元智大學資管系

第三十屆專題製作競賽報告

研究主題：

基於多模態模型微調技術與白血球影像分類之研究

姓名：邱芯彤、黃彥婷

學號：1111635、1112206

公司代號：ZI2

實習公司：元智大學資訊管理學系

指導教授：謝瑞建 副教授

中華民國 114 年 11 月

Nov, 2025

Abstract

本研究旨在提升大型語言模型於白血球影像判讀任務中的應用表現，以協助醫療資源相對不足地區進行更有效的輔助診斷。現今大型語言模型(Large Language Models, LLMs)雖具備強大的生成與理解能力，但在面對專業領域問題時，常因缺乏領域知識而產生語意模糊或內容錯誤的回答，為克服此問題，本研究結合多模態視覺語言模型 LLaMA 3.2 Vision 與 Unsloth 架構，透過輕量化參數微調(LoRA)與低位元量化(4-bit Quantization)技術，強化模型對白血球影像的分類能力與文字描述生成能力。

研究資料包含六類白血球影像，配合多風格描述語句與標籤設計，進行模型微調與圖文對齊訓練。經分類績效指標(Precision、Recall、F1-score)、混淆矩陣與平均值 ± 標準差分析驗證，微調後模型於各細胞類別皆達高分類精度，整體 F1-score 均值達 0.99 並具穩定一致性。

本研究不僅展示微調後模型可有效進行影像分類與語意生成，亦可結合醫療人員專業判斷，提供具有參考價值之診斷建議，未來有望應用於偏遠地區或醫療量能不足之場域，作為輔助檢測與教學工具。

關鍵字：多模態模型、微調、白血球分類

Contents

Chapter 1　緒論

在科技迅速發展的現代，發展出了各式各樣的技術，如生成式 AI、影像辨識、聲音模擬等等，並廣泛運用於各個領域。在 ChatGPT 發布於 2022 年之後，在人工智慧領域掀起了熱潮，成為了人們關注的焦點，其展現出在自然語言處理技術上的潛力。

不過，隨著應用的深入，人們也逐漸發現大語言模型並非萬能，其在落地應用上面臨諸多挑戰，其能力邊界也逐漸成為研究的重要課題，在一些需要專業知識的領域，模型雖然能產生相關文本，但可能缺乏深入的理解，例如，在醫學領域，雖然能提供一些疾病的基本信息，但對於複雜疾病的診斷和治療建議可能不準確[1]。因此，我們以微調 Unsloth 的 LLaMa 3.2 Vision 模型作為研究核心，期望透過專業領域資料的訓練，讓模型在醫療應用中具備更準確且深入的理解能力。

本研究目的是提升模型在白血球影像判讀的能力，進而協助專業人力資源相對不足的地區醫療機構，微調後的大型語言模型不僅能對白血球影像進行分類還能生成對應的文字描述，並與醫療人員的專業判斷結合，提供更完整且具參考價值的輔助診斷建議。然而，在運用過程中還是會面臨一些限制與挑戰，如：雖然模型能生成診斷描述，但其內容仍可能出現判斷錯誤的情形，無法完全取代醫療專業人員的臨床判斷。

Chapter 2 相關技術與研究

2.1 與本研究相關技術一

2.1.1 大型語言模型(LLM) 與 Transformer

大型語言模型(LLMs)已成為人工智慧領域中一股變革性的力量，重新塑造了人們與機器學習模型互動與使用的方式，其在多種領域中展現出卓越的能力，包含文字生成、翻譯、問答、摘要等，使其成為現代人工智慧系統的核心[2]，而這些成果其實都跟 Transformer 架構脫不了關係，此模型的設計讓系統能夠更有效地理解語句中詞語之間的關係，不只提升了處理效率也讓模型變得更聰明、更靈活。

Transformer 是一種基於注意力機制(Attention-Based Model)的模型，用來處理像機器翻譯這類序列到序列(Sequence-to-Sequence)的任務，它放棄了傳統的 RNN 結構，改採編碼器(Encoder)與解碼器(Decoder)的設計：編碼器負責讀懂輸入句子，解碼器則根據這些資訊生成對應的輸出。在每一層中，Transformer 都運用了自注意力機制(Self-Attention)，讓每個詞可以同時考慮句子中其他詞的關聯性，並搭配前饋神經網路(FNN)來加強語意的轉換能力，透過多頭注意力(Multi-Head Attention)的設計，模型能從不同角度理解語句，增強學習效果，因為它不具備處理詞語順序的能力，Transformer 也引入了位置編碼(Positional Encoding)，讓模型能理解詞語的排列順序。此外，每個子層都使用殘差連接(Residual Connection)和層正規化(Layer Normalization)來穩定訓練[3]，這讓整個模型更有效率，也更容易學習。

2.1.2 Pytorch

在說明了 Transformer 架構之後，進一步探討這類模型在實際應用中所依賴

的開發工具與平台，Transformer 本質上是一種神經網路架構，其部署與訓練仍需仰賴成熟且高效的深度學習框架，其中 PyTorch 是目前廣泛使用的選擇之一，無論在研究領域或產業應用中皆扮演關鍵角色。

