

Applying Generative Language Models to Generate Synthetic Medical Records: ChaVinci

(應用生成式語言模型生成擬真之醫療病歷文本：ChaVinci)

林璟芸 Jing-Yun Lin; 李旭清 Xu-Qing Li; 鍾聖倫 Sheng-Luen Chung

國立臺灣科技大學電機工程學系

Electrical Engineering Department

National Taiwan University of Science and Technology

Taipei, Taiwan

slchung@mail.ntust.edu.tw; fdsa3304@gmail.com; stanley890314@gmail.com

摘要

自動語音辨識技術 (ASR) 依賴大量語料庫支持，但在醫療情境中，真實病歷錄製成本高昂且受隱私限制。為解決此問題，本研究提出利用大語言模型生成擬真病歷文本的替代方案，結合 ChatGPT 與 DaVinci 的生成能力，創建了 ChaVinci 病歷文本集。ChatGPT 提供專業且一致的醫學描述，DaVinci 增加了病歷格式與內容的多樣性。本文探討病歷組成特性、NER 及 GPT 的文本生成技術，並分析大語言模型在擬真病歷中的應用，包含提示工程和生成合理性。實驗比較了真實與生成病歷，並將生成病歷應用於醫療語音庫錄製，評估不同語料庫訓練 ASR 的效能，證明其可有效支持 ASR 模型的訓練和測試，並拓展其應用。

Abstract

Automatic Speech Recognition (ASR) relies on extensive speech corpora for support, but in medical contexts, recording real medical records is costly and subject to privacy constraints. To address this issue, this study proposes using large language models to generate synthetic medical records as an alternative. By combining the generative capabilities of ChatGPT and DaVinci, we created the ChaVinci medical record set. ChatGPT provides accurate and consistent medical descriptions, while DaVinci adds diversity to the format and content of the records. This paper explores the characteristics of medical records, NER, and GPT text generation techniques, and analyzes the application of large language models in generating synthetic medical

records, including prompt engineering and generation rationality. Experiments compared real and generated medical records and applied the generated records to medical speech corpus recording, evaluating the performance of ASR trained on different corpora, demonstrating that it effectively supports ASR training and testing and expands its application.

關鍵字：擬真病歷生成、大語言模型、自動語音辨識 (ASR)

Keywords: Synthetic Medical Record Generation, Large Language Models, Automatic Speech Recognition (ASR)

1 簡介

1.1 醫療情境 ASR

醫療領域 ASR 的發展受限於語料庫稀缺。為解決此問題，本研究通過資料增強提升有限病歷的多樣性，涵蓋詞彙、句型和結構。我們對疾病診斷、臨床表現、藥物、手術等資訊進行分類和統計，並使用命名實體識別 (NER) 自動提取病歷資訊，為搜尋特定病症和治療方式提供基礎。此方法生成結構合理、內容多樣的擬真病歷，擴充現有文本庫。

新興文本生成技術為生成擬真病歷提供了新方法。我們利用 GPT-3 模型生成與真實病歷風格相符的文本，儘管可能出現診斷與臨床表現不一致的情況，但經專業人員校閱可提升準確性。我們還使用文本分析工具評估生成文本品質，確保其在醫療語境中的可行性，並支持 ASR 模型的訓練與驗證，拓展其在醫療領域的應用價值。

1.2 本文貢獻

本研究利用大語言模型生成大量擬真病歷。ChatGPT 等生成模型的崛起展現了其強大的文本生成能力，特別在醫療領域生成擬真病歷的潛力。我們使用 OpenAI 的 GPT 框架，通過詞彙與情境組合生成病歷，並將其轉換為錄音腳本，形成 ChaVinci 語料庫。這一方法克服了傳統醫療語料庫的高成本、耗時和隱私問題，為語料庫擴充提供高效解決方案，適用於住院、門診、巡房、急診室及手術室等多種醫療場景。

1.3 架構

本文結構如下：第 2 節「醫療文本」將探討病歷組成特性、命名實體識別 (NER) 及 GPT 文本生成技術的應用。第 3 節「大語言模型之文本生成技術」將詳細說明生成擬真病歷的流程、提示工程 (prompt engineering) 的應用，以及生成病歷的合理性分析。第 4 節「實驗與結果」將比較真實病歷與生成病歷，並探討生成病歷如何作為醫療語音庫腳本，以及不同語料庫訓練的 ASR 效能比較。最後，第 5 節「結論」將總結大語言模型的病歷生成技術在各種情境應用中的潛力與實踐價值。

