【Gemini】Gemini vs Vertex AI

vertex ai 與 gemini. 是什麼關係

簡單來說，**Vertex AI 是「平台」，而 Gemini 是「模型」**。

它們的關係可以類比為：**Vertex AI 是「廚房（包含所有工具和設備）」，而 Gemini 是「最頂級的食材或主廚」**。

以下是更詳細的關係拆解：

1. 定位上的不同

- **Gemini (模型)：** 是 Google 開發的最強大、最通用的**人工智慧模型家族**（包括 Gemini Ultra, Pro, Flash 等）。它具備多模態能力（能處理文字、圖片、影片、程式碼），是 AI 的「大腦」。
- **Vertex AI (平台)：** 是 Google Cloud 旗下的一個**企業級機器學習 (ML) 平台**。它的目的是幫助開發者和企業「管理、部署和擴展」AI 模型。

2. 合作方式：Gemini 運行在 Vertex AI 之上

當一個企業想要在自己的產品（例如手機 App、公司內部系統）中使用 Gemini 的能力時，他們通常會透過 Vertex AI 來達成：

- **API 調用：** 開發者透過 Vertex AI 提供的 API 來呼叫 Gemini 模型。
- **模型自定義 (Tuning)：** 企業可以將自己的私有數據導入 Vertex AI，對 Gemini 進行「微調」，讓它更懂該公司的專業知識，而這些數據不會被拿去訓練 Google 的公共模型。
- **落地應用 (RAG)：** 利用 Vertex AI 的工具（如向量搜尋），讓 Gemini 結合企業內部的文檔來回答問題（這稱為檢索增強生成）。
- **安全性與控管：** Vertex AI 提供企業級的資料安全、隱私保護和權限管理，確保使用 Gemini 時符合法規。

3. 為什麼不直接用 Gemini，而要透過 Vertex AI？

如果你只是想聊天，你可以直接去 `gemini.google.com`。但如果你是**企業或開發者**，透過 Vertex AI 使用 Gemini 有以下好處：

1. **資料隱私：** 在 Vertex AI 中處理的資料，預設不會被 Google 用來訓練模型。
2. **穩定性：** 提供服務水準協議 (SLA) 保障，適合正式產品環境。
3. **完整工具鏈：** 除了 Gemini，Vertex AI 還有「模型花園」(Model Garden)，你也可以在那裡使用 Llama (Meta) 或其他開源模型。
4. **監控與評估：** 幫助你檢查 Gemini 回答的品質，是否有偏見或錯誤。

總結

- **Gemini** 是 Google 提供的**強大 AI 引擎**。
- **Vertex AI** 是 Google Cloud 提供的**開發環境與工具箱**。
- **關係：** Vertex AI 是企業獲取、客製化並部署 Gemini 模型的主要管道。

一句話總結： Gemini 提供「智慧」，而 Vertex AI 提供「將這份智慧轉化為商業應用」所需的所有基礎設施。

使用gemini 問答，gemini cli，VertexAI 呼叫api 得到答案，比較三種模式

這三種模式代表了從「普通用戶」到「開發者」再到「企業級應用」的三種不同維度。以下是針對這三種模式的詳細比較：

---

1. Gemini 網頁版問答 (gemini.google.com)

這是最直觀的 **B2C (消費者)** 模式，類似於 ChatGPT 的網頁界面。

- **目標對象：** 一般大眾、學生、辦公室職員、內容創作者。
- **介面：** 圖形化網頁或 App 介面，支援多模態輸入（圖片、文件上傳）。
- **費用：** 免費（Gemini Flash）或 月費訂閱制（Gemini Advanced / Ultra）。
- **資料隱私：** **較低。** 除非特別設定，否則對話內容可能會被 Google 用於改進模型（去識別化後的人工審核）。
- **功能特點：**
  - 整合 Google 生態系（讀取你的雲端硬碟、郵件、地圖、YouTube）。
  - 具備較好的上下文對話管理。
  - 無法進行大規模批次處理或程式整合。

