

國立臺灣師範大學國際與社會科學學院華語文教學系

碩士論文

Department of Chinese as a Second Language

College of International Studies and Social Sciences

National Taiwan Normal University

Master Thesis

生成式 AI 在華語文分級測驗試題編寫之應用研究

—與 TOCFL 試題之比較分析

A Study on Generative AI in Developing Chinese Proficiency

Test Items: A Comparative Analysis with TOCFL Questions

施俐綺

Shih, Li-Chi

指導教授：邱詩雯 博士

Advisor: Chyu, Shih-Wen, Ph.D.

中華民國 114 年 7 月

July 2025

謝 誌

從以前查找論文時，就很喜歡看每個人的謝誌的我，現在也終於輪到我給別人看了啊……時間真的過得很快，沒想到我竟然也寫完碩士論文了。這過程好像很長、做了很多事，但回頭看，又好像只是轉瞬間的事。

這趟旅程中，首先要感謝的當然是我的指導教授—詩雯老師。在我一開始還不確定要寫什麼主題的時候，老師就拋出了一個有趣又有挑戰性的題目，讓我這一路上邊寫邊學，開了不少眼界。同時也非常感謝口考委員們繁忙中還是非常認真地看了我的全文，且溫柔地給了我很多實質又到位的建議。一路上遇到的師長都既專業、又充滿著教育愛，感動與感謝之餘，更讓我對研究有了更深刻的反思與成長。

謝謝我的爸爸、媽媽、哥哥、姐姐總是這麼溫暖地支持我、關心我，每個人都用自己的方式表達著對我的愛，讓我可以沒有太多壓力的情況下，一步一步地完成各式各樣的挑戰。也要感謝我最好的室友和朋友，謝謝你們給予的鼓勵與陪伴，在每個時候都是我的精神支柱。

很感謝這段過程中的所有人和事，讓我把這份小作品完成。雖然只是人生中的一小段路，卻是美好難忘的回憶。希望未來的我，也能繼續保持這份好奇心與韌性，探索更多未知的路！

施俐綺

2025 年 7 月於臺北

摘要

本研究旨在探討生成式人工智能於華語文分級測驗試題編寫上的應用可行性，並以 TOCFL 模擬題本為參照，分析 AI 所生成之聽力與閱讀題本在語言難度、認知層次、試題結構及作答表現上的表現與差異。本研究所使用的生成工具為 GPT-4，以 TOCFL 入門基礎級之等級設計試題，並與人工編寫之 TOCFL 題本進行比較分析。本研究共邀請 40 位華語文程度達 A1 以上的越南籍華語學習者參與施測，藉由分析受試者的答題表現進行試題的區辨度與命題品質評估，同時輔以 PIRLS 四層次提問架構分析試題提問難度與類型，再透過 Bloom 認知層次理論分析 AI 的表現，來看 AI 在華語文試題編寫上所展現的認知能力與限制。

研究結果顯示，AI 所生成的試題在語言形式與題型模仿方面表現穩定，尤其於聽力題型中展現出與人工題本相近的理解難度與得分分布，可有效檢驗出學習者的學習難點，快速生成有效且多樣的試題，展現其具備一定的命題潛力。然而，AI 閱讀題本整體難度偏低，受試者分數分布集中，導致區辨度不足。此外，AI 難以穩定掌握語言等級與提問層次，常出現超綱詞彙與語境提示不足等問題。認知層次分析亦發現，AI 多生成低層次提問（如直接提取、直接推論），在整合、評鑑等高層次命題能力上仍顯不足。

綜合而言，生成式 AI 可作為華語命題的輔助工具，具備提高出題效率與生成初稿的實用價值，但在高層次的語境邏輯以及誘答選項設計上，仍需仰賴人工審查與語言判斷，以確保試題品質與測驗效度。

關鍵字：人工智能、華語文能力測驗、華語文測驗編寫

Abstract

This study investigated the applicability of generative artificial intelligence (AI) in developing test items for Chinese language proficiency tests. Using the Test of Chinese as a Foreign Language (TOCFL) mock tests as a reference, this study analyzed the performance and differences of AI-generated listening and reading test booklets in terms of language difficulty, cognitive levels, test item structure, and test-taking performance. The generative tool employed in this study was GPT-4, which was used to design test items at the TOCFL Band A (Beginner) level. These AI-generated items were then compared and analyzed against manually compiled TOCFL test booklets.

A total of 40 Vietnamese Chinese language learners with proficiency levels above A1 participated in the assessment. By analyzing the test takers' performance, the discriminability and quality of the test items were evaluated. Concurrently, the PIRLS four-level questioning framework was utilized to analyze the difficulty and types of questions. Furthermore, Bloom's Taxonomy of Cognitive Levels was applied to analyze the AI's performance, examining the cognitive abilities and limitations demonstrated by AI in Chinese language test item development.

The study results indicated that AI-generated test items exhibited stable performance in terms of linguistic forms and item type imitation, particularly in listening comprehension, where they showed similar comprehension difficulty and score distribution to human-compiled test booklets. These items effectively identified learners' learning difficulties and rapidly generated valid and diverse test items, demonstrating AI's potential in test item development. However, the

overall difficulty of the AI-generated reading test booklets was low, leading to a concentrated distribution of test taker scores and insufficient discriminability. Additionally, AI struggled to consistently control language levels and questioning tiers, frequently producing out-of-syllabus vocabulary and insufficient contextual cues. Cognitive level analysis also revealed that AI predominantly generated low-level questions (e.g., direct retrieval, direct inference) and remained inadequate in high-level item development capabilities such as integration and evaluation.

In conclusion, generative AI can serve as an auxiliary tool for Chinese language test item development, offering practical value in enhancing item generation efficiency and producing preliminary drafts. Nevertheless, for high-level contextual logic and the design of distractor options, human review and linguistic judgment remain essential to ensure test item quality and test validity.

Keywords: Artificial intelligence, Test of Chinese as a Foreign Language (TOCFL) , Chinese language test item development

目次

謝誌.....	i
摘要.....	ii
Abstract.....	iii
目次.....	v
表次.....	vii
圖次.....	ix
第一章 緒論.....	1
第一節 研究背景與動機.....	1
第二節 研究目的.....	2
第三節 研究範圍.....	3
第四節 研究價值.....	3
第五節 名詞釋義.....	5
第二章 文獻探討.....	7
第一節 華語文測驗概述.....	7
第二節 華語文能力測驗（TOCFL）分析.....	11
第三節 華語文能力測驗（TOCFL）試題編寫流程.....	14
第四節 生成式 AI 在各領域的表現.....	15
第五節 生成式 AI 在試題編寫上的相關研究.....	20
第三章 研究設計.....	23
第一節 研究架構.....	23

第二節 研究對象	26
第三節 TOCFL 題本分析	27
第四節 AI 試題編寫模型訓練過程	40
第五節 測驗設計考題架構	52
第六節 資料收集與分析方法	54
第四章 研究結果	57
第一節 受試者表現概述	57
第二節 AI 題本與人工題本的比較	58
第三節 受試者的答題表現分析	64
第五章 結論與建議	95
第一節 研究發現與結論	95
第二節 研究限制	100
第三節 教學建議及未來研究建議	101
參考文獻	105
附錄	
AI 試題生成指令 (Prompts)	109

表次

表 2-1	TOCFL 各等級考試介紹.....	8
表 2-2	TOCFL 聽力、閱讀題型說明.....	11
表 2-3	TOCFL 聽力、閱讀通過標準.....	13
表 3-1	受試者 TOCFL 適性測驗結果統計.....	26
表 3-2	聽力測驗題型說明.....	27
表 3-3	閱讀測驗題型說明.....	29
表 3-4	聽力測驗 PIRLS 提問類型分佈與數據統計.....	34
表 3-5	聽力測驗各題型的 PIRLS 提問類型分佈.....	35
表 3-6	閱讀測驗 PIRLS 提問類型分佈與數據統計.....	35
表 3-7	閱讀測驗各題型的 PIRLS 提問類型分佈.....	37
表 3-8	聽力與閱讀測驗通過分數與題數對照表.....	38
表 3-9	AI 試題與人工試題之比較.....	45
表 3-10	AI 試題生成的圖片問題.....	51
表 3-11	閱讀測驗題本架構.....	53
表 3-12	聽力測驗題本架構.....	54
表 4-1	閱讀測驗兩題本各部分的斷詞表.....	58
表 4-2	聽力測驗兩題本各部分的斷詞表.....	61
表 4-3	兩題本閱讀測驗中的 PIRLS 類型題數統計表.....	63
表 4-4	兩題本聽力測驗中的 PIRLS 類型題數統計表.....	63
表 4-5	各題本成績統計表.....	65
表 4-6	閱讀測驗各能力組的分數差統計表.....	68
表 4-7	聽力測驗各能力組的分數差統計表.....	70
表 4-8	閱讀測驗的高答錯率試題分析.....	78

表 4-9 聽力測驗的高答錯率試題分析 88



圖次

圖 2-1	TOCFL 閱讀測驗分級方式及與歐洲共同語文參考架構 (CEFR)、美國外語教學協會指標 (ACTFL PROFICIENCY GUIDELINES)、臺灣華語文能力基準(TAIWAN BENCHMARKS FOR THE CHINESE LANGUAGE, 簡稱 TBCL) 之對應關係.....	9
圖 2-2	TOCFL 聽力測驗分級方式及與歐洲共同語文參考架構 (CEFR)、美國外語教學協會指標 (ACTFL PROFICIENCY GUIDELINES)、臺灣華語文能力基準(TAIWAN BENCHMARKS FOR THE CHINESE LANGUAGE, 簡稱 TBCL) 之對應關係.....	10
圖 3-1	研究流程	25
圖 3-2	直接提取範例題	32
圖 3-3	直接推論範例題	32
圖 3-4	詮釋、整合觀點和訊息範例題	33
圖 3-5	AI 試題的生成流程圖.....	43
圖 4-1	閱讀測驗兩題本的成绩分佈圖	73
圖 4-2	聽力測驗兩題本的成绩分佈圖	74
圖 4-3	閱讀測驗兩題本的每題答錯率分佈圖	75
圖 4-4	聽力測驗兩題本的每題答錯率分佈圖	76

第一章 緒論

第一節 研究背景與動機

近年來，隨著國際間對華語學習的關注持續升高，各類標準化語言能力測驗也逐漸發展起來，如中國的漢語水平考試（Hanyu Shuiping Kaoshi，簡稱 HSK）與我國的華語文能力測驗（Test of Chinese as a Foreign Language，簡稱 TOCFL）等，皆是目前具代表性的華語能力測驗。這些標準化測驗設計多依循歐洲共同語文參考架構（CEFR）進行語言分級，除了可幫助學習者了解自身語言能力，也作為學校或機構在入學、畢業時的重要語言依據。其中 TOCFL 的入門基礎級（A1 - A2）在我國的華語教學現場極具實用性，尤其在大專校院中，許多課程規定外籍生須於入學前或在學期間通過 A2 等級測驗，才可修習正式課程。此級別作為銜接生活溝通與學術語言的初步能力門檻，使得其測驗题目的設計品質與效度具有非常關鍵的意義。然而，TOCFL 試題的編寫過程極為繁複，需依據「臺灣華語文能力基準」（TBCL）進行詞彙選擇與語用情境設計，並經歷審題、改寫、預試、統計分析等多道程序，每一題的完成需仰賴命題教師及測驗專家的大量投入，對測驗資源開發造成極大壓力。

在此背景下，生成式人工智慧（Generative Artificial Intelligence, GenAI）技術的快速發展，為語言測驗編製帶來全新的可能性。生成式 AI 指的是能根據大量語料自動生成新內容的人工智慧系統，目前已可應用於生成文本、圖像、聲音等多種類輸出。以目前最普遍使用的 ChatGPT 為例，其所依據的 GPT-4 語言模型現今已能處理複雜語境，生成符合語法規範與語意邏輯的內容，並具備語言練習、對話模擬、語篇重寫等應用潛力（Creely, 2024；Pérez et al., 2023）。另外，多項研究指出，ChatGPT 作為語言教育輔助工具，具備高度客製化與即時回饋能力，可協助教師設計練習題、協

助學生釐清語法用法，並營造更具互動性的學習情境(Kavak, Evis, & Ekinici, 2024)。根據 Kavak 等人(2024)對 ChatGPT 的應用分析，其在語言教學中最常見的優勢包括：促進語言練習、提供個別化回饋、強化學生自信心與語言動機等；然而，研究者也提醒，生成式 AI 所產生的內容仍可能出現文化失真、語意偏差或難以辨別語境的情況，若缺乏人為判讀與語用審核，恐將影響其應用品質。

目前雖已有文獻初步探討 AI 在語言教學上的潛力，但針對 AI 能否有效應用於標準化測驗命題的研究，特別是如 TOCFL 這類具有明確語言分級、結構規範與施測對象的標準化測驗，仍較少有整合 AI 試題與官方試題進行語言層次與作答行為比較的實證研究。因此，本研究將選擇以 TOCFL 入門基礎級(A1 - A2)為例，實際應用 ChatGPT 的 GPT-4 模型生成聽力與閱讀試題，並與 TOCFL 人工試題進行語言結構與命題層次的比較，進一步透過受試者作答結果分析 AI 試題之難度、鑑別度與適切性，期望釐清生成式 AI 作為命題輔助工具的可能性與限制，為未來華語分級測驗的資源開發與應用方向提供具體參考。

第二節 研究目的

本研究旨在探討生成式人工智慧應用於華語分級測驗試題編寫的可行性與潛力。本研究將透過實際題本的編製與施測，從命題品質、語言難度、認知層次展現與受試者表現等面向，深入分析 AI 生成試題與 TOCFL 人工試題的異同，期望能初步釐清生成式 AI 在華語能力測驗試題編寫上的可行性，並探討其未來在教學與測驗領域的應用可能。

本研究不僅可作為初步了解生成式 AI 在語言測驗命題上可行性的參考，同時也希望能夠為教學現場提供一個具體的應用案例。若 AI 能有效支援標準化測驗中的命題工作，將有助於降低教師與命題團隊的負擔，亦可

提升命題的效率與彈性，讓 AI 更貼近實際的語言教學需求，發揮技術在教育上的輔助價值。

第三節 研究範圍

本研究以臺灣教育部所主辦之華語文能力測驗（TOCFL）為試題參考架構，聚焦於入門基礎級（A1-A2）之聽力與閱讀測驗，探討生成式人工智慧於該級別試題編寫的可行性與應用潛力。本研究主要採用 GPT-4 語言模型進行試題生成，並以 TOCFL 的官方模擬題本為對照基準，進行語言難度分析、命題結構比較與受試者答題表現的統計分析。本研究僅聚焦於標準化測驗中的封閉式選擇題（multiple choice），不包含口說與寫作等建構式題型，也不涉及 GPT 模型的技術訓練或微調過程。分析重點集中於生成試題的語言特徵、展現的認知層次與受試者表現之差異，並不涵蓋 AI 教學對話、課文生成或其他教學應用情境。

第四節 研究價值

本研究結合標準化語言測驗命題原則與生成式人工智慧技術，嘗試以 GPT-4 語言模型生成華語分級測驗試題，並與 TOCFL 官方模擬試題進行系統性的比較分析。過去多數 AI 應用於語言教學的研究多著重於課文生成、對話模擬或語法練習設計，較少深入探討 AI 在語言測驗命題工作中的具體操作與命題品質，特別是針對具分級標準與語言能力指標之測驗，更缺乏相關的實證研究。因此，本研究希望補足此一研究空缺，提供一套具體可行的實驗設計與分析方式，做為未來延伸應用與研究的基礎。

實務層面而言，AI 若能協助語言教師快速產出符合難度與語用場景的測驗題目，將有助於降低命題壓力與人力成本，提升教材與評量資源的彈性與效率。尤其在華語教學資源有限、教師需兼顧教學與測驗編寫的情況

下，若能導入適當的 AI 工具，將可能成為實際命題流程中的有力輔助。此外，本研究以 TOCFL A1-A2 為測驗等級，呼應目前各大學對外籍生的語言能力需求，具備高度應用性。

理論層面而言，本研究採用 PIRLS 四層次提問架構(Sabatini et al., 2024) 以及布魯姆 (B.S. Bloom) (Anderson et al., 2001) 的認知層次理論，檢視 AI 試題的語言結構與認知層次表現，並以受試者實測的結果數據資料加以驗證，補足過去以模型生成結果為主、較少結合實際學習者回饋的研究限制。透過本研究所建構的比較分析方法與施測設計，未來相關研究可進一步延伸至不同語言級別、其他技能測驗，甚至納入教師或命題專家的評估意見，發展更完整的 AI 命題評鑑架構。

綜合來說，本研究不僅在技術應用層面提供操作實例，也在語言測驗與教材研發領域提出新的研究觀點，期待能為華語教育數位轉型與 AI 輔助教學開啟更多可能性。



第五節 名詞釋義

壹、生成式人工智慧 (Generative Artificial Intelligence)

泛指能夠根據輸入資訊，自主產出新內容的人工智慧系統，常見輸出形式包括圖像生成、語音合成與自然語言處理等。本研究所指的生成式 AI 為具備語言生成能力之大型語言模型，如 OpenAI 開發的 GPT-4，可應用於文章撰寫、對話模擬及測驗試題產出等教育情境。

貳、Progress in International Reading Literacy Study 四層次提問

類型 (PIRLS Question Types)

根據國際閱讀素養研究 (PIRLS) 所發展的閱讀理解分類架構，包括：直接提取 (Retrieve)、直接推論 (Infer)、詮釋整合 (Interpret/Integrate)、評估批判 (Evaluate/Critique)。本研究以此為試題命題分析工具。

參、布魯姆 (B.S. Bloom) 認知領域教育目標分類

此為一種將學習目標依複雜度分為六個層次的系統。它從最低的「知識」層次開始，逐步提升至「理解」、「應用」、「分析」、「綜合」，最終達到最高的「評鑑」層次。此分類法為教育者規劃教學、設計課程及評估學生認知能力提供了系統框架。本研究將以此分析目前生成式 AI 的認知能力表現。



第二章 文獻探討

本章將回顧與本研究相關的研究文獻，藉此探討華語文測驗類型與試題編寫標準，並進一步分析生成式 AI 在試題編寫上的應用與研究現況。透過相關研究的整理與分析，本章的文獻回顧將有助於確認 AI 技術在華語文測驗試題編寫中的可行性，並為本研究提供理論依據與分析框架。

第一節 華語文測驗概述

壹、TOCFL 華語文能力測驗

「華語文能力測驗」(Test of Chinese as a Foreign Language, TOCFL) 為我國教育部委託「國家華語測驗推動工作委員會」(簡稱華測會)所主辦的標準化語言測驗，旨在評量非華語母語者在聽、說、讀、寫各面向的華語能力。該測驗以「華語文能力基準 (TBCL)」為設計依據，強化測驗標準與評估體系，提升測驗內容與學習者實際語言應用需求之契合度。

TBCL 為我國國家教育研究院所建置，作為華語能力標準體系之一，規劃七級語言能力指標，並針對漢字、詞語與語法點進行系統性分級。該基準不僅提供教學與學習的重要參考依據，也深度影響 TOCFL 測驗之命題原則與語言難度安排。

目前 TOCFL 採「四等八級」之分級制度，分別為「準備級」、「入門基礎級」(第 1、2 級)、「進階高階級」(第 3、4 級)，以及「流利精通級」(第 5、6 級)。各等級測驗主要涵蓋聽力與閱讀，口說與寫作則為獨立測驗項目，需另行報考。為協助理解 TOCFL 測驗結構與其分級設計，下方表 2-1 呈現 TOCFL 各等級測驗與 TBCL 能力基準之大致對照關係，而聽力測驗與閱讀測驗與各能力指標的具體對應，可參考圖 2-1 與圖 2-2。

表2-1

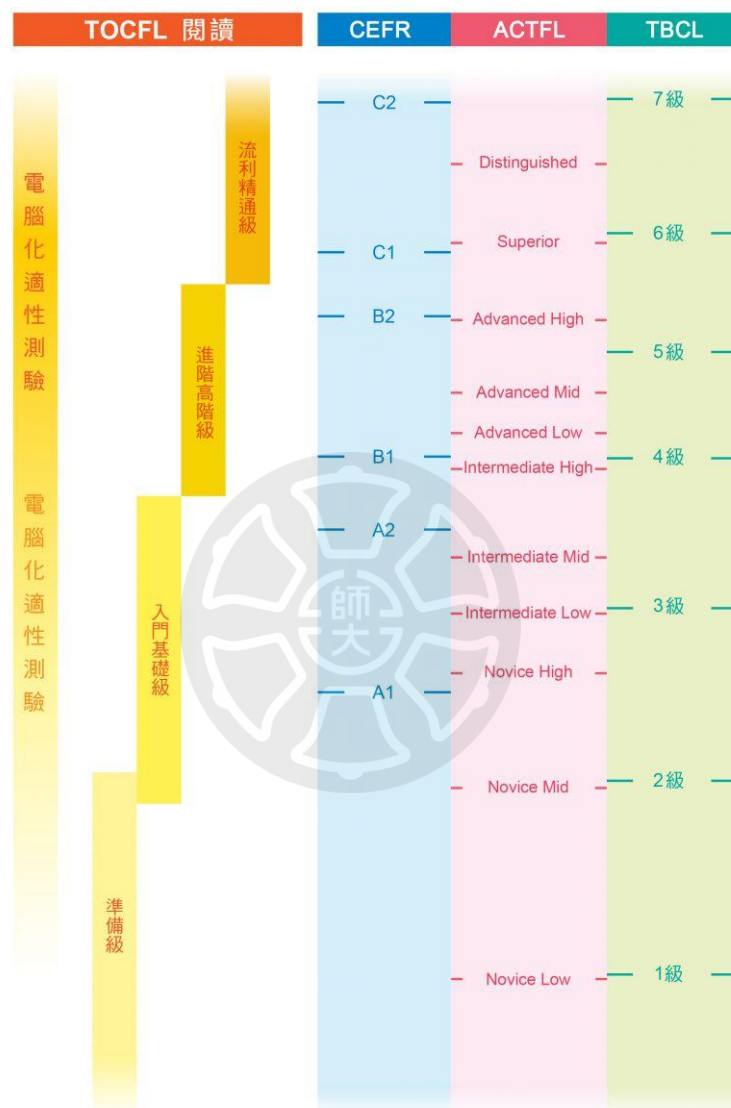
TOCFL 各等級考試介紹

等級	測驗項目	測驗題數	測驗時間	對應 TBCL 等級
準備級一級	聽力、閱讀	約 100 題	約 120 分鐘	1
準備級二級	聽力、閱讀	約 100 題	約 120 分鐘	2
入門級	聽力、閱讀	約 100 題	約 120 分鐘	2
基礎級	聽力、閱讀	約 100 題	約 120 分鐘	3
進階級	聽力、閱讀	約 100 題	約 120 分鐘	4
高階級	聽力、閱讀	約 100 題	約 120 分鐘	5
流利級	聽力、閱讀	約 100 題	約 120 分鐘	6
精通級	聽力、閱讀	約 100 題	約 120 分鐘	7

資料來源：國家華語測驗推動工作小組 (n.d.)。華語文能力測驗 (TOCFL)。取自 <https://tocfl.edu.tw/index.php/home/index> (取用日期：2024 年 10 月)

圖2-1

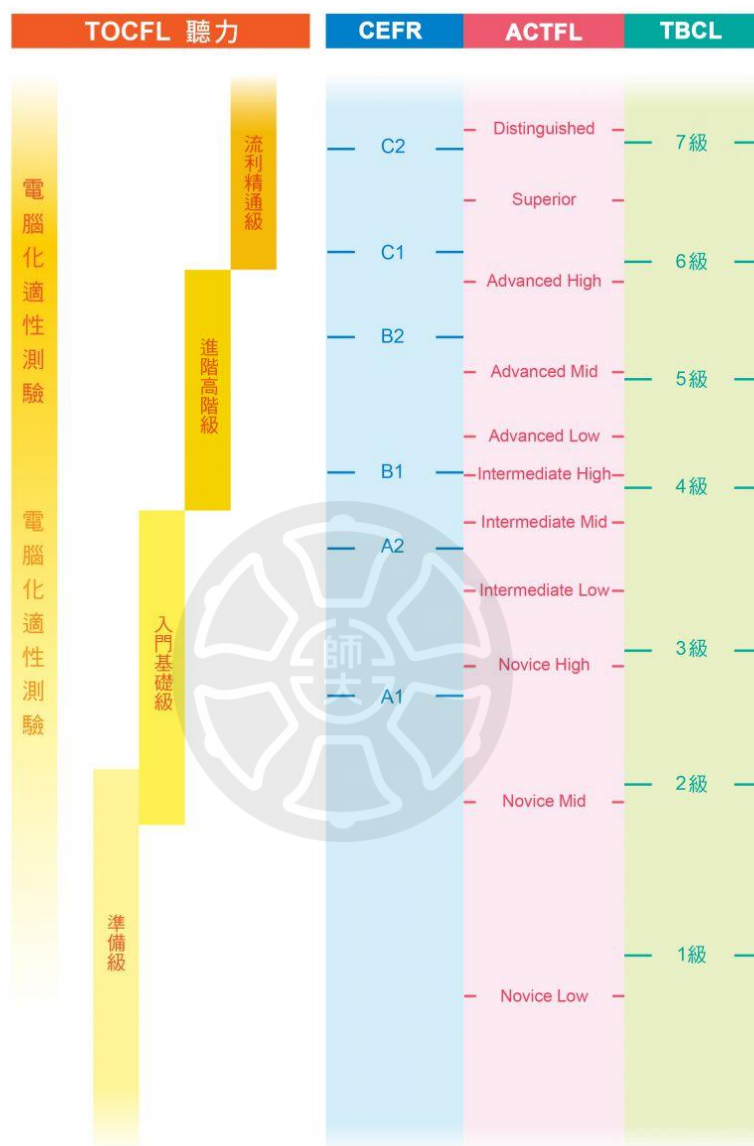
TOCFL 閱讀測驗分級方式及與歐洲共同語文參考架構 (CEFR)、美國外語教學協會指標 (ACTFL Proficiency Guidelines)、臺灣華語文能力基準 (Taiwan Benchmarks for the Chinese Language, 簡稱 TBCL) 之對應關係



資料來源：國家華語測驗推動工作小組 (n.d.)。華語文能力測驗 (TOCFL) 等級對照。取自 <https://tocfl.edu.tw/tocfl/index.php/test/cat/list/4> (取用日期：2024 年 10 月)

圖2-2

TOCFL 聽力測驗分級方式及與歐洲共同語文參考架構 (CEFR)、美國外語教學協會指標 (ACTFL Proficiency Guidelines)、臺灣華語文能力基準 (Taiwan Benchmarks for the Chinese Language, 簡稱 TBCL) 之對應關係



資料來源：國家華語測驗推動工作小組 (n.d.)。華語文能力測驗 (TOCFL) 等級對照。取自 <https://tocfl.edu.tw/tocfl/index.php/test/cat/list/4> (取用日期：2024 年 10 月)

第二節 華語文能力測驗 (TOCFL) 分析

壹、聽力測驗與閱讀測驗之考試題型

根據 TOCFL 官方網站資訊所示，該測驗編製過程皆按照標準化測驗程序進行，屬於標準化的語言能力測驗。為檢測華語學習者的日常生活語言使用能力，測驗命題皆不以任何教材為依據，測驗內容涵蓋日常生活各情境所設計，使用多元且真實的不同體裁。測驗題型依照不同測驗等級有所差異，為了能清楚瞭解 TOCFL 考試題型，本研究參考 TOCFL 官方網站所提供的模擬試題（最後更新時間為 2024 年 7 月），整理成表 2-2 以呈現聽力及閱讀測驗所有題型，並在表格中簡短說明各題型的答題說明，以及該題型出現的測驗等級。

