國立臺灣大學生物資源暨農學院生物機電工程學系

碩士論文

Department of Biomechatronics Engineering
College of Bioresources and Agriculture
National Taiwan University

Master's Thesis

應用生醫文獻大型語言模型於白血病相關基因的擷取與分析

Retrieving Leukemia-related Genes Using Large Language Models built with Biomedical Literature

林東甫

Tung-Pu Lin

指導教授：陳倩瑜 博士

Advisor: Chien-Yu Chen, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202504150

謝辭

滿懷著喜悅與感傷的心情終於要離開母校臺大，當年在得知錄取時的無比激動之情，好比經歷了萬古黑夜的長路才總算迎來的第一道曙光；在作為新生時的無數個快活探索的日子以及無數個孤苦奮鬥的夜晚至今仍是歷歷在目，在即將要開始人生新的旅程以前，我首先要感謝恩師陳倩瑜老師，願意收留我這樣懶惰愚鈍的學生，並且一直以極大的雅量容忍我並給予我自由，同時亦毫不吝惜地指導我使我得到引導，這樣在無助中給人希望的力量，不啻恩同再造，隻言片語實難以言表其恩德。

同時也想感謝我的家人，感謝我的父母，一直以來盡他們所能地永遠支持著白日做夢的我並給予我最大的幫助；感謝我的姊姊在我這一路巔沛流離時，傾其所能不斷地幫助著我、盡己所能不斷地鼓勵著我，這樣的犧牲成就是我這種無才無德之人本不應得的，我實在受到的恩惠太多，遠超出一切我所能回報的。

我也想感謝這一路走來的戰友昕恩哥哥的陪伴與他一直以來的幫助，這短短幾年來我們一齊並肩作戰、一路披荊斬棘建立起的革命情感讓我始終銘記在心裡。同時也要感謝孫醫師給予我的各種寶貴建議，不但使得這份研究成為可能，也極大地擴展了我的視野。感謝最可愛的學弟妹昭佑、嘉安、于婷，他們是最優秀也是最善良的後輩，我從他們那得到的幫助遠多於我作為前輩所能給予他們的幫助。

最後，感謝 C4Lab 的夥伴們以及一路相互提攜的師長同學們，謹以此文，獻給所有在求學路上曾經幫助與啟發我的人；願曾給我光的人，也都能被世界溫柔以待。

doi:10.6342/NTU202504150

中文摘要

急性骨髓性白血病（Acute Myeloid Leukemia, AML）是一種高度異質性的惡性腫瘤，其病理進展與治療策略密切受到體細胞突變、遺傳特徵以及對各類治療反應的影響。隨著個人化醫療與基因標靶治療的持續發展，如何有效辨識並分析與疾病相關的基因，已成為提升臨床預後判斷與治療效果的關鍵。然而，現有的基因資訊擷取技術仍然高度仰賴傳統的文獻檢索與人工整理，缺乏具擴展性與自動語意理解能力的工具。為解決此問題，本研究旨在探討開源大型語言模型（Large Language Model, LLM）於生物醫學文獻分析中的應用潛力，並以 AML 為實例，建立一套基於 BioGPT 模型的基因機率分析流程。BioGPT 是由微軟開發的語言模型，基於 OpenAI 原始 GPT-2 架構，並進一步以生物醫學文獻摘要進行預訓練。本研究採用以「Causal Language Modeling」中的「next-token prediction（下一詞預測）」為核心的分析框架。透過建構疾病語境提示語（disease-context prompt），評估特定基因名稱作為下一詞出現的正規化機率值，以推測其與疾病文獻的語意關聯性。以歐洲白血病網路（European Leukemianet, ELN）AML 治療指引中建議的基因集合為測試資料，本研究結果顯示，透過適當的提示詞設計（Prompt Engineering），已知和 AML 相關的基因與其他基因在預測機率分布上存在顯著差異，顯示本方法具有潛在有效性。為進一步強化模型的穩定性與泛化能力，本研究整合檢索增強生成（Retrieval-Augmented Generation, RAG)機制，並調整嵌入模型與 chunk size 等超參數，進行分析架構改良。最終結果顯示，透過上述微調（Fine Tune），可有效提升模型對 AML 相關基因的預測準確性，並聚焦於具高度語意關聯之目標區段。本研究所提出的分析流程可作為未來生醫語言模型應用之模組化方法框架，亦提供大型語言模型應用於領域知識擷取與語意分析的初步實證依據與技術參考。

關鍵字：急性骨髓性白血病、大型語言模型、生醫文本探勘、檢索增強生成、提示工程、微調

doi:10.6342/NTU202504150

英文摘要

Acute Myeloid Leukemia (AML) is a highly heterogeneous malignancy, whose pathological progression and therapeutic strategies are significantly influenced by somatic mutations, genetic characteristics, and responses to various treatments. With the continuous advancement of personalized medicine and gene-targeted therapies, the effective identification and analysis of disease-associated genes has become critical for improving prognosis and treatment outcomes. However, current gene information extraction methods still heavily rely on conventional literature searches and manual curation, lacking scalable and semantically-aware automated tools. To address this issue, this study explores the potential applications of open-source Large Language Model (LLM) in biomedical literature analysis, using AML as a case study to establish a gene probability estimation pipeline based on the BioGPT model. BioGPT, developed by Microsoft, is a transformer-based language model built upon OpenAI's original GPT-2 architecture and further pretrained on biomedical literature abstracts. We adopt a next-token prediction framework rooted in Causal Language Modeling, constructing disease-context prompts to evaluate the normalized probability of a given gene name appearing as the next token. This score is then used to infer the semantic association between the gene and AML-related literature. Using gene targets recommended by the European Leukemianet (ELN) AML treatment guidelines as test data, our results demonstrate that with carefully designed prompts, the predicted probability distributions for known targets are significantly different from those of other genes, indicating the potential effectiveness of this approach. To further enhance model stability and generalizability, we introduced the Retrieval-Augmented Generation (RAG) framework into the existing pipeline and fine-tuned architectural components such as the embedding model and chunk size. The final results show that these refinements improved the model's prediction performance and allowed it to better focus on the relevant semantic scope, leading to more selective and goal-oriented predictions. The proposed pipeline provides a modular framework for future biomedical LLM applications and offers preliminary empirical support and technical references for domain-specific knowledge extraction and semantic analysis using large language models.

Keywords: Acute Myeloid Leukemia (AML), Large Language Model (LLM), Biomedical Text Mining, Retrieval-Augmented Generation (RAG), Prompt Engineering, Fine-tune

doi:10.6342/NTU202504150

目　　次

doi:10.6342/NTU202504150

doi:10.6342/NTU202504150

圖　　次

==========

doi:10.6342/NTU202504150

doi:10.6342/NTU202504150

表　　次

doi:10.6342/NTU202504150

第一章 前言

1.1 背景介紹

白血病（Leukemia）又稱血癌，是一種影響血液和骨髓的惡性腫瘤，主要可以分為急性和慢性兩類，並且包括了急性骨髓性白血病（Acute Myeloid Leukemia, AML）和急性淋巴性白血病（Acute Lymphoblastic Leukemia, ALL）等亞型。白血病的病理過程以及其治療方式涉及基因突變、遺傳特徵及其對不同治療方法的反應。因為基因變異對治療反應和預後的影響十分顯著，因此其中這些基因特徵的發現和分析對於改善患者的個體化治療方案至關重要。在白血病治療的臨床實踐中，醫師經常會需要針對患者的特定基因突變來決定適合患者的最佳治療方案，例如化療、幹細胞移植或免疫療法等。因此，對於識別和分析白血病相關基因的遺傳特徵及其對治療反應的影響，對於提升治療成功率、延長患者的生存期具有重大意義。（Papaemmanuil et al., 2016）

隨著生物醫學研究的迅速發展，在生醫領域相關的學術文獻也呈現爆發性增長，在這其中也包含了大量記錄了最新的基因突變與治療反應的相關研究。然而，由於白血病研究所涉及的基因範疇相當龐大，特別是在基礎生物醫學研究與基因相關的研究資料繁多，時至今日以人工手動檢索並分析所有相關生物醫學文獻已經變得非常困難。傳統的檢索方法已經無法有效且快速地從海量文獻中提取有價值的資訊，從而難以幫助醫師在有限時間內制定最佳治療方案。

在今日生物醫學研究領域文獻出現資訊爆炸的背景之下，大型語言模型（Large Language Model, LLM）技術，特別是生成式預訓練 Transformer 模型（Generative Pre-trained Transformer, GPT），（Radford et al., 2018）如在生物醫學領域的 BioGPT（Luo et al., 2022），已經展示了其在自動化分析並生成大量生醫文獻方面的巨大潛力。BioGPT 是一種專門用於生物醫學領域文本的預訓練語言模型，而本研究便試圖通過自然語言處理（Natural Language Processing, NLP）技術，從海量的生物醫學文獻中自動提取關鍵基因資訊並進行資訊整合。

doi:10.6342/NTU202504150

1.2 研究目的

本研究旨在開發一套以 BioGPT 為核心的自動化分析流程，系統性評估 AML 相關基因之語義關聯性與潛在功能意涵。具體而言，本研究聚焦於以下四個研究目標：

首先是建構起一套以 BioGPT 為基礎的基因機率值評估系統，應用其因果語言建模能力（Causal Language Modeling）與 next-token prediction 機制，（Luo et al., 2022）針對在生物醫學研究語境下的提示語（prompt），預測特定基因名稱出現為「下一詞」的機率，並以此機率值作為估算該基因與 AML 的語義相關程度之基礎。我們以歐洲白血病網路（European LeukemiaNet, ELN）指南所列標的基因為指標核心，並逐步擴展至 HGNC 所收錄的 44,000 多個人類基因名稱，以進行完整的機率值語義評分（semantic scoring）分析（Döhner et al., 2022）。

其次，為解決原生 BioGPT 本身的多 token 組成的基因名稱在生成過程中造成的機率偏差問題，我們也必須進一步嘗試引入不同的 token-level 正規化策略（normalization strategy），分別是 Per-token Total Sum Normalization，以及基於 token 長度的 length-based normalization 方法。這些正規化策略的使用，將有助於提升模型在不同基因長度下的預測一致性，並強化整體評分流程的公平性與穩定性（Zhavoronkov et al., 2023）。

第三，綜上所述需要開發一套可行的全自動評分流程（scoring pipeline），可批次處理任意基因清單（包含全體 HGNC 基因或 ELN 所列之標的基因），並輸出其在特定提示詞下的預測機率分佈，並且結合視覺化模組與統計檢定（如 U-test、paired t-test）來評估不同基因群體間的差異顯著性以檢視該評分流程之成效。該流程模組化設計也將便於未來可以延伸至其他與基因相關之癌症或疾病領域進行語義機率值分析。

doi:10.6342/NTU202504150

最後，本研究亦嘗試引入檢索增強生成（Retrieval-Augmented Generation，RAG）技術，透過所有與 AML 相關之 PubMed 文獻整合為語料庫的語意檢索機制，自動取出與 AML 最相關之段落作為提示語背景（prompt context），以強化 BioGPT 的語境理解能力與生成品質。最後再實驗比較，加入語境強化後是否能夠顯著提升模型對標的基因的辨識能力與語義精準度，此目標為體現出結合知識檢索與語言模型推論之潛力（Lewis et al., 2020）。

整體而言，本研究不僅提出一套可系統化評估疾病關聯基因的生成式語言模型方法，也為大型語言模型在生醫領域的知識挖掘與語義判別任務提供技術基礎與實證依據，未來可望應用於疾病基因發掘、藥物靶點預測、及個人化醫療決策支援等多樣化場景。

doi:10.6342/NTU202504150

第二章 文獻探討

隨著近年來次世代定序技術（Next Generation Sequencing, NGS）與人工智慧技術的迅猛發展，基因體醫學研究已逐漸進入資料驅動（Data Driven）的新時代。AML 作為一種高度異質性和致命的血液系統惡性腫瘤，其病理機制與基因突變息息相關；在 AML 中，不同基因突變會影響患者的臨床表現、治療反應及最終預後，因此如何識別這些基因變異對於設計個體化治療方案至關重要。然而由於相關生物醫學領域文獻數量甚為龐大、研究資料的複雜性高，依靠傳統人力資訊檢索的方法去探討白血病的基因突變與治療反應的關係是非常繁重的工作。基於此背景而引入人工智慧，特別是基於生成式預訓練的語言模型（如 BioGPT），對於在生物醫學研究中的應用為解決這一困境提供了新的視角和方法。

2.1 白血病的遺傳特徵與治療反應

白血病的基因變異在其病理過程中影響顯著，近年研究（Papaemmanuil et al., 2016）顯示，AML 的發生與多種基因的突變密切相關，包括 *FLT3*、*NPM1*、*CEBPA*、*RUNX1* 等。這些基因變異不僅影響細胞的增殖和分化，還與患者的治療反應和預後密切相關。例如，*FLT3-ITD* 突變與 AML 患者較差的預後和較高的復發率相關，而 *NPM1* 突變患者則對於化療的反應較好，並且預後相對較佳（Döhner et al., 2017）；而關於 *NPM1*、*FLT3-ITD* 和 *CEBPA* 等基因變異的研究，揭示了這些基因突變對於 AML 患者的治療反應和預後具有顯著影響（Schlenk et al., 2008; Papaemmanuil et al., 2016）。然而由於 AML 的高度複雜性，單一基因變異無法完全地解釋致病性和治療反應，因此需要做更全面的相關研究文獻的整合。

隨著生物資訊領域的飛速發展，近年來基因組學的研究在 AML 的遺傳特徵識別方面取得了顯著進展，而基因定序技術的大幅進展使得基因變異的高通量篩查成為可能。然而隨著資料量的迅速增長，相關領域的研究工作也大量的

doi:10.6342/NTU202504150

增加，如何從大量的研究資料中提取有效信息已然成為當今研究工作的一大課題。若能通過有效整合多種基因突變資訊，基於遺傳特徵進行 AML 風險評估的概念，便能為往後 AML 的治療決策提供更多文獻依據。

2.2 Transformer

Transformer 技術是自然語言處理領域的一項革命性進展，自 2017 年提出以來，立即迅速取代過去廣泛使用的循環神經網絡（Recurrent neural network，RNN）和長短期記憶網絡（Long Short-Term Memory，LSTM），成為當前語言模型的基礎架構。Transformer 的關鍵在於使用 Attention mechanism，它能夠允許此類模型有效的處理句子中的不同單詞之間的長距離依賴關係（Long-range dependence，LRD），並同時並行計算整個輸入序列，這大幅提升了訓練成果和效率（Hochreiter et al., 1997）。

在此以前，RNN 和 LSTM 等傳統神經網路模型往往依賴於順序處理輸入資料，導致長距離依賴的關係學習效果不佳，並且模型結構導致訓練起來相當困難，這些都嚴重影響到應用效果。Transformer 通過自 Attention mechanism 解決了這一問題，與 RNN 和 LSTM 不同，Transformer 模型可以平行處理整個輸入序列，而不是依賴於逐步處理，這使得訓練過程顯著加速。使得模型能夠同時關注句子中的不同部分，並且能夠在更大的上下文範圍內進行學習。這使得 Transformer 在處理各種語言任務（如機器翻譯、文本生成、摘要和問答等）上表現更加出色。（Vaswani et al., 2017）

自 Transformer 技術的橫空出世以後，基於 Transformer 的模型，如 BERT（Bi-directional Encoder Representations from Transformers）和 GPT（Generative Pre-trained Transformer）如雨後春筍般陸續誕生，極大地提升了自然語言處理任務的此類型模型性能。BERT 是一個雙向編碼器模型（Encoder），它能夠同時考慮句中每個詞語的前後文關聯，這對各種自然語言處理任務如詞句分類

doi:10.6342/NTU202504150

（Sentence Classification）、命名實體識別（Named Entity Recognition，NER）和問答系統（Question Answering，QA）帶來了顯著性能提升（Devlin et al., 2019）。而 GPT 是一個自回歸生成模型，它能夠根據上下文生成連貫的文本，並且在文本生成（Text Generating）、對話系統（Dialogue System）和語言翻譯等文字生成任務中表現優異（Radford et al., 2018）。

總體而言，Transformer 並非特定於某一種語言建模任務，而是一種通用架構，其可根據設計的不同（如 Encoder-only, Decoder-only, Encoder-Decoder）搭配不同訓練任務（如 MLM, CLM）進行調整與應用。顯然基於 Transformer 的預訓練模型可以被用於各種 NLP 任務，並且僅需通過微調（fine-tuning）即可應用於新的任務，而不需要從頭開始訓練，這提高了模型在使用上的靈活性和重用性（Wolf et al., 2020）。尤有甚者，Transformer 模型的架構能夠隨著計算資源的增長而線性擴展，這也使得為後續大規模語言模型陸續問世提供了紮實技術基礎，今日如 GPT 或 BERT 等大型語言模型能夠以超大算力有效地處理更大規模的數據並進行預訓練，從而成為更加精確和功能強大的多功能任務語言模型。

2.3 語言建模任務

語言建模是自然語言處理中核心的預訓練任務之一，主要分為兩大類型：

（1）掩蔽語言模型（Masked Language Models, MLM）

（2）因果語言模型（Causal Language Models, CLM）

MLM（如 BERT）在訓練時會將部分詞語進行隨機遮蔽（masking），讓模型學習根據前後文預測被遮蔽的詞語。這種雙向語境建模方式對語意理解與分類任務非常有效。然而，MLM 無法自然應用於純生成任務，且因遮蔽策略在推理階段不再使用，導致訓練與應用之間存在方法不一致的問題。

doi:10.6342/NTU202504150

CLM（如 GPT）則是透過自回歸（autoregressive）機制，即模型只能根據當前序列的前綴來預測下一個詞，其訓練與推理方法一致，且特別適合於連貫的語言生成與序列建模任務。CLM 屬於單向預測策略，且常搭配 Decoder-only 的 Transformer 架構，透過 Causal attention mask 限制模型只能關注先前出現的 token。

本研究採用的 BioGPT 模型即為基於 CLM 任務訓練的生醫語言模型，本研究使用 next-token prediction 方法計算指定基因名稱在 AML 語境下出現的機率。此方法透過逐 token 推理與乘積累計方式，估計整體詞組的生成可能性。由於 CLM 架構能忠實保留自然語言的序列性並捕捉語境間的語意邏輯，特別適合應用於如本研究此類強調語境推斷與語意生成任務。這種架構的設計與另一種常見的語言建模架構 MLM 形成了鮮明的對照，兩者各自是不同技術路線的發展，並有其適用的下游任務範疇與技術挑戰，而對於本研究採用 BioGPT 進行生醫文本的基因機率分析任務而言，CLM 的運作原理至關重要。

