推理時間比 Meta, Mistral 模型【快】兩倍

- 由Mistral-7B Finetune 而來的MR Breeze-7B是專為繁體中文的用戶所設計，以模型參數量小為特色，在繁體中文與英文提供出色的表達能力。
- 此外，MR Breeze-7B的推論速度非常快，且在表格處理以及用戶對話的表現能力也很優秀。

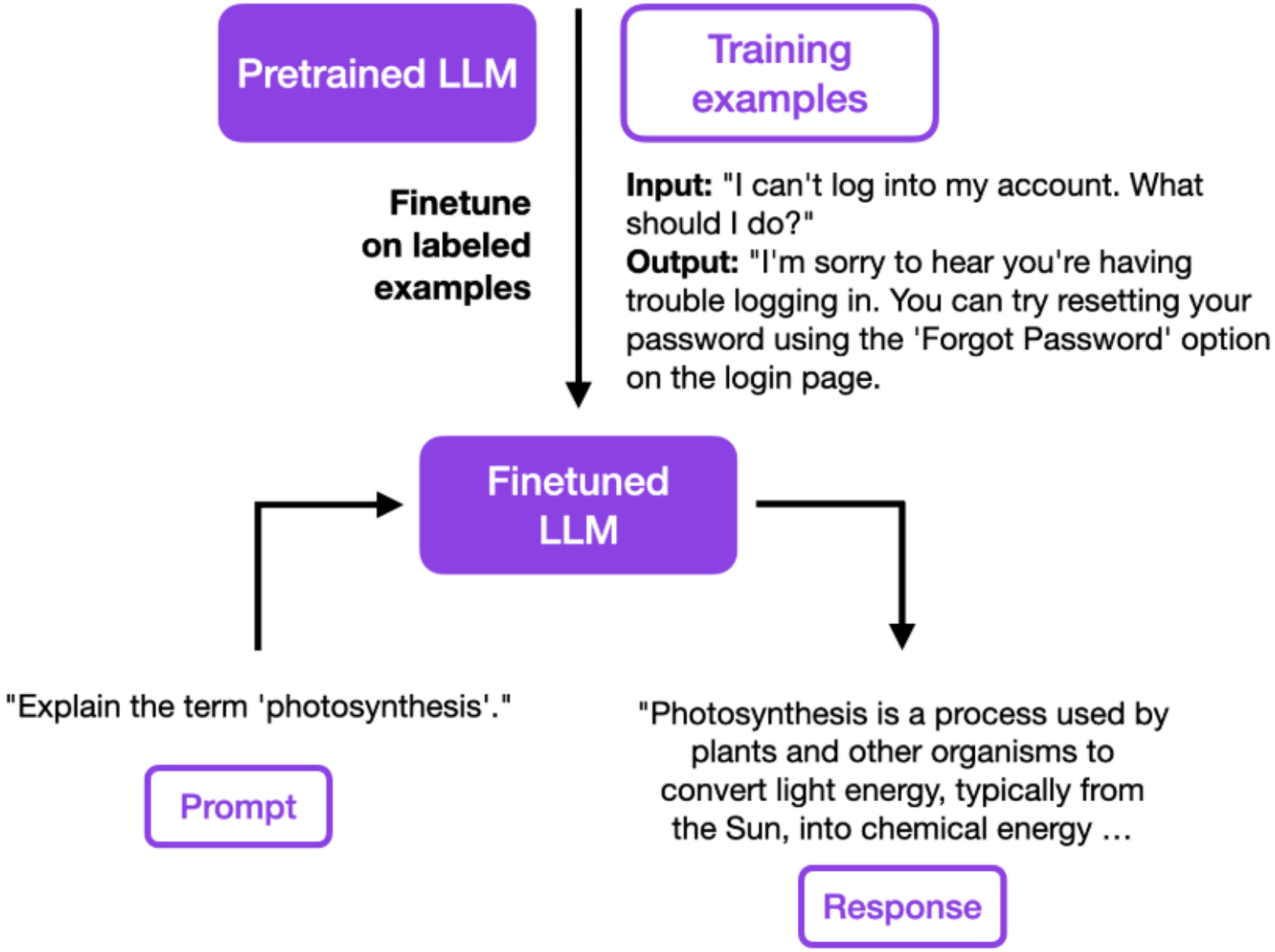
Mistral 7B在當時(2023年9月27日)是最好的70億參數規模的大模型。

法國的“OpenAI”，總統都按讚，但難以營利

Mistral AI 將獲得OpenAI 同等待遇，直接將模式資源放在微軟雲端當中販售，成為第二家在微軟Azure 雲端平台上提供商業AI 模式的公司。

Mistral AI的成立背景

- Mistral AI是一家法國人工智慧新創企業，Mistral AI於2023年初於法國巴黎成立，是由三名曾在美國AI領域公司工作的研究人員共同創立，包括曾在Google的DeepMind工作的公司CEO Arthur Mensch，曾在Meta（Facebook的母公司）工作參與了LLaMA的開發的公司CTO Timothée Lacroix和公司CSO Guillaume Lample。
- 三位共同創辦人從學生時代就互相認識，在他們看到生成式人工智慧領域的快速發展後，意識到歐洲也需要有自己的生成式人工智慧，於是在法國創立Mistral AI。


Pretrained LLM
Training examples
Finetune on labeled examples
Input: "I can't log into my account. What should I do?"
Output: "I'm sorry to hear you're having trouble logging in. You can try resetting your password using the 'Forgot Password' option on the login page.
Finetuned LLM
"Explain the term 'photosynthesis'."
Prompt
"Photosynthesis is a process used by plants and other organisms to convert light energy, typically from the Sun, into chemical energy …
Response

•Breeze-7B-Base expands the original vocabulary with an additional 30,000 Traditional Chinese tokens. With the expanded vocabulary, and everything else being equal, Breeze-7B operates at twice the inference speed for Traditional Chinese to Mistral-7B and Llama 7B.