表2-2

TOCFL 聽力、閱讀題型說明

測驗內容	題型	說明	測驗等級
聽力	單句理解	聽完一個句子後，從圖片選項中選出正確答案。	準備級
	看圖回答	聽完三個選項後，選出符合圖片內容的答案。	準備級、 入門基礎級
	問答理解	聽到一段兩句的對話以後，根據對話內容，選出符合這段對話的圖片。	入門基礎級
	對話理解	聽到一段約四句的對話以後，根據對話內容，選出符合這段對話的圖片。	入門基礎級

(續上頁)

	對話	聽完一段對話後，根據對話回答問題。	準備級、 入門基礎 級、進階高 階級、流利 精通級
聽力	段落	聽完一段話以後，根據提問，選出符合這段話的答案。	進階高階 級、流利精 通級
閱讀	單句理解	看完一個句子後，從圖片選項中選出正確答案。	準備級、 入門基礎級
	看圖釋義	看完三個選項後，選出符合圖片內容的答案。	準備級、 入門基礎級
	閱讀理解	閱讀完一篇短文以後，根據提問，選出符合這段短文的答案。	入門基礎 級、進階高 階級、流利 精通級
	選詞填空 (句子)	有一張情境圖，圖片下面有五個句子。根據圖片內容，選出合適的答案。	入門基礎級
	選詞填空 (短文)	每一段短文有幾個空格，每個空格有四個選項，根據上下文，選出正確的答案完成短文。	進階高階 級、流利精 通級

(續下頁)

完成段落 一篇短文裡有五個空格，一共有 六個可能的選項，根據文章的上
入門基礎級
六個可能的選項，根據文章的上
下文，選出合適的答案。

貳、聽力測驗與閱讀測驗之通過標準

根據 TOCFL 官方網站的說明，該測驗採用量尺分數 (Scale Scores) 及標準設定 (Standard Setting) 的結果作為判斷通過等級的依據標準。聽力及閱讀測驗各共有 50 題，滿分為 80 分，答對題數越多者，量尺分數越高。為更清楚瞭解各等級通過標準，本研究將各等級的通過分數整理成表 2-3。

表2-3

TOCFL 聽力、閱讀通過標準

測驗等級		聽力測驗通過分數	閱讀測驗通過分數
準備級	一級	38	39
	二級	59	61
入門基礎級	入門級	41	42
	基礎級	57	60
進階高階級	進階級	46	48
	高階級	61	64
流利精通級	流利級	50	52
	精通級	61	69

從表 2-3 中可見，各級別的閱讀測驗通過標準均高於聽力測驗，這可能代表閱讀試題的難度較大，或是普遍來說學習者在閱讀的表現相對較弱。基於此，本研究也將探討同一學習者在聽力測驗與閱讀測驗上的表現差異，

再進一步分析這種情況是否會隨語言程度的不同而有所變化，並且本研究也將從試題編寫角度探討影響造成考生表現差異的因素。

第三節 華語文能力測驗（TOCFL）試題編寫流程

壹、試題編寫的目的與標準

華語文能力測驗（TOCFL）旨在評估學習者的華語能力，試題的編寫須確保測驗的有效性（validity）、信度（reliability）及公平性（fairness），因此，TOCFL 的試題編寫流程需經過多重審查與測試，以確保測驗能夠準確反映考生的語言能力（TOCFL, n.d.）。以下分點說明華語文能力測驗試題編寫步驟，並作為未來 AI 試題生成參考。¹

貳、試題編寫流程

- 一、 確立測驗目標與大綱：測驗專家根據「臺灣華語文能力基準（TBCL）」制定測驗大綱，確立各級別測試的語言能力範圍、題型及測驗重點。
- 二、 試題撰寫：由受過專業培訓的試題編寫者根據測驗大綱撰寫題目，並利用「國教院的教材編輯輔助系統」檢視題目語言的詞彙等級，以確保試題符合不同等級考生的語言能力需求。
- 三、 內部審查與修正：編寫完成的試題須經過專家審查，以評估試題的語言準確性、難度適切性及文化適應性等，並進行必要修正。
- 四、 測驗預試與數據分析：修訂後的試題會進行預試，並透過統計方法分析試題的難度（difficulty）、鑑別度（discrimination）及信度（reliability），以確保試題品質

¹ 國家華語測驗推動工作委員會（n.d.）試題編寫流程。取自 <https://tocfl.edu.tw/tocfl/index.php/teach/examples>。

五、正式施測與持續修正：經過預試後，合格的試題才會納入正式測驗題庫，並根據測驗結果持續調整與優化。

TOCFL 的試題編寫流程主要包含命題說明擬定、試題初稿撰寫、審題與修訂、試題預試與分析、以及通過審查後進入題庫等階段。根據曾文璇等人 2023 年針對 TOCFL 試題設計的說明，命題者在撰題前會先透過「華語文能力基準整合應用系統」查詢對應等級之詞彙與語法範疇，並依據《華語八千詞表》控制試題中出現的詞彙等級。此外，為強調語境真實性與語用功能，試題內容多參考報章雜誌、網路文本、影音素材等，並依據 TBCL 主題分類架構設計出與學習者生活經驗相關的場景與溝通任務。若初稿出現超綱詞彙，系統平台亦會提供建議詞彙以供替換，協助命題者控制難度，維持試題品質的一致性。

本研究將參考既有的試題編寫流程與工具，結合生成式 AI 進行試題設計，旨在探討 AI 技術於華語文能力測驗命題工作的適用性與可行性。同時，也藉此確保生成題本與人工題本在命題原則上的一致性，以避免影響後續施測結果的分析準確性。

第四節 生成式 AI 在各領域的表現

現今生成式 AI 的發展已日漸成熟，為了分析生成式 AI 技術在試題編寫領域的應用，我們需要先評估目前生成式 AI 技術在模仿人類認知層次上的表現。因此本研究將利用布魯姆 (B.S. Bloom) 認知領域層次理論，分析當前生成式 AI 技術的認知領域層次表現，以及運用到各領域的成果。

壹、布魯姆 (B.S. Bloom) 認知領域教育目標分類 (新版)

根據布魯姆 (B. S. Bloom) 提出的認知領域教育目標分類 (Bloom's Taxonomy)，學習與思考的過程可分為六個層次，分別為知識 (Knowledge)、

理解(Comprehension)、應用(Application)、分析(Analysis)、綜合(Synthesis)、評鑑(Evaluation);到了2001年則有了修訂版,將認知領域層次從簡單到複雜分成記憶(Remember)、瞭解(Understand)、應用(Apply)、分析(Analyze)、評鑑(Evaluate)、創造(Create),以下分項說明(Anderson et al., 2001)。

- 一、 記憶(Remember):為最低層次的認知歷程。指的是學習者可以針對所學,從自身長期記憶提取相關資訊,此歷程中又包括再認(Recognizing)和回憶(Recalling)。
- 二、 瞭解(Understand):此層次指的是學習者能從教學內容中連結新知與舊知,以建構觀念架構來理解學習內容。此歷程又包括詮釋(Interpreting)、舉例(Exemplifying)、分類(Classifying)、摘要(Summarizing)、推論(Interring)、比較(Comparing)及解釋(Explaining)等。
- 三、 應用(Apply):此層次指的是學習者可以利用已學內容,來執行任務及解決問題。此認知歷程包括執行(Executing)和實行(Implementing)。
- 四、 分析(Analyze):此層次指的是學習者得以組織與區別其所學,並發現其中隱含資訊,進而提出假設。此歷程包括區辨(Differenting)、組織(Organizing)和歸因(Attributing)。
- 五、 評鑑(Evaluate):此層次指的是透過所學,學習者可以發展出自身標準,並藉此做出判斷。此歷程包括檢查(Checking)和批判(Critiquing)。
- 六、 創造(Create):為最高層次的認知歷程。此層次指的是學習者可以透過自身創作思考能力,重組、形成一個新的組合或結構。此歷

程包括生產 (Generation)、計畫 (Planning) 和製作 (Producing)。

貳、人工智慧生成技術與人類創造力的認知理論

DiPaola、Gabora 與 McCaig (2018) 的研究以生成圖像的深度學習 DeepDream 算法為例，主要探討如何將人類創造力的認知理論融入生成式人工智慧技術，希望藉此提高生成作品的人性化與創意性。接下來，本研究將以 DiPaola 等人的研究結果，結合布魯姆 (B.S. Bloom) 的認知層次理論，討論生成式 AI 在各層次的表現。

第一，在記憶 (Remembering) 層次上，目前生成式 AI 已有相當不錯的成果表現。DiPaola 等人發現深度學習算法能從輸入的圖像提取視覺特徵，透過數據訓練以學習、記憶不同的藝術風格與特徵，除此之外此算法也能有效地識別不同視覺特徵，如藝術品的紋理與結構等。而在瞭解 (Understand) 層次上，生成式 AI 在一定程度上可理解輸入數據間的關聯性，如該研究透過情境式聚焦 (contextual focus) 的指令，引導生成式 AI 對輸入資訊的理解；然而其理解能力仍仰賴既有的模型訓練，因此對更深層的文化理解尚有限。第二，在應用 (Apply) 層次上，生成式 AI 可以將學習到的知識模型應用在新的創作上，故能結合多個藝術特徵，創造出具審美標準的藝術圖像。但在分析 (Analyze) 上的表現較有限，目前生成式 AI 的分析較多體現在較表層的技术操作層面，難以分析文化背景下的符號學意義。第三，在評鑑 (Evaluate) 與創造 (Create) 的高層次表現都相當有限。DiPaola 等人指出生成式 AI 的評價缺乏主觀性和多樣性，仍需要大量的人工介入與監控；而生成式 AI 的創造性也依賴既有數據，多屬於「再現性創造 (reproductive creativity)」，缺乏人類的創造情感與意圖，無法實現「原創性創造 (transformational creativity)」。

參、人工智慧生成技術與決策認知模型

Malloy 和 Gonzalez (2024) 的研究中提出了一種將生成模型 (GMs) 整合到決策認知模型中的方法，藉此檢測生成式 AI 在決策認知上的表現。接下來，本研究將同樣透過 Malloy 等人的研究結果，討論生成式 AI 在各認知層次的表現。

第一，在記憶 (Remembering) 層次與瞭解 (Understand) 層次上，AI 模型已有的顯著表現。該研究的測試模型能記住大規模數據的模式與特徵，並基於上下文資訊，進行檢索和再現，如記錄視覺特徵、語言結構和詞彙語意等；然而，生成式 AI 對記憶搜尋無選擇性，容易生成不相關內容。另外，該研究結果與 DiPaola (2018) 的結果相互證實，兩者皆發現 AI 模型對於深層語意或文化背景的理解有限。第二，在應用 (Apply) 層次上，其發現 AI 模型可將記憶內容應用在新的情境任務上，但其侷限於既有模板，對於創新場景的表現較差。而在分析 (Analyze) 上，AI 模型能有效地分析數據的內部結構，如分類相似物品，但其對於數據背後的邏輯和情境的分析則相對不足。第三，同樣與 DiPaola (2018) 等人的研究結果相同，AI 模型在評鑑 (Evaluate) 與創造 (Create) 的高層次表現最為不足。Malloy 等人發現，AI 模型能生成決策建議，但無法進行自主的主觀評價，在人類價值觀的道德和倫理層面考量幾乎為零；另外，AI 模型能通過數據重組，進行多樣化的藝術和自然語言生成，但缺乏原創性的深層創意。

肆、生成式 AI 在教育領域的應用

目前各國對於人工智慧運用到教學現場的態度不一，有一部分國家對此抱持謹慎的態度，甚至禁用人工智慧技術。以義大利為例，義大利當局因考量隱私保障，於 2023 年 3 月底起限制義大利使用者使用 ChatGPT

(Mccallum, 2023)。另一部份國家則接受這項技術，並運用在教學現場上，像是香港、印度、西班牙、英國與澳洲等國，皆在國家教育中加入了人工智慧課程，並鼓勵教師與學生運用此技術（鏗科技，2023）。

近年來，我國已有許多教師發展結合生成式 AI 的課程，這樣的科技工具在教育領域學習輔助的應用上，已展現了多樣的潛力。首先，生成式 AI 的即時性可以提高學生的學習興趣，特別是在語言學習、寫作及問題解答等方面具有顯著的效果（顏榮泉，2024），並提供學生個性化的學習輔助，促進學習的個別化發展（Wu & Tsai, 2022）。例如全球教育科技品牌 ViewSonic 與我國教科書出版品牌康軒文教集團及南一教育集團合作，於 2024 年推出臺灣首款《AI 教科書》，此合作將康軒及南一的電子書整合至具生成式 AI 的 ClassSwift 即時互動教學平台，提供即時測驗、多元評量、學習歷程分析等功能；另外，我國另一教科書標竿品牌翰林出版也同樣運用 AI 科技，推出「AI e-Hanlin」智慧學習平台，提供個人化學習計劃與建議。總的來說，AI 技術運用到教育領域上的已有可見的成果，也有著無限的潛力，透過教師的引導與生成式 AI 的外部記憶系統，除了可輔助教學效率，亦可幫助學生應用對科技輔助工具的所學，強化其解決問題的能力（林穎俊，2024；顏榮泉，2024）。

然而，生成式 AI 固然便利，在目前的發展上尚有所缺失。首先，從教學角度來看，牛津大學出版社於 2021 年發表的《解決數位鴻溝》報告中顯示，全球有超過一半的教師和學習者仍缺乏數位學習技能，生成式人工智慧的出現可能使人們更難適應新技術（Oxford University Press., 2021）。美國教育出版商 McGraw Hill 於 2024 年 4 月公布全球教育調查報告「學習成效和數位科技教室」（Learning Outcome and the Digital Classroom），報告中提到有超過七成的教師希望學校能提供更多關於 AI 的教學指導，並對於 AI 將帶來的未知影響表示擔憂（游昊耘，2024）。再從學習角度來看，

顏榮泉 (2024) 指出，過度依賴生成式 AI 可能會導致學生認知負荷的降低，進一步影響其批判思維及高層次思考能力的發展 (Sparrow, Liu, & Wegner, 2011; Ward, 2021)。另外，生成式 AI 的事實準確性目前尚未發展完善，時常出現提供錯誤資訊的狀況，若學生過於依賴生成式 AI，也可能造成其對於自身產出內容有錯誤的自信，進而影響學習品質 (Ward, 2021)。

整體而言，生成式 AI 技術在輔助教學方面具備潛在優勢，但其實際效果仍取決於教師如何運用這些技術來促進有效的學習。因此在結合 AI 輔助工具的課程設計上，教師必須先理解並應用生成式 AI 的技術，針對學生學習需求與表現選用適合的 AI 來設計課程；另外，教師也應擔任學習路上重要的引導者，教師在課堂中須適時引導、監控，以降低學生因依賴技術而弱化自主思考和問題解決能力的風險 (伍柏翰、張雅綺，2024)。

第五節 生成式 AI 在試題編寫上的相關研究

生成式 AI 在教育領域的應用越來越普遍，隨著大型語言模型 (LLMs) 的發展，生成式 AI 已經可以根據文本內容快速生成測驗題目，甚至在特定情況下，試題的品質能夠接近人工編寫的題目。本節將整理現有的研究成果，分析 AI 在試題編寫方面的應用表現。

在國內外，關於此議題的研究皆已累積一定成果。其中，國內研究如國立中興大學的 ENABLE 中心曾在 2023 年進行一項 AI 自動生成試題與人工編寫試題的比較研究。該研究將生成式 AI 技術運用在英文試題編寫上，隨機混合 AI 生成題目與人工編寫之題目，並請英文老師評估考題的品質，以此分析 AI 生成的英文測驗題目的成效。其初步調查結果顯示，教師認為其生成模型有助於命題，但在誘答選項上生成模型仍需加強 (ENABLE, 2023)。除此之外，現今有許多教育平台也開始運用生成式 AI 的技術，協助教師自動生成測驗試題。如 Quizizz 教育平台，教師可以在

該平台上上傳教學資料，讓 AI 根據上傳內容生成相關的題目，更能根據指令進行適應性調整 (FlipEdu, 2023)。國外研究亦有許多成果，在程式設計教育領域上也出現了針對 AI 技術生成的多選題 (MCQs) 與人類編寫之題目的比較研究，該研究結果顯示，生成式 AI 能有效地生成語言清晰、選項明確且與學習目標高度一致的題目，其品質更與人工編寫的題目不相上下 (Singh et al., 2023)。然而，該研究也指出生成式 AI 可能會生成不完全正確或邏輯上不合理的選項，因此仍需依賴教師進行後續的審查與修正。這類研究皆顯示了 AI 在教育領域的可行性，更凸顯了 AI 技術在個人化試題與適性測驗的潛力及現階段技術仍存的限制。

第六節 小結

綜合現有研究，生成式 AI 在認知層次上已展現了在不同層次的應用中的潛力，特別是在記憶 (Remembering)、理解 (Understanding) 等較低層次上已有顯著的表現。DiPaola 等人 (2018) 指出，生成式 AI 不僅能模仿人類創造力，還能透過多層次的生成過程進行創新，顯示其在創作領域的應用潛力；Malloy 和 Gonzalez (2024) 則強調 AI 透過數據分析與環境資訊整合的特性，有助於提升決策準確性。此外，在教育輔助方面，顏榮泉 (2024) 發現生成式 AI 能夠提升學習者的參與感與個性化學習體驗，但過度依賴技術可能削弱學生的批判性思維 (Sparrow, Liu, & Wegner, 2011; Ward, 2021)。

儘管生成式 AI 已在多個領域展現應用價值，但基於演算法與參數運算的特性，使 AI 技術在較高的認知層次領域方面仍有所不足，仍需依賴人工進行後續檢驗與修正，以確保生成內容的品質與可用性 (Singh, 2023; Zhang, 2023)。在試題編寫領域，現有研究顯示 AI 生成試題雖已具備一定品質，但仍面臨幾項挑戰，包括試題難易度、試題的語言敘述及答案選項

合理性等問題 (Singh, 2023; Zhang, 2023)。AI 生成試題與人工試題的比較研究，目前大多仍集中於特定語言與學科，針對華語文測驗的相關研究尚有限，還需進一步探討 AI 在此領域的應用成效與潛在挑戰。

基於上述，本研究將運用生成式 AI 技術於華語文測驗試題的編寫，並透過對比分析 AI 生成試題與官方試題，探討 AI 技術在華語測驗領域的應用表現、試題品質及其可能的挑戰，並為未來 AI 在華語教育測驗的應用與發展提供參考。



第三章 研究設計

本研究旨在探討生成式 AI 在華語文能力測驗命題工作中的應用潛力。具體而言，本研究聚焦於 TOCFL 入門基礎級測驗，將比較由 AI 生成的試題與人工試題在語言難度、認知層次及學習者作答表現上的異同。本研究採用文件分析結合調查研究的設計，透過分析 TOCFL 官方考題的特徵作為 AI 訓練基礎，並由 GPT-4 生成對應的試題。接著，邀請某科技大學之越南籍華語學習者參與同組內比較設計的測驗，分別完成 AI 題本與人工題本（皆含閱讀與聽力測驗）。本研究將運用國教院的詞彙斷詞分析系統與 PIRLS 四層次提問架構對試題進行深度分析，並透過統計方式評估受試者在不同題本下的測驗成績與作答情形，以綜合判斷 AI 生成試題的難易度與區辨效度。本章節將詳細說明本研究的設計架構，包含研究方法、研究對象、測驗題本設計、施測流程、數據收集與分析方式。

第一節 研究架構

壹、 研究方法

本研究為達成探討生成式 AI 在華語文能力測驗命題應用之目的，主要採用文件分析法（Content Analysis）與調查研究法（Survey Method）兩種研究方法，以下分項說明。

一、文件分析法：本研究採用文件分析法，透過分析試題內容，以描述並比較 AI 生成試題與 TOCFL 人工試題在語言與命題結構上的異同。具體分析內容包括以下兩部分：

（一）TOCFL 官方模擬試題分析：在 AI 出題操作進行之前，研究先針對 TOCFL 於官方網站上所提供的模擬試題進行內容分析，統整各題型的出題方式、語言難度，以及 PIRLS 四層次提問架構

使用比例。此分析結果不僅作為 AI 模型學習與生成題目的依據，也為後續比較兩類試題品質提供標準化的參照基準。

- (二) AI 生成試題與人工試題對比分析：AI 生成試題完成後，亦以相同的分析架構進行文件分析，針對題本題目的語言使用、題型設計與提問層次等面向進行檢視，並與人工試題的分析結果進行對照，藉此評估 AI 生成試題在語言表現與命題邏輯上是否貼近 TOCFL 的命題規範。

二、調查研究法：本研究亦採用調查研究法，透過系統性方式蒐集受試者的作答數據，藉此了解 AI 生成之試題與 TOCFL 官方試題在實際施測中的表現差異。具體操作方式如下：

- (一) 作答表現數據蒐集：本研究邀請特定群體的華語學習者作為受試者，請其分別作答由 GPT-4 生成的 AI 題本與 TOCFL 官方模擬題本，並蒐集其測驗成績與各題答錯率等作答表現統計資料，作為後續分析之依據。
- (二) 品質與可行性評估：透過比較受試者在兩份題本上的作答表現，分析兩者在難易度、區辨度等面向上的差異，進一步探討 AI 生成試題的品質與其於華語測驗命題之應用可行性。

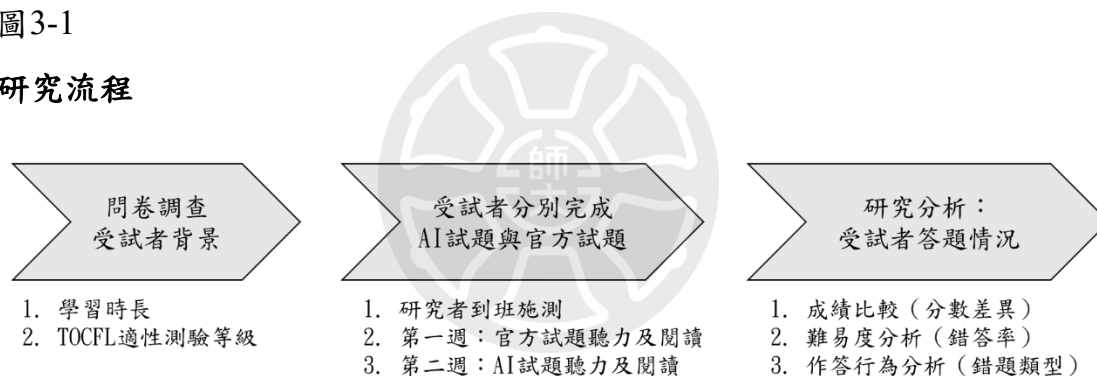
貳、 研究設計

本研究採用受試者內設計 (within-subject design)，即每位受試者均需完成 AI 生成試題與 TOCFL 模擬試題兩組測驗，以降低個體差異對測驗結果的影響。兩組試題的測驗內容皆包含閱讀測驗與聽力測驗，題本難度皆對應 TOCFL 的入門基礎等級 (A1-A2)，得分標準則皆參考 TOCFL 所提供的模擬試題測驗分數題數對照表以及等級通過標準分數表，以確保測驗公正性。

本研究流程可分為三個階段（如圖 3-1），分別為受試者背景調查、試題施測與數據分析。首先，本研究透過問卷調查受試者的華語學習經驗與測驗背景，並篩選語言能力達 CEFR A1 以上的學習者參與，以確保其能順利完成測驗，同時也有助於後續分析中比較不同語言程度受試者的作答表現。在施測流程上，所有受試者皆需完成 AI 生成試題與官方模擬試題兩組測驗。為避免一次施測造成疲勞影響或作答策略干擾，本研究安排分週進行：第一週施測官方試題，隔週再進行 AI 試題。施測皆於相同場地、同一時間進行，以維持測驗環境的一致性。最後，在資料分析階段，將針對受試者在兩份試題中的成績表現、作答正確率與題目難易度等指標進行比較，以評估 AI 試題的可行性與應用潛力。

圖3-1

研究流程



參、 研究問題

- 一、同一受試者在 AI 試題與官方試題的測驗成績是否存在差異？
- 二、AI 試題難易度是否符合 TOCFL 入門基礎級的難度？
- 三、AI 試題能否有效區分不同程度考生？

第二節 研究對象

壹、 受試者的語言學習背景

本研究的受試者為 40 位新北市某科大的越南籍華語學習者，包含大學一、到四年級的學生。受試者的中文學習時間不一，部分受試者的學習時間少於一年，而部分則已學習三至四年。此外，所有受試者皆參加過 TOCFL 適性測驗，並已通過不同等級的閱讀與聽力測驗，其通過的測驗成績範圍涵蓋 CEFR A1 至 B2 級（如表 3-1），顯示受試者具備一定的華語基礎，應能完整進行本施測。

表3-1

受試者 TOCFL 適性測驗結果統計

聽力成績級別	通過人數	閱讀成績級別	通過人數
A1	6	A1	4
A2	18	A2	26
B1	15	B1	10
B2	1	B2	0

貳、 研究對象的選取標準與排除條件

本研究選取曾參與過 TOCFL 測驗並具備正式測驗成績的華語學習者作為受試者，以確保其語言能力與研究測驗內容具有對應關聯。施測內容包括 AI 試題與人工試題的閱讀測驗與聽力測驗，共計四份題本，受試者須於施測期間完整完成全部測驗。

本次原定招募 40 位受試者，惟其中一位學生僅完成部分測驗，缺考其中一項，導致無法進行完整的題本比較分析。為維持數據分析的公平性與

一致性，故將該位受試者的資料自樣本中剔除。最終納入分析者共 39 人，樣本流失率為 2.5%。

第三節 TOCFL 題本分析

壹、 TOCFL 入門基礎級測驗題型說明


為確保後續 AI 生成試題與 TOCFL 官方人工試題的測驗題型相同，接下來本研究將根據 TOCFL 於官方網站所提供的模擬題本，分析其聽力測驗與閱讀測驗之題型，以作為後續試題生成的參考依據。

一、聽力測驗


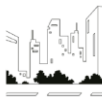

為了更清楚瞭解 TOCFL 測驗之命題內容，本研究於表 3-2 整理聽力測驗題本中所涵蓋的各類題型，並簡要說明其作答方式與題目所使用的語言特徵，同時從 TOCFL 入門基礎級模擬試題中擷取各題型之範例題，作為輔助說明之用。

表3-2

聽力測驗題型說明²

題型	測驗方式	題目語言	題目範例
看圖回答	題本上每題有一張圖片，考生聽到一個問題以及三個選項後，選出符合圖片內容的答案。	題目和選項皆為一句約 10 字左右的句子，問題大多問時間、地點、動作及事件等。	<p>9.  這位先生正在做什麼？</p> <p>(A) 他正在買菜 (B) 他正在煮飯 (C) 他正在洗衣服</p>

² 表 3-2 中範例題皆選自《TOCFL 華語文能力測驗模擬試題》(入門基礎級)，試題來源為教育部國家華語測驗推動工作小組公開提供之模擬題本，擷取自不同回版本，僅作為題型分析與研究說明之用。試題可見：[https://tocfl.edu.tw/tocfl/index.php/exam/test/page/1?pressBtn=\(%E9%A1%8C%E6%9C%AC\)](https://tocfl.edu.tw/tocfl/index.php/exam/test/page/1?pressBtn=(%E9%A1%8C%E6%9C%AC))

