

以文物詮釋資料建構生成式 AI 語料之初探

■ 羅勝文

博物館典藏與研究資料蘊含高度知識密度，具有做為生成式 AI 語料開發之潛力。本文參考上下文學習方法，提出一套基於文物詮釋資料之語料生成架構雛形。第一階段以故宮文物詮釋資料中的「說明文」與「出版紀錄」欄位為語意擴展基礎，推導並補全「質材」、「紋飾」、「色澤」等欄位內容，建立初步的語意對應關係。第二階段則進一步結合《故宮文物月刊》等研究資源，自其中擷取具代表性的詮釋語句與描述語料，建構具國立故宮博物院（以下簡稱故宮）特色的語料訓練資料集。經由格式轉換與語義調整，該資料集可投入大型語言模型之微調或檢索增強生成應用，除可支援多語系展件／展覽導覽服務外，亦能語意、圖形化等技術優化跨文物、跨時期與跨類型的語意理解與互動應用。

前言

故宮長期致力於運用數位科技，讓典藏與研究成果以更豐富的方式與觀眾互動，隨著生成式 AI（Generative AI, GAI）的興起也迎來新的發展方向。近年來，許多博物館紛紛將典藏資料轉化為 AI 訓練語料，發展智慧導覽、語意檢索、內容生成、展覽詮釋等服務，¹ 然而，這些應用成果的品質高度仰賴訓練語料的品質、多樣性與規模，² 且經過精心整理的專業資料，其效果往往優於來源龐雜的網路文本。³ 對於擁有豐富館藏與學術研究的故宮而言，若能將館藏資訊、文物詮釋資料（Metadata）、展覽圖錄與研究出版品等資源轉化為高品質語料，將能更有效發揮故宮在內容與專業知識上的優勢，建構可持續發展的 AI 服務基礎，同時為未來的智慧導覽與文化體驗創新奠定關鍵基礎。如何有效完成這項轉換，正是推動故宮

數位服務深化的重要課題，也是本文的核心探討方向。

要實現上述目標，首要步驟是確定合適的語料來源。儘管研究出版品本身便能作為訓練語料，但由於每位作者的寫作目標、研究取徑、專業術語使用皆不盡相同，加之故宮典藏數量龐大、文物類型與專業術語繁多，若僅依賴研究出版品則可能產生語意模糊、概念混用與分類不一致等問題。因此，文物詮釋資料扮演關鍵的「語意對齊」角色，使 AI 能依循文物基本資料抓取適合的語料，確保語料的來源與可回溯性。

幸而，故宮已擁有兩大類具備極高價值的潛力語料：結構化的詮釋資料和豐富的學術知識。故宮自約 2000 年起，持續系統性建置文物詮釋資料，期間歷經數次改版，重要項目涵蓋品名、分類方式、時間、空間、數量、改製、

功能等典藏脈絡，⁴ 具備高度結構化特性。另一方面，《故宮學術季刊》和《故宮文物月刊》創刊以來已累積豐富的研究成果，蘊含大量深度知識。這些資料不僅承載重要知識內容，也具備作為 AI 訓練語料的潛力。

要讓上述潛力真正轉化為可用的 AI 語料，關鍵在於如何將現有的文物詮釋資料之外進行內容擴展與語意補強。為此，本文參考上下文學習（In-Context Learning，以下簡稱 ICL）方法，先將文物詮釋資料轉化為具體的示範案例（input-label pairs），再依這些範例從研究出版品中擷取對應內容，作為 AI 生成導覽與解說文字時的參考語料。

擴展文物詮釋資料內容

文物詮釋資料具備高度結構化特性，使其成為大型語言模型（Large Language Models，以下簡稱 LLM）理想的示範案例。ICL 是 LLM 廣泛應用的推論技術，透過適當的提示設計（prompting），引導模型僅需在推論階段讀入少量示範案例，無須進行任何額外的參數更新，⁵ 即能模仿文物詮釋資料的內容結構與敘述邏輯，生成語義準確、風格一致且富含文化深度的文本。如表一所示給予 LLM 數個範例和應對方式，模型能據以回覆，LLM 能夠經由少量範

