twitch.tv/dancojocar
youtube.com/dancojocar

Lecture #13
AI

Mobile Applications
Fall 2024

Happy New Year!

2025

Generate more

Generate a set of fantastical images

Good morning

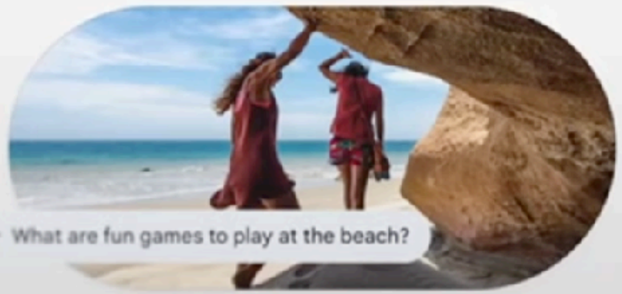
Type, talk, or share a photo

- Make it sound more casual
- Summarize this email
- Help me write
- Make a dinner reservation
- Suggest the next book I should read based on what I've already read this year.

Track info in a sheet

When receipts are added to the folder Receipts add details like vendor, date and cost to this spreadsheet

Receipt Tracker



What are fun games to play at the beach?

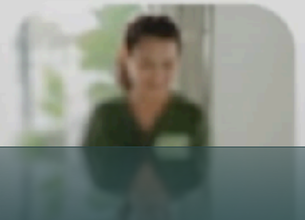
Good morning

https://www.youtube.com/watc...

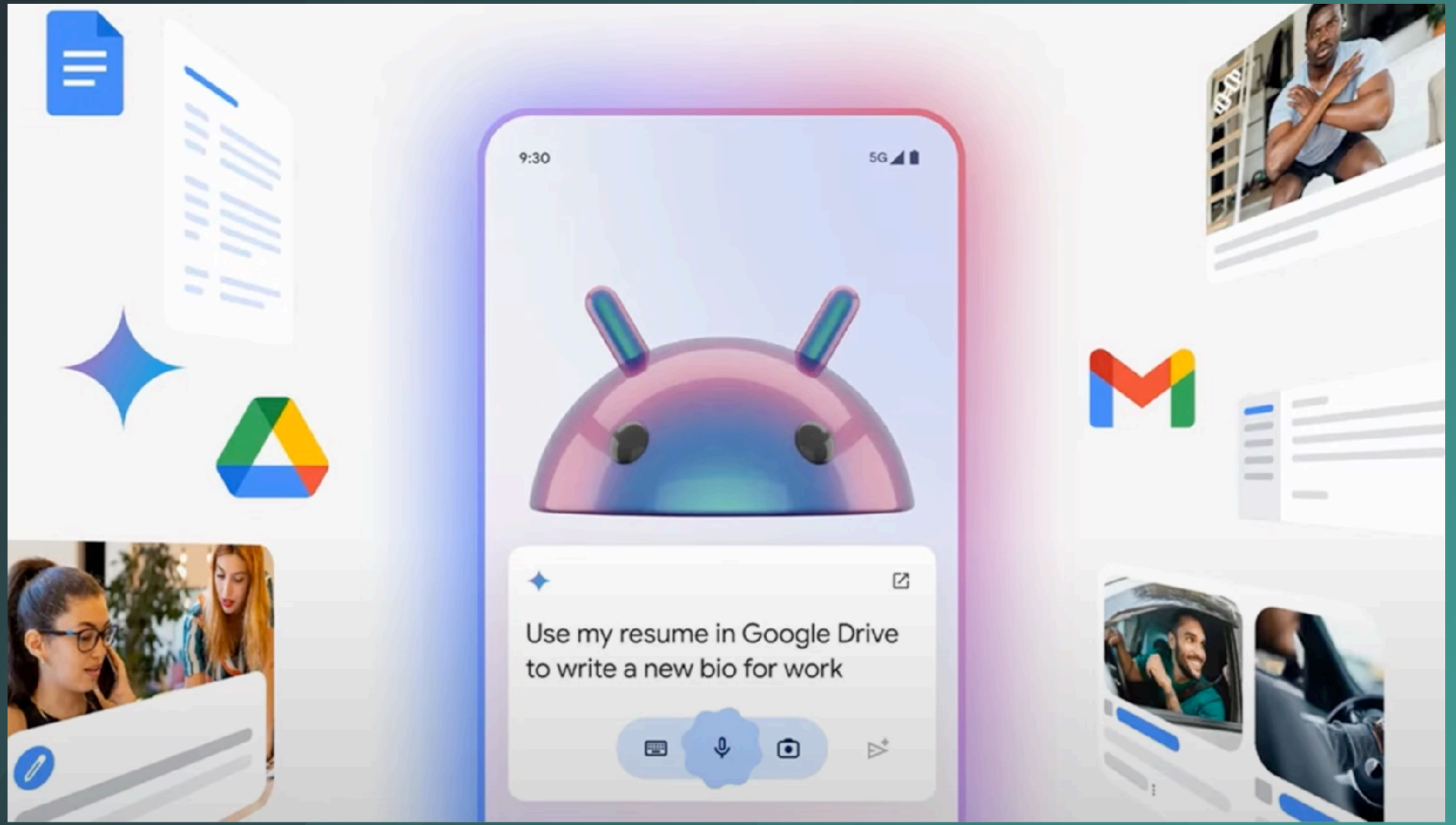
YouTube video

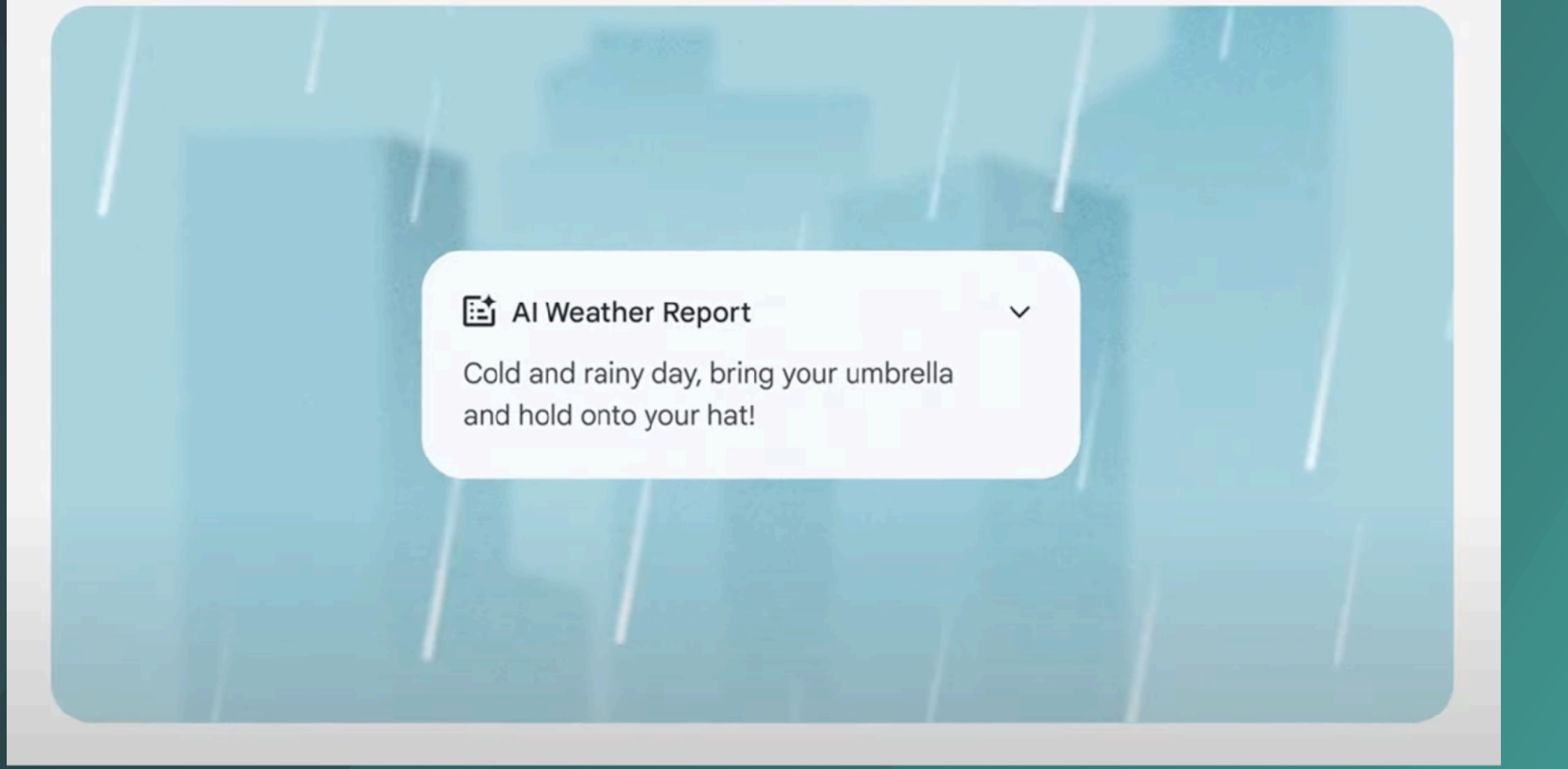
Get help with what's in this video

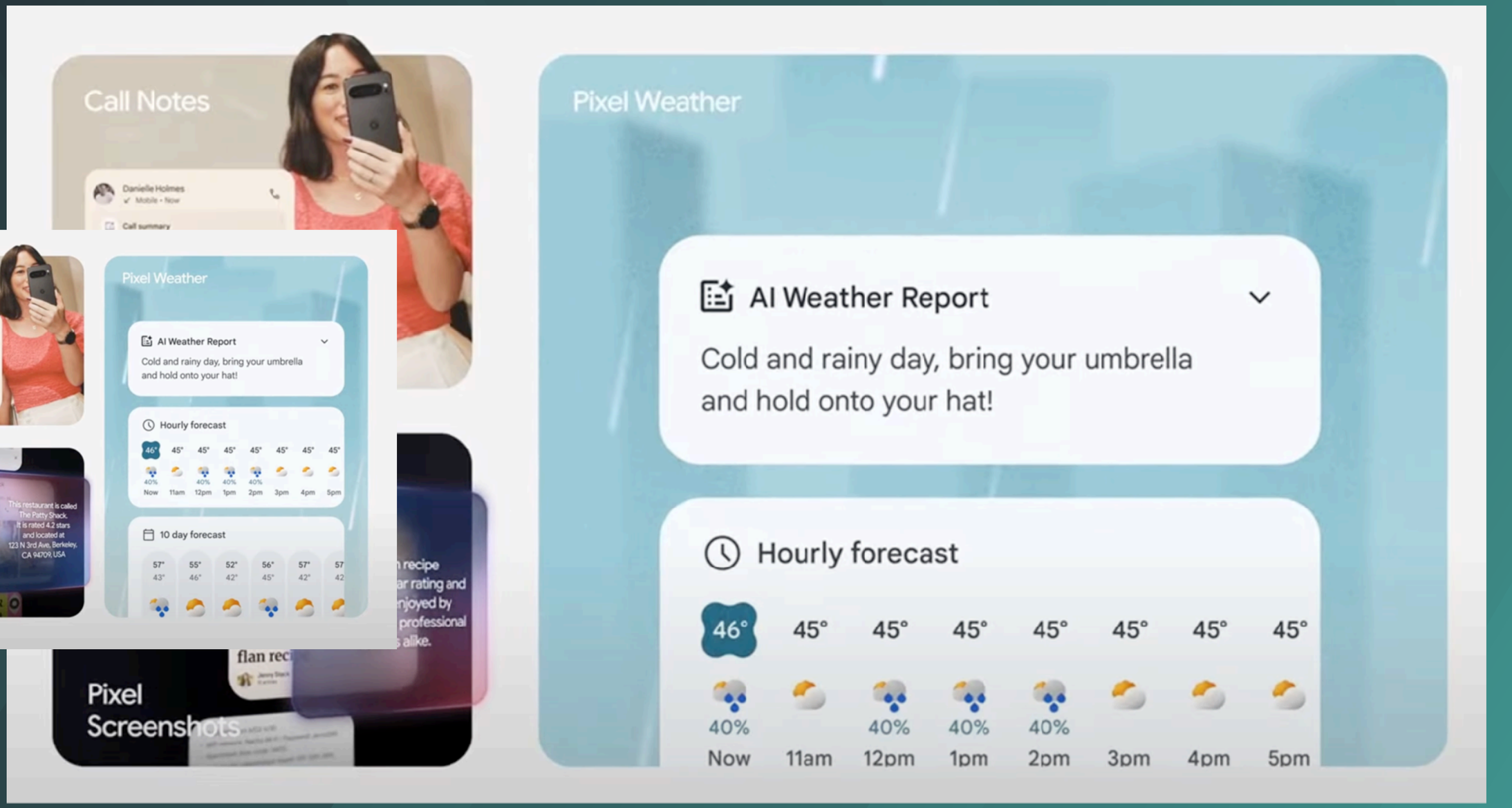
Settle a debate: how should you store bread?