<p>問答理解</p>	<p>題本上每題有三張圖片，考生聽到一段問答以後，根據對話內容，選出符合這段對話的圖片。</p>	<p>問答每句字數落在 10 到 30 字左右。問答主題多元，如購物、旅行、問路等。</p>	<p>29. (A)  (B)  (C) </p>																																																						
<p>對話理解</p>	<p>題本上每題有三張圖片，考生會聽到一段四句的對話以及一句問題，考生需根據對話內容和問題，選出符合的圖片。</p>	<p>對話每句由兩到三個子句構成，字數至多不超過 30 字。主題為日常對話，如家人、職場、朋友間的互動。問題多為統整對話重點資訊。</p>	<p>29. 男 這個週末，你想去哪兒玩？ 女 天氣那麼熱，我想去海邊走走。</p> <p>43. 男 您好，請問有什麼事嗎？ 女 你好，這張紙上寫著現在學游泳，比較便宜。 男 是的，星期一到星期五的早上比較便宜。但是晚上來學，就和平常一樣了。 女 我知道了，我可以白天來上課。 男 請問什麼時候學游泳比較便宜？</p> <p>43. (A) <table border="1" data-bbox="922 741 1043 837"><tr><th>星期</th><th>一</th><th>二</th><th>三</th><th>四</th><th>五</th></tr><tr><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td></tr><tr><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td></tr></table> (B) <table border="1" data-bbox="1075 741 1197 837"><tr><th>星期</th><th>一</th><th>二</th><th>三</th><th>四</th><th>五</th></tr><tr><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td></tr><tr><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td></tr></table> (C) <table border="1" data-bbox="1219 741 1340 837"><tr><th>星期</th><th>一</th><th>二</th><th>三</th><th>四</th><th>五</th></tr><tr><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td><td>☀️</td></tr><tr><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td><td>🌙</td></tr></table></p>	星期	一	二	三	四	五	☀️	☀️	☀️	☀️	☀️	☀️	🌙	🌙	🌙	🌙	🌙	🌙	星期	一	二	三	四	五	☀️	☀️	☀️	☀️	☀️	☀️	🌙	🌙	🌙	🌙	🌙	🌙	星期	一	二	三	四	五	☀️	☀️	☀️	☀️	☀️	☀️	🌙	🌙	🌙	🌙	🌙	🌙
星期	一	二	三	四	五																																																				
☀️	☀️	☀️	☀️	☀️	☀️																																																				
🌙	🌙	🌙	🌙	🌙	🌙																																																				
星期	一	二	三	四	五																																																				
☀️	☀️	☀️	☀️	☀️	☀️																																																				
🌙	🌙	🌙	🌙	🌙	🌙																																																				
星期	一	二	三	四	五																																																				
☀️	☀️	☀️	☀️	☀️	☀️																																																				
🌙	🌙	🌙	🌙	🌙	🌙																																																				
<p>對話</p>	<p>題本上每題有四個選項敘述，考生聽到一段四句對話和一個問題後，根據對話和問題選出相符的選項。</p>	<p>對話每句由兩到三個子句構成，字數至多不超過 30 字。主題為日常對話，如家人、職場、朋友間的互動。問題多為統整對話內容。</p>	<p>49. 男 請進，有什麼事嗎？ 女 老闆，我最近總是覺得很累、很不舒服。我想請假。 男 妳想請多久呢？ 女 如果可以的話，我想請一個月；但是如果不行，兩個星期也可以。 男 這位小姐希望老闆答應她什麼事？</p> <p>49. (A) 讓她休息一個月 (B) 讓她的工作少一點 (C) 兩星期後，可以休息幾天 (D) 一個月後，讓她離開公司</p>																																																						

二、閱讀測驗

接下來，閱讀測驗亦如同聽力測驗之分析，於表 3-2 呈現閱讀測驗題本中所涵蓋的各類題型、其作答方式與題目所使用的語言特徵，同時從 TOCFL 模擬試題中擷取各題型之範例題，作為輔助說明之用。

表3-3

閱讀測驗題型說明³

題型	測驗方式	題目語言	題目範例
單句理解	題本上有一個句子和三個圖片選項，考生看完題目句子後，需從圖片選項中選出相符的圖片選項。	句子約 10 字左右，句子內容皆為人物在某地點做某事的日常事件。	<p>9. 張先生開車的時候喜歡聽音樂。</p> <p>(A)  (B)  (C) </p>
看圖釋義	題本上有一張圖片和三個選項句子，考生需按照圖片，選出符合的句子選項。	選項句子皆約為 10 字左右，且都和圖片有關，只是改變了細節資訊。	<p>20. </p> <p>(A) 小吃店九點以前休息。 (B) 小吃店十二點開始休息。 (C) 十二點以後可以去小吃店吃飯。</p>
選詞填空	題組有一張情境圖，和五題句子，每個句子有一個空格，和三個選項。考生需根據圖片語境和句子內容，選出合適的答案填充空句子。	五題皆為一個約 15 字左右的句子，都和圖片情境有關，有些是從圖片得知答案，有些是固定句型語法題（如：要是...就...）	<p></p> <p>31. _____ 著眼鏡的小女孩在看書。 (A) 穿 (B) 帶 (C) 戴</p> <p>32. 她一邊看書，一邊 _____ 孩子吃麵。 (A) 帶 (B) 用 (C) 找</p> <p>33. 那個小女孩 _____ 有一隻狗。 (A) 旁邊 (B) 前邊 (C) 後邊</p> <p>34. 那隻狗 _____ 睡覺。 (A) 在 (B) 要 (C) 是</p> <p>35. 小女孩 _____ 小狗是好朋友。 (A) 有 (B) 跟 (C) 一起</p>

(續下頁)

³ 表 3-3 中範例題皆選自《TOCFL 華語文能力測驗模擬試題》(入門基礎級)，試題來源為教育部國家華語測驗推動工作小組公開提供之模擬題本，擷取自不同回版本，僅作為題型分析與研究說明之用。試題可見：[https://tocfl.edu.tw/tocfl/index.php/exam/test/page/1?pressBtn=\(%E9%A1%8C%E6%9C%AC\)](https://tocfl.edu.tw/tocfl/index.php/exam/test/page/1?pressBtn=(%E9%A1%8C%E6%9C%AC))

<p>完成段落 一篇短文為一題組，短文中有一個空格，一共有六個可能的選項，根據文章的上下文，選出合適的答案完成短文。</p>	<p>短文約 100 到 200 字，主題皆為日常生活，如：生病、校園生活、旅行等。</p>	<p>昨天晚上我覺得很不舒服，____(41)____，所以很早就睡覺了。今天早上起來，____(42)____。我去看病，醫生說我感冒了，給了我一些藥，____(43)____要多休息，多喝水，才會快點好。這幾天的天氣一會兒熱，一會兒冷，____(44)____。我要____(45)____，不要再感冒了。</p> <p>(A) 通告我 (B) 頭有點兒痛 (C) 很容易生病 (D) 覺得很舒服 (E) 更不舒服了 (F) 多注意自己的身體</p>
<p>閱讀理解 每題有一篇短文和一個問題，考生閱讀完短文以後，根據問題，選出符合短文的答案。</p>	<p>每篇短文字數約為 120 字到 200 字。每篇短文有一個問題，問題常為統整短文主旨，或是提取短文資訊。</p>	<p>(一) 曾有一項調查發現，很多員工生病的時候不敢請假，因為他們擔心老闆會不高興，覺得他們沒有責任感。有人認為，員工會這麼想是公司的責任。一個好的公司應該能照顧員工，而不是讓他們拿健康去換錢。因此，讓員工有幸福感，應該是未來企業努力的方向。</p> <p>46. 這篇文章說了什麼內容？ (A) 老闆應該給員工多一點兒假 (B) 常關心別人的人更有責任感 (C) 對公司有意見要勇敢說出來 (D) 照顧身體比認真工作更重要</p>

貳、 試題分析標準：PIRLS 四層次提問法

國際閱讀素養研究 (Progress in International Reading Literacy Study, 簡稱 PIRLS) 由國際教育成就評估協會 (International Association for the Evaluation of Educational Achievement, IEA) 主導，旨在測量全球各國四年級學生的閱讀理解能力。PIRLS 根據閱讀理解的認知歷程，將題目分類為四個層次 (Mullis et al., 2017)，從高到低分別為：直接提取 (Focus on and Retrieve Explicitly Stated Information)、直接推論 (Making straightforward inferences)、詮釋、整合觀點和訊息 (Interpreting, Integrating, and Evaluating) 以及評估、批判內容 (Evaluating and Critiquing Content)，以下簡要說明各層次內涵。

- 一、直接提取 (Focus on and Retrieve Explicitly Stated Information)：這類題目要求考生從文本中找到直接陳述的信息，或根據明顯的線索做出簡單的推論。通常可透過關鍵詞比對來找到答案，例如「根據文章內容，誰去了市場？」。

- 二、直接推論 (Making straightforward inferences)：這類題目涉及基本的推理能力，考生需根據文本中提供的資訊，對隱含的內容進行推測。例如「從文章中得知，主角對這次旅行的感受如何？」。
- 三、詮釋、整合觀點和訊息 (Interpreting, Integrating, and Evaluating)：這類題目為較高層次題目，要求考生不僅理解字面意義，還需整合多個段落的資訊，或分析作者的寫作意圖。例如「作者對於這個議題的整題評價為何？」。
- 四、評估、批判內容 (Evaluating and Critiquing Content)：在最高層次的題目中，考生需評估文本的品質、結構或表達方式，並根據自身經驗與文本內容進行批判性思考。例如「你是否同意作者的立場？為什麼？」。

PIRLS 的四層次提問法涵蓋從基礎理解到高層次分析的閱讀表現，可有效評估學生的閱讀素養與推理能力，該架構已廣泛應用於閱讀能力測驗與試題編寫領域。基於此，本研究將以 PIRLS 的四層次提問為依據，分析 TOCFL 入門基礎級試題的提問類型，以評估該等級的題本提問層次與出題方向，並作為 AI 生成試題的設計參考。

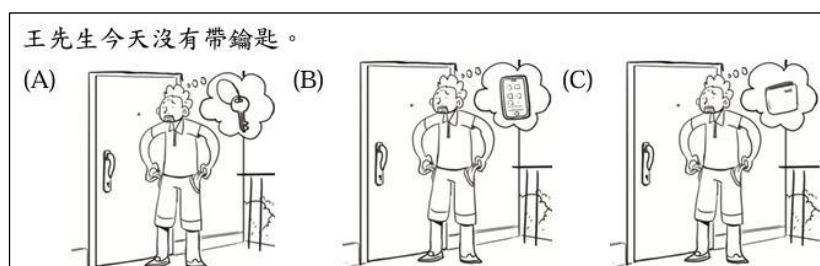
參、 試題分析說明

本研究參照 PIRLS 的四層次提問法，對 TOCFL 試題進行分析，以判定各題所屬的提問層次，以下先說明各提問層次之題目說明與 TOCFL 中的範例題目。

- 一、直接提取：考生僅需掌握文本中的關鍵字詞，即可回答出正確答案，不需經過換句話說、上下文判斷等基本推論。範例題如圖 3-2，考生僅需掌握該題的「鑰匙」，就可知道答案應為選項 A。

圖3-2


直接提取範例題



二、直接推論：考生需掌握文本或對話中的關鍵資訊，並透過基本推理來判斷正確答案，例如透過換句話說、圖表與文字對照等方式進行理解。範例題如圖 3-3，該題的文本為租房廣告，考生需根據選項內容，從文本中提取關鍵資訊，再進行基本推論以確認正確與否。例如，選項 B 指出「只有一個房間」，考生需對應廣告中的第二點「兩房」，由此推斷該選項為錯誤描述。

圖3-3

直接推論範例題

<p>公寓出租</p> <ul style="list-style-type: none">◎ 房租 16000 元◎ 兩房、一客廳、一浴室◎ 有家具、可以上網◎ 近立山小學，走路十分鐘 <p>請找王先生： 0911-234567</p> 	<p>A) 這間公寓在學校附近。</p> <p>B) 公寓裡面只有一個房間。</p> <p>C) 想租這間公寓的人上網找陳先生。</p>
--	--

三、詮釋、整合觀點和訊息：這類提問多應用於較長的文本，要求考生整合三句以上的資訊，並透過上下文的邏輯關聯，推斷文本中人物的言外之意。範例題如圖 3-4，女方並未直接表達「這裡很吵」，但透過上下文對話的邏輯推理，考生需理解女方的意思是「因為人變多了，所以現在變吵了」，因此正確答案為選項 A。

圖3-4

詮釋、整合觀點和訊息範例題

男	好久沒來這裡了，變了好多喔。	
女	以前這裡是漂亮的風景，現在這裡到處都是人。	(A) 這裡很吵
男	人多不好嗎？	(B) 這裡人不多
女	我喜歡安靜的環境。	(C) 這裡的環境很好
男	這位小姐最後的意思是什麼？	(D) 這裡的風景很美

四、評估、批判內容：這類提問側重於考生的批判性思考能力，主要應用於寫作或口語測驗。在本研究的核心範疇——閱讀與聽力測驗中，並未涉及此層次，因此接下來將不予探討。

肆、 試題分析結果

一、聽力測驗 PIRLS 提問類型分佈與數據統計

從表 3-4 可見，聽力測驗入門基礎級的題目大多屬於提問層次最低的「直接提取」，考生在作答時主要依賴對題目文字、圖片與測驗音檔的理解，無需額外推論，即可得出正確答案。此外，表 3-4 也顯示了 TOCFL 官方提供的模擬試題從題本一到題本五，難度呈現逐步提升的趨勢。然而，即使是難度相對較高的題本五，其中屬於「詮釋、整合觀點和訊息」等較高層次的題目比例仍不足 20%。由此可見，入門基礎級測驗主要關注考生對基礎日常會話的理解能力，而較少涉及更高層次的分析與整合。

表3-4

聽力測驗 PIRLS 提問類型分佈與數據統計

題本	直接提取	直接推論	詮釋、整合觀點 和訊息	總題數
題本一	41 題 (82%)	5 題 (10%)	4 題 (8%)	50 題 (100%)
題本二	38 題 (76%)	8 題 (16%)	4 題 (8%)	50 題 (100%)
題本三	29 題 (58%)	14 題 (28%)	7 題 (14%)	50 題 (100%)
題本四	26 題 (52%)	15 題 (30%)	9 題 (18%)	50 題 (100%)
題本五	27 題 (54%)	14 題 (28%)	9 題 (18%)	50 題 (100%)

二、聽力測驗各題型的四層次提問類型分佈（五個題本，共 250 題）

根據表 3-5，聽力測驗四個部分的題目在提問層次上皆呈現逐步提高的趨勢。第一部分「看圖回答」中幾乎皆為最低層次（直接提取）的題目，考生只需從圖片中擷取關鍵資訊即可作答；第二部分「問答理解」中仍以直接提取為主，但比例略低於第一部分，並開始出現少量第二層次（直接推論）的題目，即要求考生在理解內容的基礎上進行簡單推理。隨著測驗難度的提升，第三部分「對話理解」中，約有半數的題目屬於第二層次（直接推論），考生需根據文本內容進行基本的邏輯推斷，同時，在此部分也開始出現較高層次的題目，考驗考生的語境理解能力。到第四部分「對話」，測驗則以第三層次（詮釋、整合觀點與訊息）的題目為主，要求考生綜合分析對話內容，並從整體脈絡推斷出隱含資訊，而第二層次（直接推論）的題目則作為輔助，僅有少量第一層次（直接提取）的題目。

表3-5

聽力測驗各題型的 PIRLS 提問類型分佈

題型	直接提取	直接推論	詮釋、整合觀點和訊息
看圖回答 (共 80 題)	79 題 (98.75%)	1 題 (1.25%)	0 題 (0%)
問答理解 (共 78 題)	54 題 (\approx 69.23%)	24 題 (\approx 30.77%)	0 題 (0%)
對話理解 (共 52 題)	20 題 (\approx 38.5%)	19 題 (\approx 36.5%)	13 題 (25%)
對話 (共 40 題)	4 題 (10%)	16 題 (40%)	20 題 (50%)

整體而言，聽力測驗的題目層次隨著測驗部分的推進而逐步提升，這樣的設計不僅能有效區分不同程度的考生，也能透過多元題型，評估考生在不同層次上的語言理解與應用能力。

三、閱讀測驗四層次提問類型分佈與數據統計

從表 3-6 可以發現，各題本的試題主要以「直接提取」和「直接推論」較低層次的提問為主；「詮釋、整合觀點和訊息」的試題比例相對較少，顯示大部分題目都是測驗考生對於文本的基本理解能力，較少涉及更高層次的分析與評估，並且不強調考生的批判性思考能力，而是以基礎理解和簡單推論為主要測驗重點。

表3-6

閱讀測驗 PIRLS 提問類型分佈與數據統計

題本	直接提取	直接推論	詮釋、整合觀點和訊息	總題數
題本一	29 題 (58%)	15 題 (30%)	6 題 (12%)	50 題 (100%)

(續下頁)

題本二	28 題 (56%)	17 題 (34%)	5 題 (10%)	50 題 (100%)
題本三	26 題 (52%)	18 題 (36%)	6 題 (12%)	50 題 (100%)
題本四	32 題 (64%)	14 題 (28%)	4 題 (8%)	50 題 (100%)
題本五	31 題 (62%)	16 題 (32%)	3 題 (6%)	50 題 (100%)

四、閱讀測驗各題型的 PIRLS 提問類型分佈（五個題本，共 250 題）

根據表 3-7，閱讀測驗在提問層次上亦呈現逐步提升的趨勢。第一部分「單句理解」主要測驗考生對簡單句子的理解，題目大多屬於最低層次（直接提取）。第二部分「看圖釋義」的題目仍以直接提取為主，但部分題目開始要求考生結合圖像與文字內容進行基本的理解與判斷，開始出現少量第二層次（直接推論）的提問。到了第三部分「選詞填空」的題目類型大多數題目仍然是直接提取，但部分填空題屬於第二層次（直接推論），即需根據前後文推理最合適的詞語。第四部分「完成段落」略提升難度，部分題目要求考生理解整體段落的結構與邏輯關係，使得直接推論的比例略為提升。第五部分「閱讀理解」，題目層次達到最高，其中約有半數題目為第二層次（直接推論），要求考生在理解文本內容的基礎上，進行基本的邏輯推斷；詮釋、整合觀點與訊息的題目也比例顯著增加，測驗考生綜合分析文章內容的能力。

表3-7

閱讀測驗各題型的 PIRLS 提問類型分佈

題型	直接提取	直接推論	詮釋、整合觀點和訊息
單句理解 (共 75 題)	52 題 ($\approx 69.3\%$)	23 題 ($\approx 30.7\%$)	0 題 (0%)
看圖釋義 (共 75 題)	38 題 (50.7%)	26 題 ($\approx 34.7\%$)	11 題 ($\approx 14.6\%$)
選詞填空 (共 40 題)	31 題 (77.5%)	9 題 (22.5%)	0 題 (0%)
完成段落 (共 35 題)	21 題 (60%)	12 題 ($\approx 34.3\%$)	2 題 ($\approx 5.7\%$)
閱讀理解 (共 25 題)	4 題 (16%)	10 題 (40%)	11 題 (44%)

整體而言，閱讀測驗的題目層次隨著測驗部分的推進而逐步提升，這樣的設計有助於區分不同程度的考生，並透過多層次的提問方式，全面評估考生在不同層次上的閱讀理解與應用能力。

五、聽力與閱讀測驗通過標準與 PIRLS 提問類型表現的關聯

根據 TOCFL 在官方網站所提供的各等級通過分數門檻表可知，聽力測驗滿分為 80 分，測驗分數若達到 41 分以上，即為入門級 (A1)；分數若達 57 分以上，則為基礎級 (A2)。該測驗採用量尺分數 (Scale Scores) 及標準設定 (Standard Setting) 作為通過等級判定的依據。受試者的最終得分與作答正確數呈正相關，即答對題數越多，分數越高；然而，由於每次測驗的難度可能有所不同，最終得分亦會隨之變動。因此本研究再以 TOCFL 所提供的入門基礎級模擬試題的聽力與閱讀測驗成績對照表為參考，整理各等級的通過分數與對應的答對題數 (如表 3-8)。

表3-8

聽力與閱讀測驗通過分數與題數對照表

測驗項目	等級	通過分數	分數對應之答對題數
聽力測驗	入門級	41	25
	基礎級	57	38
閱讀測驗	入門級	42	24
	基礎級	60	39

根據表 3-8，考生在共 50 題的聽力測驗中需答對 38 題才能達到基礎級 (A2)，相當於 76% 的正答率；然而，從表 3-4 的數據可見，難度較低的試卷 (如題本一) 中，約有 80% 的試題即屬於最低層次的「直接提取」，表示考生僅需正確理解題目中的關鍵資訊，即可正確作答並達標。而在難度較高的試卷 (如題本五)，其最低層次的試題比例約為 50%，而第二層次 (直接推論) 的試題約佔 30%。因此即便試題較難，考生若能掌握基本的直接提取即推論能力，仍有較高機率達到基礎級標準。同樣的趨勢也出現在閱讀測驗中。根據表 3-8 的數據，考生需在閱讀測驗中答對 39 題才能達到基礎級 (A2)，相當於 78% 的正答率；然而，從表 3-6 可見，五個題本中約有八成的題目屬於較低層次 (直接提取與直接推論)。這表示考生只要能掌握這些較低層次的考題，即可順利達到基礎級的通過標準。由此可見，入門基礎級測驗主要考察考生對日常語言的理解能力，而非較高層次的分析與整合能力。

伍、 AI 試題設計的參考依據

本研究將上述分析的試題架構作為 AI 出題的參考依據，以確保 AI 試題在題目架構上與 TOCFL 試題保持一致。以下將說明試題的題型分類、考察重點及難易度分佈，確保 AI 生成試題能夠符合 TOCFL 的測驗標準，並維持測驗的有效性與公平性。

一、閱讀測驗

- (一) 題型：與 TOCFL 試題架構相同，將題本分為五個部分，分別為「單句理解」、「看圖釋義」、「選詞填空」、「完成段落」、「閱讀理解」。
- (二) 題數：TOCFL 五個題本中，第三部分和第四部份的題數在不同題本間偶有不同，因此本研究僅以題本一為標準，方便後續試題的分析。按照題本一，題本共有 50 題，其中第一、二部分各有 15 題；第三部分為 10 題；第四、五部分各為 5 題。
- (三) PIRLS 提問比例：本研究按照表 3-5 的統計數據，設定各部分的四層次提問比例如下
1. 第一部分（單句理解）：僅使用「直接提取」和「直接推論」提問，並以「直接提取」的提問為多數，「直接推論」的提問為輔，兩者比例約為 7:3。
 2. 第二部分（看圖釋義）：「直接提取」、「直接推論」和「詮釋、整合觀點和訊息」三個層次皆使用，三者佔比約為 5:4:1。
 3. 第三部分（選詞填空）：僅使用「直接提取」和「直接推論」提問，兩者比例約為 7:3。
 4. 第四部分（完成段落）：「直接提取」、「直接推論」和「詮釋、整合觀點和訊息」三個層次皆使用，三者佔比約為 6:3:1。
 5. 第五部分（閱讀理解）：「直接提取」、「直接推論」和「詮釋、整合觀點和訊息」三個層次皆使用，三者佔比約為 2:4:4。

二、聽力測驗

- (一) 題型：與 TOCFL 試題架構相同，將題本分為四個部分，分別為「看圖回答」、「問答理解」、「對話理解」、「對話」。
- (二) 題數：TOCFL 五個題本中，五個部分的題數在不同題本間偶

有不同，因此本研究同樣以題本一為標準，方便後續試題的分析。按照題本一，題本共有 50 題，其中第一部分共 25 題；第二部分共 15 題；第三部分及第四部分各 5 題。

(三) PIRLS 提問比例：本研究按照表 3-2 的統計數據，設定各部分的四層次提問比例如下

1. 第一部分（看圖回答）：僅使用「直接提取」和「直接推論」提問，並以「直接提取」的提問為大多數，「直接推論」的提問為少數，兩者比例約為 9:1。
2. 第二部分（問答理解）：僅使用「直接提取」和「直接推論」提問，兩者佔比約為 7:3。
3. 第三部分（對話理解）：使用「直接提取」和「直接推論」和「詮釋、整合觀點和訊息」三層次，三者比例約為 4:4:2。
4. 第四部分（對話）：使用「直接提取」、「直接推論」和「詮釋、整合觀點和訊息」三個層次，三者佔比約為 1:4:5。

第四節 AI 試題編寫模型訓練過程

壹、 模型介紹

本研究的試題生成模型為 OpenAI 開發的人工智慧聊天生成預訓練轉換器（Chat Generative Pre-trained Transformer，以下簡稱 ChatGPT），為確保試題生成的流暢度和正確度，本研究使用其付費的 ChatGPT Plus 版本。因 ChatGPT 已具備一定的語言理解與生成能力，本研究基於 ChatGPT 既有的語料庫與預訓練模型，再透過提示工程（Prompt engineering），訓練 AI 產出符合 TOCFL 試題架構及命題者所需的測驗內容，而非重新訓練模型或透過微調（fine-tuning）來獲得特定的功能。這樣的方式除了生成過程中

可快速調整提示詞，確保試題產出的靈活性，未來也方便再擴展到不同測驗等級或試題；然而，未經過微調的 ChatGPT 仍有一定限制，例如生成的語言可能超出測驗等級，或試題題型風格與 TOCFL 人工試題存在差異等狀況。因此，本研究將在後續章節探討如何確保 AI 試題的品質與可用性。

貳、 訓練語料與資料處理

一、 數據資料整理

- (一) 出題所用詞彙：本研究使用「華語八千詞表」，篩選其中準備級一級(N1)、準備級二級(N2)、入門級(A1)及基礎級(A2)的詞彙，並整理成供 AI 出題參考的 PDF 表格。詞彙表格內容涵蓋主題領域、詞彙、漢語拼音及詞類，以確保 AI 生成試題時能符合華語出題詞彙範圍與需求。
- (二) 各部分題型樣本及四層次提問分析：本研究採用 TOCFL 官方網站提供之模擬試題，將各題型整理為獨立的 PDF 表格。表格內容包含題目文本、選項文本、聽力音檔逐字稿，以及對應之四層次提問類型，以作為 AI 題目生成的結構參照。為使生成式 AI 在模擬學習過程中具備充足語料依據，各題型皆盡可能蒐集約 100 題作為參考樣本，藉此提升 AI 生成題目的題型一致性與提問層次準確性，確保其符合華語測驗之命題邏輯與難度標準。
- (三) 句型參考：為確保 AI 生成題目的語言自然且具多樣性，本研究另從 TOCFL 模擬試題中擷取各類句型與提問，並整理成句型參考的 PDF 檔案，以作為 AI 出題時的參考依據。

二、 建立出題 GPT：聽力測驗

- (一) 名稱：TOCFL A2 聽力試題編寫

- (二) 類別：教育
- (三) 功能：DALL·E 圖像、網頁搜尋、數據分析
- (四) 知識庫檔案：「出題所用詞彙」、「看圖釋義」、「問答理解」、「對話理解」、「對話」、「句型參考」
- (五) 後台指令：選項和答案皆只能使用「出題所用詞彙」檔案中的詞彙。按照知識庫「看圖釋義」、「問答理解」、「對話理解」、「對話」各檔案的出題方式，給予符合使用者需求的題目、選項和答案。出題語言請參考「句型參考」檔案。選項若為三張圖片，則分成三張圖片給使用者，不可放在同一張圖片。正確答案只有一個，不可以出沒有正確答案的問題。