2 醫療文本

2.1 病歷組成特性

病歷可概分成醫師跟護理師的病歷紀錄，這兩種的內容是完全不同的方式，其中，醫師一般使用全英文，大概內容包括：病人主述、過去相關病史評估，理學檢查 (Physical examination, PE)、相關檢查結果，疾病診斷及治療計畫。另一方面，護理師紀錄則是同時使用中英文，主要呈現護理過程，也就是問題解決策略，大概內容是主觀資料 (病人主述)，客觀資料 (行為觀察，身體評估及病歷相關資料)，護理問題確立，照護措施實施及評值。不同醫院護理紀錄標準會有不同的要求。一份完整的病歷除了醫師及護理師的紀錄外，還有很多其他重要的資料。

本研究所專注的病歷屬於住院護理病歷 (Inpatient medical records)，也就是住院病房之護理師交班病歷，其格式比較像是專業摘要似的筆記，除了簡稱之外，大部份是中英文夾雜的用字與句型。我們歸納 ChiMeS-14 語料庫

中的各病歷語音之譯文，得知以下規律：一般住院病歷組成的典型架構，按照醫療語境的分類，如表 1 所示，大致可歸納為以下的組成順序，其中，用粗字體的分類是為了方便讀者明瞭該句型在臨床情境上的分類，以及整份病歷不同屬性句型組成的架構。由這些分類所成所串接的分類順序則可視為該病歷的組合架構。對本例而言，即為：<觀察對象><疾病與診斷><臨床表現><評估表><手術/治療><檢驗><檢查><藥物><後處理><觀察對象>。其他的病歷，則有可能會有上述不同分類片段，如：<疾病與診斷><臨床表現><手術/治療><藥物><後處理> 等不同重複組合。

先是講明是住在第幾個房間、第幾號床的病人<觀察對象>，這病人過去的病史，是否容易跌倒、或者是其他高危險群的屬性。接下來會講是什麼的原因住院<疾病與診斷>，有哪些病徵/病狀<臨床表現>，有時會有明確帶數字的<評估表>。住院病人主要是有手術的，所以一般接下來就會講到要接受什麼<手術/治療>，而開刀之前，有時會需要麼針人體之血液、排泄物、分泌物來進行化驗或培養<檢驗>，或是透過影像等，來查看對人體各器官功能是否外否正常<檢查>。住院期間要吃什麼藥<藥物>。並且，對排定要手術的病人，需要作什麼準備，以及手術/檢查後護士要幫患者做的事<後處理>，以及要量取患者的那些基本信息<心率、能力...><觀察對象>，最後是出院。

表 1：住院護理病歷的典型架構

2.2 參考文獻

GPT 與文本生成：Transformer 的架構是由 Encoder 和 Decoder 組合而成的，其中 Encoder 將輸入的序列映射的高維度的空間，而 Decoder 則是將該 Embedding 轉換成另一個序列，這當中的序列可以是聲音、文字和圖片等資料。前文提到的 BERT 正是只使用了 Encoder 的架構，來將輸入映射至高維度空間；相對而言，如果是針對生成任務，則可以使用 Decoder 的架構。A Radford (2018) 提出了 GPT (Generative Pre-Training) 的架構，其為 Transformer 的 Decoder 架構，利用非監督式學習，從大量文本學習語意理解，其在當時多項任務中都獲得 State-of-The-Art 的效果。之後根據 GPT 的基礎，Alec Radford et al. (2019) 提出了 GPT-2，其利用 Multitask 讓模型在不同領域中學習，以實現在不同任務中取得遷移學習的知識，而證明了大型語言模型能夠在 Zero-Shot 中執行一系列任務的能力。GPT 最新的發展是，Brown (2020) 使用了約為 GPT-2

1000 倍的 45TB 資料，訓練出了帶有 175 Billion 參數量的 GPT-3，並且提出了 Few-Shot 的觀點：只要給 GPT-3 幾個例子，即使是當初沒有 Fine-Tune 過的任務，GPT-3 所生成的文本還是能夠有一定的正確率。GPT-3 在大多任務都超過了 Human baseline，展示了當資料量非常大的時候，模型能夠學習到除了訓練任務之外的資訊。最後，Raffel et al. (2020) 提出了 T5 (Text-to-Text Transfer Transformer) 的模型，通過引入一個將所有基於文本的問題，都轉換成 Text-to-Text 的格式的統一架構，使用了完整的 Transformer 架構進行遷移學習，利用 Attention 機制來理解語意以及詞與詞之間的相關性，並且比較了數十種語言理解任務的架構和目標，其中在生成摘要、問答、文本分類任務都獲得了非常好的成效。在文本生成的任務通常都是由左至右生成的，包括 GPT 和 T5 都是用這種方式生成摘要，於是 Zhang et al. (2019) 提出了一種兩階段的方法，第一階段利用基於 Transformer 的 Decoder 來生成輸出的草稿，之後依序遮罩草稿的單詞並輸入 BERT，再由其結果透過另一個 Decoder 來輸出精確的單詞。傳統的文本生成都依賴於良好的內部結構，亦即訓練資料都會有一個完整的結構，LeClair et al. (2019) 提出了新的做法，他們將每個輸入源單做一個獨立的輸入。這個做法允許模型學習不同架構的文本，即使提供沒有相應架構的輸入，此過程也能夠在許多情況獲得適當的摘要。最後，文本生成任務中難以使用一般的方法來評估文本生成的品質，於是 Van Der Lee et al. (2019) 人研究了多種評估的方式，並且總結了 General、Criteria、Sampling、Annotation、Measurement、Design 和 Statistics 的最佳實踐方法，通過標準化人類評估的執行方法來幫助 NLG (Natural Language Generation) 的研究。