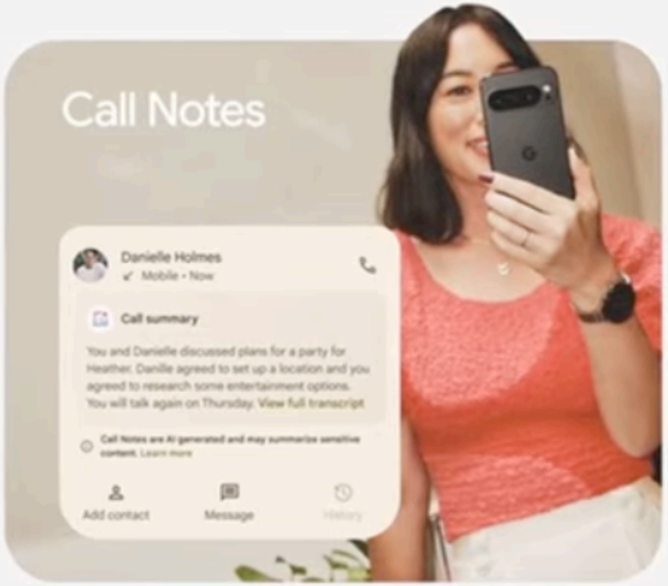
Create a 12-week study plan for learning a new

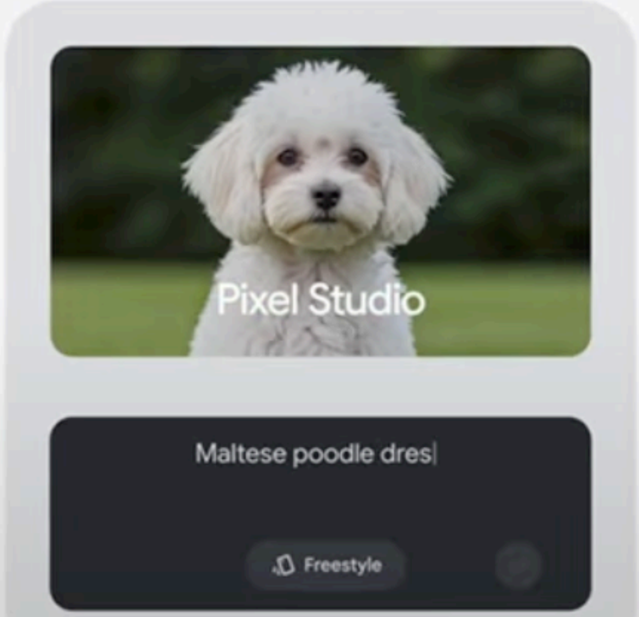

9:30
5G
Use my resume in Google Drive to write a new bio for work

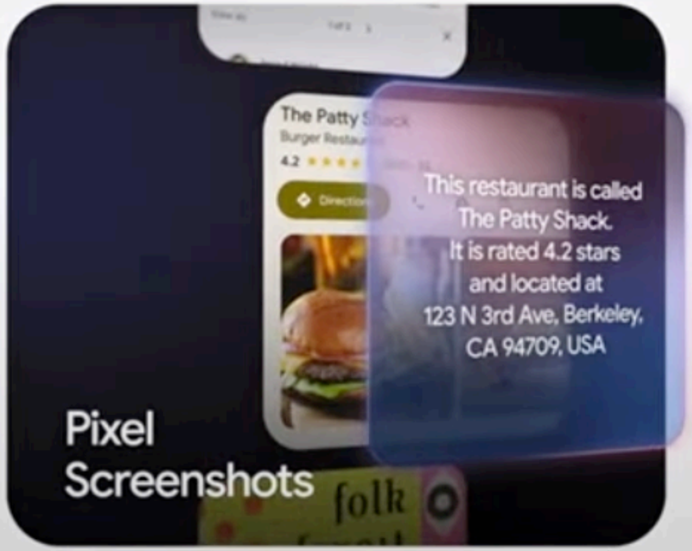

AI Weather Report
Cold and rainy day, bring your umbrella and hold onto your hat!

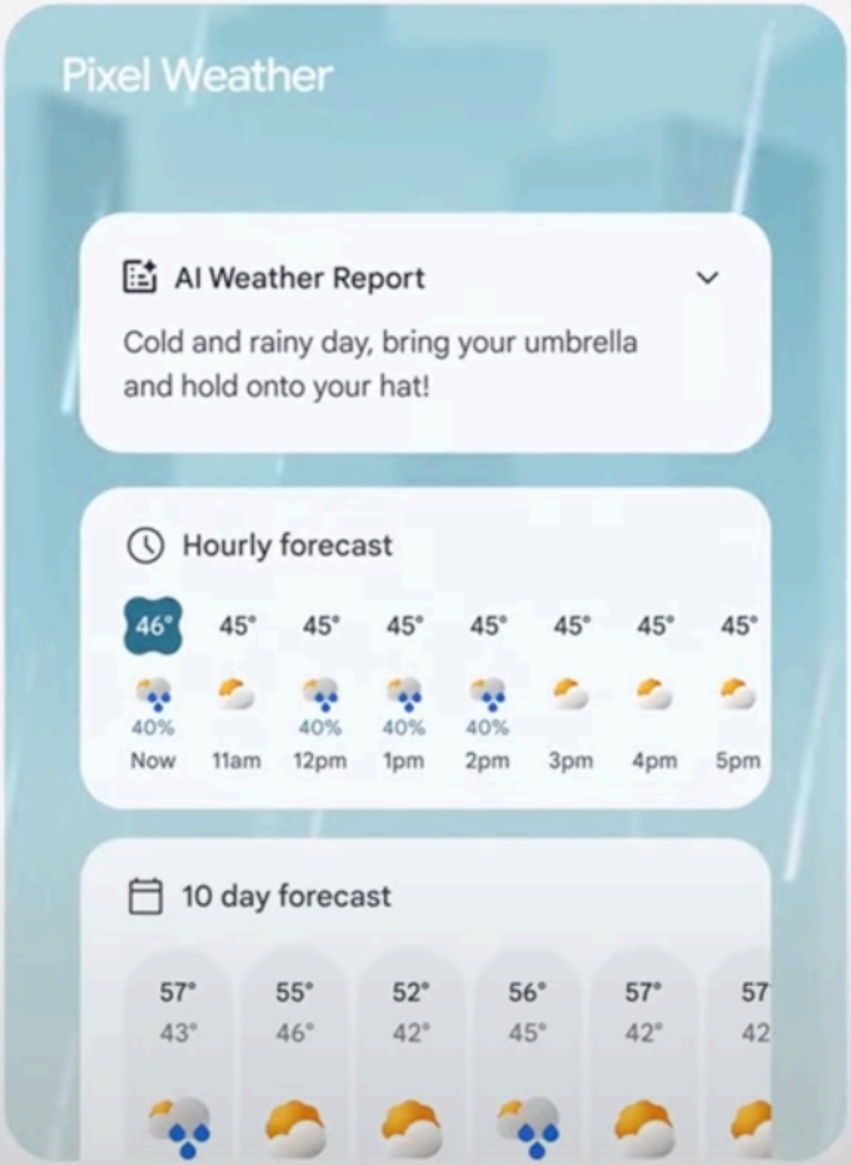

Call Notes
Danielle Holmes
Mobile • Now
Call summary
Pixel Weather
AI Weather Report
Cold and rainy day, bring your umbrella and hold onto your hat!
Hourly forecast
46° 45° 45° 45° 45° 45° 45° 45°
40% 40% 40% 40%
Now 11am 12pm 1pm 2pm 3pm 4pm 5pm
10 day forecast
57° 43°
55° 46°
52° 42°
56° 45°
57° 42°
This restaurant is called The Patty Shack. It is rated 4.2 stars and located at 123 N 3rd Ave, Berkeley, CA 94709, USA
Pixel Screenshots
recipe
rating and
njoyed by
professional
alike.
flan rec
Pixel Screenshots
Pixel Weather
AI Weather Report
Cold and rainy day, bring your umbrella and hold onto your hat!
Hourly forecast
46° 45° 45° 45° 45° 45° 45° 45°
40% 40% 40% 40%
Now 11am 12pm 1pm 2pm 3pm 4pm 5pm


G4


Call Notes
Danielle Holmes
Mobile • Now
Call summary
You and Danielle discussed plans for a party for Heather. Danille agreed to set up a location and you agreed to research some entertainment options. You will talk again on Thursday. View full transcript
Call Notes are AI generated and may summarize sensitive content. Learn more
Add contact
Message

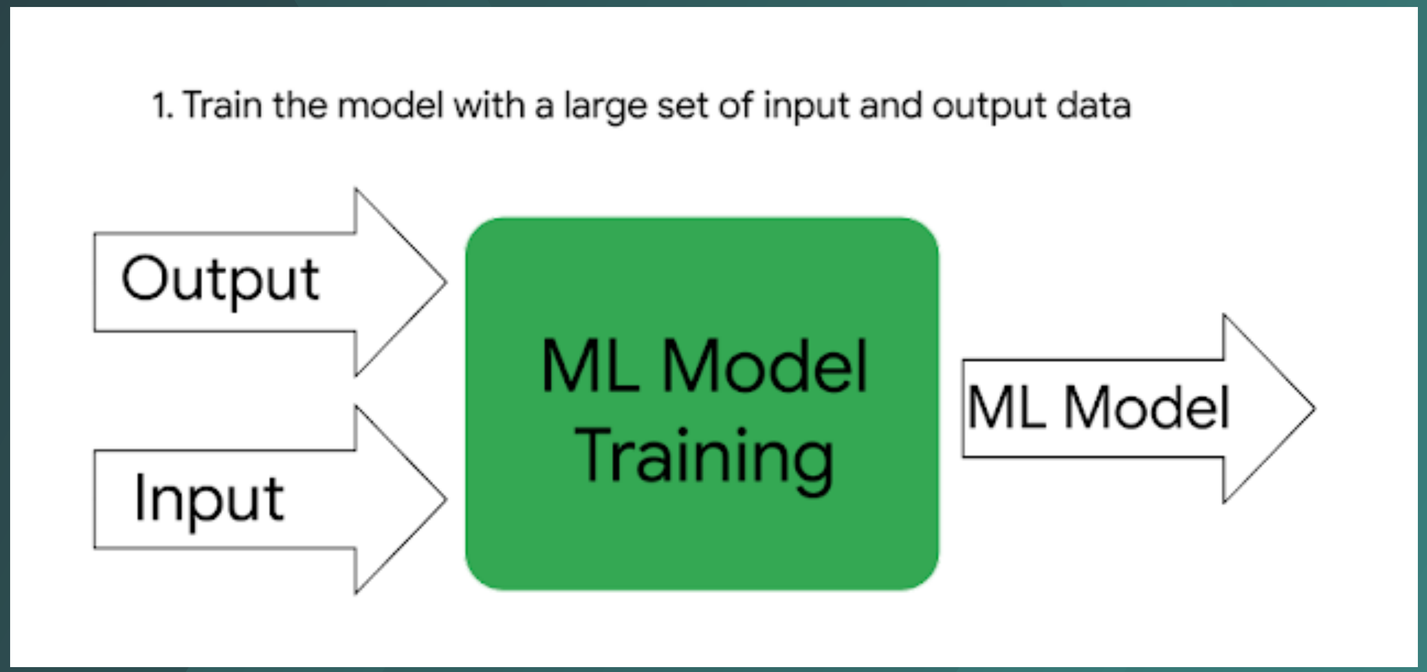

Pixel Studio
Maltese poodle dres
Freestyle


Pixel Screenshots
This restaurant is called The Patty Shack. It is rated 4.2 stars and located at 123 N 3rd Ave, Berkeley, CA 94709, USA

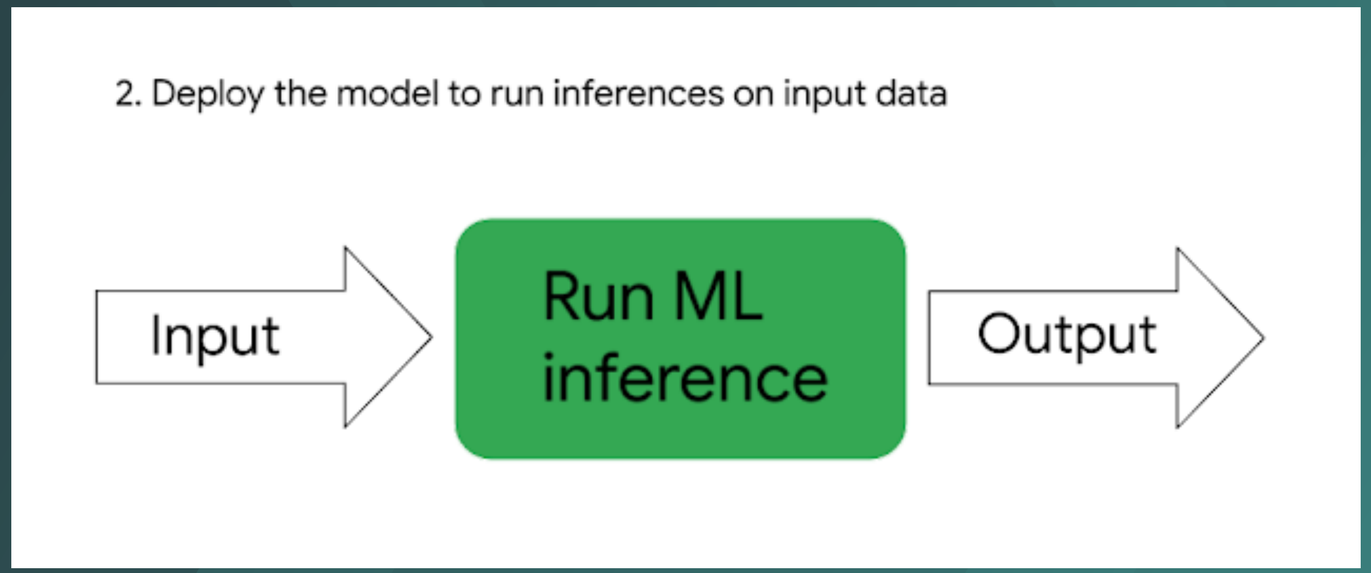

Pixel Weather
AI Weather Report
Cold and rainy day, bring your umbrella and hold onto your hat!
Hourly forecast
46° 45° 45° 45° 45° 45° 45° 45°
40% 40% 40% 40%
Now 11am 12pm 1pm 2pm 3pm 4pm 5pm
10 day forecast
57° 43°
55° 46°
52° 42°
56° 45°
57° 42°