三、建立出題 GPT：閱讀測驗

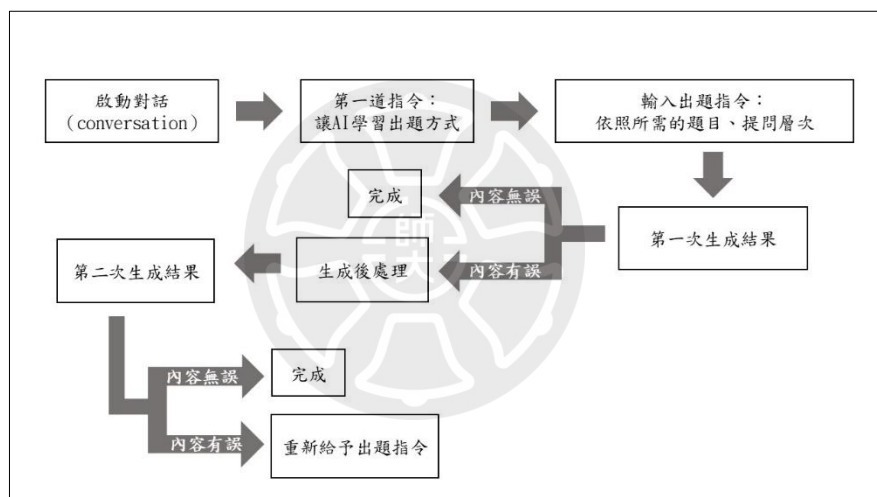
- (一) 名稱：TOCFL A2 閱讀試題編寫
- (二) 類別：教育
- (三) 功能：DALL·E 圖像、網頁搜尋、數據分析
- (四) 知識庫檔案：「出題所用詞彙」、「看圖釋義」、「問答理解」、「對話理解」、「對話」、「句型參考」
- (五) 後台指令：選項和答案皆只能使用「出題所用詞彙」檔案中的詞彙。按照知識庫「單句理解」、「看圖釋義」、「選句填空」、「選詞填空」、「閱讀理解」各檔案的出題方式，給予符合使用者需求的題目、選項和答案。出題語言請參考「句型參考」檔案。若為三張圖片，則分成三張圖片給使用者，不可放在同一張圖片。正確答案只有一個，不可以出沒有正確答案的問題。

參、 AI 試題生成方法說明：以聽力測驗為例

AI 試題的生成流程如圖 3-5 所示，大致可分為六個步驟：啟動對話、給予第一道提示、輸入出題指令、獲得第一次生成結果、進行必要的生成後處理，及取得第二次生成結果。為了呈現 AI 出題的原始表現，同時亦為避免過多人工干預導致內容偏離初始生成邏輯，本研究僅在第一次生成後進行一次修改，若第二次生成結果仍存在明顯錯誤，則不再進行修正，而是重新開始出題流程。

圖3-5

AI 試題的生成流程圖



一、依照測驗題型啟動對話(Conversation):本研究依據 TOCFL 聽力測驗題型，將對話類型分為「看圖釋義」、「問答理解」、「對話理解」與「對話」四類，並針對各類對話輸入相應的需求指令，以確保 AI 生成題目時符合測驗規範。此外，研究過程中亦進行後續修改與調整，以提升試題的準確性與適切性。

二、開始生成：以下以「對話理解」的對話為例，說明生成過程。

(一) 為了讓 AI 學習出題方式，給予第一道指令。例如：「請查看知識庫中「對話理解」檔案，學習這類題型的出題方式：三張圖片

選項、一個 4 句的對話（一男一女）和一個問題。考生要按照聽到的對話和問題，選擇符合的圖片選項。再學習這類題型在 PIRLS 四層次的出題方式。」

(二) 按照所需的題型與提問層次，輸入出題指令。例如：「請給我一個"直接提取"的"對話理解"題目。請確保答案只有一個。主題：租房。請說明三張選項圖片描述。對話設計上，每題隨機讓不同的說話的人帶出正確答案，對話句子長一點，但依然嚴格按照"出題所用詞彙"出題，不可超出範圍。」

(三) 第一次生成結果

1. 對話：男：「我們去看了幾個房子，你覺得哪一個比較好？」
女：「我覺得那個有大陽台的房子很好，不但空間大，而且陽光很好。」男：「是啊，那個房子的客廳也很舒服。不過，我也覺得靠近公園的那間不錯，環境很安靜。」女：「我還是更喜歡陽台大的房子，我們可以在陽台放幾張椅子，喝茶聊天。」
2. 問題：女生更喜歡哪一個房子？
3. 圖片選項：A.大陽台的房子 B.靠近公園的房子 C.普通的公寓
4. 正確答案：A.大陽台的房子

(四) 生成後處理

1. 根據上述例子，「陽台」一詞在國教院的詞語分級標準檢索系統中為進階 4* 等級，超過出題等級範圍，因此需再做修正。
2. 修正指令：「"陽台"超出範圍，請修正對話」

(五) 第二次生成

1. 對話：女：「我們看了幾間房子，你覺得哪一間比較好？」男：
「我覺得有客廳大窗戶的房子不錯，光線很好，看起來很明



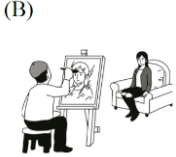

亮。」女：「嗯，那間的客廳確實很舒服。不過，我也喜歡離公車站近的房子，這樣上下班方便。」男：「也是，但我還是覺得窗戶大的房子住起來會更舒服。」

2. 問題：男生更喜歡哪一間房子？
3. 圖片選項：A.客廳有大窗戶的房子 B.靠近公園的房子 C.普通的房子
4. 正確答案：A.客廳有大窗戶的房子

肆、 AI 試題與人工試題比較

為了初步分析 AI 試題與人工試題的命題品質差異，接下來將選擇相同題型，且同一主題或同一層次提問的 AI 試題與人工試題，進一步比較與說明其異同。

表3-9 AI 試題與人工試題之比較

範例一		
題型說明	人工試題	AI 試題
1. 測驗：閱讀測驗。	題幹： 老王正畫著小天的臉。	題幹： 哥哥正在用筆寫字。
2. 題型：單句理解。	選項圖片： (A) 	選項圖片： (A) 
3. 四層次提問：直接提取。		
4. 主題：人正在做某事。	(B) 	(B) 

(C)





(C)



比較

1. 圖片：人工試題圖片簡潔，易於對比與判斷；AI 試題圖片畫面細節較多，有較多干擾資訊。
2. 題目：人工試題題幹與圖片高度對應，選項之間的差異清晰，能有效評估考生對描述內容的理解；AI 試題的題幹與圖片之間的對應度可能較低，部分選項與「用筆寫字」的行為無直接關聯，可能影響題目的測驗品質。

範例二


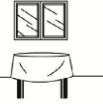




題型說明	人工試題	AI 試題
1. 測驗：閱讀測驗。 2. 題型：看圖釋義。 3. 四層次提問：直接推論。 主題：店家營業時間公告。	題幹圖片：  選項句子： (A) 小吃店九點以前休息。 (B) 小吃店十二點開始休息。 (C) 十二點以後可以去小吃店吃飯。	題幹圖片：  選項句子： (A) 這家店每天都從早上 8 點開始營業。 (B) 星期二不能去這家店。 (C) 這家店的營業時間是 24 小時。
比較	<ol style="list-style-type: none"> 1. 圖片：人工試題圖像簡單，重點資訊明確，考生不易受到干擾；AI 試題的圖像較為真實，但可能增加考生的認知負擔 2. 選項：人工試題的推論較為單一，AI 試題則增加了星期二公休的額外推論需求，考生需考慮不同時間條件的影響。 	

(續下頁)

範例三

題型說明	人工試題	AI 試題
<p>1. 測驗：閱讀測驗。</p> <p>2. 題型：看圖釋義。</p> <p>3. 四層次提問：詮釋、整合觀點及訊息。</p> <p>4. 主題：宣傳廣告。</p>	<p>題幹圖片：</p> <div style="border: 1px solid black; padding: 5px; text-align: center;"> <p>中天旅行社</p> <p>韓國的冬天又漂亮又好玩，讓我們一起去韓國去滑雪</p> <p>一個人只要19999元</p> <p>兩個人一起來，一個人可以便宜500元</p> <p>三個人一起來，一個人可以便宜1000元，再送你好喝的紅茶一盒</p> <p>有興趣的人，請打電話：02-2511-1687</p> </div> <p>選項句子：</p> <p>(A) 兩個人一起去，就有紅茶可以喝。</p> <p>(B) 一個人不到兩萬就可以去韓國旅行。</p> <p>(C) 三個人一起去的話，一共可以少給 1000 元。</p>	<p>題幹圖片：</p> <div style="border: 1px solid black; padding: 5px; text-align: center;"> <p>語言中心 中文課</p> <p>上課時間：每週 2 次，早上 9:00 到 11:00</p> <p>上課內容：拼音、對話、文化學習</p> <p>學費：一個人報名 5000 元。</p> <p>三個人一起報名，每個人可以便宜 200 元，還送學習本。</p> </div> <p>選項句子：</p> <p>(A) 要是三個人一起報名，就可以拿到免費的學習本。</p> <p>(B) 課都在下午，時間是兩點到四點。</p> <p>(C) 每個人都只要 3000 元就能報名這個中文課。</p>
比較	<p>人工試題的資訊量較大，考生需進行費用計算與條件推理，因此整體難度較高。AI 試題則更聚焦於資訊辨識，難度稍低，但同樣能測試基本的閱讀理解能力。</p>	





範例四

題型說明	人工試題	AI 試題
<p>1. 測驗：聽力測驗。</p> <p>2. 題型：問答理解。</p> <p>3. 四層次提問：直接提取。</p> <p>主題：家具擺放</p>	<p>聽力腳本：</p> <p>男：你覺得桌子放在哪兒比較好呢？</p> <p>女：放在窗戶前面吧！那兒比較亮。</p> <p>選項圖片：</p> <p>(A)</p>  <p>(B)</p>  <p>(C)</p> 	<p>聽力腳本：</p> <p>女：要把冰箱放在哪裡？</p> <p>男：先放在廚房的櫃子旁邊吧。</p> <p>選項圖片：</p> <p>(A)</p>  <p>(B)</p>  <p>(C)</p> 

比較

人工試題需要考生根據對話內容知道桌子的最佳擺放位置，題目設計較細緻，難度較高；而 AI 試題則是物品與房間的「直接匹配」，選項鑑別度明顯，考生較易判斷。

範例五

題型說明	人工試題	AI 試題
1. 測驗：聽力測驗。	聽力腳本：	聽力腳本：
2. 題型：對話理解。	男：請問十點二十分的	男：車站現在有幾班車？
3. 四層次提問：直接推論。	火車開了嗎？	女：有的車每小時一班，有的車每三十分鐘一班。
4. 主題：班次時間	女：現在都二十五分了，早就開了。 男：幾點還有車？ 女：二十分鐘以後就有車了。男：請問這位先生可以坐什麼時候的火車？	男：有一班車剛離開，現在已經三點了。 女：那下一班應該是半個鐘頭以後來。 男：請問下一班車什麼時候來？
	選項圖片： (A) 	選項圖片： (A)  (B)  (C) 

比較

人工試題的設計較多層次，考生需要理解對話內容、推算火車時間、選擇適合的班次。AI 試題的設計較簡單，考生只需掌握現在時間和距離下一班的時間後，即可推算出答案，題目難度較低。

(續下頁)

範例六

題型說明	人工試題	AI 試題
1. 測驗：聽力測驗。	對話腳本： 男：媽，你看見我的課本嗎？	對話腳本： 男：下週要考試了，你準備得怎麼樣？
2. 題型：對話。	女：書桌那裡沒有嗎？	女：我這幾天一直在看書，但題目有點多。
3. 四層次提問：詮釋、整合觀點及訊息。	男：找過了，都沒有。等一下就 要上課了，真糟糕。 女：不是放在書桌上，就是跟字典 放在一起，你去看看。 男：這位太太的意思是什麼？	男：那你覺得哪一個比較難？ 女：數學比較難，我覺得很不容易學。 男：別擔心，慢慢準備就好。 男：這位小姐的意思是什麼？
	選項： (A) 她要兒子再找找 (B) 課本不在書桌上 (C) 她把課本收起來了 (D) 書桌上放了本字典	選項： (A) 她覺得準備考試很簡單。 (B) 她需要更多時間準備數學考試。 (C) 她已經準備好所有的考試了。 (D) 她不喜歡考數學，所以沒準備。
比較	兩者都需透過對話，推敲其中一人的言下之意，但人工題目的選項干擾性較強；AI 則較容易透過關鍵字找到正確答案。	

伍、 模型訓練過程之困難與解決方法

綜上所述，AI 技術在試題編製方面已展現出一定的能力，能夠產出與人工試題品質相近的題目。然而，現階段 AI 技術仍存在一些限制，尚未達到完全精確的水準。此外，部分題目圖片仍無法透過現有 AI 技術生成，以下將進一步說明各類狀況與目前應對方法。

一、生成題目的語言與測驗程度不符

- (一) 狀況說明：儘管前期已整理「出題所用詞彙」檔案，作為 AI 生成試題的語料庫，但目前 AI 仍無法完全限定詞彙範圍，生成題目時仍可能出現不少超出範圍的詞彙。因此，在每次指令輸入時，需反覆強調詞彙範圍，以減少不符合測驗等級的用語。
- (二) 解決方法：在初步生成試題後，需經過多次的人工審核與修改，以確保詞彙難度符合測驗標準。本研究參考 TOCFL 試題編製流程，並運用國教院教材編輯輔助系統中的「分級標記」功能來檢測 AI 生成題目的用詞難度。具體而言，試題詞彙應主要落在等級 1~3，且確保等級 4 以上的詞彙使用比例與人工試題相當，再對於難度過高的超等詞，進行改寫，以確保試題內容符合測驗等級標準。

二、圖片生成的限制

- (一) 狀況說明：當前的 AI 技術已能根據指令生成相應的圖片，但在處理含有資訊文字的圖像時，仍存在技術上的限制，導致生成的內容可能不準確。在 TOCFL 考題中，常見的地圖、時間表、時刻表等需依賴圖像資訊解讀的題目，AI 在生成此類題型時，經常出現圖片資訊錯誤、不合理或文字難以辨識的情況，如表 3-10 範例說明。

表3-10

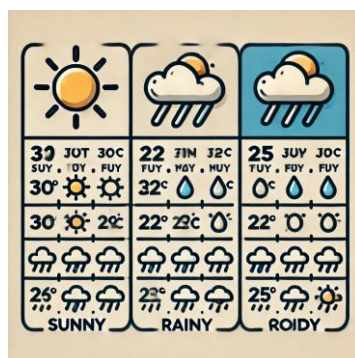
AI 試題生成的圖片問題

生成的題目與圖片描述	生成的圖片結果	問題
<p>題型：聽力—看圖釋義</p> <ul style="list-style-type: none"> ● 圖片描述：圖片中有一間教室，黑板上有一些中文字，桌子上有幾本書和兩支筆。 ● 問題：黑板上有什麼？ ● 選項： <ul style="list-style-type: none"> (A)漢字。 (B)圖片。 (C)空白。 		<p>黑板上的漢字為假字，無法辨識</p>
<p>題型：聽力—對話理解</p> <ul style="list-style-type: none"> ● 對話題幹： <ul style="list-style-type: none"> 女：車站現在有幾班車？ 男：有的車每小時一班，有的車每半小時一班。 女：有一班車剛離開，現在時間是三點。 男：那下一班應該是半個鐘頭以後來。 男：請問下一班車什麼時候來？ ● 選項圖片描述： <ul style="list-style-type: none"> (A)三點半：一個時鐘顯示三點半，時針指在3和4之間，分針指在6。 (B)四點整：一個時鐘顯示四點整，時針正對數字4，分針正對數字12。 (C)三點十五分：一個時鐘顯示三點十五分，時針指在3和4之 		<p>圖片的時間皆和其生成的選項敘述不符，也沒有正確答案，無法作答。</p>

間，分針指在 3。

題型：聽力－看圖釋義

- 圖片描述：有一張日曆，標記了三天的天氣和溫度：第一天是晴天，溫度 30°C；第二天是雨天，溫度 22°C；第三天是多雲，溫度 25°C。



生成的圖片資訊太多，無法清楚辨識。

- 問題：哪一個是對的？
- 選項：
 - (A) 第一天的溫度最高。
 - (B) 第二天的溫度最低。第三天的溫度和第一天一樣。

(二) 解決方法：此問題屬於現階段的技術限制，即使透過指令調整，仍無法生成理想的圖片。因此，本研究將依據 AI 生成的圖片描述，重新繪製符合內容的圖片，以確保 AI 試題的完整性與可用性。

第五節 測驗設計考題架構

壹、 測驗目標及內容

本測驗針對華語學習者的華語能力測驗，該測驗的等級為入門基礎級 (A1-A2)，分為閱讀測驗和聽力測驗，旨在測驗受試者的華語文能力程度。本測驗的試題編寫遵循 TOCFL 的試題編製說明，並運用國教院教材編輯輔助系統中的「分級標記」功能，以確保試題的語言難度符合 TBCL 第二級與第三級的標準，維持試題難度的合理性與一致性。

貳、 測驗題本架構及各題型四層次提問分佈

本測驗依據 TOCFL 模擬試題一題本一的考題架構進行編製，並參考

第三節第四部分所整理的四層次提問分佈比例，確保試題難度與題型設計符合標準。以下以表 3-11 及表 3-12 說明各測驗試題題目架構與四層次比例。

一、閱讀測驗：共 50 題

表3-11

閱讀測驗題本架構

題組部分	題型名稱	題數	提問層次分布
第一部分	單句理解	15 題	<ul style="list-style-type: none"> ● 11 題「直接提取」 ● 4 題「直接推論」
第二部分	看圖釋義	15 題	<ul style="list-style-type: none"> ● 7 題「直接提取」 ● 6 題「直接推論」 ● 2 題「詮釋、整合觀點和訊息」
第三部分	選詞填空	10 題	<ul style="list-style-type: none"> ● 7 題「直接提取」 ● 3 題「直接推論」
第四部分	完成段落	5 題	<ul style="list-style-type: none"> ● 3 題「直接提取」 ● 2 題「直接推論」 ● 1 題「詮釋、整合觀點和訊息」
第五部分	閱讀理解	5 題	<ul style="list-style-type: none"> ● 1 題「直接提取」 ● 2 題「直接推論」 ● 2 題「詮釋、整合觀點和訊息」

二、聽力測驗：共 50 題

表3-12

聽力測驗題本架構

題組部分	題型名稱	題數	提問層次分布
第一部分	看圖回答	25 題	● 22 題「直接提取」 3 題「直接推論」
第二部分	問答理解	15 題	● 10 題「直接提取」 5 題「直接推論」
第三部分	對話理解	10 題	● 2 題「直接提取」 ● 2 題「直接推論」 1 題「詮釋、整合觀點和訊息」
第四部分	對話	5 題	● 1 題「直接提取」 ● 2 題「直接推論」 2 題「詮釋、整合觀點和訊息」

第六節 資料收集與分析方法

除了文件分析法，本研究亦使用調查研究法，透過分析受試者在題本的分數統計以及作答行為分析，探討 AI 生成試題與人工試題的難易度、分數分佈與測驗表現的差異，並進一步分析考生的作答行為，以評估 AI 生成試題的測驗效度與適用性。

壹、 量化數據分析

一、分數比較

本研究將比較同一群受試者在 AI 試題與人工試題中的分數分佈，

觀察是否存在顯著的分數差異，並分析 AI 試題是否能夠有效評估學習者的華語能力。分數比較將採用描述性統計與配對樣本 t 檢定，以檢驗不同試題的測驗結果是否具有統計顯著性。

二、難易度分析

透過試題的平均得分與通過率，評估 AI 生成試題與人工試題之間的難度差異。本研究將以試題難度 (Difficulty Index, P 值)，計算各題目的答對比例，並比較兩組試題在不同提問層次上的難度趨勢。

貳、 作答行為分析

本研究將統計考生在 AI 試題與人工試題中的錯誤類型，觀察不同題型及提問類型（直接提取、直接推論、詮釋整合觀點和訊息）是否存在不同的作答錯誤趨勢，並分析 AI 試題是否存在較高的混淆率或誤導性選項。





第四章 研究結果

第一節 受試者表現概述

壹、 受試者基本資料

本研究的受試者華語程度涵蓋 A1 至 B2，依據 CEFR 的等級標準，A1 等級的學習者約具備 500 個詞彙量，而 B2 等級的學習者則可達 5000 個詞彙。此外，所有受試者皆已接受至少一學期以上的華語課程，具備一定的語言學習基礎。因此，在本研究所設計的測驗難度範圍內，受試者理論上應皆能達到通過標準，故將藉此分析 AI 試題的出題品質與難易度適配性。

為進一步探討不同程度學習者在 AI 試題與官方試題中的表現差異，本研究根據受試者的 TOCFL 適性測驗結果，將受試者分為三個能力組別：

- 一、 高能力組：適性測驗結果達 B1 或更高者。
- 二、 中能力組：適性測驗結果為 A2 者。
- 三、 低能力組：適性測驗結果為 A1 者。

透過此能力分組，本研究進一步分析不同程度學習者在 AI 生成試題與人工試題的作答表現，探討 AI 出題是否能有效區分考生能力，並評估其測驗品質。

貳、 施測參與情況與排除樣本說明

本研究的受試者共計 40 人，每位受試者需完成 AI 題本(閱讀與聽力)及人工題本(閱讀與聽力)，共計四個測驗題本。其中，1 名受試者因個人因素缺席其中一次測驗，為確保研究分析的公平性與數據完整性，本研究將該受試者排除於樣本之外。因此，最終回收並納入分析的有效樣本共計 39 份。

第二節 AI 題本與人工題本的比較

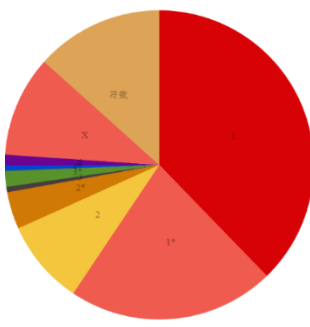
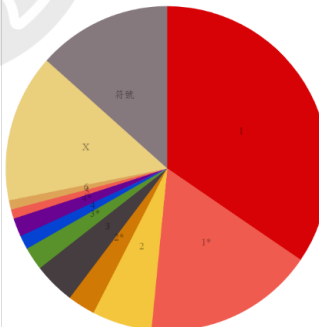
壹、題本的語言難度分析

本研究使用國家教育研究院的「教材編輯輔助系統」，並選擇中央研究院的「平衡語料庫」作為語料來源，進行題本語言的斷詞分析，以探討兩種題本在相同測驗等級下的用語選擇差異。以下分別說明兩題本在閱讀測驗與聽力測驗的斷詞。

根據表 4-1 的數據顯示，閱讀測驗中兩組題本的用詞難度大致相近，主要皆以等級 1 和 1* 的詞彙為主。其中 AI 題本的用語分布顯示，即便核心仍集中於初級詞彙，題本中仍有較多中高等級詞彙出現，可能反映出生成模型在語料選用上較不易精確控制語言難度。

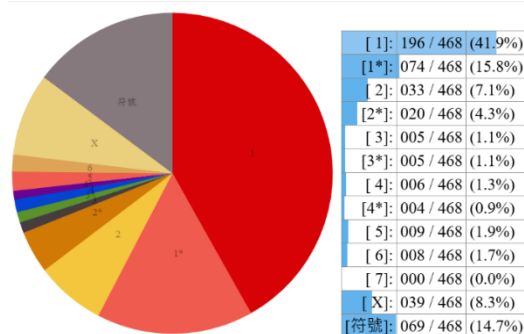
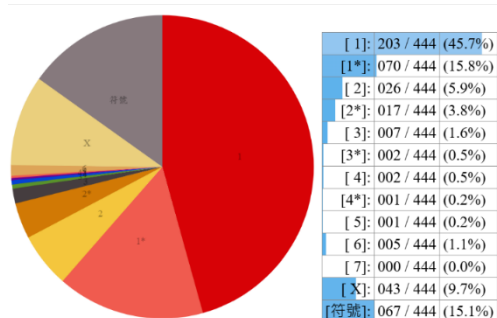
表4-1

閱讀測驗兩題本各部分的斷詞表

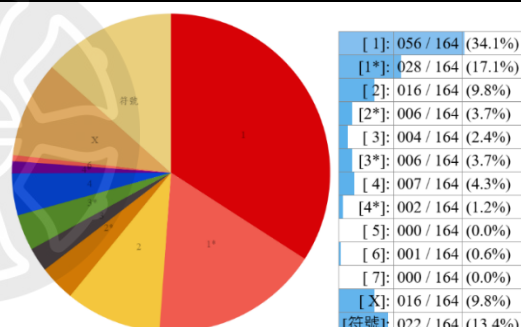
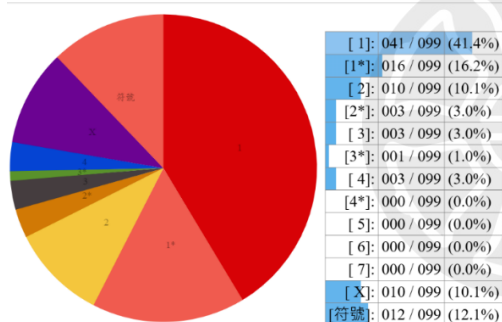
題型	人工題本	AI 題本																																																				
第一部分 單句 理解	 <table border="1"> <tr><td>[1]:</td><td>068 / 180 (37.8%)</td></tr> <tr><td>[1*]:</td><td>039 / 180 (21.7%)</td></tr> <tr><td>[2]:</td><td>016 / 180 (8.9%)</td></tr> <tr><td>[2*]:</td><td>007 / 180 (3.9%)</td></tr> <tr><td>[3]:</td><td>001 / 180 (0.6%)</td></tr> <tr><td>[3*]:</td><td>003 / 180 (1.7%)</td></tr> <tr><td>[4]:</td><td>000 / 180 (0.0%)</td></tr> <tr><td>[4*]:</td><td>001 / 180 (0.6%)</td></tr> <tr><td>[5]:</td><td>000 / 180 (0.0%)</td></tr> <tr><td>[6]:</td><td>002 / 180 (1.1%)</td></tr> <tr><td>[7]:</td><td>000 / 180 (0.0%)</td></tr> <tr><td>[X]:</td><td>019 / 180 (10.6%)</td></tr> <tr><td>[符號]:</td><td>024 / 180 (13.3%)</td></tr> </table>	[1]:	068 / 180 (37.8%)	[1*]:	039 / 180 (21.7%)	[2]:	016 / 180 (8.9%)	[2*]:	007 / 180 (3.9%)	[3]:	001 / 180 (0.6%)	[3*]:	003 / 180 (1.7%)	[4]:	000 / 180 (0.0%)	[4*]:	001 / 180 (0.6%)	[5]:	000 / 180 (0.0%)	[6]:	002 / 180 (1.1%)	[7]:	000 / 180 (0.0%)	[X]:	019 / 180 (10.6%)	[符號]:	024 / 180 (13.3%)	 <table border="1"> <tr><td>[1]:</td><td>075 / 217 (34.6%)</td></tr> <tr><td>[1*]:</td><td>037 / 217 (17.1%)</td></tr> <tr><td>[2]:</td><td>013 / 217 (6.0%)</td></tr> <tr><td>[2*]:</td><td>006 / 217 (2.8%)</td></tr> <tr><td>[3]:</td><td>009 / 217 (4.1%)</td></tr> <tr><td>[3*]:</td><td>005 / 217 (2.3%)</td></tr> <tr><td>[4]:</td><td>003 / 217 (1.4%)</td></tr> <tr><td>[4*]:</td><td>004 / 217 (1.8%)</td></tr> <tr><td>[5]:</td><td>002 / 217 (0.9%)</td></tr> <tr><td>[6]:</td><td>002 / 217 (0.9%)</td></tr> <tr><td>[7]:</td><td>000 / 217 (0.0%)</td></tr> <tr><td>[X]:</td><td>032 / 217 (14.7%)</td></tr> <tr><td>[符號]:</td><td>029 / 217 (13.4%)</td></tr> </table>	[1]:	075 / 217 (34.6%)	[1*]:	037 / 217 (17.1%)	[2]:	013 / 217 (6.0%)	[2*]:	006 / 217 (2.8%)	[3]:	009 / 217 (4.1%)	[3*]:	005 / 217 (2.3%)	[4]:	003 / 217 (1.4%)	[4*]:	004 / 217 (1.8%)	[5]:	002 / 217 (0.9%)	[6]:	002 / 217 (0.9%)	[7]:	000 / 217 (0.0%)	[X]:	032 / 217 (14.7%)	[符號]:	029 / 217 (13.4%)
[1]:	068 / 180 (37.8%)																																																					
[1*]:	039 / 180 (21.7%)																																																					
[2]:	016 / 180 (8.9%)																																																					
[2*]:	007 / 180 (3.9%)																																																					
[3]:	001 / 180 (0.6%)																																																					
[3*]:	003 / 180 (1.7%)																																																					
[4]:	000 / 180 (0.0%)																																																					
[4*]:	001 / 180 (0.6%)																																																					
[5]:	000 / 180 (0.0%)																																																					
[6]:	002 / 180 (1.1%)																																																					
[7]:	000 / 180 (0.0%)																																																					
[X]:	019 / 180 (10.6%)																																																					
[符號]:	024 / 180 (13.3%)																																																					
[1]:	075 / 217 (34.6%)																																																					
[1*]:	037 / 217 (17.1%)																																																					
[2]:	013 / 217 (6.0%)																																																					
[2*]:	006 / 217 (2.8%)																																																					
[3]:	009 / 217 (4.1%)																																																					
[3*]:	005 / 217 (2.3%)																																																					
[4]:	003 / 217 (1.4%)																																																					
[4*]:	004 / 217 (1.8%)																																																					
[5]:	002 / 217 (0.9%)																																																					
[6]:	002 / 217 (0.9%)																																																					
[7]:	000 / 217 (0.0%)																																																					
[X]:	032 / 217 (14.7%)																																																					
[符號]:	029 / 217 (13.4%)																																																					