PyTorch 由 Meta(原 Facebook)所開發，其設計強調動態計算圖(Dynamic Computation Graph)、高度模組化的架構以及與 Python 語言的高度整合，使得模型的建構、調整與實驗流程更加靈活。根據 Paszke 等人[4]的說明，PyTorch 採用命令式程式設計的風格，提供強大的 GPU 加速能力，特別適合於原型設計和研究應用。這些特性使 PyTorch 在自然語言處理任務相當合適並迅速成為 BERT、GPT、LLaMA 等，主流大型語言模型的核心開發與訓練平台，從模型設計到大規模部署，PyTorch 提供了完整的支援也成為推動大型語言模型發展的重要基礎設施。

2.1.3 多模態模型 (Multi-Modal)

隨著大型語言模型在文本理解、生成上的優秀表現，研究群體逐漸將注意力擴展至多模態學習(Multimodal Learning)的領域，多模態模型的核心目標在於整合來自不同感知的資訊，例如：文字、影像、音訊、影片等，使模型能夠更完整的理解與推理，這類型的模型提升了 AI 在複雜領域中的適應能力，也開啟了跨領域應用的新契機，如圖文生成、語音辨識、醫學影像輔助診斷等等。

2.2 與本研究相關技術二

2.2.1 量化(Quantization) 與 Unsloth

現代大型語言模型規模極大，往往擁有數十億的參數，例如：LLaMa3.2 Vision 11B 模型，然而一般使用者沒有高階的顯卡能運用在大語言模型的微調上，為了降低訓練與推論過程中的計算成本，模型量化(Model Quantization) 成為近年來的

重要研究方向之一。量化的核心概念是將模型中原本使用的 32 位元浮點數(FP32)權重，轉換為更低位元數的表示(如：INT8 或 FP16)，藉此大幅減少記憶體用量與加速運算，根據 Nagel 等人(2021)指出，使用量化技術可以在推論效率與模型精度之間取得良好平衡，有助於模型在實際應用中順利運作與部署[5]。

常見的量化方式可以分為兩種:訓練後量化(Post-Training Quantization, PTQ)與量化感知訓練(Quantization-Aware Training, QAT)兩種方法。 PTQ 是將模型訓練完再進行壓縮處理，適合快速部署；而 QAT 則是在訓練階段即模擬量化的影響，可進一步提升壓縮後的模型精度。在設備資源有限或是無法使用高效能 GPU 的情況下，透過量化技術能有效減少硬體需求，是實現大型語言模型微調與推論的關鍵技術之一。

此外，Unsloth 是一個專為大型語言模型微調而設計的高效開源框架。它結合了多種先進技術，顯著加速微調過程，同時減少記憶體使用量，並保持模型的準確性[6]。Unsloth 支援開源的量化模型，特別是 QLoRA(Quantized Low-Rank Adaptation)，這是一種將模型參數壓縮為小位元(如：4-bit)的技術，非常適合資源有限的開發環境。

2.2.2 微調(Finetuning)、Lora、Adapter

大型語言模型的廣泛運用，然而這些以大規模、通用性資料進行預訓練的大型語言模型，在面對特定專業領域(如：醫療、法律等)時，缺乏足夠的背景知識與上下文理解能力，導致回答不夠精確或出現錯誤資訊，解決這個問題常見的方法是對模型進行微調。微調是透過一組特定領域的資料對已經訓練好的預訓練模型(Pretrained Model)進行額外訓練，使模型能夠適應特定領域的資料與語境。這是一種遷移學習(Transfer Learning)策略，通常會將整個模型的參數作為起始點

(Initialization)，並允許其進一步更新權重。

傳統的全參數微調(Full Fine-tuning)需要更新整個模型的參數，對於參數量數十億的大型語言模型而言，訓練成本過高。參數高效微調(PEFT)只需更新模型中的少部分參數，或是引入少量模組，在保持大部分預訓練參數凍結的情況下有效學習特定領域資料。這類方法在節省成本的同時，依然能在專業任務中展現優異的表現。

其中，LoRA(Low-Rank Adaptation)是由 Hu et al.等人(2021)提出[7]的一種代表性 PEFT 方法，透過在模型的權重矩陣上插入低秩分解(Low-Rank Decomposition)模組，僅訓練這些小模組，原始權重保持凍結，這種方式在保持推論效率與效能的前提下，降低微調所需的參數量與 GPU 記憶體使用。

另一種常見的 PEFT 方法 Adapter 由 Houlsby et al 等人(2019)提出[8]，在 Transformer 各層之間插入小型神經網路模組，通常由一個下投影(Down-Projection)與一個上投影(Up-Projection)組成，僅需調整這些插入的模組，原始模型權重完全凍結。此技術被應用於多領域與多語言微調場景。

Chapter 3 研究方法

本研究採用衛生福利部桃園醫院血液鏡檢科所提供的白血球影像資料及各類白血球描述作為微調素材，並對 Unsloth 的 LLaMa 3.2 Vision 11B 模型進行微調，以提升模型對於血液塗片中白血球類型的識別與解讀能力。