Machine learning as a new programming paradigm

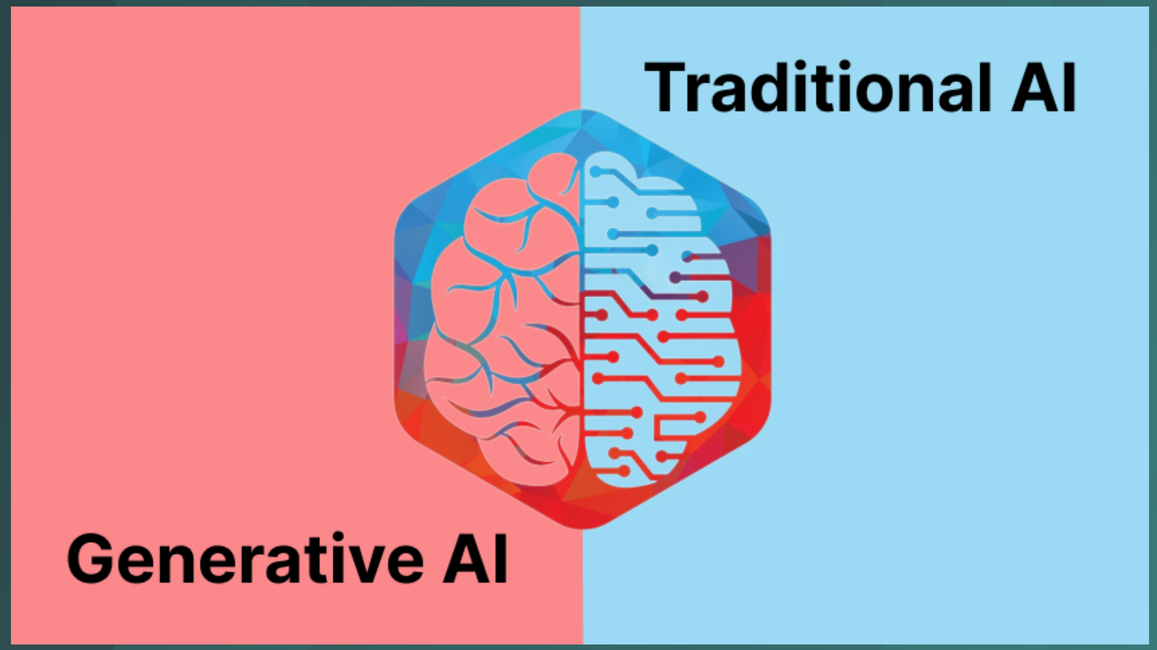
Machine learning as a new programming paradigm



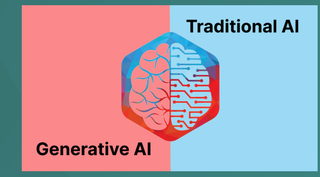
Machine learning as a new programming paradigm

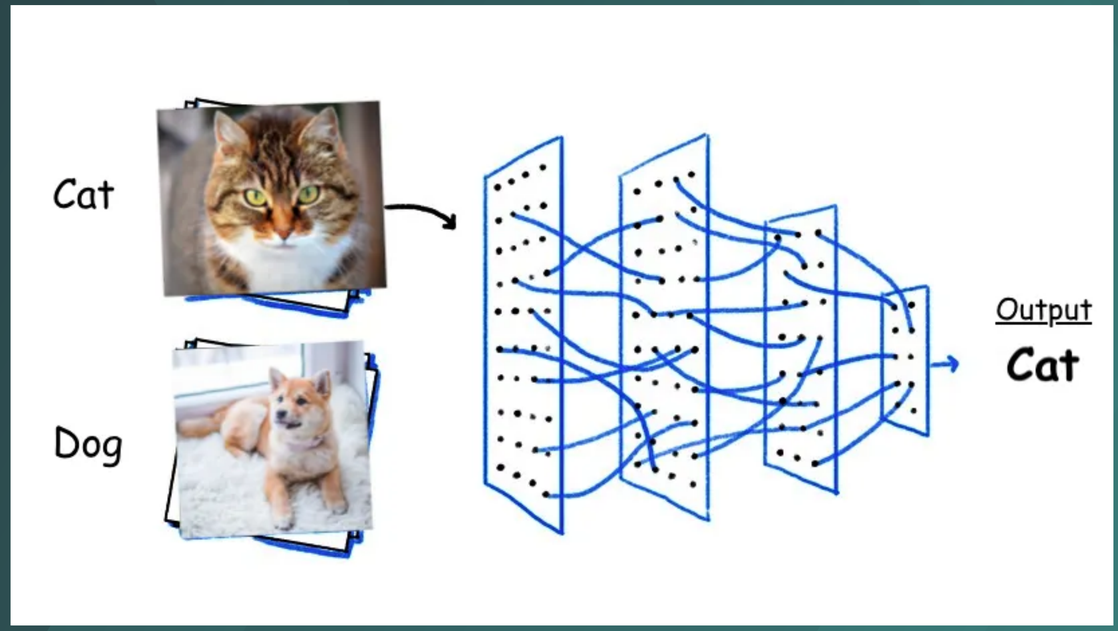


Traditional machine learning vs. generative AI on Android

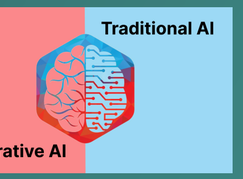


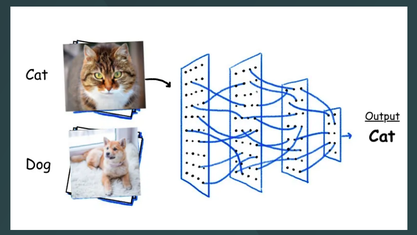
Traditional machine learning vs. generative AI on Android



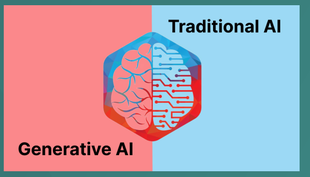


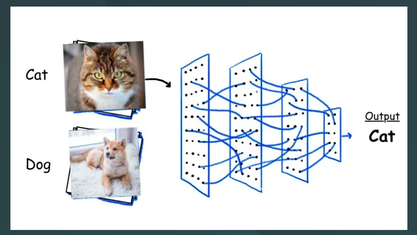
Traditional machine learning vs. generative AI on Android