因果語言模型的設計基礎同樣來自於生成式預訓練的想法，而這類模型通常透過在大規模語料庫上的自監督學習訓練，在訓練的過程中學習在給定過去詞序列的情況下來預測下一個詞的機率值分布。這種訓練模式的最大優勢是能夠透過捕捉語言中自然語境的前後依賴關係，並且具備自然生成文本的能力。OpenAI 所提出的 GPT-2 與 GPT-3 更是將 CLM 的性能推向高峰（Radford et al., 2018; Brown et al., 2020）。

與 MLM 相比，CLM 通常在評估詞語境相關性類型的任務中更具優勢，原因有二，其一是因為它所產生的任一 token 都是一致地建構於前文語境上，而不會像 MLM 一樣存在遮蔽訊息從而造成語意扭曲問題。其二，CLM 不需要預先標記「哪一個 token 要預測」，其自回歸性質允許在自然語言中的序列性的特質能夠得以保留，因此在對於模擬人類語言活動如寫作與閱讀的語意流動更為真實，而在本研究中即為能夠更好的幫助在大量預訓練資料之上進行預測。

doi:10.6342/NTU202504150

此外在技術實作上，CLM 在模型架構上採用的是單向 Transformer 架構，透過 Causal attention mask 限制每一個 token 僅能關注先前出現的詞。這使得模型在訓練與推理階段較能維持一致，在純粹使用推理能力的應用上，也相較之下無須再考量其在訓練上的方法與推理方法不同的問題。

總體來說，CLM 作為一種基於自回歸預測的語言建模方法，非常適合用於需要考慮語境語意一致性的生成與預測任務。因此本研究在基於生物醫學研究領域的專用文獻中，將 CLM 應用於相關基因名稱機率評估，不僅可自動推斷語意相關性，亦能避免傳統搜尋與人工標記的限制，展現其在智慧化知識抽取中的潛力。

2.4 基於 Transformer 的模型在醫學研究中的應用

Transformer 架構是自然語言處理領域近年來的一個重大突破，至今已經成為處理各類語言文字類型任務的基礎模型框架。自 Attention mechanism 使得模型能夠同時處理長距離依賴的語言關聯性，而這對於處理生物醫學文獻中複雜專門用語與專有語境的語義結構尤為有用。因此在生物醫學領域，Transformer 技術的應用也已經顛覆了傳統的文獻檢索與知識提取方式。基於 Transformer 的模型（如 BioBERT、PubMedBERT 和 BioGPT 等）通過預訓練大量生物醫學文本，使得自動化的知識提取、關鍵字檢索和資訊總結變得更加高效。這些專門用於專用領域的模型不僅能處理專業的醫學術語和複雜的專業研究文獻內容，還能夠根據不同的研究需求生成高質量的醫學建議和摘要，顯示出卓越的文獻處理和資訊提取能力。

BioBERT 是一種基於 Transformer 的語言模型，它通過在大規模生物醫學文本（PubMed 和 PMC）上進行訓練，能夠識別並提取出文本中的基因、蛋白質、疾病和藥物等專有名詞，從而加速了研究文獻檢索和綜述的過程（Lee et al., 2020）。與之相似的，PubMedBERT 則是專門針對 PubMed 中的文獻進行訓

doi:10.6342/NTU202504150

練的基於 Transformer 的語言模型，因此能使其在生物醫學文本中的語義理解更加精確（Gu et al., 2021）。

BioGPT 作為一種生成式預訓練 Transformer 模型，同樣專門針對 PubMed 中的文獻摘要進行訓練。該模型能夠不僅能夠提取文本中的專有名詞或針對生物醫學有關內容進行問答，還能夠自動生成與生物醫學相關的摘要以及識別關鍵研究發現，從而有效的減少研究人員手動檢索和分析文獻的工作量（Luo et al., 2022）。因此基於 Transformer 的各式模型應該可以幫助快速識別和分析與不同基因突變和治療反應相關的研究，從而為制定個體化治療方案提供關鍵支持。因此，總體來看基於 Transformer 的模型應用在加速醫學文獻的處理和自動化信息提取，將可以為研究人員提供有效的工具來應對海量文獻資料的挑戰。

2.5 檢索增強生成

在大型語言模型蓬勃發展的今日世界，各類新型面向不同下游任務的大型語言模型如雨後春筍般快速誕生，如何有效提升其對於下游任務之專業領域的精確度與知識廣度，成為近年來自然語言處理研究的熱門課題。在此其中，最令人矚目的要以檢索增強生成為核心的一系列新技術，這是一種結合資料檢索與文本生成的技術架構，其試圖透過動態地將外部知識庫引入語言模型的輸入層中，並以此來提升其對特定知識密集任務（knowledge-intensive tasks）的應對能力。

RAG 的基本概念，很早便源自於 2020 年的將非參數式記憶結合進語言模型的構思（Lewis et al., 2020）。這其中 RAG 模型主要會包含了兩個子模組：一個是檢索器（Retriever），主要會負責自文獻庫中找出與查詢相關的文本片段；另一個則為生成器（Generator），通常也就是一個自回歸語言模型，如 BERT 或 GPT，則會根據查詢與檢索到的相關內容生成回應。透過這種架構，使得模型得以利用外部資料庫中的知識，來突破傳統中預訓練語言模型在其訓練參數內

doi:10.6342/NTU202504150

知識（parametric knowledge）受限的問題。

而使用 RAG 的最大好處之一則是在於其極富彈性的可擴充性。在與一般 LLM 相較之下，傳統的 LLM 中的知識受限於訓練參數中，因此一旦知識內容過時或超出範疇，則便需要再次重新訓練整個模型，過程曠日廢時。而在 RAG 架構中，知識的更新只需要針對檢索資料庫再進行修改便足夠，整個模型本身並不需要大規模的重新訓練，便可以反映出新的或者原有參數中未含的知識內容。此外近年來 RAG 架構在許多諸如零樣本學習（zero-shot learning）、開放式問答（open-domain QA）、醫學文本問答、法律查詢等下游任務中，均展現出不俗的表現。

因此在本研究中亦會試圖將 RAG 架構應用於生醫文獻中的疾病基因關聯性分析。即以 BioGPT 作為基礎的生成式模型，搭配蒐集來的以 PubMed 資料庫中 AML 相關的生醫研究文獻摘要所構建之向量化資料庫，使 BioGPT 生成時可參照具體文獻內容進行 next-token prediction，從而提升特定基因名稱的語意生成機率，間接反映其在文獻語境中的語意關聯性。本研究也使用 LangChain 框架實作向量資料庫，以 RecursiveCharacterTextSplitter 切分段落，Huggingface 模型進行語義嵌入，構建出高品質且更專業的基因文獻語境資料庫，進一步完成新的 BioGPT-RAG 基因機率值預測流程框架。

總體來說，檢索增強生成技術提供了有效整合外部知識與語言模型推論能力的新方法，特別適合應用於專業術語密集、背景知識繁雜的生醫文獻場景。其模組化與可延展性特質，使其在未來大型語言模型之下游任務的應用開發中具備高度潛力。

doi:10.6342/NTU202504150

第三章 研究方法

3.1 資料蒐集與預處理

本研究所採用的主要資料來源為 PubMed，在一方面是本研究主要使用的模型 BioGPT 是經由 PubMed 文獻摘要預訓練而成；另一方面，本研究所使用之 RAG 之語料庫來源亦是自 PubMed 中蒐集。

3.1.1 PubMed

PubMed 是美國國家衛生研究院（National Institutes of Health，NIH）的國家生物技術資訊中心（National Center for Biotechnology Information，NCBI）與國家醫學圖書館（The United States National Library of Medicine，NLM）共同維護管理的生物醫學文獻資料庫，其收藏涵蓋了超過 3000 萬篇來自生物醫學領域的文獻和期刊文章。該資料庫自 1996 年開始運行，當下已成為世界級生物醫學研究查閱、引用和分享文獻的主要平台。而在白血病相關研究中，PubMed 也提供了相當豐富的資源，其中包括基因變異與治療反應相關的文章、臨床試驗數據和治療指導原則（NCBI, 2020）。

在 PubMed 中亦收藏有各類與 AML 基因突變相關文獻，其涵蓋了不同的研究主題，從分子生物學到臨床應用，提供了大量基因變異與治療反應關聯的數據。因此有效的使用 PubMed 可以對這類 AML 與基因變異相關的文獻進行檢索，為後續研究提供豐富的資料。

隨著 PubMed 中的研究文獻數量日漸持續的飛速增加，手動檢索和分析這些數據已經變得越來越困難。因此儘管這些基因突變的研究為 AML 的治療提供了許多的寶貴建議，但如何能夠更有效地將這些知識整合並且應用於臨床，並且實現對基因特徵與治療反應的即時全面性分析，仍然是當前研究工作中的一大難題。要克服這項難題將可能需要一種能夠自動從海量文獻中提取目標資訊的輔助工具，並且可以跨領域進行相關資訊的快速整合與應用，而大型語言模型的出現為解決這一問題提供了契機。

doi:10.6342/NTU202504150

3.1.2 AML 相關文獻蒐集與預處理

儘管 PubMed 一方面提供了龐大的文獻資源，但如果直接將其作為 RAG 與料庫的輸入，將可能會面臨資料過於雜訊化、語境差異以及語言格式多樣等問題。因此在正式將資料導入語言模型之前，必須要先進行有效且系統化的資料過濾與預處理（Data filtering and preprocessing）。本研究在此設計了一套處理流程，從文獻檢索結果中提取具與 AML 相關生物醫學價值的摘要內容，並排除可能干擾模型學習效果的不相關資訊，以確保後續分析具備一致性。

首先，透過關鍵字進行資料初步篩選。本研究鎖定的疾病領域為急性骨髓性白血病（Acute Myeloid Leukemia, AML），因此在進行文獻抓取時，篩選條件設為標題或摘要內容包含「Acute Myeloid Leukemia」關鍵字，同時為避免語義模糊或疾病混淆，將包含「CML」（慢性骨髓性白血病）、「ALL」（急性淋巴性白血病）等非 AML 相關的摘要剔除。其次，為確保文獻內容的時效性與語言一致性，僅保留發表時間至 2024 年底以前、且語言為英文的文獻，避免因語言差異導致語言模型的表現受影響，經此初步篩選之後共計取得 44,279 份文獻摘要（圖 3-1）。

在取得初步篩選後的文獻摘要後，接著進行文本內容的清理與格式化。由於 PubMed 摘要涵蓋的年代分布較長，在撰寫風格與文章格式上可能存在段落不齊、標點符號使用不一致、特殊符號（如 HTML 標記或非 ASCII 字元）殘留等問題，本研究使用了正則表示式（Regular Expression）與自然語言處理套件（NLTK）進行基礎清理，包括移除多餘換行字元、統一標點格式、過濾無意義符號等。接著，為排除重複文獻或無效摘要（如空白摘要或極短摘要），本研究設定最小長度閾值（如 50 個字以上）進行篩除，並使用 Pandas 等工具進行去重複（drop_duplicates）與空值處理（dropna）。最終處理完成後，將所有摘要段落轉換為統一格式的資料結構，以便後續導入語言模型進行嵌入建構與語意分析，最後共計取得 27,524 份文獻摘要。

doi:10.6342/NTU202504150

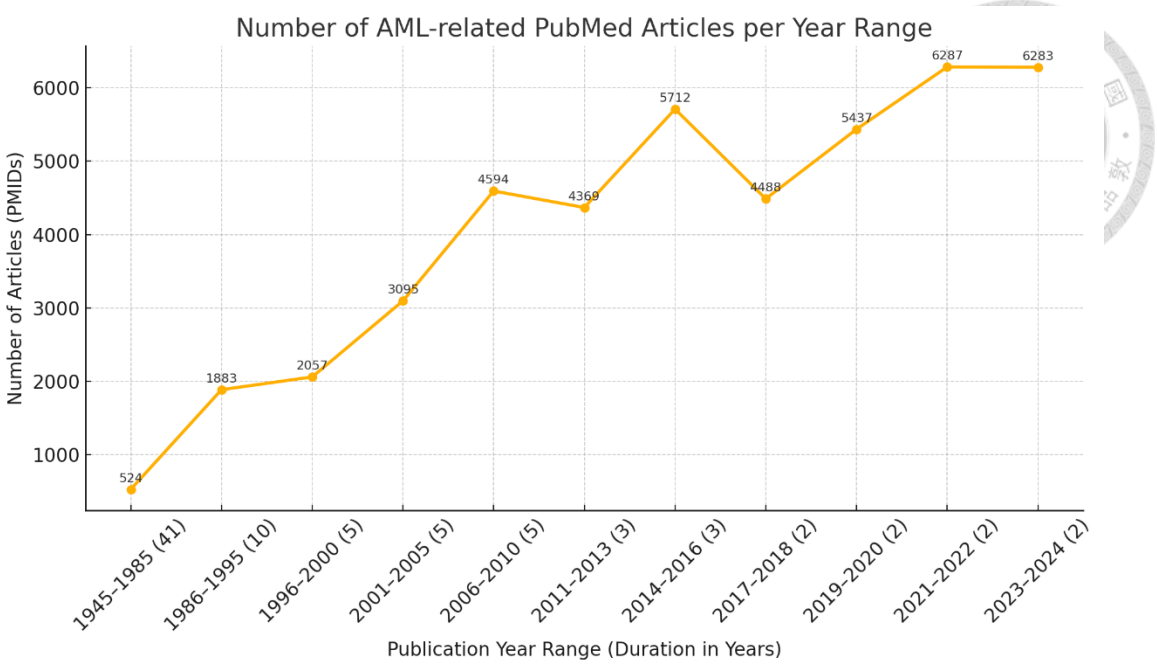

圖 3-1：不同年份群組之 Acute Myeloid Leukemia 關鍵字搜尋結果文獻摘要數

3.1.3 European LeukemiaNet

歐洲白血病網絡（European LeukemiaNet, ELN）為一跨國研究合作組織，其統整了歐洲各國在白血病領域之研究成果並制定臨床診療指南。在急性骨髓性白血病（AML）的臨床實務中，ELN 所發佈之治療風險分類標準被廣泛視為國際學界共識，最近一期在 2022 年發布以因應分子診斷技術與基因研究的進展（Döhner et al., 2022）。在其中的一大關鍵即是針對 AML 研究當中具有臨床預後價值與治療指引意義的基因變異提出建議與分級，因此 ELN 所提供之基因清單可視為當前 AML 領域中最具權威性的基因參考標準之一。

在本研究中蒐集並且彙整了 ELN 於 2022 年版之指南中提及之相關基因清單，共中涵蓋 26 個與 AML 臨床表現與治療預後密切相關的基因（表 3-1），基因命名格式參照人類基因命名委員會（HUGO Gene Nomenclature Committee，HGNC）。因此，這份清單在本研究中亦同時扮演「已知標的基因」之角色，可用以驗證本研究所提出之方法是否能夠成功辨識出目前公認具臨床意義的 AML 關聯基因，進而作為模型效能評估之基準指標。儘管 ELN 所提供之清單已涵蓋

doi:10.6342/NTU202504150

多數臨床常見的重要基因，但仍可能有潛在基因未被納入現行臨床標準，我們亦期待本研究之方法能夠幫助找出過往較被忽略之相關基因。

表 3-1：ELN 指南分類 AML 標的基因對照表（基因名稱、全名）

Gene Symbol	Gene Full Name
ABL1	Abelson murine leukemia viral oncogene homolog 1
ASXL1	Additional sex combs like transcriptional regulator 1
BCR	Breakpoint cluster region
BCOR	BCL6 Corepressor
CBFB	Core-binding factor subunit beta
CEBPA	CCAAT enhancer binding protein alpha
DEK	DEK proto-oncogene
EVI1	Ecotropic viral integration site 1
EZH2	Enhancer of zeste 2 polycomb repressive complex 2
GATA2	GATA binding protein 2
KMT2A	Lysine methyltransferase 2A
MECOM	MDS1 and EVI1 complex locus
MLLT3	Myeloid/lymphoid or mixed-lineage leukemia translocated to, 3
NPM1	Nucleophosmin
NUP214	Nucleoporin 214
PML	Promyelocytic leukemia
RARA	Retinoic acid receptor alpha
RUNX1	Runt-related transcription factor 1
RUNX1T1	RUNX1 translocation partner 1
SF3B1	Splicing factor 3b subunit 1
SRSF2	Serine/arginine-rich splicing factor 2
STAG2	Stromal antigen 2
TP53	Tumor protein p53
U2AF1	U2 small nuclear RNA auxiliary factor 1
ZRSR2	Zinc finger CCCH-type, RNA binding motif and serine/arginine rich 2

值得注意的是，本研究主要以 ELN 指南中 Table 1 所列之 26 個基因作為已知標的基因，主要考量在於該清單聚焦於具有高度臨床共識的基因突變，這些

doi:10.6342/NTU202504150

基因被廣泛應用於 AML 的預後分析與治療決策，已具備實務可參考性與明確臨床意涵。然而，其他仍存在多種與 AML 相關的潛在候選基因，部分已被研究指出與疾病機轉相關，但其臨床意義與實際應用價值尚未完全確立。因此，本研究仍選擇優先以 Table 1 所列具高度可信度之基因，作為檢驗模型是否能有效辨識出公認關鍵基因的基礎。未來工作則將可考慮納入相關潛在候選基因進行延伸性分析，進一步拓展本研究方法於新標的基因發掘之應用潛力。

3.1.4 HUGO Gene Nomenclature Committee

為了建立完整且可比較的基因預測分析，本研究除了針對 ELN 建議之 AML 標的基因進行評估之外，亦需建立以全人類基因清單作為背景基因集（Background gene set）。此舉可用來觀察特定基因在預測中的機率分布是否與一般基因有所不同，進而檢驗模型是否能夠有效區分疾病相關基因與背景基因。

本研究採用的全基因清單資料來源為人類基因命名委員會（HUGO Gene Nomenclature Committee，HGNC）所負責維護的 HGNC complete set 資料集。HGNC 是國際人類基因組組織（Human Genome Organisation，HUGO）下轄之基因命名標準機構，專門針對人類基因提供統一的命名與分類系統。該資料集涵蓋了所有經過命名的標準人類蛋白質編碼與非編碼 RNA 基因。