0.實驗模型

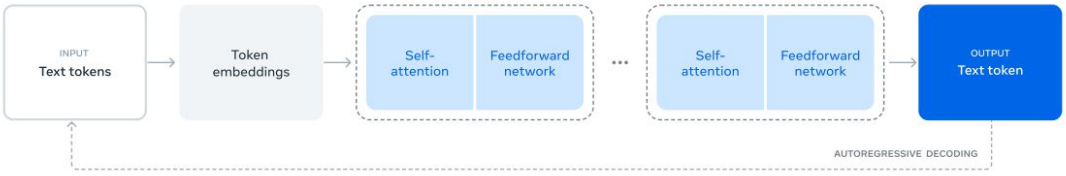
(1) **LLaMa**

LLaMA(Large Language Model Meta AI)是由 Meta AI 提出的一系列基礎語言模型，旨在透過更高效的訓練方法與設計，使中小規模的模型也能達到與現有大型語言模型相當的作用。該模型架構採用典型的 Transformer 解碼器結構(Decoder-Only Architecture)，專注於自回歸(Regression)式語言建模任務，即模型依序預測輸入序列中的下一個詞元，與過去的大型模型不同，LLaMA 在保持較小參數規模的同時，透過一系列的架構優化顯著提升了效能。[9]

在模型設計方面，LLaMA 採用了預正規化(Pre-Normalization)策略，在每個 Transformer 子層之前施加 RMSNorm，此種替代 LayerNorm 的技術可有效穩定訓練過程並降低運算成本，而前饋神經網路(FNN)中則採用 SwiGLU(Swish-Gated Linear Unit)作為激活函數(Activation Function)以強化非線性表達能力，進而提升整體模型表現，為了加強對長距離依賴(Long-Range Dependencies)關係的建模，LLaMA 採用旋轉位置編碼(Rotary Positional Embedding, RoPE)，相較於傳統絕對位置編碼，可更有效捕捉詞元間的相對位置資訊，特別是在處理長序列輸入時表現更為穩定。[9]

(2) LLaMa 3.2 Vision

LLaMA 3.2 VisionMeta 基於 LLaMA 3.1 架構所擴展的多模態大型語言模型，能同時處理圖像與文字輸入，其核心設計在於結合視覺編碼器與語言模型，實現圖文對齊而不改動語言模型本身的參數，圖片首先利用視覺編碼器提取局部與全局特徵，這些多層次的視覺資訊再經由交叉注意力機制嵌入至語言模型之中。語言模型本體則保留原有結構，共包含 40 層，其中每隔數層插入一個交叉注意力層以融合圖像語意，整體模型具備 128K Token 的長文本處理能力，並使用分組查詢注意力(GQA)機制提升推理效率，訓練方面只微調視覺模組與融合層，語言部分保持凍結，確保語言能力穩定並有效整合視覺訊息，此設計讓模型能廣泛應用於圖像描述、視覺問答等多模態任務中[10]。圖一為 LLaMa 3. 整體架構與訓練流程示意圖[11]。



圖一、LLaMa 3. 整體架構與訓練流程示意圖[11]

(3) Unsloth LLaMa 3.2 11B Vision

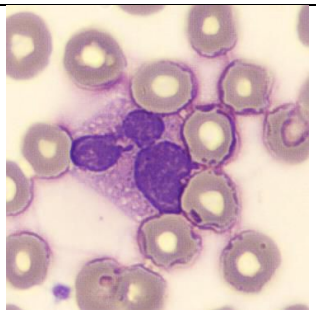
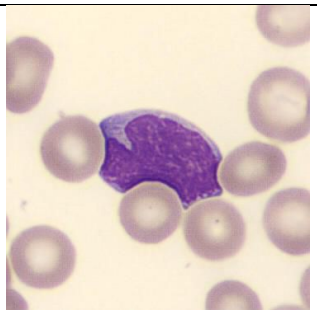
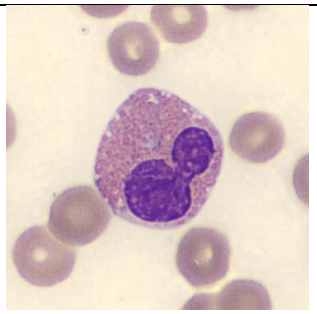
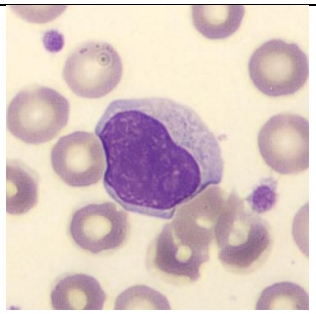
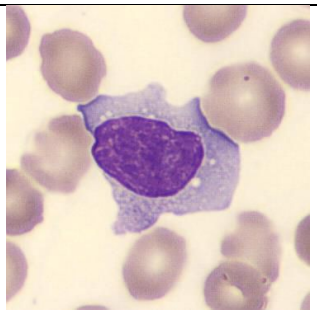
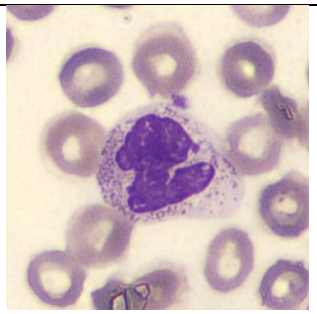
Unsloth LLaMA 3.2 Vision 是一種多模態語言模型，其建立於 Meta 的 LLaMA 3.2 11B Vision 架構之上，整合了圖像與文字處理能力並對圖文對齊、視覺問答與描述任務進行優化，模型由視覺編碼器與語言模型兩部分構成，圖片先經過 32 層局部 Transformer 和 8 層全局視覺整合層提取多層次特徵，然後透過交叉注意力注入至語言模型中，語言模型包含 40 層 Transformer，其

