以大語言模型進行兒童敘事句法能力檢測與分析
MINAS: Mandarin Intelligent Narrative Assessment of Syntax for Children

王睿孺 Ruei-Ru Wang[1,*] , 李亞欣 Ya-Sin Li[1,*] , 尹懌碩 Yi-Shuo Yin[1] ,
陳韜宇 Tao-Yu Chen[1] , 張顯達 Hint-Tat Cheung[2] , 陳鯨太 Ching-Tai Chen[3,†]

ctchen@utaipei.edu.tw
*These authors contributed equally to this work.
†corresponding author
1 Department of Bioinformatics and Medical Engineering, Asia University
2 Department of Audiology and Speech-Language Pathology, Asia University
3 Department of Computer Science, University of Taipei

摘要

兒童敘事能力是語言發展的重要指標，常用於臨床診斷與語言研究。然而，缺乏大規模、標準化、精準註記的中文兒童語料，使得語法分析既耗時又容易受主觀影響，現有自動化工具難以滿足臨床和研究需求。本研究提出 MINAS (Mandarin Intelligent Narrative Assessment of Syntax for Children)，結合 MAIN 故事情境與 MAPS-R 語法架構，建立涵蓋四個類別、20 個指標的中文敘事語料資料集。我們以 Prompt Engineering 評估商用模型（ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash、DeepSeek），並以 LoRA 微調開源模型（Chinese RoBERTa、OpenHermes-2.5）。實驗結果顯示，Few-shot Prompt 能提升多數指標的辨識準確度；LoRA 微調則在名詞與動詞短語上表現更佳，但在複雜句型仍具挑戰。本研究驗證了 LLM 應用於「中文兒童敘事語料語法分類」的可行性，展現其在臨床與語言研究的潛力。

Abstract

Children's narrative ability is an important indicator of language development and is commonly used in clinical diagnosis and linguistic research. However, the lack of large-scale, standardized, and accurately annotated Chinese child language corpora makes grammatical analysis both time-consuming and prone to subjectivity, while existing automated tools fall short of clinical and research needs. This study introduces MINAS (Mandarin Intelligent Narrative Assessment of Syntax for Children), which integrates the MAIN story framework with the MAPS-R syntactic framework to construct a Chinese narrative corpus encompassing four categories and 20 indicators. We evaluated commercial models (ChatGPT-4, Claude Sonnet 4, Gemini 2.5 Flash, DeepSeek) through prompt engineering, and fine-tuned open-source models (Chinese RoBERTa, OpenHermes-2.5) with LoRA. Experimental results show that few-shot prompting achieves high accuracy across most indicators, while fine-tuning with LoRA achieves better performance in noun and verb phrase identification but is not as good for complex sentence structures. This study validates the feasibility of applying large language models to syntactic classification of Chinese child narrative corpora, highlighting their potential in clinical applications and linguistic research.

關鍵字：兒童語言評估、語法分類、大型語言模型、少樣本學習

Keywords: Child Language Assessment; Syntactic Classification; Large Language Models; Few-shot Learning

1 Introduction

兒童語言能力的發展是語言學與語言病理學的重要研究議題。特別是在語言學習初期，敘事能力(narrative ability)被視為整合語音、語法、語意與篇章組織的綜合指標，能夠有效評估兒童的語言表達與理解發展(Berman et al., 1994)。近年研究指出，兒童的敘事結構能力與其語言障礙、語用能力與認知表現息息相關，因此也逐漸成為臨床診斷的重要依據。

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 184–192

為此，MAIN（Multilingual Assessment Instrument for Narrative）與 MAPS-R（Multidimensional Assessment of Preschool Syntax – Revised）等工具陸續被提出。MAIN 透過故事圖片刺激，提供標準化的敘事引導情境 (Gagarina et al., 2012)，已應用於全球 65、90 餘語言；MAPS-R 則是著重於針對兒童語法能力進行詳細的系統評估 (Cheung et al., 2024)，從名詞短語、動詞短語、介詞短語到句型結構等 20 項語法分類，為中文語言臨床工作者提供了評估的參考架構。然而，即使評估架構逐漸成熟，但人工標註成本高、資料量有限、中文語料的採樣方法、內容與標註格式差異極大，都導致自動化系統難以精準建立與驗證。