---

2. Gemini CLI (命令行工具)

這通常是透過開源工具（如 `gemini-cli`）或 Google 提供的 SDK 自行封裝的**開發者測試** 模式。

- **目標對象：** 工程師、系統管理員、喜歡自動化的 Geek。
- **介面：** 終端機 (Terminal)。通常底層呼叫的是 **Google AI Studio (Gemini API)**。
- **費用：** 根據 API 使用量計費。目前 Google AI Studio 提供相當優渥的**免費額度**（有頻率限制）。
- **資料隱私：** 中等。API 使用規範通常優於網頁版，但仍需視 API 條款而定。
- **功能特點：**
  - **腳本化：** 可以串接到 Shell 腳本中，批次處理本地檔案。
  - **效率：** 適合快速測試 Prompt 或在伺服器環境中進行簡單問答。
  - **侷限：** 適合單兵作戰，不適合部署為大規模商用服務。

---

3. Vertex AI 呼叫 API (Google Cloud)

這是 **B2B (企業級)** 模式，是 Google Cloud 平台的一部份。

- **目標對象：** 企業開發團隊、數據科學家、需要將 AI 整合進 App 的公司。
- **介面：** 程式碼 (Python, Node.js, Go, Java) 或 REST API。
- **費用：** **隨收隨付 (Pay-as-you-go)**。按 Token 數量計費，無免費版（僅有新帳號試用金）。
- **資料隱私：** **最高（企業級保護）。** Google 承諾**「您的數據不會被用來訓練其底層模型」**，且符合 SOC2, ISO 等合規認證。
- **功能特點：**
  - **完整生命週期管理：** 包含微調 (Tuning)、模型評估、自動化部署 (MLOps)。
  - **高可用性：** 提供服務水準協議 (SLA) 保障穩定性。
  - **整合性：** 輕易結合雲端資料庫（BigQuery）、搜尋引擎（Vertex AI Search）實踐 RAG。
  - **安全控管：** 支援 IAM 權限控管、VPC 網路安全限制。

綜合對照表

特性	Gemini 網頁版	Gemini CLI (AI Studio)	Vertex AI API
主要用途	日常對話、輔助辦公	開發測試、個人自動化	企業產品開發、商業化部署
操作難度	極低（會打字即可）	中等（需配置環境與 API Key）	高（需具備程式開發與雲端知識）
隱私安全性	低（資料可能被拿去訓練）	中（開發者條款保護）	**極高（企業級隱私，不訓練）**
自定義能力	低（僅能調整 System Instruction）	中（可調整參數如 Temperature）	**極高（可微調模型、串接 RAG）**
穩定性/SLA	無保障	無保障（可能隨時調整頻率限制）	**有 SLA 保障**
費用	免費 或 月費制	免費限額 或 隨量計費	**隨量計費（企業預算）**

我該選哪一個？

1. **如果你想寫作業、寫 Email、整理雜亂的筆記：** 請用 **Gemini 網頁版**。
2. **如果你是工程師，想寫個小程式自動整理本地 100 個文字檔：** 請用 **Gemini CLI (或 Google AI Studio API)**。
3. **如果你要開發一個手機 App 給萬名用戶使用，或公司內部要處理敏感個資：** 請務必使用 **Vertex AI**。

---

gemini 只是模型名稱，有可能改變？

所以未來gemini 被淘汰成 GGG 模型，名稱變了，但Vertex AI 仍叫Vertex AI ？

沒錯，這是軟體架構中「平台」與「內容物」的核心差異。

這可以從以下三個層次來理解為什麼 **Vertex AI** 的名稱會比 **Gemini** 更穩定：

1. 「貨架」與「商品」的關係

- **Vertex AI 是「貨架（平台）」：** 它是企業用來管理 AI 的一整套基礎設施。
- **Gemini 是「商品（模型）」：** 它是目前貨架上最強的主打產品。
- **未來：** 就算 Google 開發出下一代更強的模型叫 "GGG"，Google 只需要把 GGG 放上 Vertex AI 的「模型花園 (Model Garden)」即可。對於企業客戶來說，他們不需要更換整個平台，只需要把 API 呼叫的路徑從 `gemini-1.5-pro` 改成 `ggg-v1`。