比較 NER 技術與 ChatGPT 用於生成仿病歷時，ChatGPT 展現了更高的效率和靈活性。NER 需要對病歷資料進行標注和結構分析，過程繁瑣且耗時，而 ChatGPT 能即時生成高質量仿病歷，縮短數據準備到結果產出的時間。NER 生成的文本受原始語料限制，難以產生多樣化內容；相較之下，ChatGPT 能靈活生成多種句型和情境，提供更豐富的變化。ChatGPT 在專業性和擴展性方面也更具優勢，能生成符合臨床實踐的文本，並減少人工干

預，易於整合其他自動化工具，實現一體化流程。

3 大語言模型之文本生成技術

3.1 生成擬真病歷流程

本研究利用先進文本生成模型來生成特定風格文本。注意力機制引發了基於 Attention 的模型發展，尤其是 Transformer 架構，帶來了 Pre-train 模型的突破。這些模型，如 GPT-3、T5 和 PaLM，都需要大量算力和數據。GPT 專注於生成任務，從 2018 年起推出 GPT-2 和 GPT-3 等模型，隨著參數量和數據規模增長，GPT-2 採用 Multitask 策略應對翻譯、問答、總結等任務。GPT-3 引入少樣本(Few-shot)技術，進一步提升生成文本的精準性與實用性。

我們選用 ChatGPT 3 系列中，生成文本效果最好的 DaVinci，以及特別針對對答優化的 chatGPT-3.5 當作生成病歷的模型，生成擬真之醫療病歷文本集：ChaVinci (ChatGPT + DaVinci)。以下是其生成病歷的流程：

1. 初步準備：首先，我們從資料庫中挑選一些真實的病歷資料。這些病歷資料將作為模型的初步輸入，讓它瞭解真實病歷的結構和內容。
2. 設定 prompt：在設定 prompt 時，我們不僅將真實病歷作為參考輸入，還會添加一些指令，如：“請根據以下的真實病歷，生成一份與其不同但具有醫學意義的病歷。”
3. 模型生成：利用 ChatGPT 或 DaVinci 的生成能力，根據我們提供的 prompt 生成一份新的病歷。
4. 質量檢查：一旦生成病歷，我們需要進行質量檢查，確保生成的病歷不僅與原始病歷有所不同，還保持醫學的準確性和連貫性。
5. 轉換成語音：經過質量確認的病歷將被轉換成 TTS 音檔，為後續的語音辨識和分析提供實際應用場景。

3.2 提示工程 (prompt engineering)

要有效利用 ChatGPT-3.5 和 DaVinci 生成病歷文本，關鍵在於精心設計的 prompt。首先，應明確病歷的具體需求，如科別、症狀、診斷和處方，並保持邏輯一致。

評估 prompt 的好壞可從幾個角度進行。首先，專業性和準確性是最重要的，生成的文本需符合醫學標準。其次，好的 prompt 應能引導模型生成包含所有必要醫學信息的完整病歷，包括症狀、診斷、處方和患者背景，並確保文本前後邏輯一致，沒有矛盾。此外，prompt 的可控性和靈活性應能讓使用者調整文本風格和深度，應對不同醫療場景。最終目標是通過有效的 prompt，高效生成高質量的病歷文本，減少試錯和修改需求。

本文將探討 ChatGPT 和 DaVinci 在病歷生成中的應用與比較。首先，將介紹這兩個模型的 prompt 設計，分析如何通過設計有效的 prompt 來最大化它們的生成能力。接著，通過具體範例展示兩個模型的生成結果，並進行對比分析。最後，將綜合評估 ChatGPT 和 DaVinci 的優缺點，為使用者選擇醫療文本生成工具提供參考。

能模仿下面內文的格式生成一筆不同的資料嗎?
[病歷 1]

表 2：ChatGPT prompt

[病歷 1]
[病歷 2]
開頭提示(ex: 男性_歲診斷、診斷是、七 B 三二...)