中 8 層為交叉注意力層，用以融合圖像語意，訓練策略採用 LoRA 微調與 4-bit 量化技術，降低資源需求、提升速度[12]。

1.資料集

該資料集共包含 48,557 筆白血球影像,依照細胞類型分為六大類:Basophil 、Blast Cell、Eosinophil、Lymphocyte、Monocyte 以及 Neutrophil，每張影像皆為 RGB(全彩)影像，然後根據用途將資料集分為訓練集 45,942 筆、驗證集 1,615 筆、測試集 1,000 筆，接著我們針對這些影像進行資料前處理以利後續模型的訓練與評估。

➤ **資料類型**

		
BASO (Basophil) 嗜鹼性球細胞	BLAST (Blast Cell) 芽母細胞	EO (Eosinophil) 嗜酸性球細胞
		
LYM(Lymphocyte) 淋巴細胞	MONO(Monocyte) 單核球細胞	SEG(Neutrophil) 嗜中性球細胞

➤ 資料前處理

本次的資料前處理流程，我們將原始影像轉換為適合多模態模型輸入的格式，並結合標準化影像、語意描述與對應標籤，建立圖文對齊的訓練、驗證、測試用資料。

(1) 標準化圖片

在資料前處理階段，首先我們先對影像進行了統一尺寸調整，將所有影像縮放為 224*224 的大小以符合模型輸入的尺寸要求，接下來參考其模型的預訓練設定對影像進行標準化(Standardization)處理，使色彩通道的分布更接近模型訓練時的資料分布，有助於提升學習效果與穩定性，處理完成後，影像經反標準化(Denormalization)還原為可視格式並儲存起來並建立一個模型可讀的檔案格式放入每張影像的路徑。

(2) 描述

接下來在每張影像對應的路徑後方新增一段文字描述作為模型的文字輸入，用以強化圖文對齊的能力，每一類細胞分別搭配五種不同風格但主要內容相同的文字描述，藉此提升模型在不同語境下的理解與泛化能力(Generalization Ability)。

(3) Label(標籤)

最後，在每張影像的描述後方增加一個標籤，以數字方式表示其所屬的細胞類別，例如：Basophil 為 0、 Eosinophil 為 1，以此類推，而這些數字標籤能夠幫助我們根據影像與對應描述判斷其分類是否正確。

2.微調 Unsloth LLaMa 3.2 Vision 11B 模型

在微調過程中，我們針對模型的多模態輸入進行調整，使其能有效整合影像特徵與文字描述，進而強化其醫學影像判讀的準確性與語意理解能力，使用 LLaMa-3.2-11B-Vision-unsloth-bnb-4bit 模型進行微調[13]。

(1) 設定設備、模型及 LoRa 微調參數

本次訓練使 RTX 3060 進行微調，LLaMa3.2 Vision 11B 模型是透過 Unsloth 工具從 Hugging Face 載入，支援 4bit 量化，此模型為多模態模型支援圖像及文字。最後設定 LoRa 微調參數，為模型加入可微調的參數模組，不需要微調整個大模型就能完成訓練，節省記憶體與計算成本。

(2) 資料集載入及轉換

載入資料的 CSV 檔，檔案包含圖片路徑與對應描述及標籤，開啟影像並將其轉換為適合 LLaMa 的對話格式，為使模型能夠正確對齊圖文資料，訓練資料會轉換為包含 Image 與 Text 的多輪對話格式(Chat Format)，這是 SFT 訓練所需格式；同時也檢查圖片路徑與格式，排除無效或損壞的資料。

(3) 建立訓練器與訓練參數設定

使用 trl 套件的 SFTTrainer 建立一個支援視覺模型的訓練器。設定訓練參數，例如：批次大小(Batch Size)、訓練週期(Epochs)、優化器與學習率(Optimizer&LR)、模型儲存策略(每個 Epoch 儲存一次，保留最佳模型)。此外，使用 Unsloth 提供的專屬資料整理器(UnslothVisionDataCollator)，確保圖像與文字可正確送入模型，並且依照硬體支援自動選擇使用 Fp16 或 Bf16 進行混合精度訓練。

(4) 執行模型訓練

根據前面準備的訓練資料與設定，開始微調模型。這裡會自動執行 SFTTrainer 中的前向傳播(Forward Propagation)、計算損失(Loss)、反向傳播(Backward Propagation)、權重更新，訓練完一次執行一次驗證，模型儲存與紀錄。訓練過程中的損失會記錄下來，後續可視覺化繪製 Loss 曲線圖，觀察模型收斂情形。

3.生成描述

首先，我們將模型透過 Unsloth 提供的 FastVisionModel 架構載入，並用先前以所訓練好的權重進行微調，讓模型更適應本研究中的細胞影像分佈與語境，然後輸入指定圖片與指令提示詞(Prompt)，讓模型扮演「白血球影像分類器」的角色，藉由模型內建的 Chat Template 將影像與文字格式化為模型可接受的輸入，接著針對輸入的影像進行視覺分析，並產生其對應的描述(輸出的文字將作為該影像的描述)。