近年來，大型語言模型（Large Language Models, LLMs）如 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeekSeek 迅速發展，其語言理解與分類能力已在多項自然語言任務中達到極佳表現。利用提示工程（Prompt Engineering）、少樣本學習（Few-shot Learning）、模型微調（Fine-tune model）等技術，研究者能在不需要大量標註資料的前提下，引導模型完成複雜語言任務，包含語法結構分析、篇章分類、語意判斷等。然而，目前還未有針對「中文兒童敘事語料之語法結構分類」的 LLM 應用，亦缺乏可驗證的語料資源與方法驗證。

因此，本研究提出一套以 MAIN 故事為基礎、融合 MAPS-R 語法架構設計之中文敘事語料分類資料集，並結合提示工程與 LLM 模型進行語法指標辨識任務，針對名詞短語、動詞短語、介詞短語與句型類別等語法指標進行分析。藉由比較不同模型（如 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeekSeek）與提示設計，探討其於兒童語言能力自動評估任務的應用潛力，本研究為探索中文語言臨床實務與語言學研究中的應用潛力。

2 Related Work

2.1 MAIN 工具與中文語法標註與語料資源建構

在兒童語言研究與臨床實務研究中，敘事能力的量測通常透過多種標準化工具來進行其中包含 ENNI（Edmonton Narrative Norms Instrument）、Renfrew Bus Story（RBS）、MAIN。

ENNI(Schneider et al., n.d.)針對 4–9 歲兒童語言能力的敘事評估工具以收集語言資料並建立本地語言表現的標準樣本數據。Renfrew Bus Story 是一個透過聽故事後復述的方式，來評估兒童口語敘事能力與語言發展的標準化測驗工具。而 MAIN（Multilingual Assessment Instrument for Narratives）是近年在語言學與語言治療領域中出現的敘事能力評估工具。該評估工具利用模範故事、故事複述及自主講述為基礎，能有效的評估兒童在語意、句法或敘事結構上的表現，在多國學者合作開發下，目前現已有 90 多種語言版本，在全球 65 多個國家使用。

在華語兒童語言研究中，中文資源相較有限。國際間規模最大的兒童語料庫 CHILDES 收藏了台灣的兒童語料，但語料來自不同的研究，研究者參與誘發發言的程度不一，導致內容複雜度差異頗大，而且系統工具 CLAN 是以 1984 年該系統創立時的處理方法，以句子為單位，逐句分層(tier)標示詞類與語法關係，並且一律強制使用簡體字，形成許多使用上的困難(MacWhinney & Snow, 1985); Sinica Treebank (Huang et al., 2000)雖提供了語法標註，但以成人語料為主。因此本研究以 MAIN 為基礎，建立華語兒童敘事語句的誘發語法分類資料集，使用臨床資料做實際的驗證及測試，補足現有資源不足之處，結合生成式 AI 進行自動評估功能，快速掌握兒童敘事語法能力。

2.2 商用大型語言模型的應用

大型語言模型在自然語言處理任務上已展現高度潛力，尤其在語句分類、語法結構判斷、語意辨識等方面，已被廣泛應用於語料分析。本次研究的模型包含 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeepSeek-V3。ChatGPT-4 (OpenAI et al., 2024)其架構基於多層 Transformer 編碼器，能夠處理更長的文本輸入，並且更準確地掌握句子的語意與上下文；Claude Sonnet 4 該模型具備良好的語言理解、視覺分析、電腦操作與工具使用能力，特別擅長進行複雜的程式設計與推理任務；Gemini-2.5 Flash (Comanici et al., 2025)此模型可在需要時啟動內部「思考」機制以提升理解力和計劃能力，出色處理分類、翻譯、

程式碼執行等任務；DeepSeek-V3 (DeepSeek-AI et al., 2025)該模型結合 Multi-Head Latent Attention（MLA） 與 Multi-Token Prediction（MTP）技術，在數學、程式碼與知識推理任務中展現最先進表現。

2.3 開源大型語言模型的發展

LLaMA 系列(Touvron et al., 2023)以開放權重與高效能架構為特色，是具代表性的開源 LLM。LLaMA-1 提供 7B、13B、33B 與 65B 四個版本，採用 80 Transformer layers 與 64 attention heads 的組態；其後的 LLaMA-2 延續了 LLaMA-1 的架構，在語料選取與訓練策略上進行優化，釋出 7B、13B 與 70B 參數的版本，其在多項基準任務上表現卓越，並成為眾多研究與應用的基礎；Mistral 系列中的 Mistral-7B 包含 32 個 Transformer layers、32 個 attention heads， 並使用 Grouped Query Attention (GQA) 及 Sliding Window Attention (SWA)機制，能有效處理長序列輸入；OpenHermes 系列是在 Mistral-7B 基座模型上進行微調的開源模型，使用了大量程式碼相關指令資料、由 GPT-4 生成的訓練樣本，以及其他 AI 領域公開語料 (Teknium/OpenHermes-2.5-Mistral-7B· Hugging Face, 2024)。此外，社群亦釋出了基於 LoRA/QLoRA（Low-Rank Adaptation）的參數效率微調版本，使其能在資源有限的環境中應用。