表 3：DaVinci prompt

ChatGPT prompt 是使用指令句：「能模仿下面內文的格式生成一筆不同的資料嗎？」加上一份真實病歷，如表 2。

DaVinci prompt 則是使用使用兩份病歷，加上不同的開頭提示來生成，如表 3。

經實驗和觀察，我們發現 ChatGPT 生成的病歷往往類似於「拼湊」的結果，它傾向於對特定單詞進行替換，但整體結構保持不變。相對之下，DaVinci 所產生的病歷與[病例 1]和[病例 2]有較大的差異，展現出更高的原創性和活潑度。

在表 4，紅色標註為與 prompt 重疊的部分，我們特意對 ChatGPT 和 DaVinci 進行了相同的 prompt 測試，進一步確認了我們的觀察結果，ChatGPT 主要進行「拼湊」和模仿，而 DaVinci 展現更大的原創性，這揭示了兩模型的不同生成策略，我們可以針對兩種模型不同的特性列出優缺點，如表 5。

3.3 生成病歷的合理性

我們詳細檢視了生成的病歷在各個部分(如症狀、診斷、病名、手術、術後及用藥等)是否展現出一致性。為了驗證這一點，我們隨機對 15 份生成的病歷人工進行了仔細的評估，

<p>Prompt : 八一診斷是{acu}{te}髓性白血病，病人沒有高熱或高壓，沒有過敏史，目前{on}NPO，然後在急診病人有抽過CBCBCS 跟{blood}{ca}，然後有抽血{da}{ta}是 HB 是十六點三，HCT 是四十四點四，MCV 九十二點六，{mo}{no}{cyte}十點六，{glu}{cose}一零二，{white}{count}是兩百八十萬，{neu}{tro}{phil}是兩百六十萬，{lym}{pho}{cyte}十五萬，然後{se}{rum}{iger}兩點五，PLT 是一百三萬，EKG 是 NSR X{ray}是圓形心肥大，然後 PRN 的話有{no}{va}{min}可以給，然後因為病人有血尿兩次，所以有建議他入院治療，病人有{acu}{te}髓性白血病的{his}{tory}，在急診有{cexi}{tin}一{gram}IVD 劑，然後{vi}{tal}{sign}都{sta}{ble}，GCS 是 E 四，V{two}M 四。</p>
<p>ChatGPT generate: 九二診斷是{mye}{lo}骨折，病人沒有糖尿病或心臟疾病，沒有過敏史，目前{on}液體飲食，然後在急診病人有抽過 ABG 跟{urine}{test}，然後有抽血{da}{ta}是 RBC 是五點二，HGB 是十三點五，WBC 是七點九，{po}{ta}{ssium}五點三，{sodi}{um}一四五，{blood}{urea}十四，{lym}{pho}{cyte}九萬，{plate}{let}是二百五十萬，{mono}{cyte}十二萬，EKG 是 AF，然後 PRN 的話有{pain}{killer}可以給，然後因為病人有輕微的呼吸困難，所以有建議他入院治療，病人有{bro}{ken}骨的{his}{tory}，在急診有{amoxi}{cillin}五百{mg}IVD 劑，然後{oxy}{gen}{level}九五，BP 一二零/八十，GCS 是 E 三，V{three}M 五。</p>
<p>DaVinci generate: 他的診斷是右側乳癌，沒有高熱，沒有高壓。目前{on}軟飲食，對{ce}{fa}{zo}{lin}過敏，右側禁治療。他有乳癌的{his}{tory}，還有 CS 開過刀。他做過乳房切除術，然後術後轉本院急診。急診時有嘔吐，所以有打過一支{no}{va}{min}，外帶 JP 兩條，X{ray}是{nor}{mal}，EKG 是{si}{nus}{ta}{chy}{car}{dia}。抽血報告 VBG 是 PH 七點四三六，PCO{two}三十四點五，PO{two}七十四點六，{white}{count}是兩萬三千六百八，{neu}{tro}{phil}是九十二點一，{lym}{pho}{cyte}四點一，{glu}{cose}一四九，{amy}{lase}一七六，{tro}{po}{nin}I{nor}{mal}，PCT{nor}{mal}，PTAPTT{nor}{mal}。有 PRN 的{no}{va}{min}可以給，還有{a}{ti}{van}如果有失眠的話可以給，然後{mu}{ton}{pain}是如果有痛的話 Q{four}可以給。開完刀右側乳房下有一條 JP{drain}，另外一條是五月八號拔掉，傷口是紗布跟彈紗{co}{ver}。住院期間 JP{drain}兩條倒出來都二十多，五月八號{re}{move}掉一條，剩下一條帶回家。MBD 有說需要申請診斷書三份，申請重大傷病卡，可以用這次住院出院。</p>

