custom words
- Embeddings created based on user profiles for personalization

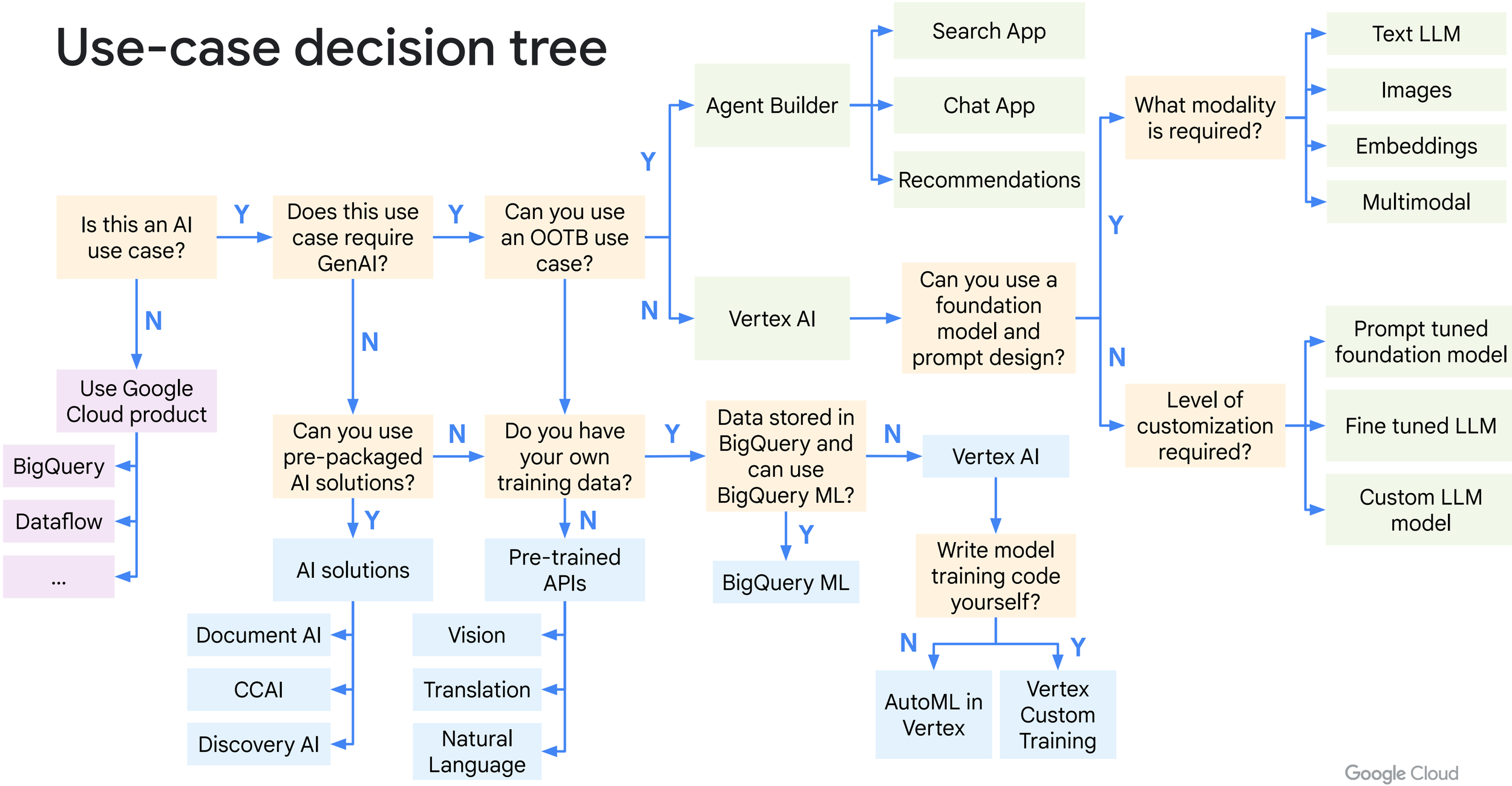


Topics

01	Agent Builder
02	Automating Business Processes with AI
03	Product Catalog Management Case Study
04	Architecting the Catalog Management Solution

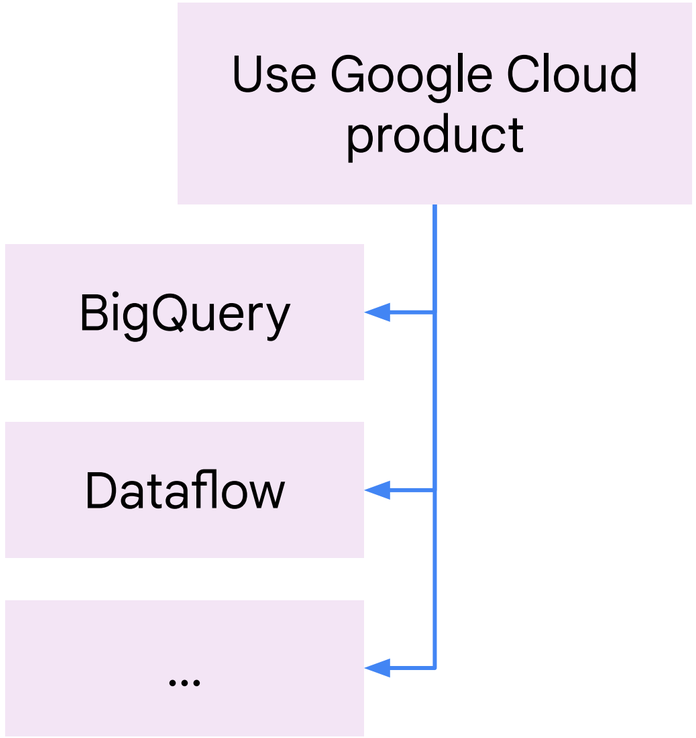
For each task where you are considering AI, consider how you would approach it both with and without AI to weigh cost and performance.

Use-case decision tree



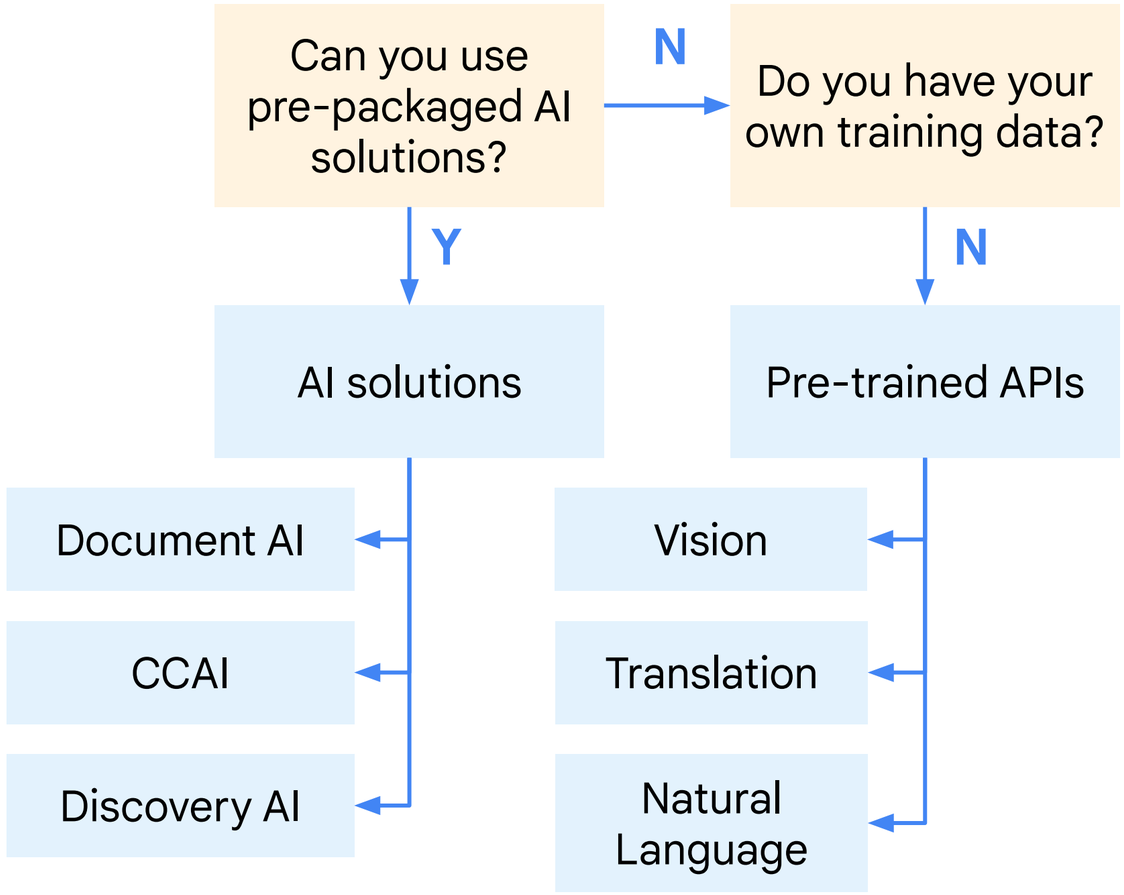
Not all automation problems are AI problems

- Sometimes there are traditional programming solutions to automating tasks
- Find the right Google product
- Use the simplest way that works



