# 范式变革与文明赓续：人工智能时代的古典学[①]

**摘要：**2025 年初，deepseek 的爆火让国内人文学术圈的 AI 应用与研究的热情由冷淡走向高涨，更让原本因脱离于现实生活而广受质疑与自证困境的古典学，在技术挑战的危机中找到了范式突破的勇气。古典学界正在形成某种共识：研究古典，是为更好地观照现代，而坚守传统，则旨在创造性地开辟未来。因此古典学的价值，不仅在于考古钩沉，还在于透过古今对话使人类共享智慧遗产并激发思想的创生。正是基于这一范式论的考虑，古典学才不得不直面 AI 时代的挑战，以建构赓续人类文明的新的学术范式。

**关键词：**古典学；人工智能；典故；典故学

## 引言：人工智能时代与古典学的基本境遇

雷•库兹维尔（Ray Kurzweil）曾指出，到 2045 年，通过把大脑新皮层与云端的人造新皮层无线相连，这样人类的智能将增加 10 亿倍。[②] 而英国牛津大学人类未来研究院院长尼克•波斯特洛姆（Nick Bostrom）则在《超级智能》一书中预测：“人把计算机变成了超级自我，但自我该如何做，才能保证人本身才是目的，而不被自己发明的手段所打败呢？”[③]而年初 deepseek 的爆火，也让国内人文学术圈的 AI 应用与研究热情由冷淡走向高涨。截止 2025 年 9 月 27 日，仅知网上有关人工智能（artificial intelligent）与人文（humanity）之关系的专题性讨论文章，就已经达到 2515 篇，其中核心期刊或 C 刊论文共计 731 篇，且处于快速攀升趋势。比如 2024 年发表达到 135 篇，仅较上年增加 24 篇，但 2025 年目前便已达到 166 篇（部分期刊的发表具有延迟性）。这些 AI 论文涉及了众多学科门类，其中来自于人文社科领域的 C 刊（含扩展版，但不包括繁体排版的集刊）文章便达 616 篇，相关数据与科睿唯安网所呈现的逐年上升趋势也基本一致（共计 6389 篇，文史哲类——含哲学、文献、宗教、艺术、历史以及科学哲学等——共 745 篇）。令人遗憾的是，不断增长的 AI 论文中，鲜有触及“古典学”的前沿思考，偶有论及，也尚未触及“古典”对于 AI 技术如何落地等实践意义[④]。这便让原本就因脱离于现实生活而广受质疑与自证困境的古典学范式，益发陷入技术挑战的危机之中。因此，本文乃就人工智能时代之古典学的存在形态、传播路向、

---

[①] **人机协作写作说明：** 本文在部分写作环节中使用人工智能工具（ChatGPT，GPT-5）提供语言与资料辅助，所有观点与结论均由作者本人独立完成并核定。

[②] Ray Kurzweil. Get ready for hybrid thinking, TED Blog, “The hierarchy in your brain: Ray Kurzweil at TED2014”, https://www.ted.com/talks/ray_kurzweil_get_ready_for_hybrid_thinking 2014-03-20（2025-06-21）.

[③] [英]尼克•波斯特洛姆：《超级智能：路线图、危险性与应对策略》，张体伟、张玉青译，北京：中信出版社，2015 年，第 XIV 页。

[④] 如曾建华：《人工智能与人文学术范式革命——来自 ChatGPT 的挑战与启示》，《北京师范大学学报(社会科学版)》2023 年第 4 期，第 78-88 页；胡易容、尹婧雯、李婉婷：《数智媒介时代的意义危机与古典符号学回归——2024 年度全球传播符号学前沿》，《新闻界》2025 年第 1 期，第 63-73 页；姜秀敏、张嘉印：《古希腊提示对人工智能图像生成的影响：一种新的教育范式》，《北京行政学院学报》2025 年第 3 期，第 31-42 页；Anna Kalargirou et al.. The Impact of Ancient Greek Prompts on Artificial Intelligence Image Generation: A New Educational Paradigm, *AI* 2025, 6(4), p. 81, https://doi.org/10.3390/ai6040081; Joshua Brecka. Creatures and creators: God, humanity, and artificial general intelligence, *Religious Studies*, DOI: 10.1017/S0034412525000289, 2025-04-30; Radek Schuster & Hamid R. Ekbia. Language in the Godless Age of AI, Social Epistemology, DOI: 10.1080/02691728.2025.2484743, 2025-05-04.

接受方式与创作研究所引发的知识谱系与思维方式的重要变革，尤其是古典学在数字化时代的知识重构与文化塑造等深层问题，试图给出专题性的讨论和应对性的方案。

# 一、何以古典：古典学的现代范式及问题

当我们讨论“古典学”（Classics 或 Classical Studies）时，我们首先要明确的就是，我们所讨论的实哪个层面上的古典学。广义上的“古典学”，乃是指专门研究人类自古以来之人文经典与传统的学问，广涉语言、文献、历史、艺术、宗教、哲学等多个关连领域的学科属性和理论范式，是典型的跨学科的综合性知识体系，甚至涵盖了文艺复兴以来的所谓人文学科（studia humanitatis）。正所谓，人文经典长存之处，就存在某种形式的古典学。[①]而狭义的古典学，通常被定义为专门研究古希腊、古罗马文明时期之原创性、范式性典籍（即经典）的学科，并长期被视为西方文明的基石和精英教育的核心，乃至成为欧洲中心主义者之理论偏见的源头。因为“Classics”之名本身就暗含偏见，从而隐性地排斥了其他古代文明，如古代中国文明与印度文明等。[②]

事实上，现代学科属性的古典学，是随着经院神学的消解与人文主义尤其是科学主义的发展而逐步确立的。18 至 19 世纪的欧洲，特别是在德国，古典学已转型为所谓 Wissenschaft，即系统研究古代文明的“科学”。为了区别于此前的古典学，德国学者甚至还创造了一个新的术语“古代学”（Altertumswissenschaft，即对古代世界的科学研究），并强调有关古代世界的一切研究都应遵循一种统一而严谨的科学范式——以实证为底色的现代文本批校技术—— “新语文学”（new philology）应运而生[③]。比如，1807 年，德国语文学家弗里德里希·奥古斯特·沃尔夫（Friedrich August Wolf，1759-1824）便在其划时代的著作《古代学的表述》中，正式将“古代学”的研究边界，从希腊、拉丁语言文学扩展到古代历史、考古学（艺术与文物）、钱币学、碑铭学、神话学和哲学等人文学术的各个“古代”领域，并试图将其作为一个固有的跨学科领域的现代学术框架予以确立[④]。这意味着，18-19 世纪的“古典学”，不再只是通过阅读名著以诉诸道德教化的文艺复兴运动，而是要求运用一切可用的学术工具来重建和理解完整的古希腊-罗马世界的知识体系。简言之，此前的古典学，着眼于经典文本的阐释（附带关注历史），而此后的“古代学”，则更倾向于对整个古代文明的全面探究与重构。

然而，自 20 世纪开始，盛极一时的古典学逐渐被严格界定的文学、历史与哲学所肢解，一方面，古典语言（希腊语和拉丁语）作为必修科目的局面不复存在[⑤]，另一方生源数量也逐年下降，从而导致相关系所被迫精简。与此同时，学科化的古典学所必然导致的认识论局限也逐渐暴露无遗，一个不容回避的事实就是，其赖以立论的证据本身也在学科化的知识谱系中变得支离破碎且不乏偏见。这在很大程度上限制了古典学者提出创新性问题的能力也让我们对于古代世界的认知，越来越依赖于一系列偶然保存下来的文献、器物，尤其是那些经

---

① 王中江：《中国古典学的场域和特性》，《中国哲学史》2025 年第 2 期，第 5-6 页。

② 参 James I. Porter, “What Is ‘Classical’ about Classical Antiquity? Eight Propositions,” *Arion* 2005, 13(1), p. 28; Martin Bernal, Black Athena, vol. I, London: Free Association Books, 1987, p. XX.