表 4：相同 prompt 分別用 ChatGPT 與 DaVinci 生成病歷

ChatGPT	
優點	1. 結構一致性：確保了病歷資料的結構一致性和格式正確性。 2. 可調整的生成策略：通過調整 prompt 可以靈活生成不同的情境，因應所生成的結果再作調整，適應各種醫療需求。 3. 隨時更新能力：GPT 可以通過「網路瀏覽功能」即時獲取最新的醫療研究和臨床數據，不斷更新和擴展其生成內容的專業性。
缺點	1. 過度依賴 prompt：生成結果高度依賴 prompt 的設置，如果 prompt 設置不當，可能會導致生成內容的質量不穩定或偏離預期。
DaVinci	
優點	1. 高度原創性：能夠生成與原始病歷有較大差異的內容，有助於增加語料庫的多樣性。 2. 細節豐富：DaVinci 能夠捕捉並表達更多的臨床細節，如病史、過去的手術記錄、具體的檢查結果和治療方案等，這些細節使得生成的文本更為全面和完整。
缺點	1. 不確定性：由於生成內容較為原創，可能出現與真實病歷情境不匹配或不正確的描述。 2. 需要更多審核：原創性較高的病歷可能需要更多的人工審核，以確保其質量和正確性。

表 5：ChatGPT 與 DaVinci 的優缺點

結果證明這些病歷在各部分都具有相對的一致性。此外，我們還將這些文本提供給 ChatGPT，請它自行評估這些病歷的一致性，作為對我們人工審核的補充。如表 6 在 ChatGPT 所生成的病歷中，從診斷到病史到狀態都是合理且有邏輯的。

ChatGPT generate analysis	
診斷	診斷為骨折。
病史	病人沒有糖尿病或心臟疾病，沒有過敏史，且有骨折的歷史。這些資訊都是清晰的。
目前狀態	病人只能攝取液體飲食。這在有骨折或其他疾病的情況下是合理的。
急診檢查	在急診中，病人接受了 ABG 和尿液檢查，以及抽血檢查，得到了多項血液數據。這是常見的做法。
檢查結果	EKG 顯示 AF。
治療	病人因呼吸困難被建議入院治療。在急診中，病人得到了 amoxicillin，這可能是因為病人可能有或是疑似有感染。病人的氧氣濃度是 95%，血壓是 120/80，這些都在正常範圍內。
狀態	GCS (Glasgow Coma Scale) 是 E3, V3, M5，這表示病人的意識狀態是相對正常的。
總體來看，病例資料在前後文上都是相當一致的。每一部分都有合理的解釋和描述，並且與病人的疾病和治療方式相符。	

表 6：ChatGPT generate 病例一致性分析

如表 7 在 DaVinci 所生成的病歷中，也可以看到診斷、病史、治療和狀態，都圍繞者診斷中的症狀。綜上所述，無論是 ChatGPT 或 DaVinci 生成的病歷，它們在症狀、診斷、手術和用藥等方面都顯示出了相對的一致性和邏輯性。這意味著這些生成的病歷不僅具有

形式上的真實感，而且在內容上也具有相當的可信度，為醫療領域的研究和應用提供了一個有力的工具。

DaVinci generate analysis	
診斷	病人有右側乳癌的診斷。
病史	病人有乳癌的歷史且已經接受乳房切除術。這是相關且符合邏輯的。
目前狀態	病人目前僅能吃軟食，對 cefa 有過敏，所以右側禁止某些治療。
急診檢查	因術後嘔吐而轉入本院急診。給予止吐藥，並且做了 X 光和 EKG 檢查。此部分資訊是合理的
檢查結果	抽血報告提供了多項數值，包括血氣、白血球數、電解質、酵素等。這些檢查在術後都是相對常見的，尤其是在病人有症狀時。
治療	有提供止吐藥，並提供止痛和助眠藥物作為需要時使用
狀態	開完刀右側乳房下有一條 JP drain。另外提及的傷口處理方式(紗布和彈紗覆蓋)也是合乎常規的。
其他	住院期間有倒出的液體量記錄，和提及有申請重大傷病卡及診斷書的需求。這部分是住院手續和病人出院後需要注意的事項。
總體來看，病歷資料在前後文上都是符合邏輯的。每一部分都有合理的解釋和描述，並且與病人的疾病和治療方式相符。	

表 7：DaVinci generate 病例一致性分析

4 實驗與結果

第四章將會深入探討以生成式 AI 技術為基礎的醫療語音辨識系統構建過程及其效能。我們將先從真實病歷與生成病歷的比較著手，探索 ChatGPT 與 DaVinci 所生成的 ChaVinci 語

指標	Mes2023	DaVinci	ChatGPT
是否需要實病例的範例	否	是	是
是否需要專業性的錄音	是	否	否
製作成本	高 (需專業人員)	低 (僅模型費用)	低 (僅模型費用)
標注的成本	高 (需人工標注)	低 (無需標注)	低 (無需標注)
靈活性	低 (受限於背景知識)	中 (調整 prompt)	高 (prompt 與網路瀏覽)
多樣性	低 (表達風格相對固定)	中 (調整 prompt)	高 (prompt 與網路瀏覽)
便捷性	低 (耗時費力)	高 (可快速生成)	高 (可快速生成)
專業性	高 (具有豐富醫學知識)	中 (取決於模型)	中 (取決於模型)
資料隱私風險	中風險 (護士撰寫擬真案例)	低風險 (生成數據)	低風險 (生成數據)

