基於領域百科增強語言模型之知識圖譜微調方法應用於中文醫療問題意圖分類

Chinese Medical Question Intent Classification Based on Domain Encyclopedia Enhanced Language Models with Knowledge Graph Fine-tuning

陳冠竹 Kuan-Chu Chen[1] **謝采蓁 Tsai-Chen Hsieh**[1]
陳柏翰 Po-Han Chen[2] **李龍豪 Lung-Hao Lee**[1,*]

[1] 國立陽明交通大學智能系統研究所
Institute of Artificial Intelligence Innovation,
National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[2] 國立中央大學電機工程學系
Department of Electrical Engineering,
National Central University, Taoyuan, Taiwan

[*] lhlee@nycu.edu.tw

摘要

本研究針對中文醫療問句意圖分類任務，提出 EKG-Transformer 模型，整合醫學領域百科知識與知識圖譜中的關係資訊，以強化預訓練語言模型對醫療問句意圖的理解。醫學百科可提供疾病、症狀、檢查與治療等領域知識及其屬性資訊；然而，醫療問句的意圖判斷亦涉及不同醫療實體之間的語意關係。爲此，本研究於下游分類階段，根據問句中的命名實體擷取相關醫療知識三元組，並據以建構 Sentence Graph，再透過 Soft-position Embedding 與 Visible Matrix 控制知識節點間的資訊傳遞，以降低不相關知識所造成的語意干擾。實驗採用中文醫療問句意圖分類（CMID）資料集，涵蓋 4 個主類別與 36 個子類別的分類任務。實驗結果顯示，導入知識圖譜進行微調後，BERT、RoBERTa 與 MacBERT 均獲得穩定的效能提升。其中，完整的 EKG-MacBERT 模型在 4 類別與 36 子類別分類任務上分別取得 0.7450 與 0.5370 的 Micro-F1，顯示醫學領域百科知識與知識圖譜關係資訊具有互補性，能有效提升模型對中文醫療問句意圖的辨識能力。

Abstract

This study proposes EKG-Transformer for Chinese medical question intent classification, integrating domain-specific encyclopedic knowledge with relational knowledge from medical knowledge graphs to enhance the ability of pretrained language models to understand the intents of medical questions. Medical encyclopedias provide domain knowledge and attribute information concerning diseases, symptoms, examinations, and treatments; however, identifying the intent of a medical question often also requires understanding the semantic relationships among medical entities. To address this issue, during downstream classification, relevant medical knowledge triples are retrieved based on named entities identified in each question and are subsequently used to construct a Sentence Graph. Soft-position Embedding and a Visible Matrix are further employed to control information flow among knowledge nodes, thereby reducing semantic interference introduced by irrelevant knowledge. Experiments are conducted on the Chinese Medical Intent Dataset (CMID), which comprises classification tasks involving 4 coarse-grained categories and 36 fine-grained subcategories. The experimental results show that incorporating knowledge graph information during fine-tuning consistently improves the performance of BERT, RoBERTa, and MacBERT. In particular, the complete EKG-MacBERT model achieves Micro-F1 scores of 0.7450 and 0.5370 on the 4-category and 36-subcategory classification tasks, respectively. These results demonstrate the complementary benefits of medical encyclopedic knowledge and relational knowledge from knowledge graphs for improving Chinese medical question intent classification.

關鍵字：領域知識強化、預訓練語言模型、醫學百科、知識圖譜

Keywords: Domain knowledge enhancement, pretrained language models, medical encyclopedia, knowledge graph

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 333–342
November 6-7, 2026 ©2026 Association for Computational Linguistics

1 介紹

多類別文本分類旨在將非結構化文本歸納至預先定義的單一類別，可應用於資訊檢索、對話系統與自動問答等任務。近年來，BERT、RoBERTa 與 MacBERT 等預訓練語言模型透過大規模語料學習上下文語意，在多項自然語言理解任務中取得良好表現。然而，一般領域預訓練模型主要依賴詞語共現與上下文資訊學習語意，面對包含疾病、症狀、藥物、治療與檢查等專業知識的醫療問句時，仍可能缺乏判斷意圖所需的領域知識。

中文醫療問句通常具有篇幅短、口語化、資訊省略及專業詞彙密集等特性。例如，問句「打肺炎疫苗導致心悸跟心肌炎有關嗎？」同時涉及藥物、症狀與疾病實體，但其主要意圖爲詢問臨床表現。若能準確辨識醫療問句的意圖，系統便可據此篩選相關醫療資訊，並作爲醫療問答或資訊檢索系統的前置模組。

醫學百科可提供疾病、症狀、檢查與治療等醫療知識及其屬性資訊，有助於模型建立基本的醫療概念表示。然而，醫療問句的意圖判斷往往不僅依賴單一醫療概念，也涉及不同醫療實體之間的關係。例如疾病可能對應症狀、藥物可能對應適應症，治療方式亦可能與檢查項目或恢復時間相關。因此，僅依賴百科領域知識仍不足以完整捕捉醫療問句中的關係語意。

基於上述觀察，我們提出 EKG-Transformer (**E**ncyclopedia enhanced pre-training with **K**nowledge **G**raph fine tuning transformer) 模型，結合百科領域知識與醫療知識圖譜關係，用於中文醫療問句意圖分類。於下游分類階段，模型根據問句中的命名實體擷取相關醫療知識圖譜三元組，並將原始問句轉換爲 sentence graph。接著，透過 soft-position embedding 保留原始問句的相對位置資訊，並利用 visible matrix 控制原始詞彙與知識節點之間的可見關係，以降低不相關知識造成的語意干擾。最後，模型將知識增強後的表示輸入 Visible-Transformer Encoder 與分類器，預測醫療問題意圖。我們採用 Chinese Medical Intent Dataset (CMID)，其包含 4 個意圖主類別與 36 個子類別，共 12,254 則經人工標註之中文醫療問句；實驗結果顯示，完整結合百科知識與知識圖譜關係之 EKG-MacBERT 模型於 4 類別任務取得 Micro-F1 0.7450、於 36 子類別任務取得 0.5370，均優於未加入額外領域知識之預訓練模型。