③ August Boeckh, *Encyklopädie und Methodologie der philologischen Wissenschaften*, ed. Ernst Bratuscheck, Leipzig: Teubner, 1877, p. 11; p. 21; p. 56.

④ Friedrich August Wolf, *Darstellung der Alterthumswissenschaft nach Begriff, Umfang, Zweck und Werth,* Berlin: Realschulbuchhandlung, 1807; repr. with afterword by Johannes Irmscher, Weinheim: VCH, 1986, pp. 5-7.

⑤ 例如英国学校尤其是高校，便率先取消了希腊语和拉丁语的强制性学习要求，到 21 世纪初，只有极少数公立学校仍开设相关课程；美国也不再将拉丁语和古希腊语作为中学阶段的标配课程，而仅仅是少数学校的选修课程。

过个人选择或现代性阐释才得以传世并被理解的所谓“经典”。于是，古代经验领域（如通俗文化、乡野生活、女性和贫民生活、边疆民族生活等）之复杂性及其与生活世界的关联性、差异性和延续性越来越被忽视，以至于原本以现代性启蒙为初衷的整体性学问——古典学——日益走向自我的对立面。一些古典学者甚至将古代理想化，视之为现代的隐然或更具优越性的替代者。正是从这一时期开始，古典学逐步脱离于现实，并因其“因循守旧”而遭遇学科式微与范式转型的困境，原本作为显学的古典学，逐步陷入大众质疑和价值自证的困境。

面对时代现实的困局，古典学者开始广泛吸纳人类学、社会学、文学理论等跨学科的多元视角，进而追问古典叙述中究竟有哪些声音被迫缺席（例如女性、下层阶级、战败者），以及这些叙述得以构建的深层原因。在大数据、人工智能等信息科技元素的加持下，有的研究者，甚至已经不同程度地运用计量分析、大数据统计乃至人工智能等科技手段进行古史研究[①]。这些方法上的突破，让原本式微的古典学重新焕发出生机，但日益量化与科学化的发展路径，也加剧了古典学内部日益割裂与孤立的碎片化倾向，从而在处理连续与变革等文明赓续问题上难以取得有效的平衡，甚至还因为过分强调不同范式之差异而走向思想的断裂[②]。如果抛开文化的传承与迭代，很多号称“冷门绝学”的古代研究，已逐步成为某种缺乏现实意义的个人性的猎奇行为。

随着学科的职业化与高度专业化，学界对何谓专长设立了严格标准，进而形成极度聚焦的研究范式。于是，学者们为了在各自圈层的竞争中胜出，不得不严格遵循“不发表就出局”（Publish or Perish）[③]的铁律。为此，博士生或“青椒”，只能选择更容易出成果的某一经典作家、作品或某处考古遗存，以皓首穷经的态度，精研只有极少数人才会阅读的微观/冷僻文献。这样做，确实能让学者在尽可能短的时间内产出较有深度的研究成果，进而在圈层立足甚至成名成家。但这种“挖掘”式研究，也进一步摧毁了学术共同体的交互生态，乃至让许多青年学者不得不选择杀鸡取卵、饮鸩止渴或“竭泽而渔”的学术道路。

吊诡的是，在西方古典学全面“衰微”[④]的同时，以其为范式的所谓“中国古典学”，正以燎原之势取代整个中国传统的治学之道。一个极为常见的现象就是：西方各“学”无需冠以“西方”之名即可彰显其学术之“正统”，而与中国相关的学术，反倒需要刻意加上“中国”之定语以彰显其“从属”地位。这意味着，即便研究内容属于中国，也必须奉行西学的方法论范式，否则就会遭遇边缘化乃至“民科”化的命运。北京大学吴飞教授便明确指出，“旧学的研究方法必须和一定程度的现代理论相结合，才能得到广泛的认可，并在社会文化中产生全局性的影响，推动学术范式的形成。”而吴飞教授所指向的“现代理论”即“西方理论”，而且他也毫不讳言，“新式主流学术的确立”，应是在“大量接受了西方理论的前提下完成的”，而且“以西方理论来解释中国的材料”，已经成为“大家普遍接受的研究方式”。而由此导致的问题就是，“中国文化的味道”变了，因此吴飞教授倡导一种更为成熟的“以中解中”的新范式，也就是张祥龙教授所谓“更关注于揭示意义和存在的发生结构及动态维持自身的方式”。[⑤]

这一系列问题，不止存在于中国，也反映在整个古典学界所隐含的文化与意识形态偏见。

---

[①] 比如 Henriot, Christian, “The AI-Augmented Research Process: A Historian’s Perspective, *arXiv preprint*, 2025, arXiv:2508.01779. https://doi.org/10.48550/arXiv.2508.01779

[②] 聂敏里便警告我们，必须清晰区分古代与现代的知识范式，否则就会“把本质迥异的知识体系混为一谈”，在实际上已经发生断裂之处创造出虚幻的连续性。其言下之意在于，“古代科学”与“现代科学”、“古代学术”与“现代学术”等术语，指的是质上截然不同的认识论框架，如若像 19 世纪的辉格派史学家那样，将二者视为一脉相承的连续体，则会在认识论上并现代古典学者批评为不够成熟。

[③] Harold J. Coolidge and Donald H. Piatt, *University Administration: The Roles of Faculties and Students,* New York: Columbia University Press, 1932, p. 270.

[④] 董平：《中国古典学的“返本”与“开新”》，《中国哲学史》2025 年第 3 期，第 5 页。

[⑤] 吴飞：《寻求现代中国学术的成熟范式》，《北京大学学报(哲学社会科学版)》2015 年第 1 期，第 28-32 页；张祥龙：《中国研究范式探义》，《北京大学学报(哲学社会科学版)》2015 年第 1 期，第 23 页。

因为古典学的兴起和发展始终植根于特定的文化语境，本身就下意识地将希腊罗马文化定位为独一无二的经典（即典范或更高级），而将其他古代文化视作次要或边缘。因此，它不可避免地体现和延续了多种传统偏见——尤以欧洲中心论、精英主义为甚。例如，19 世纪的大英帝国主义者便经常将自身比附罗马帝国，并试图通过古典教育向统治阶层灌输身份认同。[①]因此，虽然中国古典学的研究路径，仍未摆脱借用西方古典学，以研究先秦两汉为内核的中国古代文明的范式性局限，但其作为中华文化之多元、开放、自信与自主性重构的范式诉求，却很值得重视。[②]事实上，早在 20 世纪初以来，章太炎、胡适、梁启超、钱玄同、顾颉刚等学者便开始使用作为“古典学”的国学或国故学一词，并试图以之框定现代学科体系下的中国古代之文化遗产。[③]也正因为此，尽管当下的“中国古典学”之本意，乃在于逐步转向以本土方式（如出土文献）重建其学术自主性，进而振兴以经学为本源与中心的传统学术。[④]但同时，又不可避免地将“中国之学”（可简称国学），纳入到类似于西方古典学之所谓现代学术框架之中，从而与其独立、自主、多元之初衷背道而驰。之所以会造成如此分裂的局面，根本原因就在于前述缘木求鱼式的范式诉求。

一方面，西方古典学的学术范式与方法论，早已主导整个人文学术（自然包括研究中国古代之学），另一方面，以经学为中心所构建的中国传统学术谱系，又与强调中心、权威与圈层门槛的古典学范式和观念颇为契合。于是，就连古典学者最引以为傲的浩如烟海的古代中国典籍，也逐步沦为现代西方视角的研究资料。有鉴于此，中国学者始逐步意识到这样一个事实：只有发展出真正的中国学术范式，中国的古代经典才可能找到“真我”（即所谓自主知识体系），并实现其创造性转化，最终在国际学术舞台上提供一种不同于现代西方的学术视野和路径。[⑤]只不过，以争夺国际话语权为导向的中国古典学，并非以获取历史知识并重建一个逝去的历史世界为目标，而更多是为了延续并阐释一个活的传统，进而为现实的道德实践和治国理政提供理论指引。简言之，中国古典学虽同样致力于以整体性的视角去研究古代经典，但这种研究不仅是为了理解古代，更是为了取法前人以促进当下的治理、伦理和文化之发展。