表 8：醫療病例生成比較表

料庫的優勢與應用，特別是在大規模生成文本上的效率與多樣性。接著，將說明如何利用這些生成的病歷文本作為腳本，建立醫療語音資料庫，並對應到不同文字轉語音(TTS)技術的辨識性能差異。最後，我們將進行不同語料庫的 ASR 效能比較，探討 GPT 所生成的語料在靈活性、多樣性和專業性上的表現，並分析其與真實病歷和 NER 技術之間的優劣，進一步展示生成式 AI 在醫療應用中的潛力。

4.1 真實病歷與生成病歷

利用 ChatGPT-3.5 以及 DaVinci 所生成的 ChaVinci 語料庫，共含 ChatGPT 生成的 450 份病歷，以及由 DaVinci 生成的 650 份病歷組成，總計 1100 份病歷。表 8 比較了由成大醫院護理科所準備六百份病例的 Mes2023，以及 DaVinci 和 ChatGPT 所生成擬真病歷文本的各項指標。結果顯示，Mes2023 具備高度的專業性和細緻的語境捕捉能力。然而針對真實病歷採集方法的缺點是成本高昂，且撰寫過程耗時，文本的多樣性和靈活性也受到撰寫者個人表達風格和知識背景的限制。

相較而言，GPT 在靈活性、多樣性、便捷性等方面顯著優於 Mes2023。通過大量醫療數據的訓練，GPT 能夠快速生成內容豐富且語境連貫的文本，並且可以靈活調整以適應不同的情境需求。雖然在專業性上可能略遜於真人撰寫，但其生成的文本仍能達到臨床應用的要求，並且在便捷性和生成效率上具有無可比擬的優勢。這使得 GPT 成為一種極具

潛力的工具，特別是在需要大規模、快速生成醫療文本的場景中。

4.2 生成 ChaVinci 為腳本的醫療語音庫

當作生成訓練 ASR 的語音庫，我們利用 ChaVinci 語料庫中所生成擬真病歷當作錄製醫療語音的腳本，然後分析相應 ASR 的辨識績效。為了快速檢驗此概念我們採用兩種不同的文字轉語音 (TTS) 技術：Google TTS 以及 Microsoft TTS，

合成相應的語音庫。總計包括 1100 份病歷的 ChaVinci 共生成，總時長約 17 個小時半的生成語音庫，如表 9。而在音訊生成部分，根據兩種不同的文字轉語音 (TTS) 服務平台音檔分割方式也不同。

語料庫	句數 (訓練 /測試)	平均 時長	平均 字數	總時長
Chavinci	6802 (5208 /1597)	9.31 s	56.16	1055.67 min (17hr 35.67 min)

表 9：Chavinci 語料庫資料分布

第一種是使用 Google TTS 服務的 ChaVinci-g，在 ChaVinci-g 中只有使用單一聲音生成音檔，從生成的語音可以明確地辨識出它是由機器生成。其語音雖然清晰，但在語調和發音上仍顯得稍微機械化，不夠自然。

第二種是使用 Microsoft TTS 服務的 ChaVinci-m，在 ChaVinci-m 中使用到了 8 個聲

語料庫	Training				Testing			
	npsChimes14	Mes2023	ChaVinci-g	ChaVinci-m	npsChimes14	Mes2023	ChaVinci-g	ChaVinci-m
病歷總數	394	499	800		101	122	300	
句子總數	5682	1573	5208		395	1573	1594	
Characters	2234	3363	2980		1940	1633	2108	
registered C /總出現數	2234/189181 (1.18%)	3363/422069 (0.8%)	2980/198118 (1.5%)		1807/39004 (4.63%)	1511/54305 (2.78%)	1859/60946 (3.05%)	
Keyword	973	1946	2070		758	699	1345	
registered K /總出現數	973/15212 (6.4%)	1946/34296 (5.67%)	2070/22049 (9.39%)		629/3179 (19.79%)	586/4675 (12.53%)	903/6786 (13.31%)	
OOV /OOV_frequency	X	X	X	X	122/183 (66.67%)	133/186 (71.51%)	249/298 (83.56%)	
OOK	X	X	X	X	113	129	442	
CER	X	X	X	X	14.3%	13.16%	4.23%	6.15%
KER	X	X	X	X	18.72%	18.11%	9.61%	13.58%
OOK_URR	X	X	X	X	40/113 (35.4%)	42/129 (32.56%)	240/442 (54.3%)	210/442 (47.51%)
OOK_ORF	X	X	X	X	55/155 (35.48%)	51/160 (31.88%)	300/531 (56.5%)	255/527 (42.69%)