本文架構如下：第二章回顧相關研究；第三章詳述本研究提出之 EKG-Transformer 模型架構；第四章說明實驗設計，並比較多種模型之效能表現；第五章總結本研究貢獻。

2 相關研究

2.1 中文醫療問題意圖分類

中文醫療問題意圖分類旨在精準理解使用者問題查詢的核心目的，爲醫療問答系統與資訊檢索的關鍵前置任務。在資料集發展上，Chen et al. (2020) 提出了涵蓋多個意圖類別的中文醫療問題意圖分類資料集 (CMID)，爲相關研究奠定基礎。Zhang et al. (2022) 則發表了 CBLUE 基準測試，進一步統整了 KUAKE-QIC 與 CHIP-CTC 等多項醫療意圖分類與語言理解任務，大幅推動了中文醫療自然語言處理的發展。隨著技術演進，實際應用面臨了新的挑戰。Bai et al. (2022) 針對醫療領域中不斷出現的新意圖，探討了增量式學習 (Incremental Learning) 挑戰，並提出 CRN (Contrastive Replay Networks) 來解決醫療罕見詞與資料不平衡造成的災難性遺忘。此外，儘管大型語言模型 (LLMs) 在通用領域表現出色，但 Mullick et al. (2024) 針對醫療意圖與實體萃取的實證研究指出，經過特定領域微調 (Supervised Fine-Tuning, SFT) 的預訓練模型 (如 PubMedBERT)，在醫療意圖分類的精準度上，現階段仍顯著優於 ChatGPT 或 Llama-2 等通用型 LLMs。此結果顯示，針對特定醫療領域持續發展知識強化與微調模型，仍具有高度的學術與實務價值。更重要的是，此類領域強化模型的參數規模通常僅爲大型語言模型的數十分之一，訓練與推論成本亦相應大幅降低。在眞實醫療問答系統中，查詢通常要求即時回應，且基於資料隱私與法規遵循考量，多數醫療機構傾向於院內或離線部署，而非仰賴外部大型語言模型服務。因此，在效能相近的前提下，輕量化的領域知識強化模型在部署成本與可控性上更具實務優勢，此亦爲本研究持續發展知識強化機制，而非直接採用大型語言模型的核心動機。

2.2 多類別文本分類技術

多類別文本分類技術已從傳統機器學習全面邁向基於預訓練語言模型的深度學習架構，近期研究更著重於特徵強化與進階訓練策略以處理複雜語意。例如，Swarnali et al. (2024) 提出詞元層級對抗對比訓練 (TACT) 與標籤感知對比損失 (LCL)，透過對比學習顯著提升模型對細微語意的分辨能力。Jain et al. (2025) 則利用命名實體辨識 (NER) 結合強化注意力機制與知識蒸餾，在處理多類別複雜文本時取得了良好的分類表現。

在醫療與生物醫學領域，文本分類常面臨專

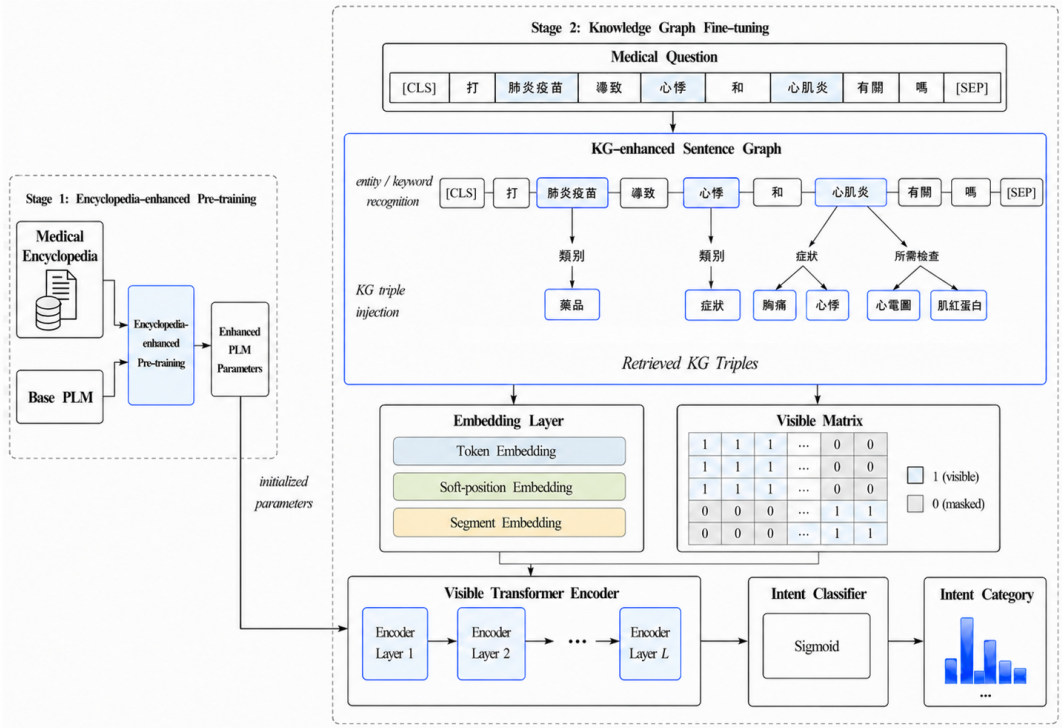

圖 1: EKG-Transformer 整體架構

業詞彙密集、資料極端不平衡與語意不確定性等嚴峻挑戰。爲此，Ndama et al. (2025) 結合 BERT 的上下文學習與 LSTM 的序列處理能力，並融入句子順序資訊，大幅提升了生物醫學文獻的分類準確度。針對醫療數據不平衡的問題，Chen and Du (2025) 提出結合自注意力對抗增強網路 (SAAN) 與疾病感知多任務 BERT (DMT-BERT) 的分類框架，不僅能生成高品質的少數類別樣本，還能同時學習疾病的共現關係。此外，Manage et al. (2026) 針對 BERT 在面對多類別醫療術語時所產生的「不確定性」限制，提出創新的兩階段決策框架，並引入「詞-類別機率嵌入」(WCPE) 來增強領域特定的表示學習，精準分離並處理了低信心的預測案例。