說明 人工題本和 AI 題本在這個部分的用詞多落在等級 1 到等級 1*，各佔了 59.5%和 51.7%，相差不多；而等級 2 和等級 2*則各佔 12.8%和 10.8%，同樣相差不多。但是在 AI 題本中，使用等級 3 到等級 4*的比例（9.6%）略高於人工題本（2.9%）。兩組的用詞最高等級皆到等級 6，比例皆各佔 1%左右；在兩組的超綱詞中，大多是屬於人名或動補結構的語詞，而比例也各為 10 多%左右。整體來說兩組的用詞難度差不多，但是 AI 題本的用語等級相較於人工題本廣泛，從等級 1 到等級 6 皆有所採用。

(續上頁)

第二
部分
看圖
釋義

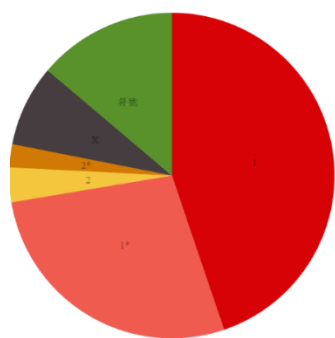
說明 人工題本和 AI 題本在這個部分的用詞多落在等級 1 到等級 1*，各佔了 61%和 54.8%，相差不多；而等級 2 和等級 2*則各佔 9.8%和 11.7%，同樣相差不多。兩組的用詞最高等級皆到等級 6，比例皆各佔 1%左右；在兩組的超綱詞中，大多是屬於人名、特殊名詞或動補結構的語詞，而比例也相差不多。整體而言和第一部分的情況類似，AI 題本的用語等級相較於人工題本廣泛，從等級 1 到等級 6 皆有所採用。

第三
部分
選詞
填空

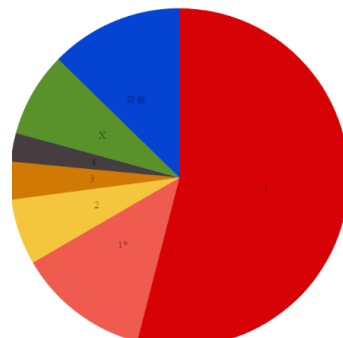
說明 人工題本和 AI 題本在這個部分的用詞多落在等級 1 到等級 1*，各佔了 57.6%和 51.9%，人工題本略多；而等級 2 和等級 2*則各佔 13.1%和 13.6%，幾乎一樣；但是在等級 3 到 4*，AI 題本（11.7%）高於人工題本（5%）；另外 AI 題本的用詞最高等級到等級 6，而人工題本僅到等級 4，由此可知 AI 題本的用語等級相較於人工題本廣泛。在兩組的超綱詞中，大多是屬於人名、特殊名詞或動補結構的語詞，而比例也相差不多。

(續下頁)

第四
部分
選句
填空



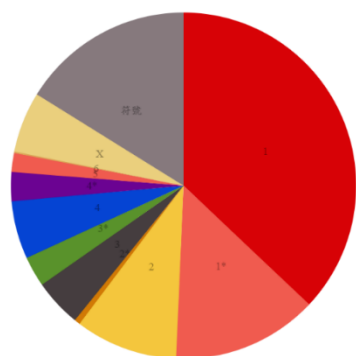
[1]:	039 / 087	(44.8%)
[1*]:	024 / 087	(27.6%)
[2]:	003 / 087	(3.4%)
[2*]:	002 / 087	(2.3%)
[3]:	000 / 087	(0.0%)
[3*]:	000 / 087	(0.0%)
[4]:	000 / 087	(0.0%)
[4*]:	000 / 087	(0.0%)
[5]:	000 / 087	(0.0%)
[6]:	000 / 087	(0.0%)
[7]:	000 / 087	(0.0%)
[X]:	007 / 087	(8.0%)
[符號]:	012 / 087	(13.8%)



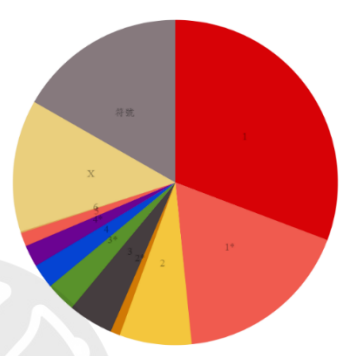
[1]:	060 / 111	(54.1%)
[1*]:	014 / 111	(12.6%)
[2]:	007 / 111	(6.3%)
[2*]:	000 / 111	(0.0%)
[3]:	004 / 111	(3.6%)
[3*]:	000 / 111	(0.0%)
[4]:	003 / 111	(2.7%)
[4*]:	000 / 111	(0.0%)
[5]:	000 / 111	(0.0%)
[6]:	000 / 111	(0.0%)
[7]:	000 / 111	(0.0%)
[X]:	009 / 111	(8.1%)
[符號]:	014 / 111	(12.6%)

說明 兩組在等級 1 和 1* 的比例相當（對照：71.6%；實驗：74%），但是人工題本的用詞分佈僅在等級 1 到等級 2*，而 AI 題本的用詞還使用了少數等級 3 和 4 的詞語。

第五
部分
閱讀
理解



[1]:	274 / 739	(37.1%)
[1*]:	101 / 739	(13.7%)
[2]:	071 / 739	(9.6%)
[2*]:	004 / 739	(0.5%)
[3]:	033 / 739	(4.5%)
[3*]:	021 / 739	(2.8%)
[4]:	040 / 739	(5.4%)
[4*]:	020 / 739	(2.7%)
[5]:	013 / 739	(1.8%)
[6]:	001 / 739	(0.1%)
[7]:	000 / 739	(0.0%)
[X]:	042 / 739	(5.7%)
[符號]:	119 / 739	(16.1%)



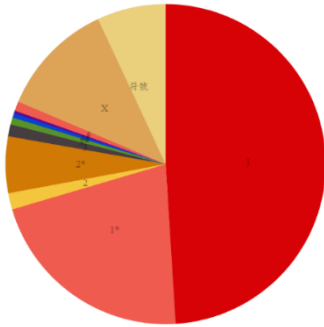
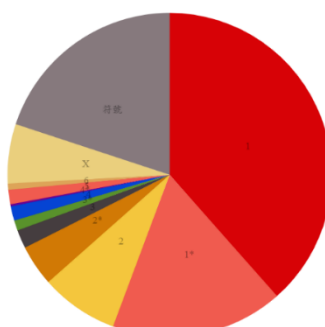
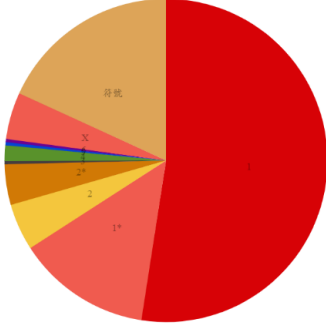
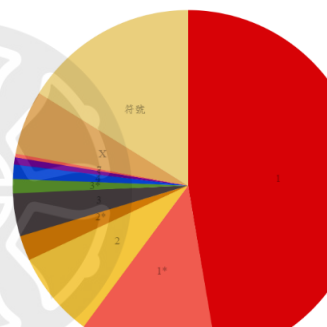
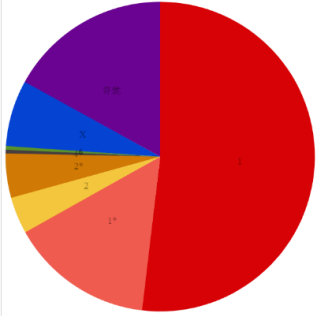
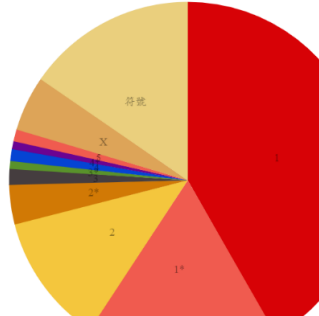
[1]:	227 / 737	(30.8%)
[1*]:	130 / 737	(17.6%)
[2]:	053 / 737	(7.2%)
[2*]:	007 / 737	(0.9%)
[3]:	033 / 737	(4.5%)
[3*]:	022 / 737	(3.0%)
[4]:	018 / 737	(2.4%)
[4*]:	016 / 737	(2.2%)
[5]:	010 / 737	(1.4%)
[6]:	001 / 737	(0.1%)
[7]:	000 / 737	(0.0%)
[X]:	097 / 737	(13.2%)
[符號]:	123 / 737	(16.7%)

說明 這個部分兩組的用語等級分佈相似，詞語大多落在等級 1 和 1*，並都同樣使用了少數的等級 3 到 5 的詞語，佔比也都相近。

根據表 4-2 數據，聽力測驗的結果也與閱讀測驗呈現相似趨勢。兩組題本的整體用詞難度接近，主要以等級 1 和 1* 的詞彙為主，僅少量涉及等級 5 和 6，且比例極低。而相比人工題本，AI 題本較常使用等級 3 至 4* 的詞彙，雖整體佔比有限，但仍反映出生成模型在嚴格控制語言難度的限制。

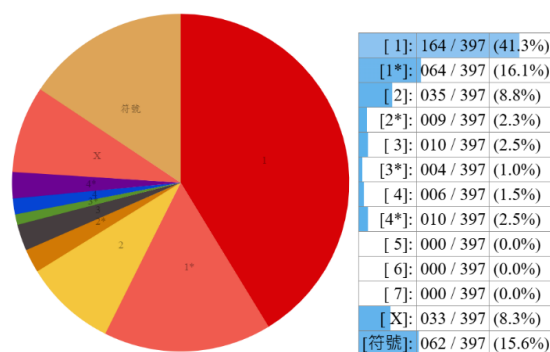
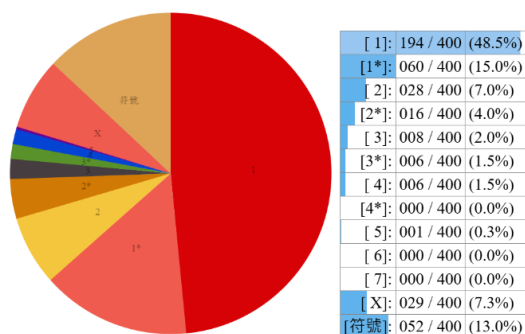
表4-2

聽力測驗兩題本各部分的斷詞表

題型	人工題本	AI 題本																																																				
第一部分 看圖 釋義	 <table border="1"> <tr><td>[1]:</td><td>206 / 420 (49.0%)</td></tr> <tr><td>[1*]:</td><td>090 / 420 (21.4%)</td></tr> <tr><td>[2]:</td><td>007 / 420 (1.7%)</td></tr> <tr><td>[2*]:</td><td>024 / 420 (5.7%)</td></tr> <tr><td>[3]:</td><td>005 / 420 (1.2%)</td></tr> <tr><td>[3*]:</td><td>003 / 420 (0.7%)</td></tr> <tr><td>[4]:</td><td>002 / 420 (0.5%)</td></tr> <tr><td>[4*]:</td><td>001 / 420 (0.2%)</td></tr> <tr><td>[5]:</td><td>004 / 420 (1.0%)</td></tr> <tr><td>[6]:</td><td>000 / 420 (0.0%)</td></tr> <tr><td>[7]:</td><td>000 / 420 (0.0%)</td></tr> <tr><td>[X]:</td><td>049 / 420 (11.7%)</td></tr> <tr><td>[符號]:</td><td>029 / 420 (6.9%)</td></tr> </table>	[1]:	206 / 420 (49.0%)	[1*]:	090 / 420 (21.4%)	[2]:	007 / 420 (1.7%)	[2*]:	024 / 420 (5.7%)	[3]:	005 / 420 (1.2%)	[3*]:	003 / 420 (0.7%)	[4]:	002 / 420 (0.5%)	[4*]:	001 / 420 (0.2%)	[5]:	004 / 420 (1.0%)	[6]:	000 / 420 (0.0%)	[7]:	000 / 420 (0.0%)	[X]:	049 / 420 (11.7%)	[符號]:	029 / 420 (6.9%)	 <table border="1"> <tr><td>[1]:</td><td>188 / 488 (38.5%)</td></tr> <tr><td>[1*]:</td><td>084 / 488 (17.2%)</td></tr> <tr><td>[2]:</td><td>038 / 488 (7.8%)</td></tr> <tr><td>[2*]:</td><td>020 / 488 (4.1%)</td></tr> <tr><td>[3]:</td><td>009 / 488 (1.8%)</td></tr> <tr><td>[3*]:</td><td>005 / 488 (1.0%)</td></tr> <tr><td>[4]:</td><td>007 / 488 (1.4%)</td></tr> <tr><td>[4*]:</td><td>001 / 488 (0.2%)</td></tr> <tr><td>[5]:</td><td>007 / 488 (1.4%)</td></tr> <tr><td>[6]:</td><td>003 / 488 (0.6%)</td></tr> <tr><td>[7]:</td><td>000 / 488 (0.0%)</td></tr> <tr><td>[X]:</td><td>029 / 488 (5.9%)</td></tr> <tr><td>[符號]:</td><td>097 / 488 (19.9%)</td></tr> </table>	[1]:	188 / 488 (38.5%)	[1*]:	084 / 488 (17.2%)	[2]:	038 / 488 (7.8%)	[2*]:	020 / 488 (4.1%)	[3]:	009 / 488 (1.8%)	[3*]:	005 / 488 (1.0%)	[4]:	007 / 488 (1.4%)	[4*]:	001 / 488 (0.2%)	[5]:	007 / 488 (1.4%)	[6]:	003 / 488 (0.6%)	[7]:	000 / 488 (0.0%)	[X]:	029 / 488 (5.9%)	[符號]:	097 / 488 (19.9%)
[1]:	206 / 420 (49.0%)																																																					
[1*]:	090 / 420 (21.4%)																																																					
[2]:	007 / 420 (1.7%)																																																					
[2*]:	024 / 420 (5.7%)																																																					
[3]:	005 / 420 (1.2%)																																																					
[3*]:	003 / 420 (0.7%)																																																					
[4]:	002 / 420 (0.5%)																																																					
[4*]:	001 / 420 (0.2%)																																																					
[5]:	004 / 420 (1.0%)																																																					
[6]:	000 / 420 (0.0%)																																																					
[7]:	000 / 420 (0.0%)																																																					
[X]:	049 / 420 (11.7%)																																																					
[符號]:	029 / 420 (6.9%)																																																					
[1]:	188 / 488 (38.5%)																																																					
[1*]:	084 / 488 (17.2%)																																																					
[2]:	038 / 488 (7.8%)																																																					
[2*]:	020 / 488 (4.1%)																																																					
[3]:	009 / 488 (1.8%)																																																					
[3*]:	005 / 488 (1.0%)																																																					
[4]:	007 / 488 (1.4%)																																																					
[4*]:	001 / 488 (0.2%)																																																					
[5]:	007 / 488 (1.4%)																																																					
[6]:	003 / 488 (0.6%)																																																					
[7]:	000 / 488 (0.0%)																																																					
[X]:	029 / 488 (5.9%)																																																					
[符號]:	097 / 488 (19.9%)																																																					
說明	兩組的詞語難度分佈大致相同，皆有 1 到 5、6 級的字詞，並大部分皆使用 1 和 1* 等級的字詞。																																																					
第二部分 問答 理解	 <table border="1"> <tr><td>[1]:</td><td>168 / 320 (52.5%)</td></tr> <tr><td>[1*]:</td><td>043 / 320 (13.4%)</td></tr> <tr><td>[2]:</td><td>015 / 320 (4.7%)</td></tr> <tr><td>[2*]:</td><td>013 / 320 (4.1%)</td></tr> <tr><td>[3]:</td><td>001 / 320 (0.3%)</td></tr> <tr><td>[3*]:</td><td>000 / 320 (0.0%)</td></tr> <tr><td>[4]:</td><td>005 / 320 (1.6%)</td></tr> <tr><td>[4*]:</td><td>000 / 320 (0.0%)</td></tr> <tr><td>[5]:</td><td>001 / 320 (0.3%)</td></tr> <tr><td>[6]:</td><td>001 / 320 (0.3%)</td></tr> <tr><td>[7]:</td><td>000 / 320 (0.0%)</td></tr> <tr><td>[X]:</td><td>015 / 320 (4.7%)</td></tr> <tr><td>[符號]:</td><td>058 / 320 (18.1%)</td></tr> </table>	[1]:	168 / 320 (52.5%)	[1*]:	043 / 320 (13.4%)	[2]:	015 / 320 (4.7%)	[2*]:	013 / 320 (4.1%)	[3]:	001 / 320 (0.3%)	[3*]:	000 / 320 (0.0%)	[4]:	005 / 320 (1.6%)	[4*]:	000 / 320 (0.0%)	[5]:	001 / 320 (0.3%)	[6]:	001 / 320 (0.3%)	[7]:	000 / 320 (0.0%)	[X]:	015 / 320 (4.7%)	[符號]:	058 / 320 (18.1%)	 <table border="1"> <tr><td>[1]:</td><td>144 / 304 (47.4%)</td></tr> <tr><td>[1*]:</td><td>039 / 304 (12.8%)</td></tr> <tr><td>[2]:</td><td>024 / 304 (7.9%)</td></tr> <tr><td>[2*]:</td><td>007 / 304 (2.3%)</td></tr> <tr><td>[3]:</td><td>012 / 304 (3.9%)</td></tr> <tr><td>[3*]:</td><td>004 / 304 (1.3%)</td></tr> <tr><td>[4]:</td><td>004 / 304 (1.3%)</td></tr> <tr><td>[4*]:</td><td>000 / 304 (0.0%)</td></tr> <tr><td>[5]:</td><td>002 / 304 (0.7%)</td></tr> <tr><td>[6]:</td><td>000 / 304 (0.0%)</td></tr> <tr><td>[7]:</td><td>001 / 304 (0.3%)</td></tr> <tr><td>[X]:</td><td>018 / 304 (5.9%)</td></tr> <tr><td>[符號]:</td><td>049 / 304 (16.1%)</td></tr> </table>	[1]:	144 / 304 (47.4%)	[1*]:	039 / 304 (12.8%)	[2]:	024 / 304 (7.9%)	[2*]:	007 / 304 (2.3%)	[3]:	012 / 304 (3.9%)	[3*]:	004 / 304 (1.3%)	[4]:	004 / 304 (1.3%)	[4*]:	000 / 304 (0.0%)	[5]:	002 / 304 (0.7%)	[6]:	000 / 304 (0.0%)	[7]:	001 / 304 (0.3%)	[X]:	018 / 304 (5.9%)	[符號]:	049 / 304 (16.1%)
[1]:	168 / 320 (52.5%)																																																					
[1*]:	043 / 320 (13.4%)																																																					
[2]:	015 / 320 (4.7%)																																																					
[2*]:	013 / 320 (4.1%)																																																					
[3]:	001 / 320 (0.3%)																																																					
[3*]:	000 / 320 (0.0%)																																																					
[4]:	005 / 320 (1.6%)																																																					
[4*]:	000 / 320 (0.0%)																																																					
[5]:	001 / 320 (0.3%)																																																					
[6]:	001 / 320 (0.3%)																																																					
[7]:	000 / 320 (0.0%)																																																					
[X]:	015 / 320 (4.7%)																																																					
[符號]:	058 / 320 (18.1%)																																																					
[1]:	144 / 304 (47.4%)																																																					
[1*]:	039 / 304 (12.8%)																																																					
[2]:	024 / 304 (7.9%)																																																					
[2*]:	007 / 304 (2.3%)																																																					
[3]:	012 / 304 (3.9%)																																																					
[3*]:	004 / 304 (1.3%)																																																					
[4]:	004 / 304 (1.3%)																																																					
[4*]:	000 / 304 (0.0%)																																																					
[5]:	002 / 304 (0.7%)																																																					
[6]:	000 / 304 (0.0%)																																																					
[7]:	001 / 304 (0.3%)																																																					
[X]:	018 / 304 (5.9%)																																																					
[符號]:	049 / 304 (16.1%)																																																					
說明	兩組在等級 1 到等級 2* 的使用比例差不多，但在等級 3 到 4 的使用上，AI 題本的佔比略高於人工題本。																																																					
第三部分 對話 理解	 <table border="1"> <tr><td>[1]:</td><td>135 / 260 (51.9%)</td></tr> <tr><td>[1*]:</td><td>039 / 260 (15.0%)</td></tr> <tr><td>[2]:</td><td>010 / 260 (3.8%)</td></tr> <tr><td>[2*]:</td><td>012 / 260 (4.6%)</td></tr> <tr><td>[3]:</td><td>000 / 260 (0.0%)</td></tr> <tr><td>[3*]:</td><td>000 / 260 (0.0%)</td></tr> <tr><td>[4]:</td><td>000 / 260 (0.0%)</td></tr> <tr><td>[4*]:</td><td>001 / 260 (0.4%)</td></tr> <tr><td>[5]:</td><td>000 / 260 (0.0%)</td></tr> <tr><td>[6]:</td><td>001 / 260 (0.4%)</td></tr> <tr><td>[7]:</td><td>000 / 260 (0.0%)</td></tr> <tr><td>[X]:</td><td>018 / 260 (6.9%)</td></tr> <tr><td>[符號]:</td><td>044 / 260 (16.9%)</td></tr> </table>	[1]:	135 / 260 (51.9%)	[1*]:	039 / 260 (15.0%)	[2]:	010 / 260 (3.8%)	[2*]:	012 / 260 (4.6%)	[3]:	000 / 260 (0.0%)	[3*]:	000 / 260 (0.0%)	[4]:	000 / 260 (0.0%)	[4*]:	001 / 260 (0.4%)	[5]:	000 / 260 (0.0%)	[6]:	001 / 260 (0.4%)	[7]:	000 / 260 (0.0%)	[X]:	018 / 260 (6.9%)	[符號]:	044 / 260 (16.9%)	 <table border="1"> <tr><td>[1]:</td><td>117 / 280 (41.8%)</td></tr> <tr><td>[1*]:</td><td>049 / 280 (17.5%)</td></tr> <tr><td>[2]:</td><td>033 / 280 (11.8%)</td></tr> <tr><td>[2*]:</td><td>010 / 280 (3.6%)</td></tr> <tr><td>[3]:</td><td>004 / 280 (1.4%)</td></tr> <tr><td>[3*]:</td><td>002 / 280 (0.7%)</td></tr> <tr><td>[4]:</td><td>003 / 280 (1.1%)</td></tr> <tr><td>[4*]:</td><td>002 / 280 (0.7%)</td></tr> <tr><td>[5]:</td><td>003 / 280 (1.1%)</td></tr> <tr><td>[6]:</td><td>000 / 280 (0.0%)</td></tr> <tr><td>[7]:</td><td>000 / 280 (0.0%)</td></tr> <tr><td>[X]:</td><td>014 / 280 (5.0%)</td></tr> <tr><td>[符號]:</td><td>043 / 280 (15.4%)</td></tr> </table>	[1]:	117 / 280 (41.8%)	[1*]:	049 / 280 (17.5%)	[2]:	033 / 280 (11.8%)	[2*]:	010 / 280 (3.6%)	[3]:	004 / 280 (1.4%)	[3*]:	002 / 280 (0.7%)	[4]:	003 / 280 (1.1%)	[4*]:	002 / 280 (0.7%)	[5]:	003 / 280 (1.1%)	[6]:	000 / 280 (0.0%)	[7]:	000 / 280 (0.0%)	[X]:	014 / 280 (5.0%)	[符號]:	043 / 280 (15.4%)
[1]:	135 / 260 (51.9%)																																																					
[1*]:	039 / 260 (15.0%)																																																					
[2]:	010 / 260 (3.8%)																																																					
[2*]:	012 / 260 (4.6%)																																																					
[3]:	000 / 260 (0.0%)																																																					
[3*]:	000 / 260 (0.0%)																																																					
[4]:	000 / 260 (0.0%)																																																					
[4*]:	001 / 260 (0.4%)																																																					
[5]:	000 / 260 (0.0%)																																																					
[6]:	001 / 260 (0.4%)																																																					
[7]:	000 / 260 (0.0%)																																																					
[X]:	018 / 260 (6.9%)																																																					
[符號]:	044 / 260 (16.9%)																																																					
[1]:	117 / 280 (41.8%)																																																					
[1*]:	049 / 280 (17.5%)																																																					
[2]:	033 / 280 (11.8%)																																																					
[2*]:	010 / 280 (3.6%)																																																					
[3]:	004 / 280 (1.4%)																																																					
[3*]:	002 / 280 (0.7%)																																																					
[4]:	003 / 280 (1.1%)																																																					
[4*]:	002 / 280 (0.7%)																																																					
[5]:	003 / 280 (1.1%)																																																					
[6]:	000 / 280 (0.0%)																																																					
[7]:	000 / 280 (0.0%)																																																					
[X]:	014 / 280 (5.0%)																																																					
[符號]:	043 / 280 (15.4%)																																																					
說明	兩組在等級 1 到等級 2* 的使用比例差不多，但在等級 3 到 5 的使用上，AI 題本的佔比略高於人工題本。																																																					

(續下頁)