表 10：npsChimes14、Mes2023、以及 ChaVinci-g 和 ChaVinci-m 指標對比

音生成音檔，在訓練集中聲音種類有 4 女 1 男（聲音 id 為 #35），而在測試集中有 2 女 1 男（聲音 id 為 #38），聲音種類在資料集中沒有重複，相較 Google 的 TTS，Microsoft 所生成的音檔更靠近人聲。

4.3 不同語料庫訓練的 ASR 效能比較

我們針對四個不同的語料庫進行比較，分別是 Mes2023、npsChimes14、ChaVinci-g 和 ChaVinci-m。使用 Conformer ASR 架構，在每個語料庫上進行了訓練和測試，表 10 為實驗結果。

在表 10 中，我們比較了 npsChimes14、Mes2023、與 GPT 生成的 ChaVinci 病歷文本的多樣本。其中涵蓋多項指標，**病歷總數**表示語料中包含的病歷文本的總數，而**句子總數**則是語料中所有句子的總數，這些句子以句號分割計算。**Characters** 代表語料中的唯一 token (token_unique)，並不計算出現的次數。這裡的 token 是一顆顆中文字和英文單音節。**registered C / 總出現數**表示在訓練集中已經出現的 token_unique (即 registered C) 占測試集中所有 token 出現的總次數的比例。**Keyword** 是指醫療文本中的專業醫療詞彙，如藥品名稱或診斷名稱。**registered K / 總出現數** 表示在訓練集中已出現的 keyword_unique 占測試集中所有 keyword 出現總次數的比例。

OOV (Out of Vocabulary) 指訓練集中未出現、只在測試集中出現的 token_unique，而 **OOV_frequency** 指測試集中 OOV 出現的總次數。**OOK** (Out of Keyword) 則指訓練集中未出現、只在測試集中出現的關鍵詞。**OOK_URR** (OOK Unique Recognition Rate) 為在測試及中有被正確辨識出 OOK_unique/OOK_unique 數量。**OOK_ORF** (OOK Overall Recognition Frequency) 為在測試集中有被正確辨識出 OOK 總次數/OOK 出現的總次數。

在語音識別準確性方面，**CER** (Character Error Rate) 是用來衡量字元錯誤率的指標，而 **KER** (Keyword Error Rate) 則用來衡量系統對關鍵詞的辨識準確度。

我們以語料庫中涉及的單字與關鍵詞，當作衡量多元性的指標，與測試集文本中未曾出現在訓練集文本中的單字以及關鍵詞，即 OOV 與 OOK，當作生成式工具的原創性指標。在本表中，npsChimes14 是耗時三年所搜集的病歷，其訓練集中不重複的單字為 2234，所涉及的關鍵詞有 973 個。為了擴增多元性，Mes2023 是動用 28 位成大護理師，特別安排涵蓋 16 個醫學科別的病歷，而得到不重複的單字 3363 個與關鍵詞 1946。另外 npsChimes14 與 Mes2023 測試集中，所有出現不重複的單字分別為 1940 與 1633 個，而未曾出現在訓練集中的單字 OOV 則各有 122 與 133；而測試集

的關鍵詞部份共出現 758 與 699 個，而未曾出現在訓練集中的 OOK 各別有 113 與 129。

相較而言，以 npsChimes14 為文本範例所生成的 ChaVinci 病歷，在不動用護理師的情況下，所生成病歷數與前兩個語料庫相當的情況下，訓練文本包括 2980 個單字、2070 個關鍵詞；而測試文本包括 2108 單字與 1345 個關鍵詞，而最重要的是 OOV 含 249 個，OOK 有 442。優於 Mes 所對應的 133 與 129。ChaVinci 系列的關鍵詞數量在訓練和測試集中均超過其他語料庫，顯示 GPT 能生成豐富且連貫的醫療專業詞彙。此外，ChaVinci 在 OOV 數量上也最多，表明 GPT 生成的文本具備更多樣性和新穎性，能靈活適應不同醫療情境和詞彙需求。

在醫療文本生成中，相較於真人撰寫和 NER 技術，我們可以看到 GPT 在靈活性、多樣性和專業性方面展現出了無可比擬的優勢，使其在醫療仿病歷的生成上更具競爭力。首先，從靈活與多樣性上看，GPT 生成的文本顯著優於由真人撰寫的文本與 NER 技術。真人撰寫的病歷雖然能夠準確反映專業醫療知識，但受限於個人的語言表達風格和專業背景，其生成的文本往往有一定的局限性；而 NER 技術主要依賴於預先設計好的標籤和規則，這使得它在生成新穎或未曾見過的語料時，容易受到原始數據的限制，缺乏足夠的多樣性；相比之下，GPT 能夠通過大量醫療數據的訓練，學習並模擬多種語言風格和表達方式，透過不同的 prompt 的調整使得生成的文本能夠適應不同的語境需求，並具備更高的靈活性。此外，GPT 還能透過「網路瀏覽功能」從網路上獲取最新的醫療研究和臨床實踐資料，不斷更新和擴展其生成的內容，這是 NER 技術難以達到的。