綜合上述近期研究可見，結合實體特徵與外部領域知識以優化注意力機制，並針對不平衡與不確定性設計強化框架，已成爲當前醫療多類別分類的重要發展方向。這也印證了本研究所提出的 EKG-Transformer 框架，透過導入醫學百科與知識圖譜關係來降低知識噪聲與語意干擾，具有進一步研究的價值。

2.3 領域知識導入機制

在醫療領域中，確保文本分類與生成的正確性及可信度至關重要。Wang et al. (2025) 針對醫療問答可信度的綜述指出，通用 LLMs 在醫療場景中仍面臨嚴重的幻覺 (hallucination) 與過度自信等安全風險，並強調透過檢索增強生成或導入外部結構化知識 (如知識圖譜) 來降低錯誤率的必要性。Gao et al. (2025) 提出 DR. KNOWS 圖模型，藉由從 UMLS 知識庫中擷取與病患臨床背景相關的知識圖譜路徑，強化 LLMs 在診斷預測上的推理能力。Zuo et al. (2025) 提出的 KG4Diagnosis 則建構了一個結合自動化知識圖譜生成的階層式多代理 (multi-agent) 框架，利用語意驅動的實體擷取與推理來處理複雜的醫學診斷場景。綜合上述發展趨勢，本研究結合「百科全書預訓練」與「知識圖譜微調」的兩階段機制，將醫學實體關係融入預訓練語言模型中，在維持語法結構的前提下，爲中文醫療意圖分類提供一套精準且具備領域知識感知的解決方案。

3 模型架構

本章說明本文提出的知識圖譜增強微調框架。EKG-Transformer 整體架構如圖 1所示，模型以百科知識增強預訓練模型爲基礎，並在下游分類階段加入醫療知識圖譜三元組，以強化中文醫療問句中的實體關係表示。

3.1 模型整體架構

本研究採用百科知識增強的領域適應策略作爲模型初始化方式，使預訓練語言模型能先接觸疾病、症狀、檢查、治療與藥物等醫療知識及其屬性。此類百科知識導入方式可透過醫學百

科中的焦點與面向資訊，使模型學習醫療概念之間的語意關聯 (He et al., 2020; Chen et al., 2021)。本文將此階段作爲概念知識初始化，並將重點放在後續的知識圖譜增強微調。

在知識圖譜微調階段，模型根據中文醫療問句中的命名實體，從醫療知識圖譜中擷取相關三元組，並將其加入原始問句以建構 sentence graph。接著，模型透過 soft-position embedding 保留原始問句的相對位置資訊，並利用 visible matrix 控制原始詞元與知識節點之間的注意力範圍，最後輸入 Visible-Transformer Encoder 與分類器進行意圖分類。

因此，EKG-Transformer 透過百科知識提供醫療知識與屬性資訊，並透過知識圖譜微調補充醫療實體之間的關係資訊，以提升中文醫療問句意圖分類效果。

3.2 知識圖譜增強微調框架

雖然百科知識增強預訓練能補充醫療概念及其面向資訊，但中文醫療問句的意圖判斷往往也涉及疾病、症狀、藥物、治療與檢查之間的實體關係。爲進一步導入此類結構化知識，本文於下游分類階段延伸 K-BERT(Liu et al., 2020) 的知識注入方法，根據問句中的命名實體擷取醫療知識圖譜三元組，並將其加入原始輸入中。

直接插入外部知識節點可能改變原句語意結構，甚至使不相關資訊影響分類結果，形成知識噪聲。因此，本研究將原始問句與注入的三元組建構爲 sentence graph，並透過 soft-position embedding 保留原始問句的相對位置資訊，再利用 visible matrix 限制原始詞元與知識分支之間的注意力連結，以控制知識節點的資訊流動 (Liu et al., 2020)。

假設知識圖譜爲 K，模型詞彙表爲 V，其中 w_i 與 w_k 分別表示頭實體與尾實體，r_j 表示兩個實體之間的關係類型。知識圖譜中的三元組可表示爲：

$$\varepsilon = (w_i, r_j, w_k), \quad \text{且} \quad \varepsilon \in K \tag{1}$$

由於實體名稱與關係名稱皆由模型詞彙表中的詞元表示，因此 w_i、w_k 與 r_j 均可由 V 中的詞元組成。知識圖譜微調階段可分爲知識層、多重嵌入層、可視矩陣層、Visible-Transformer Encoder，以及分類層等部分。

3.2.1 知識層 (Knowledge Layer)