為客觀評估模型產出的描述品質，我們引用 SentenceTransformer 模型(All-MiniLM-L6-v2)將生成描述與對應的人工參考描述轉換為語意嵌入(Semantic Embedding)向量，並計算其餘弦相似度(Cosine Aimilarity)作為語意一致性的指標，此相似度指數可用於後續描述篩選、品質控制或人為驗證之依據。

4.白血球分類

(1) 定義六個類別的對應表，對應表是將資料集 CSV 檔中的 Label 數字(0~5)對應到真實的白血球類別名稱。

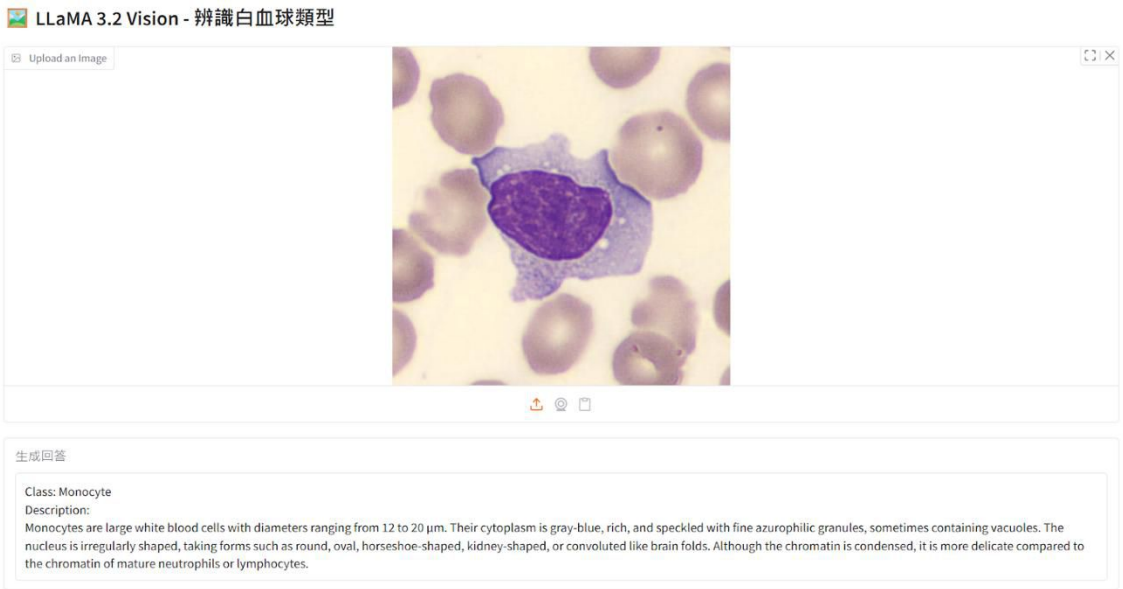
(2) 從生成文字中取出白血球的預測類別，透過推論的方式生成描述文字並找到最早出現的白血球類別名稱，作為模型的預測標籤，如果完全找不到，就預設是 "Neutrophil"，因為嗜中性球細胞的資料量最多。這個做法是因為訓練資料集的圖片文字描述開頭會先提供此影像的類別，再進行的詳細特徵說明，部分資料會提及該類白血球與其他類別的差異，因此有一些生成描述可能包含兩種以上的細胞類別名稱，但會選出最可能的類別作為預測結果。

Chapter 4　實驗結果與系統展示

4.1　系統展示

Gradio 是一個開源 Python 套件，用來快速建立機器學習模型的互動式網頁介面，並支援多種輸入輸出，包含圖片、文字、聲音、影片等，讓開發者可以輕鬆展示模型效果，也可以建立公開網址分享給他人測試。

本系統採用 Gradio 作為前端展示框架，提供簡潔的操作介面，使用者只需上傳白血球影像，系統將圖像與提示語(Prompt)包裝為多模態輸入，傳入後端微調好的 LLaMa-3.2-11B-Vision-unsloth-bnb-4bit 模型進行推論，即可在前端顯示「白血球類型」與「生成描述」，如圖二顯示。



