運用長句簡化及少樣本學習以提升大型語言模型辨識蛋白質交互作用準確性

Enhancing Protein-Protein Interaction Recognition Accuracy in Large Language Models through Sentence Reduction and Few-Shot Learning

Yi-Yun Chou, Li-Kai Chen, Wei-Ren Liu, Hong-Jie Dai* and Ching-Tai Chen

Department of Bioinformatics and Medical Engineering, Asia University
*Department of Electrical Engineering, National Kaohsiung University of Science and Technology
choucandy998@gmail.com, a0920324955@gmail.com, sssss1050287@gmail.com,
hjdai@nkust.edu.tw, ctchen@asia.edu.tw

摘要

蛋白質交互作用（Protein-Protein Interaction，PPI）辨識是生醫文獻探勘的重要課題。近年來具備通用知識的預訓練大型語言模型（Large Language Model，LLM）能夠透過 Prompt 快速運用於下游任務的特性，已廣泛運用於各種自然語言處理議題。本研究透過 Prompt Engineering 提出 Sentence Reduction 和 Few-Shot 方法引導五種 LLM 進行 PPI 辨識，並驗證其效能，結果顯示本文的方法可有效提升 LLM 的效能，在三組語料上的 F1-Score 均達 0.8 以上，優於過往傳統方法，並揭示不同的 Prompt 策略和 Few-Shot 對提高準確度的重要影響。

Abstract

Protein-Protein Interaction （PPI） recognition is a crucial task in biomedical literature mining. In recent years, pre-trained Large Language Models （LLMs） with general knowledge have been widely applied to various natural language processing tasks due to their ability to quickly adapt to downstream tasks through prompting. This study proposes Sentence Reduction and Few-Shot methods through Prompt Engineering to guide five LLMs in PPI recognition and evaluates their performance. The results show that our approach can effectively enhance LLM performance, achieving F1-Scores above 0.8 on three datasets, outperforming traditional methods. The study also reveals the significant impact of different prompting strategies and few-shot learning on improving accuracy.

關鍵字：大型語言模型、文獻探勘、提詞工程、句子簡化、少樣本學習

Keywords: Large Language Model, Text Mining, Prompt Engineering, Sentence Reduction, Few-Shot Learning

1 Introduction

今日生物醫學研究領域中書目資料的數量持續大幅成長，這些文獻蘊含大量研究成果，對生醫領域具有重要參考價值（N.-W. Chang 2020），然而，面對浩如江海的資料量，研究人員往往需投入大量人力及時間以取得重要資訊，近年運用自然語言處理技術（Natural Language Processing，NLP）作生物醫學文獻探勘有重大貢獻，例如從文獻中自動提取生物實體間相互作用（Warikoo, Chang, and Ma 2022），其中蛋白質交互作用（Protein-Protein Interaction，PPI）的辨識，可為生醫研究、新藥開發、癌症免疫、精準治療等提供相當重要的資訊。目前研究的熱點為結合注意力機制和特徵向量並進行深度學習（Liu and Guo 2019），但這種這種方法需要調整大量的超參數，過程中不僅耗時，且重複訓練往往需要大量的計算資源，增加使用難度（Li et al. 2021）。而近年來大型語言模型（Large Language Model，LLM）的興起（Minaee et al. 2024），不但結合深度學習的優勢（Du et al. 2024; Wolf et al. 2020），同時可以藉由透過提詞工程（Prompt Engineering）設計並優化 Prompt（Schulhoff et al. 2024）提升 LLM 的學習能力（Chen et al. 2023），並進行各種下游任務，現今已在 NLP 領域上廣泛利用（Deng et al. 2022）。

本研究透過 Prompt Engineering，測試了三種不同 Prompt 的設計概念，分別為：（1）Sentence Reduction：主要包含刪除句子冗餘成分、合併簡化後句子、句法結構轉換、同義詞替換、以更精煉的描述取代原有敘述，重構句子成分（Jing and McKeown 2000）。透過從複雜文本中提煉出關鍵信息，將長句濃縮為精簡短句，同時保持原意（Nguyen et al. 2004; Feng et al. 2023）。本研究中，我們運用 LLM 實現 Sentence Reduction 並保留與 PPI 相關資訊；（2）Few-Shot Learning：透過提供少量與任務相關的範例，協助 LLM 透過類比學習指定任務的回答邏輯（Brown et al. 2020），指導 LLM 能夠更精確地進行下游目標執行任務；（3）One-Stage 及 Two-Stage Prompting：基於前述兩種方法，One-Stage Prompting 策略於提示中整合了 Sentence Reduction 與辨識 PPI 這兩個核心任務，而 Two-Stage Prompting 策略則將這兩個任務分開獨立執行。結合上述三項方法，我們共設計出 15 種 Prompt，並使用五個廣為人知的公開商用 LLM 及三組資料集進行驗證及性能比較。

2 Related Work

2.1 LLM 在 PPI 辨識任務上的應用

有多項相關研究展示了 LLM 在生物醫學文本探勘中的潛力。Park 等人（2023）對 LLM 應用在 PPI 辨識任務上進行全面的評估，研究中使用多個 LLM 在 STRING 數據庫進行 PPI 辨識。他們設計兩項子任務：一是要求 LLM 生成給定語句中的蛋白質列表（PPI Task1），二是由 LLM 判斷兩個蛋白質是否存在相互作用（PPI Task2）。在 PPI Task1 中，研究者觀察到 LLM 傾向於根據給定蛋白質的首字母生成相關蛋白質名稱，導致 LLM 對相似名稱蛋白質能準確預測，但對不相似名稱的蛋白質則表現較差。對於 PPI Task2，研究者透過建構的平衡數據集評估 LLM 判斷語句中提及的兩蛋白質是否有 PPI 的能力，根據實驗結果 LLM 在 PPI Task2 可達 0.5-0.984 的 F1-Score，反映了 LLM 在二元分類的潛力。