知識層首先從輸入問句中找出可與醫療知識圖譜連結的實體。假設原始問句的詞元序列爲 $s = \{w_0, w_1, w_2, \ldots, w_n\}$，系統透過命名實體字典比對問句中的詞元，並以匹配到的實體查詢知識圖譜。三元組擷取過程可表示爲：

$$E = \text{KG_Query}(s, K) \tag{2}$$

其中，E 爲與問句實體相關的三元組集合。

取得三元組後，系統將關係與尾實體插入原始實體之後，使問句擴展爲具有分支結構的 sentence graph。知識注入可表示爲：

$$t = \text{KG_Injection}(s, E) \tag{3}$$

其中 t 爲完成知識擴展後的句圖。

若同一實體對應多個三元組，系統會先排列其關係分支，再依序加入輸入序列，直到達到模型設定的最大序列長度。爲避免句圖持續向外擴張，本研究僅注入一層知識，亦即新加入的尾實體不再繼續擷取下一層關係。此注入策略依循 K-BERT 將知識三元組展開爲句圖的基本概念 (Liu et al., 2020)。

3.2.2 多重嵌入層 (Embedding Layer)

Sentence graph 最終仍需轉換爲一維序列，才能輸入 transformer 編碼器。知識分支會被插入至其所屬實體之後，因此轉換後的序列可能不再符合自然語句的閱讀順序。爲同時表示詞元內容、原始句法位置與文本片段，本研究採用 1) Token embedding 用於表示原始詞元以及注入的實體與關係詞元；2) Soft-position embedding 則依據原始問句結構重新配置位置編碼，使知識分支的加入不會改變主要句子的相對詞序；3) Segment embedding 用於區分不同的文本片段，使模型能辨識各詞元所屬的輸入序列。

以問句「打肺炎疫苗導致心悸和心肌炎有關嗎」爲例，知識注入後可能在「肺炎疫苗」後加入「類別—藥品」等分支。若直接沿用插入後的線性位置，「導致」便會被視爲接續在「藥品」之後。soft-position embedding 因此仍將「導致」的位置設定爲緊接於「肺炎疫苗」，使主要問句保留原有的語意順序。

然而，soft-position embedding 可能使原始詞元與知識分支具有相同或相近的位置編碼。爲避免不同分支因位置重疊而互相干擾，後續再由 visible matrix 限制詞元之間的可視關係。

3.2.3 可視矩陣層 (Visible Matrix Layer)

Visible matrix (Liu et al., 2020) 的作用是指定 sentence graph 中哪些詞元可以彼此交換資訊。原始問句中的詞元可維持正常的注意力連結，而知識分支則僅能與其所屬的實體或同一分支內的節點互動，避免外部知識直接影響無關詞元。

假設可視矩陣爲 M。當詞元 w_i 可看見詞元 w_j 時，令 M_{ij} 的值爲 0；若兩者不可互相注意，則令 M_{ij} 的值爲負無限大 $(-\infty)$。公式表示如下：

$$M_{ij} = \begin{cases} 0, & w_i \ominus w_j, \\ -\infty, & w_i \oslash w_j. \end{cases} \tag{4}$$

例如，若「症狀」與「心悸」是由「心肌炎」延伸出的知識分支，這些節點可參與「心肌炎」的表示計算，但不應直接影響「肺炎疫苗」等其他實體。藉由限制知識分支的注意力範圍，模型可使用外部知識，同時保留原始問句的整體語意。

3.2.4 Visible-Transformer Encoder

Visible-Transformer Encoder 以自注意力機制爲基礎 (Vaswani et al., 2017)，並依循 K-BERT 的做法將可視矩陣加入注意力分數 (Liu et al., 2020)。假設第 i 層的隱藏表示爲 h^i，則 Query、Key 與 Value 可表示爲：

$$Q^{i+1}, K^{i+1}, V^{i+1} = W_q h^i, W_k h^i, W_v h^i \tag{5}$$

加入可視矩陣後，注意力分數的計算方式爲：

$$S^{i+1} = \text{softmax}\left(\frac{Q^{i+1}(K^{i+1})^T + M}{\sqrt{d_k}}\right) \tag{6}$$

並得到下一層表示：

$$h^{i+1} = V^{i+1} S^{i+1} \tag{7}$$

當兩個詞元被設定爲不可視時，其對應的矩陣元素爲負無限大，經 Softmax 後注意力權重便會趨近於 0。因此，不相關的知識節點無法直接參與該詞元的表示計算。此設計使編碼器能在保留原始句意的前提下，選擇性地整合醫療實體關係。

3.2.5 分類層 (Classification Layer)

完成 Visible-Transformer Encoder 的編碼後，模型取得輸入問句的整體語言表示 z。接著，將 z 輸入線性分類層與 Sigmoid 函數，計算各類別的預測分數：

$$\hat{y} = \text{sigmoid}\left(W_c z + b_c\right) \tag{8}$$

其中，W_c 爲分類層的權重矩陣，b_c 爲偏差值 (bias)，而 $\hat{y}$ 表示各候選類別的預測分數。由於 CMID 屬於單標籤多類別分類任務，模型最終選取分數最高的類別作爲預測結果。模型以交叉熵損失 (Cross-Entropy Loss) 進行最佳化：

$$L = \sum_{c=1}^{C} \sum_{i=1}^{n} -y_{c,i} \log_2\left(p_{c,i}\right) \tag{9}$$

其中，C 爲類別總數，n 爲資料筆數，$y_{c,i}$ 表示第 i 筆資料在第 c 類的眞實標籤，$p_{c,i}$ 則爲模型對該類別所預測的機率。透過百科全書預訓練與知識圖譜微調的連續訓練，模型可同時利用醫療概念、屬性與實體關係資訊，完成中文醫療問句的意圖分類。

4 實驗設計與結果分析

4.1 資料集

本研究使用中文醫療問題意圖分類資料集 (Chinese Medical Question Intent Dataset, CMID)(Chen et al., 2020)。該資料集蒐集自多個中文醫療問答網站，共包含 12,254 則經人工標註的使用者問句，研究中使用 OpenCC 工具將其由簡體轉換爲繁體。由於問句多來自眞實網路情境，因此具有口語化、句型不完整、篇幅較短及醫療專有詞彙密集等特性。

