

TÀI LIỆU TÓM TẮT – GEMINI 3 PRO API (TIẾNG VIỆT)

1. Giới thiệu

Gemini 3 Pro là mô hình đa phương thức mới của Google, hỗ trợ văn bản, hình ảnh, video và tài liệu dung lượng lớn (lên đến hơn 1M tokens). Dùng được qua Gemini API hoặc Vertex AI.

2. Model ID thông dụng

- gemini-3-pro-preview
- gemini-3-pro-latest (tùy thời điểm Google cập nhật)
- gemini-3.0-pro (bản nhúng sẵn trên Google Cloud)

3. Gọi API (REST)

POST <https://generativelanguage.googleapis.com/v1/models/{model}:generateContent>

BODY:

```
{
  "contents": [{
    "parts": [{"text": "Xin hãy tóm tắt nội dung sau..."}]
  }]
}
```

4. Gọi API bằng JavaScript (Node.js)

```
import { GoogleGenerativeAI } from "@google/generative-ai";

const genAI = new GoogleGenerativeAI(process.env.GEMINI_API_KEY);

const model = genAI.getGenerativeModel({ model: "gemini-3-pro-preview" });

const result = await model.generateContent("Xin chào");

console.log(result.response.text());
```

5. Gọi API bằng Python

```
import google.generativeai as genai

genai.configure(api_key="YOUR_API_KEY")

model = genai.GenerativeModel("gemini-3-pro-preview")
```

```
resp = model.generate_content("Xin chào")  
print(resp.text)
```

6. Tham số quan trọng

- temperature: điều chỉnh mức sáng tối
- top_p, top_k: kiểm soát phân phối
- max_output_tokens: giới hạn số dài trả lại
- thinking_level: điều khiển mức "lập luận sâu"
- media_resolution: phân giải khi xử lý ảnh/video

7. Upload file (PDF, hình ảnh, video)

Gemini API hỗ trợ upload và phân tích:

POST <https://generativelanguage.googleapis.com/v1/files>

Sử dụng multipart/form-data.

8. Streaming Response

API hỗ trợ stream dữ liệu:

```
model.generate_content_stream()
```

9. Giới hạn & quota

- Token input realtime (hạn 1M tokens)
- Tốc độ và hạn mức tùy gói Google Cloud / AI Studio
- API key cần bảo mật, tránh đưa vào client frontend

10. Tài liệu chính thức

- <https://ai.google.dev/gemini-api/docs>
- <https://ai.google.dev/gemini-api/docs/gemini-3>
- <https://cloud.google.com/vertex-ai/docs/generative-ai>