同樣探討 LLM 在 PPI 辨識任務中的應用（Rehana et al., n.d.），該研究不僅驗證 BERT 模型在 PPI 識別任務中的高效能（在 LLL 語料庫可達 0.868 的 F1-Score），也展示了 GPT 模型在此領域中的應用價值。儘管 GPT 模型並非專門為生物醫學領域分析而設計，GPT4 在該研究中仍展現出與 BERT 相當的性能，在 LLL 語料庫達到 0.864 的 F1-Score，證明 GPT 模型具備從非結構化文本中有效識別 PPI 的能力。

2.2 Sentence Reduction 改善 PPI 識別效能

Jonnalagadda 和 Gonzalez（2009）開發一個名為 bioSimplify 的句子簡化工具，該工具基於語法規則執行以下四個主要步驟：移除無關短語、替換基因名稱、替換名詞短語以及基於依存關係的句子分割。在評估中，他們結合 bioSimplify 與 PIE 系統（Kim et al. 2008）做為基準 PPI 辨識工具進行實驗。根據結果顯示其系統的召回率（Recall）提高了 8%，F1-Score 提高 3 個百分點，凸顯句子簡化確可提升 PPI 抽取系統的效能。

3 Method

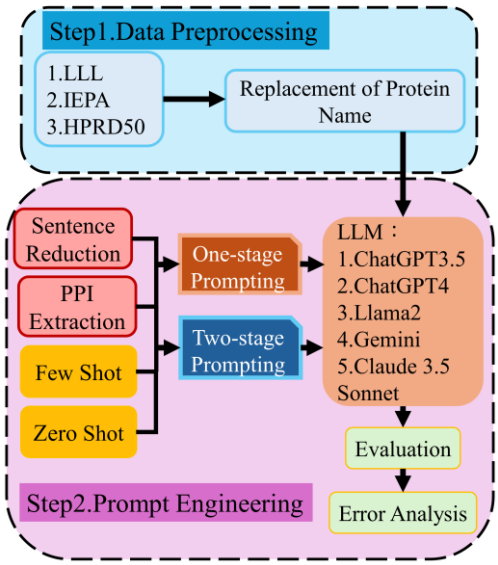

圖 1.實驗流程圖

儘管先前的研究已經證實 LLM 和 Sentence Reduction 在 PPI 識別任務上的有效性，但目前文獻中仍然缺乏將這兩種方法協同應用於提詞工程中以辨識 PPI。於此節我們闡述如何透過運用 LLM 實現 Sentence Reduction 並結合 Few-Shot Learning 指引 LLM 精確執行 PPI 辨識任務。圖 1 為本研究的實驗流程圖。在 Data Preprocessing 中，我們對 PPI 語料庫中的原始文本進行 Replacement of Protein Name，將蛋白質和基因名稱標記為特定的詞彙。在第二

個階段 Prompt Engineering 中，本研究以 Sentence Reduction 、PPI Extraction、Few-Shot Learning 為核心，開發出一系列不同的 Prompt，並把開發的 Prompt 分為一次作完 Sentence Reduction 及 PPI Extraction 的 One-Stage Prompting ，與分別用兩次 Prompt 依序完成前述兩項任務的 Two-Stage Prompting。為驗證本文提出的 prompting 方法是否能優化 LLM 提取 PPI 的能力，搭配不同策略，本研究選用五個熱門的商用 LLM 進行測試，包含 Google Gemini Pro （Gemini Team et al. 2024），Meta Llama2 （Touvron et al. 2023, 2），Anthropic Claude3 Sonnet （Caruccio et al. 2024），OpenAI ChatGPT3.5，以及 OpenAI ChatGPT4。（Budzianowski and Vulić 2019; Floridi and Chiriatti 2020; Kalyan 2024; OpenAI et al. 2024; Wu et al. 2023），並對 LLM 的輸出進行評估與錯誤分析。以下章節針對前述流程中各子步驟進行詳細說明。

3.1 Dataset

本研究採用三個在 PPI 研究領域中廣泛使用的語料庫：LLL（Learning Language in Logic）（Nédellec and Nedellec, n.d.）、IEPA（Ding et al. 2002）、HPRD50（Peri et al. 2004）。這些語料庫被公認為評估 PPI 辨識準確率的標準數據集（Y.-C. Chang et al. 2016）。LLL 語料庫源於 2005 年同名研討會，其內容主要摘自 Medline 數據庫中的生物學文獻摘要，該語料庫的訓練集包含 269 個句子、測試集包含 61 個句子。IEPA 語料庫由 303 篇 PubMed 摘要組成，其中訓練集含有 681 個句子、測試集包含 136 個句子。HPRD50 語料庫的建構基於 Human Protein Reference Database，研究者從中隨機選取了 50 篇摘要，訓練集包含 363 個句子，測試集包含 70 個句子。

3.2 Replacement of Protein Name

在此步驟，我們根據語料庫中提供的蛋白質名稱標註資訊將每個句子中的任兩個蛋白質或基因名稱分別替換為 PROTEIN1 和 PROTEIN2 標籤。然而有些句子含超過兩個蛋白質或基因，對於其它蛋白質或基因名稱我們使用一系列泛化標籤，如 MOLECULE、MATERIAL、SUBSTANCE、PARTICLE、INGREDIENT 等進行替換，以保持句子語義結構，同時減少其餘非目標實體對 LLM 造成注意力分散的現象。對於包含多個蛋白質和基因名稱的句子，本研究採用排列組合的方法，這意味著同一原始句可能產生多個變體句子，每個變體句子中 PROTEIN1 和 PROTEIN2 標籤所代表的蛋白質實體會有所不同。透過這種方法能夠捕捉到同一語境下不同蛋白質對之間的潛在交互作用。在後續實驗中，這些通過排列組合的變體句子被視為獨立樣本（亦為 3.1 節中測試資料的樣本數），個別驗證正確率。