然而，这类宏大构想一旦脱离了历史语境，便很可能导向另一个令人忧虑的结果，即在“复兴传统文化”等相关政治倡议的推动下，所谓“中国古典学”将从原来的西学附庸，逐步走向另一种功利性的学科化乃至建制化的附庸——甚至进一步丧失其学术的独立性和自主性，逐步沦为政治宣传的工具。例如某些新开设的古典学学院、专业、学科以及学术项目和刊物，虽然其基本的范式导向仍在于将本土的知识体系与全球的学术规范相融合，但呈现的研究成果，却多是以全球战略的政治文化为主导的“资料汇编”，很少能看到批判的锋芒，更缺乏持久而深刻的文化洞见——而这才应成为当下中国古典学重构的内在支撑。

因此，要重构现代中国之古典学，我们首先要做的，就是对古典学进行去殖民化或去中心化处理，诸如通过当下仍然强势的“接受”视角，完成古典学的多元视角的建构，并以此为前提，开放、流动地讨论不同时代和文化（包括非西方文化）中的古典遗产如何被接受和重释等问题。其次，就是要尽可能地避免古典学的意识形态化或工具化。最后也最为重要的

---

① James Bryce, *The Ancient Roman Empire and the British Empire in India: The Diffusion of Roman and English Law throughout the World*, London: Oxford University Press, 1914, pp. I-II; Christopher Stray, “Classics in the Curriculum before the 1960s,” in James Morwood (ed.), *The Teaching of Classics*, Cambridge: Cambridge University Press, 2003, pp.1-5.

② 这方面较有代表性的观点，见刘钊、陈家宁 ：《论中国古典学的重建》，《厦门大学学报（哲学社会科学版）》2007 年第 1 期, 第 5-13 页。

③ 参见顾颉刚：《古史辨》（第一册），上海：上海古籍出版社，1982 年，第 23-24、102-105 等页；《古史辨》（第二册），第 320、331 等页。

④ 裘锡圭：《出土文献与古典学重建》，载复旦大学出土文献与古文字研究中心编：《出土文献与古典学重建论集》，上海：中西书局，2018 年，第 13-37 页。

⑤ 参谢乃和：《中国古典学的学理逻辑与构建路径》，《中国史研究》2025 年第 1 期，第 5-22 页。

则是，要建立区别于当下文、史、哲之学科分工的独立性、整体性、专业性的古典学学科，从而避免原本作为整体研究对象的古典学问，被肢解为破碎的知识系统，并确保各个子系统不至于形成孤立、封闭、僵化的话语体系。可喜的是，自 2025 年起，中国国家社科基金申报中，首次将古典学作为一级学科纳入到申报体系之中，至此，一个具有文明互鉴、文化包容与方法多元的古典学学科才算是真正完成了独立的学科建制，但也面临着前所未有的技术冲击——来自 AI 的挑战。

## 二、范式重建：面向 AI 时代的古典学

裘锡圭先生曾指出，古典学（与任何史学领域一样）会周期性地因“观念、方法的更新或重要新材料的发现”而重塑，有时甚至经历“剧烈的变化”，相当于某种范式的重建。[①]而在古典学的变革浪潮中，人工智能的发展所带来的挑战和机遇，无疑是空前的。正如网络盛传为 MIT 教授谢丽·特克尔（Sherry Turkle）所警示的那样：“技术不仅仅让我们便于行事，也改变了我们自身：不仅改变我们的所作所为，也改变了我们是谁。”[②]一方面，以 AI 驱动的文本挖掘、语言建模、图像识别和知识图谱等工具，正逐步介入古典学的研究，从而不断拓展古典学者的智识与能力边界，以往难以想象的艰巨工作（如跨领域的观念史、知识史等）如今正成为可能。另一方面，学界也日益意识到，不加审视地接入 AI，可能带来难以估量的后果，甚至威胁整个古典学的人文本质。尤其当我们将 AI“代理者”视为潜在的共同研究者时，关于知识阐释的真实性、深刻性、伦理性以及由此带来的文化语境的丰富性等问题便将一一凸显。因为 AI 在帮助古典学者加速甚至完成文献资料的誊抄、校勘、标注与检索等繁琐工作的同时，也在悄无声息地改变古典学原本具有范式性意义的方法论与价值属性，还可能从根本上重构整个古典学的研究范式，从而彻底改变我们诠释古典智慧、定义专业知识乃至定位人类研究者角色的方式。换言之，AI 所带来不止是现代意义的技术革命，还可能是一场颠覆人类文明之根基的哲学危机。这一史无前例的发展歧路，让我们不得不面对这样一个问题：如何将 AI 技术引入古典学研究，以构建新的足以捍卫人类尊严的古典学范式？

我们首先可以考虑的路径，就是对 GPT-4、grok-4、deepseek-R1 等在现代语言处理方面已取得较大成功的 LLM 进行针对性的微调，进而结合检索增强生成（Retrieval-Augmented Generation，RAG）[③]技术，或直接将其套用于相应的古典学知识库，使之具备古典语言、文字和文本方面的识别、校勘、标注以及释读等专业“技能”。[④]比如，依托腾讯 ima 知识库的混合检索架构，可将中华书局点校本《十三经》《二十五史》《新编诸子集成》《中国古典文学基本丛书》等精校典籍，再按 TEI-All 学术标准进行语义切分与元数据标引，建立可回溯的“正文—校勘—注释”三元分组索引。在此基础上，我们还可通过增量式持续注入新出版本数据，即可在不改动模型参数的前提下，显著提升系统对古籍自动标点、校勘比对、注释唤出与现代翻译的精度与可验证性。笔者发现，某些经过微调和强化学习后的模型，甚至还

---

[①] 裘锡圭：《中国古典学重建中应该注意的问题》，载复旦大学出土文献与古文字研究中心编：《出土文献与古典学重建论集》，第 1 页。

[②] Sherry Turkle, “The Documented Life,” The New York Times (Sunday Review), 15 Dec. 2013 (2025-6-28).

[③] RAG 是一种结合信息检索技术与生成式人工智能的自然语言处理模型架构。其核心思想是通过从外部知识库动态检索相关信息，辅助大型语言模型（LLM）生成更准确、上下文相关的文本内容，从而解决传统 LLM 的知识滞后、幻觉（虚构信息）及专业领域知识不足等问题。

[④] 参赵浜、曹树金：《生成式 AI 大模型结合知识库与 AI Agent 开展知识挖掘的探析》，《图书情报知识》2024 年 11 月 4 日网络首发稿，第 4-13 页；丘子靓等：《基于大语言模型的文史知识库构建研究》，第 58-75 页。

能实现多语言处理与对齐，或进行跨文本互文性分析、自动识别文本间的影响与引用关系等，如补全《孟子》阙文、为《楚辞》中的冷僻典故作注。这种基于 AI 大模型的智能知识库，不仅极大地降低了知识库建设的技术门槛和训练成本，也为兼通数字技术的古典学者们建设“中国古典知识库”的设想提供了新的可能路径。[①]

然而，古典语言尤其是古代汉语，其表达具有很强的歧义性，当文本内容涉及典故时，更会造成因文本信息的嵌套、压缩与丰富而导致的理解错位、迁移或次生。这些信息化的语义传递，无疑会给 AI 的深度学习带来巨大挑战，乃至造成不可避免的“幻觉”。加之，机器虽能“在特定任务领域里表现卓越”，却“不具备主观体验、道德判断与对生命意义的反思能力”[②]，因此，无论 LLM 在古典语料上训练得如何完美，它也不可能真正领悟诸如“仁”（benevolence）或“道”（the Way）这类概念所蕴含的哲学和道德意义，充其量只能基于神经网络所生成的模型思维层面的上下文“理解”，其本质是“对语言的高性能预测”[③]，而非人类基于具身经验所获得的系统性省思和悟解。换言之，AI 可以模仿所有词语的用法，甚至生成关于它们的貌似合理的句子，但始终无法体验其背后所承载的人文关切。例如，即使我们训练出一个专业的古典汉语 LLM，其可供学者快速获得初步译文或在全 corpus 范围检索相关典故，但它也只能依据有关典故的知识解释这一典故，而无法体验典故中所承载的人类的复杂情感与丰富联想，尤其是当某个典故本就涉及人的具身体验和价值判断时（如望梅止渴、仁者爱人），AI 的“理解”就越可能缺乏真实性和有效性。由此可见，我们引入 AI（这个“人工第三者”）**“本身并不会让行为变得更好或更坏”，其影响多“取决于具体情境、目标以及实施方式”**[④]。