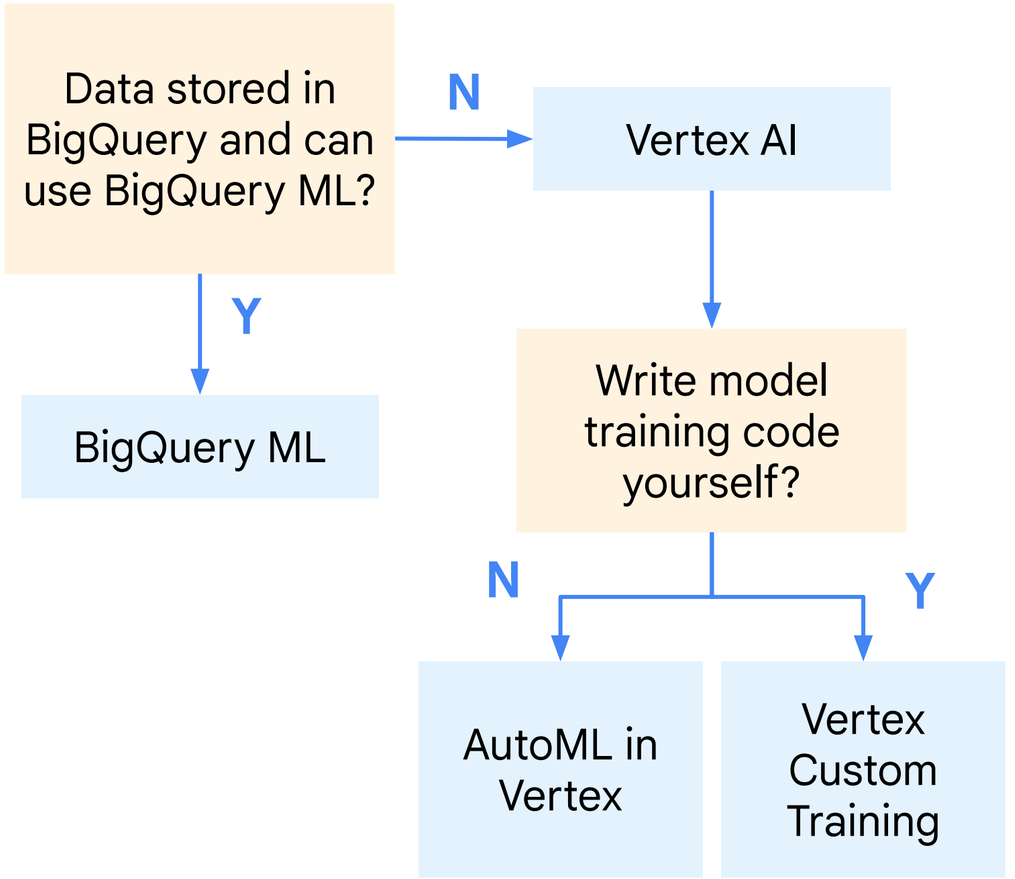
Not all AI problems are generative AI problems

- Smaller ML models will often perform an AI task as well as a foundational model
- Smaller models are cheaper and will likely provide better performance
  - Translate API for translation
  - Vision API for object detection, classification, and labels
  - Natural Language API for sentiment analysis
  - Etc.
- Google has out of the box solutions for common AI problems
  - Document AI
  - CCAI



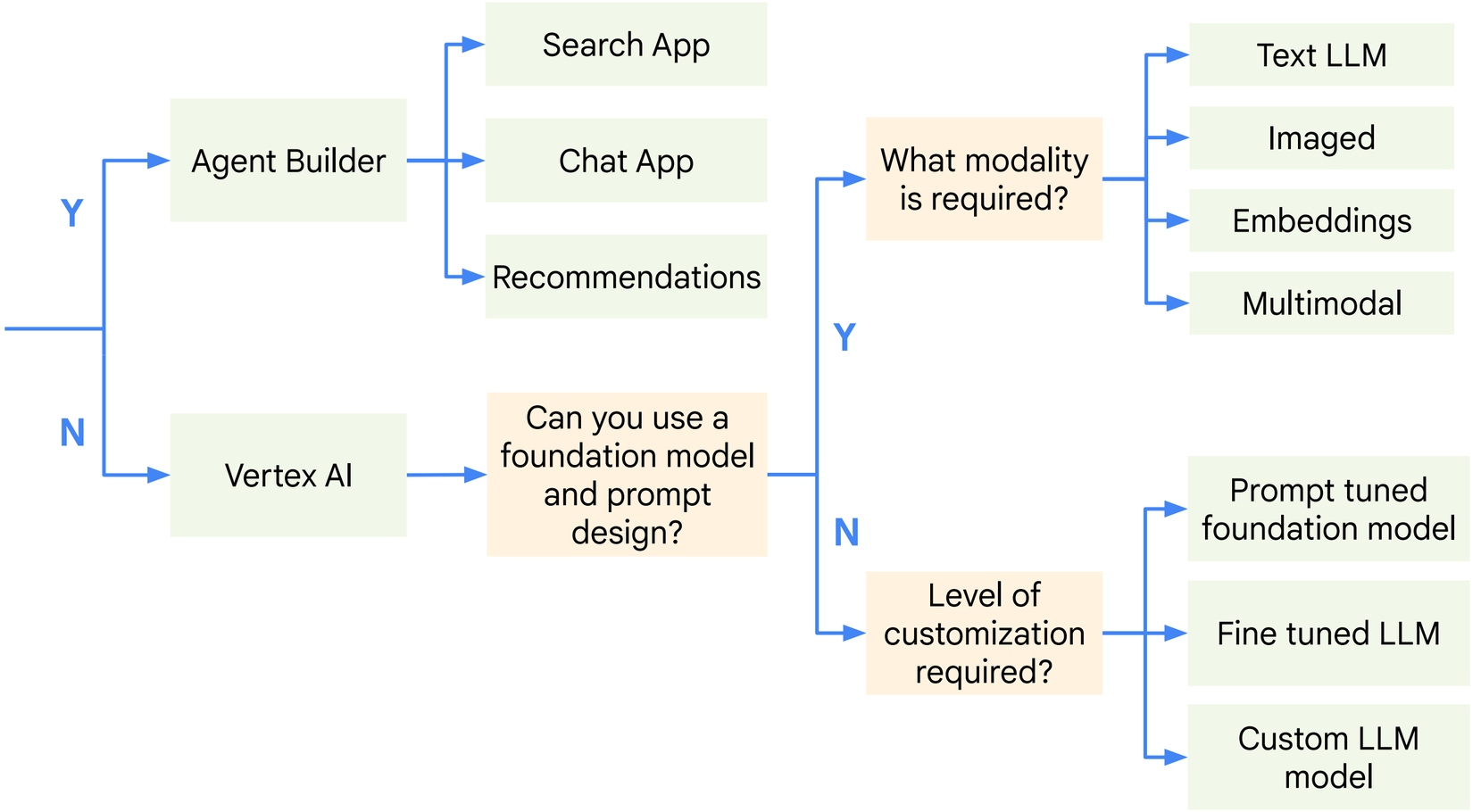
You can use BigQuery ML, Auto ML, or Vertex AI training to create your own ML models

- Requires your own training data
- Requires machine learning expertise

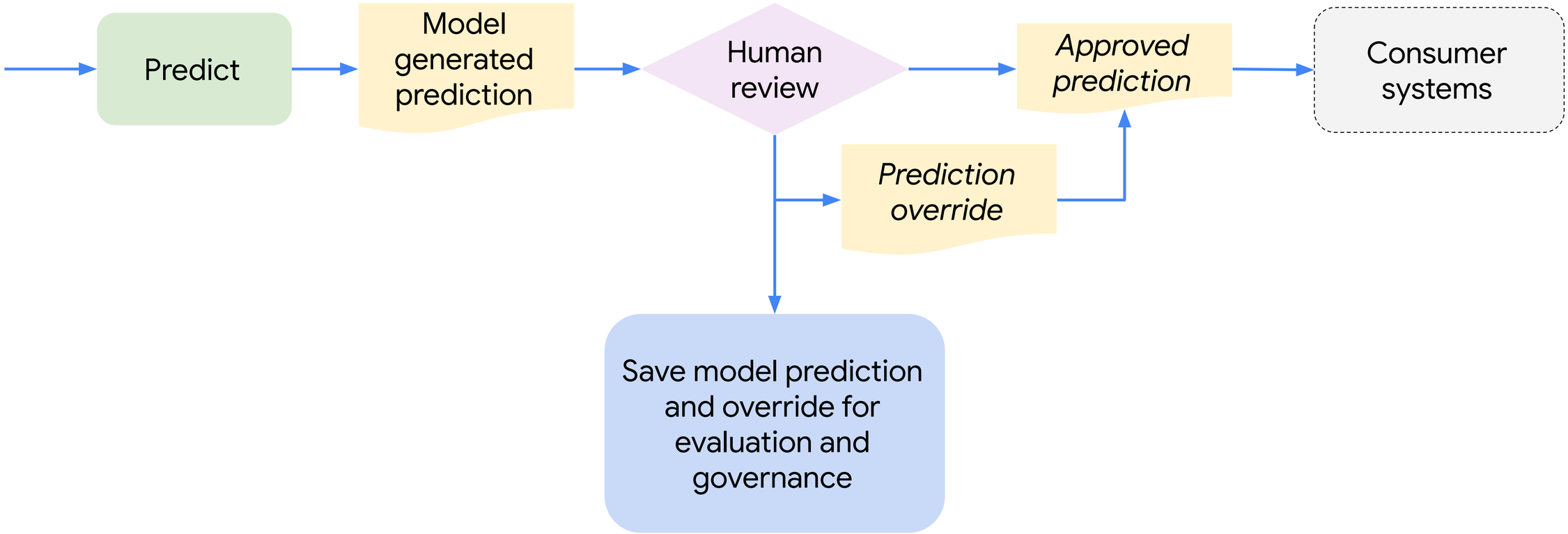


For GenAI problems, use the right generative AI model

- Gemini 1.5
- Gemini Pro
- Gemini Pro Vision
- PaLM
- Chat PaLM
- Codey
- Embeddings
- Imagen
- Etc.