2. 歷史已經證明過這點

Google 的 AI 模型名稱換得很快，但平台的名稱相對穩定：

- **以前的模型：** PaLM 2、LaMDA、BERT。
- **現在的模型：** Gemini (Pro, Flash, Ultra)。
- **平台演進：** 早期叫 Cloud AI Platform，後來整合並定名為 **Vertex AI** 後就一直沿用至今。

對企業來說，**「穩定性」**是第一考量。如果 Google 每換一個模型就換一次平台名稱，企業的工程師會瘋掉，因為所有的文件、權限設定和自動化流程都要重寫。

3. Vertex AI 裡不只有 Gemini

這也是為什麼 Vertex AI 不會輕易改名最重要的原因。在 Vertex AI 的「模型花園」裡，除了 Google 自家的 Gemini，還提供：

- **Llama** (Meta 的模型)
- **Claude** (Anthropic 的模型)
- **Mistral** (Mistral AI 的模型)
- **Stable Diffusion** (圖片生成模型)

因為 Vertex AI 是一個**「多模型平台」**，它的名字必須具備包容性。即使 Gemini 以後改名叫 GGG，或是 Google 決定不再開發自己的模型（假設），Vertex AI 依然可以作為一個託管其他公司模型的平台而存在。

總結

- **Gemini** 是品牌行銷與 AI 技術的名稱（容易隨技術迭代而改變）。
- **Vertex AI** 是雲端服務與企業解決方案的名稱（為了維護企業信任，通常會保持長期一致）。

你的直覺很對：
如果五年後 Google 推出了 **"GGG"** 模型，你大概率還是會進到 **Vertex AI** 的控制台，點選 **"GGG"** 來使用它。

【google gen ai SDK】 與 【 vertext ai SDK】差異

簡單來說：**Google Gen AI SDK 是給「個人/開發者快速原型」用的；而 Vertex AI SDK 是給「企業/正式生產環境」用的。**

以下是詳細的差異分析：

---

1. 背後運作的平台不同

- **Google Generative AI SDK (`google-generativeai`):**
  - 對接的是 **Google AI Studio**。
  - 這是一個輕量級的開發工具，讓你可以快速拿到 API Key 並開始測試模型。
- **Vertex AI SDK (`google-cloud-aiplatform`):**
  - 對接的是 **Google Cloud Platform (GCP)**。
  - 這是一個企業級的機器學習平台，整合了所有的雲端資源。

2. 身份驗證方式 (Authentication)

- **Google Gen AI SDK:**
  - 使用 **API Key**。

    - 非常簡單，就像你用 OpenWeather 或其他小工具一樣，貼上一行 Key 就能動。
- **Vertex AI SDK:**
    - 使用 **IAM (Identity and Access Management)**、**服務帳戶 (Service Account)** 或 **OAuth2**。
    - 你需要安裝 Google Cloud CLI (`gcloud`)，設定專案 ID，處理權限認證。雖然麻煩，但對於大型團隊來說，這才安全。

3. 資料隱私與安全

- **Google Gen AI SDK (AI Studio):**
    - 如果你使用的是**免費版**，Google 擁有在去識別化後，使用你的輸入/輸出來改善模型的權利。
    - 適合測試、Demo、不具敏感性的個人專案。
- **Vertex AI SDK:**
    - **預設不會將你的資料用於訓練模型。**
    - 提供企業級的加密、VPC 網路安全限制、以及合規性認證（如 HIPAA, GDPR）。
    - 適合處理公司客戶資料、機密文件。

4. 功能廣度

- **Google Gen AI SDK:**
    - 專注在「生產生 AI」本身（聊天、生成圖片、Embedding）。
    - 功能比較單一、純粹。
- **Vertex AI SDK:**
    - 除了生成 AI，它還包含整個 **MLOps 流程**。
    - 你可以進行模型監控（看有沒有胡說八道）、模型比對、自動化流水線 (Pipelines)、部署到專屬端點 (Endpoint)。
    - 支援更多模型（Llama, Claude 等等）。