Traditional machine learning vs. generative AI on Android

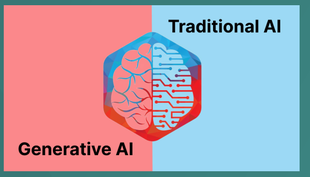


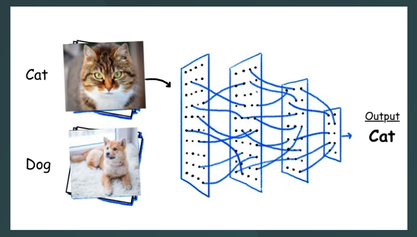


LiteRT

https://ai.google.dev/edge/litert

Traditional machine learning vs. generative AI on Android

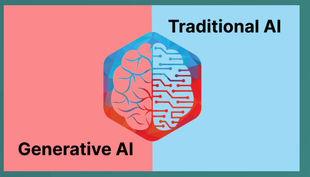


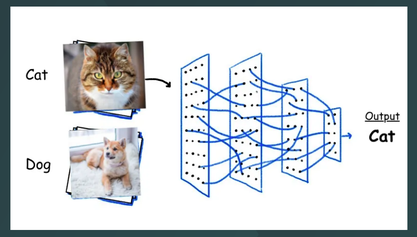


LiteRT

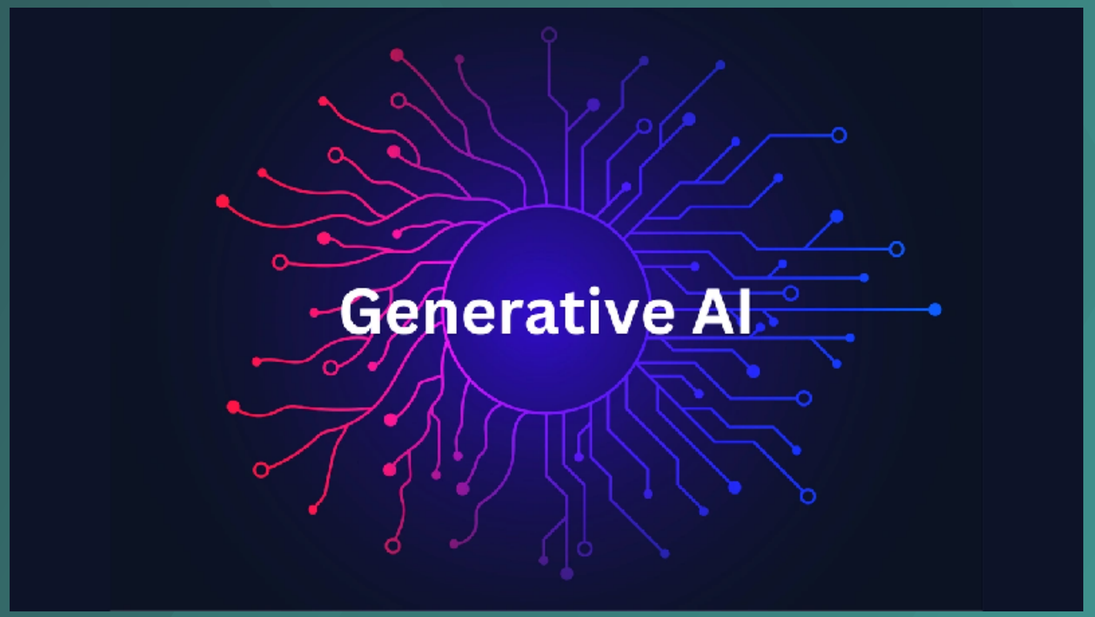
https://developers.google.com/ml-kit

Traditional machine learning vs. generative AI on Android

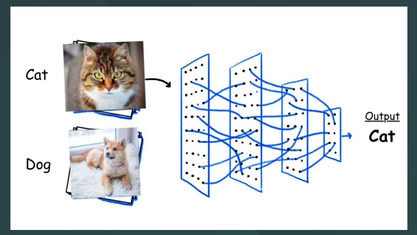


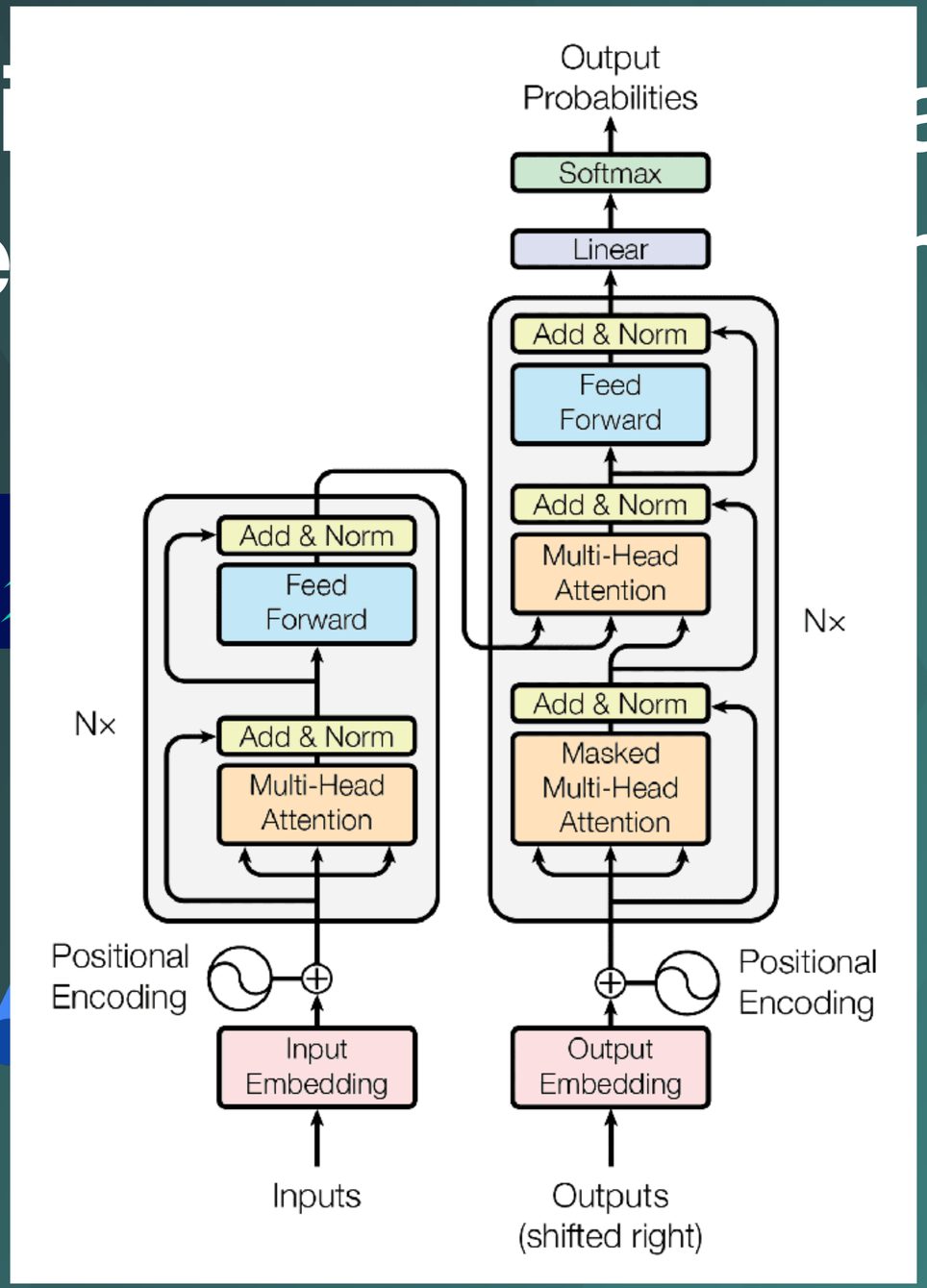


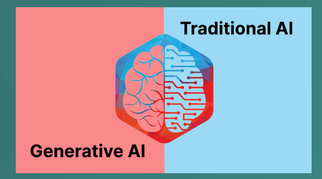
LiteRT

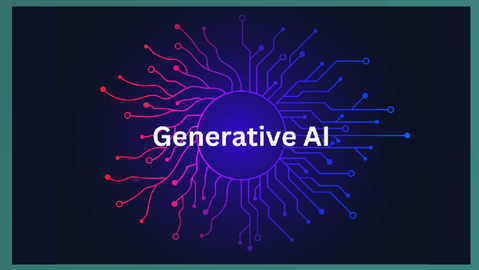


Traditi arning
vs. ge ndroid



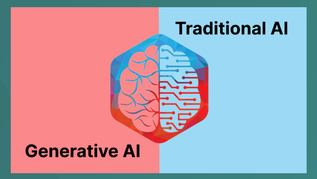


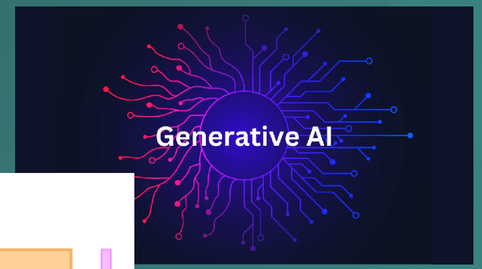


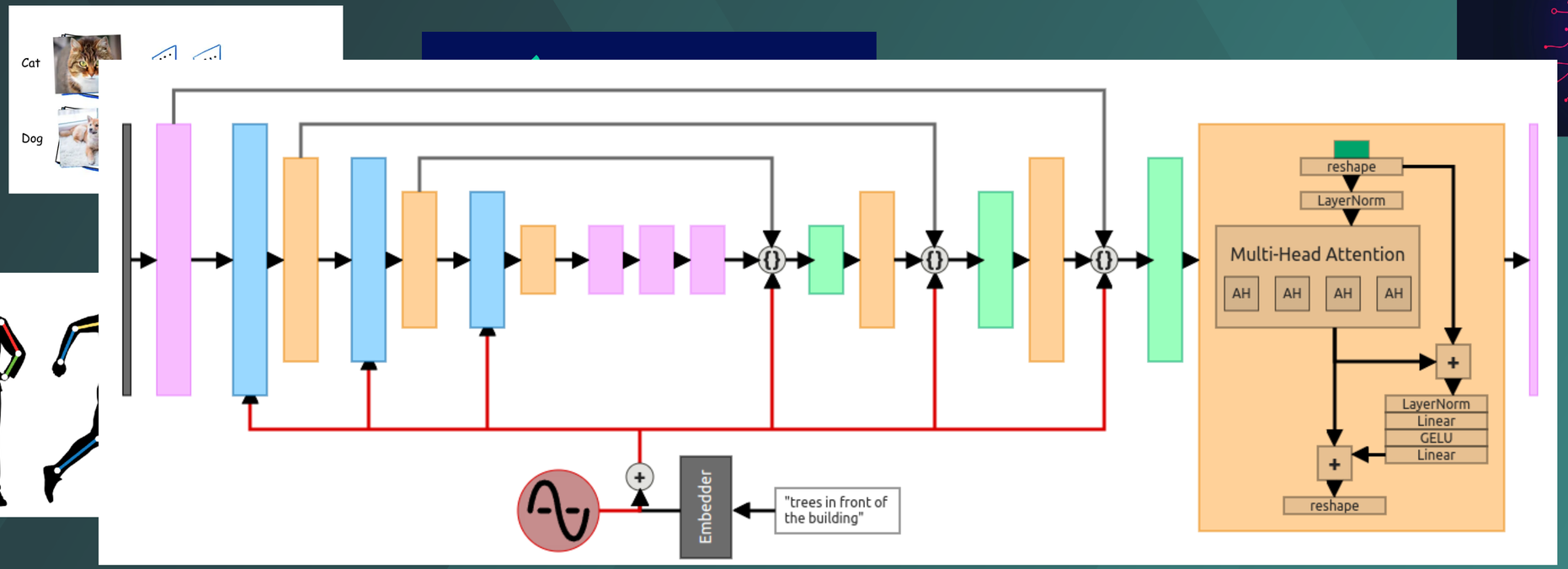


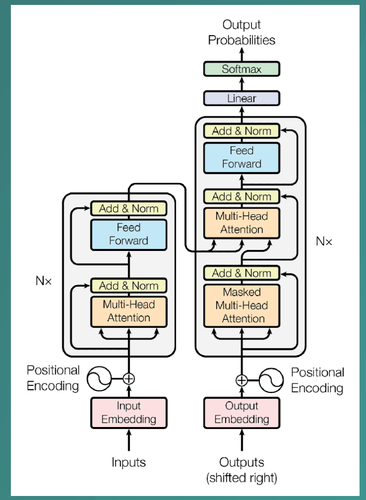
https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)

Traditional machine learning vs. generative AI on Android



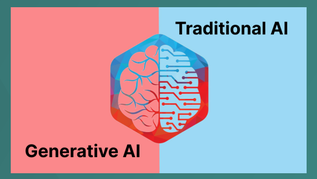


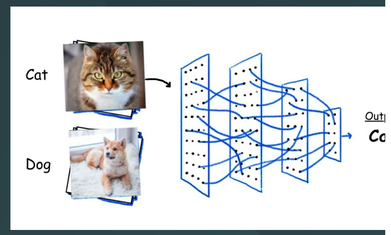




https://en.wikipedia.org/wiki/Diffusion_model

Traditional machine learning vs. generative AI on Android





2.0 Flash

Our newest multimodal model, with next generation features and improved capabilities

- Input audio, images, video, and text — get text, image, and audio responses
- Features low-latency conversational interactions with our Multimodal Live API

1.5 Flash

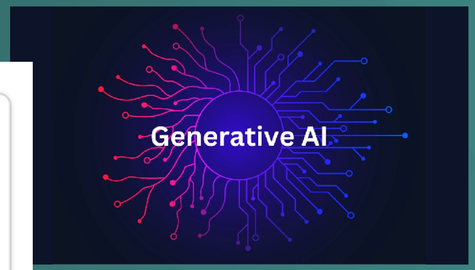
Our most balanced multimodal model with great performance for most tasks

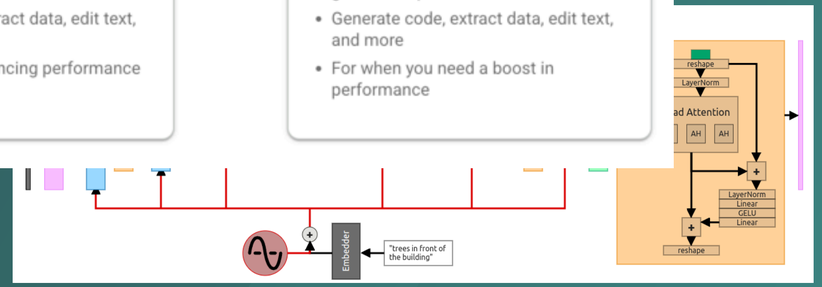
- Input audio, images, video, and text, get text responses
- Generate code, extract data, edit text, and more
- Best for tasks balancing performance and cost

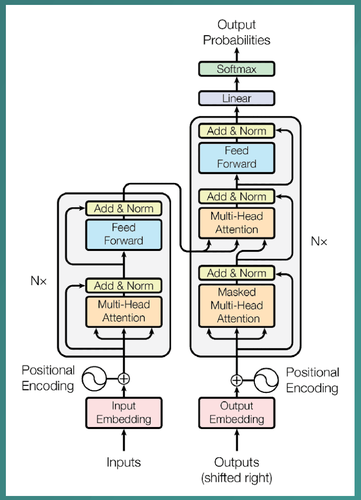
1.5 Pro

Our best performing multimodal model with features for a wide variety of reasoning tasks

- Input audio, images, video, and text, get text responses
- Generate code, extract data, edit text, and more
- For when you need a boost in performance