因此，相较于 LLM 的微调与训练，知识图谱（Knowledge Graph, KG）与推理算法相结合的 AI 工具，则是助力古典学范式转型的更为可靠的实验路径。所谓知识图谱，是一种用图谱结构表示现实世界实体及其关系的语义网络，其本质是一个结构化的知识网络，具有结构化、语义化、动态性与多源融合（如整合文本、数据库、传感器等多模态数据）等特征。在学术研究中，知识图谱发挥着知识表示（knowledge representation）、知识获取（knowledge acquisition）、知识推理（knowledge reasoning）、知识集成（knowledge integration）和知识存储（knowledge storage）等作用，尤以 Google Knowledge Graph、维基百科、百度百科、万维网等最为典型。在知识图谱的构建中，必须借助推理算法，才能将碎片化的史料、文本、时空信息重新拼接成一幅可验证的“数智化”拼图。比如，王兆鹏等学者所创建的“唐宋文学编年地图”，便是一个以节点性的代表作家生平、经典文本、交游关系、概念等为线索，并可支持自动标点、平仄标注、笺注、翻译等九大功能的“数智化”地理知识图谱。

而且，知识图谱已经“作为一种类型的先验知识”，被辅助输入到很多深度神经网络模型之中，成为“用来约束和监督神经网络的训练过程”的知识嵌入机制[⑤]。一旦知识图谱搭

---

[①] 参刘石,孙茂松：《关于建设“中国古典知识库” 的思考》，《人民政协报》，2020 年 8 月 24 日第 8 版；陈力：《数字人文视域下的古籍数字化与古典知识库建设问题》，《中国图书馆学报》2022 年第 2 期，第 42-45 页。

[②] Mehmet Latif Bakış. "Can Artificial Intelligence Replace the Subject of Consciousness? A Comparison of Human and Machine." *Beytulhikme: An International Journal of Philosophy*, vol. 15, no. 1, 2025, pp. 93–118.

[③] 参 Geoffrey Hinton.“Will AI outsmart human intelligence? - with 'Godfather of AI' Geoffrey Hinton” *YouTube*, 2025 年 7 月 23 日 Will AI outsmart human intelligence? - with 'Godfather of AI' Geoffrey Hinton（2025-7-27）

[④] Yuval Haber, Inbar Levkovich, Dorit Hadar-Shoval, and Zohar Elyoseph, “The Artificial Third: A Broad View of the Effects of Introducing Generative Artificial Intelligence on Psychotherapy,” JMIR Mental Health 11 (2024): e54781.

[⑤] 清华大学人工智能研究院、北京智源人工智能研究院、清华—工程院知识智能联合研究中心：《人工智能之知识图谱研究报告（2019 年第 2 期）》，清华大学人工智能研究院，2019 年，第 11 页。

建完成，再以 NLP 技术将文本所蕴含的各种关系抽取出来，AI 就能进行真正的算法推理。斯时，我们只需给出简明的指令提示词，如：“找出宋代所有评论孟子‘仁政’的文本，并列出撰写这些文本的官员和他们所侍奉的君主。”那么 AI 就能跨越文学、哲学、传记三张知识网，给出系统性的解答。首先，AI 会锁定《孟子》里“仁政”的节点，进而顺着“评论→作者→官职→君主”的边-路拉回结果，最后给出针对性的解释。过去通过纸质阅读进行研究的人文学者，往往只能凭模糊记忆列举出朱熹、张栻等最为知名的学者，但在 AI +知识图谱的帮助下，学者便能在几分钟甚至更短的时间里，将度正、黄干这些容易被忽视的边缘人物一并提取出来，还能进一步建立他们与庆元党禁或吕祖谦等不同圈层人物之间的谱系线索。而且，即便是最高明的历史学家也只能做定性描摹的制度-社会脉络，也可在知识图谱与 AI 的加持下，形成一张以数据铺就的可视化的全景图。

不过，要想让学者真正信任 AI 的输出，我们还必须让它像写论文一样给出准确、精详的注解。对此，Matthew 等人在 XAI 的研究中已有所强调，认为事后解释、模型透明、交互式可视化缺一不可。[①]用古典学的话说，就是： 即便只让 AI 翻译一句诗，系统也要标出影响措辞的平行文本或关键词；如果把一批文章归为“道家无为组”，它就得彰显出“无为”“自然”这些高权重特征。如此，当学者追问“GPT/Kimi，你为何觉得这条注释相关”时，它们就能立刻利用自己算出的共现内容或语义链给出证据，并完全符合学术规范。更为重要的是，古典学的材料不止文字符号，还包括诸如甲骨、金文、手稿、碑刻、绘画、器物等实物材料，只有将这些全部录入由学者主导完成的知识图谱，我们从 AI 处所获取的知识才是真正可信的。目前，由国家牵头的相关数字化工程已经陆续启动，比如“数字敦煌资源库”便涵括了敦煌莫高窟的 492 个洞窟、6 万多号写本的高清影像和全文转录[②]。 当我们将有关实物材料的图像全部喂给 AI，那么我们就可以期待这样一种人机协作的研究范式：我们请 AI 通过视觉模型比对全部甲骨碎片，让其对拼接错误的甲骨进行拆解重拼，进而借助甲骨文字库和专业知识，以统计概率的形式判定尚未释读的甲骨文字；我们还可以将《鹿王本生》壁画和对应的经文写本对齐，让 AI 去发现第 12 格画面和“布施忍辱”段落在关键词上可能出现的高频共振，从而让跨媒介的可视化叙事成为“一键生成”的画面或视频。

不难发现，知识图谱+多模态 AI 工具，将使文字、器物、图像在一场古典学的对话中，促成其范式的转型。于是，原本无人问津的古典文本与知识，竟可被编码、存储并在新的媒介中复活，乃至重构出一个绚烂而真实的古典的生活世界——这在哪怕三年前还是一种“元宇宙”式的想象，而如今正在成为一种改变古典学范式的变革力量。这场技术革命所带来的高效与创新，正让古典学的固有范式发生深刻的变革，也由此引发了一系列亟待解决的范式重建问题。

首当其冲的，便是如何实现 AI 对古典文本的语境化理解，并与研究者的具身化理解有效融合。诚然，相较于古典学者，AI 有着惊人的记忆和模式匹配能力，但目前的所有 AI（包括最新发布的 grok-4）仍然缺乏对历史语境、作者意图或文化共鸣的真正理解。比如，当一位古典学者读到唐诗中的“孤雁”时，他会立刻意识到这不只是一个词语，而是一个复合了无数典故的“文化意象”，其在中国文学传统里往往象征着孤独与流放，并为之深深感动。而 AI 也许能说出有关“孤雁”的所有知识，甚至还能准确统计出“雁”在唐诗宋词中的出现频次，却无法给出令人感动的“同情之理解”，因为它缺乏应有的经验和文化背景，从而难以体会语言之外的深情。正因为这一点，AI 反而可以充当不知疲倦且无比高效的研究助手，却无需担心其取代人的主体性——至少当下基于硅基算法的 AI 还做不到，足以成为学者具身化理解的一种补充形式——语境化理解——凭借上下文作贝叶斯式推理所获得的理

---

[①] Matthew. “Towards A Rigorous Science of Interpretable Machine Learning”, *arXiv*:1702.08608, 2017, pp.2-3.

