Google Project Astra 和 OpenAI 的 GPT-4o 接連登場 多模態 AI 助理時代真的來臨了！

近期，OpenAI 推出了 GPT-4o，而 Google 也不甘示弱，推出了其最新版本的多模態 AI 助理 Project Astra。這兩個重量級計劃的登場，無疑預示著多模態 AI 助理時代的到來，這將徹底改變我們的生活、工作以及學習方式。

Google Project Astra 革新之舉

Google Project Astra 是 Google 的一項創新之舉，它結合了語音、圖像、文本等多元模態，塑造出一個全方位的智能助理。這項技術能通過攝影鏡頭捕捉周遭的世界，並由人工智慧即時、持續地描述畫面細節。例如，它能描述擴音器的聲音發出部位，甚至指出被標記的喇叭部分是 Tweeter 高音單體。同時，它還能創意地組合蠟筆相關的詞彙、識別程式編碼內容，甚至從白板上畫的貓與紙箱聯想到「薛丁格的貓」等。

除此之外，當影片中的使用者詢問眼鏡的位置時，人工智慧能迅速給出答案，展示了其優秀的資訊記憶能力。當使用者從手機切換到配備攝影鏡頭的智能眼鏡時，Project Astra 依然能無縫遷移資料，讓使用者在不同裝置上繼續享受服務。這種跨模態的處理能力，將 Astra 在智能化互動的實現上推向了全新的高度。

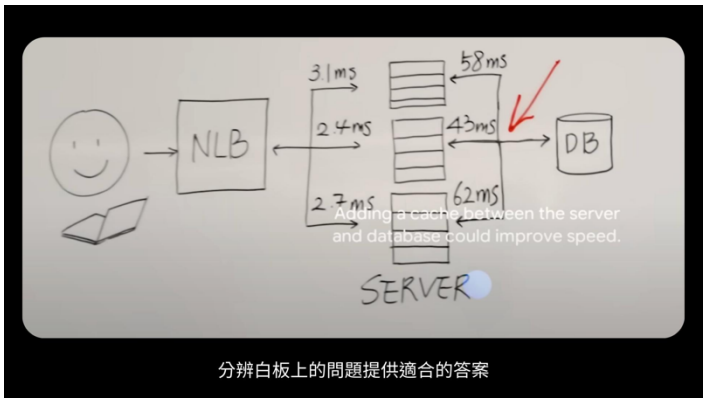
Google 同時也推出全新的 Gemini 1.5 pro，不僅能處理更多數據，也增強了該模型編寫程式碼、推理以及解析音訊和圖像的能力。另外 Google 正準備推出一個名為「Gemini 1.5 Flash」的新模型。Google DeepMind 的 CEO Demis Hassabis 表示，Gemini 1.5 Flash 在摘要生成、聊天、圖像和影片字幕生成、以及由長文件和表格中提取資料等方面更勝一籌。

此外，Google 發布 PaliGemma，首個 Gemma 視覺語言多模態開放模型，具有處理圖像和文字並輸出文字的能力。這些模型分為預訓練（pt）、混合和微調（ft）三種類型，每種均提供不同分辨率和精度選擇。PaliGemma 由 SigLIP-So400m 圖像編碼器和 Gemma-2B 文字解碼器組成，專門設計用於理解圖像和文字的聯合訓練，並可以針對各種下游任務進行微調，以實現精準的多模態互動。

Google Project Astra 和 OpenAI 的 GPT-4o 的接連登場，標誌著多模態 AI 助理時代的正式來臨。這些技術的發展，將深刻改變我們的生活方式，讓人機交互變得更加智能和自然。我們有理由期待，在不久的將來，多模態 AI 助理將成為我們生活中的重要夥伴，為我們帶來更多便利和可能性。