•Breeze-7B-Instruct can be used as is for common tasks such as Q&A, RAG, multi-round chat, and summarization.

聯發科推生成式AI服務平臺和450億參數繁中大語言模型BreeXe

BreeXe是在輝達的Taipei -1 GPU超級電腦DGX H100上進行訓練

A project by the members (in alphabetical order): Chan-Jan Hsu 許湛然, Chang-Le Liu 劉昶樂, Feng-Ting Liao 廖峰挺, Po-Chun Hsu 許博竣, Yi-Chang Chen 陳宜昌, and the supervisor Da-Shan Shiu 許大山.

Llama 3.2 for EdgeAI

- 聯發科技與 Meta在新發布的 Llama 3.2 量化模型（包括 1B 與 3B 版本）上持續合作，該模型將由聯發科技 **#天璣9400旗艦晶片** 率先支援，並逐步部署至其它天璣系列行動晶片上。
- 相較於使用浮點數據的標準模型，Llama 3.2 量化模型（1B/3B）採用整數數據，將模型大小縮減60%，有效節省存儲空間及內存數據頻寬，實現了準確性、性能和內存佔用之間的平衡。同時，此量化模型還可帶來超過 2 倍的效能提升，並展現了卓越的能效優勢。
- 對於開發者而言，Llama 3.2 量化模型（1B/3B）不僅能夠輕鬆被整合到應用程式中，大幅節省開發時間及成本，還可以幫助開發者實現與大模型相似的諸如文本摘要、自然語言聊天機器人、工具應用及檢索增強生成（RAG）等功能，僅在裝置端就能完成資料處理，有效保護使用者隱私。

聯發科技豐富的生成式 AI 應用場景

軟體開發	GAI 應用	
需求分析與規格設計	分析需求規格	分析技術文件,加速需求分析與整理
	生成技術文件	產生技術文件，減少文件寫作時間.
	編修校對文件	提供客戶前的文件品質審查
程式編碼	編寫程式代碼	協助程式代碼編寫
	修復程式問題	協助 bug 偵測與修復
	審查程式代碼	協助 code review 需求
測試與驗證	編寫測試程式	自動生成測項, 不須人工再進行修改
	分析 CR 問題	主動分析客戶提問問題
	精簡測試案例	自動生成測試案例, 縮短測試時間
	提升測試涵蓋率	擴大白箱測試涵蓋率
客戶服務	推薦客戶 FAQ	主動推薦 FAQ 給客戶加速解決問題
	回答客戶問題	自動客服機器人

功能領域	GAI 應用	
IT	自動會議記錄	自動執行語音轉檔，彙整會議重點與代辦事項
人資	招募小幫手	履歷篩選 & 自動匹配功能
財務	付款作業優化	審核非結構化內容, 流程自動判斷
	信用報告撰寫	根據外部資料庫, 自動撰寫報告
	報銷流程	自動辨識發票
法務	專利翻譯	專利撰寫過程中加速翻譯流程
	合約修改	提供修改建議，加快合約審閱
生產	外包良率分析	針對外包生產良率快速檢索分析
銷售	業務小幫手	即時查詢客戶訂單、出貨等資訊
稽核	稽核小幫手	針對潛在的資安疑慮進行主動稽核


晶片驅動臺灣產業創新方案
2024 GenAI 產業高峰論壇
AI生成·無限可能
09:00-09:30 來賓報到
09:30-09:35 主持人開場
王志仁 / 數位時代總編輯
09:35-09:40 貴賓致詞
09:40-09:50 共創台灣GenAI新世代
劉佩玲 / 台大智活中心 主任
09:50-10:20 AI時代·台灣機會
施立峰 / 前Google台灣 董事總經理
10:20-10:40 擁抱生成式 AI：創新應用的機會與挑戰
游直暐 / Appier 執行長暨共同創辦人
10:40-11:00 生成式AI創新與提升生產力的新途徑
葉家順 / 聯發科 人工智慧暨數據工程 協理
11:00-11:20 中場休息
11:20-11:35 百工百業用AI：2024 GenAI學習地圖
蔡明順 / 台灣人工智慧學校 校務長
11:35-11:50 台灣 in！打造屬於我們的 ChatGPT
李育杰 / 中央研究院資通安全專題中心 執行長暨 TAIDE計畫主持人
11:50-12:05 用AI製程讓製造智能
龔文郁 / 台灣大學機械工程學系 終身特聘教授
12:05-12:10 閉幕結語