從成本考量，真人撰寫病歷耗時且昂貴，NER 技術需標注與結構分析，過程繁瑣。而 GPT 能在短時間內生成大量不同風格的仿病歷，無需標注或微調模型。在專業性與語境捕捉上，GPT 表現更佳。真人撰寫受限於個人風格和負荷，品質不一致；NER 依賴標籤和規則，難以捕捉細微語境差異，生成文本易失去連貫性。GPT 經過大規模語料訓練，能生成更連貫、符合臨床實踐且專業用語準確的文本。

總結來說，GPT 在醫療文本生成中的靈活性、多樣性、便捷性和專業性，使其在仿病歷生成上遠優於真人撰寫和 NER 技術。GPT 不僅能夠快速生成高質量且多樣化的醫療文本，還能適應不斷變化的醫療情境，提供更準確的語境理解和專業術語應用。這些優勢使得 GPT 在醫療應用中的潛力無限，特別是在需要高效生成專業文本的場景中，GPT 的應用前景無疑是最為廣闊的。

5 結論

前述採用大語言模型的病歷生成技術可應用於更多的情境中，如以下表 11。

這些利用 GPT 用來生成病歷範例文本的來源，一共有：(1) psChiMeS14：來自衛福部台北醫院護理站合作的 516 份病歷；(2) Mes2023：來自成大醫院護理科合作的六百份病例；還有 (3) 網路上醫療網的爬文。

首先，psChiMeS14 被用來當作 DaVinci 以及 chatGPT-3.5 生成仿真病歷的範本：ChaVinci2023 (ChaVinci)。當作成功的概念驗證 (Proof of Concept) 我們的觀察是 DaVinci 生成的效果相對比較好，這可能是同屬於 chatGPT-3.0 訓練的模型，DaVinci 本就是文本產生功能中最優的，而 chatGPT-3.5 是特別針對問答應用微調。當有效果較優的文字轉語音 (Text2Speech) 模組搭配，這些生成的病歷有望用來當作錄製訓練 ASR 的腳本，以提升 ASR 的辨識效果。

其次，Mes2023 被用來當作 chatGPT-3.5 與 chatGPT-4o 生成仿真病歷的範本。主要是要生成特定科別的擬真病歷資料補足之前收集病歷僅集中在部份科別的限制，這些生成的病歷並由醫護人員進行其中 200 份生成病歷合理性的修改：Vet-Mes2023。另外，我們也在這 600 份的實際病例範例上，產生更多元性，即包括更多醫科別專業術語的病例，以作為訓練特別針對醫療情境之自然語言處理用的 Medical Bert 的訓練文本：GPT-Mes2023。此外，我們生成部份擬真病歷，實槌：「生成雙層式 NER 標註，並進一步協助雙層式模板置換」概念的不太可行性。最後，不限於我們自行收集護理交班情境的文本，我們也從網路上面所取得其他醫療網的文本，並以 chatGPT-4o 生成用來測試 ASR 的測試語音的

產出項目	生成擬真病歷目的	模型	輸入文本範例	輸出份數
Chavinci	驗證 chatGPT 生成擬真病歷文本之可行性	<ul style="list-style-type: none"> DaVinci (Model: text-davinci-003) GPT (Model: gpt-3.5-turbo-instruct) 	psChiMeS14	<ul style="list-style-type: none"> DaVinci:650 份 GPT:450 份
Vet-Mes2023	生成特定醫科別的擬真病歷，並提供成大醫院護理師審核與修改	GPT (Model: gpt-3.5-turbo)	Mes2023	200 份病歷
GPT-Mes2023	生成擬真交班病歷當作訓練文本，微調醫療情境 MedCKIPBERT	GPT (Model: gpt-4o)	Mes2023	7460 句
SNOMED-CT-Mes2023	概念驗證「結合 SNOMED CT 之雙層 NER 技術嘗試」之可行性	GPT (Model: gpt-4 網頁板)	Mes2023	各 5 份
AI-Gen-Mes	生成病房交班情境外之一般醫療語音之錄音腳本	GPT (Model: gpt-4o)	一般醫療文本 (網上蒐集)	354 份病歷

表 11：生成技術更多的應用情境

錄音腳本：AI-Gen-Mes。這被用來作為微調醫療情境自然語言處理用 Bert 的訓練樣本。

Zhang, H., Xu, J., & Wang, J. (2019). Pretraining-based natural language generation for text summarization. *arXiv preprint arXiv:1902.09243*.

References

- Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- LeClair, A., Jiang, S., & McMillan, C. (2019). *A neural model for generating natural language summaries of program subroutines*. Paper presented at the 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE).
- Radford, A. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., . . . Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 1-67.
- Van Der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Kraemer, E. (2019). *Best practices for the human evaluation of automatically generated text*. Paper presented at the Proceedings of the 12th International Conference on Natural Language Generation.