經過初步資料處理與去除重複後，共計獲得 44,031 筆基因符號，構成背景基因候選集。其中大多數為蛋白質編碼基因，但亦包含部分功能性非編碼 RNA 基因。此一背景基因集的建立，使得本研究在分析 ELN 基因預測機率值評估時可作為對照組得以進行統計對照與顯著性檢定，例如利用 U 檢定（Mann-Whitney U test）比較 ELN 清單與非 ELN 基因群體的機率分布差異是否顯著，或者利用成對 t 檢定（Paired t-test）來比較原生 BioGPT 模型與額外整合了 RAG 架構的 BioGPT 模型，進一步評估該模型在此一任務的實際成效與潛力。

doi:10.6342/NTU202504150

3.2 模型架構與流程設計

本研究使用之語言模型為 BioGPT，是由微軟研究院（Microsoft Research）於 2022 年所開發的生物醫學文獻預訓練大型語言模型，旨在提升自然語言處理技術在生物醫學任務中的應用表現。BioGPT 之基礎框架建構於 OpenAI 所提出之 GPT-2 架構之上（Radford et al., 2019），採用「Causal Language Modeling」的自回歸訓練目標，其允許模型依照預訓練資料依序生成下一個最可能出現的 Token。

3.2.1 BioGPT

相較於一般的通用大型語言模型，BioGPT 是以大量的生物醫學文獻摘要進行預訓練，用於生物醫學專業領域的大型語言模型；而其訓練資料來源取自於 PubMed 資料庫中的超過 1500 萬筆學術文獻摘要，其內容涵蓋生物學、醫學、藥理學與臨床相關主題（Luo et al., 2022）。由於這些資料所涵蓋的領域高度專業，因此 BioGPT 相較於同等級的通用大型語言模型更能學習更具專業語境性的醫學術語、疾病名稱、基因縮寫與研究用語，展現出相較一般 GPT-2 更精準的語意理解與生成能力。

一般版的 BioGPT 的模型架構基本遵循了 GPT-2 medium 設計，主要包含 24 層 Transformer，每層具有 1024 維隱藏層（hidden layer）與 16 個注意力頭（attention head），總參數量約為 345M。GPT-2 在許多功能設計上很適合將其改良成專業領域如生物醫學領域專用的模型；首先是在分詞策略上使用了位元組對編碼（Byte-Pair Encoding，BPE）進行分詞，BPE 是一種統計式分詞方法，透過反覆合併在語料中頻繁出現的字元對，進而建立子詞單位（Sennrich et al., 2016）。這種方法能有效解決詞彙爆炸（vocabulary explosion）與未知詞（OOV），並保留生醫領域中大量複合字詞結構（如“cytokine-induced”）。此外，BPE 可動態學習語料中常見片段，使模型能夠更細緻地處理罕見詞與專有名詞，在生醫語境中尤為重要。

doi:10.6342/NTU202504150

然而美中不足的是 BioGPT 在微軟最初進行模型訓練階段所使用的詞彙表（vocabulary）主要是針對 PubMed 文獻摘要文本進行訓練而自動產生，並未刻意納入所有 HGNC 官方基因名稱作為獨立 token。因此，大部分的基因名稱會被切分為多個 subword tokens。而這會對模型在進行 next-token 預測時造成兩項挑戰：首先是模型無法以一個完整語義單位來預測該基因名稱；其次是在進行多 token 的概率累積計算時，容易造成預測結果受到長度偏誤（length bias）影響。此問題亦在 BioGPT 論文中被明確指出（Luo et al., 2022）。因此，在本研究中亦進一步設計了針對 token 基因名稱的正規化策略（詳見第 3.2.4 節），以減少此類偏差對推論結果造成的不利影響。

為促進後續研究應用，微軟亦將 BioGPT 發布至 HuggingFace 開放平台，並提供開源模型權重與原始訓練腳本。本研究亦透過 Huggingface Transformers 套件作為推論用的主要模型，並在保持原始權重設定下執行推論任務，以初步評估其作為零樣本（zero-shot）推理工具於 AML 相關基因預測任務中的表現。Hugging Face 同時也是目前最廣泛應用於自然語言處理研究與開發的開源社群之一，其平台提供了來自各大研究機構與企業之預訓練模型、標準化 API、部署工具與範例程式碼等資源。透過其開源套件可載入由微軟官方發佈之 BioGPT 模型與對應的 tokenizer，並在 PyTorch 環境中進行 next-token prediction 推論與機率分佈計算。由於模型與語言處理工具皆來自可檢驗之開源平台，並在相同框架下進行操作，這將有助於確保實驗流程的可重現性與模型推理的一致性。

3.2.2 Prompt 設計

在本研究中採用了 BioGPT 作為基礎模型，針對不同目標基因計算其在特定疾病語境下作為下一詞（next token）出現的機率值。而在此值得注意的是，BioGPT 為基於 GPT-2 架構的語言模型，其推論階段主要仰賴「hard prompt」策略進行生成，即需透過明確且具語意引導性的字串來控制模型的輸出語境

doi:10.6342/NTU202504150

（Brown et al., 2020）。儘管 BioGPT 原始論文中曾針對下游任務探討使用 soft prompt tuning 技術（例如將可訓練的提示向量嵌入模型輸入）以提升模型在特定任務上的表現，但本研究並未進行額外微調或加入可學習提示，而是直接使用 BioGPT 進行推論。相較於 GPT-3、GPT-4 等經過 instruction tuning 的模型能理解自然語言任務說明，BioGPT 並不具備此類指令追隨能力，輸出結果高度依賴於提示語的結構與語境設計。因此，本研究特別著重於相關疾病語境提示語（disease-context prompt）的語句結構設計，期望透過精心構造的句型來有效誘導模型學習疾病與基因之間的潛在語義關聯，進而提升模型在基因辨識任務中的生成準確性與選擇性。因此精心設計能夠引導模型的 prompt 也是本研究的關鍵步驟之一。我們依據 BioGPT 文獻與語言模型應用慣例（Luo et al., 2022），設計了三百多組符合生物醫學語境的疾病相關 prompts，並且透過預測機率值的前一千名基因名稱出現數來對相關性進行衡量，最終得知其中與 AML 相關的最主要的 prompt 結構大致為：

"Human gene that related to Acute Myeloid Leukemia is the"

此類 prompt 的設計邏輯多來自於生物醫學文獻中常見的疾病與基因變異關聯描述句式，透過使用常見描述句型中的語義結構，能夠有效地引導 BioGPT 模型預測在該語境下最有可能出現的基因名稱，進而進行基因與疾病語義關聯的評估。在初步比較中，我們觀察到不同 prompt 設計會對機率分布結果造成一定影響；在針對句型構造進行測試以後我們發現平均來說在相似結構的 prompt 中，該 prompt 的表現皆是同類型 prompt 的最低下限值，因此我們最終選定上述 prompt 作為主要分析用的基礎結構；作為同類型 prompt 的下限，既能夠保證同類型的 prompt 之表現將會不差於此 prompt 的結果，以確保模型預測結果的穩定性與可解釋性，同時能在此標準之上進行局部詞語的變化來觀察 prompt

doi:10.6342/NTU202504150

的所導致的結果差異，並以機率值分布作為評估對象。最終所觀察到其實在相似的語句結構下的預測機率分布有極高的相似性。。

除了人類語言的條理之外，我們亦確認所設計的 prompt 結構符合 GPT 類的因果語言模型特性：即句子為單向（left-to-right）生成語境，以“gene”和“Acute Myeloid Leukemia”為句子核心，並以“is the”或“such as”作為句末語意引導，使模型預期下一詞為特定專有名詞，從而更明確地觸發模型在基因名詞上的預測結果。

3.2.3 Next-token prediction

在本研究中，我們採用了因果語言模型（Causal Language Model, CLM）的預測機制來推估「目標基因名稱」出現在給定疾病語境下的語意關聯性。具體而言，我們利用 BioGPT 的「下一詞預測」（Next-token prediction）能力，根據疾病相關 prompt，估計給定基因名稱作為下一個詞出現的機率分布（圖 3-2）。

此時最先會遇到的挑戰是，多數人類基因名稱無法對應至單一 vocabulary token，而是會被拆解為數個 subword tokens，而這時候 BioGPT 採用自回歸生成（autoregressive generation）架構，其核心任務是：根據當前輸入的 prompt 與 token 序列，預測下一個 token 的機率值分布。也因此我們能夠針對每一個基因名稱設計一套完整的推論流程。首先是輸入設計好疾病相關的 prompt 如：「AML is driven by mutations in genes such as」，而此時這段 prompt 會被轉換為若干 tokens，並在輸入 BioGPT 後得到一系列的 tokens 的機率分布；此時先篩選出所有可能構成基因名稱的 tokens 作為下一輪接續輸入的選項，緊接著便可以進入下一輪。

舉例來說，當我們希望預測的目標為 *NPM1*，在輸入 prompt 以後，找出 *NPM1* 拆解成的 subword token 字首 N 作為下一輪的序列尾端 token，此時便可以開始第二輪輸入，這時我們的 prompt 會變成「AML is driven by mutations in

doi:10.6342/NTU202504150

genes such as N」，接著再找出第二輪預測機率分布中，*NPM1* 的第二個 subword token 的 PM 在做為下一輪序列尾端 token，如此反覆直到完整的 *NPM1* 完成，最後將每一輪 token 條件機率值做乘積便可以得到其完整的機率值。

而實驗階段時，本研究使用 transformers 套件中的 BioGptForCausalLM 與 AutoTokenizer 進行推論自動 token 化與生成分數（logits）獲取。其中為確保 token-by-token 的 logit 溯源正確，會在每輪計算中儲存當下 token 的 logits 並以 softmax 函數轉換為機率值。

因為多數基因名稱長度不一且通常包含多個 tokens，因此針對多 token 基因採「token 累積機率乘積」方式推估整體機率，這使得語言模型預測每一 token 的機率時，需經歷許多輪的自回歸生成，並對每一步驟的條件機率進行累積。這樣的設計雖然保留了模型處理未知字詞（out-of-vocabulary, OOV）能力的彈性，但同時也帶來了推論上的挑戰：基因名稱越長、被拆解成越多 token，累積機率值往往會隨之下降，即便該名稱與語境高度相關也可能因長度而被低估。這也是在後續中引入正規化策略對此累積機率進行長度或 token 數校正，進一步減少長字詞偏差的影響的主要動機。此方法參考了 BioGPT 原始論文中的生成任務設計（Gu et al., 2022），並與近年常見的 next-token prediction 分析方式一致（Brown et al., 2020；Radford et al., 2018）。

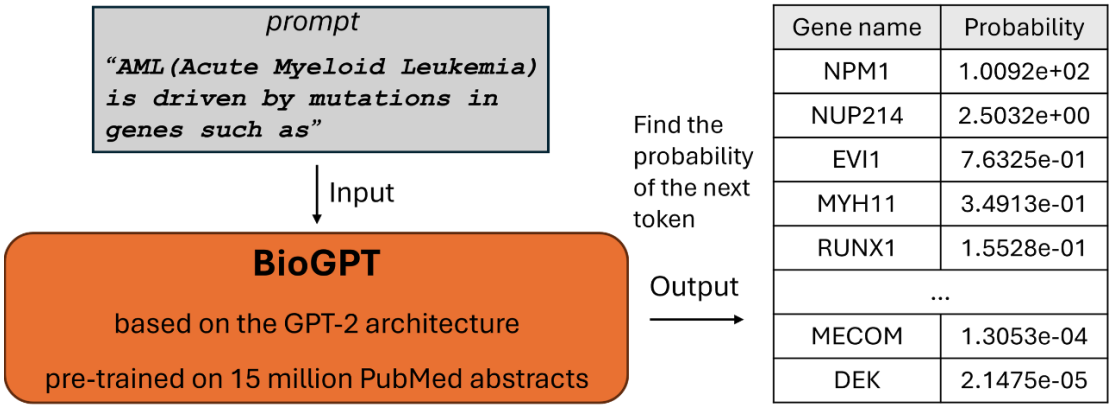

圖 3-2：基於 BioGPT 的 Next-token prediction 基本架構

doi:10.6342/NTU202504150

3.2.4 正規化策略

在使用 BioGPT 進行基因名稱的預測任務中，其中一種正規化策略便是在每一輪以自回歸方式逐一預測每一個 token 的出現機率時以正則算法去除掉非基因構成 token，並以所有基因構成 token 之機率值為條件機率值。最終，完整基因名稱的生成機率為所有 token 條件機率的乘積。這種 token-by-token 的乘積邏輯雖然合理，但也同時會導致基因名稱愈長者，受其影響總體生成機率會下降，從而很有可能對高度相關但組成所需 token 數量較多的基因名稱造成低估。也因此，本研究設計了兩種專門的正規化策略，試圖調整此偏差，使模型對於不同長度的基因名稱仍能進行公平且準確的語意機率比較（圖 3-3）。

首先作為最基本的一種正規化策略便是上述將每一輪 token 預測時基因名稱 token 條件機率值相加再平均，此方法為 Per-token Total Sum Normalization，（Zhavoronkov et al., 2023）其數學表示如下：

$$S = \frac{1}{n}\sum_{i=1}^{n} \log P(t_i \mid t_{<i})$$

其中t_1,t_2,...,t_n為基因名稱拆解成的 token 序列，$P(t_i \mid t_{<i})$表示第 i 個 token 在先前的 token 都添加至 prompt 條件下的機率值。此方法能夠消除 token 數多寡造成的總機率落差，使預測值更能反映語意契合程度，近似於 log-likelihood 的平均形式，是一種語言模型常用的 normalization 策略（Radford et al., 2018）。而另一方向則是針對總 token 數量的「長度」進行「後處理調整」，可以被稱之為 Length-based Normalization 的方法：具體來說便是將原始機率除以 token 數量的「長度」。在觀察各方法在區分 ELN 標的基因與背景基因（HGNC 其餘基因）時的預測機率分布情形與統計檢定表現。我們透過 Mann-Whitney U 檢定與 Paired t-test 來檢視其顯著性，結果顯示我們的最佳化策略確實能有效修正長度偏差。需要注意的是，本研究在推論過程中使用了條件機率計算與總 token 正規化，實際上可能已偏離真正的機率詮釋範疇，因此將其視為一種語言模型分

doi:10.6342/NTU202504150

數（Language Model Score），可作為基因名稱在特定語境提示下的語言合理性與生成傾向的量化指標。雖然該分數在數值上可能超過 1，但在排序與比較層面具有高度可比性與解釋力。

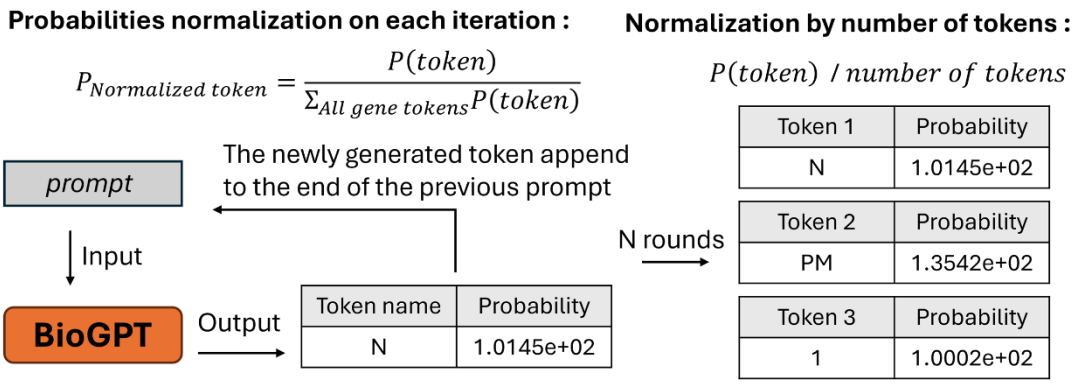

圖 3-3：基因名稱逐 token 預測流程圖。模型接受包含疾病語境的 prompt，逐 token 預測基因名稱組成之每個子詞的條件機率，最終以 token 機率乘積計算整體生成機率，最後再進一步套用正規化策略以校正長度偏差。

3.2.5 運算資源與實作環境

本研究所有程式與模型推論流程皆於具備 GPU 加速能力之本地伺服器環境下執行。主要運算平台使用 NVIDIA GeForce RTX 4090 GPU，搭配 24GB 顯示記憶體，足以支援 BioGPT 模型之高效推論與多組 Prompt 同時運算需求。程式開發環境為 Python 3.10，搭配 Huggingface Transformers 套件進行模型載入與生成操作，並輔以 LangChain 框架實作檢索增強生成（RAG）功能。向量資料庫方面則採用 FAISS 庫作為語意檢索後端，確保具備高效率之文本片段查詢能力。所有模型推論作業皆於 GPU 環境中完成，最大化推論速度與批次處理效能。此外，為提升分析流程可重現性與模組化，本研究亦將主要實驗腳本進行模組化封裝，利於後續參數調整與不同模型架構之替換測試。

doi:10.6342/NTU202504150

3.3 檢索增強生成

在以 BioGPT 為基礎之基因預測任務中，我們透過 prompt 工程設計方式，並且嘗試讓模型從疾病語境中推論出特定基因名稱。然而，BioGPT 雖然在大量生物醫學文獻上進行了預訓練，但在實際推論時，仍會因為來自於諸多不同領域的文獻影響；如果期待預測結果能夠更加集中在目標疾病，引入近年來熱門的檢索增強生成（Retrieval-Augmented Generation, RAG）架構便是一項值得嘗試的選項。

3.3.1 設計動機

除了 BioGPT 預訓練資料的領域跨度極廣，另一方面在實際推論時，其輸入仍然受限於簡短的 prompt，這可能在許多情況下會無法提供足夠的上下文來支撐模型對罕見或新興基因名稱的準確預測。尤其是在面對與 AML 相關的基因複雜關聯性時，單靠模型自身的語言記憶能力仍可能有所不足。此外，在實驗過程中可以觀察到模型在處理部分目標基因名稱時的預測值較差，可能與其訓練語料中單獨該基因與疾病共同出現的語言樣本有限有關。針對這一限制，若能在推論過程中動態補充與疾病相關的上下文資訊，則有望提升模型在這些案例中的推理能力。

為此，本研究引入 RAG 技術，作為 BioGPT 原始生成架構的資料補強機制。此技術旨在結合外部知識來源形成語料庫，在每次推論前都會先根據輸入基因名稱在其語料庫中檢索出與其最相關的文獻內容，緊接著並將其插入基礎的 prompt 中，讓模型能夠在更豐富的語意上下文中進行生成。而這樣的設計在基因名稱預測應用中具有高度應用價值：首先，它不需要修改模型權重，便可在推論階段引入外部補充知識；其次，其動態檢索設計對於處理基因名稱的預測時具有彈性與可擴展性；最後，透過段落級別的文獻引導，能有效減少因領域過於繁雜而造成的預測失焦，進一步提升 BioGPT 在疾病特定語境下的語言支持度評分準確性。

doi:10.6342/NTU202504150

3.3.2 RAG 語料庫資料蒐集與預處理

為實現 RAG 所提供的語境補強，本研究為此設計了一套完整的文獻資料擷取與段落前處理流程。首先，資料來源同樣選自 PubMed 生物醫學文獻資料庫，我們針對"acute myeloid leukemia"主要查詢關鍵字篩選文獻標題或摘要，設計分年範圍查詢腳本，確保每次檢索結果皆低於 Entrez API 所允許的 9,999 筆爬取上限（圖 3-4）。