例「即時學習」，模仿輸入資料的結構與用語，展現出高度的語言適應能力。

然而，文物詮釋資料多以精簡方式呈現資訊，限制了語言模型可用的推論線索。例如，「翠玉白菜」的「質材」欄位雖精確標示為「礦物」、「玉石珠寶」、「輝玉」（翡翠），但此類零星詞條缺乏足夠語境，導致 LLM 僅憑欄位資訊難以推論出具詮釋價值或可作為訓練語料的深度內容。為彌補此限制，本文運用文物詮釋資料中的描述性內容（如文物簡介），針對關鍵欄位進行語義內容擴展，以強化語境，使模型能更有效掌握欄位間的語義關聯並生成更具知識深度的語料（圖 1）。

為實際驗證語義擴展策略的可行性，本文以「翠玉白菜」為案例，選定「質材」、「紋飾」、

表一 ICL 的輸入與標籤對範例 作者整理製表

給 LLM 的輸入與標籤對	
輸入	標籤
這部電影真的很感人。	正面
天氣太糟了，一整天都心情不好。	負面
我今天去上班了。	中立
要 LLM 處理的新輸入	
輸入	預期答案
午餐很好吃。	正面

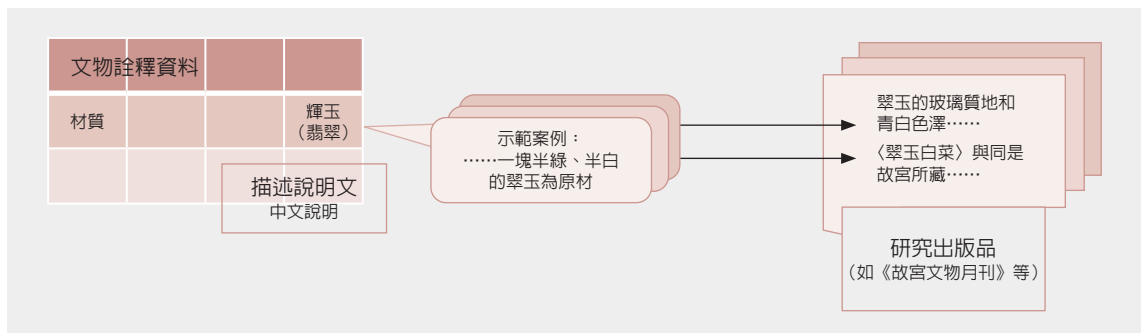


圖 1 語料生成架構雛形 作者整理繪製

「色澤」等三個欄位做為語意擴展對象，擴展內容參考自「說明文」欄位的「描述說明文」和「出版紀錄」欄的「中文說明」（表二）作為語境補強來源。

「描述說明文」和「中文說明」常混合「質材」、「紋飾」、「色澤」等資訊（見表二），本文首先將其逐一切分成每段約 30 字的語料段落，接著以「欄位——原內容」作為示範案例，引導商用 LLM⁶ 將語料段落分類至較適當的欄，提示設計如下：

請根據以下段落內容，判斷它最適合對應哪個欄位，並提取其特徵。

可選欄位包括：

- 質材：礦物、玉石珠寶、輝玉（翡翠）
- 紋飾：動物、昆蟲、蟲斯、器身、雕刻
- 色澤：以綠色部份琢為菜葉，白色處施作菜身

初步的分類結果並非絕對正確，仍須結合人工審核以確保準確性，本文依據各欄位的原內容調整分類結果或剔除不適當的段落，同時增設「其他描述」欄位，以容納部份未被 LLM 分類但具備關鍵資訊的內容段落。擴展後的文物詮釋資料（表三）將做為下一階段的示範案例。

建立語料訓練資料集

表二已可用於語料訓練，若能再擷取學術論文、研究報告等非結構化知識將能進一步充實訓練資料集。本文接著自《故宮文物月刊》擷取與文物詮釋欄位相關之段落，以〈寶石盆景〈翠玉白菜〉的聯想與賞析〉⁷一文為例，先將全文切分為數個段落，每段約 60 至 80 字，接著就表三的「欄位——擴展後的內容」建構示範案例集合，透過提示設計引導商用 LLM 進行分類，處理流程與前述階段大致相同，最終共產出 55 筆