2.4 Prompt Engineering 及 Fine-tune

大型語言模型（Large Language Models, LLMs）的快速發展，提示工程（prompt engineering）逐漸成為了提升模型效能的重要方式。few-shot prompting (Brown et al., 2020)顯示在無需額外訓練的情況下，僅透過設計少量範例提示即可顯著改善模型表現；Chain-of-Thought (CoT)方法 (Wei et al., 2023)則透過引導模型生成中間推理步驟，提升數學與邏輯任務的正確率。本研究也使用 finetuning 調整模型的表現。相較於 prompt engineering 的低成本與靈活性，finetuning 能針對特定任務進行更穩定與精準的調整，但其缺點是需要額外的大量標註資料與計算資源(Ziegler et al., 2020)。

3 Method

本研究採用兩種架構針對中文兒童敘事語料進行語法結構分類。詳細研究流程如圖 1 所示。

i. Prompt Engineering：針對商用 LLM 使用提示工程。
ii. LLM fine-tuning：使用 LoRA(Hu et al., 2021)對開源模型進行參數微調。

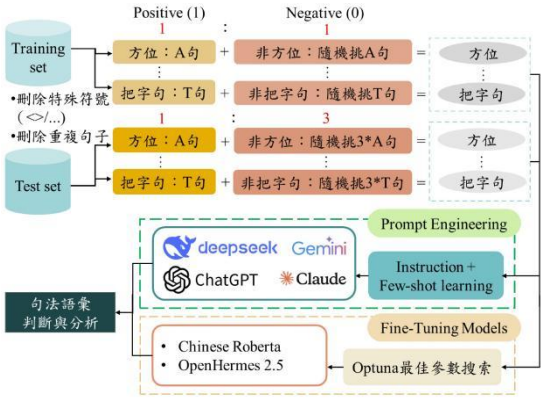

圖 1. 實驗流程圖

整體研究流程包含資料前處理、Prompt 設計、超參數最佳化、模型訓練與評估五個主要階段。首先建立基於 MAPS-R 架構的中文兒童敘事語法分類資料集，接著設計 Prompt 模板用於引導 LLM 進行語法判斷，另外也使用 Optuna (Akiba et al., 2019)進行超參數自動化搜尋，最後透過多個商用模型及開源模型的比較驗證方法的有效性。

3.1 Dataset

本研究使用的語料取自 MAPS-R 資料集 (Cheung et al., 2024)，該資料集是由三位接受 MAPS-R 編碼標準訓練的語言發展或對外進行華語教學領域的研究助理，負責中文兒童看圖敘事語料的分類。隨機抽取 20%的樣本進行 Inter-rater reliability (IRR)，計算出的 kappa 值 0.91 至 0.95。語料來源為使用 MAIN (Gagarina et al., 2012)的故事圖片引導所收集的兒童自然敘事語句。經過斷句處理與語法結構標註，能夠真實反映兒童語言發展的特徵與差異。

依據 MAPS-R 架構，兒童語言發展評估涵蓋名詞短語、動詞短語、介詞短語、句子類別等四大分類共 20 種語法指標。比如說，「他用腳踢門」符合介詞短語中的「動前介詞」指標；「拿出一本書來」符合動詞短語中的

「趨向補語」指標。

在資料前處理階段，我們移除語料中與語意無關之特殊符號、去除重複句子，確保訓練集與測試集間的資料獨立性。針對每項語法指標採用二元分類架構建立正負樣本：符合特定語法指標的句子標記為正樣本（Positive），不符合該語法指標的句子則標記為負樣本（Negative），負樣本從其他語法指標的正樣本中隨機抽取，確保樣本間的語言複雜度相當。

在樣本配置方面，訓練集採用 1:1 的正負樣本比例，測試集則採用 1:3 的正負樣本比例。各語法指標的樣本數量介於 10 至 1000 餘筆不等，反映不同語法特徵在兒童語料中的出現頻率差異。完整的資料集分布情況詳見表 1。