圖二、Gradio 系統展示

4.2 分析結果

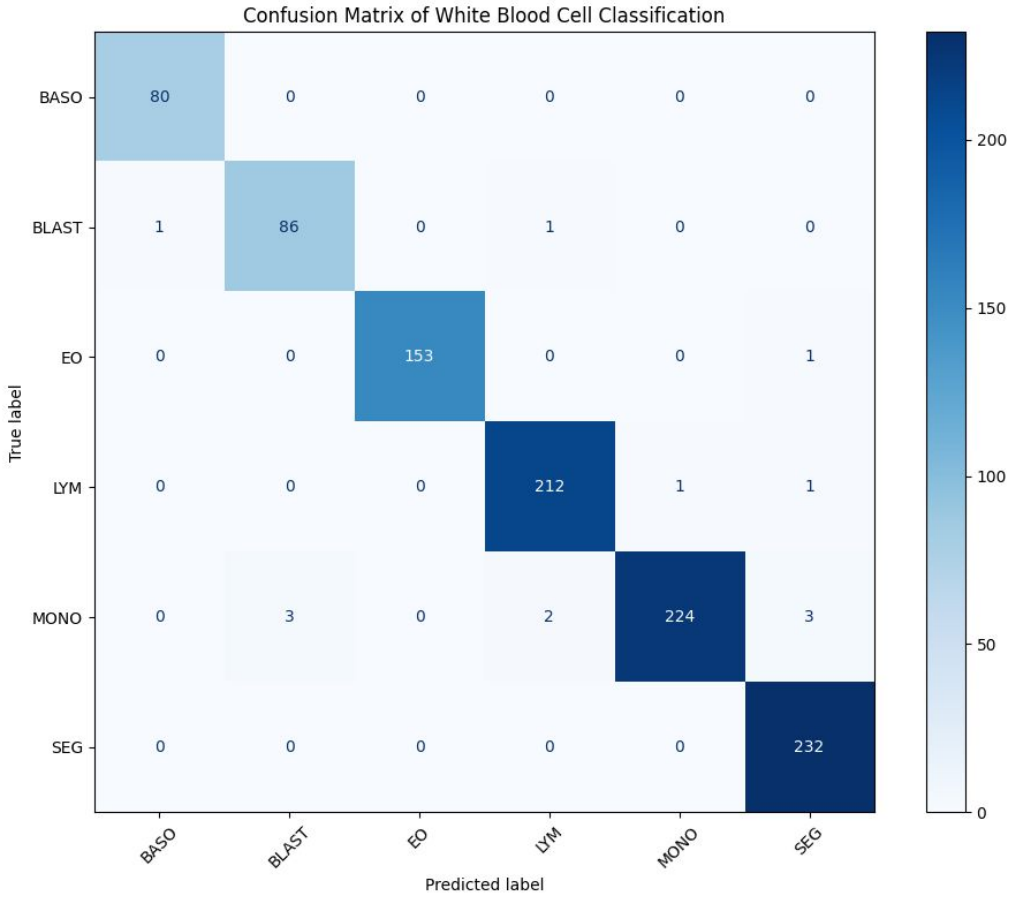
為了呈現與分析模型的預測表現，我們採用了混淆矩陣(Confusion Matrix)、平均值與標準差(Mean ± Standard Deviation)以及分類績效報告(Performance Report)指標，全面評估模型在各類別上的辨識能力與整體準確性。

1.混淆矩陣 (Confusion Matrix)

混淆矩陣(Confusion Matrix)是機器學習中評估分類模型性能的基本工具，特別適用於監督式學習的分類，它以表格形式顯示模型的預測結果與實際標籤之間的對應關係，幫助我們深入了解模型在各個類別上的預測情況。

本研究繪製的混淆矩陣，如圖三所示，從圖中可觀察到模型在各類別皆呈現良好之分類效果，特別是 SEG、MONO 與 LYM 等類別，分別有 232、224 與 212 筆樣本被正確分類，顯示模型能穩定辨識主要細胞類型。

在分類準確性方面，混淆矩陣主對角線的數值皆顯著高於非對角元素，代表模型對各類別具有良好之預測準確性。錯誤分類樣本數極少，僅見於部分相近類別之間，例如：有 3 筆 MONO 被誤分為 BLAST，1 筆 EO 誤判為 SEG，以及少量 LYM 樣本被誤分為 MONO 或 SEG。整體而言，混淆矩陣結果顯示本模型具備優異的分類能力，不只能有效區分各類白血球，也呈現高度的一致性與穩定性。



圖三、本研究繪製的混淆矩陣

2.平均值加減標準差 (Mean ± Standard Deviation)

在統計學中，平均值(Mean)與標準差(Standard Deviation)是描述資料集中趨勢與離散程度的基本指標。平均值表示資料的中心位置，而標準差則衡量資料點相對於平均值的分散程度，這兩者常以「Mean ± Standard Deviation」的形式呈現，提供對資料分布的直觀理解。

本研究進一步採用「平均值 ± 標準差(Mean ± Standard Deviation)」的方式呈現各類別的表現指標，下表 Mean ± Standard Deviation 表格中數值代表該類別在多次測試下的平均準確度與變異程度、附上對應樣本數(Support)以供參考。

從分析結果可見，模型對 EO 與 BLAST 表現最穩定，其準確率分別為 0.9658 ± 0.0483 與 0.9436 ± 0.0697，標準差偏低，顯示模型在這些類別上預測波動小、表現穩定，而 LYM 與 MONO 類別雖平均準確率分別達到 0.9262 與 0.9207，但標準差略高於 EO，可能反映這些類型在某些測試條件下略具挑戰性。

相較之下，SEG 為所有類別中標準差最高(0.1276)，平均準確率亦為最低(0.8495)，顯示模型在該類型分類時表現較不穩定，可能受到細胞形態多樣性或影像品質變異影響，模型在大多數細胞類型上能維持高平均準確率與穩定性。

➤ Mean ± Standard Deviation 表格

	Mean ± Standard Deviation	樣本數
BASO (Basophil)	0.9357 ± 0.0470	80
EO (Eosinophil)	0.9658 ± 0.0483	154
LYM (Lymphocyte)	0.9262 ± 0.0715	214
MONO (Monocyte)	0.9207 ± 0.0802	232
SEG (Neutrophil)	0.8495 ± 0.1276	232
BLAST (Blast cell)	0.9436 ± 0.0697	88