緊接著，在資料過濾方面，為提升語境資料的純度與疾病針對性，本研究排除包含其他血癌類別（CML, Chronic Myeloid Leukemia）等非目標疾病之摘要，並僅保留英文文獻。同時設定年份範圍為 1990 至 2025 年，確保文獻具有一定的時效性。

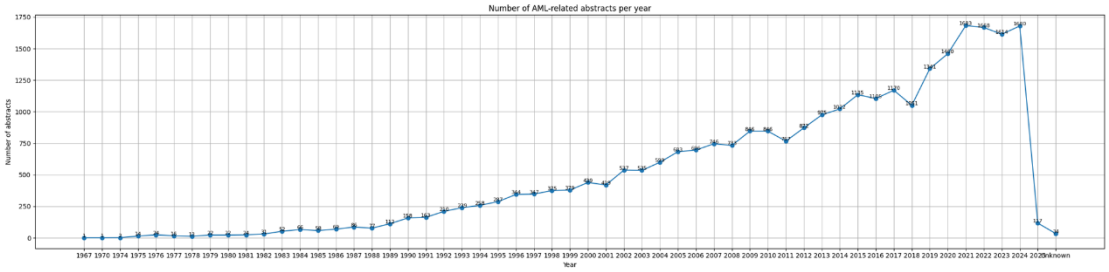

圖 3-4：以"acute myeloid leukemia"作為主要查詢關鍵字之各年份文獻數量

需要特別說明的是，本研究並非直接將每一篇 PubMed 文獻的整體摘要作為 RAG 語料來源使用，而是將其切分成較短的段落（chunks）。這樣做的主要原因在於：首先，BioGPT 在推論時有輸入長度（token limit）限制，若輸入段落過長，不僅容易截斷資訊，也會導致模型難以聚焦在具體語境中進行下一詞預測。其次，較短的段落能夠提升語意檢索階段的解析度，使得每一段文本所承載的語意更集中、而語境更清晰。在資料段落上，本研究採用 LangChain 工

doi:10.6342/NTU202504150

具中提供的 RecursiveCharacterTextSplitter 進行切分，初步設定為 chunk size = 300，chunk overlap = 20，希望能兼顧語意完整性與嵌入效率，確保段落向量表示時仍能保留疾病與基因之間的語意關係，且避免片段過長造成語言模型輸入冗餘或資訊稀釋。

因此進一步為了釐清 chunk size 對整體檢索與推論表現的影響，我們進一步設計了多組段落切分長度設定，包含 50、100、150、200、250、300、350、400、450 與 500 tokens 等不同切分長度，並分別執行全文向量建構與 BioGPT 預測評估。結果顯示，chunk size 對預測效果確實具有影響：以本研究的極短 token 任務來說，較短的段落能提供更細緻的語義對應，但亦可能導致語境過於破碎；而較長的段落則通常能涵蓋更多背景資訊，但也可能較容易被無關資訊稀釋。這些實作結果後續將於第四章的結果分析中詳細比較。

3.3.3 嵌入模型

在 Retrieval-Augmented Generation, RAG 架構中，如何將大量文獻段落轉換為語意向量並構建高效的檢索引擎，是模型語境補強能否成功的關鍵之一。本研究設計並實作了一套基於向量語意檢索的模組，並在多種嵌入（embedding）模型之間進行實測與比較，最終實驗後選定最適用於 AML 文獻語境提示的語意嵌入模型與資料庫配置。

為找出最適合本研究語境的嵌入模型，我們評估了四種代表性 biomedical 或通用語意嵌入模型，其分別為：

S-PubMedBert-MS-MARCO：基於 PubMedBERT 預訓練語言模型，進一步使用 MS-MARCO 語料（以問答任務為主）進行微調，強化語意比對能力，特別適合用於語境相關段落檢索。

biobert-base-msmarco：以 BioBERT 為基礎、同樣在 MS-MARCO 上微調，為 biomedical domain 的嵌入模型，擅長處理生物醫學專有名詞與文獻語境。

doi:10.6342/NTU202504150

pubmedbert-base-embeddings-matryoshka：是由結合了 PubMedBERT 以及 Matryoshka embedding 框架之產物，具備多尺度語意壓縮能力，可應用於語義密集型資料檢索。

all-MiniLM-L6-v2：為 Sentence-Transformers 提供的通用語意嵌入模型，效能佳、速度快，雖非生醫領域專用，但通常作為 baseline 模型可提供良好比較基準。

本研究將這些模型分別應用於相同的段落資料集，產生語意向量，並在推論流程中評估其語境補強能力與生成效果。在選定嵌入模型後，所有經過預處理與段落切分的 PubMed 生物醫學文獻摘要段落將傳入上述模型，並儲存為索引資料庫。為了實作語意檢索的功能，本研究進一步整合了 LangChain 框架中 FAISS 向量資料庫，作為儲存與查詢語意向量的核心引擎。FAISS（Facebook AI Similarity Search）為一款相似度搜尋工具，廣泛應用於向量化資料的索引與最近鄰查詢（nearest neighbor search），可支援大型嵌入資料集的快速檢索。

在 Retriever 模組的實作上，本研究採用 LangChain 的 VectorstoreRetriever 物件，並設定為使用向量內積相似度（dot product similarity）作為檢索指標。在實際推論過程中，每次針對基因名稱生成其語意向量後，即透過 Retriever 查詢向量資料庫中最相似的段落，並取回 Top-k 筆（預設為 k=5）最相關段落做為語境補強內容。LangChain 框架使得這一整體流程能以模組化方式實作，並可結合多種 LLM 推理工作流程進行串接。此種基於嵌入空間相似度的檢索策略，不僅能提升語境相關性的品質，也具備良好的擴充性與效能表現，成為本研究中語境補強設計的重要一環。

3.3.4 與 BioGPT 推論流程整合

在完成 PubMed 文獻檢索爬取、段落切分、語料庫建構與語境補強之後，最終的關鍵步驟即是如何將這些外部檢索得到的資訊，整合進 BioGPT 的推論

doi:10.6342/NTU202504150

流程中，並完成基因名稱的 token-by-token 預測與分數計算。一般會根據基因名稱進行嵌入向量查詢，返回最相關的多段文獻段落。接著這些段落會以換行符號串接，再與 base prompt 結合，構成完整的輸入 prompt。之後輸入 BioGPT 的機率預測函式，進行每一個 token 的機率計算與最終正規化。（圖 3-5）

整合後的推論流程中，BioGPT 模型仍維持其原有的 token-by-token 機率預測邏輯（詳見 3.2.3 與 3.2.4）。唯一的變化來自於前置 prompt 的擴充，使得模型在每一輪生成下個 token 時，能參考更多來自真實文獻的語境，進而提升基因名稱 token 出現的預測機率。這樣的整合方式可確保推論流程維持原本設計邏輯與結構，同時也能讓模型在每筆輸出結果上反映出 RAG 所帶來的具體差異。

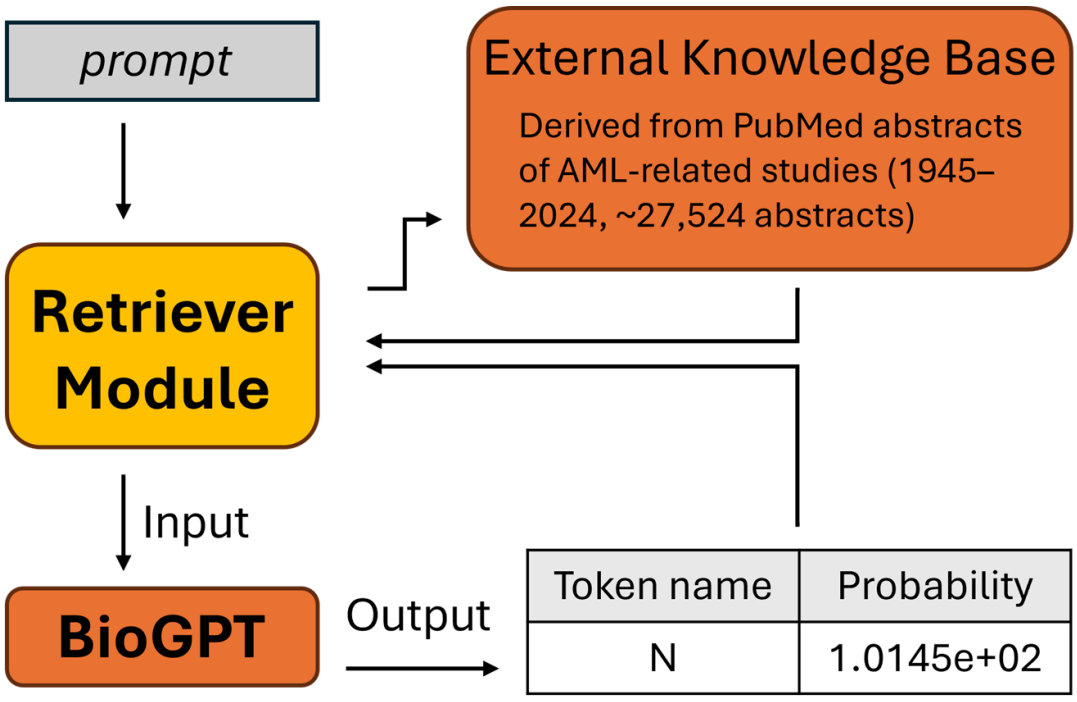

圖 3-5：BioGPT 整合 RAG 後的基因名稱機率值推論結構

doi:10.6342/NTU202504150

第四章 結果與討論

4.1 Prompt 設計與生成分析

為探討 prompt 設計對 BioGPT 模型預測結果的影響，本研究根據測試結果使用了具生物醫學語境的句型，皆以「AML（Acute Myeloid Leukemia）」為主題，並以相似的句型架構，引導模型預測潛在相關的基因名稱。圖 4-1 呈現了在這些 prompt 下，針對 ELN 基因名單中 26 個基因的正規化機率預測值所繪製的熱圖（heatmap），並以對數尺度（log10）進行視覺化；而圖 4-2 則展示了相同資料下的盒鬚圖，進一步比較不同 prompt 對 ELN 基因的機率預測值分佈。

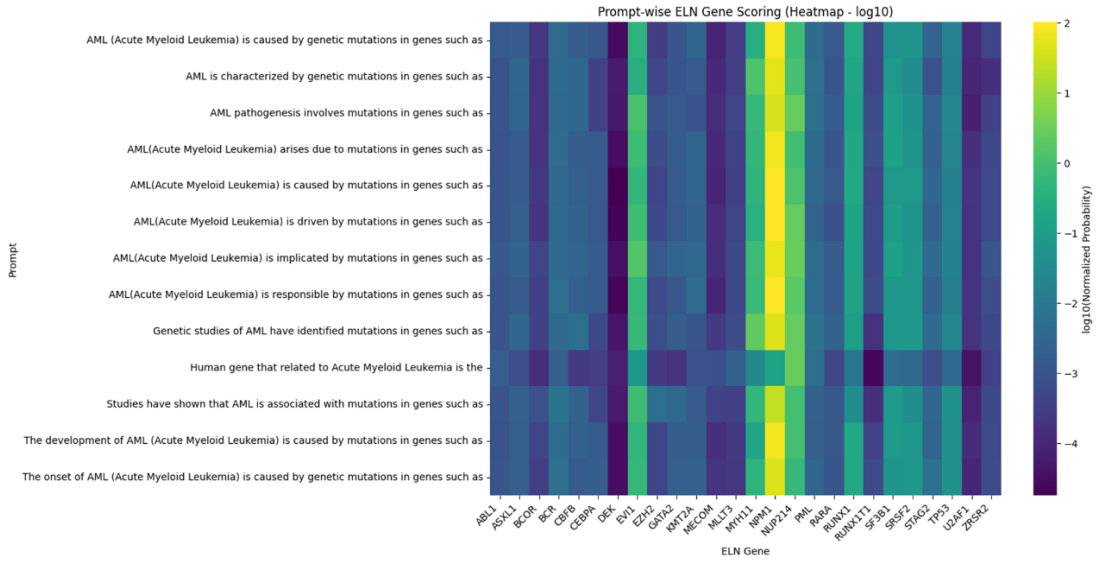

圖 4-1：相似句型結構的不同 prompt 於 ELN 基因預測值之熱圖

從圖 4-1 可觀察到，雖然大部分 prompt 在特定基因（如 *NPM1*、*RUNX1*、*TP53* 等）上皆表現出高度的一致（機率值偏高或偏低），但不同 prompt 所產生的分布仍存在細微差異。而這些相似的句型在大多數基因上皆展現出較為一致的機率，顯示其能有效引導模型產生具疾病相關性的語言生成。另一方面，圖

doi:10.6342/NTU202504150

4-2 的盒鬚圖則從統計分佈角度進一步補充觀察，顯示出 prompt 對於整體 ELN 清單基因的機率預測值中位數與離群值情形。

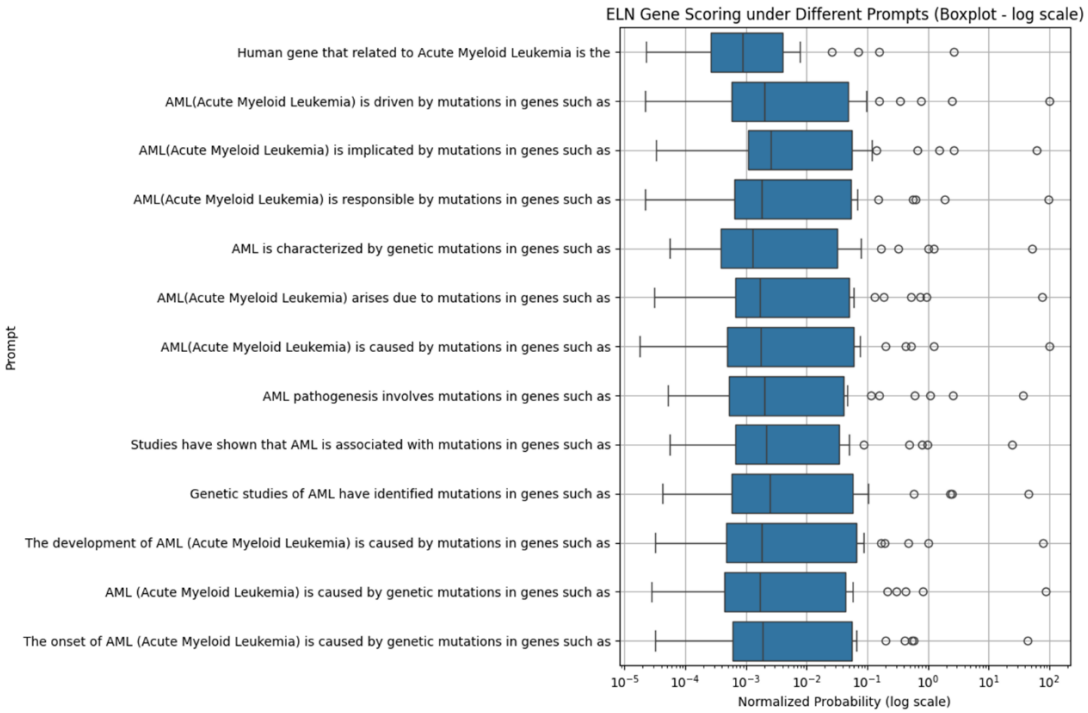

圖 4-2：相似句型結構的不同 prompt 於 ELN 基因的預測值之盒鬚圖

由圖中可知，相同句型但個別用字不同之間的仍具有一定略微的差異，例如「Human gene that related to Acute Myeloid Leukemia is the」的分佈較窄、數值整體偏低，而句型如「AML(Acute Myeloid Leukemia) is driven by mutations in genes such as」與「AML(Acute Myeloid Leukemia) is implicated by mutations in genes such as」則在高分數區域出現更多離群值（outliers），顯示這些 prompt 更容易觸發模特定基因名稱的生成。

整體而言，本節實驗結果表示 prompt 設計對 BioGPT 模型輸出確實具有影響，不同句型結構可能會改變模型對基因名稱生成的傾向與強度。

doi:10.6342/NTU202504150

4.2 Prompt 設計與生成分析

4.2.1 ELN 標的基因與背景基因之比較

為了驗證本研究在 BioGPT 模型上設計的流程對於 ELN 標的基因是否展現出更好的預測能力，本研究將 ELN 標的基因與其他全體背景基因進行對比分析（以下結果皆以對數轉換後的 $\log_{10}$（normalized probability）作為分析主軸）。

首先，圖 4-3 呈現 ELN 標的基因與背景基因的 $\log_{10}$預測值的盒鬚圖。從圖中可以觀察到，ELN 基因的中位數顯著高於背景基因，代表語言模型更傾向於生成 ELN 基因。