表二 〈翠玉白菜〉其及描述說明文（中文說明）作者整理製表

	描述說明
	翠玉白菜是國立故宮博物院最受喜愛的藏品之一。工匠順應翠玉天然的色澤，以濃重的深綠色表現層層包覆的菜葉；……
	這件與真實白菜相似度幾乎百分百的作品，是由翠玉所琢碾而成，親切的題材、潔白的菜身與翠綠的葉子，……

資料來源：取自國立故宮博物院「故宮典藏資料檢索」網站：<https://digitalarchive.npm.gov.tw/Collection/Detail/3?dep=U> 及「故宮精選賞析」網站：<https://theme.npm.edu.tw/selection/Article.aspx?sNo=04001080>，檢索日期：2024 年 7 月 29 日。

訓練樣本，其中屬「質材」欄位者 2 筆、「紋飾」9 筆、「色澤」14 筆，其餘則歸類於「其他描述」欄位。表四中列舉各欄位範例各兩則，以茲說明。這些資料經整理後可用於 AI 訓練，助力優化故宮的數位服務。

段落字數之設定對欄位分類結果具有關鍵影響。當段落過長（如逾 120 字），語意涵蓋範圍廣泛，分類結果多集中在「其他描述」欄位，難以對應至如「質材」、「色澤」、「紋飾」等；反之，段落過短（如少於 50 字）雖有助於提升分類分布的平均性，然其語義支撐不足，易造成語料破碎化，降低訓練效能。本文認為 60 至 80 字數區間較為適當，並視分類結果和學術論文的語句自然語意切分，以兼顧段落的語義完整性與分類的對應精度。

此外，「其他描述」比例偏高係因示範案例選擇「材質」、「紋飾」、「色澤」等欄位，其他重要欄位尚未建立對應，未來增列更多欄位後將有助提高「語意對齊」比例，降低「其他描述」筆數。超出文物詮釋資料範疇的語義

表三 擴展後的文物詮釋資料

作者整理製表

欄位	原內容	擴展後的內容
質材	礦物、玉石 珠寶、輝玉 (翡翠)	• 這株翠玉白菜，就是以一塊半綠、半白的翠玉為原材。
		• 翠玉白菜以一塊半白半綠的翠玉為原材，運用玉料自然的色澤分布琢磨。
		• 其中不乏裂璽、斑塊，若是在光素的璧、鐲等裝飾用玉上，這便是一種瑕疵。
紋飾	動物、昆蟲、 螽斯、器身、 雕刻	• 菜葉頂端的螭螭和蝗蟲，瞬間帶進了鮮活的田園氣息。
		• 昆蟲則象徵多產；祈願新婦能子孫衆多。
		• 一株菜綠梗白、鮮活欲滴，葉片上還停留了兩隻昆蟲兒的白菜，吸引了大家的注意。
色澤	以綠色部份 琢為菜葉， 白色處施作 菜身	• 此器利用玉之原色施以巧琢，以綠色部分琢為菜葉，白色處施作菜身，酷似實物。
		• 工匠順應翠玉天然的色澤，以濃重的深綠色表現層層包覆的菜葉；白色部分雖有裂痕及雜質，但在巧妙的安排下，轉化為新鮮、飽含水分的白菜莖部。
		• 用翠玉瑯為白菜，依玉之原色施以巧琢，以綠色部份琢為菜葉，白色處施作菜身。
其他 描述	新增	• 隨著時間層累著衆人的體會與解釋，翠玉白菜的自然色澤、人為形制、象徵意念，終於獲得和諧而圓滿，遂成一件不可多得的珍品。
		• 從編號可知，此件翠玉白菜原是永和宮的陳設器。它本來「植栽」於一個四瓣花形的掐絲琺瑯花卉紋盆中，是所謂「盆景」的一部分。
		• 這是匠人在預設的條件下，透過充分理解質材特色，以激發個人創意，協調天然與人為的創作方式，其結果往往出人意表，卻又似乎理所當然，恰如其分，所以稱巧，是所謂「巧雕」。