[②] 敦煌研究院：《数字敦煌资源库》，2023 版。

解。因此，真正值得警惕的，恰恰是古典学者可能会因为过度信任 AI 而逐渐失去其专业评估能力，乃至落入“垃圾进，垃圾出”的陷阱。而且，这种误识与过度自信的风险，正随着 AI 时代之文学越来越成为某种“‘推论’和‘表现’的混合体”[①]这一事实而不断提升。正如欧阳友权所言，AI 带来“新的生产力”的同时，也因缺乏生命体验而难以给出触动人心的阐释，如若低估人类的共情与直觉，人文学科的转型势将偏离其根本轨道。这也是古典学范式转型中可能遭遇的重要问题，即人文品质的流失以及由此产生的“空心学术”危机。

比如我们在使用 AI 保存和振兴非物质文化遗产的同时，很容易将 AI 的拼接、映射和重构，简单地视作其对古典世界的“还原”，甚至将这种“幻觉”带入到后续的研究之中。为了证明这一点，我们不妨参考 Schechtman 所做的一场人机交互与自我认同之关系的“实验”。他用 GPT-3 与自己少年时期的“AI 分身”进行对话，结果实验发现，这种既“奇异”又“有启发性”的交互，会混淆经由想象力区隔的不同自我概念。[②]而且，随着 GPT 等 AI 技术的演进，这种真实与重构的“区隔”将会变得越来越困难，尤其是当我们面对一个由 AI 重构的古典世界时，我们的想象与知识，真的能比 AI 更具有可靠性吗？我们会不会陷入一场对 AI 幻觉的幻觉之中呢？显然，前述顾虑已然涉及我们作为学者的责任与知识伦理。假如一篇 AI 翻译的古典文本发表后却被发现错漏百出，那么，该由谁承担其误导读者且损害学术信誉的责任呢？答案不言而喻。因为，在 AI 仍只是人类“智能”的延伸工具，或作为学者拓展“认知”领域的“学术助理”的前提下，我们并不能为自主性削弱而降低责任感的行为寻找任何借口[③]。因此，一种新的学术范式正在成为学界共识：即研究者必须公开 AI 的介入方式——例如注明“本分析由 X 算法处理 Y 数据而得”，以便提升可复现性；许多学术期刊也在陆续出台相关规定，强制要求研究者如同披露统计方法一般，详细说明 AI 使用的真实情况。

正如 David M. Berry 所指出，在“算法时代”与“后意识”语境下，机器生成的内容，会不可避免地模糊人与机器的主体性界限，进而冲击我们的真实性理解乃至真理观念，最终难免落入“自动化价值创造”的陷阱——人类工作的本质，只是在机器“创造”的基础上进行剩微调——其结果必然是创造力的衰竭[④]。尽管古典学的创造空间很小，但文本、事实的真伪鉴别却至关重要，AI 生成的以假乱真的内容一旦超过某个阈值，那么后续的 AI 便会以假为真，直到完全扭曲甚至颠覆原本清晰的事实。当学者也习惯于 AI 的生成与“总结”时，这种恶性循环就会形成系统性的“伪学”，甚至一段原本只需熟悉《庄子》文风便可轻易分辨的“伪《庄子》”文本，也足以欺骗 AI 的检测与学者的监督（这个我就曾遭遇过），而被当真货当成新发现的文献。因此，研究者是否具备更高的洞察力与敏感度，能否对文献的真伪作出最终裁决，仍将成为古典学新范式之关键。

但令人不得不担忧的是，当前的科技导向似乎正在形成一种以 AI 为中心的准宗教话语，在 17 世纪就已被科学逐出学术语言的神学，似乎又在以另一种神圣维度进入公众的生活世界，主导人们的观念结构。当这些“准宗教”的神圣话语随着大型语言模型的崛起而不断占据范式的中心时，生成式技术对人的解构，进而被赋予类语言甚至准神圣地位的整体趋势便难以逆转[⑤]。因此，如果连古典学的“典故性”也被生成式 AI 的答案洪流所淹没，那么我们历经千百年学术争鸣所确立的人文属性也将遭遇存续的危机。[⑥]

---

[①] 参[美]雷·韦勒克、奥·沃伦：《文学理论》，刘象愚、邢培明等译，北京：生活·读书·新知三联书店，1984 年，第 26 页。

[②] Schechtman, “The AI Self,” *AI & Society 37* (2022): 1039.

[③] Telakivi et al., “AI Extenders and Human Agency,” *Ethics and Information Technology* 23 (2021): 687.

[④] Berry, “The Computational Turn,” in *Critical Theory and the Digital*, 2014, pp. 129–130.

[⑤] Schuster & Ekbia, “Language in a Godless Age,” *Philosophy & Technology* 36 (2023): 5.

[⑥] Peter Stewart 则从认识论角度指出，AI 无法应对人类生活固有的非遍历（non-ergodic）现象：生命、语言与社会充满不可重复、开放式过程，其不可预知性与 AI 预测弱势成正比（Stewart, “AI and the

总之，只有在确保学术实践的原创性、价值性与人文本质的前提下，AI 技术才应被“增强而非削弱”[①]。同理，以 AI 驱动的新古典学范式，也必须以守护人类主体性为导向，不断增强学者的研究效能，如发现新的知识模式、实现跨领域的学术协作等，从而进一步促成优秀传统文化的保存与传播，最终达到提升人们的身份认同、文化品位与审美能力。如果可能，我们还必须进一步省思如下问题：当人工语料库或 AI 语料库与知网、web of science 等所有大型专业学术数据库实现即时互联，当每个研究者都可以拥有自己的专属大模型，当《古典学专项提示词集成》得以集解出版，当古典学研究同样可以被当成科学实验去做，当学术的数据化、图谱化、可视化、大众化、传媒化都在多模态 AI 工具的整合下得以“一键生成”，那我们古典学者又该做些什么？还能做些什么呢？我们的古典学还能叫做古典学吗？我们的古老文明又该沿着何种路径得以赓续呢？它还能继续发挥范导人类文明的作用吗？对于这一系列问题，我们或许只能选择一个答案，那就是：让接入 AI 的古典学，成为一种足以教成 AI 而不是仅仅成为 AI 之“知识库”的古典学。

## 三、“教成”AI：文明赓续的古典学路径

如前所述，即使抛开某些远大的未来愿景，AI 既可以成为一种存在性威胁，也可以成为某种延续人类文明的向量性工具，关键是人类当下的选择与筹划。而在此筹划中，古典学作为人类文明之基石，自然担负着“教成”AI 的重大使命。我们不仅要防止 AI 可能削弱人类的能动性、促成权力的过度集中乃至侵蚀文明之基石，更要尽可能地训导 AI，让其成为永远不会伤害人类的超级智能。为此，笔者提出以“典故”为研究节点的去中心化的古典学路径。需要重申的实，笔者所谓典故，指的是语言中所蕴含的历史文化典引和故事隐喻，它凝结了某民族或文化体的集体记忆和价值观念。故笔者所预设的“教成 AI 的古典学”，即以典故化知识/语词的引用与相关文本输出为核心，系统探究人工智能时代文明如何传承、延续并深刻影响当下与未来的古典学范式。

那么什么语词可以称之为典故呢？《辞源》的解释是：“典制和掌故”，或“诗文中引用的古代故事和有来历出处的词语”。[②]然而，典故作为一个多歧性概念[③]，却因受限于《辞海》以及源自《辞海》的语文学阐释，而忽视了典故同时还是“一个个具有哲理或美感内涵的故事的凝聚形态”和极具“艺术感染力的符号”[④]。维特根斯坦曾言：“即便狮子能说话，我们也无法理解它”[⑤]，因为理解语言取决于共享的生活形式。同理，典故的释义需要共享其文化背景，缺乏文明记忆的土壤，典故便只是晦涩难解的“只言片语”。作为文明赓续的基本单元， 典故在人类文明的发展长河中，始终扮演着文化 DNA 的角色，更是接续古今、勾连中外等不同文明的重要精神纽带。可以说，每一则典故都是特定历史情境下，各个民族或社群的经验凝结，并在反复传述和引用中，逐渐成为超越时空界限的文化符码。因此，当我们使用或言说一个语词时，我们不只知道这个语词的意涵，同时还能够依据已有经验和知识，下意识地“抓取”乃至“重构”被压缩在这个语词中的相关背景知识和大量相关信息的时候，

---

Non-Ergodic,” *Social Epistemology* 36 (2022): 562；并参 Landgrebe & Smith, “The Limits of Machine Prediction,” *Philosophical Studies* 178 (2021): 2885–2886）。因此，在开放世界与高度情境依赖的人文领域，人类判断仍不可或缺，技术“凯旋主义”须被警惕。

[①] Haber et al., “AI in Psychotherapy,” *Journal of Medical Ethics* 49 (2023): 325.