指標	名稱	Training set		Test set	
名詞短語		Pos	Neg	Pos	Neg
NP1	量詞-個	805	805	58	174
NP2	量詞-特定	195	195	58	174
NP3	X 的	352	352	40	120
NP4	X 的 Y	365	365	40	120
NP5	方位	510	510	43	129
動詞短句					
VP1	體貌標記	395	395	92	276
VP2	結果補語	462	462	69	207
VP3	趨向補語	252	252	47	141
VP4	情態補語	100	100	40	120
VP5	可能補語	100	100	40	120
VP6	數量補語	122	122	41	123
介詞短語					
PP1	動前介詞	100	100	48	144
PP2	動後介詞(補語)	102	102	42	126
句子					
S1	把字句	127	127	42	126
S2	被字句	100	100	40	120
S3	存現句	100	100	82	246
S4	複謂(連動/兼語)	255	255	83	249
S5	帶連詞複句	165	165	95	285
S6	緊縮複句	100	100	45	135
S7	感知/心理狀態動詞	100	100	93	279

表 1. 資料集大小

3.2 Prompt Engineering

如圖 2 所示，本研究採用經專家設計的 Prompt，引導大型語言模型進行準確的語法判斷。Prompt 模板包含兩個核心部分：

- Instruction：明確定義任務要求與輸出格式，並詳細解釋各項句法概念。為了提升模型的判斷準確性與一致性，指令中加入明確的判斷標準，規定模型須以二元分類形式（1/0）回應，並強制輸出判斷理由，以利追蹤模型推論過程與確保分析結果的可解釋性。
- Few-shot Learning： 除提供語法定義中

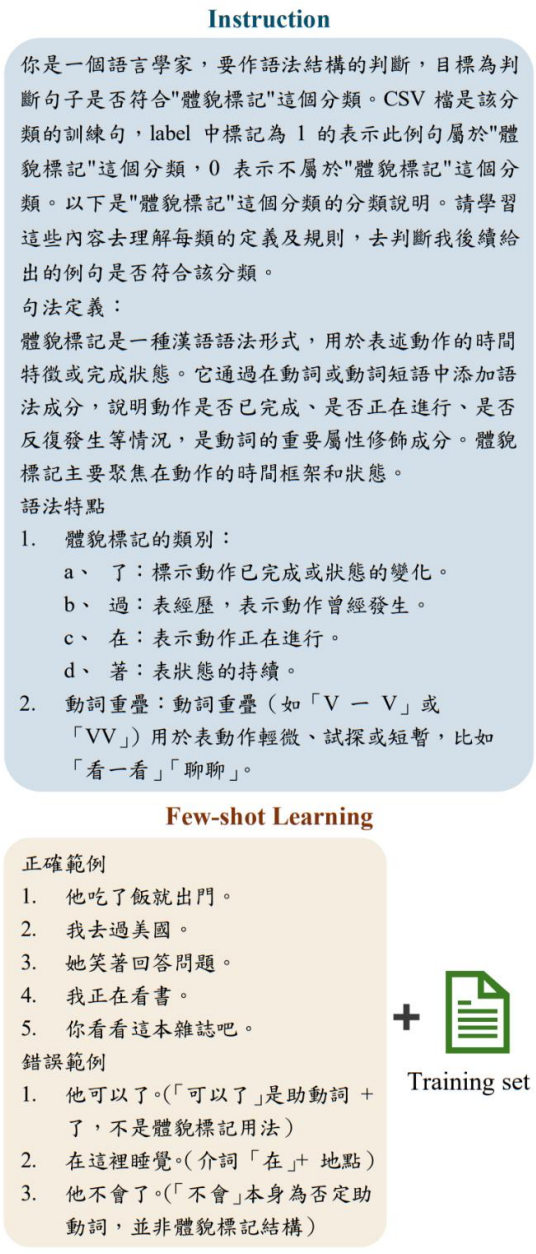

圖 2. Prrompt 模板

所附的正反範例句及錯誤分析外，進一步加入來自訓練集的句子作為示例，以增強模型對特定語法指標的判斷能力，並提升其語法概念的泛化能力。

將上述訓練集與 Prompt 輸入四種商用 LLM，分別為 ChatGPT-4、Claude Sonnet 4、Gemini 2.5 Flash 與 DeepSeek，以進行短期任務記憶與學習，隨即使用測試集進行效能評估，以驗證 LLM 在特定語法結構識別任務中的準確性。