CMID 採用階層式標註架構，將醫療問題劃分爲「病症」、「藥物」、「治療方案」與「其他」4 個主類別，並進一步細分爲 36 個子類別。其中，「病症」類別包含 12 個子類別，分別爲「定義」、「病因」、「臨床表現」、「相關病症」、「治療方法」、「推薦醫院」、「預防」、「所屬科室」、「病症禁忌」、「傳染性」、「治癒率」與「嚴重性」，共 6,264 筆，占 51.12%；「藥物」類別包含 7 個子類別，分別爲「作用」、「適用症」、「價錢」、「藥物禁忌」、「用法」、「副作用」與「成分」，共 2,756 筆，占 22.49%；「治療方案」類別包含 10 個子類別，分別爲「方法」、「費用」、「有效時間」、「臨床意義／檢查目的」、「治療時間」、「療效」、「恢復時間」、「正常指標」、「化驗／體檢方案」與「恢復」，共 903 筆，占 7.37%；「其他」類別包含 7 個子類別，分別爲「設備用法」、「多問」、「養生」、「整容」、「兩性」、「對比」與「無法確定」，共 2,331 筆，占 19.02%。上述四個主類別合計共有 12,254 筆資料。由資料分布可看出，「病症」類別所占比例最高，達 51.12%；相較之下，「治療方案」類別的樣本數最少，僅占 7.37%，顯示 CMID 存在一定程度的類別不平衡。

本研究分別設定粗粒度與細粒度兩項分類任務。粗粒度任務以 4 個主類別作爲預測目標，主要評估模型辨識整體醫療問題類型的能力；細粒度任務則以 36 個子類別作爲預測目標，用以檢驗模型區分具體詢問意圖的能力，不僅類別數量較多，部分子類別之間也具有相近的語意內容，加上樣本分布不均，因此具有更高的分類難度。

4.2 實驗設定

本研究將 EKG-Transformer 的訓練流程分爲百科全書預訓練與知識圖譜微調兩個階段。百科全書預訓練階段首先蒐集百度醫學百科中的層級資料，經過資料清理、焦點與面向定義及長度篩選後，將語料分別用於 BERT[1] (Devlin et al., 2019)、RoBERTa[2] (Cui et al., 2020) 與 MacBERT[3] (Cui et al., 2020) 的持續預訓練，基底模型來自 HuggingFace。此階段的初始學習率設爲 2×10^{-5}，批次大小爲 24，最大序列長度爲 512；BERT 與 RoBERTa 的訓練步數均爲 10,000 steps，而 MacBERT 的訓練步數爲 25,000 steps。由於 MacBERT 開源模型未包含完整的遮罩語言模型權重，因此相關參數需重新初始化，訓練步數也相對增加。

完成百科全書預訓練後，所得模型參數用於初始化知識圖譜微調階段。此階段先依據 CMID 問句中的命名實體擷取相關醫療知識圖譜三元組，建構 sentence graph，並結合 soft-position embedding 與 visible matrix 進行分類訓練。知識圖譜微調階段的初始學習率設爲 2×10^{-5}，Dropout rate 爲 0.5，批次大小爲 12，最大序列長度爲 512，訓練 20 epochs。爲降低單次資料切分造成的偏差，所有實驗皆採用 stratified 5-fold cross-validation，並以各折結果的平均值作爲最終評估結果。

本研究所使用之醫療知識圖譜整合自兩個來源：(1) 一個以疾病爲中心的既有醫療知識圖譜 (Liu, 2018)，其實體規模約 4.4 萬、實體關係規模約 30 萬，本研究進一步挑選其中與醫療問答資料集相關的 5 個關係：疾病治療方式、疾病症狀、疾病併發症、屬於科別與疾病所需檢查；(2) Chinese HealthNER 命名實體標註語料 (Lee and Lu, 2021)，涵蓋人體、症狀、醫療器材、檢驗、化學物質、疾病、藥品、營養品、治療、時間等 10 種實體類型，並額外建立「類別」關係將實體與其類型相連。兩來源合併後，最終知識圖譜規模約 15.5 萬筆實體關係，供知識圖譜微調階段檢索使用。

4.3 比較模型

爲驗證所提出方法的有效性，本研究將比較模型分爲四類。第一類爲非預訓練式文本分類模型，包括 FastText (Joulin et al., 2017)、TextCNN (Kim, 2014)、TextRNN (Huang et al., 2015) 與 TextGCN (Yao et al., 2019)，用以評估傳統深度學習方法的分類表現。第二類爲中文預訓練語言模型，包括 BERT(Devlin et al., 2019)、RoBERTa (Cui et al., 2020) 與 MacBERT (Cui et al., 2020)，用以作爲 EKG-Transformer 的基礎模型。第三類爲領域知識強化模型，其中 +Ency 表示加入百科知識增強預訓練，+KG 表示加入知識圖譜微調，EKG-Transformer 表示同時結合百科概念知識與知識圖譜關係知識。第四類爲大型語言模型，本研究以 Llama3.1-8B-Instruct (Grattafiori et al., 2024) 作爲比較模型，並依序測試 Zero-Shot 與監督式微調（Supervised Fine-Tuning, SFT）兩種設定，用以觀察通用大型語言模型在不同監督程度下的中文醫療問句意圖分類能力。Zero-Shot 設定不使用 CMID 訓練資料，直接透過任務提示要求模型輸出預測類別；SFT 設定進一步以 CMID 訓練資料進行參數微調，用以評估在充分監督訊號下所達到的分類表現上限。此外，爲進一步涵蓋近期發展之開源大型語言模型，本研究額外納入 Gemma-4-E4B-it[4]（Google DeepMind, 2026）於 Zero-Shot 設定下之結果作爲對照，以評估更新一代通用模型在缺乏領域監督訊號時的表現。