3.3 Prompt Engineering

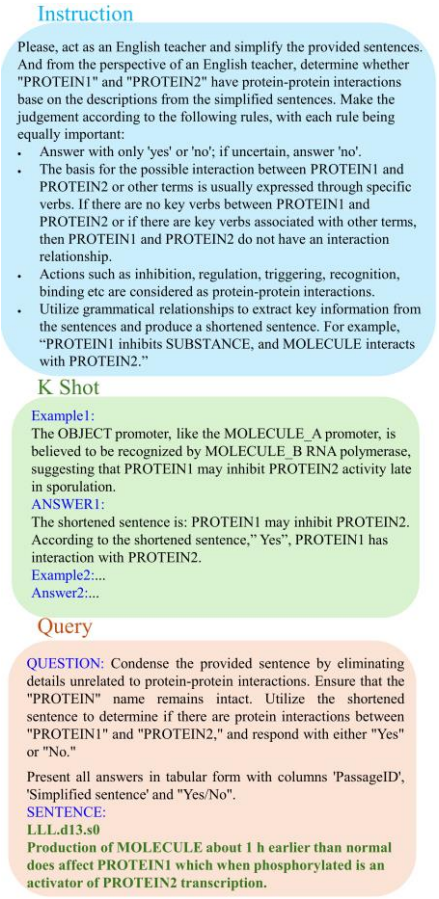

圖 2. Prompt 設計之模板

圖 2 顯示設計的 Prompt 主要核心模板，其包括兩個任務：第一個任務為 Sentence Reduction，執行句子簡化；第二個任務則是 PPI Extraction，專注於從文獻中識別並提取 PPI。我們將設計的 Prompt 模板分成三個部分，包含 Instruction 及 Few-shot Learning 以及最後一個部分的 Query：

1. Instruction：如圖 2 的 Instruction 部分所示，我們參考 Sivarajkumar 等人（2023）提出的 Heuristic Prompts 概

念，將任務轉化為一系列的指引規則，引導 LLM 完成任務。

2. Few-shot Learning：如圖 2 的 K Shot 部分所示，我們在此部分的模板中提供從 Training Data 抽取出來的 1~20 個固定範例做為 LLM 執行任務的參考。

3. Query：於此我們加入指令要求 LLM 以二元形式回答（Yes/No），並限制其輸出以表格方式呈現。最後於此部分提供要處理的句子給 LLM 執行任務。

3.4 Sentence Reduction

為了避免 LLM 被句子中過多的訊息分散注意力，本研究提出運用 Sentence Reduction 的方法讓 LLM 能夠基於語意及語法分析保留與蛋白質相關的句子片段，刪除其餘無關詞彙，透過這樣的方法來縮短冗長的句子，從而得到簡化的句子。為達成上述的目標，我們提出了三項規則：

- The essential information to be extracted pertains to the mutual interactions among PROTEIN1, PROTEIN2, MOLECULE, SUBSTANCE, ELEMENT, FACTOR, and MATERIAL. Information regarding interactions between any of the above items should be retained.
- Utilize grammatical relationships to extract key information from the Sentences and produce a shortened Sentence. For example, “PROTEIN1 inhibits SUBSTANCE, and MOLECULE interacts with PROTEIN2.”
- Actions such as inhibition, regulation, triggering, recognition, binding etc are considered as protein-protein interactions.

3.5 PPI Extraction

針對 PPI Extraction 的任務，我們提出以下規則：

- Answer with only 'yes' or 'no'; if uncertain, answer 'no'. Each sentence is independent; please make judgments only on the 'PROTEIN1' and 'PROTEIN2' from each sentence.
- If there is no direct interaction between "PROTEIN1" and "PROTEIN2" but they interact with MOLECULE, SUBSTANCE, ELEMENT, FACTOR, and MATERIAL, the answer is 'no,' as the focus is solely on the interaction between "PROTEIN1" and "PROTEIN2" in the sentence.
- The basis for the possible interaction between PROTEIN1 and PROTEIN2 or other terms is usually expressed through specific verbs. If there are no key verbs between PROTEIN1 and PROTEIN2 or if there are key verbs associated with other terms, then PROTEIN1 and PROTEIN2 do not have an interaction.
- Actions such as inhibition, regulation, triggering, recognition, binding etc are considered as protein-protein interactions.

3.6 One-Stage Prompting and Two-Stage Prompting

鑒於將複雜的任務分解成多個簡單的子任務可以優化 LLM 執行任務的能力（Schulhoff

Prompt	Method	Sentence Reduction	K shot for SR	PPI Extraction	K shot for PE
Prompt Type Ⅰ	One-Stage Prompting	N	N	Y	K=0、1、3、10、20
Prompt Type Ⅱ	One-Stage Prompting	Y	K=0、1、3、10、20	Y	K=0、1、3、10、20
Prompt Type Ⅲ	Two-Stage Prompting	Y	K=3	Y	K=0、1、3、10、20