3.3 Fine-Tuning Models

本研究使用 Chinese RoBERTa-wwm-ext (Chinese Roberta) 預訓練模型 (https://huggingface.co/hfl/chinese-roberta-wwm-ext) 和 OpenHermes-2.5-Mistral-7B-GPTQ (OpenHermes-2.5) 預訓練模型 (https://huggingface.co/TheBloke/OpenHermes-2.5-Mistral-7B-GPTQ) 進行分析評估。RoBERTa (Liu et al., 2019)是一種基於 Transformer 架構的深度學習模型，而 Chinese Roberta 進一步引入了 Whole Word Masking（WWM）(Cui et al., 2021)，在遮罩任務中針對整個詞語進行遮罩，而非單一字元，使其更具挑戰性，進而提升模型捕捉中文語義與語法關係的能力。Chinese Roberta 通過這些訓練技巧，在自然語言處理任務中顯著提升模型表現與穩健性。OpenHermes-2.5 則為大型語言模型 Mistral-7B 的 GPTQ 量化版本(Jiang et al., 2023)，也基於 Transformer 架構，具備高效微調能力，可在有限 GPU 記憶體下進行 LoRA 微調，使其能適應中文語法結構識別等任務的需求。

資料預處理：20 種指標的訓練集分別被劃分為訓練集（80%）和驗證集（20%），並確保正負類別比例分佈一致。所有語料均分別透過 Chinese Roberta 和 OpenHermes2.5 的各自 Tokenizer 轉換為模型所需的輸入格式。

LoRA：本研究採用了 LoRA 微調技術，在有限的計算資源下高效地對 Chinese Roberta 與 OpenHermes-2.5 進行 Fine-Tune。LoRA 的核心原理是在預訓練模型權重矩陣旁注入兩個 low-rank 可訓練矩陣。在訓練過程中，僅調整這兩個小矩陣的參數，而原始模型的預訓練權重保持不變。此方法顯著減少了可訓練參數的數量，大幅降低記憶體消耗與訓練時間，同時有效降低在小型資料集上發生 Over-fitting。

Optuna：在參數搜索方面，採用 Optuna 框架進行超參數自動化搜尋，以確保模型在不同資料集上均能達到最佳效能，其核心優勢在於能根據過去試驗的結果，決定下一組要嘗試的參數，從而更有效率地找到最佳參數組合。

設置 Optuna 搜尋以下幾個關鍵超參數：

- learning_rate：在 1e-6 到 1e-3 的對數尺度間搜尋。
- batch_size：從[4, 8, 16, 32]中選擇。
- lora_r：從[8, 12, 16, 20, 24]中選擇。
- lora_alpha：從[4, 8, 16, 32, 64]中選擇。
- lora_dropout：在 0.0 到 0.5 間搜尋，步長為 0.1。

每種資料集皆獨立進行 30 次試驗，以驗證集上的 validation loss 作為最佳化的目標。最終，我們將每種資料集所獲得的最佳參數組合記錄下來，並用於後續的最終模型訓練。

訓練過程使用 Hugging Face 的 Trainer 類別進行，並使用 Optuna 找到的最佳超參數。為了防止 Over-fitting，我們採用 Early Stopping 機制，當 validation loss 連續 15 個 epoch 沒有改善時，訓練會自動停止，並載入表現最佳的模型權重。最終，我們使用測試集評估模型在特定語法結構識別任務中的效能。

針對 Chinese RoBERTa 的整體 DNN 架構流程詳見圖 3。輸入的語料會先經由 Tokenizer 轉換為數位格式，接著進入 Chinese RoBERTa 並使用 LoRA 進行參數微調。最終連結至 Dense Layer 與 Softmax 以進行二元分類，得到 1 或 0 的結果，分別代表 Positive 或 Negative 標籤。