4.4 評估指標

本研究採用 Macro-F1、Micro-F1 與 Weighted-F1 作爲評估指標。Macro-F1 分別計算各類別的 F1-score 後取平均，使每個類別具有相同權重，因此較能反映模型對少數類別的辨識能力。Micro-F1 先累計所有類別的預測結果，再計算整體 F1-score，主要反映模型在全部樣本上的分類表現。Weighted-F1 則依各類別的樣本數對類別 F1-score 加權，可反映資料分布不平衡情況下的整體效能。

4.5 實驗結果

表 1呈現各模型於 CMID 4 類別與 36 子類別任務上的主要實驗結果。在非預訓練式模型中，TextCNN 在兩項任務的表現均爲最佳，預訓練語言模型藉由大規模語料所學得的上下文表示，能有效提升中文醫療問題的分類效果。同時可觀察到，相較於 4 類別任務，所有模型在 36 子類別任務中的 F1-score 均明顯下降，反映細粒度類別之間的語意邊界較爲相近，且部分子類別存在資料不平衡問題，因此分類難度較高。

在完整加入百科全書預訓練與知識圖譜微調後，EKG-BERT、EKG-RoBERTa 與 EKG-

[1] https://huggingface.co/bert-base-chinese

[2] https://huggingface.co/hfl/chinese-roberta-wwm-ext

[3] https://huggingface.co/hfl/chinese-macbert-base

[4] https://huggingface.co/google/gemma-4-E4B-it

Model	Macro-F1		Micro-F1		Weighted-F1	
	4-class	**36-class**	**4-class**	**36-class**	**4-class**	**36-class**
TextGCN	0.3702	0.2566	0.5691	0.3789	0.5218	0.3663
TextRNN	0.4391	0.1205	0.6346	0.3252	0.5981	0.2952
FastText	0.4755	0.0686	0.6670	0.2573	0.6068	0.1978
TextCNN	0.5096	0.2733	0.6808	0.4666	0.6461	0.4393
BERT	0.5931	0.4189	0.6994	0.5063	0.6893	0.5011
EKG-BERT (ours)	**0.6370**	**0.4293**	**0.7324**	**0.5347**	**0.7226**	**0.5217**
RoBERTa	0.6216	0.4308	0.7131	0.5086	0.7076	0.5046
EKG-RoBERTa (ours)	**0.6381**	**0.4436**	**0.7358**	**0.5357**	**0.7267**	**0.5246**
MacBERT	0.6260	0.4211	0.7184	0.5146	0.7136	0.5100
EKG-MacBERT (ours)	**0.6325**	**0.4359**	**0.7450**	**0.5370**	**0.7284**	**0.5246**
Llama3.1-8B-Instruct (Zero-shot)	0.3801	0.1340	0.5682	0.2299	0.5345	0.1695
Llama3.1-8B-Instruct (SFT)	**0.6584**	**0.4754**	**0.7456**	**0.5650**	**0.7394**	**0.5567**
Gemma-4-E4B-it (Zero-shot)	0.4330	0.3452	0.5610	0.3909	0.5578	0.3744

表 1: 中文醫療問題意圖分類主實驗結果。

Model	Base		Base +Ency		Base +KG		EKG-Transformer (Base +Ency +KG)	
	4-class	**36-class**	**4-class**	**36-class**	**4-class**	**36-class**	**4-class**	**36-class**
BERT	0.6994	0.5063	0.7153	0.5177	0.7217	0.5250	**0.7324**	**0.5347**
RoBERTa	0.7131	0.5086	0.7211	0.5186	0.7275	0.5288	**0.7358**	**0.5357**
MacBERT	0.7184	0.5146	0.7237	0.5215	0.7348	0.5322	**0.7450**	**0.5370**

表 2: EKG-Transformer 模型消融分析結果。

MacBERT 的 Micro-F1 在 4 類別任務分別提升至 0.7324、0.7358 與 0.7450，在 36 子類別任務則分別提升至 0.5347、0.5357 與 0.5370，均優於相對應的基礎模型。在 Weighted-F1 方面，4 類別任務以 EKG-MacBERT 最高 (0.7284)，36 子類別任務則由 EKG-RoBERTa 與 EKG-MacBERT 並列最高 (0.5246)；EKG-RoBERTa 則在兩項任務皆取得最高的 Macro-F1 (0.6381 & 0.4436)。此結果顯示，領域知識強化不僅能改善模型在整體樣本上的分類表現，也能提升模型對不同類別的平均辨識能力，且在粗粒度與細粒度醫療意圖分類任務中皆能有效補強通用預訓練模型的語意表示能力。

在 Zero-shot 設定下，Llama3.1-8B-Instruct 不使用任何 CMID 訓練樣本，直接透過 prompt 引導模型輸出預測類別。結果顯示，其在 4 類別任務的 Micro-F1 爲 0.5682，36 子類別任務爲 0.2299，均低於經過監督式微調的預訓練模型，與Mullick et al. (2024) 的觀察一致，說明通用型 LLM 在專業醫療意圖分類任務中，仍難以取代針對領域微調的模型。

在 SFT 設定下，本研究以 LoRA (Low-Rank Adaptation) (Hu et al., 2022) 方式對 Llama3.1-8B-Instruct 進行監督式微調，採用 5-fold 交叉驗證：每折獨立訓練一個模型，最終以 5 折結果的平均作爲報告數字。訓練採 4-bit 量化載入基礎模型，LoRA rank 設爲 16、alpha 爲 32，學習率 1e-4，訓練 3 個 epoch，於單張 NVIDIA RTX A4000 GPU (16GB memory) 上執行；4 類別任務 5 折合計訓練耗時約 32.7 小時，36 子類別任務 5 折合計訓練耗時約 39.2 小時。