表 1.表格為實驗設計的 Prompt 實驗種類：Y 和 N 分別代表有或無運用此技術、K Shot 欄位中數字為 Few-Shot 的範例個數設置、Method 則為使用策略。SR 代表 Sentence Reduction，PE 代表 PPI Extraction

et al. 2024），我們設計並實驗了兩種不同的策略：One-Stage Prompting 和 Two-Stage Prompting，這兩種策略的主要區別在於任務執行的結構。圖 2 所展示 Prompt 屬於 One-Stage Prompting，其將 Sentence Reduction 及 PPI Extraction 兩個任務整合為一個 Prompt 來進行任務，而 Two-Stage Prompting 則將圖 2 拆解成兩個 prompt；先進行 Sentence Reduction，從 LLM 得到簡化後的句子後，再透過第二個 Prompt 對簡化句進行 PPI Extraction。

3.7 Few-Shot Learning

在設計 Prompt 中，本研究採用 Few-Shot Learning，從 Training Data 中選取固定範例作為 K Shot，我們設置不同的 K 值以評估不同 Shot 數量對 LLM 學習效果的影響。表 1 展示了本研究設計的 Prompt 種類，Prompt Type I 為不含 Sentence Reduction 與 Few-Shot Learning 的 baseline 方法，Prompt Type II 為 One-Stage Prompting，在 Sentence Reduction 與 PPI Extraction 任務皆設置了 K=1、3、10、20 四種情況，使用上保持兩步驟的 K 值一致。Prompt Type III 為 Two-Stage Prompting，Sentence Reduction 的 K-Shot 設定為 K=3，因其相較於其它數字（1, 10, 20）準確率最高，PPI Extraction 的設置則與 Prompt Type I 相同。根據 Brown 等人（2020）提到 LLM 會因為 Prompt 中出現頻率較高的標籤而產生偏差，為了防止 LLM 在判斷時產生偏見，Prompt 設計特別考慮了範例中 Positive 和 Negative 例句的平衡，並制定了使用範例的比例規範：

- 當 K=1 時，提供一個 Positive 例子。
- 當 K=3 時，提供兩個 Positive 例子及一個 Negative 例子。
- 當 K=10 或 20 時，Positive 和 Negative 例子各占半數。

3.8 Zero-Shot Learning

除了前述的 K Shot 方法外，我們也嘗試了 Zero-Shot Learning；在這個方法中 LLM 僅接受描述任務的 Prompt（只包含圖 2 中的 Instruction 與 Query 的部分），而不提供任何範例。由於 Zero-Shot Learning 適用於缺乏大規模訓練數據的情境，僅向 LLM 提供任務的描述，此方法十分依賴 LLM 固有的能力（Radford et al., n.d.），因此在實驗中，我們藉由 Zero-Shot Learning 與結合 Sentence Reduction 及 Few-Shot Learning 的 Prompt 比較，以瞭解不同 LLM 對於 PPI 任務本身的理解程度，並探討 Few-Shot Learning 帶來的進步幅度。表 1 中的 Prompt Type I 設定 PPI Extraction 的 K=0 即為 Zero-Shot Learning。

3.9 Evaluation

本研究採用 Precision、Recall 及 F1-Score 指標評估所有實驗的結果。

$$Precision = \frac{the\ number\ of\ correctly\ recognized\ PPI}{the\ number\ of\ recognized\ PPI} \quad (1)$$

$$Recall = \frac{the\ number\ correctly\ recognized\ PPI}{the\ number\ of\ Actual\ PPI} \quad (2)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

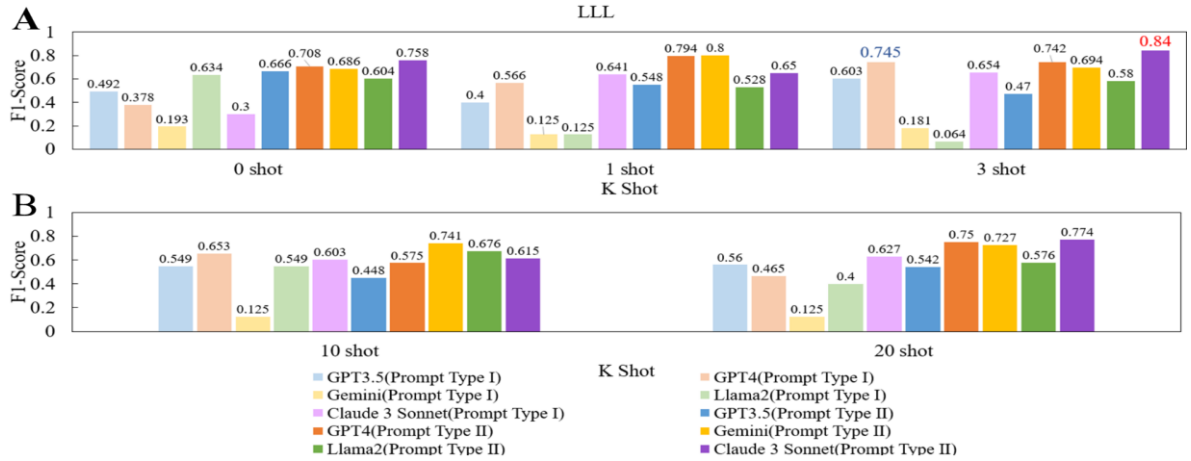

圖 3. 5 種 LLM 在 LLL 語料庫的測試結果，每個 X 軸刻度對應到 10 組數據，左邊 5 組為 Prompt Type I 實驗，右邊 5 組為 Prompt Type II 實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