Do you need a human in the loop?



If you can improve prediction quality, then the need for humans in the loop becomes less critical

Prompt Engineering

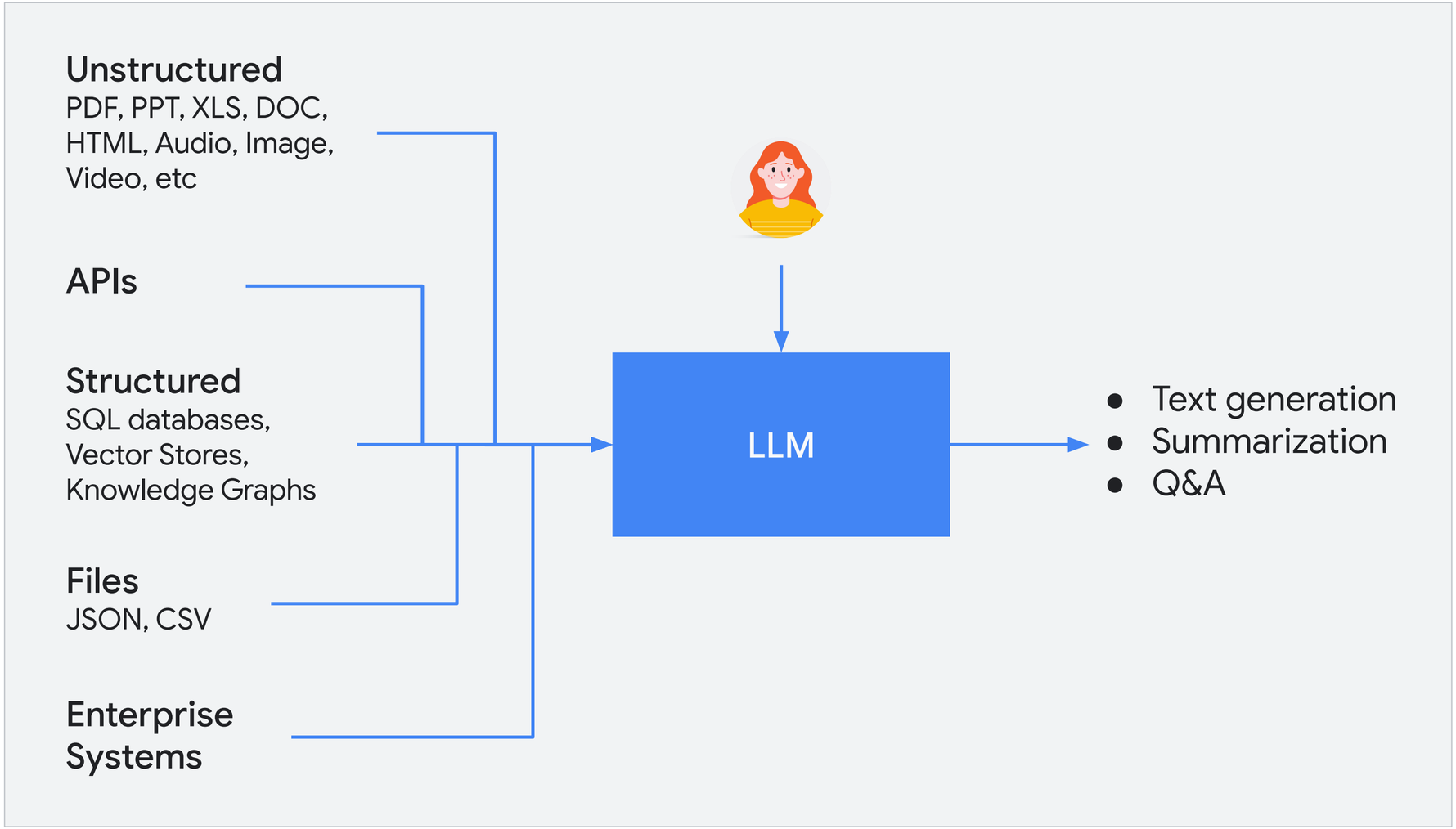
- Relatively simple
- Inexpensive
- Effective for most use cases

Fine tuning

- Requires training data
- More expensive
- Time consuming
- Effective, assuming the training data is high quality

How can you provide the model with external data?

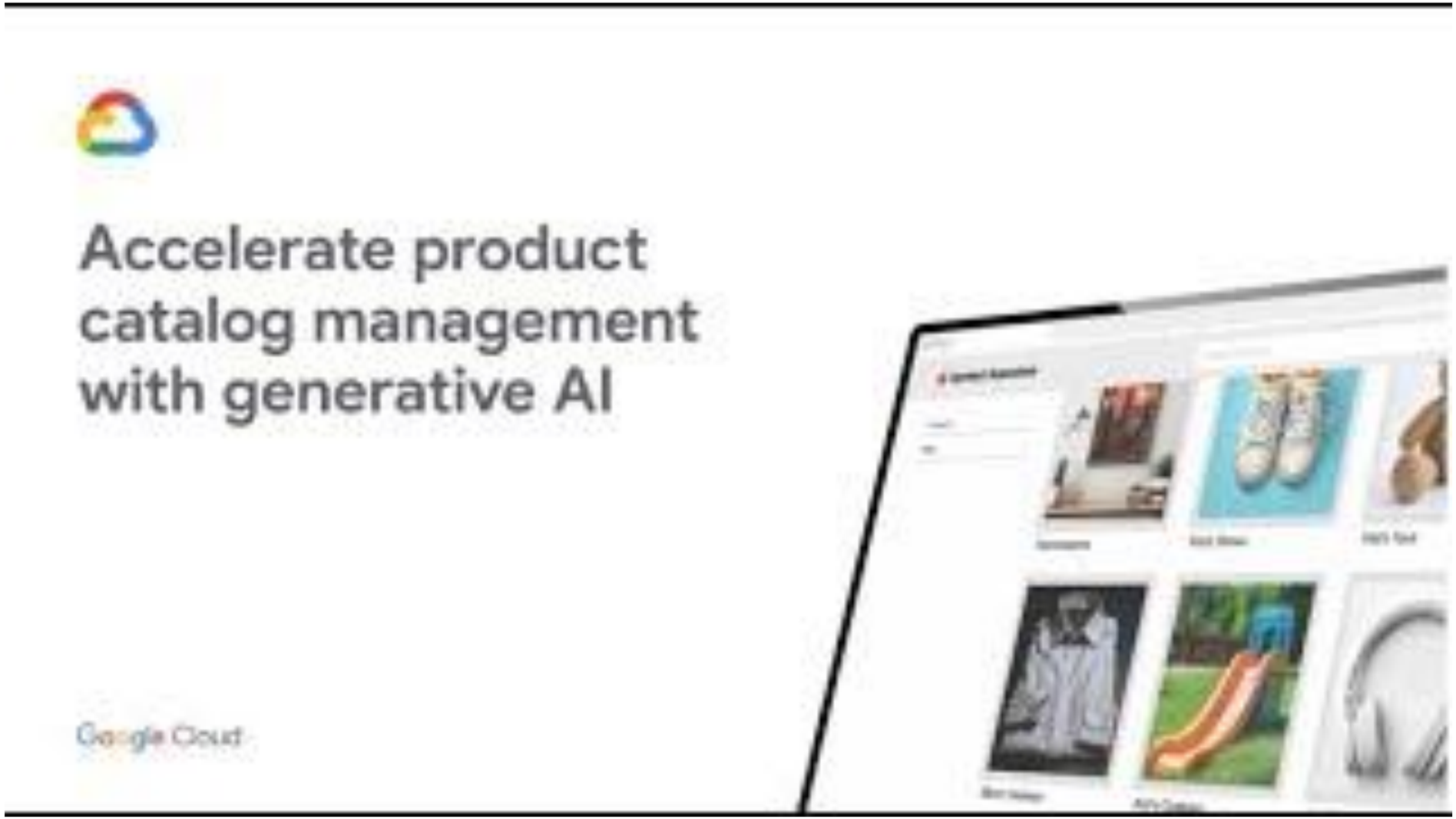
- Add data to the prompt
- Make an API call or database lookup
- Use embeddings and Vector Search (RAG)
- Use Gemini function calling