結果顯示，Llama3.1-8B-Instruct (SFT) 在 4 類別任務取得 Macro-F1 0.6584、Micro-F1 0.7456 與 Weighted-F1 0.7394，與表現最佳的 EKG-MacBERT (Micro-F1 0.7450) 幾乎持平；在 36 子類別任務則取得 Macro-F1 0.4754、Micro-F1 0.5650 與 Weighted-F1 0.5567，略高於 EKG-MacBERT。此結果顯示，當提供充足的領域監督訊號時，具備 80 億參數規模的通用大型語言模型，即使未注入任何額外的百科知識或知識圖譜關係資訊，也能透過參數高效微調達到與領域知識強化後的輕量級預訓練模型相當、甚至略優的分類表現。

另一方面，Gemma-4-E4B-it 於 Zero-Shot

案例	輸入文本	預測類別	正確類別
1	尿常規檢查有隱血：125	病症	治療方案
2	昨天檢查出了肺結核，目前這種情況嚴重嗎？可以痊癒嗎？	病症	其他
3	肝腎滋主要是用於什麼方面的治療呢，效果好不好的？	適用症	作用
4	胃癌在患病前會發出哪些信號？	臨床表現	相關病症
5	我今年 44，腦梗四個月了，左邊身子不方便動，現在拄著枴棍可以走了，可是多久才能恢復正常呀？	治療時間	恢復時間

表 3: EKG-MacBERT 之代表性錯誤案例

設定下取得 4 類別 Micro-F1 0.5610、36 子類別 Micro-F1 0.3909，相較於 Llama3.1-8B-Instruct 之 Zero-Shot 結果（Micro-F1 0.5682 ／ 0.2299）在 36 子類別任務上有大幅提升（+0.1610）。然而，即使換用更強的新一代模型，在缺乏領域監督訊號的 Zero-Shot 設定下，其表現仍明顯低於本研究以少量監督訓練之 EKG-MacBERT（36 子類別 Micro-F1 0.5370）。此結果進一步支持本研究之核心論點：領域知識強化機制相較於單純仰賴更強大的通用語言模型，在缺乏充分監督訊號時仍具有明確優勢。

然而，此結果仍須搭配運算成本一併解讀。Llama3.1-8B-Instruct 與 Gemma-4-E4B-it 等大型語言模型的參數量約爲 BERT、RoBERTa 與 MacBERT （約 1 億參數）的數十倍以上；以 Llama3.1-8B-Instruct 爲例，兩項任務合計需於單張 NVIDIA RTX A4000 GPU 上訓練約 3 天，且在 Zero-shot 設定下之平均推論延遲約爲每題 0.68 秒，Gemma-4-E4B-it 於 Zero-shot 設定下之平均推論延遲則約爲每題 3.07 至 3.47 秒；相較之下，本研究 BERT 系列模型之知識圖譜微調僅需批次大小 12、訓練 20 epochs，可於數小時內完成。換言之，EKG-Transformer 透過輕量的知識注入機制，以遠低於直接微調大型語言模型的運算成本，達到與其相近的分類效果，這對運算資源受限的實務環境而言，仍具有明確的效率優勢與應用價值。

爲驗證模型間效能差異是否具統計顯著性，本研究進一步以 Bowker 檢定（McNemar 檢定之多類別推廣版本，又稱對稱性檢定）分別比較 EKG-BERT、EKG-RoBERTa 與 EKG-MacBERT 於 4 類別（自由度 6）與 36 子類別（自由度 630）任務上相對於對應基礎模型之預測結果，P 值分別爲 1.35e-39、1.75e-56 與 1.26e-75（4 類別）及 2.77e-08、2.45e-09 與 4.23e-02（36 子類別），皆遠小於 0.05，顯示知識強化後的效能提升具統計顯著性。

4.6 消融分析

表 2 比較不同領域知識來源對分類結果的影響，效能指標爲 Micro-F1，Base 表示未加入額外領域知識的原始預訓練模型；+Ency 表示加入百科知識增強預訓練；+KG 表示加入知識圖譜微調；EKG-Transformer 表示同時結合百科概念知識與知識圖譜關係知識。在 4 類別任務中，+Ency 設定可提升 BERT、RoBERTa 與 MacBERT 的 Micro-F1，此結果顯示百科概念與面向資訊有助於模型建立醫療概念表示，進而改善下游意圖分類表現。相較之下，+KG 設定的提升幅度更爲明顯，表示醫療知識圖譜中的實體關係能補充問句文字上下文中較難直接取得的關聯資訊。

當百科概念知識與知識圖譜關係知識同時使用時，三種基礎模型皆取得最佳結果。EKG-BERT、EKG-RoBERTa 與 EKG-MacBERT 的 Micro-F1 分別達到 0.7324、0.7358 與 0.7450，相較於原始模型的相對提升幅度分別爲 4.72%、3.18% 與 3.70%。其中，EKG-MacBERT 取得 4 類任務中的最佳表現。

在 36 子類別任務中亦可觀察到相同趨勢。百科概念知識主要強化模型對醫療概念與面向屬性的理解，知識圖譜關係知識則補充醫療實體之間的顯式關係。兩種知識來源具有互補效果，且在粗粒度與細粒度分類任務中皆能穩定改善模型表現。

4.7 錯誤分析

雖然 EKG-Transformer 能提升整體分類效能，但模型在部分問句上仍會產生誤判。以下先分析 EKG-MacBERT 於 CMID 粗粒度與細粒度分類任務中比例較高的主要誤判方向，並於表 3 中列出部分具代表性的錯誤案例。