表四 欄位及其所屬訓練資料集

作者整理製表

欄位	訓練資料內容
質材	• 翠玉的玻璃質地和青白色澤，應是玉匠雕琢白菜的原始靈感，但是像〈翠玉白菜〉這件「俏色」與「巧雕」兩者絕妙配合的藝術品，可謂「天下無雙，古今鮮對」，它的珍貴稀有應列入將來是否晉級為「國寶」的首要標準。
	• 〈翠玉白菜〉與同是國立故宮博物院所藏的另外兩株玉白菜，〈翠玉小白菜〉（圖2）和〈翠玉白菜花插〉（圖3）也反映了清末宮廷對玉器的新品味——出產於雲南和緬甸的翠玉取代了新疆的和闐玉，與〈翠玉白菜〉在年代上息息相關的慈禧太后（1831-1908），便是當時最有名的愛「玉」者。
紋飾	• 故宮所藏南宋許迪（十二世紀）的〈野蔬草蟲圖〉（圖4）是現存最古老的白菜畫，畫裡的「野蔬」，葉大而圓，色澤淡綠，葉梗粉白，形色很接近「臺灣小白菜」，應是宋人所說的「揚州菘」或「白菘」。
	• 螽斯和蝗蟲是明清藝術中多種草蟲的通稱，用以寓意「多子多孫」，宮廷藝術家多取在室內蓄養過冬的螭螭和蚰蚰代表螽斯、絡緯或蝗蟲。
色澤	• 以盆景方式展示的〈翠玉白菜〉有如尚在生長的真菜，顯得更具鄉土氣息，從正、反、側面都可以看到青白相持的菜葉、潔白的菜身和清晰的葉脈。
	• 又或者玉匠別有用心，藉由盆景〈翠玉白菜〉提醒養尊處優的帝后貴妃，時時不忘民間疾苦？白菜在文學傳統中用以象徵「清白」可能始於杜甫（712-770），〈行官送菜〉一詩先提到苦苣和馬齒葉，後感嘆「青青嘉蔬色，埋沒在中園」，「嘉蔬」指的就是白菜，暗喻詩人自己為官清廉卻不得重用。

續表四 欄位及其所屬訓練資料集

作者整理製表

欄位	訓練資料內容
其他描述	<ul style="list-style-type: none"> 學者專家已指出，植在花盆裡的〈翠玉白菜〉，屬於清初康熙乾隆三朝時始盛行，主要用於宮中陳設和帝王祝壽的「寶石盆景」。
	<ul style="list-style-type: none"> 如果〈翠玉白菜〉在清代即見著錄，它的正名應是「晚崧」或「□崧」，因此〈翠玉白菜〉也可以寓意「凌冬不凋」的清白君子。



圖2 清 翠玉白菜 國立故宮博物院藏
故玉 002662



圖3 清 翠玉白菜花插 國立故宮博物院藏
故玉 007852



圖4 南宋 許迪 野蔬草蟲 冊
國立故宮博物院藏 故畫 001246

也可透過語意標註、主題分類等方式補充。整體而言，本文目標在於建立一套從文物詮釋資料與研究出版品中擷取訓練語料的流程，未來在擴充示範案例、標註語意或補充主題分類後，語料分類的精確度可望持續提升，更完整反映故宮文物詮釋知識的深度與脈絡。

挑戰與應用

本文所建構之語料生成架構雛形，係以故宮既有之文物詮釋資料為基礎，進一步自本院出版之學術論文中擷取具知識密度之段落，以轉化為可供 LLM 訓練之語料。選擇 ICL 作為主要策略，主要是基於語料特性與實務限制所致，首先，文物詮釋資料的格式清楚且結構穩定，具有明顯的模式可供模型參照，讓模型在無需額外微調的情況下即可學習輸出格式與語意關

聯；其次，本文重點在於驗證 LLM 在低資源條件下，從研究出版品抓取適當語料的可行性，而非建置完整的微調或人工評分流程。在時間與資源有限的情況下，ICL 能以最低成本快速完成初步驗證，並可作為後續若進行更完整訓練（例如微調或人工評分介入）時，設定基準與初步評估的堅實基礎。