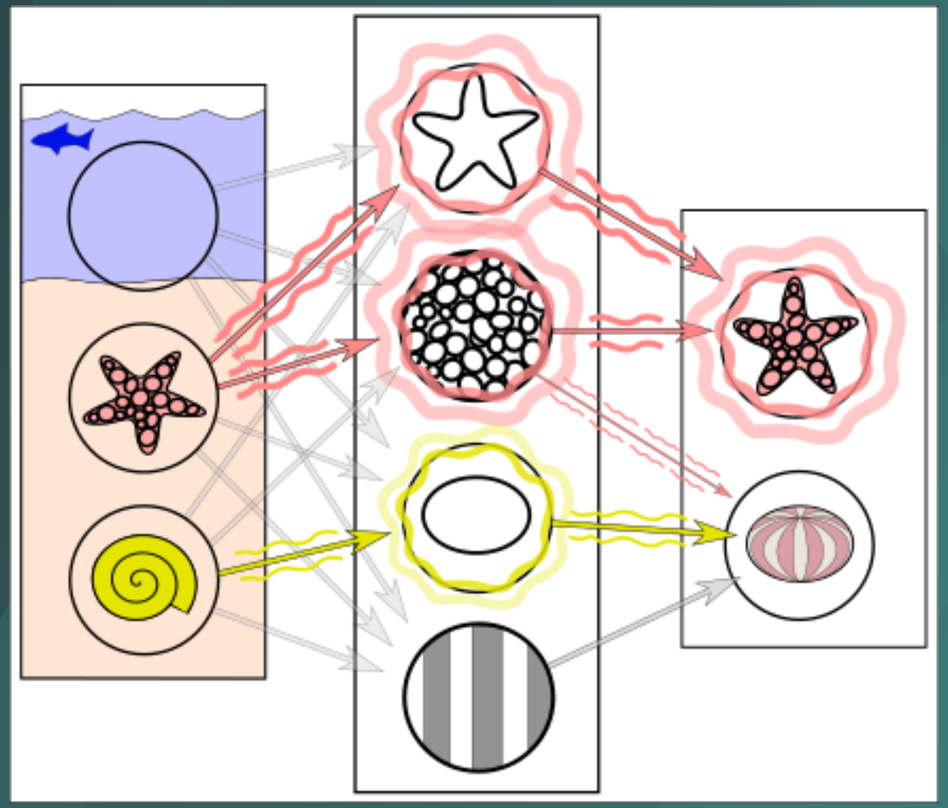




https://ai.google.dev/gemini-api/docs/models/gemini

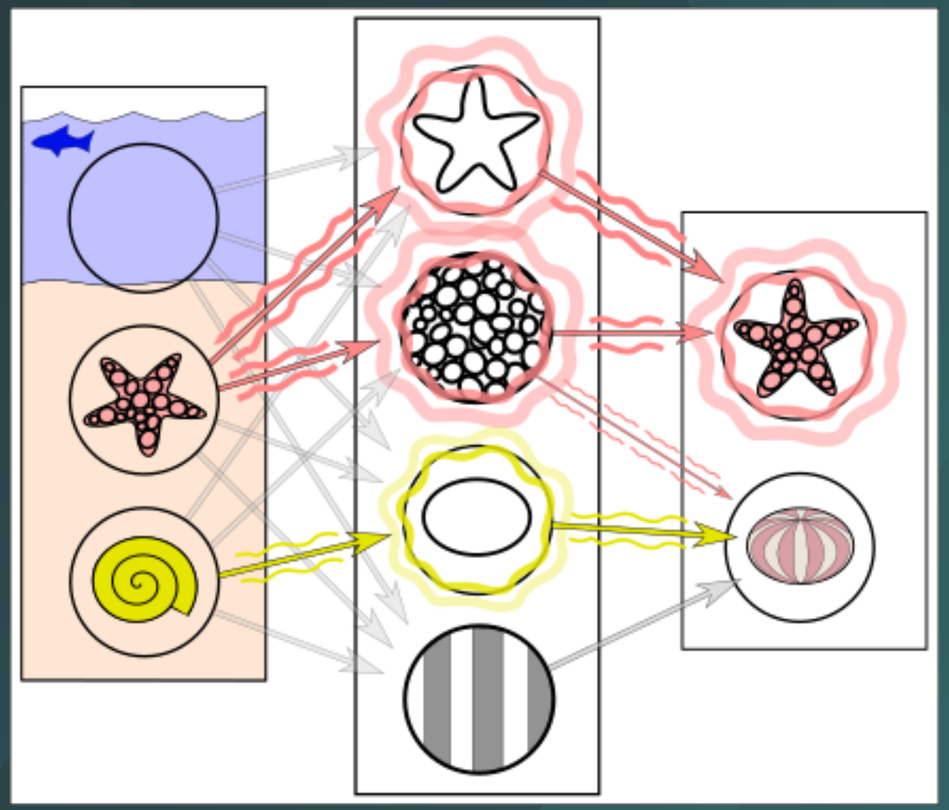
Understanding large language models

Understanding large language models



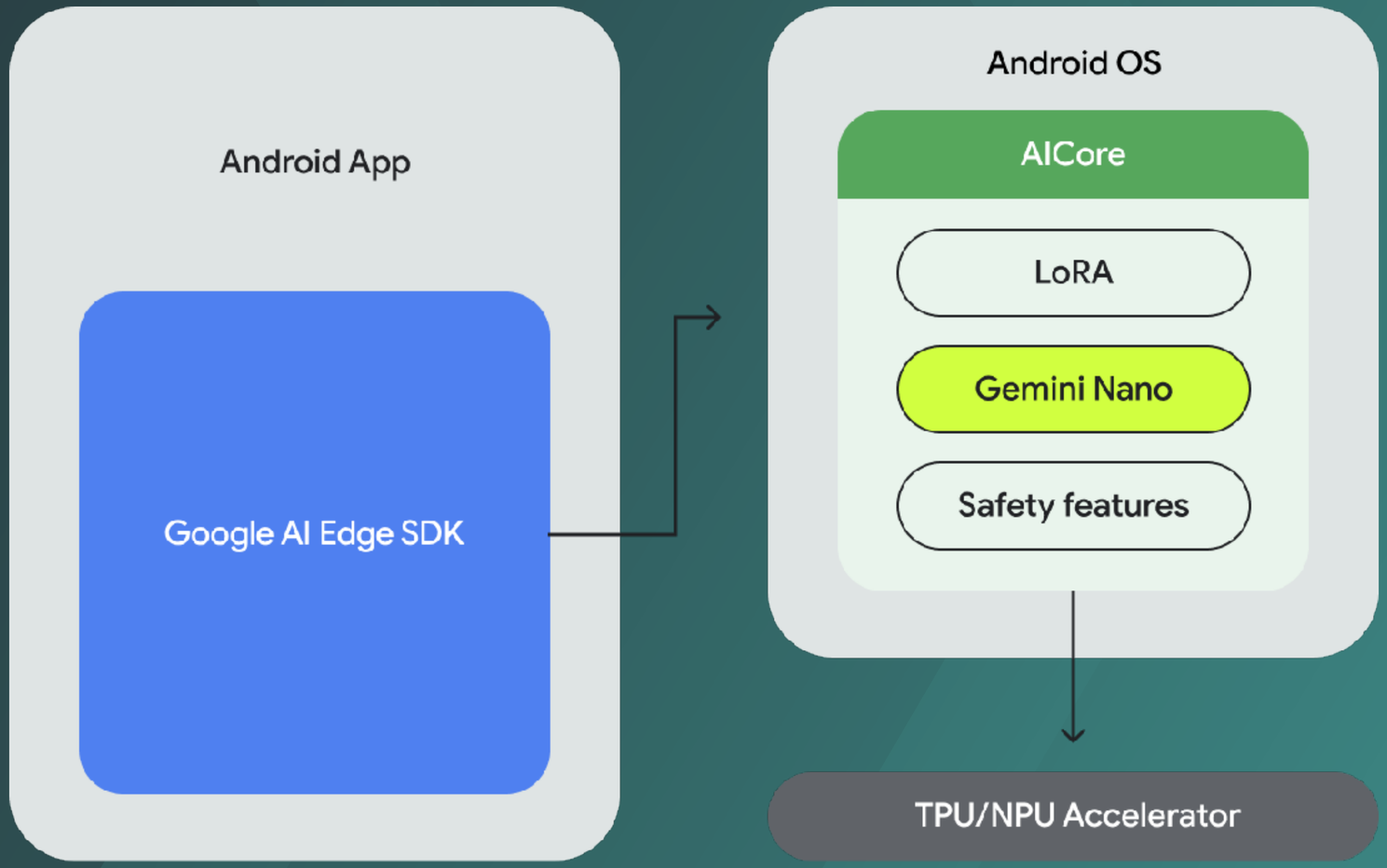
https://en.wikipedia.org/wiki/Neural_network_(machine_learning)

Understanding large language models



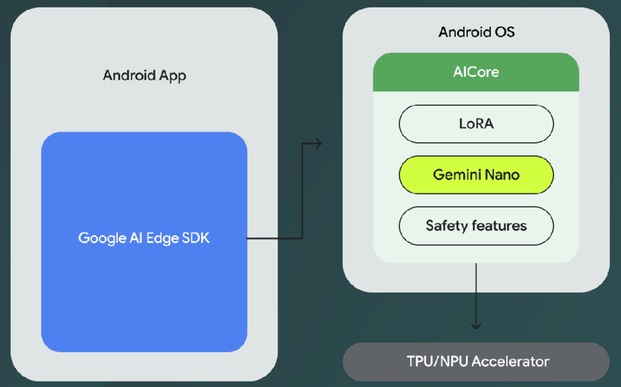
- Trained on massive amounts of text data.
- It learns:
  - Patterns.
  - Grammar.
  - Semantic relationships between words and phrases.
- Enabling it to predict and generate text that mimics human language.

Classes of models and their capabilities



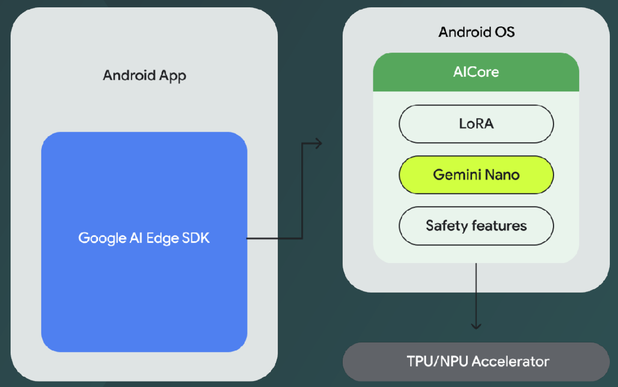
https://developer.android.com/ai/gemini-nano

Classes of models and their capabilities



https://ai.google.dev/gemma/docs

Classes of models and their capabilities



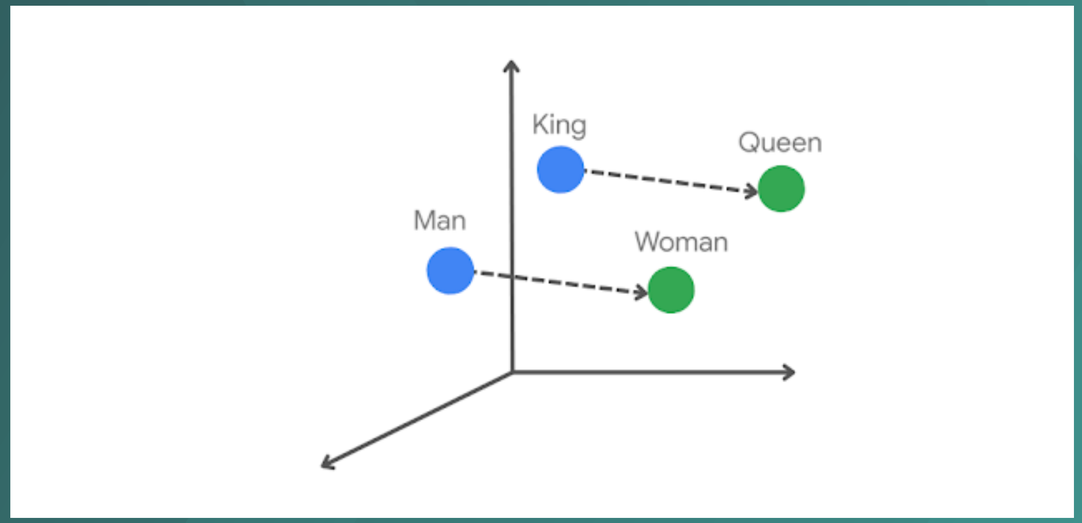
https://deepmind.google/technologies/gemini/pro/

Key concepts

- Context Window
  - The amount of tokens (converted from text, image, audio or video)
  - 100 tokens is equal to about 60-80 English words
  - Gemini 1.5 Pro currently supports 2M input token (It is enough to fit the seven Harry Potter books… and more!)

Key concepts

- Context Window
- Embeddings
  - Multidimensional numerical representations of tokens
  - Accurately encode their semantic meaning and relationships within a given vector space
  - Words with similar meanings are closer together, while words with opposite meanings are farther apart.



Key concepts

- Context Window
- Embeddings
- Top-K, Top-P and Temperature
  - Control the creativity of the model and the randomness of its output
  - Top-K filters tokens for output. Eg. Top-K of 3 keeps the three most probable tokens.
  - Tokens with the highest probabilities are selected until their sum equals the Top-P value
  - Temperature defines the randomness to select the tokens left

Key concepts

- Context Window
- Embeddings
- Top-K, Top-P and Temperature
- Fine-tuning
  - Iterating over several versions of a prompt to achieve an optimal response
  - By re-training it with data specific to your use-case

Android’s on-device GenAI

Benefits of on-device execution

Benefits of on-device execution

Benefits of on-device execution

Local processing

Offline availability

Potentially reduced latency

Benefits of on-device execution

Local processing

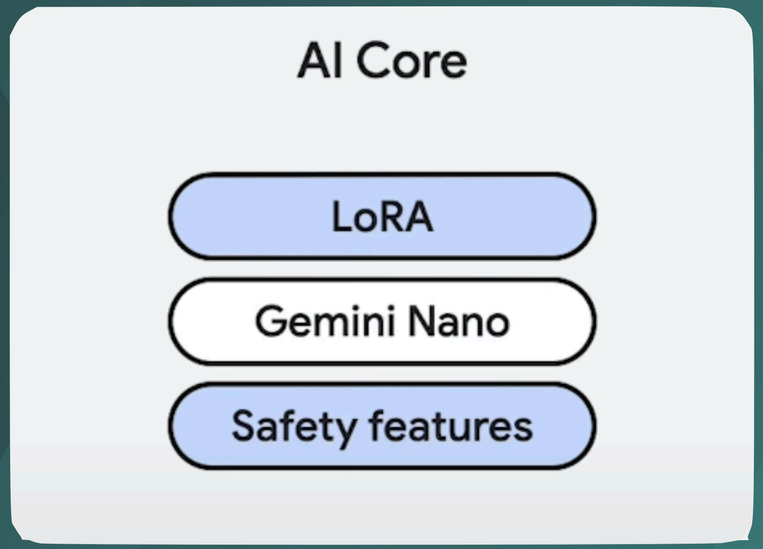
Offline availability

Potentially reduced latency

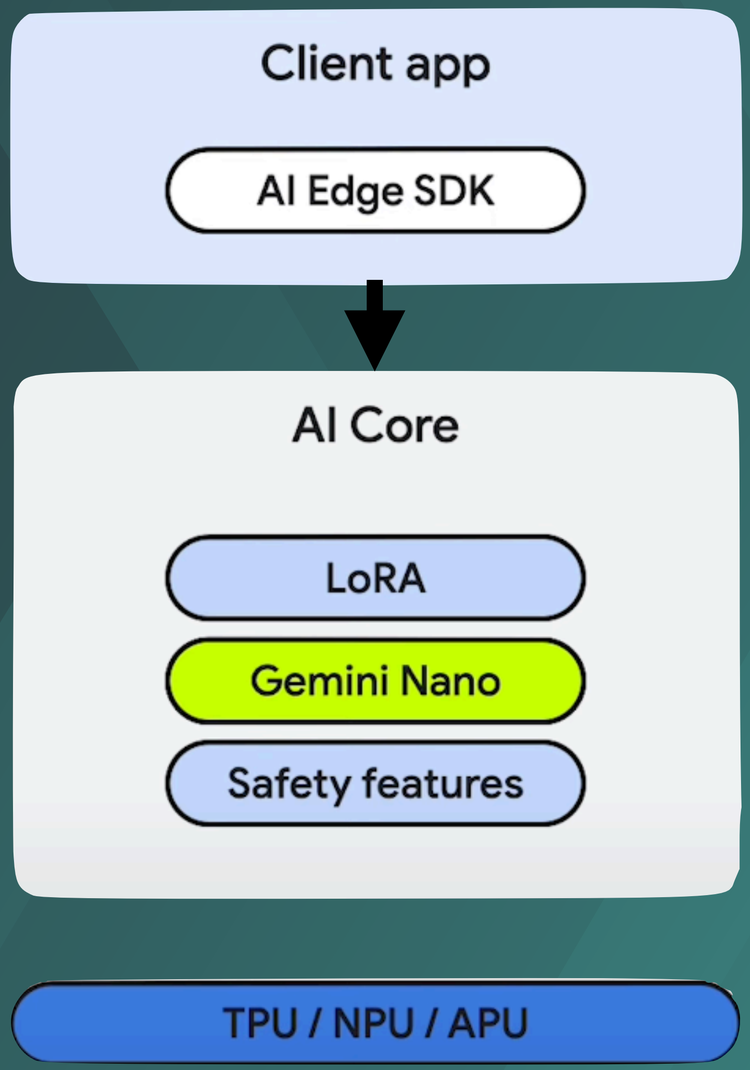
No additional cost

Gemini Nano

Gemini Nano



Gemini Nano


Client app
AI Edge SDK
AI Core
LoRA
Gemini Nano
Safety features
TPU / NPU / APU

Gemini Nano

Pixel Screenshots

Talkback

Messages

```
implementation("com.google.ai.edge.aicore:aicore:V.V.V")
```

https://developer.android.com/ai/gemini-nano/experimental

```
implementation("com.google.ai.edge.aicore:aicore:V.V.V")
```

```
val generationConfig = generationConfig {
  context = ApplicationProvider.getApplicationContext()
  workerExecutor = workerExecutor
  callbackExecutor = callbackExecutor
  temperature = 0.2f
  topK = 16
  candidateCount = 1
  maxOutputTokens = 256
}
```

```
val downloadCallback = object : DownloadCallback {
  override fun onDownloadDidNotStarte: GenerativeAIException) {
    // Log download start failure
  }
  override fun onDownloadPending() {
    // Log download pending
  }
  override fun onDownloadStarted(totalBytesToDownload: Long) {
    // Log download started
  }
  override fun onDownloadFailed(failureStatus: String, e: GenerativeAIException) {
    // Log download failed
  }
  override fun onDownloadProgress(totalBytesDownloaded: Long) {
    // Log download progress
  }
  override fun onDownloadCompleted(){
    // Download completed,
  }
}
```

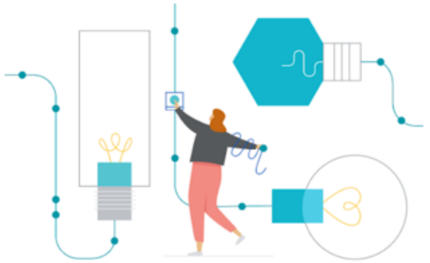
```
val model = GenerativeModel(
  generationConfig = generationConfig,
  downloadConfig = DownloadConfig(downloadCallback) // optional
}
```

```
scope. launch {
  // Single string input prompt
  val input = "I want you to act as an English proofreader.
  I will provide you texts, and I would like you to review
  them for any spelling, grammar, or punctuation errors.
  Once you have finished reviewing the text, provide me with
  any necessary corrections or suggestions for improving the
  text: These arent the droids your looking for."
  val response = generativeModel.generateContent(input)
  print (response.text)
}
```

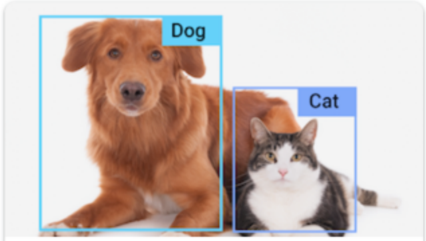

```
// Or multiple strings as input
val response = generativeModel.generateContent(
  content {
    text("I want you to act as an English proofreader.")
    text("I will provide you texts and I would like you to review")
    text("them for any spelling, grammar, or punctuation errors.")
    text ("Once you have finished reviewing the text,")
    text("provide me with any necessary corrections or suggestions")
    text("for improving the text:")
    text("These arent the droids your looking for.")
  })
  print (response. text)
}
```

On-device ML for everyone

Enjoy a new way to explore and evaluate on-device ML solutions.