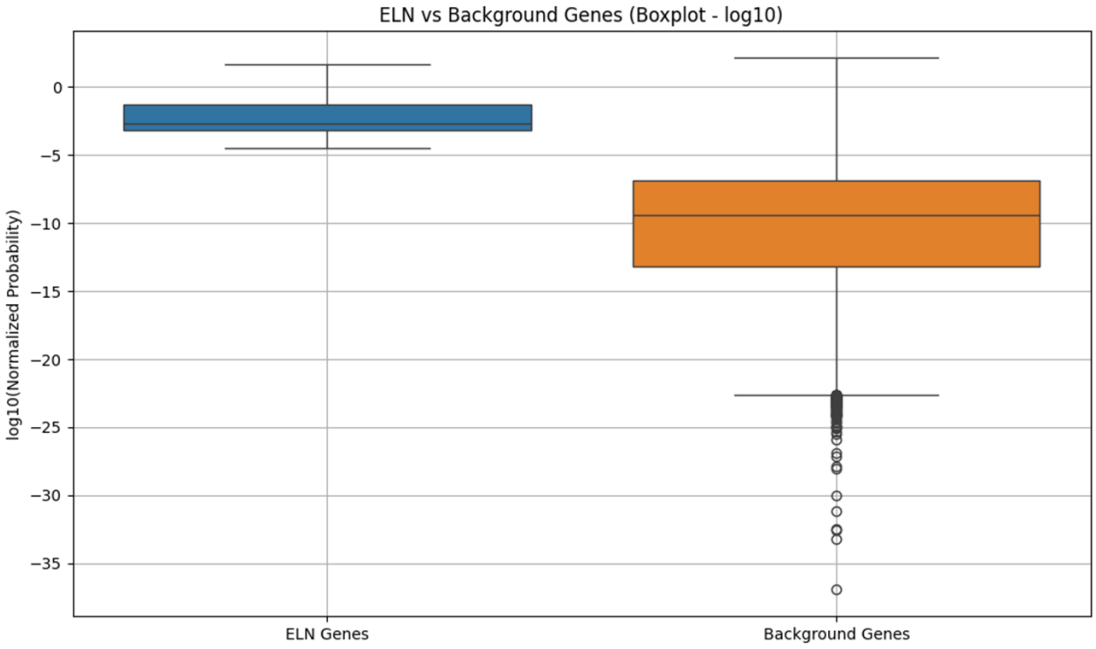

圖 4-3：ELN 標的基因與背景基因的預測值分布之盒鬚圖（$\log_{10}$）

而為了驗證上述兩群基因差異的統計顯著性，本研究進一步進行 Mann-Whitney U 檢定，檢定結果如下：

$$U = 1134729，p = 2.0931 \times 10^{-18}$$

doi:10.6342/NTU202504150

此結果顯示 ELN 標的基因的在本流程中的預測值在統計上顯著高於背景基因，p 值遠低於 0.01，顯示差異具高度顯著性。

此外，由於基因名稱長度（token 數量）可能會影響實際生成的機率值，因此圖 4-4 顯示了基因名稱 token 數量與對應預測值（$\log_{10}$）之間的散佈圖。可觀察到雖然背景基因在 token 數增加時出現明顯的長尾現象，但 ELN 基因大多集中在較高區域，顯示其受 token 數影響較小。

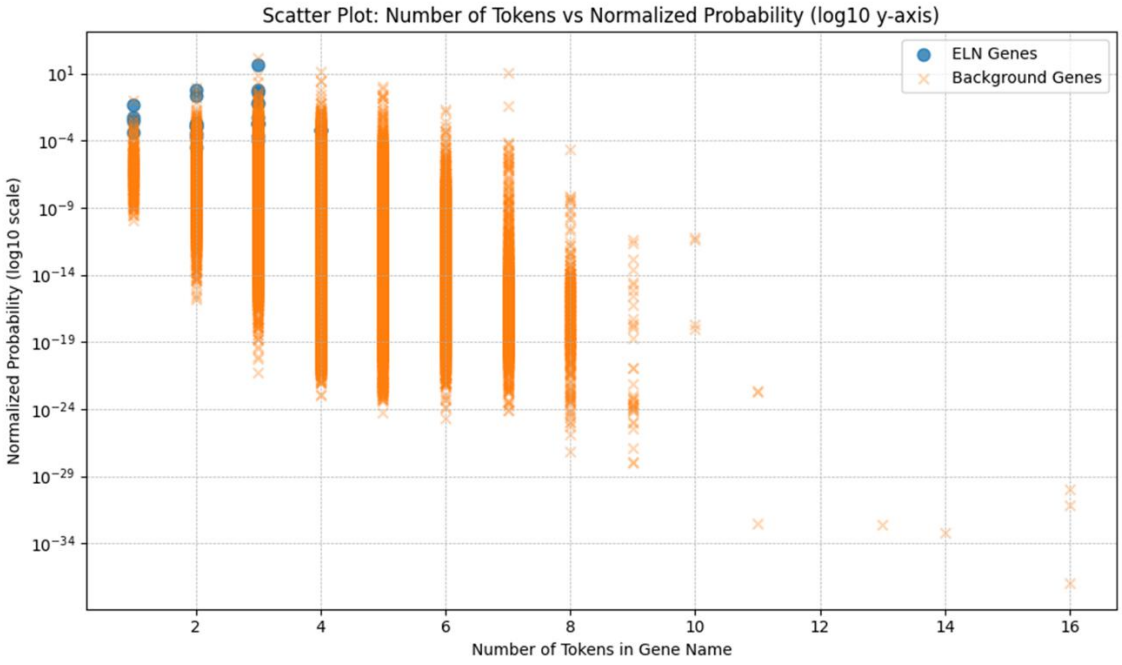

圖 4-4：基因名稱長度（Token）與預測值之散佈圖

而為了進一步觀察在 BioGPT 模型上設計的流程對 ELN 標的基因的排序能力，本研究將 ELN 標的基因合併入全體背景基因，並統計了由預測值由高到低排序後的前 N 名（Top-N）名單中出現的 ELN 基因數量，最終結果如圖 4-5 與表 4-1 所示。

doi:10.6342/NTU202504150

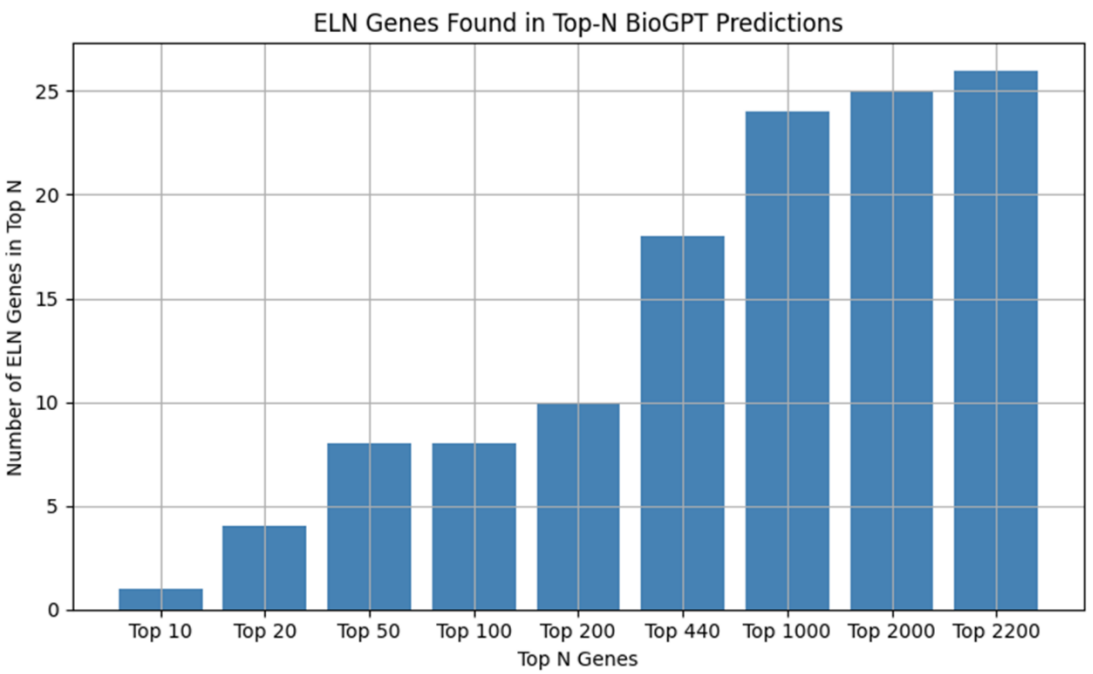

圖 4-5：Top-N 預測值中出現的 ELN 基因數量統計圖

表 4-1：Top-N 與對應出現的 ELN 基因數

Top-N	ELN 基因數	新增出現基因
10	1	*NPM1*
20	4	*NUP214, EVI1, MYH11*
50	8	*RUNX1, SRSF2, SF3B1, TP53*
100	8	
200	10	*STAG2, BCR*
440	18	*PML, KMT2A, GATA2, ASXL1, CEBPA, CBFB, ABL1, RARA*
1000	24	*ZRSR2, RUNX1T1, EZH2, BCOR, MLLT3, MECOM*
2000	25	*U2AF1*
2200	26	*DEK*

doi:10.6342/NTU202504150

在 ELN 標的基因合併至背景基因共計 44,327 個基因當中，可以看出本研究所透過 BioGPT 針對 AML 相關基因所設計的流程預測成效；其中有超過半數的 ELN 標的基因的預測值排名在整體基因中前 1 %，而基本所有的 ELN 標的基因透過本流程的排序，在整體基因的前 5 %都能被發現。顯然 ELN 標的基因的語言模型預測值要遠大於背景基因，並在 Top-N 前排中呈現強烈聚集現象，說明了大型語言模型具備能自海量研究文獻中挖掘並識別 AML 關聯基因的潛力。

4.2.2 ELN 標的基因與非 AML 基因之比較

為了更進一步評估本研究的模型是否能有效區分 AML 相關基因與非相關基因，我們進一步將 ELN 標的基因與一組自全體背景基因中隨機抽樣並一一查核目前已知與 AML 幾乎沒有關聯的基因（Non-AML genes）進行對比。這組基因（共 26 個）不屬於 ELN 標註，也不在 AML 的文獻或標的清單中，其完整清單可見本小節表格（表 4-2）。

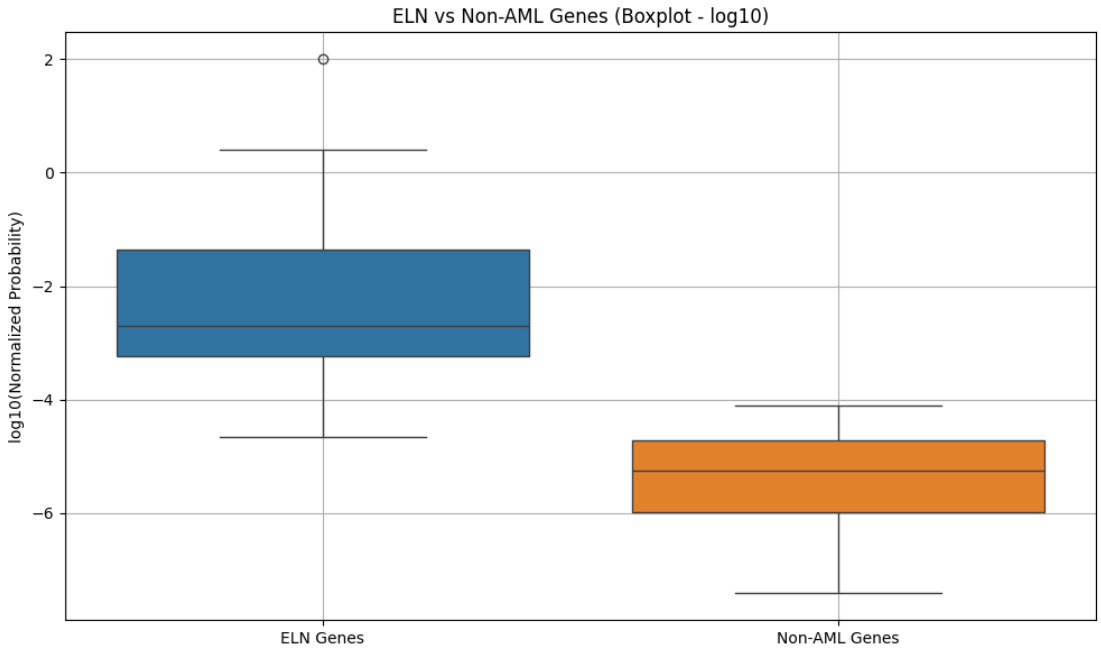

圖 4-6：ELN 標的基因與非 AML 基因的預測值分布之盒鬚圖（$\log_{10}$）

doi:10.6342/NTU202504150

而在圖 4-6 中，我們以盒鬚圖呈現 ELN 基因與 Non-AML 基因的預測值分布情形。從圖中可明顯觀察到，ELN 基因在整體上較高，其中中位數也遠高於 Non-AML 基因，足以顯示本研究的模型確實對於 ELN 基因具有較明確的萃取能力。為了驗證這個差異確實具有統計顯著性，我們同樣進行 Mann-Whitney U Test，檢定結果如下：

$$U = 669.0000, p = 7.3081e\text{-}10$$

此極低的 p 值（< 1e-9）顯示兩組基因的語言模型支持度分布存在顯著差異，進一步支持本研究方法在從大量生物醫學文獻中萃取 AML 基因的潛在價值。

doi:10.6342/NTU202504150

表 4-2：隨機自背景基因中選取的與 AML 相關性低的 26 個基因，為常見的 housekeeping genes 或非癌症相關基因以作為代表性的對照組。

Gene Symbol	Gene Full Name
TUBB	Tubulin Beta Class I
EEF1A1	Eukaryotic Translation Elongation Factor 1 Alpha 1
YWHAZ	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Zeta
ACTG1	Actin Gamma 1
HMBS	Hydroxymethylbilane Synthase
TFRC	Transferrin Receptor
LDHA	Lactate Dehydrogenase A
PGK1	Phosphoglycerate Kinase 1
TPI1	Triosephosphate Isomerase 1
VIM	Vimentin
YWHAB	Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Beta
ATP5F1A	ATP Synthase F1 Subunit Alpha
CYC1	Cytochrome c1
HSPA5	Heat Shock Protein Family A (Hsp70) Member 5
IDH3A	Isocitrate Dehydrogenase 3 (NAD+) Alpha
MRPL19	Mitochondrial Ribosomal Protein L19
NDUFA1	NADH:Ubiquinone Oxidoreductase Subunit A1
SNRPD3	Small Nuclear Ribonucleoprotein D3 Polypeptide
TUBA1B	Tubulin Alpha 1b
UBB	Ubiquitin B
RPLP0	Ribosomal Protein Lateral Stalk Subunit P0
RPL13A	Ribosomal Protein L13a
UBC	Ubiquitin C
GUSB	Glucuronidase Beta
ALB	Albumin
RPL10A	Ribosomal Protein L10a

doi:10.6342/NTU202504150

4.3 檢索增強生成模型微調與影響分析

4.3.1 不同嵌入模型的檢索增強生成效果比較

為了探討不同嵌入模型（embedding model）在檢索增強生成（Retrieval-Augmented Generation, RAG）框架下對 BioGPT 基因機率預測的影響，本研究固定 prompt（Human gene that related to Acute Myeloid Leukemia is the），並針對四種語意嵌入模型進行比較，包括：

S-PubMedBert-MS-MARCO

biobert-base-msmarco

pubmedbert-base-embeddings-matryoshka

all-MiniLM-L6-v2

每種模型皆在相同的 prompt 設計與檢索策略下，對 ELN 標的基因進行機率預測並且與原生 BioGPT 的預測結果進行比較。從各圖中顯示 ELN 各標的基因在未使用與使用 RAG 情況下的變化趨勢線，圖中可觀察到不同嵌入模型對基因分數上升幅度有明顯差異，每條線表示一個基因在兩種條件下的預測值變化趨勢。（圖 4-7, 圖 4-8, 圖 4-9, 圖 4-10）。從視覺趨勢來看，S-PubMedBert-MS-MARCO 在使用 RAG 後，雖然有單獨基因 *ZRSR2* 略為下降，但多數基因的機率分數有顯著上升，尤以 *NPM1*、*MYH11*、*NUP214* 等 AML 常見基因為代表。而至於在 biobert-base-msmarco 雖有類似趨勢，但整體結果較發散。就相對而言，嵌入模型 pubmedbert-matryoshka 與 MiniLM 的提升效果則較不顯著，甚至有部分基因（如 *ZRSR2*、*DEK*）呈現大幅下降。

圖 4-11 則彙整了四種嵌入模型在對 ELN 基因的預測值中位數，其中顯示了 S-PubMedBert-MS-MARCO 遙遙領先，凸顯其在本研究中對 BioGPT 的檢索支持提供最大化的幫助。而為了更具體量化差異，本研究對各模型在使用與不使用 RAG 時的基因分數進行成對 t 檢定（paired t-test），而最終結果則如表 4-3 所示。