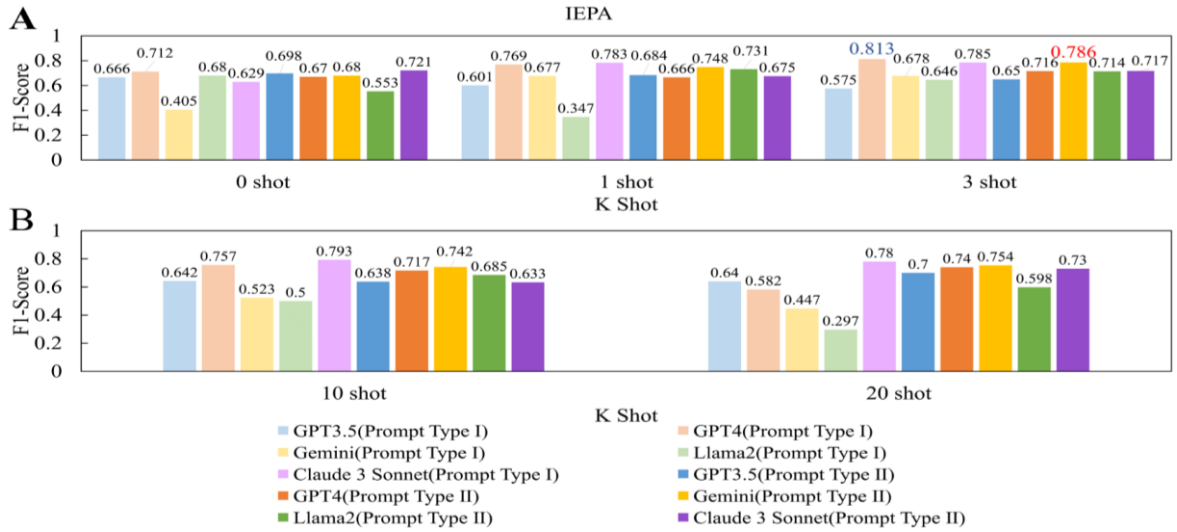

圖 4. 5 種 LLM 在 IEPA 語料庫的測試結果，每個 X 軸刻度對應到 10 組數據，左邊 5 組為 Prompt Type Ⅰ 實驗，右邊 5 組為 Prompt Type Ⅱ 實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

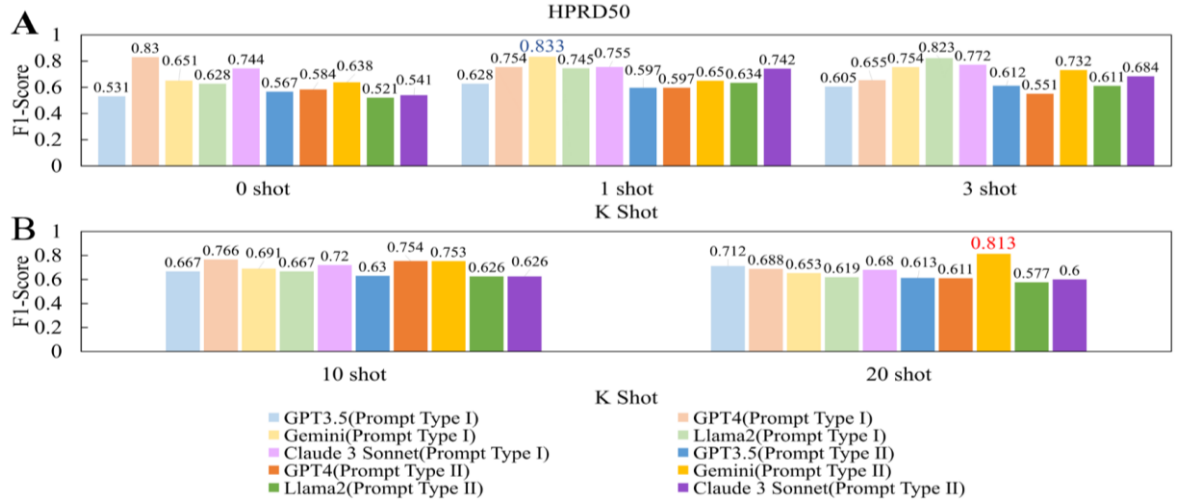

圖 5. 使用 5 種 LLM 在 HPRD50 語料庫的測試結果，每個 X 軸刻度對應到 10 組數據，左邊 5 組為 Prompt Type Ⅰ實驗，右邊 5 組為 Prompt Type Ⅱ實驗，F1-Score 在兩個實驗中的最高分分別以藍色、紅色文字標示，A 為 K=0、1、3 的測試結果，B 為 K=10、20 的測試結果

4 Result and Discussion

4.1 Sentence Reduction 對 LLM 識別 PPI 的優化程度

圖 3 到 5 為比較表 1 中 Prompt Type I（無 Sentence Reduction）和 Prompt Type II（應用 One-Stage Prompting 和 Sentence Reduction）分別在三個 PPI 語料庫上的實驗結果。LLL 語料庫的實驗結果如圖 3 所示，Prompt Type II 的 F1-Score 普遍高於 Prompt Type I，最高達 0.84，而 Prompt Type I 的結果大多低於 0.75，代表 Sentence Reduction 能顯著提升模型在 LLL 語料上的效能，提升幅度介於 0.01 至 0.5 之間。

圖 4 展示了 IEPA 語料庫的實驗結果。根據我們的統計，IEPA 是三個語料庫中語句平均長度最長的語料庫（LLL、IEPA 及 HPRD50 三個語料庫的測試集句子平均 token 數目分別為 34.31、35.4 與 25.73），令人意外的是，Prompt Type I 的 F1-Score 反而高於 Prompt Type II，最高甚至超過 0.8，顯示現今 LLM 在理解專業知識領域語句的卓越能力，然而 Prompt Type II 的結果在不同的 LLM 表現較為穩定，以 Llama2 為例，Prompt Type I 在 K=1 時 F1-Score 為 0.68，但在 K=3 時驟降至 0.064，波動極大。相比之下，Prompt Type II 的 Llama2 表現穩定維持在 0.55 以上，最高分與最低分差距僅 0.2，代表 Sentence Reduction 有