3. 績效報告 (Performance Report)

(1) 準確率 (Accuracy)

準確率是指所有分類（無論是正確或錯誤）正確的比例。由於準確度會納入混淆矩陣的所有四種結果（TP、FP、TN、FN），因此在平衡資料集的情況下，兩個類別的範例數量相近，準確度可做為粗略評估模型品質的指標[14]。其數學定義如圖四：

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

圖四、準確度公式[14]

(2) 精確度 (Precision)

精確度是指模型所有正向分類中，實際為正向分類的比例[14](預測為某類中，有多少是真的)。在實際正例數量極低（例如：總共只有 1 到 2 個例子）的資料不平衡情況下，精確度就沒有那麼有意義，也不太適合作為指標[14]。其數學定義如圖五：

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

圖五、精確度公式[14]

(3) 召回率 (Recall)

召回率是指所有實際正類正確歸類為正類的比例[14](真正屬於某類的樣本中，有多少被正確預測出來)，也稱為喚回率。在實際正例數量極低的不平衡資料集中，召回率比準確度更有意義，因為它可評估模型正確識別所有正例的能力。[14]數學定義如圖六：

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

圖六、召回率公式[14]

(4) **F1 分數 (F1-Score)**

F1 分數是精確度與喚回度的調和平均數（一種平均值)。這項指標可平衡精確度和召回率的重要性，對於類別不平衡的資料集而言，比準確度更為理想。當精確度和召回率都獲得滿分 1.0 時，F1 也會獲得滿分 1.0。更廣義來說，當精確度和召回率的值相近時，F1 也會接近這兩個值，當精確度和召回率相差甚遠時，F1 會與較差的指標相似。[14]從數學角度來看，這項值的計算方式如圖七：

$$\text{F1} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

圖七、F1 分數公式[14]

(5) **樣本數 (Support)**

每個類別中真實樣本的數量(即該類別的樣本總數)，常用來判斷各類別數據是否平衡。

(6) **宏觀平均 (Macro Average)**

代表對所有類別的評估指標(如：Precision、Recall、F1-Score)進行簡單平均，不考慮各類別的樣本數，因此能反映模型在各類別上的平均表現。

(7) **加權平均 (Weighted Average)**

對各類別的指標加權平均，權重為該類別的樣本數(Support)，因此能反映模型在整體資料分布上的實際表現。

根據下表 Performance Report-1 和 Performance Report-2，模型在六種白血球類型中均表現出高度準確性與穩定性，各類別的 F1-score 介於 0.98 至 1.00 之間，整體 Macro 與 Weighted 平均值亦皆達 0.99，顯示模型具備強大的泛化能力與分類精度。特別是 EO、SEG、LYM 等類型，模型分類表現接近完美；只有 MONO 與 BLAST 在召回率上略低，可能與細胞形態間特徵相似或樣本間異質性所致。Macro Average 顯示即使在樣本數不同的情況下，各類型也能穩定預測；而 Weighted Average 結果同樣為 0.99，說明模型對常見與少見類別皆能良好掌握。

➤ Performance Report - 表格 1

	Precision	Recall	F1-score	Support
BASO (Basophil)	0.99	1.00	0.99	80
EO (Eosinophil)	1.00	0.99	1.00	154
LYM (Lymphocyte)	0.99	0.99	0.99	214
MONO (Monocyte)	1.00	0.97	0.98	232
SEG (Neutrophil)	0.98	1.00	0.99	232
BLAST (Blast cell)	0.97	0.98	0.98	88

➤ Performance Report - 表格 2

	Precision	Recall	F1-score	Support
Accuracy	X	X	0.99	1000
Macro avg	0.99	0.99	0.99	1000
Weighted avg	0.99	0.99	0.99	1000

Chapter 5　結論

本研究成功將 LLaMa 3.2 Vision 11B 模型應用於白血球影像分析，並透過 LoRa 技術進行參數高效微調(PEFT)，大幅降低記憶體使用量與訓練資源需求，由於所使用的預訓練模型已具備成熟語言與視覺理解能力，實驗中僅需進行三輪(3 Epochs)訓練，損失值(Loss)便能有效下降至 0.02，顯示模型具備良好的收斂特性與學習效率。

根據 Performance Report 的結果，模型整體分類準確率達 0.99，並在六種白血球類別中皆表現出穩定且高準度的預測能力，顯示其具備實際應用潛力，特別是結合圖像分類與語意描述的多模態生成能力，可作為輔助診斷工具，對於專業檢測人力短缺或資源受限之偏鄉醫療單位具有實質貢獻。

但亦存在部分限制，像是訓練過程耗時長達四天且需配備高效能計算資源，可能對部分使用者形成門檻，未來可進一步探索更輕量化的模型架構與混合訓練策略以兼顧效能與可部署性，同時建議擴展資料類型與樣本規模，提升模型於臨床實務中的泛化能力與應用廣度。

總而言之，本研究證明透過特殊性資料設計與高效微調技術，即可使大型語言模型在特定醫學影像任務中達到高精度的預測與語意理解能力，LLaMa 3.2 Vision 模型經由 LoRa 微調後，成功整合視覺辨識與語言生成功能、展現出優異的多模態能力，不僅能提升影像判讀效能，更可作為專業判斷的補充資訊來源，此成果為大型語言模型於智慧醫療應用開啟了可行之路，特別適用於資源有限地區，未來具有進一步拓展與實務部署的潛力。