doi:10.6342/NTU202504150

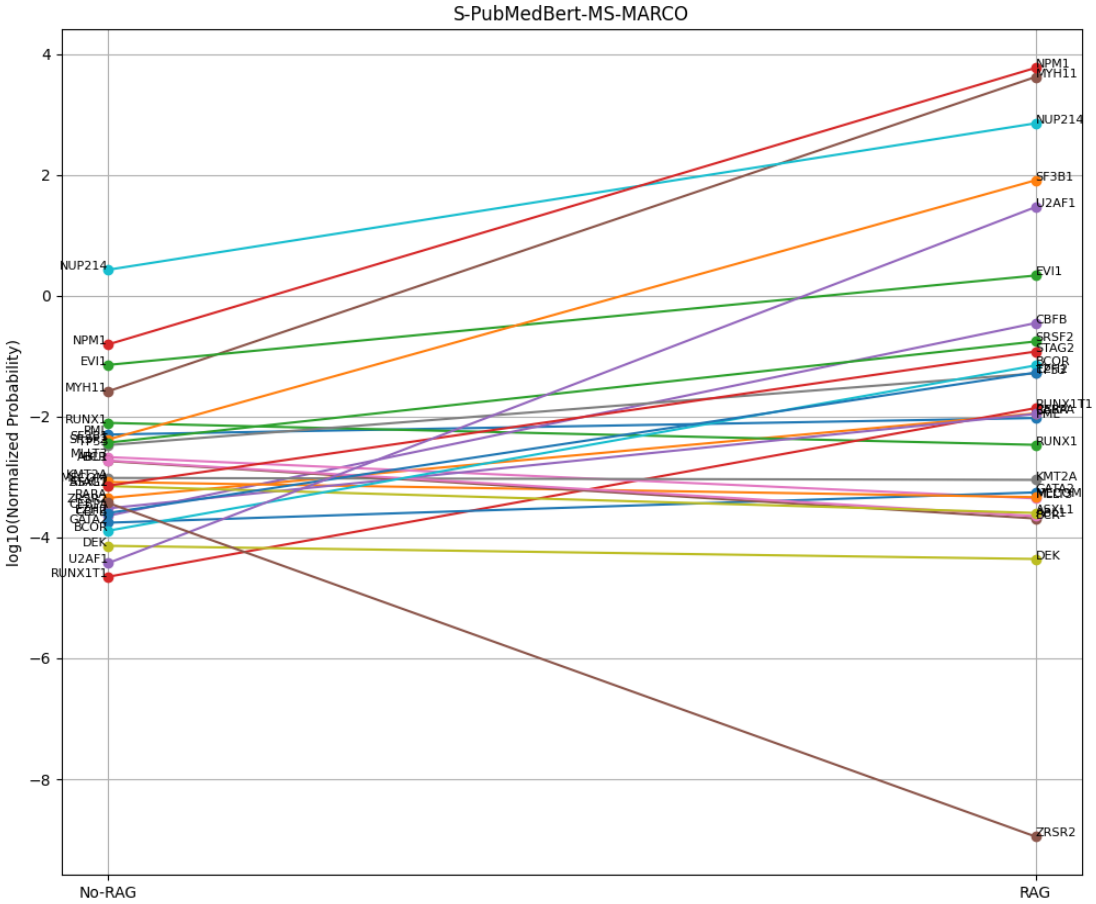

圖 4-7：S-PubMedBert-MS-MARCO 嵌入模型對 ELN 基因 RAG 使用前後變化

doi:10.6342/NTU202504150

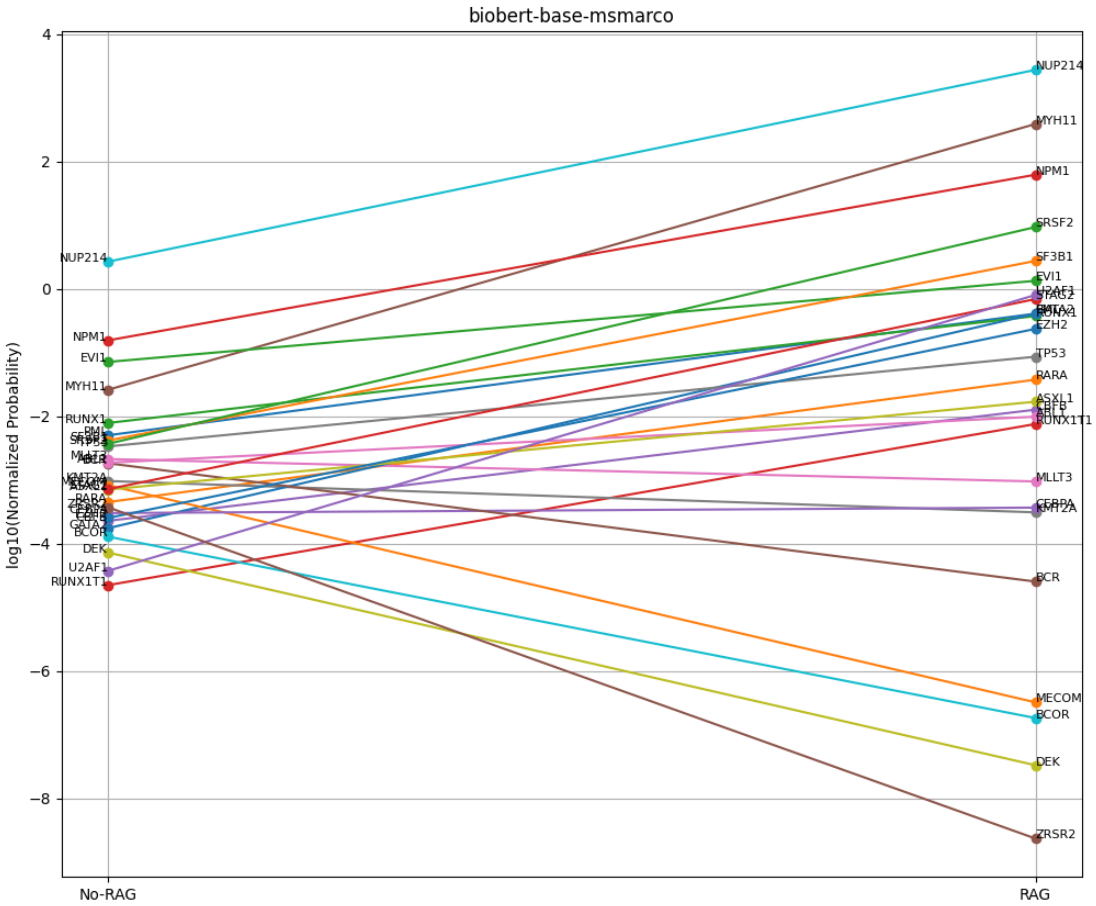

圖 4-8：biobert-base-msmarco 嵌入模型對 ELN 基因 RAG 使用前後變化

doi:10.6342/NTU202504150

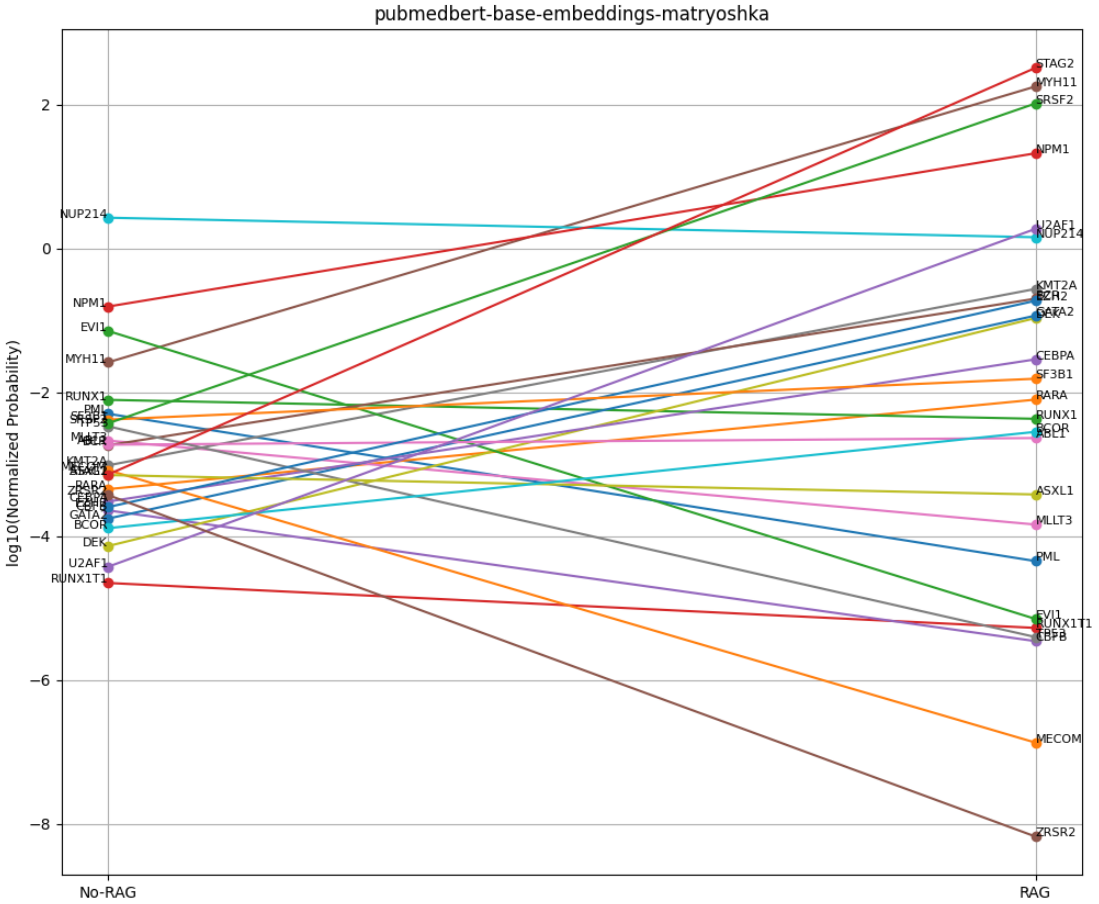

圖 4-9：pubmedbert-base-embeddings-matryoshka 嵌入模型對 ELN 基因 RAG 使用前後變化

doi:10.6342/NTU202504150

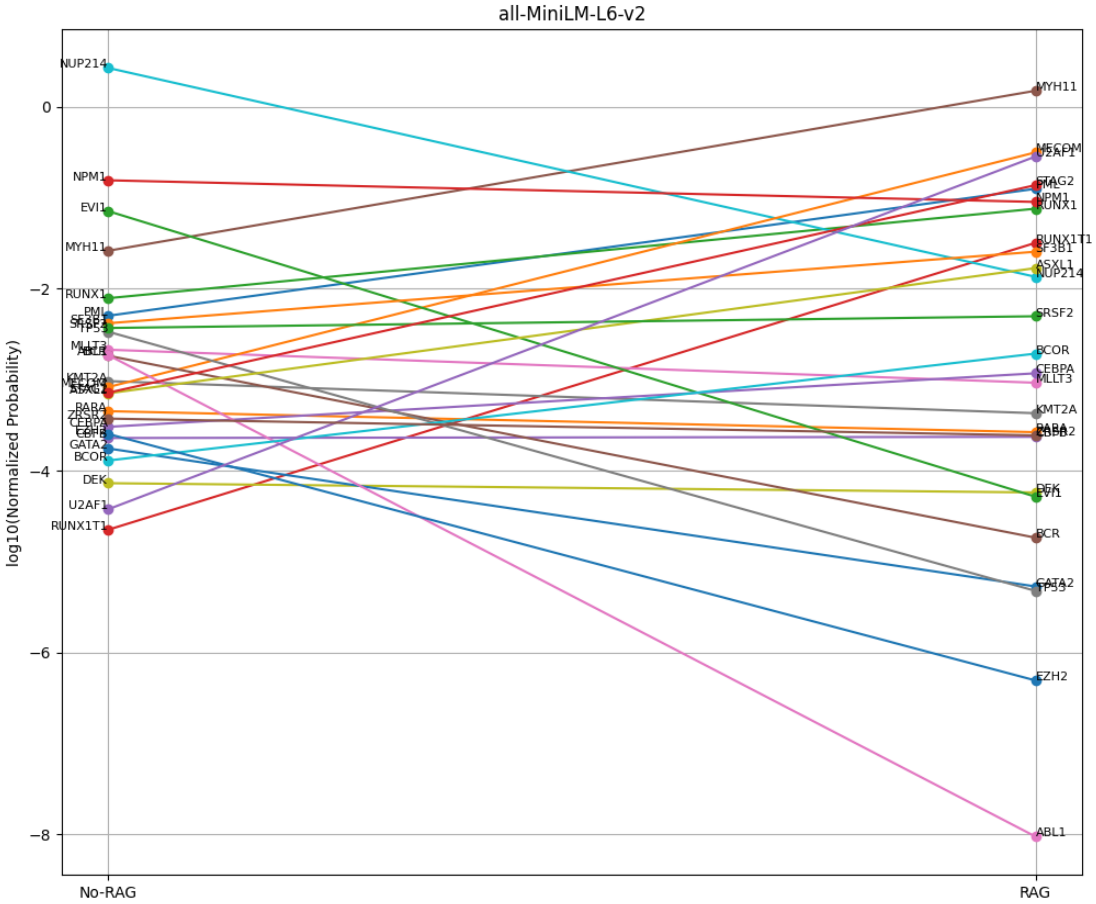

圖 4-10：all-MiniLM-L6-v2 嵌入模型對 ELN 基因 RAG 使用前後變化

doi:10.6342/NTU202504150

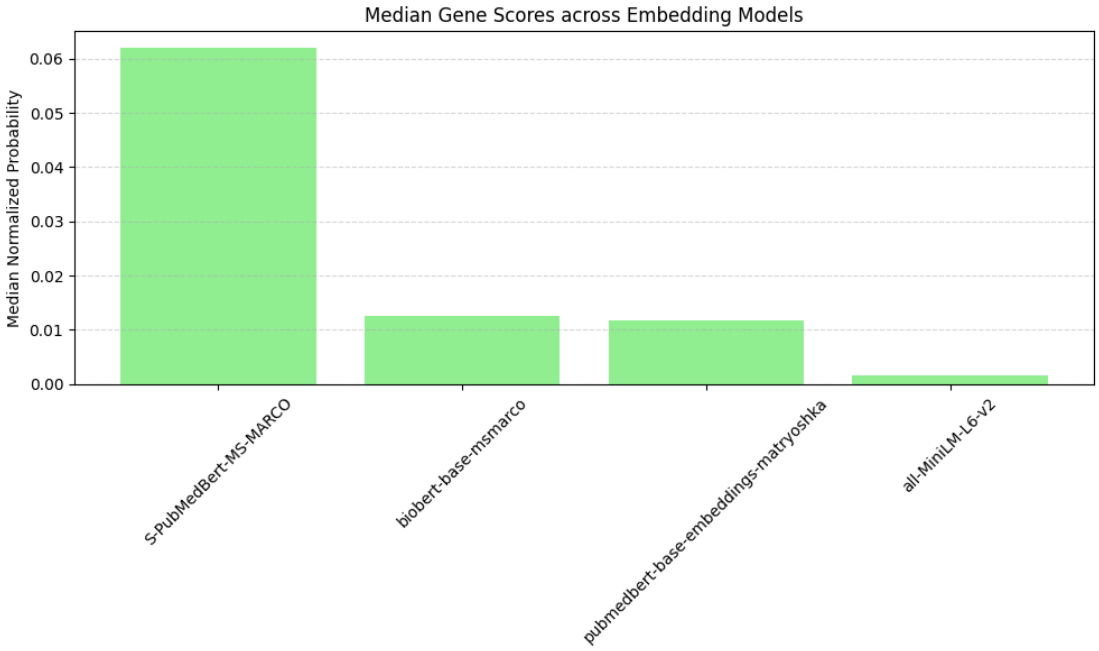

圖 4-11：不同嵌入模型在 RAG 條件下對 ELN 基因的預測值中位數比較

表 4-3：四種嵌入模型在 ELN 基因使用 RAG 前後的成對 t 檢定統計結果

Embedding model	t value	p value
S-PubMedBert-MS-MARCO	2.8287	0.0045
biobert-base-msmarco	2.0768	0.0241
pubmedbert-base-embeddings-matryoshka	1.2179	0.1173
all-MiniLM-L6-v2	-0.1087	0.5429

綜合上述分析，S-PubMedBert-MS-MARCO 嵌入模型不僅在視覺化趨勢、數值分布與統計檢定上皆展現出最佳表現，為本研究中最具代表性的檢索語意嵌入模型。估計其預訓練語料與下游微調任務高度對應生物醫學領域檢索情境，可能是導致其在本研究的目標上呈現優異效果的關鍵。

doi:10.6342/NTU202504150

4.3.2 不同 chunk size 下的檢索增強生成效果

在檢索增強生成架構中，檢索語料的切分方式同樣也對模型的輸出有很大影響。為了探討不同的 chunk size 對 BioGPT 預測 ELN 基因機率值的影響，本研究固定 prompt（Human gene that related to Acute Myeloid Leukemia is the）與嵌入模型（使用 S-PubMedBert-MS-MARCO 模型），變動檢索參數 chunk size =（從 50 到 500，每次間隔 50），並以不使用 RAG 的 BioGPT 輸出作為對照，呈現兩者於 26 個 ELN 標的基因上之預測值的配對變化趨勢（圖 4-12）。

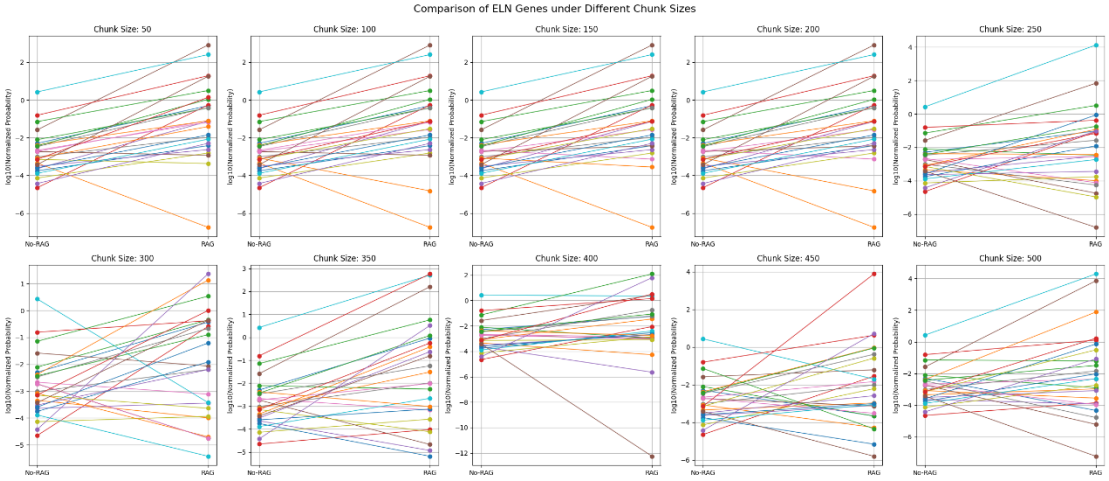

圖 4-12：不同 Chunk Sizes 下的 ELN 基因預測值在使用 RAG 前後對照（chunk size = 50～500）。

從圖可見，當 chunk size 增大時，模型對 ELN 標的基因的預測能力呈現明顯發散甚至下降趨勢。絕大多數 ELN 標的基因在小 chunk size 時皆呈現增強趨勢，顯示對於本研究所需要的短 token 長度來說，適當長度的上下文切片有助於提升模型在目標 ELN 基因上的敏感度與生成能力。然而，隨 chunk size 增大，若檢索段落過長，則可能會對 prompt 中明確指示的語義造成稀釋效應，導致 BioGPT 模型判斷能力下降。

doi:10.6342/NTU202504150

我們同時檢視了在不同 chunk size 下的 ELN 標的基因預測值中位數（圖 4-13），具體來說，chunk size 設為 50 與 100 時，中位數分別為 0.075 與 0.070，表現為最佳；而當 chunk size 達 200 以上時，中位數僅剩 0.01，模型輸出的效果呈現顯著削弱，相信這極大可能是與上下文資訊過長導致的資訊稀釋使得對基因短 token 數的預測能力下降有關。