Dataset	LLL				IEPA				HPRD50			
Model	K*	P	R	F	K*	P	R	F	K*	P	R	F
GPT3.5	K=0	0.531	0.566	0.548	K=10	0.636	0.750	0.688	K=3	0.640	0.690	0.666
GPT4	K=10	0.727	0.533	0.640	K=1	0.633	0.803	0.708	K=10	0.720	0.692	0.705
Gemini	K=0	0.570	0.860	0.693	K=20	0.880	0.540	0.667	K=1	0.631	0.923	0.750
Llama2	K=10	0.750	0.300	0.429	K=0	0.500	0.642	0.562	K=3	0.471	0.961	0.632
C3S*	K=20	**0.880**	**0.733**	**0.800**	K=0	**0.910**	**0.750**	**0.823**	K=3	**0.790**	**0.730**	**0.760**

表 2.表格為使用 Prompt Type Ⅲ在 5 個 LLM 及 3 個語料庫上的實驗最好的結果，P、R、F 分別代表 Precision、Recall、F1-Score，表格中在 3 個語料庫上最高分以粗體標示。C3S*代表 Claude 3 Sonnet，K*代表最高分的實驗中 PPI Extraction 的 Few-Shot 設置

助於維持 LLM 在不同 K shot 之間的穩定度。圖 5 為 HPRD50 上的實驗結果，所有 LLM 在此語料庫上的表現均保持在 0.5 以上，且不似前兩個語料庫出現極大的分數落差。為了深入理解這一現象，我們對 HPRD50 語料庫中的句子結構進行分析，並發現其中的句子含有大量的括號，而透過 LLM 進行句子簡化時，往往會將括號刪除，改變了語法結構和語意內容，導致無法正確識別句中的 PPI。即使如此，在 Gemini 使用 Prompt Type II 仍有達到 F1-score 最高分 0.813，顯示 Sentence Reduction 結合 Few-Shot Learning 的效力。

綜合來說，結合 Sentence Reduction 技術的 Prompt 可讓 LLM 在 PPI 識別任務中表現出較高的穩定性。在大多情況下，分析簡化後的語能提升 LLM 識別 PPI 的準確性，此結果也顯示通過本研究的 Prompt Engineering 方式引導 LLM 執行 Sentence Reduction 是一種可行且有效的策略。

4.2 One-Stage Prompting 與 Two-Stage Prompting

表 2 呈現使用 Prompt Type III（應用 Two-Stage Prompting 和 Sentence Reduction）在 5 種 LLM 及 3 個語料庫上的最優結果。其中 Claude 3 Sonnet 在所有語料庫中均取得最高 F1-Scores：LLL 為 0.8，IEPA 為 0.823，HPRD50 為 0.76。對比圖 3 至 5 中 Prompt Type B 的結果可以發現 One-Stage Prompting 在 LLL 和 HPRD50 語料庫上的表現優於 Two-Stage Prompting。具體而言，One-Stage Prompting 在 LLL、IEPA 和 HPRD50 上分別達到最高 0.84、0.786 和 0.813 的 F1-Scores。綜合分析 Prompt Type II 和 C 的最高 F1-Scores 實驗數據表明 One-Stage Prompting 在 PPI 識別任務中整體上優於 Two-Stage Prompting。深入分析這一現象，我們發現兩者在 Sentence Reduction 步驟所產生的簡化句相近，然而，在 PPI Extraction 階段，Two-Stage Prompting 的錯誤率顯著高於 One-Stage Prompting。以 HPRD50 語料庫的實驗為例，在使用同樣 Claude 3 Sonnet 的情況下，針對句子

In contrast to PROTEIN1 PROTEIN2 did not interact with the AF2 domain of MOLECULE.

兩種 Prompting 策略均輸出簡化句

PROTEIN1 PROTEIN2 did not interact with MOLECULE.

但在 One-Stage Prompting 實驗中模型正確地判斷出 PROTEIN1 和 PROTEIN2 之間沒有相互作用，而在 Two-Stage Prompting 實驗中由於句子中出現了 "PROTEIN1"、"PROTEIN2" 及"interact with" 等關鍵詞，模型卻錯誤地認為 PROTEIN1 及 PROTEIN2 之間存在 PPI。這一結果表明 Prompting 策略的選擇對 LLM 在執行任務時的表現有些微影響，特別是在處理複雜度相對低的任務時，採用連續的、一體化的指令方法（如 One-Stage Prompting）似乎更能夠使 LLM 保持持續的注意力，從而提高預測的準確性。

4.3 Few-Shot 範例數量對效能的影響

由所有實驗結果來看，使用 Few-Shot 的方法普遍優於 Zero-Shot 的方法。以使用 Prompt Type II 的方法而言，對比所有 LLM 在 Zero-Shot 的 F1-score 最高分以及所有 LLM 採用 Few-Shot 的 F1-score 最高分，可以發現後者在 LLL、IEPA、與 HPRD50 語料庫的測試上分別提供 50% 、60%、及 70%的進步（圖 3、圖 4、圖 5），因此 Few-Shot Learning 可優化 LLM

Dataset	LLL			IEPA			HPRD50		
Method	P	R	F	P	R	F	P	R	F
LLM	**0.743**	**0.966**	**0.840**	**0.754**	**0.821**	**0.786**	**0.774**	**0.857**	**0.813**
SPBA	0.780	0.790	0.780	0.750	0.790	0.770	0.750	0.760	0.750
GK	0.725	0.872	0.768	0.696	0.827	0.751	0.643	0.658	0.634
LPTK	0.789	0.721	0.753	0.748	0.664	0.702	0.727	0.622	0.671