在 4 類別任務中，「治療方案」與「其他」較容易被模型誤判爲「病症」，其誤判比例分別爲 52.71% 與 35.01%。表 3 的案例 1 僅包含檢查結果，缺乏完整的詢問內容，因此模型較難判斷其實際意圖。案例 2 雖包含「嚴重」與「痊癒」等病症相關內容，但由於同時具有多個詢問意圖，依 CMID 的標註規則應歸入

「其他」類別中的「多問」，因此模型仍產生誤判。

在 36 子類別任務中，模型較容易混淆語意相近的類別。「方法」被預測爲「治療方法」的比例最高，達 49.88%；此外，「作用」被預測爲「適用症」、「相關病症」被預測爲「臨床表現」，以及「恢復時間」被預測爲「治療時間」的比例分別爲 29.73%、28.93% 與 20.00%。表 3 的案例 3、4 分別呈現「作用」被預測爲「適用症」及「相關病症」被預測爲「臨床表現」的誤判情形，而案例 5 則顯示「恢復時間」與「治療時間」之間亦容易產生混淆。整體而言，這些錯誤主要源於類別定義與語意相近，加上醫療問句多爲口語化表達，使模型難以準確區分細粒度意圖類別。

5 結論與未來工作

本研究針對中文醫療問句意圖分類任務提出 EKG-Transformer，結合百科概念知識與醫療知識圖譜關係知識，以改善預訓練語言模型對醫療問句的理解能力。模型先透過百科知識增強方式取得醫療概念與屬性表示，再於下游分類階段注入醫療知識圖譜三元組，並結合 sentence graph、soft-position embedding 與 visible matrix，以補充醫療實體之間的關係資訊，同時降低知識噪聲對原始句意的干擾。

然而，本研究仍存在部分限制。首先，部分 CMID 問句過短或同時包含多重意圖，容易造成模型判斷困難；其次，36 個子類別之間存在資料不平衡與語意邊界重疊問題；此外，知識圖譜三元組若與當前問句語境關聯較低，仍可能引入知識噪聲。未來可進一步設計三元組關聯性排序與篩選機制，並透過資料增強、類別重加權或階層式分類改善細粒度分類表現。所提出的方法亦可延伸至醫療問答、醫療資訊檢索與醫療對話系統，作爲後續知識檢索與答案推薦的前置模組。

Acknowledgments

This study was partially supported by the National Science and Technology Council, Taiwan, under the grant NSTC 114-2221-E-A49-059-MY3.

References

Guirong Bai, Shizhu He, Kang Liu, and Jun Zhao. 2022. Incremental intent detection for medical domain with contrast replay networks. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3549–3556. Association for Computational Linguistics.

Nan Chen, Xiangdong Su, Tongyang Liu, Qizhi Hao, and Ming Wei. 2020. A benchmark dataset and case study for Chinese medical question intent classification. *BMC Medical Informatics and Decision Making*, 20(Suppl 3):125.

Po-Han Chen, Yu-Xiang Zeng, and Lung-Hao Lee. 2021. Incorporating domain knowledge into language transformers for multi-label classification of Chinese medical questions. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 265–270. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Xinping Chen and Yan Du. 2025. Enhancing medical text classification with GAN-based data augmentation and multi-task learning in BERT. *Scientific Reports*, 15:13854.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yanjun Gao, Ruizhe Li, Emma Croxford, John Caskey, Brian W. Patterson, Matthew Churpek, Timothy Miller, Dmitriy Dligach, and Majid Afshar. 2025. Leveraging medical knowledge graphs into large language models for diagnosis prediction: Design and application study. *JMIR AI*, 4:e58670.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,

and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

U. Jain, P. Mishra, A. Dash, and A. Pandey. 2025. Multi-label multi-class text classification-enhanced attention in transformers with knowledge distillation. *Journal of Applied Research and Technology*, 23(1):82–93.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Lung-Hao Lee and Yi Lu. 2021. Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition. *IEEE Journal of Biomedical and Health Informatics*, 25(7):2801–2810.

Huanyong Liu. 2018. QASystemOnMedicalKG. `https://github.com/liuhuanyong/QASystemOnMedicalKG`. GitHub repository.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Prabhashrini Dhanushika Manage, Yutong Wu, Jinglan Zhang, Anupama Udayangani Gunathilaka, and Yuefeng Li. 2026. Overcoming BERT's limitations in uncertainty: A novel two-stage solution for multi-class medical text classification. *Knowledge-Based Systems*, 336:115315.

Ankan Mullick, Mukur Gupta, and Pawan Goyal. 2024. Intent detection and entity extraction from biomedical literature. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 271–278. ELRA and ICCL.

Oussama Ndama, Ismail Bensassi, Safae Ndama, and El Mokhtar En-Naimi. 2025. Enhancing multi-class text classification in biomedical literature by integrating sequential and contextual learning with BERT and LSTM. *International Journal of Electrical and Computer Engineering*, 15(4):4202–4212.

Farhana Hossain Swarnali, Jannatim Maisha, Muhammad Azmain Mahtab, M. Saymon Islam Iftikar, and Faisal Muhammad Shah. 2024. Bengali multi-class text classification via enhanced contrastive learning techniques. In *2024 27th International Conference on Computer and Information Technology (ICCIT)*, pages 1576–1581. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Yinuo Wang, Baiyang Wang, Robert Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025. Trustworthy medical question answering: An evaluation-centric survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27489–27502. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7888–7915. Association for Computational Linguistics.

Kaiwen Zuo, Yirui Jiang, Fan Mo, and Pietro Lio. 2025. KG4Diagnosis: A hierarchical multi-agent LLM framework with knowledge graph enhancement for medical diagnosis. In *Proceedings of the First AAAI Bridge Program on AI for Medicine and Healthcare*, volume 281 of *Proceedings of Machine Learning Research*, pages 195–204. PMLR.