Topics

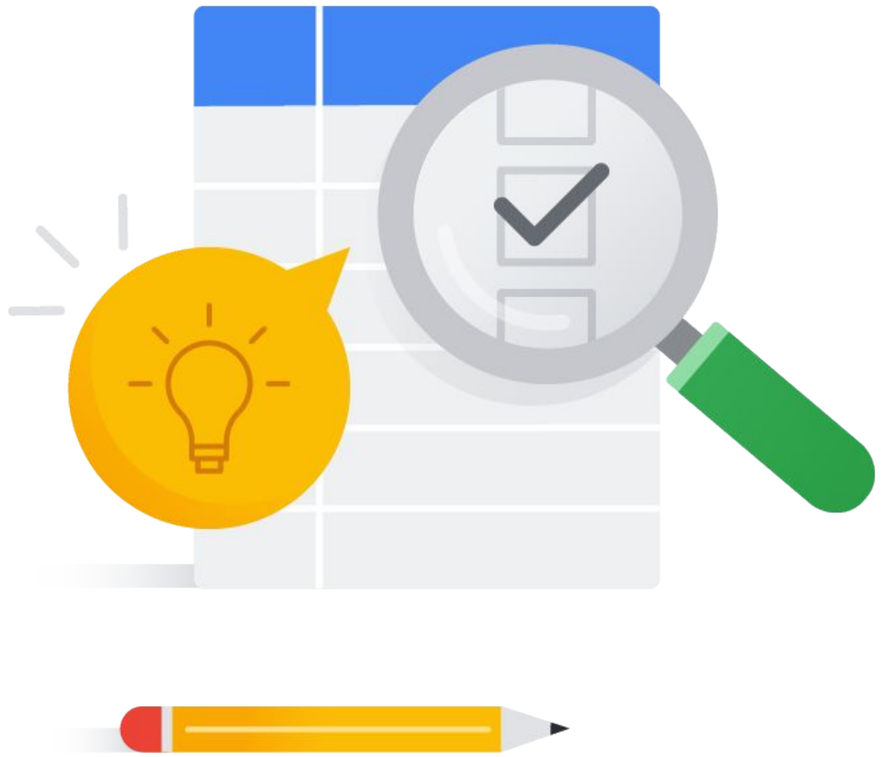
01	Agent Builder
02	Automating Business Processes with AI
03	Product Catalog Management Case Study
04	Architecting the Catalog Management Solution

Product Catalog management with Generative AI



Discussion: Based on the video, What are the tasks required for adding a product to the online catalog?

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____



Task 1

Upload the product photo and link it to text by SKU

Can AI help with this task?

Task 1

Upload the product photo and link it to text by SKU

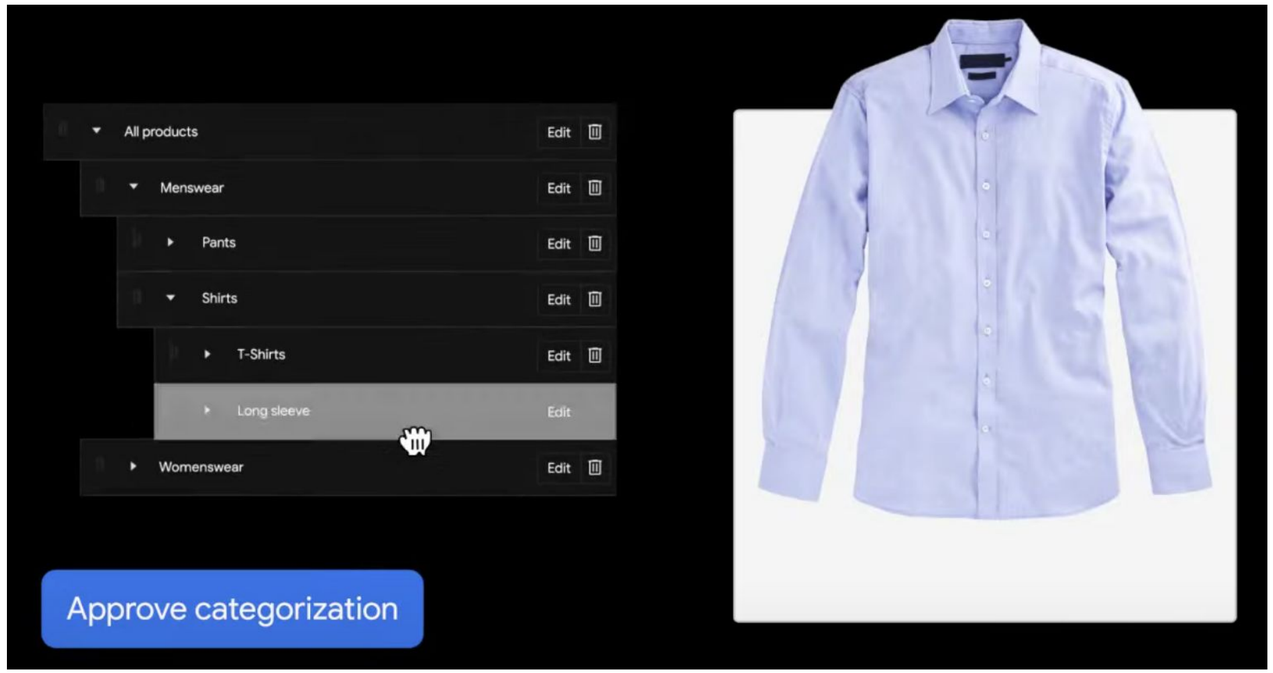
Can AI help with this task?

No, this is not a task for AI.

Task 2

Categorize the product

Can AI help with this task?
If so, how would you use it?



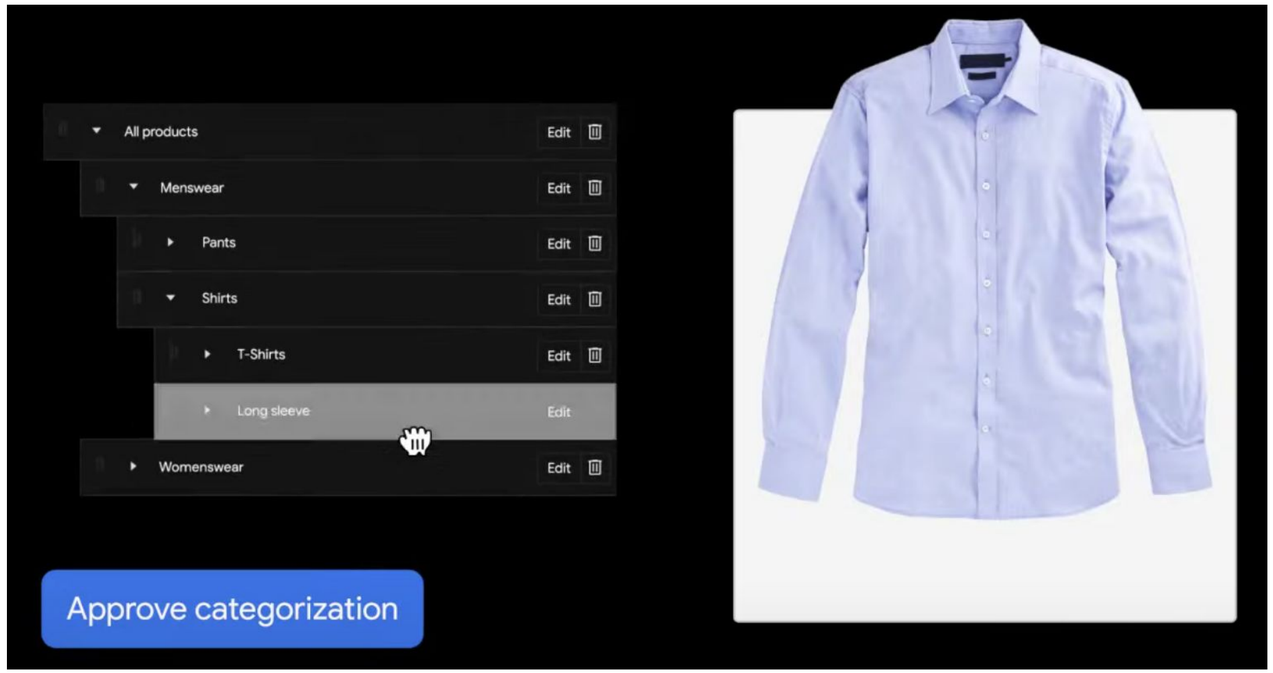
Task 2

Categorize the product

Can AI help with this task?
If so, how would you use it?

Possibly. Some options include:

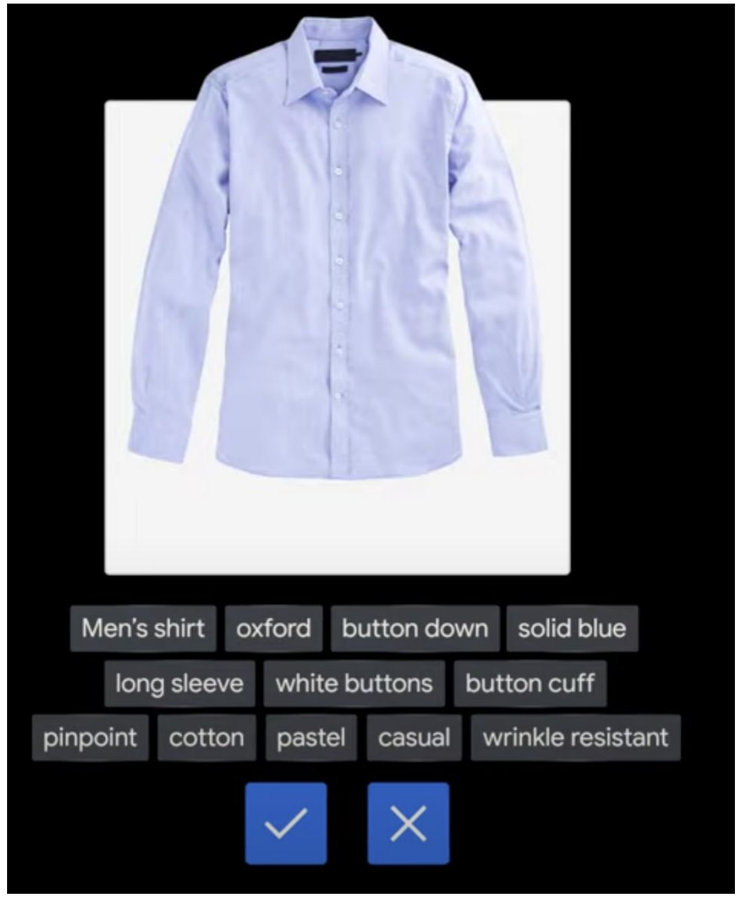
- Get embeddings of the image and find the nearest product cluster
- Visual QA to ask, “Is this a men’s shirt?” etc.



