GenAI with the Gemini API in Vertex AI

Olga Mirensky
Lead Platform Engineer

GenAI with the Gemini API in Vertex AI

PyPi Package: `google-genai`

GitHub: `googleapis/python-genai`

Google Gen AI Python SDK to integrate Google's generative models into Python apps.

Gemini Developer API

individual devs, small projects and prototyping.
Auth with API Key

Vertex AI

Enterprise, production-ready, integrated with GCP environment

A Few Topics

```
from google import genai
from google.genai.types import (
    CreateBatchJobConfig,
    CreateCachedContentConfig,
    EmbedContentConfig,
    FunctionDeclaration,
    GenerateContentConfig,
    HarmBlockThreshold,
    HarmCategory,
    Part,
    SafetySetting,
    Tool,
)
```

Chat

```
client = genai.Client(vertexai=True, project=PROJECT_ID, location=...)

chat = client.chats.create(
    model=MODEL_ID,
    config=GenerateContentConfig(
        system_instruction=system_instruction,
        temperature=0.5,
    ),
)

response = chat.send_message("prompt”)
```

Safety, tools, response type, response schema, cache, etc

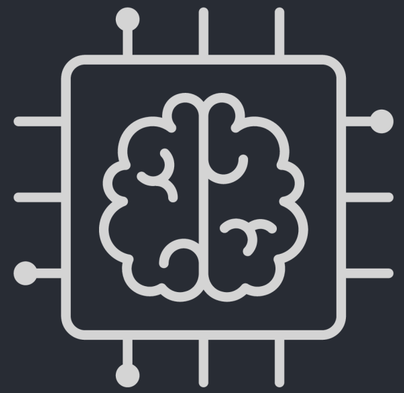
Batch Predictions

```
# Specify model and source file only, destination and job display
name will be auto-populated

job = client.batches.create(
    model='gemini-2.5-flash',
    src='bq://my-project.my-dataset.my-table',
                        # or "gs://path/to/input/data"
)

print(job)
```

Function Calling

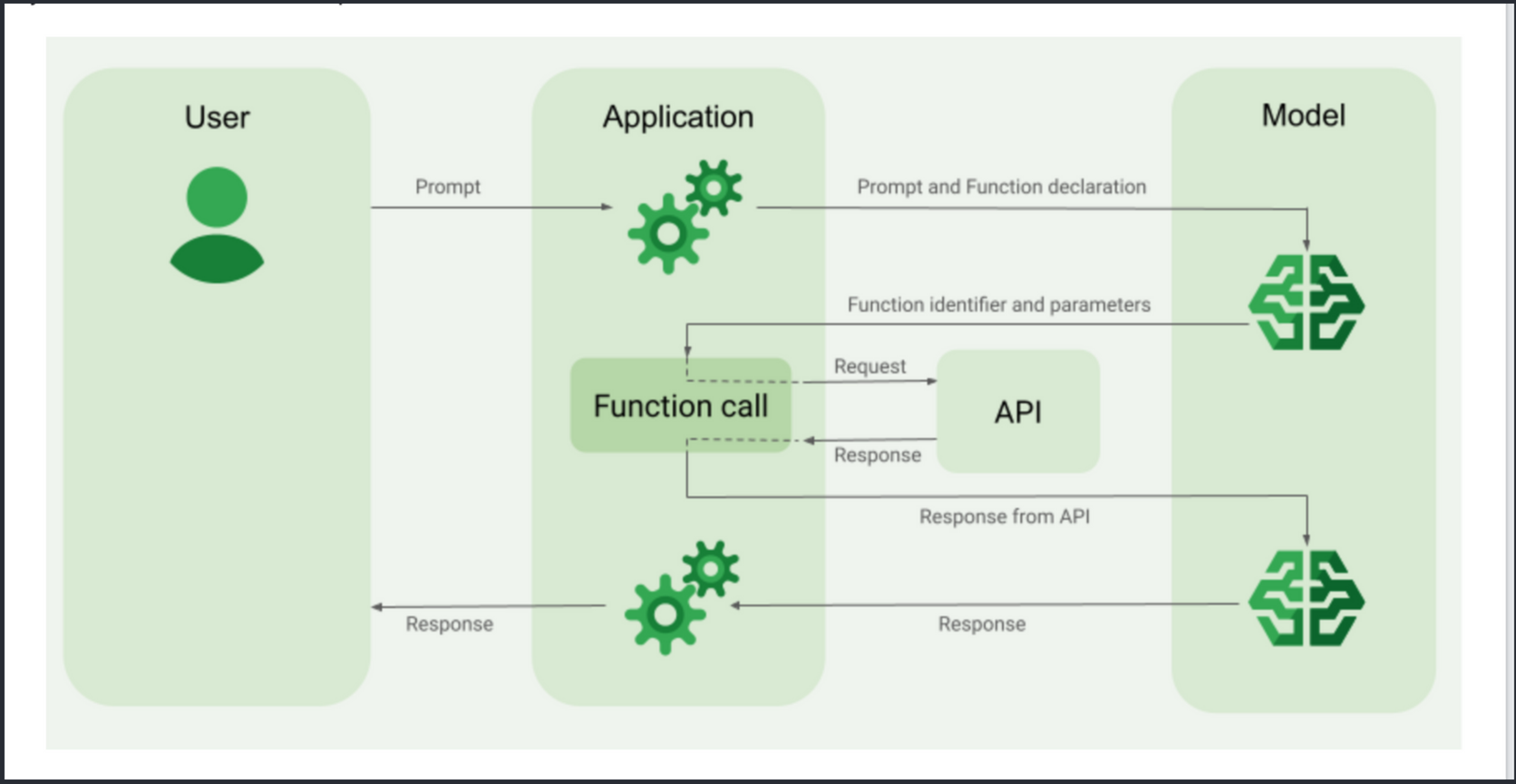
`FunctionDeclaration`
(spec)

User free-text form



here's what user wants. In your code call this function <function_name> with these params <param1, param2>

Function Calling (cont.)



https://cloud.google.com/vertex-ai/generative-ai/docs/multimodal/function-calling

Token Count

```
rsp = client.models.compute_tokens
```

rsp = client.models.generate_content(...)
rsp.usage_metadata

Cache

Implicit Caching (default, no cost)

Explicit Caching

70-90% cost saving on tokens served from cache (vs regular tokens)

Pricing (info only)

Gemini 2.5				Cached		Batch	
Model	Type	Price (/1M tokens) <= 200K input tokens	Price (/1M tokens) > 200K input tokens	Price (/1M tokens) <= 200K **cached** input tokens	Price (/1M tokens) > 200K **cached** input tokens	Price (/1M tokens) <= 200K input tokens with batch API	Price (/1M tokens) > 200K input tokens with batch API
Gemini 2.5 Pro	Input (text, image, video, audio)	$1.25	$2.5	$0.125	$0.250	$0.625	$1.25
	Text output (response and reasoning)	$10	$15	N/A	N/A	$5	$7.5

https://cloud.google.com/vertex-ai/generative-ai/pricing

Links

Lab:
https://www.skills.google/course_templates/959

GitHub Repo with Lab Notebooks:
https://github.com/GoogleCloudPlatform/generative-ai/