VISION



Object Detection
Track and label objects in images.

See demo

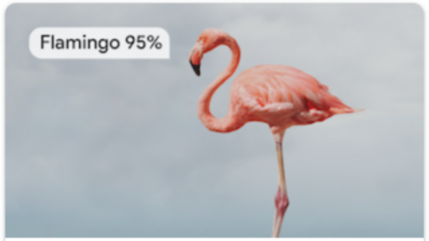


Image Classification
Identify content in images.

See demo

Image Segmentation
Locate objects and create image masks with labels.

See demo

Interactive Segmentation
Segment the object of interest in an image.

See demo

Gesture Recognition
Identify and recognize hand gestures.

See demo

Hand Landmark Detection
Detect hand landmarks.

See demo

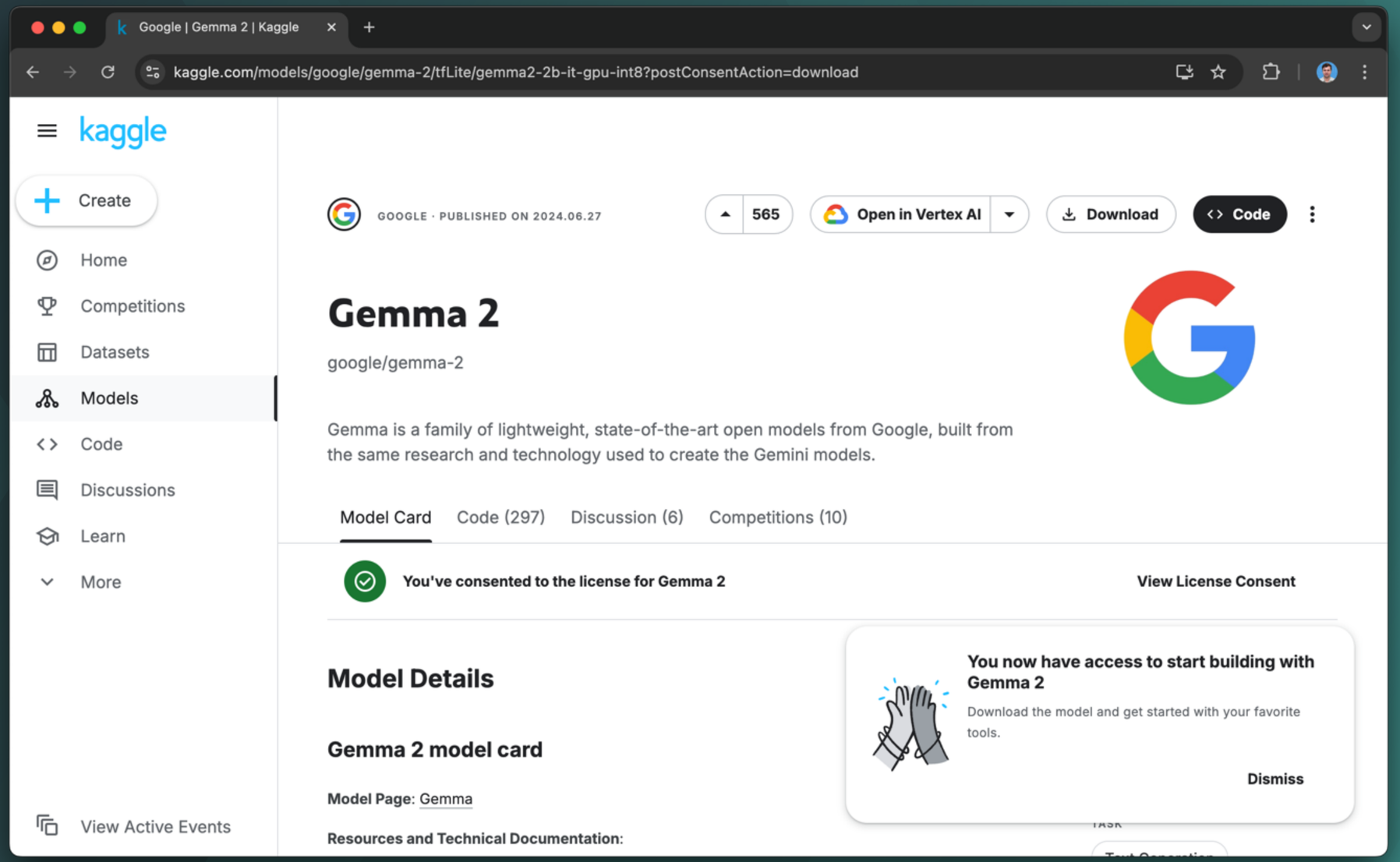
https://mediapipe-studio.webapps.google.com/home

```
implementation ("com.google.mediapipe:tasks-genai:V.V.V”)
```

https://ai.google.dev/edge/mediapipe/solutions/genai/llm_inference

```
implementation ("com.google.mediapipe:tasks-genai:V.V.V”)
```

```
val options = LlmInference.LlmInferenceOptions.builder()
.setTopK(5)
.setTemperature (0.9f)
.setMaxTokens ( 1028)
.setModelPath (MODEL_PATH)
.setLoraPath(LORA_PATH)
.setResultListener { partialResult, done ->
  _partialResults.tryEmit(partialResult to done)
.build()

llmInference = LlmInference.createFromOptions(context, options)
```


Google | Gemma 2 | Kaggle
kaggle.com/models/google/gemma-2/tfLite/gemma2-2b-it-gpu-int8?postConsentAction=download
kaggle
Create
Home
Competitions
Datasets
Models
Code
Discussions
Learn
More
View Active Events
GOOGLE · PUBLISHED ON 2024.06.27
565
Open in Vertex AI
Download
Code
Gemma 2
google/gemma-2
Gemma is a family of lightweight, state-of-the-art open models from Google, built from the same research and technology used to create the Gemini models.
Model Card
Code (297)
Discussion (6)
Competitions (10)
You've consented to the license for Gemma 2
View License Consent
Model Details
Gemma 2 model card
Model Page: Gemma
Resources and Technical Documentation:
You now have access to start building with Gemma 2
Download the model and get started with your favorite tools.
Dismiss

```
$ adb shell rm -r /data/local/tmp/llm/ # Remove any previously loaded models
$ adb shell mkdir -p /data/local/tmp/llm/
$ adb push output_path /data/local/tmp/llm/model_version.bin
```

DEMO

```
$ adb shell rm -r /data/local/tmp/llm/ # Remove any previously loaded models
$ adb shell mkdir -p /data/local/tmp/llm/
$ adb push output_path /data/local/tmp/llm/model_version.bin
```

1. To open the Device Explorer, select **View > Tool Windows > Device Explorer** or click the **Device Explorer** button in the tool window bar.
2. Select a device from the drop-down list.
3. Interact with the device content in the file explorer window:
    - Right-click a file or directory to create a new file or directory.
    - Save, upload, delete, or synchronize the selected file or directory to your machine.
    - Double-click a file to open it in Android Studio.

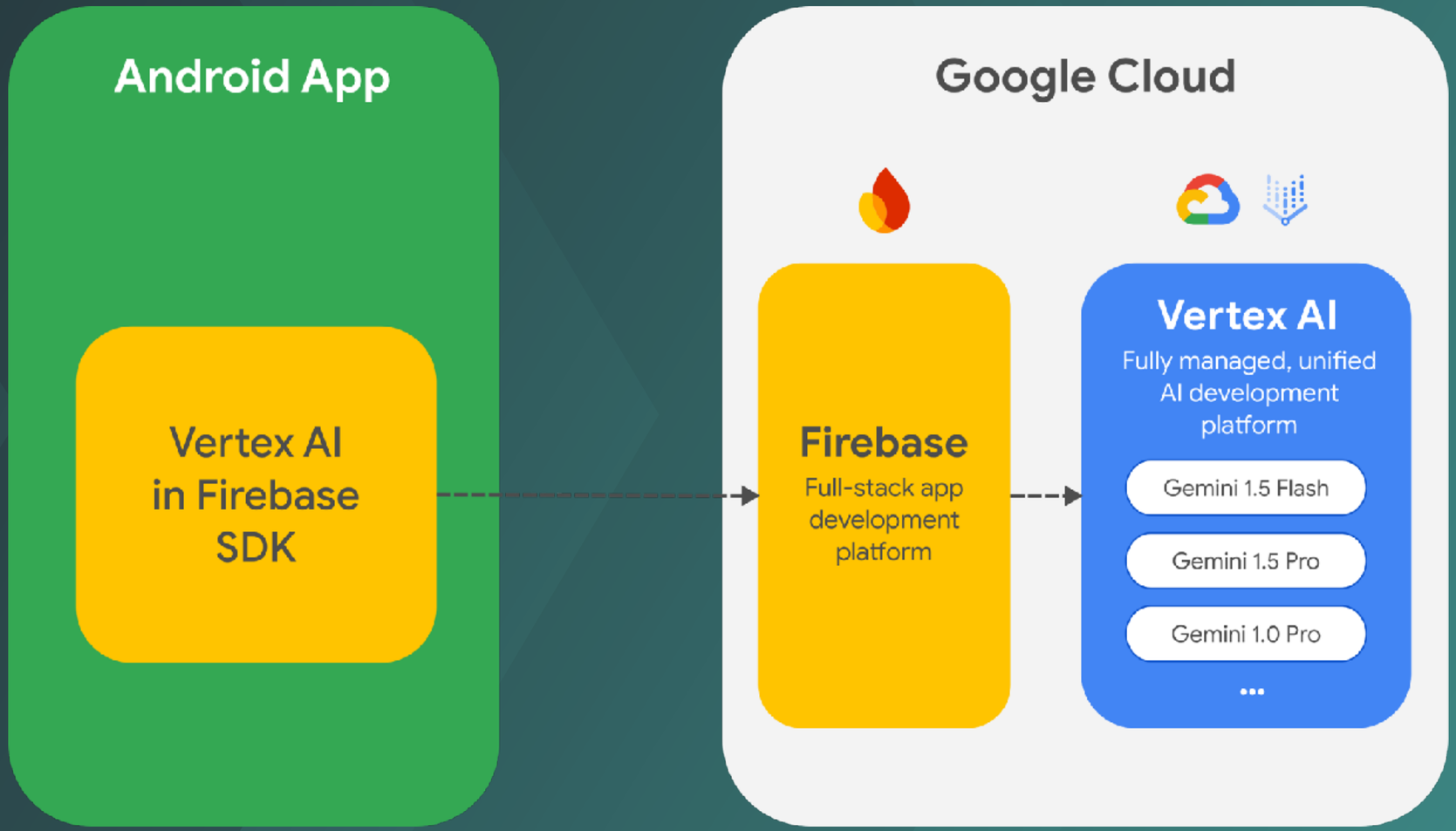
Cloud

- Apps and Services
- Gemini models
- TPUs

Devices

- Gemini, AI assistant
- Apps
- Android OS
- Gemini Nano
- Silicon

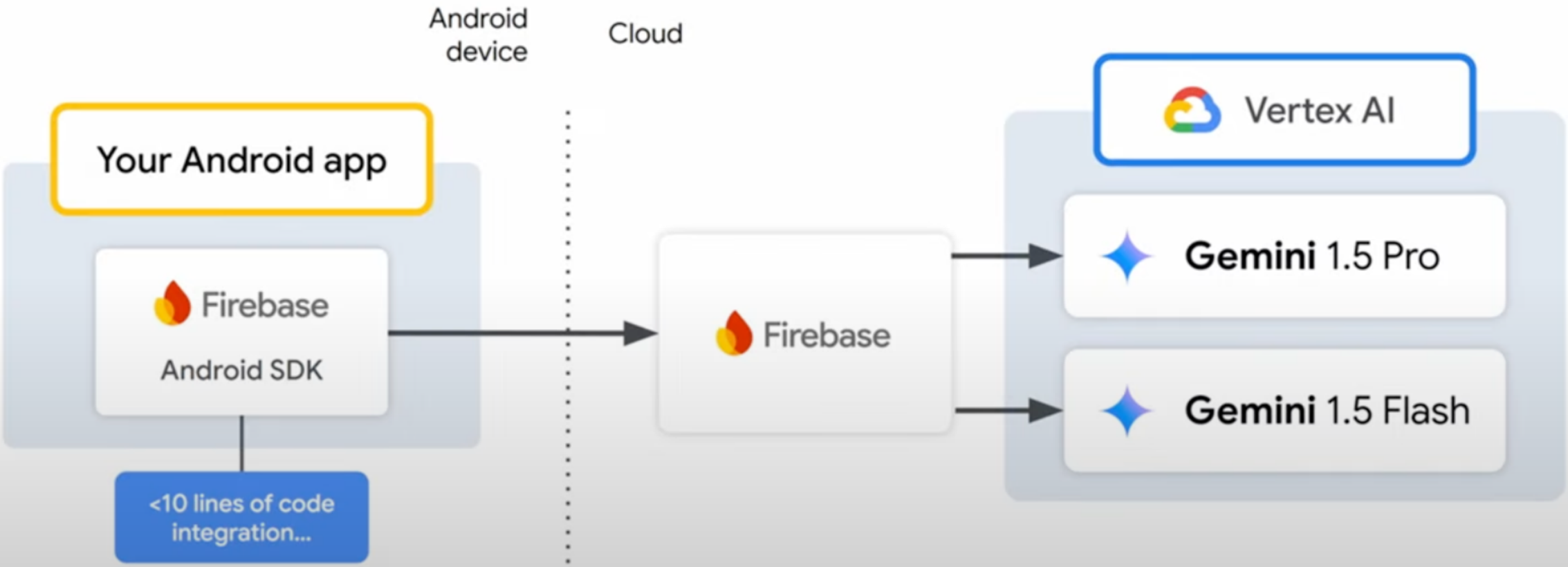
Vertex AI in Firebase



https://developer.android.com/ai/vertex-ai-firebase

Vertex AI in Firebase

Easily access Gemini Cloud models to your Android app:



1. Gradle imports

```
dependencies {
    // [...]

    // Import the BoM for the Firebase platform
    implementation(platform("com.google.firebase:firebase-bom:33.5.1"))

    // Add the dependency for the Vertex AI in Firebase library
    implementation("com.google.firebase:firebase-vertexai")
}
```

DEMO

2. Instantiate model

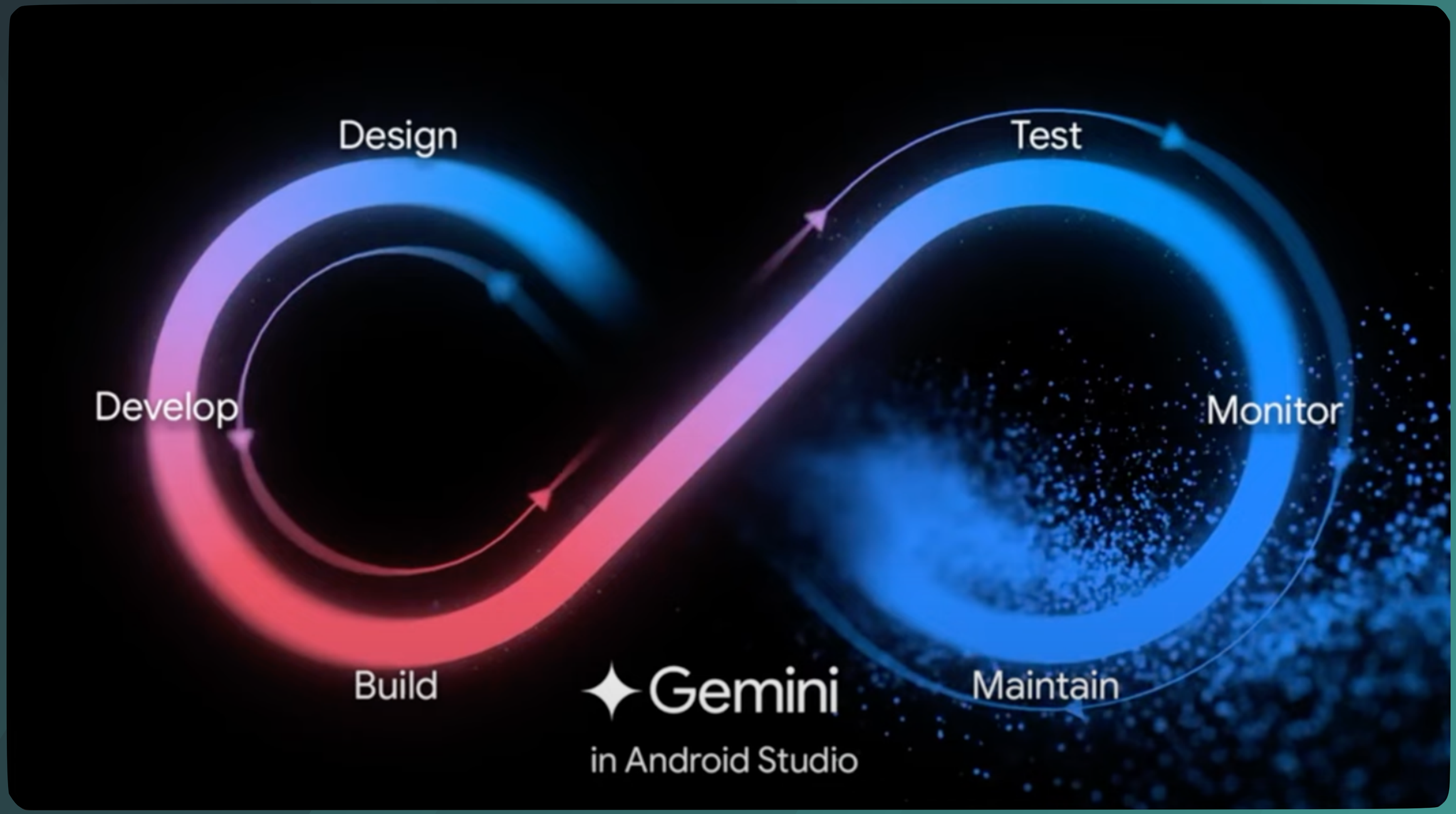
```
val generativeModel = Firebase.vertexAI.generativeModel("gemini-1.5-flash")
```

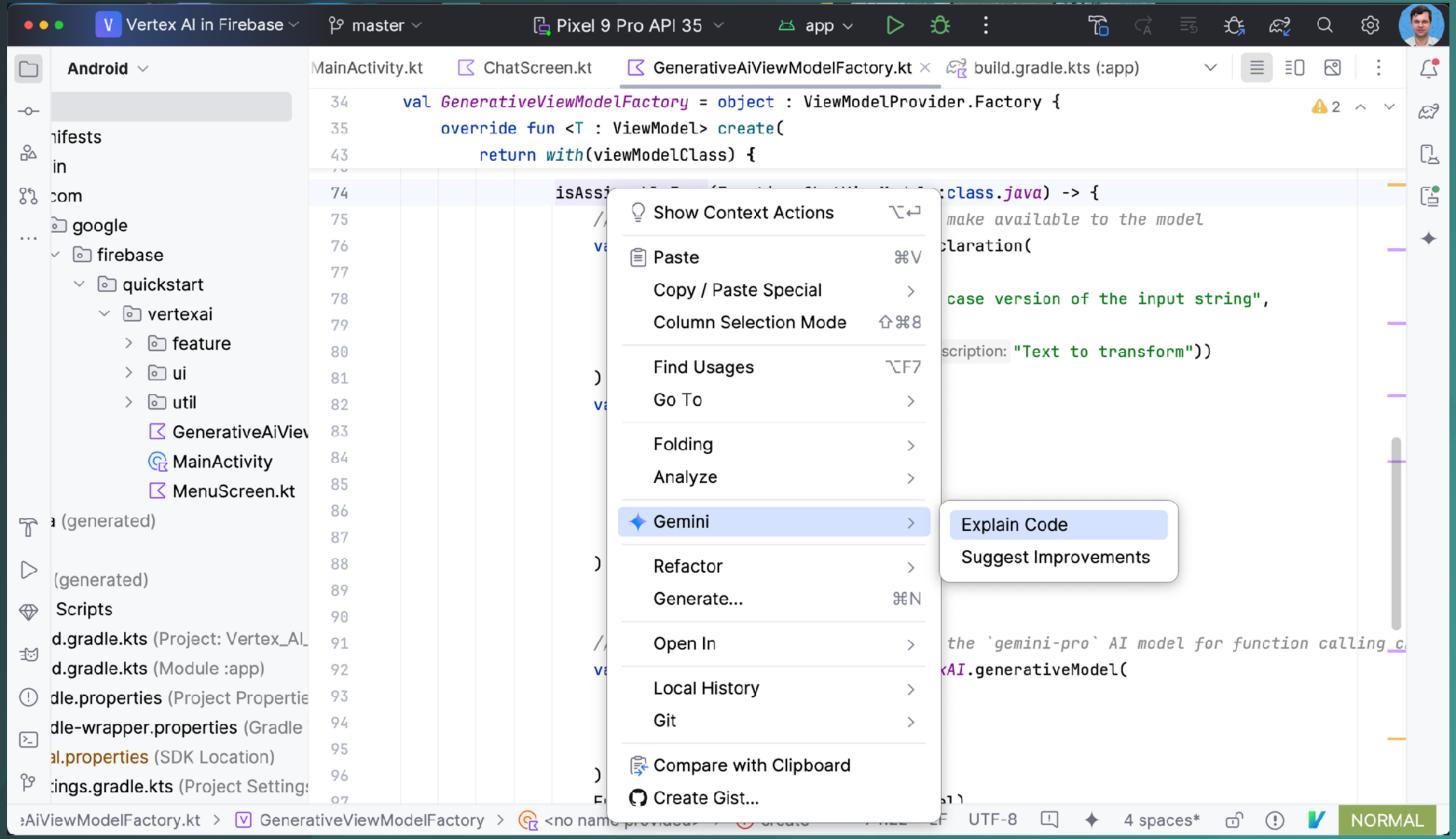
3. Generate content

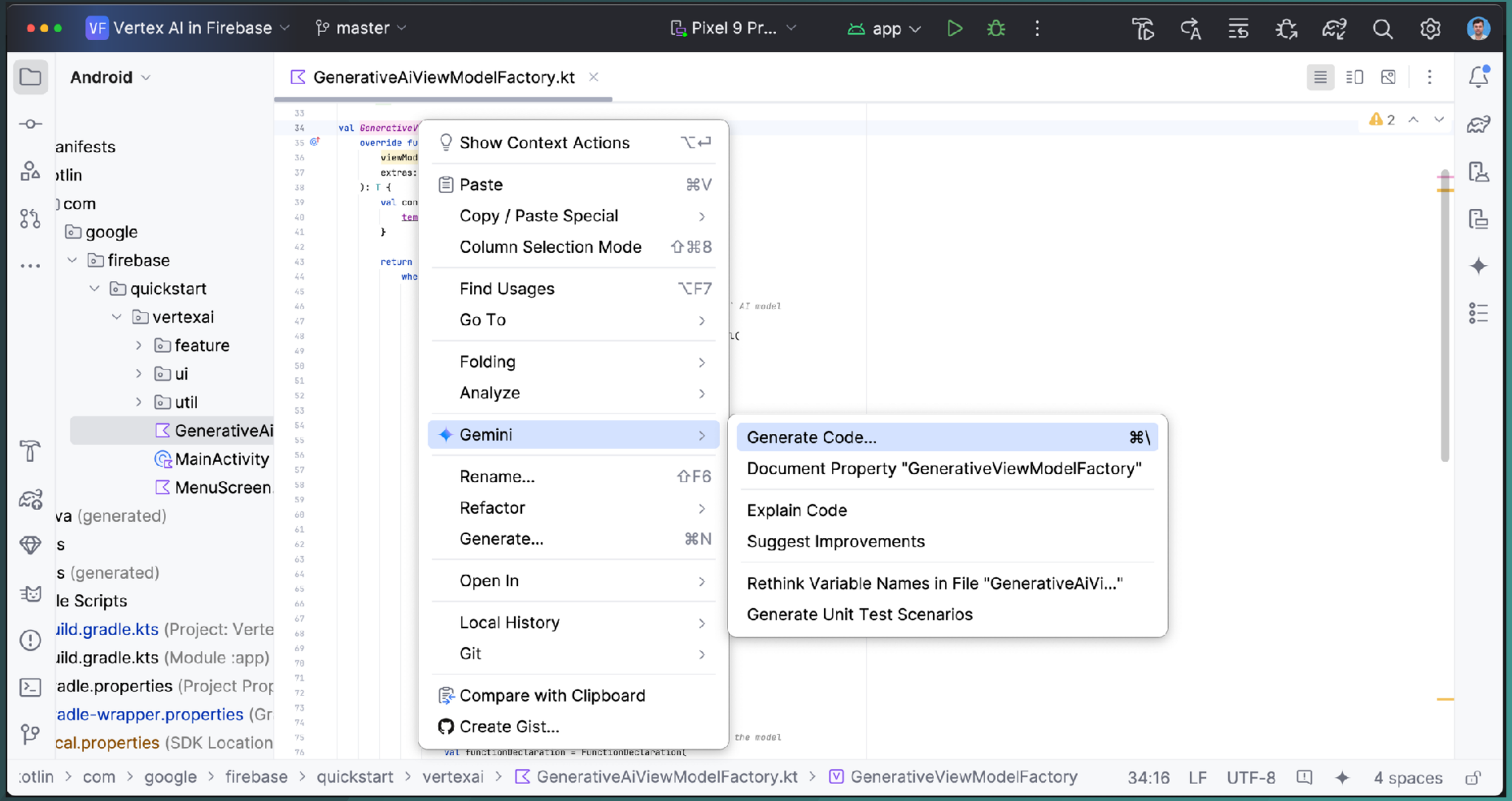
```
// Create a prompt
val prompt = "Write a story about a magic backpack."

// Call generateContent with the prompt
someScope.launch {
    val response = generativeModel.generateContent(prompt)
    print(response.text)
}
```