Task 3

Create tags (labels) for the product

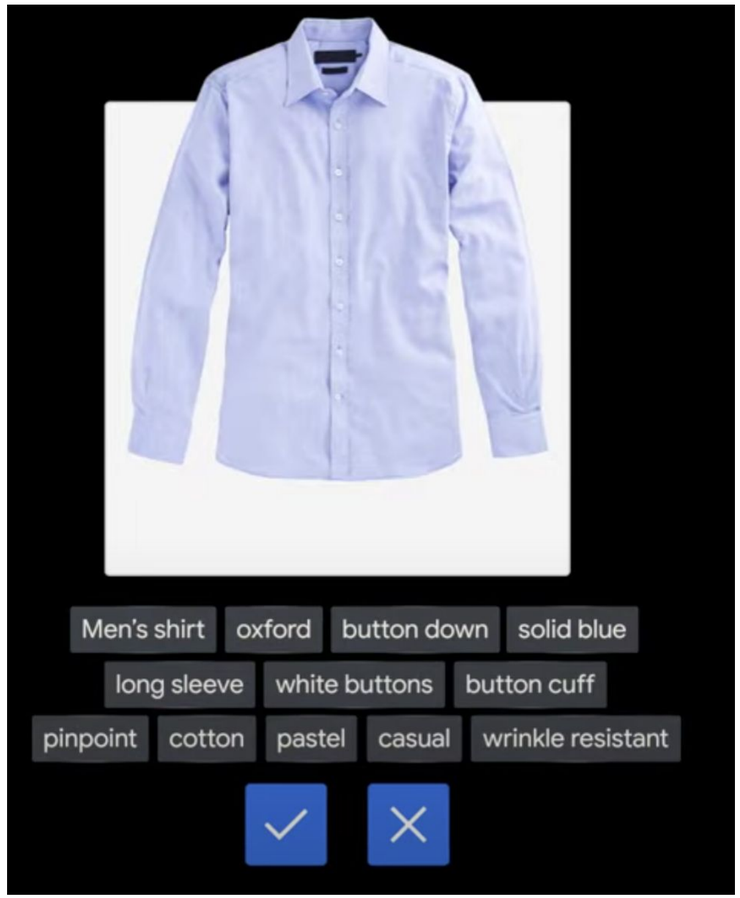
Can AI help with this task?
If so, how would you use it?



Task 3

Create tags (labels) for the product

Can AI help with this task?
If so, how would you use it?



Some options include:

- Image captioning and extracting nouns and adjectives
- Visual QA to ask, “Please provide a comma-separated list describing attributes of this shirt.”

Task 4

Generate a name for the product

Can AI help with this task?
If so, how would you use it?

Task 4

Generate a name for the product

Can AI help with this task?
If so, how would you use it?

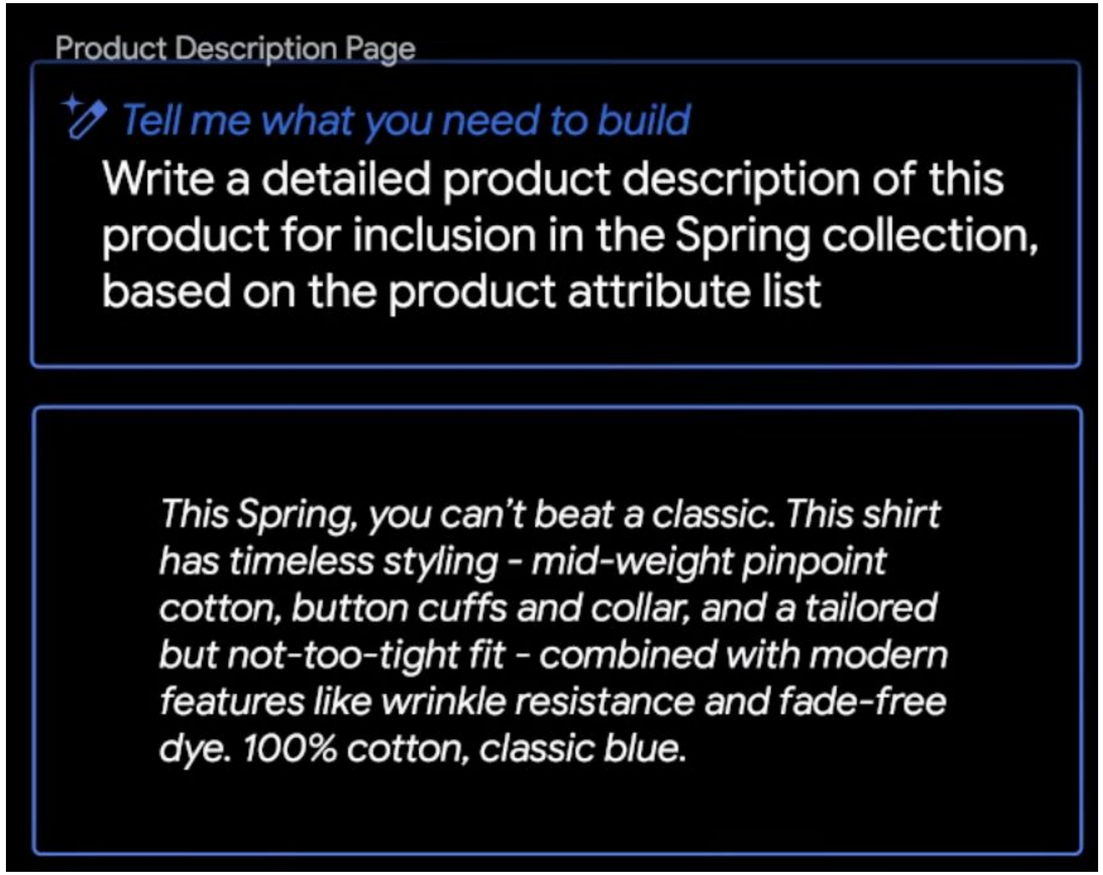
Possibly. One option includes:

- Gemini with context and examples of other product names

Task 5

Write a product description

Can AI help with this task?
If so, how would you use it?



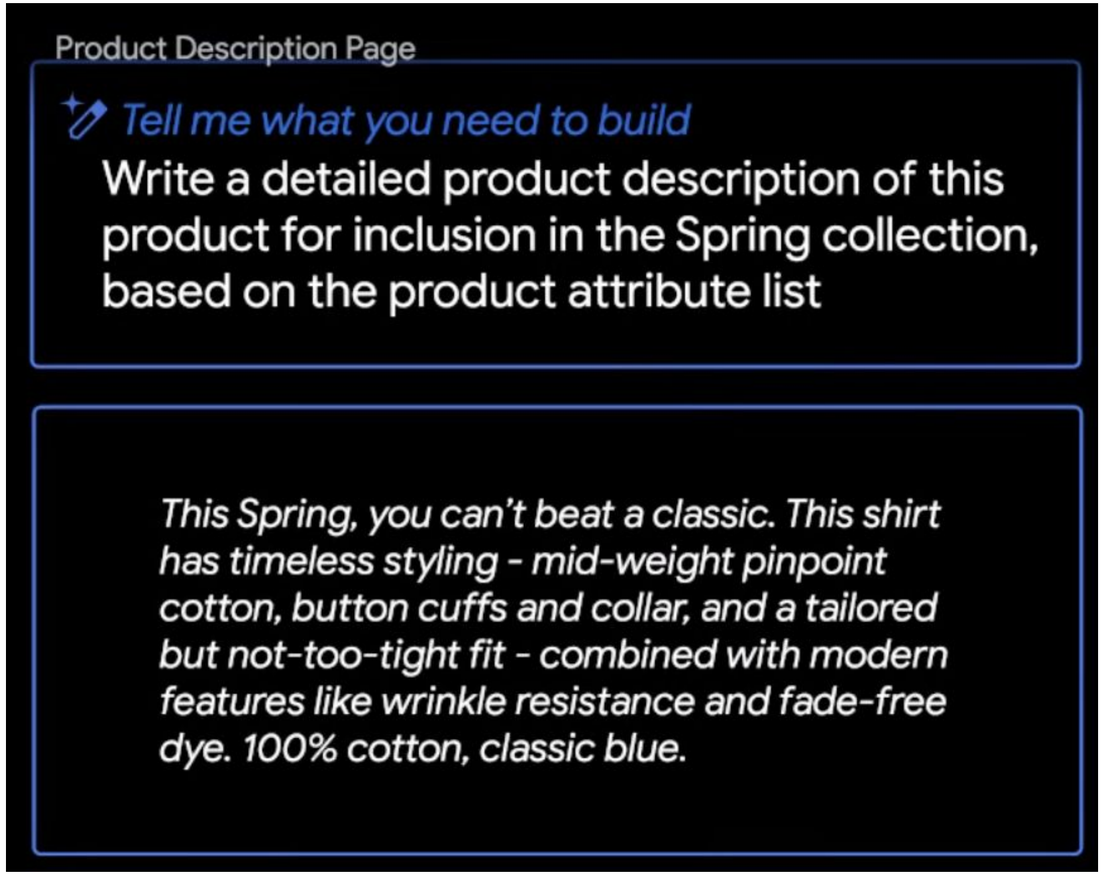
Task 5

Write a product description

Can AI help with this task?
If so, how would you use it?