第四
部分
對話



說明 兩組的詞語難度分佈大致相同，皆有 1 到 5、6 級的字詞，並大部分皆使用 1 和 1* 等級的字詞。

整體而言，AI 題本的用詞雖與人工題本同樣，以等級 1 和 1* 詞彙為主，但其詞彙等級分佈較為廣泛，涵蓋至等級 3 至 6 的詞語比例略高，顯示出 AI 出題系統在語料選用上較不集中於入門級詞彙，語言使用的範圍也較具彈性。此現象雖可能有助於試題多樣性，但亦需注意是否會對初級學習者造成理解負擔，或降低題本在基礎等級測驗中的適切性。

貳、 題本的四層次提問類型分布

為確保題本難度相當，本研究於第三章分析了 TOCFL 模擬試題的四層次提問類型及其統計分佈，並在生成 AI 試題時，盡可能使其提問類型的比例與統計結果相符。表 4-3 與表 4-4 將呈現受試者測驗之 AI 題本與人工題本在四層次提問類型上的題數統計。

根據表 4-3 和表 4-4 的統計資料，AI 題本在各部分均涵蓋不同層次的提問類型，並在設計過程中參考官方題本的統計比例，力求在提問類型的分佈上與其保持相近，以確保 AI 題本與官方題本在難度設定上一致，從而減少因試題設計差異所可能產生的影響，提升後續研究結果的有效性與信效度。

表4-3

兩題本閱讀測驗中的 PIRLS 類型題數統計表

	第一部分 單句理解		第二部分 看圖釋義		第三部分 選詞填空		第四部分 完成段落		第五部分 閱讀理解		合計	
	人工	AI	人工	AI	人工	AI	人工	AI	人工	AI	人工	AI
直接提取	8 (53%)	10 (67%)	9 (60%)	8 (53%)	6 (60%)	4 (40%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	23	22
直接推論	7 (47%)	5 (33%)	4 (27%)	5 (34%)	4 (40%)	6 (60%)	4 (80%)	4 (80%)	1 (20%)	2 (40%)	20	22
詮釋與整合	0 (0%)	0 (0%)	2 (13%)	2 (13%)	0 (0%)	0 (0%)	1 (20%)	1 (20%)	4 (80%)	3 (60%)	7	6
各部分題數	15 (100%)	15 (100%)	15 (100%)	15 (100%)	10 (100%)	10 (100%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	--	--

表4-4

兩題本聽力測驗中的 PIRLS 類型題數統計表

	第一部分 看圖釋義		第二部分 問答理解		第三部分 對話理解		第四部分 對話		合計	
	人工	AI	人工	AI	人工	AI	人工	AI	人工	AI
直接提取	21 (84%)	16 (64%)	8 (53%)	10 (67%)	1 (20%)	2 (40%)	0 (0%)	0 (0%)	30	28

(續下頁)

直接推論	4 (16%)	9 (36%)	4 (27%)	3 (20%)	2 (40%)	2 (40%)	1 (20%)	2 (40%)	11	16
詮釋與整合	0 (0%)	0 (0%)	3 (20%)	2 (13%)	2 (40%)	1 (20%)	4 (80%)	3 (60%)	9	7
各部分題數	25 (100%)	25 (100%)	15 (100%)	15 (100%)	5 (100%)	5 (100%)	5 (100%)	5 (100%)	--	--

整體而言，AI 題本在四層次提問類型的分布上與人工題本相近，顯示本研究在生成題目時已盡可能參考 TOCFL 模擬試題的比例，維持題型與提問層次的基本結構。然而，提問類型雖然在數量上對應一致，實際是否能呈現出相應的認知挑戰與語言理解層次，仍需進一步結合受試者的作答表現進行分析。因此接下來本研究將透過受試者的答題表現與錯誤題型分析，探討 AI 題本在編製不同理解層次題目上的命題能力與鑑別度，以及該題本在學習診斷上的潛力與限制。

第三節 受試者的答題表現分析

壹、 量化數據分析

一、 單一測驗題本的成績分析－統計

為了呈現各題本測驗的在不同語言程度受試者中的成績表現，表 4-5 呈現各測驗題本的施測人數、通過 A2 等級人數、最高得分、最低得分、整體平均分數以及標準差。為更清楚瞭解答題情況，以下先說明各測驗分數對照：閱讀測驗（人工考題與 AI 考題）滿分 80 分，共 50

題，60分（39題）以上通過A2等級；聽力測驗（人工考題與AI考題）滿分80分，共50題，60分（40題）以上通過A2等級。

表4-5

各題本成績統計表

測驗類型	能力程度組別	施測 人數	通過 人數	最高分	最低分	平均 分數	標準差
人工試題 (閱讀)	低能力組	3	3	69	60	65.3	4.726
	中能力組	26	26	78	63	73.3	4.722
	高能力組	10	10	80	65	73.2	4.984
AI 試題 (閱讀)	低能力組	3	3	76	61	68.6	7.505
	中能力組	26	26	80	69	76.7	4.546
	高能力組	10	10	80	69	76.3	3.164
人工試題 (聽力)	低能力組	5	5	80	71	75.6	3.881
	中能力組	18	18	80	75	77.9	1.731
	高能力組	16	16	80	75	79.4	1.258
AI 試題 (聽力)	低能力組	5	5	80	68	76.0	4.446
	中能力組	18	18	80	77	78.4	1.722
	高能力組	16	16	80	75	79.0	1.461

二、單一測驗題本的成绩分析—發現

(一) 閱讀測驗

1. 試題難度分析

(1) 中能力組與高能力組：AI 閱讀試題的平均分數（中能力組為76.7分，高能力組為76.3分）均顯著高於人工閱讀試題（中能力組73.3分，高能力組73.2分）。這顯示對於中高能力程度的受試者而言，AI 生成的閱讀試題整體難度較低，使他們更容易取得較高分數。

(2) 低能力組：AI 閱讀試題的平均分數（68.6分）也略高於人工閱讀試題（65.3分），同樣反映出AI 題本對低能力組而言，難度亦較低的現象。

2. 試題鑑別度分析

(1) 人工試題：各能力組的標準差（低能力組為 4.726，中能力組為 4.722，高能力組為 4.984）相對接近且穩定。這表示人工閱讀試題在各能力組中，其分數分散程度大致相似。

(2) AI 試題：在各能力組中呈現不同趨勢

- 低能力組：AI 試題的低能力組標準差為 7.505，遠高於人工閱讀的 4.726。這表明在低能力組中，AI 閱讀題可能鑑別力較強，可以有效區分出受試者的表現差異。
- 中能力組：AI 試題的中能力組標準差為 4.546，與人工題本的中能力組標準差（4.722）相近。顯示對於中能力組而言，兩題本的鑑別力較為相當。
- 高能力組：AI 試題對高能力組的標準差僅為 3.164，顯著低於人工閱讀的 4.984，再加上高能力組在 AI 試題的平均分數為 76.3，顯示了 AI 閱讀試題對高能力組的鑑別度較弱，分數高度集中在高分區，題目無法有效區分出高能力學習者之間的細微差異。

（二）聽力測驗

1. 試題難度分析：在所有能力組別中，AI 聽力試題的平均分數（低能力組 76.0 分，中能力組 78.4 分，高能力組 79.0 分）與人工聽力試題的平均分數（低能力組 75.6 分，中能力組 77.9 分，高能力組 79.4 分）極為接近。這顯示了 AI 題本在聽力測驗題型的模仿與難度控制上展現了較高的精準度，其生成的聽力試題在難度層面與官方試題幾乎無差異。

2. 試題鑑別度分析：AI 聽力試題與人工聽力試題在中、高能力組的標準差均普遍較低且相互接近。例如，高能力組的標準差在 AI 試題與人工試題分別為 1.461 和 1.258，這代表兩者題本在中高能力組而言，鑑別度皆較低。另外，高能力組在 AI 試題與人工試題的平均分數分別為 79.0 與 79.4，分數普遍集中在高分區域，代表試題可能難以有效區分出高能力學習者在聽力理解上的細微差距。

(三) 小結

綜合分析結果顯示，AI 在聽力試題的生成上展現出一定的模仿能力，不論是平均難度或分數的分散程度，都與人工試題相當接近。不過在閱讀試題的部分，AI 題本整體難度偏低，對低能力組的區辨效果較好，但在高能力組的鑑別度明顯下降，呈現出整體分數集中在高分區的趨勢，這與人工題本的表現有明顯差異。

三、受試者的成績比較

閱讀測驗中共 39 位受試者，由於其中四位受試者（代號為 B7、B17、B26 以及 C1）在作答過程中出現跳題情況，導致成績無法完整對應於兩份題本的分析，為維持資料處理的一致性，故將其數據排除，最終納入分析者共 35 人。本研究將受試者按照其適性測驗結果的等級分數分成低能力、中能力、高能力三組，其中低能力組有 3 人；中能力組有 23 人；高能力組有 9 人。聽力測驗中共 39 位受試者，其中低能力組有 6 人；中能力組有 18 人；高能力組有 16 人。

為分析同一受試者在 AI 題本與人工題本的分數差異，以下將統計同一受試者在不同題本的成績分數差，並按照不同能力組做分類，分析在不同能力的受試者中，成績是否有所差異。個人分數差的計算方式為（人工試題分數-AI 試題分數），經計算後，閱讀測驗的分數差統

計如表 4-6 所示；聽力測驗的分數差統計則如表 4-7 所示。

表4-6

閱讀測驗各能力組的分數差統計表

能力分組類別	受試者代號	人工試題分數	AI 試題分數	兩試題分數差	分數差的標準差	t 值
低能力組 共 3 人	A1	69	69	0	4.932	-1.17
	A2	60	61	-1		
	A3	67	76	-9		
中能力組 共 23 人	B1	78	79	-1	5.241	-3.78
	B2	74	80	-6		
	B3	74	78	-4		
	B4	65	78	-13		
	B5	78	79	-1		
	B6	71	79	-8		
	B8	69	60	9		
	B9	76	79	-3		
	B10	74	80	-6		
	B11	71	76	-5		
	B12	78	80	-2		
	B13	71	79	-8		
	B14	79	79	0		
	B15	78	79	-1		
	B16	74	76	-2		
	B18	71	79	-8		
	B19	76	74	2		
	B20	78	78	0		
	B21	63	76	-13		
	B22	74	74	0		
	B23	76	79	-3		
	B24	67	80	-13		
	B25	78	78	0		

(續下頁)

高能力組 共 9 人	C2	78	78	0	3.632	-2.85
	C3	79	79	0		
	C4	71	74	-3		
	C5	65	69	-4		
	C6	69	79	-10		
	C7	69	76	-7		
	C8	80	80	0		
	C9	71	76	-5		
	C10	76	76	0		

(一) 閱讀測驗兩題本分數差異分析

1. 低能力組：此組的分數差標準差為 4.932，代表 AI 閱讀試題對此組受試者的成績影響顯示出不一致性，例如受試者 A3 的 AI 題本分數（76 分）比人工題本分數（67 分）高出 9 分，而受試者 A1、A2 則沒有太大變化，這種在少數個體中呈現出的多樣化分數變化，進一步說明了 AI 閱讀題本對低能力組的某些個體會產生較大的成績差異。此組的配對樣本 t 值為 -1.17，表示 AI 閱讀試題分數平均略高於人工試題分數。然而，由於樣本數較小 ($n=3$, 自由度 $df=2$)，在常用的顯著水準 ($\alpha=0.05$) 下，此分數差異不具備統計顯著性。這意味目前尚無法從統計學上斷定 AI 試題對於低能力組的整體成績產生了普遍的影響。
2. 中能力組：該組的分數差標準差為 5.241，是三個能力組中分數差標準差最大的。這顯示在中能力組中，受試者在兩份試題之間成績變化的幅度有較大的分散度。另外，從表 4-6 可觀察到多數受試者的 AI 試題分數更高，再結合表 4-5 的平均分數來看 (AI 閱讀平均分 76.7 分，人工閱讀平均分 73.3 分)，證實了 AI 閱讀試題對於中能力組來說可能整體難度較低，使得受試者普遍能獲得較高的分數。此組的配對樣本 t 值為 -3.78。

在樣本數為 23 (自由度 $df=22$) 的情況下，此 t 值達到統計顯著性，表明 AI 閱讀試題的平均分數顯著高於人工閱讀試題。然而，分數差標準差較大也同時反映了這種難度變化對中能力組個體影響的不一致性，即為部分受試者分數大幅提升，而另一些則變化較小，甚至可能略有下降，這可能暗示 AI 在針對中能力程度的命題穩定性或預測性上仍有提升空間。

3. 高能力組：此組分數差的標準差為 3.446，是所有組別中分數差標準差最低的。這顯示了在高能力組中，受試者在人工試題與 AI 試題之間的分數變化非常小。此組的配對樣本 t 值為 -2.85。在樣本數為 10 (自由度 $df=9$) 的情況下，此 t 值達到統計顯著性，表明 AI 閱讀試題的平均分數顯著高於人工閱讀試題。這說明 AI 題本對於高能力者的命題品質與人工題本具有高度的一致性，且同樣能帶來分數的提升，但也可能提示其對高能力者存在天花板效應 (Ceiling Effect) (Taylor, 2010)，未能有效拉開較程度學習者之間的細微差距。

表4-7

聽力測驗各能力組的分數差統計表

能力分組類別	受試者代號	人工試題分數	AI 試題分數	分數差	分數差的標準差	t 值
低能力組 共 5 人	D1	79	77	2	3.697	-0.24
	D2	75	80	-5		
	D3	71	68	3		
	D4	73	75	-2		
	D5	80	80	0		

(續下頁)

中能力組 共 18 人	E1	79	79	0	2.333	-0.909
	E2	75	80	-5		
	E3	80	77	3		
	E4	77	79	-2		
	E5	79	77	2		
	E6	77	73	4		
	E7	77	77	0		
	E8	80	80	0		
	E9	80	79	1		
	E10	77	79	-2		
	E11	79	80	-1		
	E12	79	79	0		
	E13	75	79	-4		
	E14	79	80	-1		
	E15	77	77	0		
	E16	75	79	-4		
	E17	79	79	0		
	E18	79	79	0		
高能力組 共 16 人	F1	79	79	0	1.147	1.31
	F2	80	80	0		
	F3	80	79	1		
	F4	80	77	3		
	F5	80	80	0		
	F6	80	80	0		
	F7	75	75	0		
	F8	79	79	0		
	F9	79	79	0		
	F10	79	80	-1		
	F11	79	80	-1		
	F12	80	77	3		
	F13	80	80	0		
	F14	80	79	1		
	F15	80	80	0		
	F16	80	80	0		

(二) 聽力測驗兩題本分數分析

1. 低能力組：低能力組(共 5 位受試者)的分數差標準差為 3.697，這顯示其在聽力測驗中，兩份題本間的成績變化幅度呈現中等程度的分散。從個別數據來看，分數差有正有負(例如受試者 D1 為 2，D2 為-5)，表明 AI 聽力試題對低能力者的影響程度各異。此組的配對樣本 t 值為-0.24。由於樣本數較小($n=5$ ，自由度 $df=4$)，在常用的顯著水準 ($\alpha=0.05$) 下，該差異不具備統計顯著性。這說明在統計上，AI 聽力試題與人工試題在低能力組的平均分數並無顯著差異。
2. 中能力組：此組分數差標準差為 2.333，在三組中相對較低。這表明該組受試者在兩份聽力試題之間的分數變化幅度相對集中且一致。此組的配對樣本 t 值為-0.91。在樣本數為 18(自由度 $df=17$) 的情況下，此 t 值不具備統計顯著性。這進一步驗證了 AI 聽力命題在該能力層級上，能夠較好地維持與人工題本的平均分數一致性。
3. 高能力組：該組分數差標準差為 1.147，是所有組別中最低的。這顯示高能力組受試者在人工試題與 AI 試題之間的分數變化極為微小且高度穩定。此組的配對樣本 t 值為 1.31。在樣本數為 16(自由度 $df=15$) 的情況下，此 t 值不具備統計顯著性。極低的標準差和統計上不顯著的 t 值，共同驗證了 AI 聽力試題在命題品質上與人工題本具有高度的相似性，分數表現幾乎一致，但在區分頂尖能力者細微差異方面也可能同樣面臨天花板效應 (Taylor, 2010) 的挑戰。

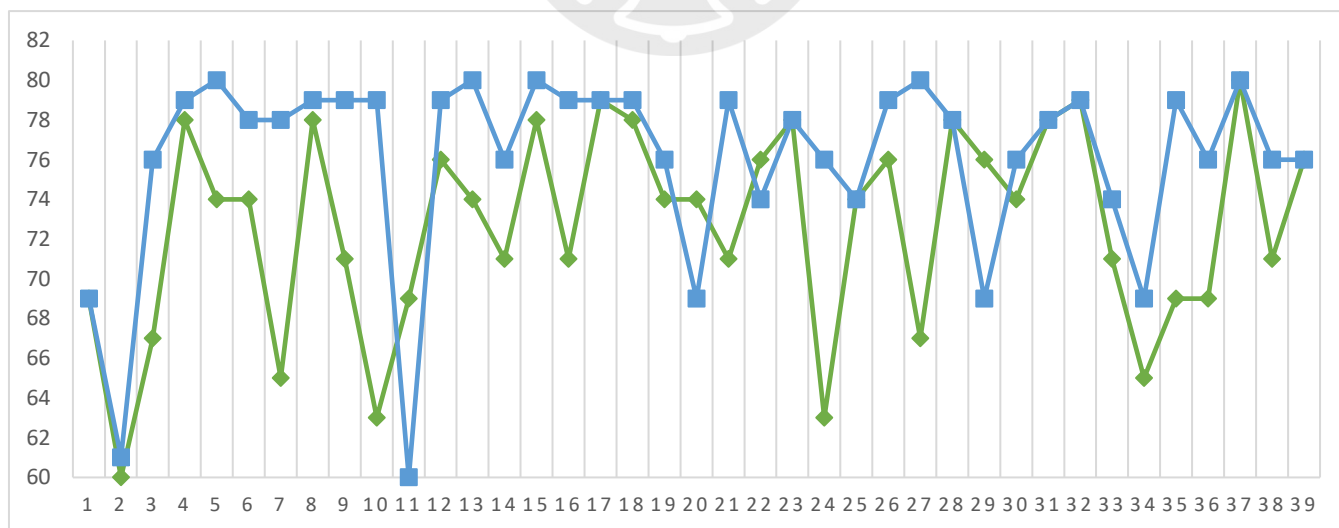
四、兩題本的成績分佈比較

在上個部分中我們可以看到同一受試者在兩個題本的表現差異，接下來將以折線圖的方式呈現數據，藉此討論兩題本在測驗表現的整體趨勢。

從圖 4-1 及圖 4-2 可觀察到，AI 題本的整體成績分布較為集中，受試者普遍表現良好，分數較穩定且整體偏高，顯示該題本可能整體相對容易，使大多數受試者能取得較高分數。相較之下，官方題本的分數變異幅度較大，部分受試者成績顯著偏低，顯示其題目在難度差異設計上較具有起伏，能有效反映受試者間的能力差異。這樣的結果顯示，官方題本應具備較佳的鑑別度，能夠辨識不同語言程度的學習者，而 AI 題本則可能在設計上偏向過於簡單，尚需進一步調整題目難度，才能提升區分不同程度學習者的鑑別力。

圖4-1

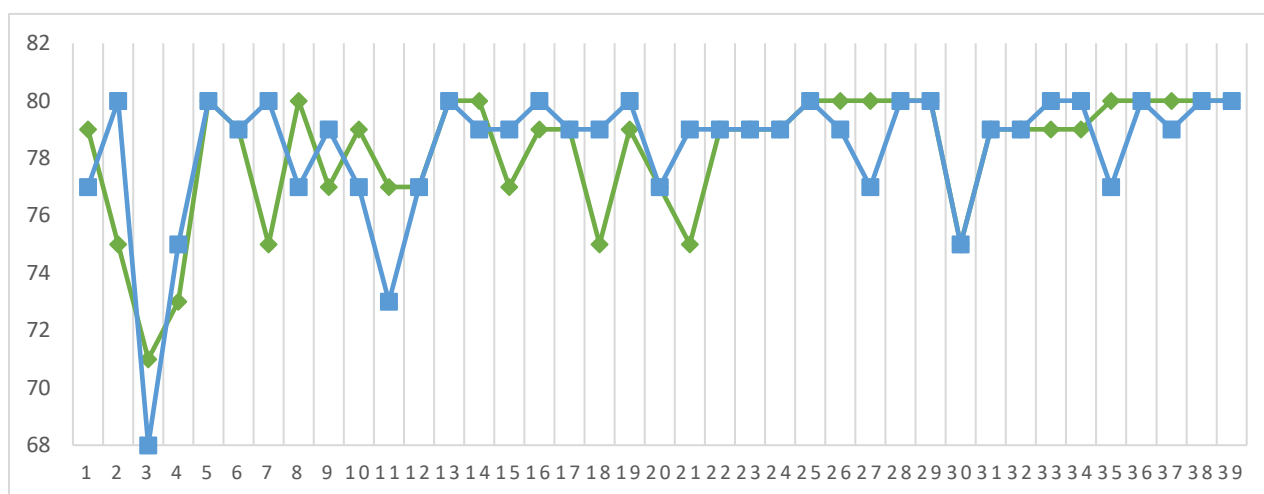
閱讀測驗兩題本的成績分佈圖



註：本圖顯示 39 名受試者的閱讀測驗表現，橫軸為學生編號（1-39），縱軸為成績（60-80）。綠色（◆）為人工考題，藍色（■）為 AI 考題。

圖4-2

聽力測驗兩題本的成绩分佈圖



註：本圖顯示 40 名受試者的聽力測驗表現，橫軸為學生編號（1-39），縱軸為成績（68-80）。綠色（◆）為人工考題，藍色（■）為 AI 考題。

五、 兩題本的難易度分析—答錯率的比較

為更清楚比較 AI 題本與人工題本在難易度上的差異，本研究進一步統計兩組題本中每一題的答錯率，並以直條圖呈現答錯率分布情形，以視覺化方式輔助分析題目在整體難易度上的表現。

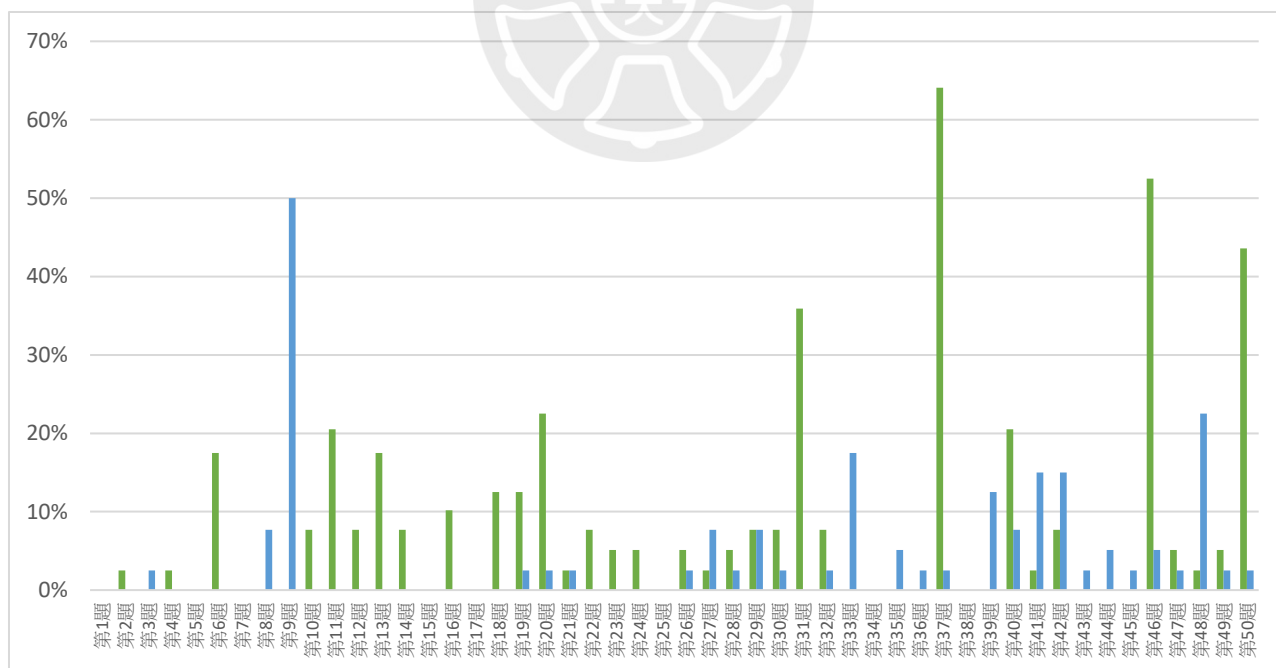
從圖 4-3 的閱讀測驗答錯率的分布趨勢來看，人工考題幾乎每題答錯率皆非零，且其中部分題目的答錯率明顯偏高，顯示其題目難易度有一定變化，能有效區分不同能力層次的考生。相較之下，AI 考題的答錯率分布較為極端，答錯率為 0 的題目佔多數，如聽力測驗 50 題中有高達 34 題為零答錯率；而閱讀測驗 50 題中也有 22 題為答錯率零的題目，僅有少數題目出現較高的答錯率。此種分布現象可能反映出 AI 題目的答案設計較為直接或明確，或是選項誘答力不足，整體試題難度偏低，導致無法充分區分出不同能力等級的受試者。由此可見，與人工題本相比，在鑑別度方面

AI 題本尚有改進空間，需再針對答對率過高的題目進行難度調整，以提升測驗對受試者語言能力差異的辨識能力。

而從圖 4-4 所呈現的聽力測驗答錯率分布來看，AI 題本與人工題本在多數題目的答錯率皆接近零，顯示整體題目對受試者而言難度偏低，大部分受試者皆能順利作答。然而，值得注意的是，AI 題本中出現數題答錯率顯著偏高的情況，與整體表現形成明顯對比。在題本整體難度不高的前提下，這些高錯率題目可能反映出題目設計上的特殊狀況，例如選項間的區辨度不足、題幹語言表達不夠清晰，或是試題有效檢測出受試者尚未熟練的語言點。因此，為進一步釐清這些題目錯誤率偏高的原因，並評估 AI 題本是否具備檢測學習弱點的潛力，本研究將於下節深入探討相關題目的設計特徵與可能影響因素。