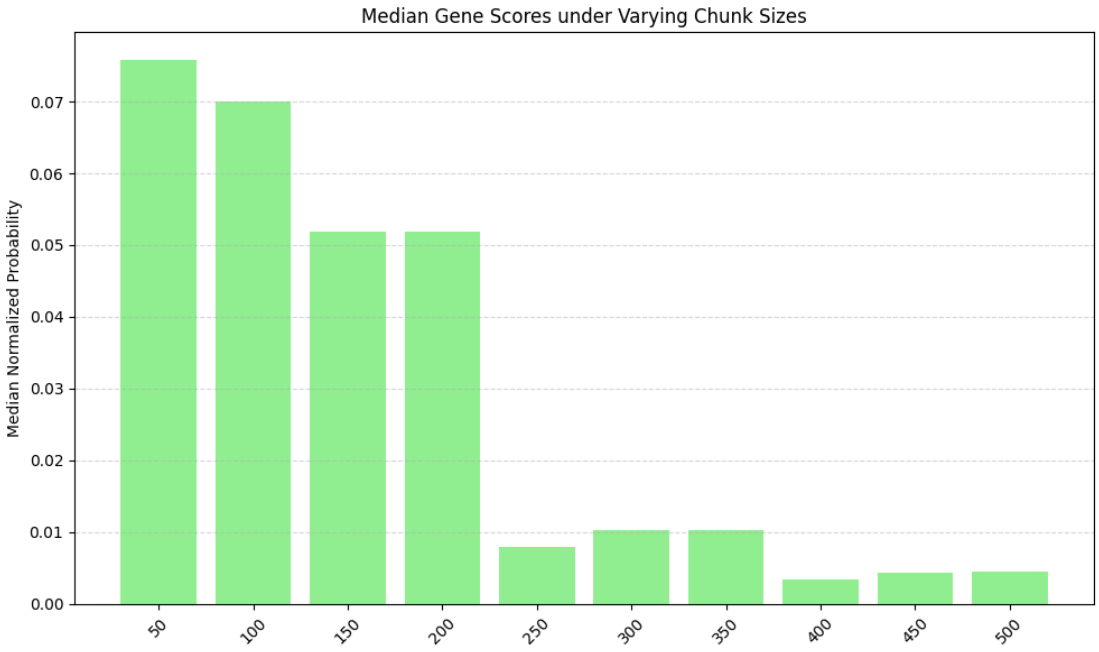

圖 4-13：不同 Chunk Sizes 下的 ELN 基因預測值的中位數

緊接著本研究更進一步從統計檢定觀察 paired t-test 的結果（圖 4-14, 圖 4-15），chunk size 設為 50 時達成最顯著差異（$t = 5.6458$, $p = 3.54\times10^{-6}$），顯示在此設定下，BioGPT 在有 RAG 的情況下，所產生的預測值差異最大且具有統計顯著性。t 值與 p 值在 chunk size=200 前皆維持極顯著差異性（$p < 0.001$），但隨 chunk size 增大而快速衰減，而當 chunk size 超過 350 時，效果上的顯著性逐漸喪失。總結而言，本節實驗結果指出 chunk size=50 時為最佳區間。

doi:10.6342/NTU202504150

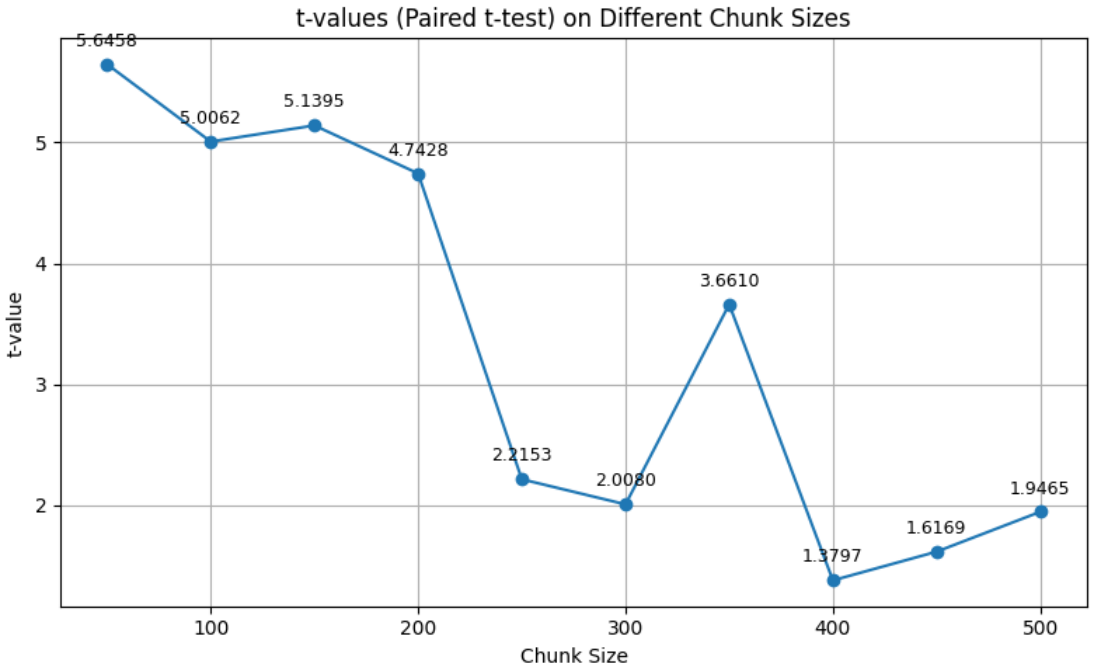

圖 4-14：不同 Chunk Sizes 下的使用 RAG 前後的 ELN 基因預測值之 t 值

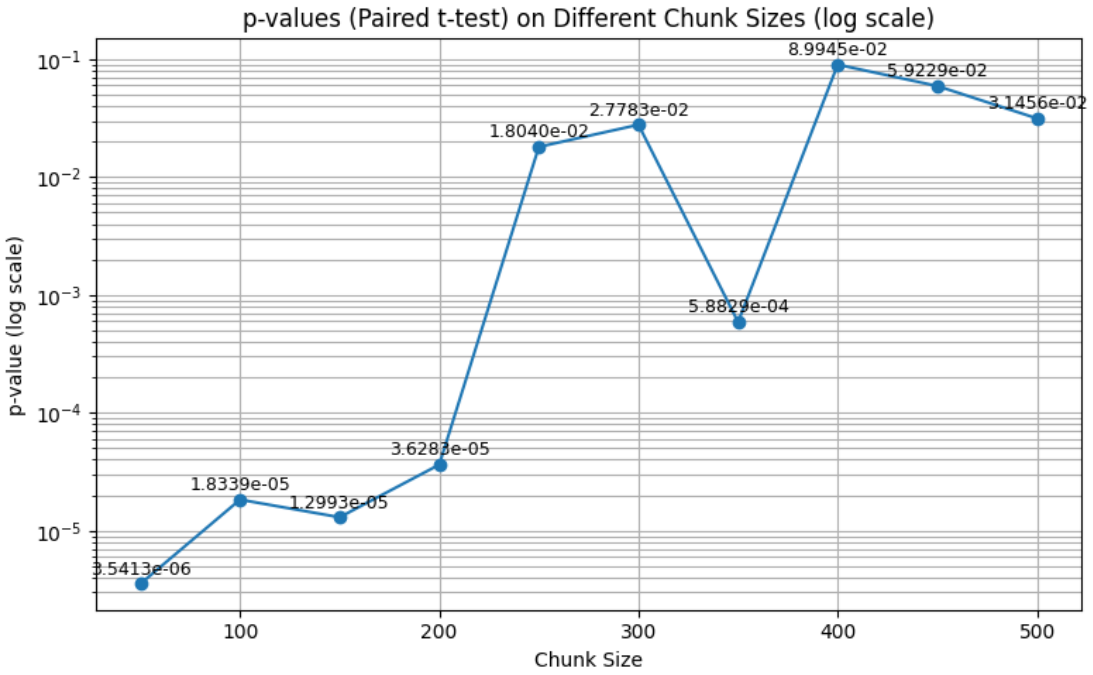

圖 4-15：不同 Chunk Sizes 下的使用 RAG 前後 ELN 基因預測值之 p 值（log）

doi:10.6342/NTU202504150

4.3.3 RAG 在不同 prompt 下對預測的影響分析

本節旨在探討不同 prompt 設計與是否應用檢索增強生成（RAG）技術，對 BioGPT 模型在 ELN 標的基因預測結果上的影響。為了系統性比較各 prompt 的表現，我們針對 26 個 ELN 基因分別進行評分，並對比是否使用 RAG 的效果。圖 4-16 與圖 4-17 分別呈現未使用 RAG（原生 BioGPT）與使用 RAG 的情況下，各 Prompt 對 ELN 標的基因所生成的預測值分布的盒鬚圖，以便觀察不同 Prompt 模板的相對分布差異。整體而言，絕大多數 Prompt 在 RAG 模式下均展現出分布較高、波動範圍較小的現象，顯示引入 RAG 實質有助於模型穩定地強化相關基因的生成。值得注意的是，使用 RAG 的模型具有比較長的左尾，在這裡可以明顯看出這是唯一例外逆勢下降的基因 *ZRSR2* 所造成的結果，可以在先前與後續比較圖中都可以觀察到類似的現象。

圖 4-18 與圖 4-19 則進一步呈現不同 prompt 對 ELN 標的基因的得分熱圖。從圖中可觀察到，在使用 RAG 後的整體有顯著提升（預測值更高），除了特定基因表現大幅提升，整體 ELN 標的基因都在多數 prompt 下穩定的表現，反映了檢索段落對於該些關鍵基因能有效提供上下文資訊，進而提高模型的一致性與預測力。同樣值得注意的是，如同先前的結果，可以明顯看出唯一例外逆勢下降的基因 *ZRSR2*。

綜上所述，RAG 即使在不同 prompt 的影響下，對 ELN 基因的評分結果仍有顯著幫助。因此適當設計的 prompt 結合具針對性的檢索內容，能夠顯著提高模型對 AML 相關基因的辨識力與預測能力。後續模型應用與系統建置上，將 prompt 設計與資料檢索機制納入共同優化或許可以是一個不錯的方向，以獲得更穩定且具生物醫學意義的推論結果。

doi:10.6342/NTU202504150

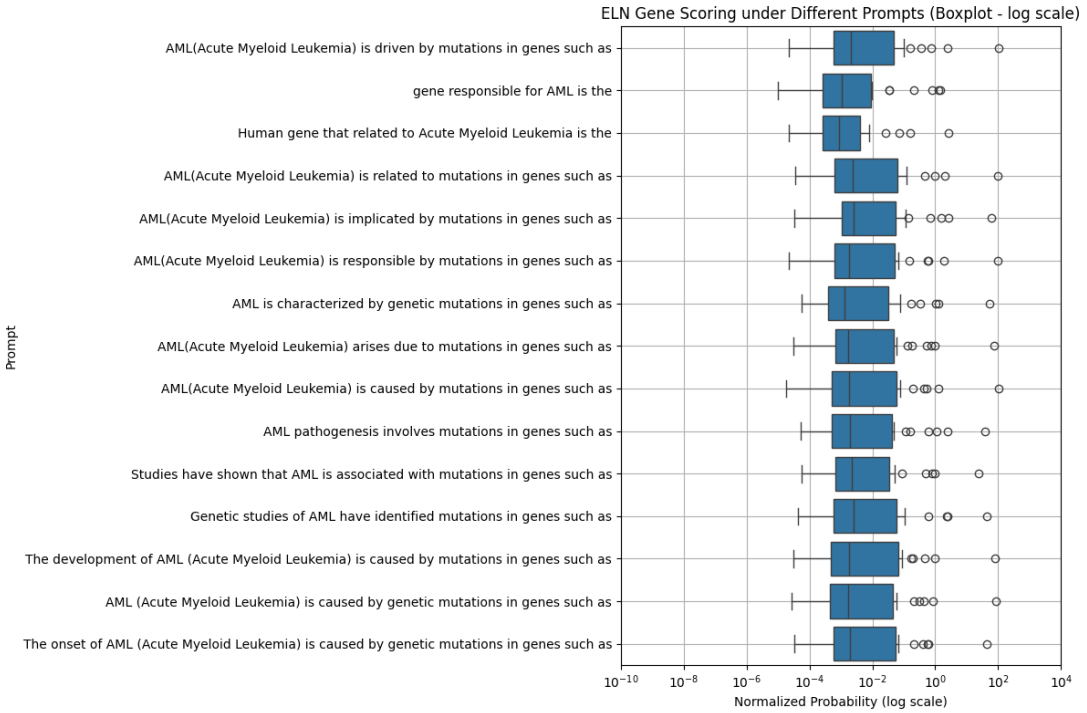

圖 4-16：不同 prompt 在未使用 RAG（原生 BioGPT）於 ELN 基因預測值分布

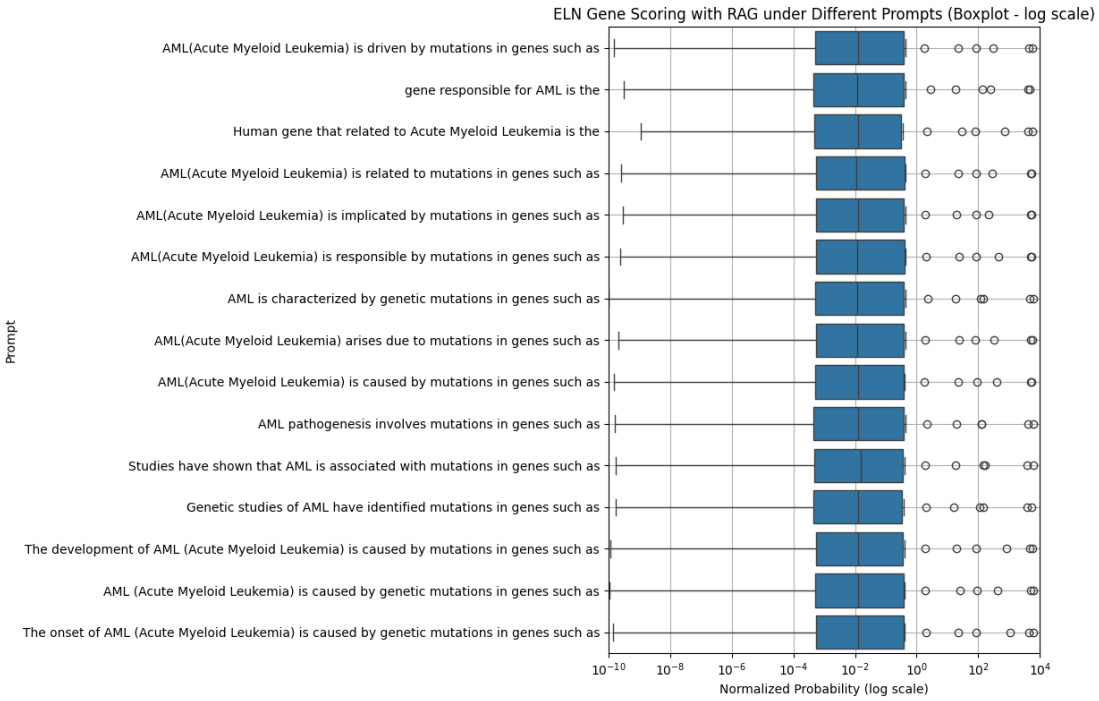

圖 4-17：不同 prompt 在使用 RAG 架構時 ELN 基因預測值分布

doi:10.6342/NTU202504150

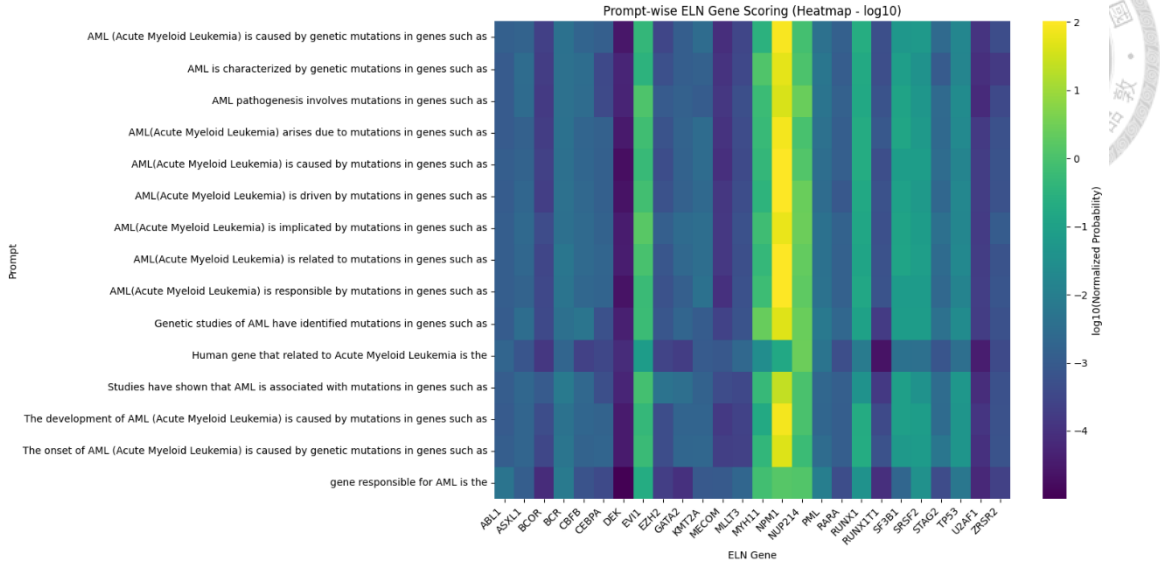

圖 4-18：不同 prompt 在未使用 RAG（原生 BioGPT）於 ELN 基因預測值熱圖

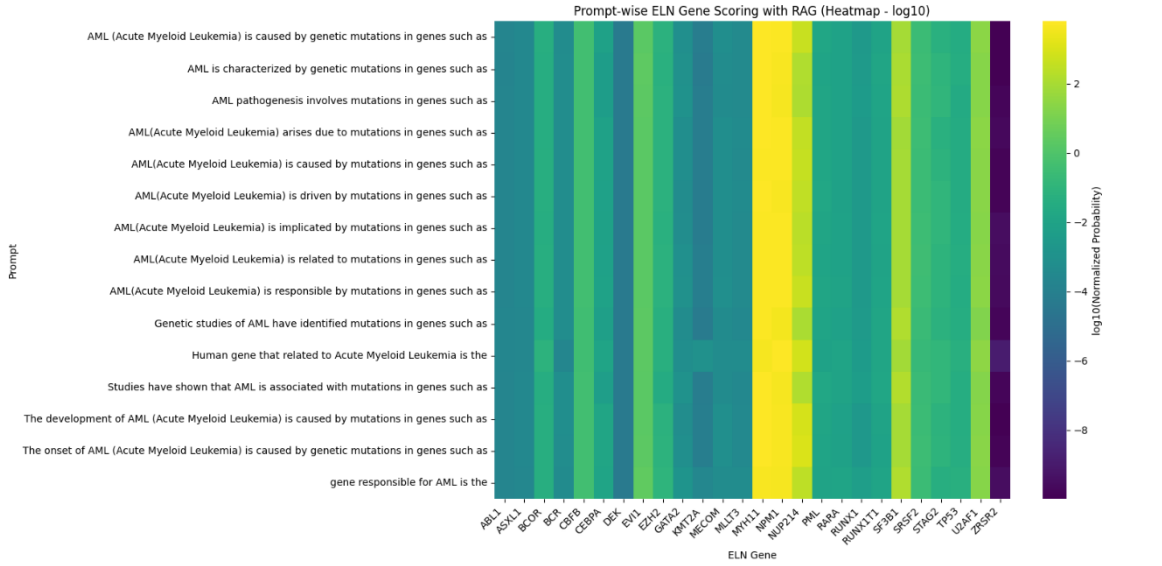

圖 4-19：不同 prompt 在使用 RAG 架構時 ELN 基因預測值熱圖

doi:10.6342/NTU202504150

4.3.4 RAG 對 ELN 標的基因與非 AML 基因預測的影響

綜合上述的成果，本研究得以評估檢索增強生成（RAG）架構對目標基因預測效能的實質影響，本研究針對 ELN 標的基因與非 AML 基因（選自背景基因）進行分組分析，並比較兩組基因在使用 RAG 前後之語言模型的預測值變化情形。在此使用 prompt（Human gene that related to Acute Myeloid Leukemia is the）與嵌入模型（使用 S-PubMedBert-MS-MARCO 模型），檢索參數 chunk size =50，並且以不使用 RAG 的原生 BioGPT 輸出作為對照。