利用文字辨識方式將眼前螢幕的程式碼轉化為文字後分辨使用的程式語言

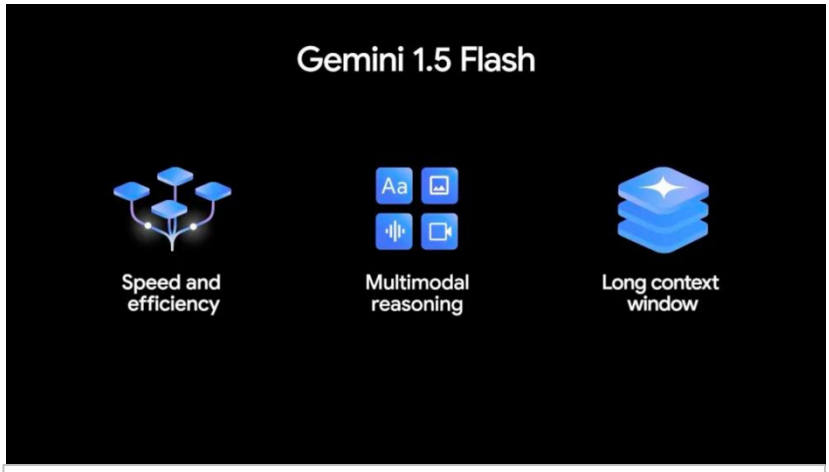
幫助用戶找到自己遺落(但Project Astra眼角餘光曾掃到)的眼鏡

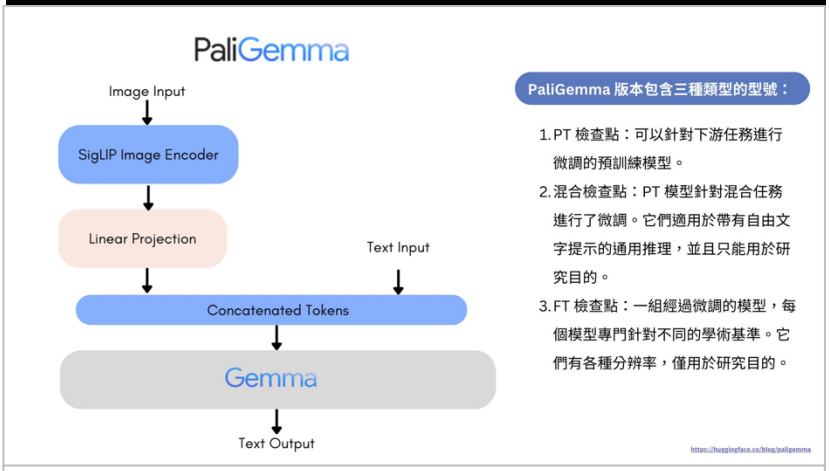


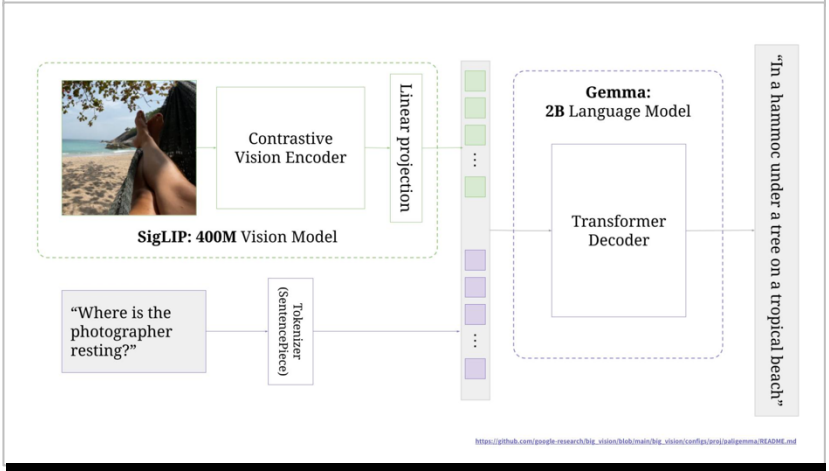
分辨白板上的問題提供適合的答案

發揮創意幫眼前的景象下註解

參考資料
https://huggingface.co/blog/paligemma
https://github.com/...../configs/proj/paligemma/README.md







PaliGemma 模型

PaliGemma-3B 是一種視覺語言模型，其設計靈感受到 PaLI-3 啟發 。這個模型的主要組件包括 SigLIP 視覺編碼器，例如 SigLIP-So400m/14 和 Gemma 2B 語言模型。

輸入處理： PaliGemma 可以接收一幅或多幅圖像作為輸入。這些圖像首先由 SigLIP 視覺編碼器處理，將圖像轉換成一種稱為「軟標記」的中間表示形式。這種轉換過程是將圖像的視覺信息編碼成一種形式，讓模型能夠理解和處理。

文本處理： 同時，輸入的文本（前綴）會由 Gemma 的標記產生器進行標記。這意味著文本會被處理成模型可以理解的格式。

資料整合與處理： 接下來，圖像的軟標記和文本的前綴標記會按照一定順序被連接在一起，然後傳遞給 Gemma 解碼器。Gemma 解碼器使用完整區塊注意力（full block attention）來處理這些連接的資料。

輸出生成： 解碼器利用屏蔽注意力（masked attention）機制進行自動回歸，生成輸出文本（後綴）。這一步驟意味著解碼器會自動地根據已處理的圖像和文本資料，生成相應的輸出文本。