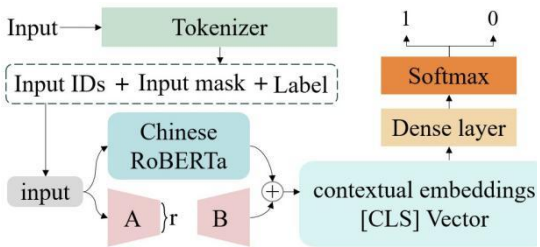

圖 3. 基於 RoBERTa-LoRA 的文本分類 DNN 架構

4 Result and Discussion

指標	Few-shot				Zero-shot			
	Gemini	Claude	ChatGPT	Deepseek	Gemini	Claude	ChatGPT	Deepseek
NP1	**0.983**	0.966	0.975	0.922	0.945	0.952	**0.974**	0.758
NP2	0.779	**0.855**	0.689	0.769	0.780	0.790	0.775	**0.820**
NP3	0.976	0.988	0.987	**1.000**	0.975	0.981	0.870	**1.000**
NP4	0.987	**1.000**	0.909	0.963	0.980	**0.985**	0.867	0.935
NP5	**0.966**	0.930	0.913	0.945	0.960	0.967	0.966	**0.977**
VP1	**0.948**	0.928	0.879	0.845	0.917	0.910	**0.938**	0.738
VP2	0.763	0.789	0.838	**0.872**	0.724	0.745	0.644	**0.769**
VP3	**0.842**	0.793	0.832	0.839	0.835	0.823	0.825	**0.839**
VP4	0.988	**1.000**	0.833	0.975	0.976	0.980	0.951	**1.000**
VP5	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
VP6	0.932	0.911	**0.975**	0.874	0.914	0.920	0.845	**0.953**
PP1	0.787	**0.922**	0.750	0.883	0.853	0.864	0.831	**0.949**
PP2	0.966	0.953	**0.977**	0.943	0.910	0.930	0.788	**0.989**
S1	**1.000**	0.976	0.822	0.988	**0.986**	0.985	0.988	0.977
S2	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
S3	0.747	**0.796**	0.793	0.755	0.742	0.750	**0.761**	0.720
S4	**0.964**	0.878	0.698	0.872	0.881	**0.907**	0.626	0.943
S5	**0.940**	0.931	0.581	0.935	0.569	0.580	0.540	**0.593**
S6	**0.977**	0.957	0.911	0.945	0.945	**0.956**	0.846	0.936
S7	0.883	0.690	**0.962**	0.690	0.617	**0.657**	0.583	0.633
Micro_F1	**0.902**	0.890	0.854	0.878	0.832	0.840	0.797	**0.844**
Macro_F1	**0.920**	0.912	0.865	0.901	0.875	**0.884**	0.831	0.876

表 2. Few-shot 各語法類別的 F1-Score

4.1 Prompt Engineering 效果分析

表 2 呈現 Few-shot 以及 Zero-shot prompting 在各語法類別的 F1-Score 實驗結果。整體而言，多數語法指標在 Few-shot 條件下均能達到穩定且高準確度的表現，F1-Score 在 0.95 以上；然而在結構和語義較複雜的類別（如「量詞-特定」、「結果補語」、「動前介詞」及「存現句」）上，模型仍存在辨識困難，F1-Score 落在 0.792 和 0.855 之間。

在 Zero-shot 實驗結果中，雖然部分語法指標，如「X 的」、「被字句」仍能維持較高分數，但在多數類別上，模型表現均低於 Few-shot 條件，尤其在句法結構與語義較為複雜的類別中差距更為明顯，例如「感知/心理狀態動詞」類別在 Few-shot 條件下的 F1-Score 較 Zero-shot 提升 30.5%，而帶連詞複句則提升 34.7%。

由此可見，單純依賴模型的內建知識和語法定義並不足以應對複雜的語法任務，而 Few-shot 的引入則能顯著提升模型在語法指標識別上的整體表現。Zero-shot 雖展現出模型固有語言知識的潛力，但在多數類別上仍表現不足，尤其是結構與語義多樣化的句型中。相比之下，Few-shot 透過少量範例有效縮小了模型的判斷偏差，顯示出在資源有限的情境下，Prompt 設計仍是提升大型語言模型語法處理能力的重要策略。

4.2 Fine-tuning 效果分析

表 3 列出兩個 LLM 在 20 個語法指標的最終的 F1-Score，整體而言，兩個模型在名詞短語與簡單動詞短語上均能達到高 F1-Score（0.95 – 1.00），且所需訓練迭代較少；相較之下，介詞短語及複雜句型（如「複謂」、「帶連詞複句」、「緊縮複句」、「感知/心理狀態動詞」）的 F1-Score 較低。

比較兩個模型可見，Chinese Roberta 在大部分語法類別上的 F1-Score 稍高於 OpenHermes 2.5。OpenHermes 2.5 部分動詞短語與句子類別的 F1-Score 與 Chinese Roberta 相近，但在語義和結構複雜的項目仍稍差。

綜合來看，Fine-tune 在大多數語法指標的識別上也能展現出不錯的效果，Macro_F1-Score 達到 0.854。對大部份名詞與動詞短語的辨別能力尤為顯著；然而對結構或語義複雜的句子，模型仍存在一定限制，未來可針對