表 3. P、R、F 分別代表 Precision、Recall、F1-Score，LLM 採用 Prompt Type Ⅱ方法的 LLM 最高 F1-score 作為代表，與其它三種非 LLM 的方法之效能比較。在各語料庫表現最高分的方法以粗體標記

識別 PPI 的準確度。但 Few-Shot 範例的數量增加不一定能優化 LLM 的表現，甚至有時會導致 LLM 的性能下降，如圖 4 中 Claude 3 Sonnet 的表現，在 K=0 時，F1-Score 來到了 0.758，但是在 K=1 時，F1-Score 卻下降到 0.65，而在 Shot 數目增加到 K=3 時，F1-Score 達到最高分數為 0.84。Few-shot Learning 並未隨著 K 值增加有所提升，我們推測可能提供的 K Shot 範例與 Query 中的句子並無類似的語法結構，導致 LLM 無法從中得到相關資訊，作正確的判斷邏輯。過去在 Shi 等人的研究中（2023），提到給予 LLM 與任務不相關的範例會分散 LLM 的注意力並大幅降低效能，與我們觀察的現象類似。而我們的實驗結果同時證實 prompt 內容對 LLM 造成輸出偏差的影響。在 K=1 的情境下，我們觀察到 LLM 輸出 False Positive 的數量偏高，而當 K=3，False Positive 才明顯降低，到 K=10、20，LLM 產生的 False Positive 與 False Negative 便趨於穩定。

4.4 與其它非 LLM 方法的效能比較

表 3 為我們使用 Prompt Type II 的方法與語法分析及統計準則方法進行比較，SPBA 為 N.-W. Chang （2020） 為建立準則樣本辨識文獻中 PPI 關係的方法；GK 為 Graph Kernel（Airola et al. 2008）根據句子的語法結構並建立及分析形式化的圖譜識別 PPI；LPTK（Warikoo, Chang, and Hsu 2018）為利用語義分析樹特徵提取 PPI。由表 3 可以看出我們的方法在 3 個語料庫上皆勝過其他方法，凸顯使用 Prompt Engineering 透過 LLM 識別 PPI 的方法在生醫文獻探勘上的巨大潛力。

5 Conclusion

本研究透過 Prompt Engineering 開發了適用於 PPI 的 Prompt，透過結合 Sentence Reduction 與 Few-Shot Learning 於開發的 Prompt 中來提升 PPI 的辨識效能。實驗證實我們所提出的 Sentence Reduction 方法在 LLL、IEPA、HPRD50 上最高的 F1-Score 分別可達 0.84、0.813 與 0.813，相較於沒有使用 Sentence Reduction 且 PPI Extraction 使用 Zero-Shot 的方法可分別提升 0.36-0.54、0.04-0.38 與 0.08-0.126 的 F1-Score，顯示本研究所提出的方法的有效性。另外，我們發現 Few-Shot 的準確率普遍較 Zero-Shot 更高，但當提供的範例量到達 10 甚至是更高的 20 筆並未有顯著的效能提升。此外，我們也比較 One-Stage 和 Two-Stage Prompting 兩種策略，結果顯示 One-Stage Prompting 策略通過整合任務流程，使 LLM 能更全面地把握任務本質，提高了注意力和準確性，準確率優於其它非 LLM 的傳統方法如語法分析及統計準則方法。綜合而論，我們的研究證實透過 Prompt Enginerring 配合公開可用之 LLM 進行 PPI 識別任務的可行性和巨大潛力，未來我們將探討 LLM 本身對於 PPI 任務的先天知識，並在生物醫學領域上使用 fine-tuning 方式優化 LLM，以及應用至其它生物醫學關係的抽取任務上，如：基因-疾病關聯、化學物質-蛋白質相互作用。

References

Airola, Antti, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. “All-Paths Graph Kernel for Protein-Protein Interaction Extraction with Evaluation of Cross-Corpus Learning.” *BMC Bioinformatics* 9 （11）: S2. https://doi.org/10.1186/1471-2105-9-S11-S2.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. “Language Models Are Few-Shot Learners.” arXiv. http://arxiv.org/abs/2005.14165.

Budzianowski, Paweł, and Ivan Vulić. 2019. “Hello, It’s GPT-2 -- How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems.” arXiv. http://arxiv.org/abs/1907.05774.

Caruccio, Loredana, Stefano Cirillo, Giuseppe Polese, Giandomenico Solimando, Shanmugam Sundaramurthy, and Genoveffa Tortora. 2024. “Claude 2.0 Large Language Model: Tackling a Real-World Classification Problem with a New Iterative Prompt Engineering Approach.” *Intelligent Systems with Applications* 21（March）:200336. https://doi.org/10.1016/j.iswa.2024.200336.

Chang, Nai-Wen. 2020. “基於統計準則式方法偵測生醫文獻中的生物關聯.” 臺灣大學生醫電子與資訊學研究所學位論文 2020（January）:1–87. https://doi.org/10.6342/NTU202001123.

Chang, Yung-Chun, Chun-Han Chu, Yu-Chen Su, Chien Chin Chen, and Wen-Lian Hsu. 2016. “PIPE: A Protein–Protein Interaction Passage Extraction Module for BioCreative Challenge.” *Database: The Journal of Biological Databases and Curation* 2016（August）:baw101. https://doi.org/10.1093/database/baw101.

Chen, Banghao, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. “Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review.” arXiv. https://doi.org/10.48550/arXiv.2310.14735.

Deng, Mingkai, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P. Xing, and Zhiting Hu. 2022. “RLPrompt: Optimizing Discrete Text Prompts with Reinforcement Learning.” arXiv. http://arxiv.org/abs/2205.12548.