② 辞海编辑委员会：《辞海》，上海辞书出版社 1999 年版，第 831 页。

③ 据吴直雄先生统计分析，典故的定义多达 10 余种，其中最值得遵循的当属《辞源》“典故”条。见吴直雄：《界定典故多歧义<辞海>定义应遵循——论典故的定义》，《南昌大学学报（人社版）》2003 年第 3 期。

④ 葛兆光：《论典故——中国古典诗歌中一种特殊意象的分析》，《文学评论》1989 年第 5 期。

[⑤] 路德维希·维特根斯坦：《哲学研究》，第 2 部分第 11 节，商务印书馆，2016 年，第 215 页。

那么这个语词就是作为一个典故而被言说或使用。

正是借助一张张由“典故”链条交错而生的知识网络，前人的智慧、美德和教训，才得以源源不断地交互、拓展、传递，直至成为文明赓续的根脉。比如，汉文化圈的儒家学者，但凡领悟了“程门立雪”之类典故所内蕴的尊师重道等儒家价值观念，就可能将其融入自身的行为准则并内化为醇厚的精神气度，从而在言行举止中体现出温文尔雅、恭谨谦逊的儒者品质。故在此意义上，典故可以被视为一种集体记忆的节点，只需寥寥数字，便可唤起整个群体对共同历史的联想与情感共鸣。如果我们将文字视为某种信息媒介，那么典故便可作为一种知识集约的信息模型——以最少的信息当量传递尽可能丰富的思想情感，以高度压缩的形式承载文化的“基因序列”。因为，一个典故往往能“集中表达许多类型化的知识”，同时“引发无限的审美联想与情感体验”，这正是人类文明难以被生成式 AI 所完全取代的文化自信之所在。[①]比如，当我们说出“问鼎中原”这个短语时，其中包含的远不止地理概念层面的资源争夺，而是周、秦之变过程中的数百年权力变迁的历史故实，以及中心与边缘、野蛮与文明之间不断角逐等一系列复杂的文化心理与精神内涵，而且这些故事、观念、知识、思想与情感的复合，又可通过两千多年的中华历史得以不断丰富和塑造。因此，这四个字所具有的信息压缩与承载功能，是很难用有限的语言来描述的，而且其在话语交互中的意义涌现，更会因人而异，从而使这一典故得以成为古老之中华文明不断延续、革新的内在活力。正是无数个日用而不知的微小典故，共同编织出我们文化记忆的巨大网络，维系着一个古老文明的血脉相连般的身份认同和连续性。

然而，随着人工智能时代的到来，基于“典故”的交互形式，遭遇了贝叶斯推理式的话语生成机制的挑战。所谓贝叶斯推理，是一种基于贝叶斯定理的动态概率更新方法，即先设定先验信念，获得新证据后则按似然加权修正，再输出后验概率，从而使认知随着数据的变更而持续演进。显然，这与我们借助典故网络实现意义涌现和链接的理解形式，具有根本性的不同。因此，基于这种大语言模型的生成式 AI，是永远也不可能真正理解典故的，甚至还会产生幻觉性误解，这就必然导致智能机器生成的信息将逐步与人类文化记忆脱节。更糟糕的一种情况是，AI 可能创造出貌似古雅而实则虚假的“伪典故”，从而造成集体记忆的“幻觉”，甚至对当下和未来的整个话语环境造成不可逆转的“数据污染”，最终导致人类文明传承的彻底断裂或异化。

因此，我们要做的正是避免上述问题的同时，进一步促进人工智能为人类文明的传承与发展带来新的契机。一种可能的技术路径是引导 AI 深度融入人类的文化记忆体系，促使人与 AI 共同奏响文明发展的“深度复调”。这里所说的“深度复调”，借用了音乐理论中的复调（polyphony）概念，即多声部各自独立而又和谐共鸣的演奏形式。对应于文化领域，即多重意义、多种声音在同一开放、发展的文化空间中的共鸣与对话。其实，典故本身便具有复调性，因为，即便是同一典故，只要遭遇新的语境，便会与时代之音产生互渗，从而使其原始意涵获得丰富或重新诠释的同时，又会引导时代之音向着某种原初性的文明路径发展，或至少获得人文性的纠偏。文这便意味着，古老文明的声音与当代的声音相互作用，形成更为丰富的意义之和声，并为 AI 所“听见”和“记住”，深度影响甚至决定着它的“言说”。换言之，如果 AI 能够真正地理解典故并创造性地运用典故，便极有可能成为文明复调合唱中的新声部，并与人类一同完成文明的传承与创造。比如，AI 在回答科技伦理问题时，会因引用（听到）“潘多拉的盒子”而领会（记住）其所隐喻的不确定的风险，那么它的“声音”（言说）便会与古希腊神话与整个人类的向善之音产生共鸣，进而在辅助决策或程序的独立运行中，遵循具有复调属性的人文主义的文化框架，从而避免科技主义的失控。

但是，要达到这一目的，就必须让人工智能不再停留于简单复制人类“言说”的“生成”

[①] 详见拙著：《人工智能与人文学术范式革命》，北京：中国书籍出版社&光明日报出版社，2025 年，第 37-45 页。

阶段，而是具备典故化的深度理解与情境创生能力——即对任何一个语词的“言说”都能关联到所有文化情境与整个知识图谱之中。这实际是要求，AI 具有一定的“文化想象力”，能够参与人类文化意义的生产，同时还能确保其将人类的任何语词作为“典故”引用，而非只是一种针对人类言说的机械/概率式的“模仿”。

而当前最先进的人工智能大语言模型，主要依赖对海量文本语料的深度学习和统计学式输出，无法对人类语言的潜在语义作先验性把握。这便决定了 AI 对典故的学习或理解，本质上是一种模式模仿：模型通过训练数据中大量出现的语言模式，学习到某些短语与特定上下文的高频关联，而非真正理解这些短语所承载的文化内涵。例如，若训练语料中多次出现“破釜沉舟”且周围常伴随“决心”“胜利”等词，模型便会在输出时据此推断“破釜沉舟”表示决绝的行动。这种推断的本质不过是一种关联性统计，模型并不知道项羽背水一战等关联性的历史故事，更不会由此生出某些令人意外的“联想”。这种局限意味着：AI 目前对典故的掌握仍停留在“似懂非懂”，其本质仍是对字符串模式的拟合。正因为此，LLM 对于常见典故或成语，往往因其训练数据足以覆盖，故可正确使用和分析，但一旦遭遇冷僻典故或背景复杂的经典引文，它便无法准确理解，甚至会“胡言乱语”“张冠李戴”——即模型将词面相似却语义不同的典故相混淆。正因为此，当下几乎所有的 LLM 都很容易忽略典故应用时所需考虑的语境和语气，故其生成内容往往缺乏人类使用典故时的精妙和灵活。这类现象在 ChatGPT-plus、Grok-4 以及 deepseek-R1 等模型的实际表现中都屡见不鲜，它们常常在实时交互中生硬地塞入不合语境的名人名言或典故，以为这样便能增强说服力，实际上正好暴露出其对人文语境的陌生和无知。