Reference

[1] 復旦大學：2024 大語言模式的能力邊界與發展思考報告 https://hao.cnyes.com/post/119391

[2] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, Jianfeng Gao, Large Language Models: A Survey, https://arxiv.org/pdf/2402.06196, 2025

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, https://arxiv.org/pdf/1706.03762, 2023

[4] Paszke, Adam and Gross, Sam and Massa, Francisco and Lerer, Adam and Bradbury, James and Chanan, Gregory and Killeen, Trevor and Lin, Zeming and Gimelshein, Natalia and Antiga, Luca and others, https://arxiv.org/pdf/1912.01703, 2019

[5] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, Tijmen Blankevoort, A White Paper on Neural Network Quantization, https://arxiv.org/abs/2106.08295, 2021

[6] 大幅提升 LLM 微調速度：Unsloth 介紹 https://tenten.co/learning/what-is-unsloth-llm-fine-tune-tool/

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, LoRA: Low-Rank Adaptation of Large Language Models, https://arxiv.org/abs/2106.09685, 2021

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, Sylvain Gelly, Parameter-Efficient Transfer Learning for NLP, https://arxiv.org/abs/1902.00751, 2019

[9] Touvron, Hugo and Lavril, Thibaut and Izacard, Gautier and Martinet, Xavier and Lachaux, Marie-Anne and Lacroix, Timothée and Rozière, Baptiste and Goyal, Naman and Hambro, Eric and Azhar, Faisal and others, LLaMA: Open and Efficient Foundation Language Models, https://arxiv.org/pdf/2302.13971, 2023

[10] Meta-llama/Llama-3.2-11B-Vision https://huggingface.co/meta-llama/Llama-3.2-11B-Vision

[11] Llama Team, AI @ Meta, The Llama 3 Herd of Models https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

[12] Unsloth/Llama-3.2-11B-Vision https://huggingface.co/unsloth/Llama-3.2-11B-Vision

[13] Fine-Tuning Llama 3.2 Vision https://www.datacamp.com/tutorial/fine-tuning-llama-3-2-vision

[14] 分類：準確率、喚回度、精確度和相關指標 https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall?hl=zh-tw

附錄 A. 專題工作內容

● **邱芯彤、黃彥婷**

我們於謝瑞建教授的醫療資訊暨遠距醫學實驗室進行專題研究，參與大型語言模型研究，以及將結果運用於醫學領域。製作專題期間透過定期與教授開會回報研究進度、自主與組員線上討論，提升本人於大型語言模型應用之專業能力，並對 AI 應用更深入的理解。工作內容包括：

(1) **環境安裝：**建立適合開發的虛擬環境，了解並利用不同套件之應用(例如：Pytorch、Transformer、Python 等等)。

(2) **遠端連接：**使用 MobaXterm 連接至實驗室電腦進行研究，跨越距離與空間的障礙，增加研究便利性。

(3) **資料集調整：**標準化白血球的影像，然後將輸入模型的圖片調整程式適合 LLaMa 的輸入格式。

(4) **模型微調：**運用 GPU、LoRa 技術做 PEFT 訓練輔助模型進行微調，減少訓練的記憶體用量，並提升處理速度。

(5) **推論與評估：**推論-了解模型的架構，依照其對話架構進行圖片輸入與描述生成，之後再利用其描述分類白血球影像。評估-使用混淆矩陣、平均值加減標準差和績效報告來評估與分析準確率。

附錄 B. 專題心得與建議

● 邱芯彤

這一年的專題研究中，我們遇到了不少挑戰，但也因此收穫良多。去年暑假在實驗室學長姐的帶領下接觸 RAG 技術，這是我第一次踏入大語言模型的應用領域。一開始面對大量技術上的專有名詞和開源工具有點手足無措，也曾因模型效果不如預期而感到挫折，但在老師的教導與同學的合作，以及不斷嘗試下，我慢慢學會如何查詢資料、找出問題並解決。這些過程雖然辛苦，看到結果時的成就感，讓所有努力都變得有意義。

一路從陌生到熟悉，我開始不再只是使用工具，而是開始理解背後的原理與邏輯，並激發我深入探索這個領域的興趣。這次專題製作不僅讓我學習到專業知識，也增強了解決問題的能力，更重要的是我獲得在遇到困難不放棄的毅力。

● 黃彥婷

在加入實驗室的這段期間，從暑假開始學習教授和學長姐推薦的教材到一步步開始進入狀況，邁向實作的第一步是由學長姐教我們製作 RAG 的應用，對於只會用 AI 來問問題的我來說，真的讓我感到很有趣也很新奇，後來也決定往相關方面繼續研究下去。

但在真正開始自己實作時，總是會遇到一些狀況和挑戰，像是：第一次用遠端，裡面的許多功能都不太了解、在環境安裝時總是出現無法理解的錯誤等等，不過在教授、學長姐和同學的幫助下，我也漸漸地感受到自己的進步，雖然還是會因為做不出來而感到挫折，但跟解決問題和研究成功後的喜悅相比，這樣的感覺是挫折無法比擬的，因此讓我更有動力繼續研究，也很感謝幫助我的大家！