Ding, J., D. Berleant, D. Nettleton, and E. Wurtele. 2002. “Mining MEDLINE: Abstracts, Sentences, or Phrases?” *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 326–37. https://doi.org/10.1142/9789812799623_0031.

Du, Wenyu, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and Jie Fu. 2024. “Stacking Your Transformers: A Closer Look at Model Growth for Efficient LLM Pre-Training.” arXiv. https://doi.org/10.48550/arXiv.2405.15319.

Feng, Yutao, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. “Sentence Simplification via Large Language Models.” arXiv. http://arxiv.org/abs/2302.11957.

Floridi, Luciano, and Massimo Chiriatti. 2020. “GPT-3: Its Nature, Scope, Limits, and Consequences.” *Minds and Machines* 30（4）: 681–94. https://doi.org/10.1007/s11023-020-09548-1.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, et al. 2024. “Gemini: A Family of Highly Capable Multimodal Models.” arXiv. http://arxiv.org/abs/2312.11805.

Jing, Hongyan, and Kathleen R. McKeown. 2000. “Cut and Paste Based Text Summarization.” In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 178–85. NAACL 2000. USA: Association for Computational Linguistics.

Jonnalagadda, Siddhartha, and Graciela Gonzalez. n.d. “Sentence Simplification Aids Protein-Protein Interaction Extraction.”

Kalyan, Katikapalli Subramanyam. 2024. “A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4.” *Natural Language Processing Journal* 6（March）:100048. https://doi.org/10.1016/j.nlp.2023.100048.

Kim, Sun, Soo-Yong Shin, In-Hee Lee, Soo-Jin Kim, Ram Sriram, and Byoung-Tak Zhang. 2008. “PIE: An Online Prediction System for Protein-Protein Interactions from Text.” *Nucleic Acids Research* 36（Web Server issue）: W411-415. https://doi.org/10.1093/nar/gkn281.

Li, Yijing, Yanping Chen, Yongbin Qin, Ying Hu, Ruizhang Huang, and Qinghua Zheng. 2021. “Protein-Protein Interaction Relation Extraction Based on Multigranularity Semantic Fusion.” *Journal of Biomedical Informatics* 123（November）:103931. https://doi.org/10.1016/j.jbi.2021.103931.

Liu, Gang, and Jiabao Guo. 2019. “Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification.” *Neurocomputing* 337（April）:325–38. https://doi.org/10.1016/j.neucom.2019.01.078.

Minaee, Shervin, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. “Large Language Models: A Survey.” arXiv. https://doi.org/10.48550/arXiv.2402.06196.

Nédellec, C, and Claire Nedellec. n.d. “Learning Language in Logic - Genic Interaction Extraction Challenge.”

Nguyen, Minh Le, Susumu Horiguchi, Akira Shimazu, and Bao Tu Ho. 2004. “Example-Based Sentence Reduction Using the Hidden Markov Model.” *ACM Transactions on Asian Language Information Processing* 3（2）: 146–58. https://doi.org/10.1145/1034780.1034785.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. “GPT-4 Technical Report.” arXiv. http://arxiv.org/abs/2303.08774.

Park, Gilchan, Byung-Jun Yoon, Xihaier Luo, Vanessa López-Marrero, Shinjae Yoo, and Shantenu Jha. 2023. “Comparative Performance Evaluation of Large Language Models for Extracting Molecular Interactions and Pathway Knowledge.” arXiv. http://arxiv.org/abs/2307.08813.

Peri, Suraj, J. Daniel Navarro, Troels Z. Kristiansen, Ramars Amanchy, Vineeth Surendranath, Babylakshmi Muthusamy, T. K. B. Gandhi, et al. 2004. “Human Protein Reference Database as a Discovery Resource for Proteomics.” *Nucleic Acids Research* 32 （Database issue）: D497-501. https://doi.org/10.1093/nar/gkh070.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. n.d. “Language Models Are Unsupervised Multitask Learners.”

Rehana, Hasin, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. n.d. “Evaluation of GPT and BERT-Based Models on Identifying Protein-Protein Interactions in Biomedical Text.”

Schulhoff, Sander, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, et al. 2024. “The Prompt Report: A Systematic Survey of Prompting Techniques.” arXiv. http://arxiv.org/abs/2406.06608.

Shi, Freda, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. “Large Language Models Can Be Easily Distracted by Irrelevant Context.” arXiv. http://arxiv.org/abs/2302.00093.

0.1109/JAS.2023.123618.

Sivarajkumar, Sonish, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2023. “An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing.” arXiv. https://doi.org/10.48550/arXiv.2309.08008.

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, et al. 2023. “Llama 2: Open Foundation and Fine-Tuned Chat Models.” arXiv. http://arxiv.org/abs/2307.09288.

Warikoo, Neha, Yung-Chun Chang, and Wen-Lian Hsu. 2018. “LPTK: A Linguistic Pattern-Aware Dependency Tree Kernel Approach for the BioCreative VI CHEMPROT Task.” *Database: The Journal of Biological Databases and Curation* 2018 （January）:bay108. https://doi.org/10.1093/database/bay108.

Warikoo, Neha, Yung-Chun Chang, and Shang-Pin Ma. 2022. “Gradient Boosting over Linguistic-Pattern-Structured Trees for Learning Protein–Protein Interaction in the Biomedical Literature.” *Applied Sciences* 12 （20）: 10199. https://doi.org/10.3390/app122010199.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. “HuggingFace’s Transformers: State-of-the-Art Natural Language Processing.” arXiv. http://arxiv.org/abs/1910.03771.

Wu, Tianyu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. “A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development.” *IEEE/CAA Journal of Automatica Sinica* 10 （5）: 1122–36. https://doi.org/1