显然，由于 AI 模型缺乏常识约束和对知识的真实理解，故其在标识、解释、引用或生成典故等方面的最大瓶颈，仍然是“拼贴复制”的“幻觉”（hallucination）问题，即在典故化场景下的张冠李戴或凭空杜撰，从而输出似是而非的内容。尽管这类问题已经受到 AI 圈的普遍重视，但幻觉率仍不同程度地存在于各类大模型。这种并不可靠甚至完全错误的传统文化生成文本，一旦借助 AI 数据库和网络的瞬时性传播，便不只是普及了错谬或并不存在的观念和知识，还将干扰公众的真实文化记忆，极大地消解历史的实在感。而且，AI 时代的“纠错”成本远远超过纸媒时代的“辟谣”和“订正”，甚至可能将人类引向无法溯源的认知深渊。因此，通过基于典故化知识库的精确语境和源头考据训练，让 AI 不仅能使用作为文化符号的语词，更可能真切地“理解”或“抵达”而非关联性地“触及”语词之文化要义，并可为通用大模型降低幻觉率提供某种可能路径。

但真正的问题也在于此，目前我们虽有设想，却无法突破技术瓶颈以获得切实可行的路径，而首先要解决的问题，就是如何提升 AI 对典故等人文知识的“理解”深度。当前较为普遍的改进思路是结合外部知识检索与知识图谱，赋予模型实时查证、推断的能力，这也正是当下某些 AI 的“深度思考/研究”模式下的工作原理。当模型需要生成涉及某一确切知识的回答时，必先令其调用一个典故化的知识库，或借助搜索引擎查询该知识的原始出处和使用语境，只有知识库与相关信息之间确立起一个完整的知识图谱后，才能据此生成答案。这实际上，是让 AI 将任意一个知识或语词当做一个有关于人类与全部知识的“典故”进行处理，同时借用符号推理来弥补纯神经网络节点不足所形成的盲点。[①]例如，当 AI 接到“如何评价越王勾践”的提问后，它在深度思考后，会先从互联网这个巨大的知识库中检索吴越争霸、卧薪尝胆等相关史典，然后在回答中据实叙述，而不是凭记忆临时编造，这样的生成模式，确实很大程度上降低了因幻觉所产生的讹误率。但不得不承认的是，如果这些知识库的知识本身就存在问题，或者相关知识过于冷僻，它就可能偏听偏信，从而成为某些“噪音”的传声筒。因此，知识图谱的引入便可使模型对语词或知识的理解，从纯文本的相关性选择

[①] 参[奥]迪特·芬塞尔（Dieter Fensel）：《知识图谱：方法、工具与案例》，郭涛译，北京：清华大学出版社，2023 年，第 82-91 页。

（它往往选择在统计学上占据量化优势的回答）提升至结构化语义（基于不同言说和知识链条进行逻辑推断）的水平，从而令知识/语词与出自专业知识库的历史、人物、事件、价值等知识进行逻辑关联，进而生成更可信的内容。

当然，这里还要考虑大模型在深度思考模式下的“望文生义”现象，这一点在 deepseek-R1 的“深度思考推理”上表现得格外明显，即便是当前月活量最大（突破 5 亿）也最“靠谱”的 ChatGPT plus，也难以避免类似的“幻觉”。比如，笔者将新加坡国立大学劳悦强教授有关“君子”观念及相关学术方法论范式的讲座笔记与 PPT 上传给 ChatGPT plus 版，并给出明确的提示词：“请根据上列两个文档，为我整理一份主题明确、内容完整、知识准确的讲座内容纪要。”在其“深度研究”的进一步提示下，我又补充了提示词：“生成文章式综述（适合发表或归档），同时包括：发言人引用；概念演变的时间轴梳理；对“君子”思想史演变的核心观点总结。”很快，它便生成了一篇 3000 多字的专题文章《“君子”概念的思想史演变讲座纪要综述》。[①]我对文章进行了逐字逐句的校阅和编辑，修改了多处问题，继而发布在我的《典故学与后人文时代》公众号。然而，就在文章被转载 10 次、阅读 95 次后，也始终没被指出任何问题的情况下，治学严谨的劳悦强教授，却只是“粗略看了一下”便发现了文章的诸多问题：

> 刚刚粗略看了一下，“小爱”同学的记录和整理不太准确，有好些话根本不是我说的。概念溯源部分尤其失实，如果能够删掉最好。上古晚期和两汉以后，我根本没有讨论过，其中有不少错误。比如说，“谦谦君子”出自《易经》，而非《诗经》；所引两章《论语》，我没有提过。“君”字从口，口不是指众人，而是指握杖之人发号司令之口。关于英文用词，我说的是 denotation（字义、词义）【我没有说 notation】、connotation（衍义）、associated meaning（延伸义）和 cluster of meanings（意义群）【我没有说 class meaning】。如果能够在公众号上修改，必须修改，避免误导你的粉丝。[②]

此次交互实验足以说明，仅仅提供知识库和具备联网检索等功能，还不足以确保模型输出的准确性，还有必要改进其语境感知能力，否则其制造的似是而非的典故化“知识”更具有欺骗性。因此，在当前技术条件下，可通过引入更长上下文窗口（比如将 PPT 替换成论文正文，目前 plus 仅支持 10 个上传文本），来保证模型在输出文本时能更充分地考虑前后文所需要的意义，而非统计学所决定的词语；同时还需利用人类反馈的强化学习（RLHF）机制对模型进行适配性调试，使其在不确定某个典故化的“语词/知识”的具体指向时，宁可只是解释说明也不贸然“引申/联想”，比如当同一文档中包含多篇不同作者文章时，最好明确其引用范围的具体页码，否则张冠李戴的可能性依然很大；最后，在算法判别输入中，可设计一个{是否引用包含特定“语词/知识”的内容并需要特别处理}的补充指令——这一点在 ChatGPT 的“深度研究”模式下已经完善，但国内的 LLA 还普遍缺失这一环节——从而让模型能敏锐识别出用户话语或自身生成中哪些是典故化的表达，并作准确的调用和特殊处理（如检索、附加说明），最大程度地降低因“幻觉”产生“偏差”的概率。

最后也最为关键的一环，是确保 LLA 能实现跨模态学习，并基于典故化知识进行合理的联想，如通过图像、音频等可视化模态辅助典故的学习。目前，这一路径仍在尝试阶段，其有效性仍待进一步验证。但是，要让人工智能将每一个语词作为“典故”进行掌握和理解，进而以一种“他者”的身份而非工具参与到人类文明的传承与创造之中，这对于人文学者而言，确实有些“疯狂”，因为它要求当下的 AI 从业者，必须对现有的大语言模型作出结构性的“改良”。但我们不妨再疯狂一些，再为 AI 界提出一种“典故识别—生成融合模型”（ARGFM，后文简称“融合模型”），并尝试将这一模型与“典故”的检测与生成有机结合，使 AI 在处理任何语言任务时，都像识别、应用典故一样严谨又充满联想和创造。

---

[①] 详见 ChatGPT 交互记录：讲座内容纪要整理。

[②] 此为我与劳悦强教授的微信聊天内容，经劳悦强教授授权，特征引于此。

笔者所设想的融合模型，将由典故识别模块和文本生成模块两个主要的子模块组成，二者通过共享的知识库和交互接口实现实时性的联结，从而形成一个对外开放（仅限于融合接口和权限层）的闭环系统（如数据管道、再训练触发条件、融合权重更新策略等形成“闭环”的整体迭代流程）。识别模块主要负责从输入或生成内容中检测典故化信息，而生成模块则负责产出连贯的自然语言文本。这一架构的独特之处在于，识别与生成并非割裂运行，而是在整个对话或写作过程中循环互动。模型的工作流程可简述为，当用户提出问题或要求生成文本时，识别模块先扫描输入语句中的所有语词，并将其代入所有可能的语境之中，再从中引用典故化成分（如成语、诗词、历史人名地名等）作为预生成草稿，进而再对照查询典故知识库以获取更为详尽和格式化的释义、语源以及相关联想，然后将这些结构化的信息一一反馈给生成模块。而生成模块则据此不断调整文本创作，包括以典故化的形式解释用户输入里的语词乃至文字，并在输出中以“用典”式的“言说”态度，对所涉及的全部语词进行反复核验和征引，直到确保正确使用的同时又能有所开创和丰富，且对任何不确定性都要做实时请求和澄清。而且，生成的文本还需经过识别模块的二次校验，检查其中引用的语词是否在任何联想情境中都能准确、恰当且符合交互者的预期，必要时还需再度调用整个人类知识库给出修正。如此往复，直到模型确信每一个语词的使用和处理都像诗人用典一样准确无误又丰富灵动，才能输出最终答案。