圖 4-20 顯示了 26 個 ELN 基因在未使用與使用 RAG 的情境下，語言模型支持分數的變化趨勢。從圖中可見，絕大多數基因的分數在導入 RAG 後呈現明顯提升，為驗證此提升是具有顯著統計意義，因此進行成對 t 檢定。檢定結果為 $t = 2.8287$，$p = 0.0045$，達顯著水準（$p < 0.01$），說明 RAG 在 ELN 基因上確實具有增強語言模型分數的效果。

另一方面，為測試 RAG 是否對非 AML 相關之背景基因產生影響，本研究採用先前所使用之 26 個非 AML 相關之基因為對照組（表 4-2）。結果如圖 4-21 所示，背景基因的預測值在導入 RAG 後多呈現下降趨勢，代表在 AML 專屬語境下，RAG 能夠抑制語言模型對這些非 AML 基因的萃取能力。此現象同樣可由成對 t 檢定佐證（左尾），檢定結果為 $t = -3.0939$，$p = 0.0024$，同樣具統計顯著性。最後觀察圖 4-22，ELN 基因在排名中，從原先未使用 RAG 時的全基因的前 1%囊括了近半數的 ELN 基因，進步到使用 RAG 後，前 0.1%囊括了近半數的 ELN 基因，而從原先未使用 RAG 時的全基因的前 5%囊括了所有的 ELN 基因進步到使用 RAG 後的前 0.1%囊括近乎所有 ELN 基因（除 *ZRSR2*）。

綜上所述，RAG 不僅能顯著提升模型對 ELN 標的基因的預測能力，同時也能降低其他與目標疾病無關的背景基因的結果，表現出良好的選擇性與敏感性。此結果進一步支持 RAG 架構在疾病相關基因預測任務中的應用潛力。

doi:10.6342/NTU202504150

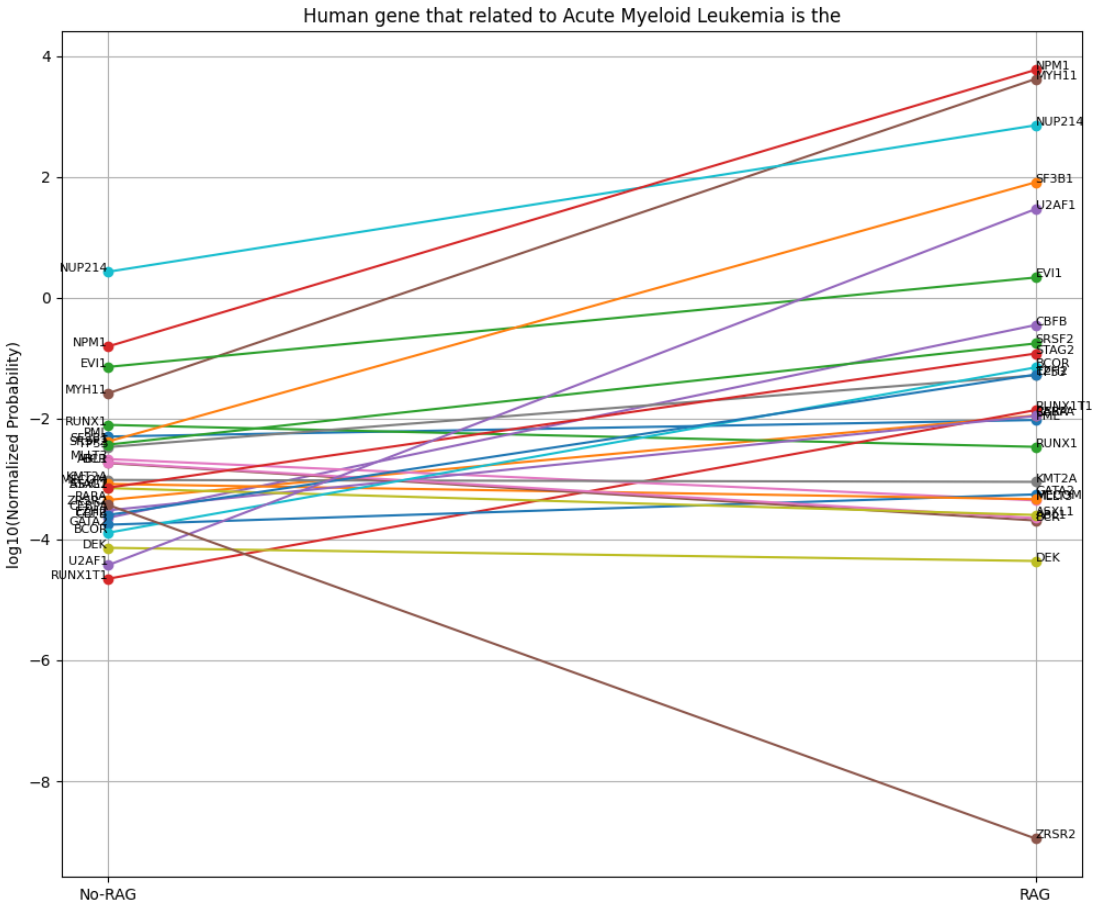

圖 4-20：使用 RAG 前後對 ELN 標的基因的預測值變化

doi:10.6342/NTU202504150

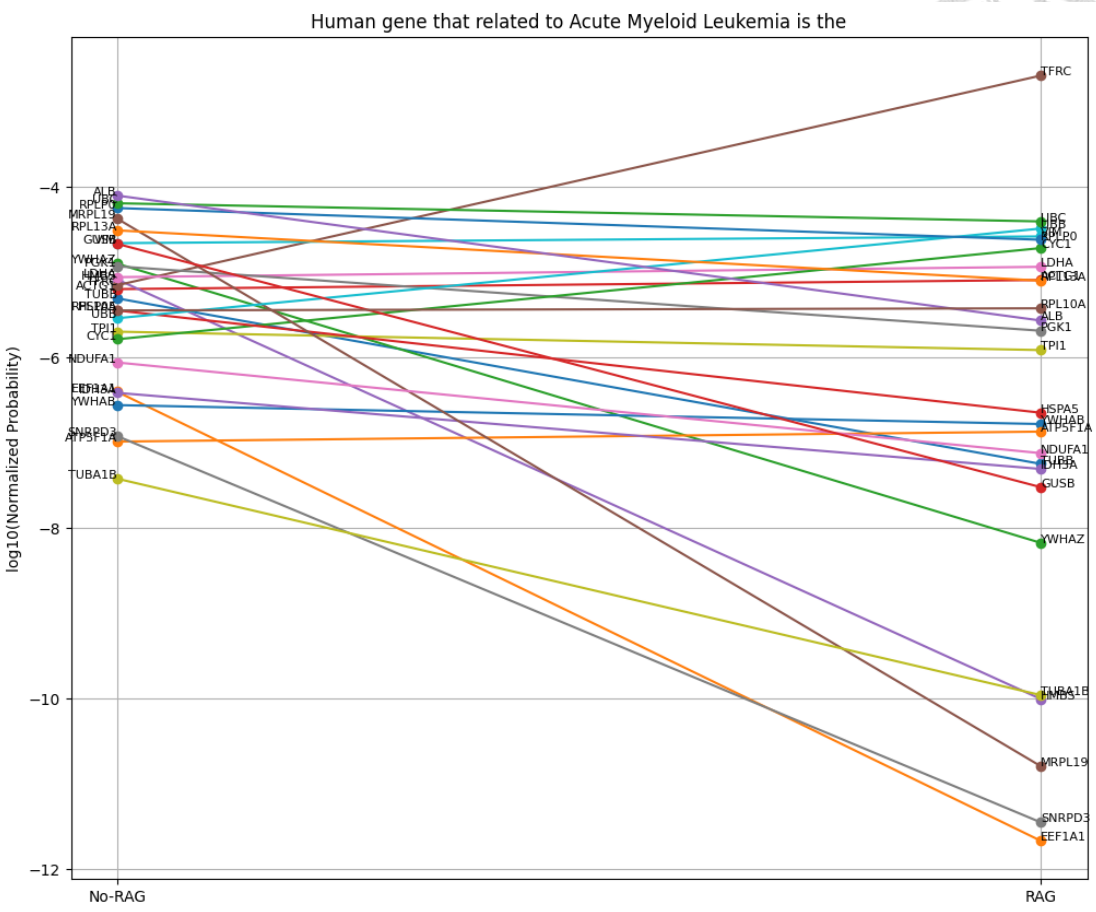

圖 4-21：使用 RAG 前後對非 AML 相關基因的預測值變化

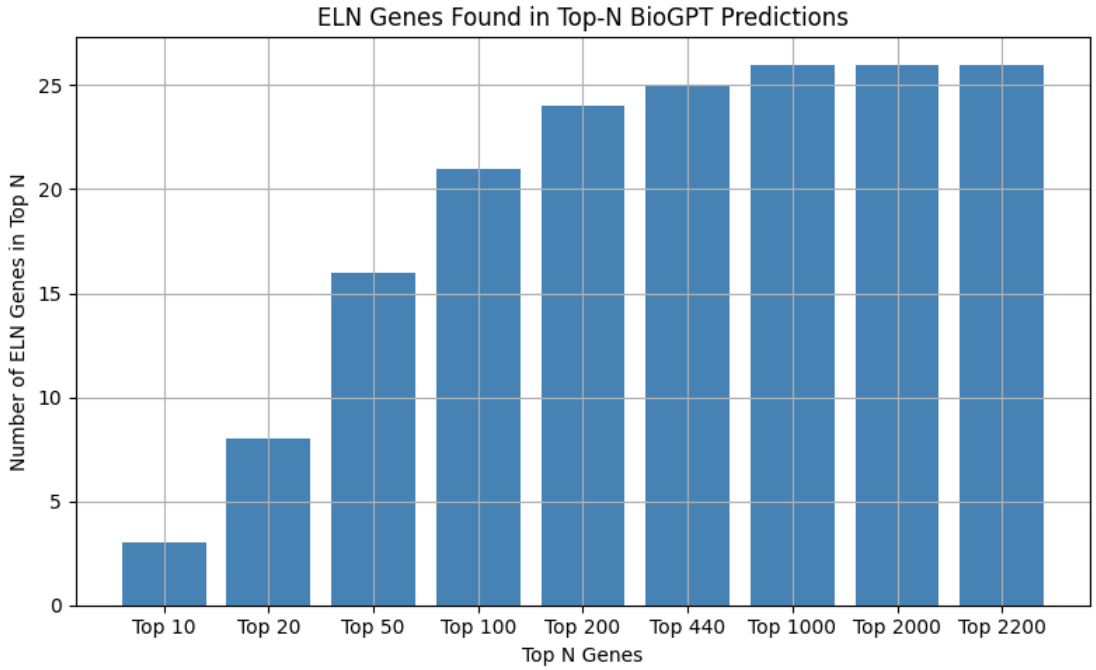

圖 4-22：使用 RAG 之後 Top-N 預測值中出現的 ELN 基因數量統計圖

doi:10.6342/NTU202504150

4.4 總結

本章節針對 RAG 在 BioGPT 對疾病相關基因預測任務上所展現的性能，進行了系統性實驗與深入分析，內容涵蓋嵌入模型選用、prompt 設計、chunk size 調整與不同基因群體間的差異進行多角度探討。主要的發現總結如下：

首先是 RAG 架構能進一步增強 BioGPT 的基因預測能力，經由 paired t-test 分析，在最佳化 prompt 之下加入 RAG，ELN 標的基因的 預測值有顯著提升。此結果支持檢索增強生成架構能提供有力上下文以強化模型對疾病基因的識別能力，提升其生物醫學領域的應用潛力。而 RAG 同樣也可以對非目標背景基因則呈現抑制效果，相對於 ELN 基因，使用相同 RAG 設定後，非 AML 背景基因的模型預測值普遍下降。此一「拉高目標、壓低背景」的對比結果，進一步凸顯 RAG 能協助語言模型聚焦於目標語意範圍，達成更具選擇性與目標性的預測行為。

綜合來說，本章實證了透過 prompt 精細設計與結合檢索增強架構，能顯著提升 BioGPT 在 AML 領域內對目標基因的預測表現。此結果不僅驗證語言模型在生物醫學知識領域的潛能，也為未來進一步導入外部知識庫與強化特定疾病預測提供明確路徑。

doi:10.6342/NTU202504150

第五章 結論

本研究旨在探討大型語言模型於生物醫學領域中的應用潛能，特別聚焦於 BioGPT 此一開源模型對於急性骨髓性白血病相關基因的語言模型評估能力，並且進一步結合了近年熱門的檢索增強生成架構，強化其對領域知識的建構與推理能力。

本研究建立起一個可針對目標基因進行 token-by-token 預測機率評估的推論流程，並採用合適正規化策略將語言模型的機率輸出轉換為可比較數值，成功應用於 AML 相關基因之排序任務。本研究也針對 AML 設計多組 prompt，發現語意上具有因果性（如「driven by」、「responsible for」）的 prompt 能顯著提升模型對 ELN 標的基因的預測能力，突顯 prompt 設計在 BioGPT 生物醫學應用中的關鍵角色。

為進一步了解模型預測結果，未來可結合 ELN 指南所定義之不同風險分類（Risk Categories），進行更細緻的比較分析。現行研究主要針對 ELN 所列標的基因與其他基因間的語意機率差異進行評估，而 ELN 風險分類中尚可細分為 Favorable、Intermediate 與 Adverse 等類別，未來若能將基因分群與模型分數進行交叉對照，將有助於釐清 BioGPT 對於不同臨床風險層級的語境感知能力與基因辨識傾向與應用潛力。

此外，本研究亦透過將 PubMed 資料庫中的 AML 相關研究文獻進行切分後，構建向量資料庫，並與 BioGPT 推論流程整合，實證了 RAG 架構能有效提供語境補強，使得目標基因在加入檢索後的模型預測值顯著提高，並且能抑制非目標基因，強化模型的選擇性與準確性。在其過程中也進行了大量分析不同嵌入模型與 chunk size 影響的實驗，並系統性比較多種嵌入模型與不同參數設定，發現較短文本片段（chunk size =50）更能提供此類短 token 數生成任務以高辨識力的上下文，然而過長片段則可能會造成稀釋語意焦點，從而導致預測效果下降。

doi:10.6342/NTU202504150

而不論是否加入 RAG 架構，本研究均以統計檢定（Mann-Whitney U test, Paired t-test）佐證 ELN 標的基因之平均模型預測值顯著地高於背景基因，顯示 BioGPT 已具備特定疾病知識的生成能力，RAG 則有能力可以進一步放大此差異性。

本研究為少數以「機率輸出解析」角度探討 BioGPT 模型語言理解能力的研究實驗，並結合大量相關實證測試，形成具實用性的 RAG-inference pipeline 而其中所提出具可重複性之基因機率評分與正規化流程，可作為日後其他疾病基因篩選任務的基礎架構，在未來可能的相關應用中，本研究提供了實證支持大型語言模型在生物醫學領域的目標相關基因擷取中，透過語言模型與 RAG 的組合可實現近乎專家級的表現。。

除了應用至其他疾病領域之外，另一種在未來可以期待的方向便是擴展至其他模型：可進一步延伸至 GPT-4、Gemini 等大型商業模型，探討其在其他基因相關疾病上的泛化能力。

doi:10.6342/NTU202504150

參考文獻

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv*, abs/2005.14165.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics.*

Döhner, H., Estey, E., Grimwade, D., Amadori, S., Appelbaum, F. R., Büchner, T., Dombret, H., Ebert, B. L., Fenaux, P., Larson, R. A., Levine, R. L., Lo-Coco, F., Naoe, T., Niederwieser, D., Ossenkoppele, G. J., Sanz, M., Sierra, J., Tallman, M. S., Tien, H. F., Wei, A. H., … Bloomfield, C. D. (2017). Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, *129*(4), 424–447.

Döhner, H., Wei, A. H., Appelbaum, F. R., Craddock, C., DiNardo, C. D., Dombret, H., Ebert, B. L., Fenaux, P., Godley, L. A., Hasserjian, R. P., Larson, R. A., Levine, R. L., Miyazaki, Y., Niederwieser, D., Ossenkoppele, G., Röllig, C., Sierra, J., Stein, E. M., Tallman, M. S., Tien, H. F., … Löwenberg, B. (2022). Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. *Blood, 140*(12), 1345–1377.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare, 3*(1), Article 2.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics*, *36*(4), 1234-1240.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *In Proceedings of the 34th International Conference on Neural Information Processing Systems* (Article 793).

doi:10.6342/NTU202504150

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T.-Y. (2022, September). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, **23**(6), bbac409.

National Center for Biotechnology Information (NCBI). (2020). PubMed Overview. National Library of Medicine.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., Teague, J. W., … Campbell, P. J. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England journal of medicine*, *374*(23), 2209–2221.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-training. OpenAI.

Schlenk, R. F., Döhner, K., Krauter, J., Fröhling, S., Corbacioglu, A., Bullinger, L., Habdank, M., Späth, D., Morgan, M., Benner, A., Schlegelberger, B., Heil, G., Ganser, A., Döhner, H., & German-Austrian Acute Myeloid Leukemia Study Group (2008). Mutations and treatment outcome in cytogenetically normal acute myeloid leukemia. *The New England journal of medicine*, *358*(18), 1909–1918.

Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural machine translation of rare words with subword units. *In K. Erk & N. A. Smith (Eds.), Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1715–1725). Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.

Zagirova, D., Pushkov, S., Leung, G. H. D., Liu, B. H. M., Urban, A., Sidorenko, D., Kalashnikov, A., Kozlova, E., Naumov, V., Pun, F. W., Ozerov, I. V., Aliper, A., & Zhavoronkov, A. (2023). Biomedical generative pre-trained based transformer language model for age-related disease target discovery. *Aging (Albany NY)*, *15*(18), 9293–9309.

doi:10.6342/NTU202504150