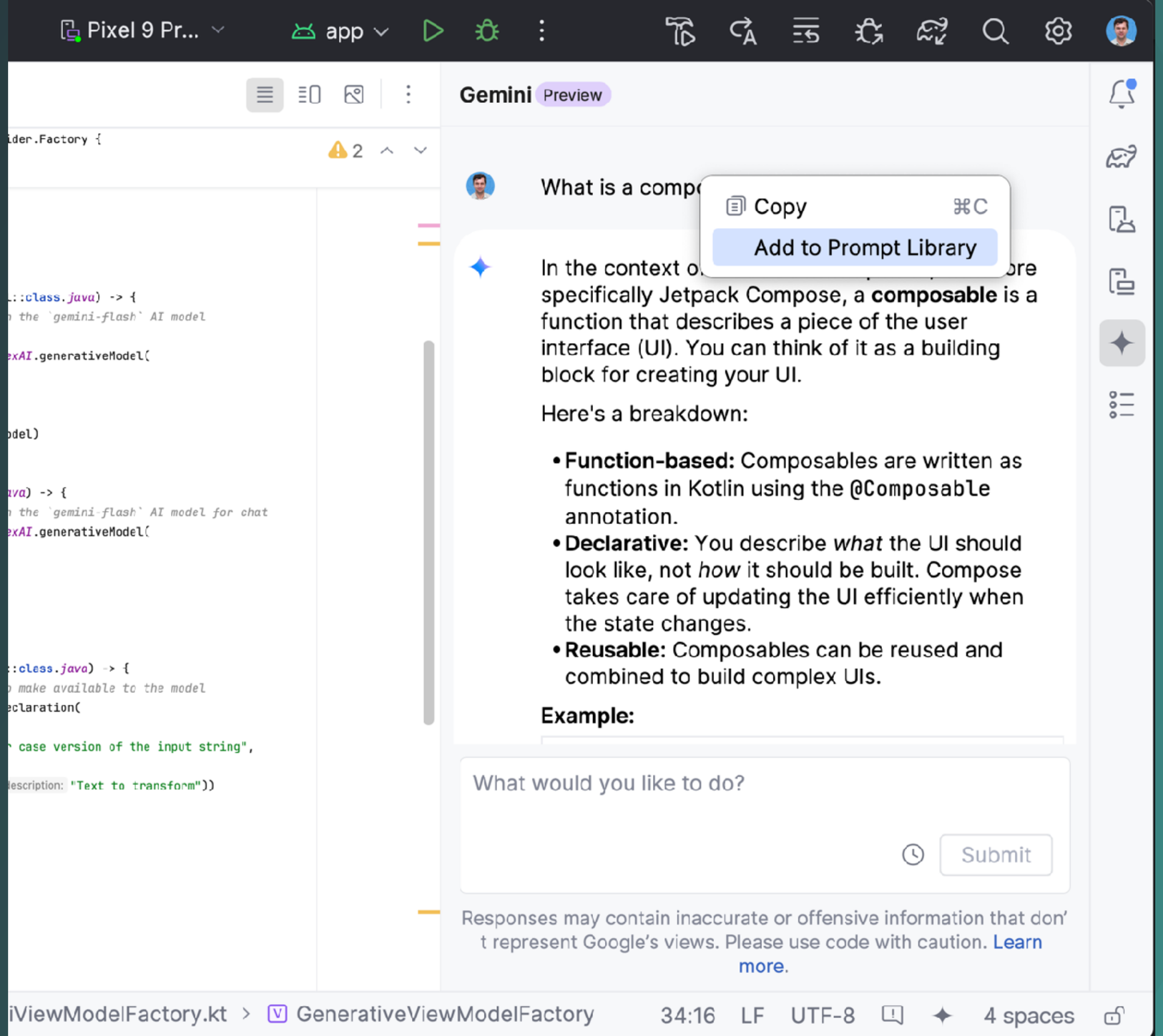
Gemini in Android Studio

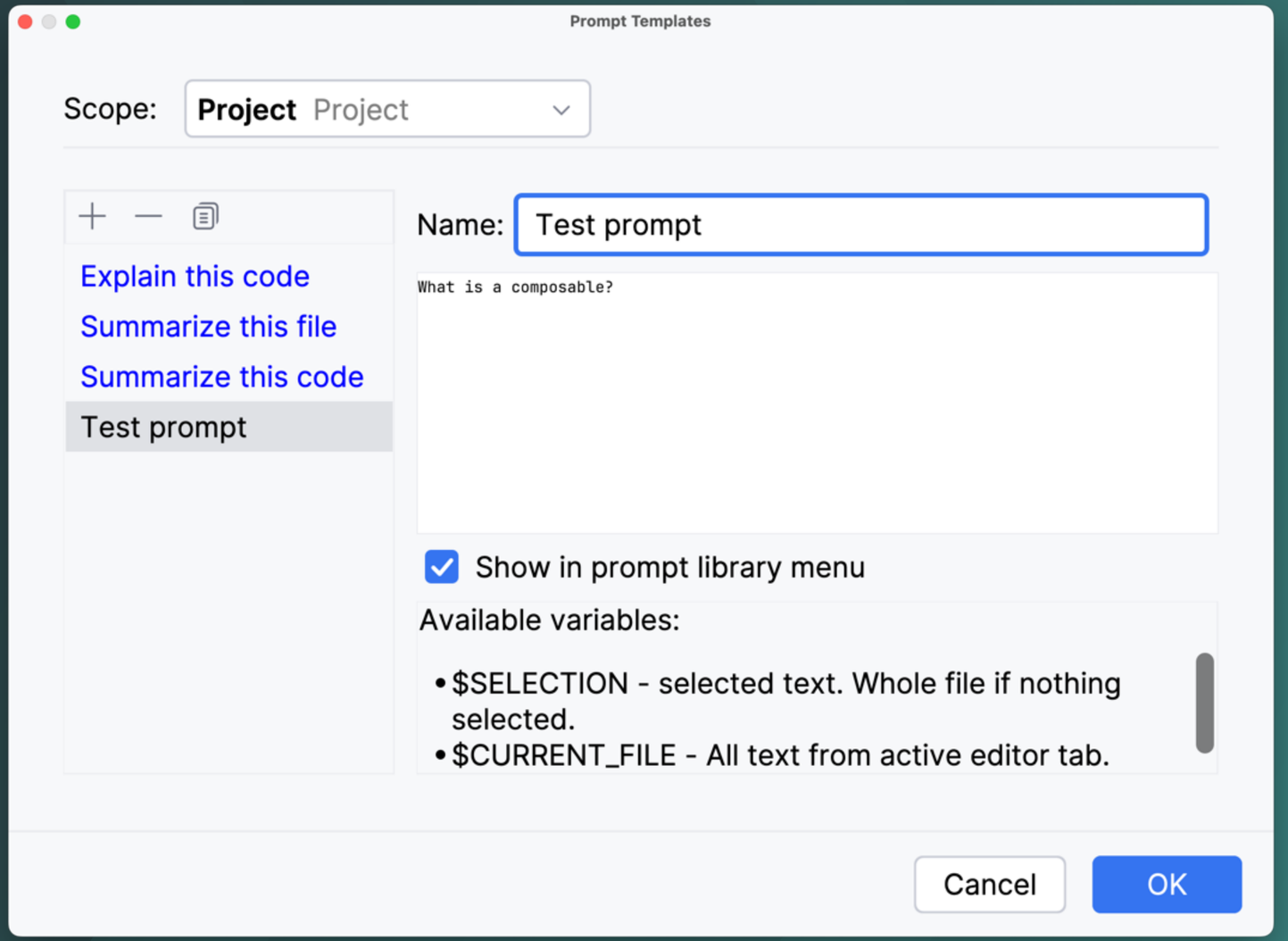
Accelerate your development experience

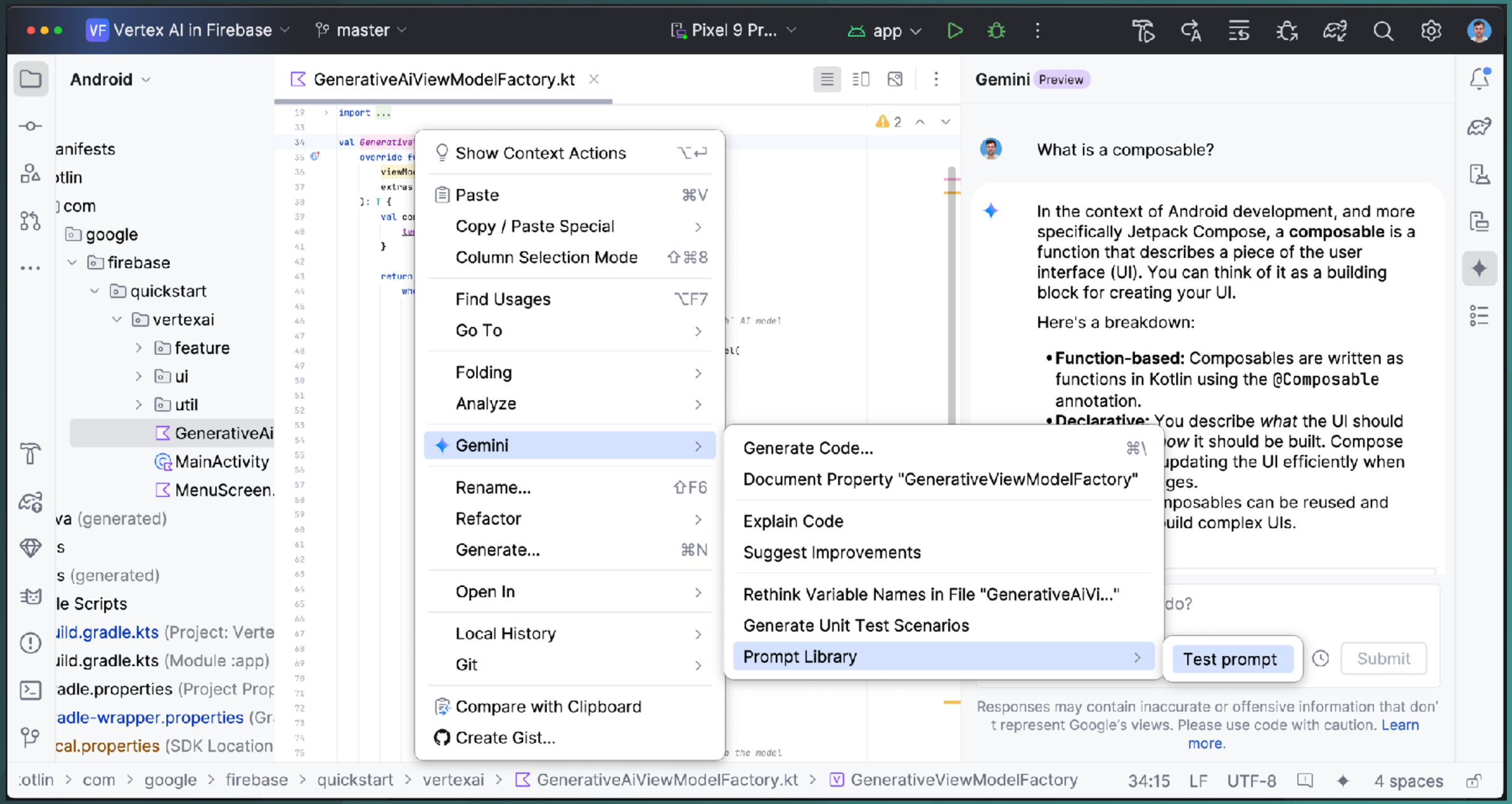

Design
Test
Develop
Monitor
Build
Maintain
Gemini
in Android Studio


Vertex AI in Firebase
master
Pixel 9 Pro API 35
app
Android
MainActivity.kt
ChatScreen.kt
GenerativeAiViewModelFactory.kt
build.gradle.kts (:app)
val GenerativeViewModelFactory = object : ViewModelProvider.Factory {
override fun <T : ViewModel> create(
return with(viewModelClass) {
Show Context Actions
Paste
Copy / Paste Special
Column Selection Mode
Find Usages
Go To
Folding
Analyze
Gemini
Refactor
Generate...
Open In
Local History
Git
Compare with Clipboard
Create Gist...
Explain Code
Suggest Improvements
NORMAL


Vertex AI in Firebase
master
Pixel 9 Pr...
app
Android
GenerativeAiViewModelFactory.kt
Show Context Actions ⌥↵
Paste ⌘V
Copy / Paste Special
Column Selection Mode ⇧⌘8
Find Usages ⌥F7
Go To
Folding
Analyze
Gemini
Rename... ⇧F6
Refactor
Generate... ⌘N
Open In
Local History
Git
Compare with Clipboard
Create Gist...
Generate Code... ⌘\
Document Property "GenerativeViewModelFactory"
Explain Code
Suggest Improvements
Rethink Variable Names in File "GenerativeAiVi..."
Generate Unit Test Scenarios
com
google
firebase
quickstart
vertexai
feature
ui
util
MainActivity
MenuScreen
com > google > firebase > quickstart > vertexai > GenerativeAiViewModelFactory.kt > GenerativeViewModelFactory
34:16
LF
UTF-8
4 spaces


Pixel 9 Pr...
app
Gemini
Preview
What is a compo
Copy
⌘C
Add to Prompt Library
In the context o
ore specifically Jetpack Compose, a composable is a function that describes a piece of the user interface (UI). You can think of it as a building block for creating your UI.
Here's a breakdown:
Function-based: Composables are written as functions in Kotlin using the @Composable annotation.
Declarative: You describe what the UI should look like, not how it should be built. Compose takes care of updating the UI efficiently when the state changes.
Reusable: Composables can be reused and combined to build complex UIs.
Example:
What would you like to do?
Submit
Responses may contain inaccurate or offensive information that don't represent Google's views. Please use code with caution. Learn more.
iViewModelFactory.kt
GenerativeViewModelFactory
34:16
LF
UTF-8
4 spaces


Prompt Templates
Scope:
Project Project
Explain this code
Summarize this file
Summarize this code
Test prompt
Name:
Test prompt
What is a composable?
Show in prompt library menu
Available variables:
• $SELECTION - selected text. Whole file if nothing selected.
• $CURRENT_FILE - All text from active editor tab.
Cancel
OK


Vertex AI in Firebase
master
Pixel 9 Pr...
app
Android
GenerativeAiViewModelFactory.kt
Gemini
Preview
What is a composable?
In the context of Android development, and more specifically Jetpack Compose, a composable is a function that describes a piece of the user interface (UI). You can think of it as a building block for creating your UI.
Here's a breakdown:
Function-based: Composables are written as functions in Kotlin using the @Composable annotation.
Show Context Actions
Paste
Copy / Paste Special
Column Selection Mode
Find Usages
Go To
Folding
Analyze
Gemini
Rename...
Refactor
Generate...
Open In
Local History
Git
Compare with Clipboard
Create Gist...
Generate Code...
Document Property "GenerativeViewModelFactory"
Explain Code
Suggest Improvements
Rethink Variable Names in File "GenerativeAiVi..."
Generate Unit Test Scenarios
Prompt Library
Test prompt
Submit
Responses may contain inaccurate or offensive information that don't represent Google's views. Please use code with caution. Learn more.
com
google
firebase
quickstart
vertexai
feature
ui
util
MainActivity
MenuScreen
GenerativeViewModelFactory
34:15
LF
UTF-8
4 spaces

Lecture outcomes

- Gemini Nano
- Google AI Edge SDK
- MediaPipe
- Cloud AI