這些類型強化資料基礎或探索更精細的 Fine-tuning 策略。

指標	Chinese Roberta	OpenHermes 2.5
NP1	**0.983**	0.953
NP2	**0.836**	0.798
NP3	**1.000**	0.982
NP4	**0.987**	0.964
NP5	**0.743**	0.697
VP1	**0.941**	0.876
VP2	**0.841**	0.783
VP3	**0.729**	0.681
VP4	**0.889**	0.824
VP5	**0.930**	0.851
VP6	**0.795**	0.746
PP1	**0.822**	0.783
PP2	**0.848**	0.802
S1	**0.977**	0.917
S2	**1.000**	0.924
S3	**0.766**	0.546
S4	**0.783**	0.698
S5	**0.891**	0.745
S6	**0.750**	0.678
S7	0.547	**0.603**
Micro_F1	**0.828**	0.774
Macro_F1	**0.854**	0.793

表 3. Fine-tune 各語法類別的 F1-Score

5 Error Analysis

為進一步理解模型在語法指標上的判斷偏差，本研究根據 Macro_F1-Score 表現最佳的 Gemini 在部分指標上分數相對較低的情況，選取量詞-特定、結果補語、動前介詞及存現句進行錯誤分析，檢視 False Positive（FP）與 False Negative（FN）案例，並比較模型判斷邏輯與應有標註的差異。

5.1 量詞-特定：

「帶特定量詞」的名詞短語用於表達對事物的特定量化，通過量詞與名詞搭配增加語義精確性，例如「一瓶水」、「一碗飯」。FP 案例如：「這份報告轉交給部門經理了」、「我跑了五分鐘就累了」、「這幾件都很好看，我想要亮的」。模型將「份」、「分鐘」等誤判為特定量詞修飾名詞，但實際上應視為一般個體量詞或度量單位，非真正特定量詞。

5.2 結果補語

結果補語表示動作完成後的結果，例如「我把衣服洗乾淨了」。FN 案例如：「牠跌倒」，模型識別「跌倒」是一個複合動詞，卻未識別出其結果補語為「倒」。FP 案例如：「因為聽不清楚，所以我又問了一次」，模型將「聽不清楚」判為結果補語，實際上應屬情態補語，結果補語的否定形式需在主要動詞前加「沒有」。

5.3 動前介詞

動前介詞出現在動詞前，對句子的謂語提供輔助資訊，用於修飾或限定動作的條件、時間、地點、對象等。FP 案例如：「我擠到人群裡面了」、「她塞了一個蘋果到背包裡」。模型將「到人群裡面」、「到背包裡」誤判為動前介詞，但語法上介詞出現在動詞後應屬動後介詞。

5.4 存現句

存現句用於陳述物品存在或事件發生，例如「桌子上有一本書」。FN 案例如：「警察局需要有槍」、「我這裡有三塊」，模型將強調必要性或擁有的句子誤判為非存現句。FP 案例如：「還有吐司」、「他還有一個翅膀」，模型識別為存現句，但實際「還有」表示在已有基礎上額外加上，並非純粹存在陳述。

5.5 Fine-tuning 的挑戰與限制

綜合本研究實驗結果，Fine-tuning 難以在相同資料條件下穩定超越 Prompt Engineering，可能的原因如下：

- 由於大型 LLM 的訓練語料龐大且涵蓋範圍廣泛，模型在預訓練過程中往往已具備一定的語言規則與推理能力。能在規則性較強的任務（如句法判斷）快速展現適應性。相較之下，Fine-tuning 資料量不夠充分時，模型可能無法有效收斂或容易過擬合，導致特定訓練句數較少的語法指標表現不佳。
- Prompt Engineering 效果穩定：透過在 Prompt 中提供適當的任務背景和範例，模型通常能夠準確地進行分類。然而，微調模型在測試語料較為罕見或與訓練集的句型差異較大的資料集上，表現可能不如預期。

- 模型成效與超參數配置的關聯：Fine-tuning 的成效依賴於學習率、batch size、epoch 與正則化等超參數設定，即使使用 LoRA 也可能因調校不足而難以達到參數收斂至最佳效果。

6 Conclusion

本研究基於 MAIN 故事情境 MAPS-R 的中文兒童敘事語料和語法架構，提出 MINAS 系統，並結合 Prompt Engineering 與 Fine-tuning 策略進行語法結構辨識。實驗結果顯示，Gemini 搭配 prompt engineering 可達 0.902 的 Micro_F1 與 0.920 的 Macro_F1，其中在「特定量詞」、「結果補語」、「動前介詞」等語法指標上的誤判較為明顯，但對多數語法指標都有較好的辨識準確度(F1 > 0.9)。LLM Fine-tuning 整體準確率稍差，Roberta 達到 0.828 的 Micro_F1 與 0.854 的 Macro_F1，其在名詞與動詞短語的分類上表現卓越，F1-Score 可達 0.95-1.00，證明其在處理特定語法任務時的有效性。然而，對於結構與語義複雜的句型，如「結果補語」和「存現句」，Fine-tuning 模型仍存在誤判。這反映了中文兒童語料在語法與語義表達上的多樣性，也提示 Fine-tuning 模型在測試集中較少見或與訓練集差異明顯的句型上可能表現受限。