圖4-3

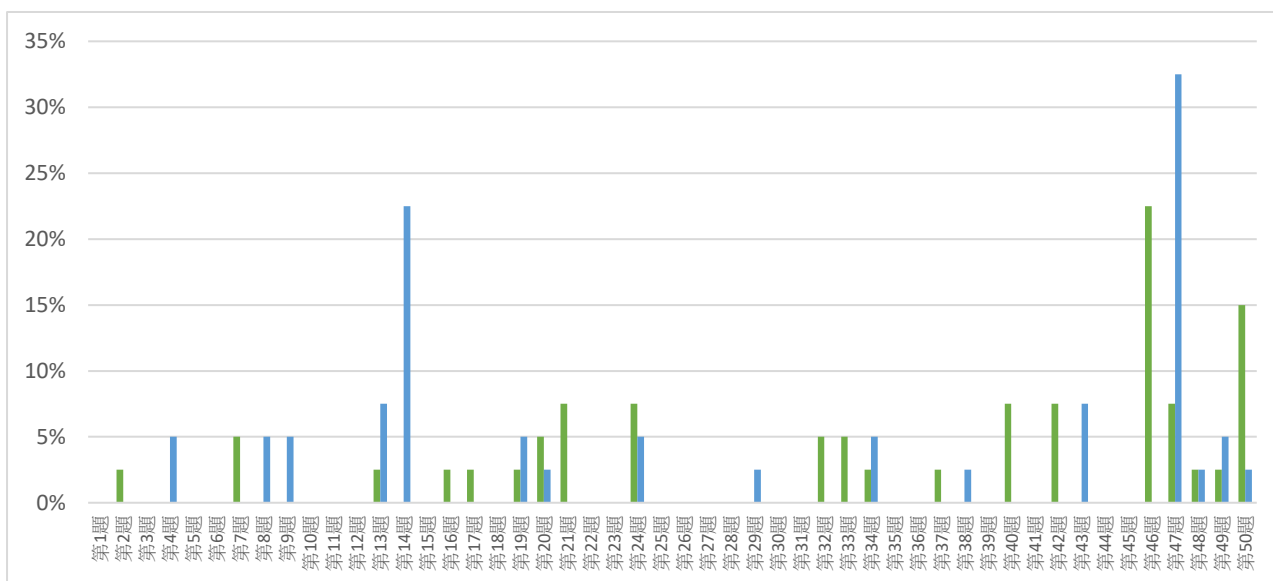
閱讀測驗兩題本的每題答錯率分佈圖



註：縱軸為答錯率，橫軸為題號，共 50 題。綠色為人工題本；藍色為 AI 題本。

圖4-4

聽力測驗兩題本的每題答錯率分佈圖



註：縱軸為答錯率，橫軸為題號，共 50 題。綠色為人工題本；藍色為 AI 題本。

貳、 作答行為分析

為了進一步探討 AI 題本在測驗診斷功能上的潛力，本研究將高於平均答錯率的題目視為高答錯率題目，嘗試分析高答錯率的題目，觀察這些題目是否具有某些共通特徵，如語言點、題型設計或提問層次等。若能發現特定類型的題目特別容易造成受試者錯誤，則顯示 AI 題本具備與人工題本相似的能力，能有效辨識考生的學習弱點，進而提升其在教學應用上的實用性。

一、 閱讀測驗的高答錯率題目分析

在閱讀測驗中，AI 題本的平均答錯率為 7.7%，高答錯率題目有六題，分別為第 9 題（答錯率 50%）、第 33 題（答錯率 17.5%）、第 39 題（答錯率 12.5%）、第 41 題（答錯率 15%）、第 42 題（答錯率 15%）、第 48 題（答錯率 22.5%）。人工題本的平均答錯率則為 13.7%，高答錯




率題目有九題，分別為第 6 題(答錯率 17.5%)、第 11 題(答錯率 20.5%)、第 13 題(答錯率 17.5%)、第 20 題(答錯率 22.5%)、第 31 題(答錯率 35.9%)、第 37 題(答錯率 64.1%)、第 40 題(答錯率 20.5%)、第 46 題(答錯率 52.5%)、第 50 題(答錯率 43.6%)。下表將呈現閱讀測驗 AI 題本與人工題本的高答錯率題的題目及選項，以及推測造成高答錯率的可能原因。

從表 4-8 的錯題比較來看，AI 題本約能達到人工題本的品質，能辨識出部分受試者未掌握的語言點與語用能力，具有檢測學習弱點的潛力。如 AI 題本的第 48 題為篇章主旨理解題，檢驗出受試者容易因細節判斷錯誤，顯示其在整合與歸納上的能力仍有不足；這與人工題本中第 50 題(段落觀點誤判)所顯現的診斷功能相似。這些題目反映出 AI 題本能有效識別出受試者推理判斷能力低落、段落整合困難等不同類型的學習弱點。顯示 AI 題本在測驗診斷上具有一定的潛力。

儘管 AI 題本具備學習診斷的潛力，分析結果也顯示其在出題上的問題。首先，在語言難度控制方面，AI 題本多次出現超出入門基礎等級的詞彙，如第 9 題的「會議」、第 33 題的「蔬菜」、第 39 題的「慶祝」皆為第 4 級詞彙，超出了 TBCL 第二與三級的測驗設定，且超綱詞又屬判斷該題的關鍵字詞，導致受試者誤答的情況。其次，題幹設計有時語境提示不足，例如第 41、42 題的情境順序不夠明確，讓學生難以從上下文推理事件先後，影響答題準確度。最後，圖像與文字之間的關聯度也有待提升，如第 33 題圖中麵包畫面過於明顯，干擾了對「蔬菜區」的判斷。這些設計上的問題顯示 AI 在試題語料選擇、語境建構的精準度及選項的設計上仍需調整，才能更貼合測驗目的。

表4-8


閱讀測驗的高答錯率試題分析

題本類型	題號	題目題幹	PIRLS 類型
AI 題本	9	<p>9. 今天早上，張先生開了一個重要的會議，跟老闆介紹了新計畫。</p> <p>(A)  (B)  (C) </p>	直接提取
<p>所有答錯受試者中，選項 A 的選擇率為 100%。(正解為 B)</p>			

第 9 題的高答錯率可能來自多重因素。首先，題幹中的關鍵詞「會議」為 TBCL 第四級詞彙，已超出本測驗的目標等級範圍，部分受試者可能尚未學習該詞，難以準確理解題意。其次，選項 A 圖片中兩人正在交談並手持文件，畫面呈現出某種溝通情境，容易被誤認為是開會，干擾性強。再者，題幹語境提示不足，未提供能幫助學生辨識「會議」情境的線索，如「簡報」、「報告」、「主持人」等關鍵詞或語意補足語，使考生無法將聽到的內容與面對聽眾報告的情境準確連結。

題本類型	題號	題目題幹	PIRLS 類型
AI 題本	33	<p>33. 有些人在蔬菜區挑 _____，看看是不是新鮮的</p> <p>(A) 麵包 (B) 水果 (C) 青菜</p>	直接提取
<p>所有答錯受試者中，選項 A 的選擇率為 43%；選項 B 的選擇率為 57%。(正解為 C)</p>			

第 33 題的高答錯率可能與詞彙難度與圖像干擾有關。首先，關鍵詞「蔬菜區」中的「蔬菜」為第四級詞彙，已超出本測驗設定的範圍，考生若未學過該詞，可能無法正確理解題幹內容，進而選擇已學過且常出現在超市的詞「水果」。其次，儘管正確選項為 C，符合「蔬菜區」的情境，但圖片左側出現了明顯的麵包圖像，考生容易被視覺焦點吸引，誤判為應選擇 A「麵包」。

題本類型	題號	題目題幹	PIRLS 類型
AI 題本	39	 <p>39. 開會的時候，同事們有時候聽，有時候低頭 _____</p> <p>(A) 吃飯 (B) 寫字 (C) 慶祝</p> <p>所有答錯受試者中，選項 C 的選擇率為 100%。(正解為 B)</p>	直接推論


第 39 題的誤答可能與詞彙難度與圖片理解有關。雖然題幹與圖片皆呈現「開會」的情境，且三個選項中僅有 B 符合畫面中桌上的細節，因此為正確答案，但題幹中的關鍵詞「慶祝」為第四級詞彙，超出本測驗設定範圍，可能造成部分考生無法理解語意，進而誤選其他選項。

題本類型	題號	題目題幹	PIRLS 類型
AI 題本	41 42	<p>新學期開始了，有一位新同學來到教室門口，<u> (41) </u>。她跟老師說明自己想學中文，<u> (42) </u>。老師讓她介紹自己，<u> (43) </u>，說自己是從東部搬來的。她也說了自己的興趣，<u> (44) </u>。大家都很喜欢這位新同學，<u> (45) </u>。</p> <p>(A) 她喜歡拍照，有很多鳥和貓的照片 (B) 她以前學了兩年多的中文，現在還想繼續學 (C) 同學們都覺得她是一個很有趣的人 (D) 她跟大家說了自己的名字 (E) 老師熱情地歡迎她進來 (F) 老師一點也不歡迎她進來</p> <p>41 題：所有答錯受試者中，選項 B 的選擇率為 100%。 (正解為 E) 42 題：所有答錯受試者中，選項 E 的選擇率為 100%。 (正解為 B)</p>	直接推論


此題組的錯誤率偏高，可能與語境連貫性與選項事件的時間邏輯有關。第 41 與 42 題分別描述老師歡迎新同學與新同學說明想學中文的原因，依照 AI 命題時的思路，應先由老師歡迎，接著才是新同學回應，故正確答案應為 E（老師歡迎）接 B（新同學說明）。然而，在現實語境中，兩者事件順序交換似乎也說得通；再加上 AI 題目題幹情境描述較為簡略，未能明確交代事件的順序關係，使得考生在缺乏清楚語境引導的情況下，較難正確推斷事件的邏輯先後，進而產生誤答。

題本類型	題號	題目題幹	PIRLS 類型
AI 題本	48	<p>(三)</p> <p>最近，很多人關注環境問題。科學家說，空氣污染和水污染是很大的問題。如果我們不愛護環境，地球上的動物和植物可能會慢慢消失。有些城市開始禁止塑膠袋，鼓勵人們用環保袋。人們說：「雖然這樣不方便，但我們的環境可以變好。」現在，越來越多人支持這些改變，也希望地球變得更乾淨。</p> <p>48. 這個文章想告訴我們什麼？</p> <p>(A) 空氣污染和水污染是唯一需要解決的環境問題。</p> <p>(B) 只要每個城市不用塑膠袋，地球上的動物和植物就會很安全。</p> <p>(C) 我們只需要使用環保袋，就能完全解決環境問題。</p> <p>(D) 為了保護環境，我們應該支持對環境好的生活方式。</p> <p>所有答錯受試者中，選項 A 的選擇率為 55%；選項 B 的選擇率為 22%；選項 C 的選擇率為 22%（正解為 D）</p>	詮釋、整合 訊息




此題考驗受試者是否能掌握整篇文章的主旨。選項 D 是文中主要傳達的核心理念，但半數受試者卻誤選了選項 A；另外一半的受試者則平均分佈在選項 B 和 C。推測可能是因為誘答選項的句子內容在文中曾以相似句子出現，使考生誤以為其為正確答案。此現象反映出部分學生在處理篇幅較長的閱讀文本時，易受局部資訊干擾，難以整合整體內容來歸納出文章重點，顯示其在較高層次的閱讀理解（如整合與詮釋）方面仍有待加強。這題有助於辨識學生在篇章理解能力上的學習弱點。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	6	6. 小心！你杯子裡的水快要滿了！ (A) (B) (C)	直接推論
			
<p>所有答錯受試者中，選項 A 的選擇率為 100%。(正解為 C)</p>			


人工題本第 6 題關鍵語法為「快要……了」，屬於等級三的語法點，符合本測驗的等級範圍。然而，多數受試者誤選為選項 A，顯示受試者對「快要」所表達的即將發生的語意掌握不足。此結果顯示，該題能有效檢測出受試者在語法上的理解弱點。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	11	11. 妹妹看完信以後，心情很愉快。 (A) (B) (C)	直接提取
			
<p>所有答錯受試者中，選項 A 和 C 的選擇率各佔 50%。(正確答案為 B)。</p>			

第 11 題句中關鍵詞「信」與「心情」皆為第 3 級，符合入門基礎級範圍，但「愉快」屬於第 4 級詞彙，對部分受試者而言可能尚未熟悉，導致理解情緒表達時出現困難。另外，句中的「看完」為第 3*級語法 (V 完)，略超出測驗範圍，部分受試者可能無法正確理解語句中的「看完信」與情緒變化的關係。


題本類型	題號	題目題幹	PIRLS 類型
人工題本	13	<p>13. 因為媽媽怕高，所以沒和爸爸一起爬過山。</p> <p>(A)  (B)  (C) </p>	直接推論
<p>所有答錯受試者中，選項 A 的選擇率為 71%；選項 B 的選擇率為 29%。(正解為 C)</p>			

這題的語言難度符合入門級水準，關鍵詞如「怕」、「高」、「爬山」以及語法「過」皆屬於測驗等級範圍內，故可排除因詞彙或語法超程度導致答錯的可能。然而，題幹情境與選項圖像並非直接對應，而需經過一定的推理解。考生需從句中「因為媽媽怕高，所以沒有和爸爸一起爬過山」推論出正確圖像應為只有爸爸與他人爬山、媽媽缺席的畫面。若考生只注意「爬山」這一動作，容易誤選了選項 A 中登山者的單一畫面，導致答錯。可見此題除了語言理解外，還需整合文本內容並進行推理判斷，因此能有效檢測出學生在語境理解與圖文推理能力上的強弱。


題本類型	題號	題目題幹	PIRLS 類型
人工題本	20	<p>20. </p> <p>(A) 小吃店九點以前休息。 (B) 小吃店十二點開始休息。 (C) 十二點以後可以去小吃店吃飯。</p>	詮釋、整合 訊息
<p>所有答錯受試者中，選項 B 的選擇率為 89%；選項 A 的選擇率為 11%。(正解為 C)</p>			

第 20 題的關鍵詞「以前」、「以後」、「休息」以及表示時間的「__點」皆屬於等級 1 詞彙，符合測驗的程度，故可排除詞語難度過高的可能。從圖片可知，小吃店的營業時間為中午 12 點以後，考生需將


圖中「9:00~12:00 休息」資訊正確理解並轉換為文字，判斷選項是否符合實際意思。然而，三個選項雖然語言結構簡單，但與圖示資訊相反，部分考生可能未充分理解時間邏輯。另外「以前」與「以後」一向是初學者易混淆的語義對應詞，皆可能為造成誤答的原因。故此題能有效檢測學生在基礎時間語義判斷與語意推理能力上的掌握情形。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	31	 <p>31. _____ 著眼鏡的小女孩在看書。 (A) 穿 (B) 帶 (C) 戴</p> <p>所有答錯受試者中，選項 B 的選擇率為 100%。(正解為 C)</p>	直接提取

此題關鍵詞如「戴」、「穿」、「帶」以及「眼鏡」皆屬於測驗等級內詞彙，語言難度本身並不超出受試者的程度。從圖片與句子內容可推知，「戴著眼鏡的小女孩在看書」為正確答案應選擇 C「戴」。然而，多數受試者誤選 B「帶」，顯示出此類同音異義詞（如「帶」與「戴」）對學習者來說具高度混淆性。這類題型能有效檢測出學生在近音詞辨識與實際語境運用上的學習盲點。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	37	 <p>37. 大家都 _____ 她慶祝。 (A) 幫 (B) 讓 (C) 對</p> <p>所有答錯受試者中，選項 B 的選擇率為 44%；選項 C 的選擇率為 56%。(正解為 A)</p>	直接提取

此題圖文明確，圖片中清楚呈現出慶祝生日的場景，與題目句子的語境高度一致，考生幾乎無需進行推理即可作答。然而，從語言難度層面觀察，「幫」雖為等級 2 詞彙，屬於測驗範圍內，但「讓」與「對」為等級 3，「慶祝」更為等級 4，皆超出入門基礎級的詞彙範圍。語言難度可能導致學生無法正確理解句子語意與語法。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	40	 <p>40. 她希望明年能 _____ 德國去玩。 (A) 到 (B) 去 (C) 來</p> <p>所有答錯受試者中，選項 B 的選擇率為 63%；選項 C 的選擇率為 37% (正解為 A)</p>	

第 40 題語言難度皆落在入門級與基礎級等級範圍內，關鍵詞如「希望」「明年」「去」與「德國」等均為常見詞彙，題幹也不需依賴圖像推理，所需的資訊皆可從語句本身獲得。另外，本題的關鍵語法結構為「到＋地點＋去＋動詞」，為等級 2*語法，亦無超出範圍。因此根據答錯情況，顯示受試者可能尚未熟悉該語法的搭配使用。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	46	<p>(一)</p> <p>曾有一項調查發現，很多員工生病的時候不敢請假，因為他們擔心老闆會不高興，覺得他們沒有責任感。有人認為，員工會這麼想是公司的責任。一個好的公司應該能照顧員工，而不是讓他們拿健康去換錢。因此，讓員工有幸福感，應該是未來企業努力的方向。</p> <p>46. 這篇文章說了什麼內容？</p> <p>(A) 老闆應該給員工多一點兒假</p> <p>(B) 常關心別人的人更有責任感</p> <p>(C) 對公司有意見要勇敢說出來</p> <p>(D) 照顧身體比認真工作更重要</p> <p>所有答錯受試者中，選項 A、B 和 C 的選擇率各佔 48%、19%、33%。(正解為 D)</p>	詮釋、整合 訊息

這題為篇章理解題，考驗的是考生整合段落重點、推論作者主旨的能力，整體語言難度多在等級 1、2，亦使用了少數 4、5 級的詞。從錯誤選項分析可見，受試者容易受段落前半段文字影響，誤以為重點是「老闆應該給員工多一點兒假」(A) 或「鼓勵員工說出來」(C)。這代表受試者可能聚焦在某一句話的細節，未能綜合段落整體的語意，忽略最後一句「讓員工有幸福感，應該是未來企業努力的方向」的主旨句。

題本類型	題號	題目題幹	PIRLS 類型
人工題本	50	<p>(五)</p> <p>以前，電影院的門口常常出現一個牌子，要觀眾別帶外面買的食物進去，如果想吃東西，只能買電影院裡賣的食物。後來，新聞說，觀眾其實可以拒絕配合這些電影院的規矩。說到這個問題，我認為，電影院的要求不是沒有道理，因為電影結束後，他們還得打掃那些垃圾。其次，如果有人帶了一些有奇怪味道的食物進電影院，也很容易影響其他看電影的人。</p> <p>50. 作者對在電影院裡吃東西的行為怎麼看？</p> <p>(A) 應該要避免吃有特別味道的東西</p> <p>(B) 覺得吃什麼東西都是個人的自由</p> <p>(C) 認為電影院的要求沒有任何道理</p> <p>(D) 看電影的人都有責任要打掃垃圾</p> <p>所有答錯受試者中，選項 C 和 D 的選擇率各佔 59% 和 41%。(正解為 A)</p>	詮釋、整合 訊息

第 50 題語言難度大致皆屬於測驗等級內，選項 C 和 D 中的部分描述曾出現在文章中，導致多數受試者誤選 C 或 D。這顯示了受試者在辨別作者觀點時，容易將文中提及的其他觀點誤認為作者立場，未能準確整合與歸納。這也說明該題能有效檢測受試者是否具備篇章閱讀的整合與判斷能力。

二、聽力測驗的高答錯率題目分析

在聽力測驗中，AI 題本的平均答錯率為 7.3%，高答錯率題目有三題，分別為第 13 題（答錯率 7.5%）、第 14 題（答錯率 22.5%）、第 47 題（答錯率 32.5%）。人工題本的平均答錯率則為 5.9%，高答錯率題目有兩題，第 46 題（答錯率 22.5%）、第 50 題（答錯率 15%）。表 4-9 將呈現聽力測驗 AI 題本與人工題本的高答錯率題的題目及選項，以及推測造成高答錯率的可能原因。

從表 4-9 的錯題題型分析來看，與閱讀測驗相比，AI 題本在聽力測驗上，更具備辨識學習者語言弱點的潛力。如 AI 題本中的第 13 題、第 14 題和第 47 題皆出現較高答錯率，這些題目能反映出考生在語意推理、空間語言理解、基本詞彙應用與對話內容整合等能力上的不足；這些錯誤模式與人工題本中出現的錯誤情形（如第 46 題和第 50 題）

高度相似。因此 AI 聽力題本在學習診斷功能上與人工題本具有相似性，並且能有效揭露學習者的語言理解弱點。


雖然 AI 題本在聽力測驗上已有相當不錯的表現，但出題上仍與閱讀測驗有相同的問題，如：語言難度控制不夠精確（第 13 題的「棵」）、語境提示不充分（第 47 題的語意推論）等的狀況。

表4-9

聽力測驗的高答錯率試題分析

題本類型	題號	題目	PIRLS 類型
AI 題本	13	 <p>腳本： 大樓旁邊有什麼？ A. 一棵樹。 B. 一台電視。 C. 一隻狗。</p> <p>所有答錯受試者中，選項 B 的選擇率為 100%（正解為 A）</p>	直接提取

第 13 題圖片中清楚呈現「樹」，多數受試者仍誤選「一台電視」，故排除該錯誤因圖片干擾的可能。推測造成此現象的原因有二：1.量詞「棵」為第四級等級詞彙，部分受試者對其不熟悉。2.學生可能對空間語言「旁邊」的理解仍不夠熟練，未能將聽力內容與圖像正確對應。

題本類型	題號	題目	PIRLS 類型
AI 題本	14	 <p>腳本： 李先生的職業是什麼？</p> <p>A. 護士 B. 老師 C. 工人</p> <p>所有答錯受試者中，選項 A 的選擇率為 100%。 (正解為 C)</p>	直接提取

第 14 題圖片設計與選項安排都清楚合理，圖中元素如安全帽、藍圖、工具等能清楚傳遞「工人」的職業意象；而「護士」與「工人」也都屬於 A1~A2 等級的常見職業詞彙，因此並無明顯用詞難度或圖片誤導的問題。受試者多誤選「護士」的情況，可能反映出他們對基本職業詞彙的辨識能力仍有不足，或是平時在語言輸入中較少接觸職業名詞。

題本類型	題號	題目	PIRLS 類型
AI 題本	47	<p>47.</p> <p>(A) 她覺得準備考試很簡單。</p> <p>(B) 她需要更多時間準備數學考試。</p> <p>(C) 她已經準備好所有的考試了。</p> <p>(D) 她不喜歡考數學，所以沒準備。</p> <p>腳本：</p> <p>男：下週要考試了，你準備得怎麼樣？</p> <p>女：我這幾天一直在看書，但題目有點多。</p> <p>男：那你覺得哪一個比較難？</p> <p>女：數學比較難，我覺得很不容易學。</p> <p>男：別擔心，慢慢準備就好。</p> <p>男：這位小姐的意思是什麼？</p> <p>所有答錯受試者中，選項 A 的選擇率為 8%；選項 C 的選擇率為 92%。(正解為 B)</p>	直接推論

第 47 題的設計整體來說沒有問題，考點明確，考生需整合對話內容，判斷女方對數學準備的態度與需求。從誤選情況來看，透露出考生未能從女方「數學很難」推論出「需要花更多時間準備數學」這一層含義，顯示在資訊整合與語意推理方面能力尚待加強。

題本類型	題號	題目	PIRLS 類型
人工題本	46	<p>46.</p> <p>(A) 她要兒子再找找</p> <p>(B) 課本不在書桌上</p> <p>(C) 她把課本收起來了</p> <p>(D) 書桌上放了本字典</p> <p>腳本：</p> <p>男：媽，你看見我的課本嗎？</p> <p>女：書桌那裡沒有嗎？</p> <p>男：找過了，都沒有。等一下就要上課了，真糟糕。</p> <p>女：不是放在書桌上，就是跟字典放在一起，你去看看。</p> <p>男：這位太太的意思是什麼？</p> <p>所有答錯受試者中，選項 B 和 D 的選擇率各佔 44% 和 56%。(正解為 A)</p>	詮釋、整合訊息

人工題本第 46 題整體語言難度皆在等級內，僅有「糟糕」為第五級，但並非答題關鍵字，故應不影響受試者判斷。多數受試者選擇了 B 和 D，可能原因如下：首先，選項 D「書桌上放了本字典」雖然正確描述了對話中的一部分資訊，但未能完整表達出女方的意圖，因此容易成為誤導選項；其次，選項 B「課本不在書桌上」則與對話中的推論相矛盾，反映部分受試者可能未能理解對話的隱含訊息。故這題反映了學生對於語境理解層次的弱點。

題本類型	題號	題目	PIRLS 類型
人工題本	50	50. (A) 沒收到衣服 (B) 沒錢吃東西 (C) 沒錢買衣服 (D) 請母親寄衣服	直接推論
腳本：			
男：喂，媽，請寄一些錢給我，我要買衣服。			
女：衣服不是都帶去了嗎？為什麼還要買？			
男：我變胖了，那些衣服都穿不下了。			
女：你才去念了半年的書就變胖了，我想那裡的東西一定很好吃。			
男：這個兒子打電話跟母親說什麼？			
所有答錯受試者中，選項 A 的選擇率為 16% ；			
選項 D 的選擇率為 84%。(正解為 C)			

這題的語言難度皆在等級 1、2，故排除詞語太難的可能。然而，多數受試者選擇了 D，可能原因在於對話中提到了「衣服穿不下」這個資訊，導致考生直覺聯想到需要「衣服」，進而誤選 D。這反映出部分考生雖能抓到關鍵詞，卻未能正確理解整體語意與人物行動的關聯。故這題也反映了受試者對語境掌握的不足。

三、 小結

本研究結果顯示，AI 題本在用詞難度方面與人工題本整體相近，然而在實際分析中可觀察到 AI 題本中較常出現中高等級詞彙，如表 4-1 中第一部分題型裡，AI 題本使用等級 3 到等級 6 的比例約 11.4%，遠高於人工題本的 4%，顯示其在語言難度的控制上仍有進一步調整的空間。雖然 AI 題本在提問類型的分佈上已盡可能貼近人工題本，並

力求維持題型與認知層次的一致性，受試者的答題表現仍反映出兩者之間的差異。具體而言，AI 題本的閱讀測驗平均分數普遍高於人工題本，顯示其整體難度可能略顯偏低；聽力測驗部分則呈現分數相近的情況，僅在區辨不同能力組別受試者方面，尤其是對中能力組的區辨效果，仍有待進一步檢視。此外，AI 題本整體成績分布較為集中，多數受試者表現穩定且得分偏高，再次驗證其題目難度可能偏低的觀察；相較之下，人工題本的分數變異幅度較大，反映出其題目難易度設計較具挑戰性，鑑別度亦相對較佳。從答錯率分析亦可看出類似趨勢，人工題本呈現出較明顯的難度差異，有助於辨識不同程度的學習者；而 AI 題本則較常出現答錯率集中於特定範圍的現象，顯示其鑑別度尚待提升。儘管如此，AI 題本在診斷學習弱點方面已展現與人工題本相當的潛力，僅在語言難度控制、語境提示設計與圖文對應關聯性等細節上，仍有進一步改善的空間。



第五章 結論與建議

第一節 研究發現與結論

本研究旨在探討生成式人工智慧運用於華語文能力聽力與閱讀測驗試題編寫之可行性，並以 TOCFL 題型為參照，評估 AI 試題在語言難度、題型結構與認知層次等方面的整體表現與實用潛力。

首先本研究以 TOCFL 試題為參考依據，訓練 AI 模型以生成 AI 試題題本，並分析 AI 試題上的表現與限制；接著透過 AI 試題與 TOCFL 人工試題的比較，進一步檢視 AI 生成的試題題目在難度、區辨度及認知層次上的表現。最後，從第四章的分析結果已呈現出 AI 試題與人工試題在語言難度、測驗題型結構、提問方式及受試者答題表現上的具體差異，顯示了 AI 所生成題目的應用價值與潛力。接下來，為回應第三章所提出之研究問題，以下將依據第四章的研究結果，分別從命題流程與受試者作答表現兩個面向，說明 AI 試題的整體表現，再進一步探討其在認知層次上的分布與特徵，並統整 AI 試題於語言教學測驗上的應用潛力與限制。

壹、研究問題之回應

一、同一受試者在 AI 試題與人工試題的測驗成績差異

為探討 AI 命題與人工命題的表現差異，本研究透過對同一受試者在兩份試題（人工與 AI）間的分數差異進行分析，並考量不同能力組別（低、中、高）的成績表現特徵與統計顯著性（配對樣本 t 檢定結果），以提供綜合性的比較與評估。

在表 4-6 與表 4-7 的資料分析中，顯示了 AI 命題在聽力測驗部分展現了高度的平均分數一致性與穩定性。各能力組的配對樣本 t 檢定結果均未達統計顯著水準（低能力組 $T = -0.24, p > 0.05$ ；中能力組 T