然而在實際操作過程中，尚面臨數項挑戰，亟需進一步探討與克服。首先，減少人為干預、提升自動化程度。現有文物詮釋資料對於 ICL 所需之線索資訊仍屬不足，致使文物詮釋欄位的擴展結果（見表三）需經人工審閱與修訂，刪除不恰當內容，或重新歸類至較為適切之欄位。其次，由於不同學術論文撰寫風格不一，段落長度與敘述邏輯差異甚大，若以固定字數進行切分，常需視分類結果再行調整，亦增加

人力成本與執行負擔。本文認為，未來可考慮引入語義分析、關鍵字萃取與自動分類模組，以強化欄位判斷與語料標註的準確性，進而提升語料建置效率。

第二，優化故宮學術論文之讀取與處理效率。由於部分學術論文之排版格式較為特殊，且原始電子檔經過影像壓縮處理，難以順利轉換成可解析的文字，時常出現文字亂碼、欄位錯置等問題。本文曾嘗試數種解析工具，惟多無法正常辨識內容，只能逐行逐段剪貼至新建檔案後方能解決。為提升學術論文處理效率，未來可考慮導入光學文字辨識（optical character recognition）模組，以處理圖像化或壓縮後的學術資料，進一步擴增語料來源的可讀性與應用範圍，並降低人工作業成本。

第三，本文僅以《故宮文物月刊》為案例，係方便說明與展示方法流程，但所採用的語料整理與結構化方法完全適用《故宮學術季刊》，

其提供更豐富的專業知識與論述能確保未來 AI 服務在不同層次知識上的精確性。

本文所建構的語料生成架構雛形，正是將故宮豐富典藏與學術研究轉化為「高品質 AI 語料」的關鍵實踐，所產出的訓練資料集兼具知識密度與語意精確度，將確保 AI 應用和服務能建立在專業的知識基礎上，省去重新編寫訓練語料的成本。據此，未來可朝以下三個方向推展：第一，針對單一展件或特定展覽，訓練具備故宮識別性與詮釋深度的導覽內容；第二，藉由 LLM 的跨語言能力拓展多語系導覽服務，強化國際觀眾的理解力與文化參與度；第三，整合語意、圖形化等技術改善典藏檢索服務，優化跨文物、跨時期與跨類型的語意理解與互動應用。總結而言，透過這些深化的 AI 應用，故宮的深厚文化資產將得更智慧、更友善地跨越時空限制，實現與廣大觀眾個人化的連結與互動。

作者任職於本院數位資訊室

註釋：

1. Zili Cai, "AI Empowering Museums: Current Applications, Challenges, and Future Trends," in *Proceedings of the 2025 4th International Conference on Humanities, Wisdom Education and Service Management (HWESM 2025)*, ed. Khan I. Alam et al. (Paris: Atlantis Press, 2025), 301-311, accessed August 1, 2025, <https://www.atlantis-press.com/proceedings/hwesm-25/126012062>.
2. Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj and Marah I. Abdin, "On the Diversity of Synthetic Data and its Impact on Training Large Language Models," *arXiv [Preprint]* (2024), accessed November 19, 2025, <https://arxiv.org/pdf/2410.15226>.
3. Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," (paper presented in *Conference on Fairness, Accountability, and Transparency (FAccT '21)*, Virtual Event, Canada, March 1, 2021).
4. 張志光，〈故宮藏品管理系統升級之器物詮釋資料需求探討〉，《博物館與文化》，28 期（2024.12），頁 209-246。
5. S.Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi and L. Zettlemoyer, "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?," *arXiv*, Published in Conference on Empirical Methods in Natural Language Processing (2022), accessed August 1, 2025, <https://arxiv.org/pdf/2202.12837>.
6. 本文選用 Llama 4 Scout 17B-16E 模型，該模型具備理解長篇文字與生成連貫說明內容之能力，適用於文物詮釋與導覽文字生成之研究。
7. 王柏樺，〈寶石盆景〈翠玉白菜〉的聯想與賞析〉，《故宮文物月刊》，444 期（2020.3），頁 46-58。