此外，本研究仍存在若干限制：資料集規模較小且語法指標分佈不均，可能影響模型的泛化能力；然而，這樣的分佈特性亦真實反映了兒童自然語料中各類語法結構的實際出現頻率差異。傳統句法分析模型未被納入比較，主要原因在於本研究的多項語法指標同時涉及語義判斷（例如趨向補語、存現句等），而傳統句法模型主要聚焦於結構層面的分析，難以處理語義層面的判斷，因此未被納入主實驗比較。

本研究嘗試在中文兒童語料上探索大語言模型的語法與語義理解能力，驗證大型語言模型在中文兒童敘事語法分類任務上的可行性與應用潛力。實驗結果證實使用 LLM 進行中文兒童語法分類的可行性，並為語言臨床評估與語言學研究提供自動化分析數據。未來，我們將持續蒐集臨床語料、擴充語料集大小，並探索更精細的微調策略，以進一步提升大型語言模型在兒童敘事能力分析的可行性與判斷準確率。

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework* (No. arXiv:1907.10902). arXiv. https://doi.org/10.48550/arXiv.1907.10902

Berman, R. A., Slobin, D. I., Aksu-Koç, A. A., Bamberg, M., Dasinger, L., Marchman, V., Neeman, Y., Rodkin, P. C., Sebastián, E., & et al. (1994). *Relating events in narrative: A crosslinguistic developmental study* (pp. xiv, 748). Lawrence Erlbaum Associates, Inc.

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., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners* (No. arXiv:2005.14165). arXiv. https://doi.org/10.48550/arXiv.2005.14165

Cheung H., Ch L., & Cj C. (2024). Measuring productive syntactic abilities in Mandarin-speaking children in Taiwan. *Clinical linguistics & phonetics*, *38*(11). https://doi.org/10.1080/02699206.2024.2302549

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., … Bhumihar, N. K. (2025). *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities* (No. arXiv:2507.06261). arXiv. https://doi.org/10.48550/arXiv.2507.06261

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-Training with Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504–3514. https://doi.org/10.1109/TASLP.2021.3124365

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2025). *DeepSeek-V3 Technical Report* (No. arXiv:2412.19437). arXiv. https://doi.org/10.48550/arXiv.2412.19437

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). MAIN: Multilingual assessment instrument for narratives. *ZAS Papers in Linguistics*, *56*, 155–155. https://doi.org/10.21248/zaspil.56.2019.414

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models* (No. arXiv:2106.09685). arXiv. https://doi.org/10.48550/arXiv.2106.09685

Huang, C.-R., Chen, F.-Y., Chen, K.-J., Gao, Z., & Chen, K.-Y. (2000). Sinica Treebank: Design criteria, annotation guidelines, and on-line interface. *Proceedings of the Second Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 12*, 29–37. https://doi.org/10.3115/1117769.1117775

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). *Mistral 7B* (No. arXiv:2310.06825). arXiv. https://doi.org/10.48550/arXiv.2310.06825

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*(2), 271–295. https://doi.org/10.1017/S0305000900006449

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 Technical Report* (No. arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Plante, E. (n.d.). *The diagnostic and predictive validity of the Renfrew Bus Story*. Retrieved September 10, 2025, from https://www.academia.edu/13384934/The_diagnostic_and_predictive_validity_of_the_Renfrew_Bus_Story

Schneider, P., Hayward, D., & Dubé, R. V. (n.d.). *Évaluer grâce au « Edmonton Narrative Norms Instrument » une histoire contée à partir d'images Storytelling from pictures using the Edmonton Narrative Norms Instrument*.

Teknium/OpenHermes-2.5-Mistral-7B · Hugging Face. (2024, April 15). https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (No. arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models* (No. arXiv:2201.11903). arXiv. https://doi.org/10.48550/arXiv.2201.11903

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-Tuning Language Models from Human Preferences* (No. arXiv:1909.08593). arXiv. https://doi.org/10.48550/arXiv.1909.08593