Possibly. One option includes:

Gemini with a high temperature and a context of the upcoming season (spring, holidays, etc.) and example product descriptions.



Task 6

Write product HTML metadata

Can AI help with this task?
If so, how would you use it?

Task 6

Write product HTML metadata

Can AI help with this task?
If so, how would you use it?

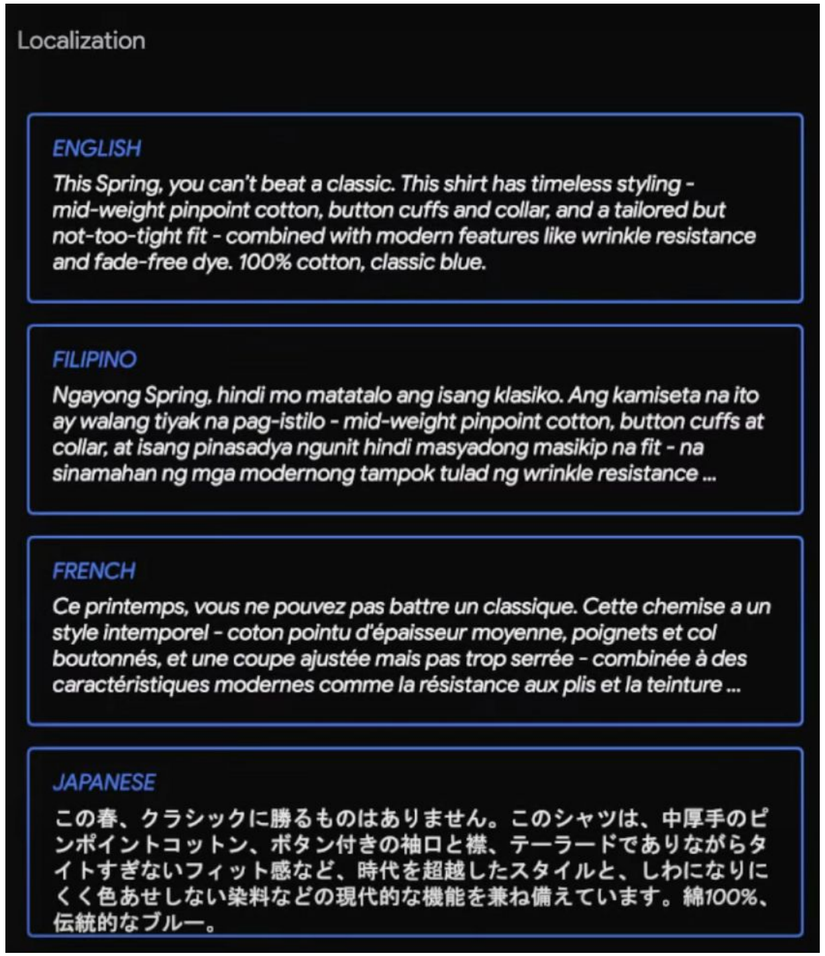
Possibly. One option includes:

- Using Gemini to generate valid HTML given the list of SEO keywords and product tags

Task 7

Translate product content into other languages

Can AI help with this task?
If so, how would you use it?



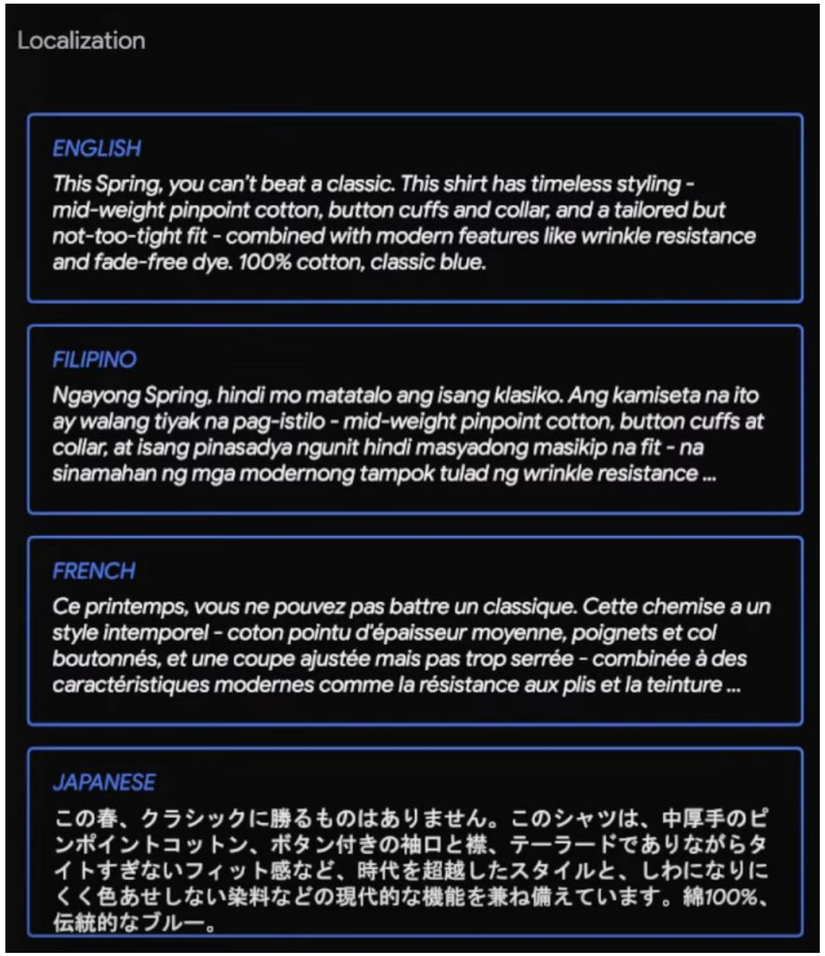
Task 7

Translate product content into other languages

Can AI help with this task?
If so, how would you use it?

Possibly:

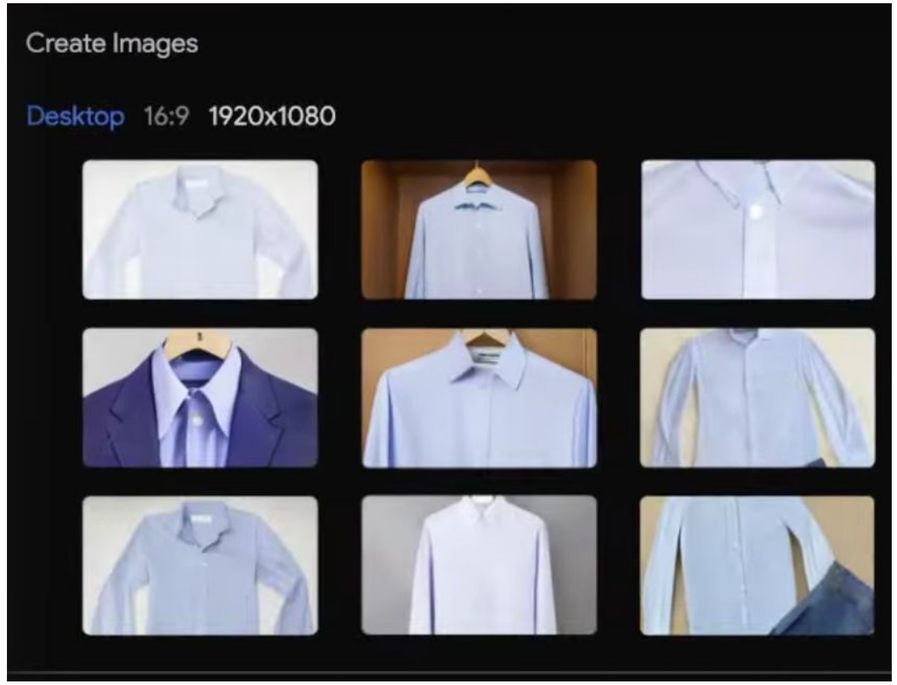
- Feed the description to Gemini and ask for the translation with a low temperature
- Use the Translation API



Task 8

Create alternate images

Can AI help with this task?
If so, how would you use it?