= -0.91, $p > 0.05$ ；高能力組 $T = 1.31, p > 0.05$ ），這表明 AI 聽力試題與人工題本在各能力層級的平均分數表現上幾乎無統計顯著差異。尤其在高能力組，極低的分數差標準差（1.147）進一步驗證了 AI 聽力題本的高度穩定性，但也提示可能存在天花板效應，未能有效區分高能力學習者之間的細微差異。

相對地，在閱讀測驗部分，AI 命題則顯現出一些不同特徵。儘管各組分數差的平均值均顯示 AI 試題分數普遍高於人工試題，但其統計顯著性與變異性因能力組別而異。中能力組的 t 值為 -3.78 ($p < 0.05$)，高能力組的 t 值為 -2.85 ($p < 0.05$)，兩者均達到統計顯著性，表明 AI 閱讀試題整體難度可能較低，導致這兩個組別的受試者普遍獲得較高分數。然而，閱讀測驗低能力組的 t 值為 -1.17 ($p > 0.05$)，未達統計顯著水準，且低能力組（4.932）與中能力組（5.241）的分數差標準差較大，顯示 AI 閱讀題本對這些能力組別的影響存在較大的不一致性。

二、 AI 試題之難易度、鑑別度與 TOCFL 試題的符合度

AI 試題在聽力與閱讀測驗中，其難易度、與 TOCFL 入門基礎級(A1-A2) 的符合度，以及在區分不同能力程度考生方面的鑑別度，均呈現出不同的情形。

在聽力測驗部分，AI 試題展現出高度的難易度符合性與整體穩定性。各能力組別（低、中、高）的平均分數與人工題本皆高度相近，配對樣本 T 檢定結果均未達統計顯著水準（低能力組 $T = -0.24$ ；中能力組 $T = -0.91$ ；高能力組 $T = 1.31$ ），這表明 AI 聽力試題在各能力層級的平均分數表現上，與官方題本幾乎無統計顯著差異，符合 TOCFL 入門基礎級的難度設定與語言設計原則。AI 在設計資訊明確、語句簡潔的聽力題型上展現了穩定性，分數差標準差普遍較低（低能力組

3.697，中能力組 2.333，高能力組 1.147)。然而，這種高度的一致性和分數集中現象，也同時指出其在區分不同能力程度考生方面的鑑別力有限，尤其在高能力組，極低的標準差和 t 值未能顯著區分出細微的能力差異，提示可能存在天花板效應。

相對之下，閱讀測驗的 AI 試題則顯現出較多的評量功能挑戰。研究結果顯示，AI 閱讀題本的整體難度顯著偏低，其平均分數(76.13 分)明顯高於官方題本(72.63 分)，且中能力組 ($t=-3.78, p<0.05$) 和高能力組 ($t=-2.85, p<0.05$) 的 t 值均達到統計顯著水準，進一步證實 AI 閱讀試題對這兩個組別而言普遍過於簡單。這種整體難度偏低的特性，直接導致了其鑑別度明顯不足。

三、 小結

總體來看，AI 命題雖然在模仿測驗格式與基本語言樣式上具備潛力，但在精準控制語言難度與確保有效鑑別度方面仍面臨挑戰。特別是 AI 在生成題目時，仍偶爾出現超出設定等級的詞彙或語法結構，這直接影響了個別題目的難易度與其在語言能力程度上的鑑別力；此外，語境設計過於簡略也可能影響考生的推理判斷。這些問題皆指向當前 AI 模型在未經特定語言等級調整下，難以完全取代人類命題者在掌握語言難度、選項誘答力與整體邏輯合理性上的判斷與經驗。這也正是傳統命題流程中需透過多輪人工編修與審查的重要環節 (TOCFL, 2021)。在先前的研究如 Singh 2023 和 Zhang 2023 也皆指出，AI 在生成試題時雖能有效掌握表面語言結構，但在較高層次的選項合理性與語境邏輯上仍有待發展，尚需靠人工後續再檢驗修正。

為提升 AI 試題的評量功能與鑑別效度，未來研究應著重於針對特定測驗等級開發專屬 AI 模型或進行微調 (fine-tuning)，並結合自動詞彙與語法分級工具，以期達到更精準的語言難度控制。同時，仍需強調

人工審核與修正的重要性，以確保 AI 試題不僅能快速產出，更能符合測驗的嚴謹評量標準，未來若能結合自動詞彙標記與分級工具，將有機會成為教師及教材開發者的輔助命題工具。

貳、生成式 AI 的布魯姆 (Bloom) 認知層次對照分析

在完成 AI 生成試題的分析後，接下來將探討其在不同認知層次上的展現情形。本研究參考布魯姆 (Bloom) 修訂版認知領域分類理論，從記憶 (Remember)、理解 (Understand)、應用 (Apply) 至分析 (Analyze)、評鑑 (Evaluate) 及創造 (Create) 等層次，觀察 AI 在生成試題時所展現的認知處理上之表現。

從語言模型的訓練特性來看，AI 在記憶 (Remember)、理解 (Understand) 以及應用 (Apply) 等層次上已展現出相對穩定的表現。本研究所使用的 GPT-4 模型，具備透過大規模語料學習自然語言模式與語境邏輯的能力，使其在文本改寫、資訊統整等語言任務中表現出高度準確性。基於此特性，AI 能根據參考題本，有效統整並模仿各類題型的出題方向、測驗重點與語境情境，生成結構完整、語言清晰的聽力與閱讀試題。舉例而言，AI 能迅速辨識人工題本中所強調的資訊類型 (如地點、動作、人物等)、常見的題幹情境 (如職場、購物、廣告等)，以及對應的提問方式 (如直接提取、直接推論等)，並據此設計出相似的題幹與選項。這顯示 AI 在記憶、理解及應用層次上的語言生成能力已具有一定的實用價值。

然而，在分析 (Analyze)、評鑑 (Evaluate) 及創造 (Create) 等較高層次的命題能力方面，AI 的表現仍顯有限。本研究所參考的人工題本中，部分試題涉及語境推論、對話整合與語意隱含判讀等認知要求，對命題者在題幹設計與選項安排上的邏輯性與鑑別力具有高度要求，亦須瞭解學習者

可能出現的理解偏誤。相較之下，AI 雖能模仿出題形式與語用場景，但在內容設計上常出現邏輯關係不明、上下文連貫性不足，或選項誘答力薄弱等問題，影響試題的有效區分功能。此一觀察亦與先前相關研究結果相符。ENABLE (2023)、Singh 等人 (2023) 均指出，生成式 AI 雖擅長語言樣式模仿與文本生成，然在命題情境下，仍常出現選項設計不嚴謹、鑑別度不足等問題，顯示大型語言模型目前尚不具備高階推理與深層認知處理的能力。

總的來說，AI 在記憶、理解與應用層次的展現相對穩定，具備一定實用性；但在較高層次的分析與評鑑能力上仍有限，須依賴人工補強，才能確保試題品質與鑑別力。

參、 結論

本研究為生成式 AI 技術運用於華語文測驗試題編寫上的實證研究，測驗等級設定為 TOCFL 入門基礎等級，以 TOCFL 模擬題本為對照，進行試題生成、施測與分析。本研究透過 GPT-4 模型設計聽力與閱讀試題，並邀請具 A1 至 B2 等級的華語學習者進行施測，進而比較 AI 題本與人工題本在語言難度與受試者表現上的差異。

研究結果顯示，AI 在出題效率與語言形式模仿上已有高度的表現，能有效產出符合 TOCFL 入門基礎級結構的題型；然而，AI 雖能生成形式正確的題目，但其選項干擾性與題目鑑別力不足，導致部分受試者分數過度集中，顯示其區分學習者能力的效度仍有待提升。認知層次分析亦發現，AI 雖能處理基本語言資訊的重組與理解，卻較難以產出涉及推論、整合與評鑑的高層次題目內容，反映出其語意判斷與測驗邏輯掌握的侷限。

綜合而言，生成式 AI 可作為華語教學與測驗命題的輔助工具，特別是

在試題草稿生成方面上具有應用潛力。然而，仍需仰賴人工命題者的審查與調整，才可達成標準化測驗的內容效度要求。未來若能結合 AI 技術與教育專業，建立一套命題機制，將有助於提升華語文測驗的發展效率與創新可能性。

第二節 研究限制

本研究針對生成式 AI 在華語文能力測驗命題上的應用進行初步探討，並以 TOCFL 題型為參照建立試題分析與施測流程，但仍存在若干限制，說明如下：

壹、AI 命題輸出結果具不穩定性

生成式 AI 的試題輸出受其訓練語料庫與使用者所輸入的提示語設計（prompt）所影響。因此本研究所呈現的命題表現僅能反映在特定模型與設定下的生成狀況，尚不足以代表所有生成式 AI 的命題能力，研究結果應作為目前應用情境下的初步觀察與參考。

貳、受試者樣本有限

本研究受試者為同一所大學之越南籍華語學習者，語言背景、學習情境與學習歷程相對一致，且數量僅 39 位，尚無法代表整體華語學習者的普遍情況。若未來能納入來自不同國籍、學習背景與語言程度的學習者，並增加樣本數，將有助於提升樣本的代表性。

參、研究測驗設定僅在單一等級

本研究為初步探討的小規模研究，試題設計上僅先針對 TOCFL 入門基礎級，也就是初級程度的測驗進行分析，尚未涵蓋中高級以上語言能力測驗層面，故未來可增加評估 AI 在高級語言程度上命題能力之研究。

肆、施測情境為非正式考場，可能影響作答表現

本研究施測過程按照 TOCFL 正式測驗的時間與測驗順序，並以教室的廣播系統統一播放聽力測驗音檔，以確保施測程序的統一。然而施測場地為受試者的華語課教室，而非正式的測驗環境，可能導致受試者因監測鬆散、心理壓力較低等因素，影響其作答專注度與真實表現，進而對數據結果造成影響。

第三節 教學建議及未來研究建議

本章欲根據前述研究發現與分析結果，針對華語教學現場、測驗命題實務及 AI 工具應用上，提出相關建議，供教育工作者及未來研究者參考。本章分為兩部分進行說明，分別為教學領域上的建議，以及相關研究的建議。

壹、教學實務

生成式 AI 在華語文測驗的試題編寫上具備初步應用潛力，尤其在語言結構模仿方面表現穩定。然而，試題內容品質、鑑別力與語境適切性仍需人工介入與審查。根據研究結果，提出以下四點建議：

一、將 AI 作為輔助命題的工具，並搭配人工審查以確保品質

AI 可快速生成形式正確的試題，能有效協助命題者在試題初稿撰寫上的工作。然而，由於 AI 在選項干擾性、語意連貫與語用邏輯上的掌握有限，建議命題者仍應結合人工審查與編修程序，以強化生成試題的鑑別力與整體合理性。

二、將 AI 命題工具優先應用於聽力題型與基礎閱讀理解題目

本研究結果顯示，AI 所生成的聽力試題在語句結構、題型模仿與

受試者表現方面，與人工題本表現相近。因此若教學現場欲應用 AI 於命題上，應優先考慮應用於聽力與初級閱讀理解題型，作為課堂補充資源或課後練習。

三、 開發可標示語言等級與提問層次的命題平台

命題者在使用 AI 進行試題編寫時，常面臨語言難度難以掌握且生成的題目提問層次過於單一的問題，進而影響試題的鑑別度。為回應此問題，本研究採用 PIRLS 四層次提問類型作為命題與分析架構，並實際應用於 TOCFL 試題與 AI 試題的分類與比較中，結果顯示該分類方式能有效對應 TOCFL 題型設計的邏輯與層次要求，具備良好的應用潛力。基於此，若未來將開發 AI 命題平台，建議可結合 TBCL 詞彙分級系統與 PIRLS 提問層次標記功能，協助命題者判讀生成內容的語言層級與認知難度，亦可提升 AI 命題的品質與提問深度。例如，平台可設計自動標示每一題所使用詞彙的等級標籤，提醒命題者是否出現超綱詞彙，並提供相近語意的替代詞建議，以協助調整語言難度；同時也可輔以提問層次的判讀功能，供使用者在設計題幹時能掌握認知層次分布，進而提升題組整體的鑑別力與試題多樣性。

四、 命題時需關注語境與文化適切性

AI 雖能生成具生活化語境的題幹，但仍可能產出語境不明確或與學習者經驗脫節的內容。故建議命題者在使用 AI 試題時，應進行語境審查並適時調整，以確保題幹內容符合目標學習者的語言經驗與文化理解範圍。

貳、 未來研究建議

根據前述的研究結果與研究限制，為了讓未來相關研究能更深入探討生成式 AI 在華語文測驗命題上的更多可能性，本研究提出以下幾點研究

建議，供後續未來研究參考。

一、 比較不同語言模型的命題品質

本研究以 GPT-4 為主要生成模型，未來可進一步比較不同大型語言模型（如 Claude、Gemini、Mistral 等）在試題語言風格、內容品質與命題穩定性上的差異，探索不同模型在教育領域中的適配性與限制。

二、 再調整施測題目難度以提升結果區辨力

本研究的研究結果中，AI 與人工聽力題本的受試者得分接近，亦可能與題本難度偏低有關，導致考生普遍獲得高分，影響測驗的區辨效果。因此，未來研究可進一步調整 AI 題目的難度設計，以提升試題的鑑別力與適切性。具體而言，可考慮設計兩組或多組難度分級明確的 AI 題本（如 Band A-C 三種版本），並邀請該程度的受試者進行施測比較，觀察不同難度題本對成績分布與能力區辨的影響；另外亦可針對特定題型（如閱讀理解、圖文對照理解等）進行微調控制，例如調整題幹長度、選項相似度或關鍵詞彙的等級，進一步分析這些變項對 AI 試題區辨力的影響。透過系統化的難度設計與量化比較，將有助於更全面評估 AI 試題在測驗功能上的應用潛力。

三、 擴大受試者樣本數並增加不同語言背景

為提升研究結果的廣泛性與代表性，未來研究可納入不同國籍、不同母語背景、不同學習時長及語言程度之學習者群體，進一步探討 AI 試題對不同背景學習者的適切性與效度。

四、 發展結合 AI 與人工之命題流程

本研究結果顯示，生成式 AI 雖已能模仿華語能力測驗的基本題型形式，但在語言難度控制、提問層次展現與題目鑑別力方面，仍難以完全取代人工命題者之判斷與專業。因此，未來研究可進一步設計並

比較不同的 AI-人工協作命題流程，探討如何有效結合 AI 的生成效率與人類命題者的語言教學經驗，以提升試題品質與應用可行性。具體而言，可嘗試多種協作模式，並研究各模式的優勢與困難。模式一，AI 先於人工：先由 AI 生成題幹與選項，教師再進行難度標註與語境修正；模式二，人工先於 AI：先由教師撰寫題幹架構作為指令語，將指令語交給 AI 以補足選項內容。研究可從試題品質、鑑別度、語言等級控制與認知層次涵蓋等指標進行分析，亦可納入現場教師使用經驗與回饋，以評估不同協作流程的實務運作情形。透過系統化比較與實證分析，將有助於建立一套兼具效率與專業性的 AI-人工協同命題機制，為華語測驗應用開創更多可能。

綜合以上結論與建議，生成式 AI 雖已具備初步模仿語言測驗題型的能力，但在實際教學與測驗應用中，仍需搭配人工專業進行審查與調整。未來若能結合教育者的經驗與 AI 技術的生成效率，發展出一套標準化、可調整的協作命題流程，將有助於拓展華語測驗的多元化與智能化發展潛力。

參考文獻

壹、中文文獻

- 丁振卿、黃紫娟、林昱辰 (2024)。生成式 AI 輔助寫作研究。 *冷凍空調&能源科技*, 146, 24-3。
- 伍柏翰、張雅綺 (2024)。應用生成式 AI 工具於國高中教育之影響與因應策略。 *臺灣教育評論月刊*, 13 (11), 39-44。
- 林穎俊 (2024)。透過生成式 AI 幫助學生成為更好的學習者。 *師友雙月刊*, 644, 56-61。
- 曾文璇、蔡怡珊、林玲英、陳柏熹 (2023)。使用華語文語料庫與能力基準整合應用系統設計 TOCFL 測驗試題。 *華文世界*, 132, 115-123。
- 顏榮泉 (2024)。從認知處理觀點評論生成式 AI 對學習的影響。 *臺灣教育評論月刊*, 13 (3), 144-153。

貳、英文文獻

- Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. *New York: Longman.*
- Creely, E. (2024). Exploring the role of generative AI in enhancing language learning: Opportunities and challenges. *International Journal of Changes in Education.*
- Dipaola, S., Gabora, liane, & Mccaig, graeme. (2018). Informing Artificial Intelligence Generative Techniques Using Cognitive Theories of Human Creativity. *Procedia Computer Science*, 145, 158–168.
- ENABLE. (2023). A comparative study on AI-generated test items and human-crafted questions in English proficiency assessment. *National Chung Hsing University ENABLE Center.*
- Kavak, V. I., Evis, D., & Ekinci, A. (2024). The use of ChatGPT in language education. *Experimental and Applied Medical Science*, 5 (2),

72–82.

- Malloy, T., & Gonzalez, C. (2024). Applying generative artificial intelligence to cognitive models of decision making. *Sec. Cognitive Science, 15*.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). PIRLS 2016 international results in reading. *International Association for the Evaluation of Educational Achievement (IEA)*.
- Oxford University Press. (2021). Addressing the Deepening Digital Divide. Sabatini, J., Kennedy, A., Wry, E., & von Davier, M. (2024). PIRLS 2026 reading assessment framework. In M. von Davier & A. Kennedy (Eds.), *PIRLS 2026 Assessment Frameworks. Boston College, TIMSS & PIRLS International Study Center*.
- Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333 (6043)*, 776-778.
- Liu, Q., Zhang, Y., Ding, S., Lin, Y., Li, J., Wu, X., & Zhang, Y. (2023). *Generative AI for Language Education: A Systematic Review*.
- Taylor, T. (2010). Ceiling effect. In *Encyclopedia of research design (Vol. 0, pp. 133-134)*. SAGE Publications, Inc.
- Tschirner, E. (2016). Listening and reading proficiency levels of college students. *Foreign Language Annals, 49 (2)*, 201–223.
- Ward, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences, 118 (43)*.
- Wu, J. Y., & Tsai, C. C. (2022). Harnessing the power of promising technologies to transform science education: Prospects and challenges to promote adaptive epistemic beliefs in science learning. *International Journal of Science Education, 44 (2)*, 346-353.
- Zhang, Y., Wu, L., & Chen, K. (2023). AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays.

參、線上文獻

國家華語測驗推動工作小組 (n.d.)。華語文能力測驗 (TOCFL)。取自 <https://tocfl.edu.tw/index.php/home/index>

鋸科技 (2023)。70%美國教師認可生成式人工智慧進課堂，其爆發潛力如何？。鋸科技 **Semedia**。取自 <https://www.bbc.com/news/technology-65139406>

游昊耘 (2024)。全球調查教師最大挑戰：保持課堂參與度、學生行為與心理健康問題。翻轉教育 **Flipped Education**。取自 <https://flipedu.parenting.com.tw/article/009230>

ViewSonic (2024)。ViewSonic 攜手康軒文教集團、南一教育集團推出首款《AI 教科書》。廣告雜誌。取自 <https://adm.com.tw/article/66c7ed06fd897800015af77e>

TechNews (2024)。翰林教育科技：導入 AI 技術打造高效學習平台。TechNews 科技新報。 <https://technews.tw/2024/06/14/ai-chanlin-smart-learning-platform/>





附錄一 AI 試題生成指令 (Prompts)

壹、指令編號說明

- (一) R 為閱讀測驗，L 為聽力。
- (二) 該測驗第一部分題型標記為 01，閱讀測驗共五個部分，故有 01 至 05；而聽力測驗有四個部分，故有 01 至 04。
- (三) 若該部分的模型訓練時，使用了不只一個指令，則再於後方標記 -1、-2。
- (四) 綜上所述，指令編號若為 R01-1，便為用於閱讀測驗第一部分的第一個指令。

貳、生成時間

本研究已盡力記錄所有 AI 試題生成的相關資訊。然而，由於 AI 模型工具的限制，試題的確切生成時間未能被精確記錄。儘管如此，本研究已詳實記錄了所使用的完整生成指令，此為確保 AI 內容可再現性的關鍵因素。

指令編號	指令內容 (Prompt)	指令目的	生成時間
R01-1	請查看知識庫中「單句理解」檔案，學習這類題型的出題方式，也學習這類題型在 PRILS 提問法(直接推論、直接提取、詮釋與整合資訊)的出題方式	訓練模型統整、模仿命題題型方式。	約為 2024 年 11 月
R01-2	請給我一題「單句理解」「直接提取」的題目。選項為三張圖片。題目主題：「校園生活」。出題請嚴格按照「出題所用詞彙」的檔案，並參考「句型參考」檔案的句子。	生成所需試題。	約為 2024 年 11 月

R02-1	請查看知識庫中「看圖釋義」檔案，學習這類題型的出題方式，也學習這類題型在 PRILS 四層次(直接推論、直接提取、詮釋與整合資訊、評估和批判內容)的出題方式。	訓練模型統整、模仿命題題型方式。	約為 2024 年 11 月
R02-2	請給我一題「看圖釋義」「直接推論」的題目。題目請嚴格按照"出題所用詞彙"檔案的詞彙，並參考"句型參考"檔案的句子，做出多樣性的題目。三個選項的句型請完全不一樣，並確保答案的唯一性。圖片場景：商店當月店休時間。選項：描述營業時間。	生成所需試題。	約為 2024 年 11 月
R03-1	請查看知識庫中「選詞填空」檔案，1. 學習這類題型的出題方式：一個題組一張圖片、5~6 題題目，每題都是一個句子和三個選項，挖空的部分會是一個字或兩個字的詞。2. 學習這類題型在 PRILS 四層次(直接推論、直接提取)的出題方式。直接提取：選擇符合句子語法的詞 直接推論：除了合乎語法，還需按照題組圖片選擇符合的詞	訓練模型統整、模仿命題題型方式。	約為 2024 年 11 月
R03-2	請給我一題「選詞填空」題組，有「直接提取」和「直接推論」的題目，隨機分配順序和題數。題目共 5 題。請避免再出現前面的出題錯誤。這個題組的答案會是：ACBAC，並和前幾題類型、選項題目所用詞彙和語法都不同，但仍要嚴格使用"出題所用詞彙"檔案中的詞彙，不可超出範圍。圖片場景：朋友聚會。題目：參考圖片場景，有些問圖片資訊的題目；有些問讓學生選符合語法的題目。	生成所需試題。	約為 2024 年 11 月

R04-1	請查看知識庫中「選句填空」檔案： 1. 學習這類題型的出題方式：一段文字、5 題題目，列出五個選項讓考生配對至正確的空格，挖空的部分都會是一個短句子。 2. 學習這類題型在 PRILS 四層次(直接推論、直接提取、詮釋、整合觀點和訊息)的出題方式。 直接提取：從單句即可判斷出答案 直接推論：從複句判斷答案 詮釋、整合觀點和訊息：從段落判斷答案	訓練模型統整、模仿命題題型方式。	約為 2024 年 11 月
R04-2	請給我一個「選句填空」的題組。選項要用 PRILS 四層次的不同層次，並隨機分配。題幹主題：日常生活	生成所需試題。	約為 2024 年 11 月
R05-1	請查看知識庫中「閱讀理解」檔案： 1. 學習這類題型的出題方式：一段約 10 個句子的短文、1 題選擇題題目和 4 個選項。 2. 學習這類題型在 PRILS 四層次中"直接提取"和"直接推論"的出題方式。	訓練模型統整、模仿命題題型方式。	約為 2024 年 11 月
R05-2	請以「購物」為主題，設計一個為「直接提取」的「閱讀理解」題目。出題請嚴格按照"出題所用詞彙"，不可超出範圍。正確答案只能有一個，選項句子有多樣性，參考「句型參考」檔案的句子。短文中有情節轉折，題目提問關於事情發生的原因，或是關於短文哪個敘述是正確的。	生成所需試題。	約為 2024 年 11 月
L01-1	請查看知識庫中「看圖釋義」檔案， 1. 學習這類題型的出題方式：一張圖片、一個問題和三個選項。考生要按照圖片，選擇符合的答案 2. 學習這類題型在 PRILS 四層次(直接推論、直接提取)	訓練模型統整、模仿命題題型方式。	約為 2024 年 12 月

	<p>的出題方式。 直接提取：直接從圖片判斷 直接推論：透過圖片線索做簡單的推論</p>		
L01-2	<p>請給我一個「直接提取」的「看圖釋義」聽力題，主題和句子類型多樣化，參考「句型參考」檔案。嚴格按照「出題所用詞彙」檔案的詞，不可超出範圍。</p>	<p>生成所需試題。</p>	<p>約為 2024 年 12 月</p>
L02-1	<p>請查看知識庫中「問答理解」檔案， 1. 學習這類題型的出題方式：三張圖片選項、一個一問一答的對話(一男一女)。考生要按照聽到的對話，選擇符合的圖片選項 2. 學習這類題型在 PRILS 四層次(直接推論、直接提取)的出題方式。直接提取：直接從對話判斷圖片為何 直接推論：透過對話線索做簡單推論(例如：推測地點為圖片選項哪一個)</p>	<p>訓練模型調整、模仿命題題型方式。</p>	<p>約為 2024 年 12 月</p>
L02-2	<p>請給我一個"直接提取"的"問答理解"聽力題。嚴格按照"出題所用詞彙"檔案的詞，不可超出範圍。句子和主題類型多樣化，參考「句型參考」檔案。請直接生成選項圖片。</p>	<p>生成所需試題。</p>	<p>約為 2024 年 12 月</p>
L03-1	<p>請查看知識庫中「對話理解」檔案 1. 學習這類題型的出題方式：三張圖片選項、一個 4 句的對話(一男一女)和一個問題。考生要按照聽到的對話和問題，選擇符合的圖片選項 2. 學習這類題型在 PRILS 四層次(直接推論、直接提取)的出題方式。 直接提取：可以直接從對話判斷圖片為何 直接推論：透過對話線索做簡單推論</p>	<p>訓練模型調整、模仿命題題型方式。</p>	<p>約為 2024 年 12 月</p>
L03-2	<p>請給我一個「直接提取」的「對話理</p>	<p>生成所需試</p>	<p>約為 2024</p>

	解」聽力題。請確保答案只有一個，並嚴格按照「出題所用詞彙」出題，不可超出範圍。請直接生成三張圖片選項，圖片風格：簡約卡通。對話設計上，每題隨機讓不同的說話的人帶出正確答案。	題。	年 12 月
L04-1	請查看知識庫中「對話」檔案 1.學習這類題型的出題方式：一個 4 句的對話(一男一女)、一個問題和四個選項句子。考生要按照聽到的對話和問題，選擇符合的選項 2. 學習這類題型在 PRILS 四層次(直接推論、直接提取、詮釋/整合資訊)的出題方式。直接提取：可以直接從對話單句判斷答案為何 直接推論：透過對話線索做簡單推論 詮釋/整合資訊：統整整段對話才可判斷出答案	訓練模型統整、模仿命題題型方式。	約為 2024 年 12 月
L04-2	請給我一個「直接提取」的「對話」題目。請確保答案只有一個。對話設計上：每題隨機讓不同的說話的人帶出正確答案，一個句子大概 7 到 20 個字，務必要嚴格按照「出題所用詞彙」的檔案中的詞彙出題，不可超出範圍。	生成所需試題。	約為 2024 年 12 月