因此，要确保前述融合模型能够正常运行，还必须为之建构一个开放的知识库接口，使其知识生成能随时获得典故化的多模态知识库的丰富和完善，且始终处于迭代更新状态。该知识库接口首先应基于一个开放、扩展、互联的知识图谱——如维基百科、谷歌学术、知网与 web of science 等具有海量权威、专业知识，且处于持续开放、扩充、迭代状态的知识库所形成的关联网络——且其每一节点，都能涵盖各类典故化的知识及其变体，而其每一条边线则表明语词所关涉的全部历史事件、人物、情感寓意等错综复杂的意义关系。例如，一旦触及节点“卧薪尝胆”，即可关联“勾践”“越国”“吴越争霸”等典故化的语词和关联信息；而触发节点“Achilles’ heel”时，又可关联到希腊神话、阿喀琉斯、致命弱点等典故化的语词和关联信息。知识库的数据来源可包括维基百科等开放网络知识，也应包括知网、科睿唯安等权威专业学术知识，同时还应包括 Z-library、读秀等涵括各类电子图书的开放或非开放的知识库，而且所有知识还必须经由专家型 AI 与人类专家的共同校订以确保其意涵的准确性和使用的规范性。因此，当识别模块利用自然语言处理技术（如模式匹配和深度序列标注）标识出某一文本中的特定语词片段时，就可在前述知识库中进行检索匹配以确保其正确、合规。而对于未直接收入库中的新出语词或知识，则通过语义相似计算寻找相近概念或典故进行标识，并确保识别单元具有上下文判断能力。

如果可能，生成模块还应在预训练的 LLM（如 Transformer 架构）的基础上，融合一套典故化的信息增强机制。具体而言，就是在文本的生成过程中，明确要求生成模块在接收到识别模块所标记的典故化信息时，需立即将其视作附加约束或提示词内容。比如用户问：“为何人们会说‘覆水难收’？”此时，识别模块会在第一时间将“覆水难收”识别为典故，并给出“比喻事情无法挽回，语出《后汉书》”等初步释义，而生成模块则据此在回答中会进一步梳理和解释该成语的字面意思、来源故事（如姜太公、朱买臣等）以及引申义或隐喻义，而不会仅做字面翻译式的泛泛而谈。另外，在需要 AI 自主使用典故化的语词时，生成模块需援引知识库给出增强典故属性的表达。例如，在写作议论文时，模型可检索相关主题的典故以佐证观点，但这一过程必须受控于语境适配度评分规则，以防生硬堆砌产生明显的 AI 味。生成模块还要具备典故替代方案选项，尤其是面对要求避免晦涩用语的跨文化交流等特定场景时，模型可选择不用传统意义上的典故，而改用直白的典故化的用语，至少要用目标文化中的等价用语进行典故化的语境替换。如果用户还强调整个生成单元的连贯性和准确性，那么就要将这些提示词当做一种典故化的用语进行前述形式的解读和理解，从而

保证生成文本的通顺和灵动。同时，任何形式的引用（不论是直接语词还是间接引用观点）都要作典故化的标识与生成，从而自动、精确地给出来源。这方面，或许还可以通过加入典故化标签的训练形式来实现，比如在模型训练中明确标注典故化的语词类型，让模型学会处理这些特殊标签，从而给模型制定一种相当于“文化标记语言”的“言说”范式。

此外，上述融合模型的另一个值得一提的设想是识别模块与生成模块之间的交互迭代。如果用一个拟人化的描述，则可将识别模块视为 AI 大脑中的“文化顾问”，可时时监督语言输出是否遵从典故化原则；而生成模块便是通常意义上的“作者”，负责援引来自各个知识库的语词进行用典性的“创作”。而且“作者”每一次“言说”，都要经由“文化顾问”检查以确保每一个语词符合“引典”规范，且要提取所有关联史料给出初始提示词和补充提示词。“作者”严格依据“作者”的“建议”调整文本后再交由“顾问”审阅，如此循环往复至少 3 轮（如何权威期刊的三审三校），直到顾问再无异议后才能输出最后的文本。这一机制的本质不仅将所有语词视为对人类知识库的“引用”，而且每个“引语”都会在一个典故化的知识谱系中反复审验，直到其满足用典的所有规范后才能进入最后的输出阶段，从而在有效降低幻觉率的同时，提升“言说”的类人性，并让 AI 在内、外双重交互语境中不断规训，直到训练出观念、想象层面的“人性”。

综上可知，融合模型并非仅为准确复述已有用语，更在于提供一种语言生成观念的变革。如果这一设想被证明有效，则不仅能推动 AI 界的变革，也将实现人类知识库的反哺，使 AI 不断学习人类的“言说”方式，从而在实现自身演进的同时，还可通过捕捉、记录、传播人类的文化元素直至实现文化创造的类人存在。这意味着人工智能不再是人文知识的消化吸收工具——数据库加语言模型的简单组合，而是真正形成对人类知识的典故化理解与应用能力，从而让人机协作、文明赓续的智能对齐成为可能。它甚至还将引导人工智能拥有人类文化的灵魂，在价值观上与我们同频共振。这些意义共同指向一个目标：构筑人工智能与人类文明和谐共生的未来图景。正如有论者所言，真正重要的问题在于如何利用 ChatGPT 等 AI 技术，“进一步构建探索知识本源及不同问题域动态关系的人机协作未来范式”。典故学取向的融合模型正是这样一种未来范式的尝试。

# 结论

人类不可能记住一切，而只会选择对自我认同有用的内容。典故的形成也是如此：众多故事里只有少数被凝练为典故，其余被遗忘。AI 时代如何在海量信息中选择、压缩并传递知识，是摆在我们面前的挑战。一味追求“大模型”并无限制地喂入数据，不仅难以保障准确性，还会加重能源消耗，产生“数据焦虑”。典故化提供了一个思路：通过识别和抽取文化记忆中的典故节点，将人类知识结构化为可调用的符号网络，这种网络既包含传统文化，也包含现代科学典范。通过动态更新，网络中某些节点会淡出，新的节点会加入，形成适应时代的新典故库。这样做既可减少 AI 训练的冗余信息，又能在代际间保留文化核心。尤其是在全球化的语境下，典故化还意味着跨文化翻译。不同文明有各自的典故体系，如何让 AI 理解并尊重这些差异，是一项挑战。如果 AI 只能简单套用某一文化的典故，可能造成误解甚至文化冒犯。典故化模型必须具备文化敏感性：在跨文化交流中，根据对话对象选择恰当的典故或提供背景解释。这需要全球范围内的典故知识库，并借助文化比较学等学科的支持。只有在尊重他者文化的基础上，文明赓续才能真正达到交流与互鉴。因此，笔者所谓“教成 AI 的古典学”，其要义便在于通过典故学这一桥梁，把人类文明的深厚积淀教给人工智能，从而探寻文明赓续的新途径。对此，笔者提出了典故识别—生成融合模型这一构想，进而讨论了该模型在人机协作、文化传承和 AI 对齐方面的意义。可以想象，当人工智能能够理解

并创造性地运用典故时，它便不只是冰冷的计算机器，而开始融入人类的文化脉络，与我们共享历史记忆和价值共识。这样的 AI 对于人类而言，应该是可信赖的知识伙伴、敏锐的文化翻译者，也可能是推动文明继续前行的助力者。当然，要实现这一愿景仍有许多挑战。技术层面，它如何构建高质量的典故知识库、确保识别模块的高精度低误报、平衡生成文本的创造性与规范性，都是需要深入研究的问题。而文化层面的问题就更为复杂了，因为不同文明体系的典故如何映射，以及 AI 如何避免在多元文化环境中误用某些具有争议性的典故等问题，都需要被审慎对待。正因为此，笔者乃将“典故学”视为人文学术（而非只是古典学）应对 AI 挑战以赓续文明传承的一种未来范式。