5. 配額與穩定性 (Quota & SLA)

- **Google Gen AI SDK:**
    - 免費額度很高，但**每分鐘請求數 (RPM) 限制較嚴格**。
    - 不保證服務水準 (No SLA)。
- **Vertex AI SDK:**
    - 付費使用，配額可以根據企業需求申請調高。
    - 有 **SLA (Service Level Agreement)** 保證，如果系統掛了，Google 要賠錢或提供補償，這對商業營運至關重要。

快速比較表

特性	Google Gen AI SDK	Vertex AI SDK
主要平台	Google AI Studio	Google Cloud (Vertex AI)
驗證方式	API Key (簡單)	IAM / Service Account (嚴謹)
上手速度	極快 (1 分鐘)	較慢 (需設定 GCP 專案)
資料隱私	免費版可能被用於訓練	**絕對不會被用於訓練**
適用場景	個人作品、快速原型、Side Project	企業級應用、商業產品、處理個資
庫名稱 (Python)	`google-generativeai`	`google-cloud-aiplatform`

我該如何選擇？

- **選 Google Gen AI SDK，如果：**
    - 你是學生或獨立開發者。
    - 你想在 5 分鐘內寫出一個會說話的機器人 Demo。
    - 你預算有限，想利用免費額度。
- **選 Vertex AI SDK，如果：**
    - 你正在為公司開發產品。
    - 你的 App 每天會有成千上萬的人使用（需要高併發）。
    - 你需要處理公司內部的敏感文件或客戶隱私。
    - 你需要把 AI 跟 BigQuery、Cloud Storage 等 Google 雲端服務串聯。

一句話總結：
Google Gen AI SDK 是進入 Gemini 世界的**快速捷徑**；而 **Vertex AI SDK** 是在 Google 雲端蓋大樓的**標準地基**。

單獨使用 Gemini API 與 Vertex AI 的差異

下表摘要列出 Gemini API 和 Vertex AI 的主要差異，協助您決定哪個選項適合您的用途：

https://docs.cloud.google.com/vertex-ai/generative-ai/docs/migrate/migrate-google-ai?hl=zh-tw

功能	Gemini API	Vertex AI
端點名稱	`generativelanguage.googleapis.com`	`aiplatform.googleapis.com`
註冊	Google 帳戶	Google Cloud 帳戶 (須同意條款並提供帳單資訊)
驗證	API 金鑰或 OAuth (如已連結至 Google Cloud 專案)	Google Cloud 服務帳戶
使用者介面遊樂場	Google AI Studio	Vertex AI Studio
API 和 SDK	伺服器和行動/網頁用戶端 SDK • 伺服器：Python、Node.js、Go、Dart、ABAP • 行動/網路用戶端 (透過 Firebase AI Logic)：Android (Kotlin/Java)、Swift、網路、Flutter 和 Unity	伺服器和行動/網頁用戶端 SDK • 伺服器：Python、Node.js、Go、Java、ABAP • 行動/網路用戶端 (透過 Firebase AI Logic)：Android (Kotlin/Java)、Swift、網路、Flutter 和 Unity
免費使用 API 和 SDK	是，視情況而定	新使用者可享 $300 美元的抵免額 Google Cloud
配額 (每分鐘要求數)	視型號和定價方案而定 (請參閱詳細資訊)	因型號和地區而異 (請參閱詳細資訊)
企業支援	否	• 客戶加密金鑰 • 虛擬私有雲 • 資料落地 • 資料存取透明化控管機制 • 可擴充的應用程式託管基礎架構 • 資料庫和資料儲存空間
MLOps	否	Vertex AI 上的完整 MLOps (例如：模型評估、模型監控、模型登錄)

修訂版本 #4
★由 treeman 建立於 22 ?Q?G?? 2025 14:48:54
由 treeman 更新於 23 ?Q?G?? 2025 11:45:51