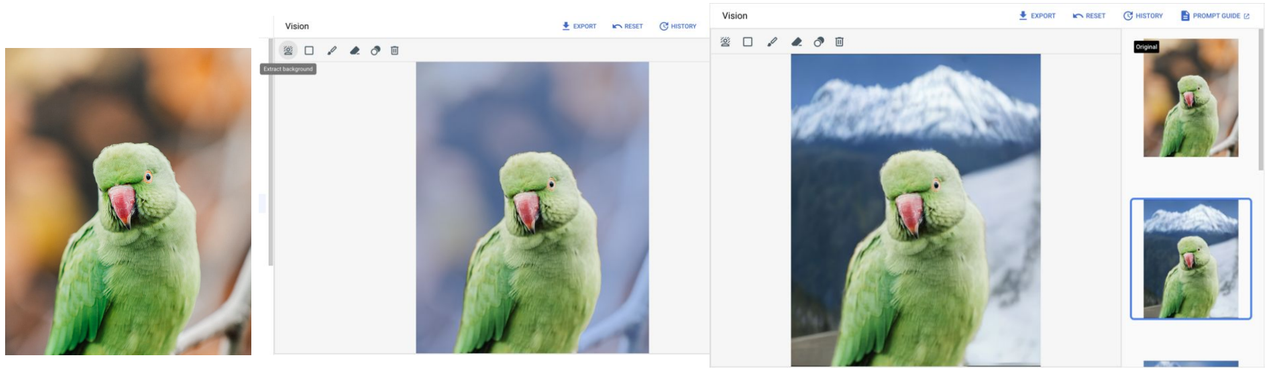
Task 8

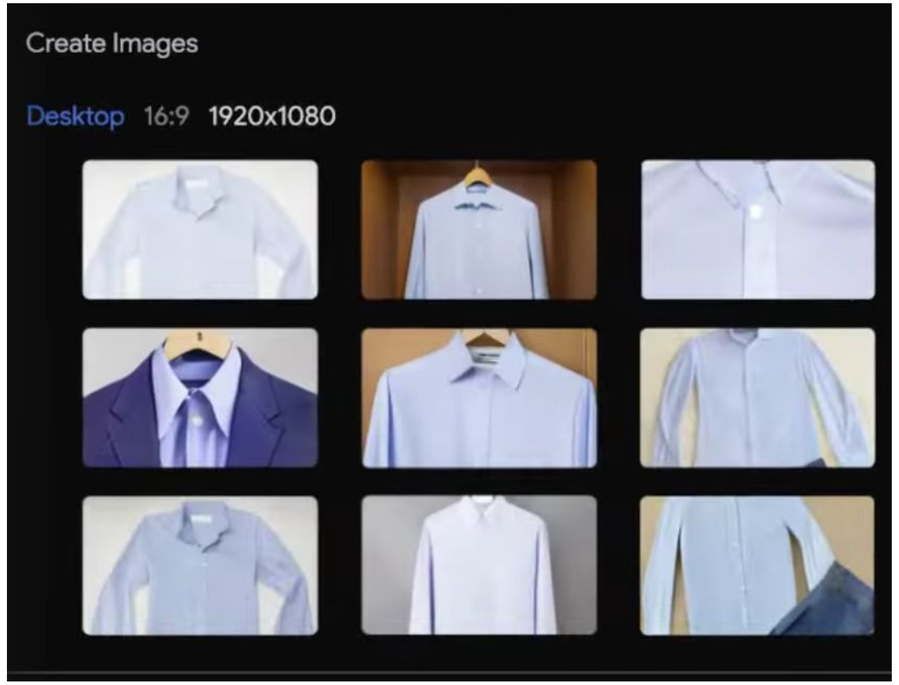
Create alternate images

Can AI help with this task?
If so, how would you use it?

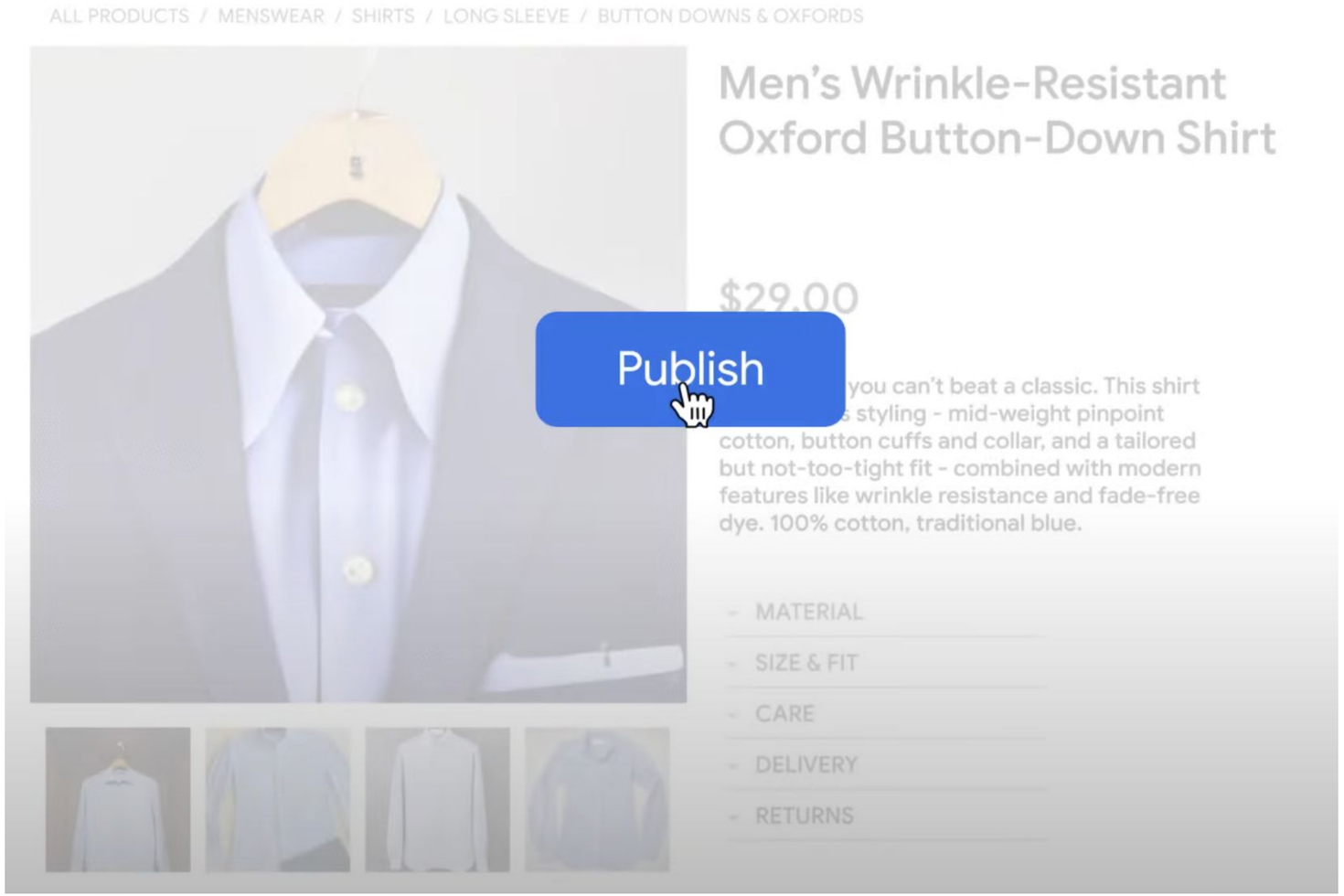
Possibilities:

- Imagen using masked-based or product background editing:

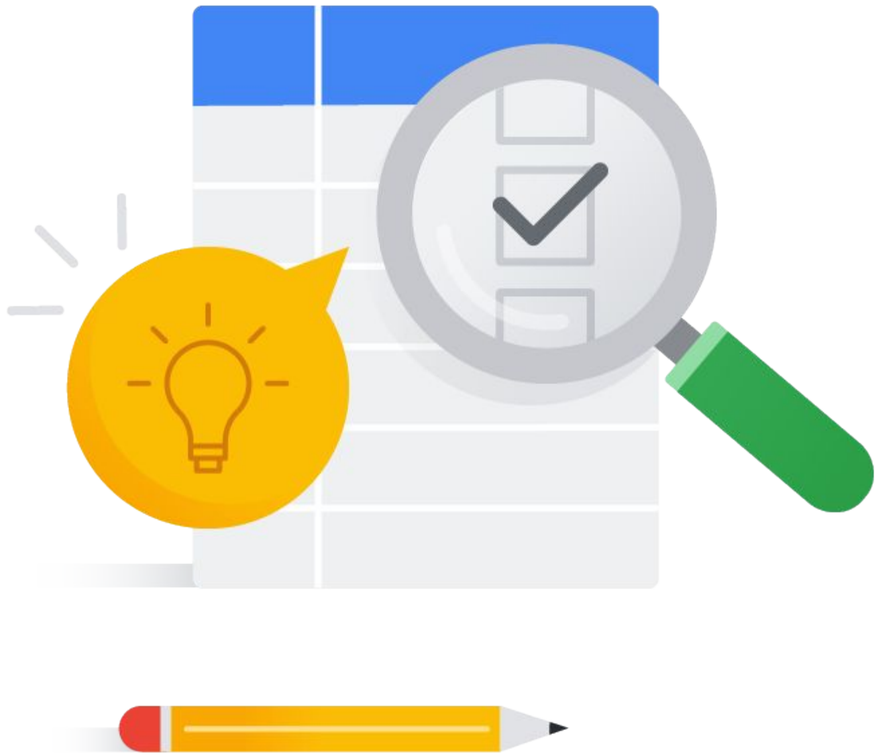




After the tasks are completed, publish the product



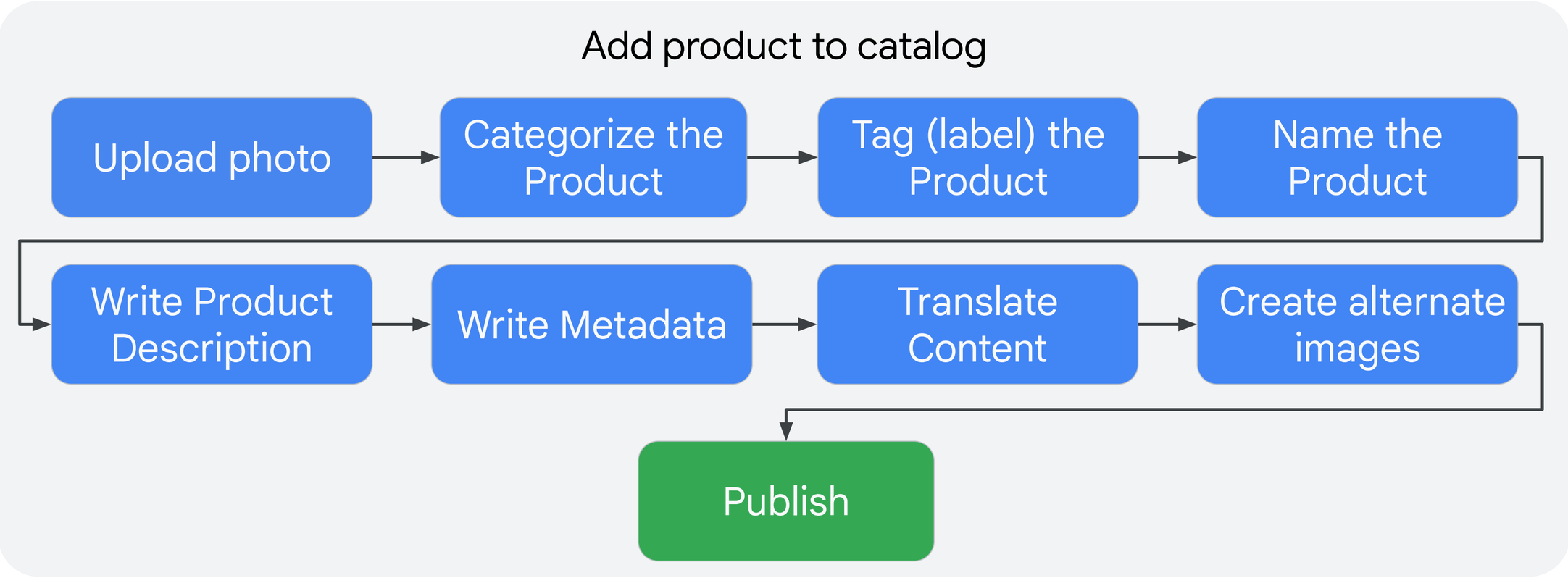
Discussion: At what steps in this process was there a human in the loop verifying the accuracy of the AI?



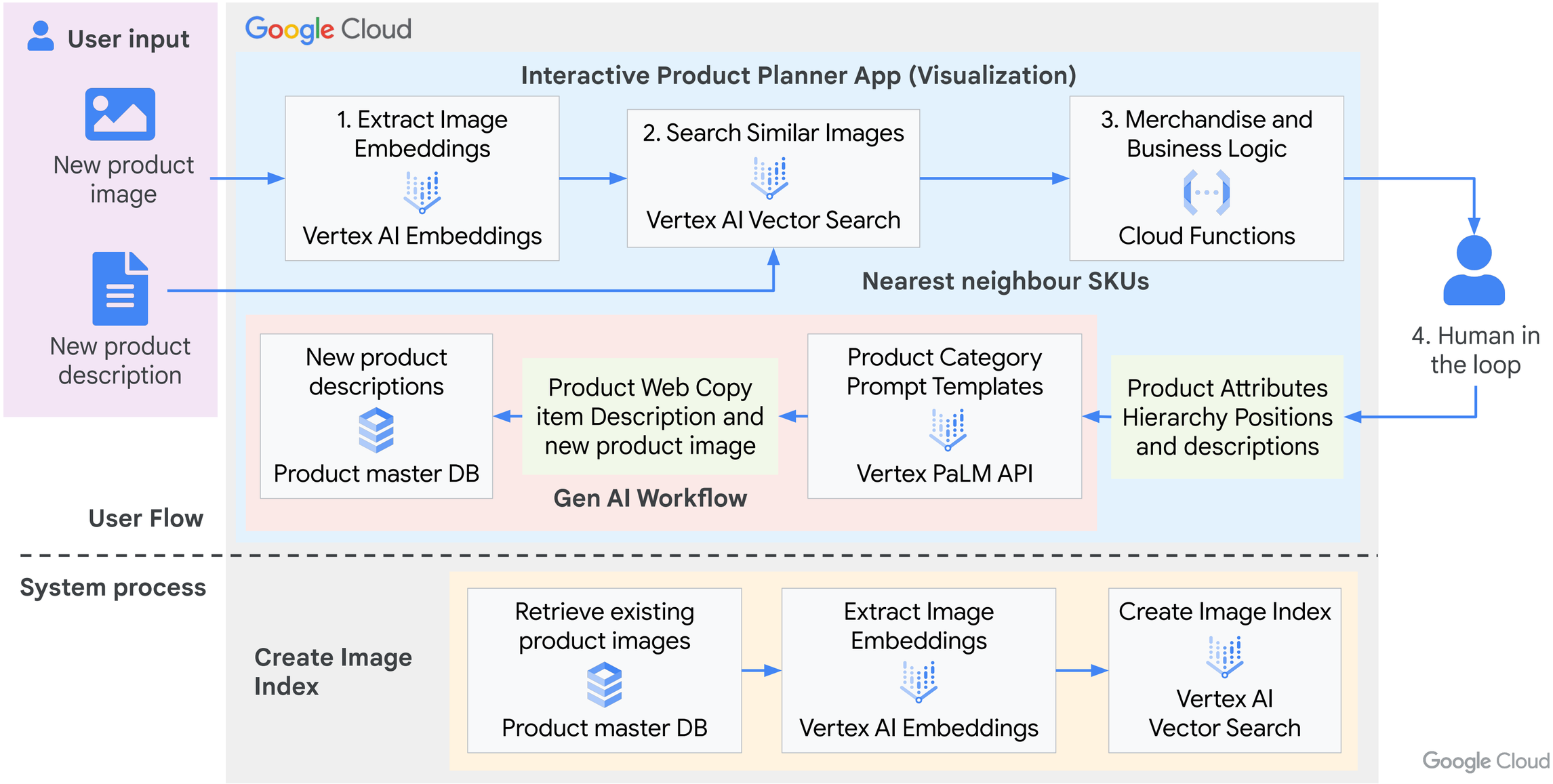
Topics

01	Agent Builder
02	Automating Business Processes with AI
03	Product Catalog Management Case Study
04	Architecting the Catalog Management Solution

The first step to automating a complex process is to break it into a set of smaller, discrete tasks



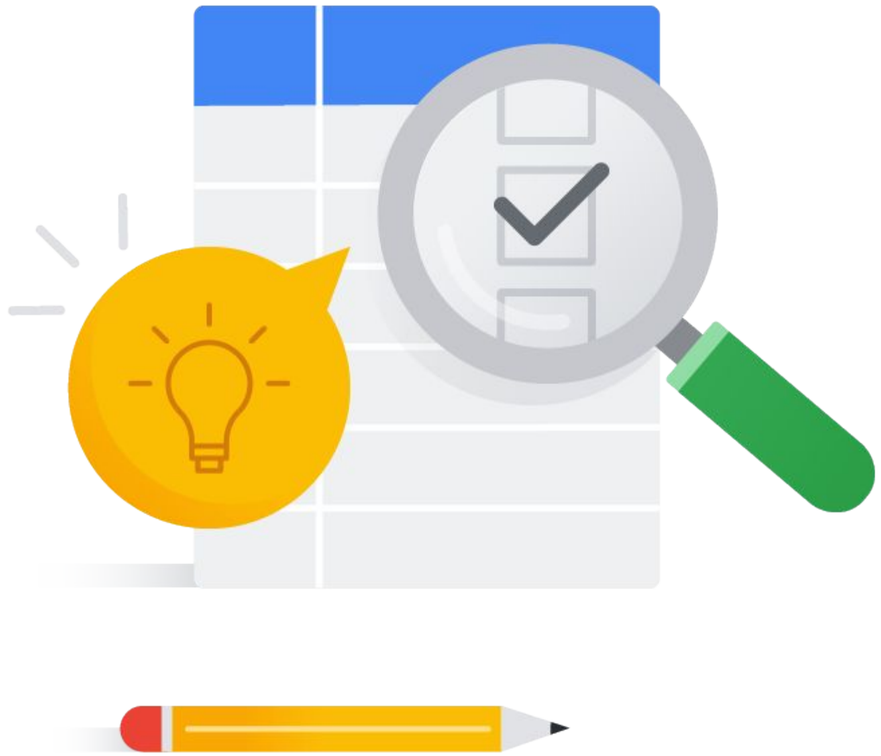
Reference Architecture for Product Catalog



Lab

1 hour

Lab: Using Vertex AI Search as a RAG



In this module, you learned to ...

01 Break complex processes into discrete tasks

02 Determine if and how AI can be used to help humans perform their tasks

03 Architect AI solutions that leverage Google Cloud tools


Google Cloud