

基于大语言模型的文史知识库构建研究^①

丘子靓 / 北京师范大学

胡韧奋 / 北京师范大学 (通讯作者)

莫凯洁 / 北京师范大学

王予沛 / 北京师范大学

刘智颖 / 北京师范大学

摘 要：高质量的文史知识库是开展数字人文研究的基石。近年来，大语言模型凭借其强大的语言理解与生成能力，为人文学科知识的深度加工带来了新的机遇。本文以典故知识分析为案例，通过典故知识结构化整理、用典判断和典故识别三项任务，评估了大语言模型在文史知识提取和加工方面的能力。实验结果显示，经合理的提示设计，大语言模型能够有效地完成文史知识整理，并表现出一定的文史知识判断能力；此外，小样本学习、检索增强及微调策略均能显著提升模型处理文史知识问题的能力。据此，文章讨论了大语言模型在文史知识加工领域的应用策略，包括模型选择、调用方式、提示与微调等引导机制，并提出了基于大语言模型构建文史知识库的基本路径。进一步地，文章展望了大语言模型在数字人文领域的未来应用方向。

关键词：大语言模型 知识库 典故 数字人文 文史知识

引 言

数据科学的研究范式正在推动“新文科”的发展，人文学科研究者尝试使用

^① 本文研究受到国家语言文字工作委员会重大项目“古籍整理智能化关键技术研究”(ZDA145-9)、北京市社会科学基金重点项目“古典文献的智能化分析与关联技术研究”(21DTR037)资助。

数据量化人文知识,利用技术解决人文问题。近年来,大数据挖掘^①、网络分析^②和自然语言处理^③等技术被引入文史研究,形成了极具交叉学科特色的数字人文方向。以古典文学领域为例,刘石(2020)提出,随着数字人文技术的发展,数据分析技术和方法越来越具有针对性和有效性,其能更清晰地揭示隐藏在文学史背后的作家与社会之间、作家与作家之间、文本与文本之间的直接与间接、显性与隐性的多种关联,能以全知型的视角系统整体地还原和呈现文学史的立体景观,改变传统的思维方式和研究范式。^④然而,现有技术的应用仍然存在可拓展的空间:目前的数据挖掘更多关注文本浅层特征,如字、词级的形式特征,较少涉及深层语义信息或抽象概念;有监督模型的训练往往依赖大量人工标注数据,而人文领域的数据往往对标注员有高于一般标注任务的知识要求,进一步提高了标注的难度与成本。

近年来,以ChatGPT为代表的大语言模型发展迅速,表现出了极强的语言理解和生成能力,为人文知识的挖掘和加工提供了新的契机。在无需标注数据额外训练的前提下,大语言模型不仅能生成合乎语法及逻辑的内容,还能够完成角色模拟、故事和诗歌写作等创意任务。因而,相关技术正在广泛地改变各领域的学术研究方式。《自然》(*Nature*)杂志针对全球1,659名科研工作者的调查显示,当被问及十年内AI工具在自己的研究领域内有多大用处时,超半数受访者认为“非常重要”或“必不可少”。^⑤在人文社科领域,越来越多的研究表明,大语言模型在数据生产与标注方面可以扮演重要角色。在数据生产层面,大语言模型可以基于其所学习的数据进行社会行为模拟,具有替代人类被试的潜力,例如通过模拟国际冲突中的参与国决策及其后果,^⑥推演社会固有集体现象的发生。^⑦在

① 诸雨辰、胡轸奋:《清人“勘救”说再审视——以〈全唐诗〉15,290首律诗为样本》,《中国诗歌研究》2019年第1期;应申等:《基于唐宋文人足迹集聚性分析的中心文化城市变迁》,《地球信息科学学报》2020年第5期。

② 严承希、王军:《数字人文视角:基于符号分析法的宋代政治网络可视化研究》,《中国图书馆学报》2018年第5期;陈松:《为学作记——从网络分析和文本分析视角看宋代地方官学碑记的作者与主题》,《数字人文》2020年第4期;赵薇:《社会网络分析与“〈大波〉三部曲”的人物功能》,《山东社会科学》2018年第9期。

③ 宋丽、刘颖、马艳军:《用计量风格学方法考察〈水浒传〉的作者争议问题——以罗贯中〈平妖传〉为参照》,《中文信息学报》2022年第8期;诸雨辰、李绅:《〈汉书·艺文志〉目录分类再审视》,《数字人文》2021年第3期。

④ 刘石:《大数据技术与古代文学经典文本分析研究》,《数字人文》2020年第1期。

⑤ Richard Van Noorden, and Jeffrey M. Perkel, "AI and Science: What 1,600 Researchers Think," *Nature*, vol. 621, no. 7980, 2023, pp. 672-675.

⑥ W. Hua et al., "War and Peace (Waragent): Large Language Model-Based Multi-Agent Simulation of World Wars," arXiv:2311.17227, 2023.

⑦ G. De Marzo, L. Pietronero, D. Garcia, "Emergence of Scale-Free Networks in Social Interactions Among Large Language Models," <https://arxiv.org/abs/2312.06619>, 2023.

数据标注层面，模型可以辅助完成数据的整理、归类及分析。Ziems等针对语言学、文学、心理学、历史学、政治学、社会学等人文社会学科关注的问题设计了二十余个评测任务，以评估大语言模型在相关领域文本数据分析上的表现。^①其实实验发现，在标签分类任务上，大语言模型与人类标注一致性尚可，在开放的生成式任务上，模型表现往往能超过众包平台的标注员。因此，虽然模型还无法完全替代人类来加工文本，但可以作为机器标注员加入人类团队，以降低标注成本，提升标注效率和多样性。

由此可见，大语言模型在数字人文领域有很强的应用潜力。然而，在模型落地应用的过程中，仍然面临一系列问题有待讨论：第一，在文史知识加工中，大语言模型擅长哪些工作？又存在哪些不足？第二，如何选择合适的模型及其调用方式，以实现高效的数据处理？第三，模型虽有很强的通用能力，但并非领域专家，如何通过提示设计、微调等策略引导模型，使其更好地辅助专家完成专业性任务？

为回答上述问题，本文拟以典故知识分析为案例，系统评测大语言模型在典故知识结构化整理、用典判断和典故识别三项任务中的表现。典故以高度凝练的形式承载了丰富的中华文化知识。以此为切入点，不仅能有效挑战大模型的知识储备和理解能力，还能为文史知识库的构建提供思路方法的支持，进而促进数据驱动的人文研究的发展。具体来说，在上述三个任务上，实验将对比GPT-4o、GPT-4o-mini、Qwen-Max（通义千问）、Qwen-Turbo这四个国内外评测表现最优且效率出色的模型，并引入提示设计、检索增强、微调等多种模型应用策略，以期得出大语言模型在数字人文领域的实践应用建议，并提出利用大语言模型构建文史知识库的方法路径。

一、典故知识分析任务设计

刘石提出，目前古典文学的大数据分析还处于检索的基础阶段，要实现技术对文本的深入分析，如史论验证、作品归属和风格变化等问题，则需要进一步拓展当前的文学资源建设与知识挖掘技术，“以上研究设想的实现，建立在两个基础之上。其一，古代文学经典文本数据的结构化……其二，利用大数据技术构建多样化文本分析系统”^②。受其思路启发，本文着眼于典故知识库的构建需求，设

①C. Ziems, W. Held, O. Shaikh, "Can Large Language Models Transform Computational Social Science?" *Computational Linguistics*, vol. 50, no. 1, 2024, pp. 237-291.

②刘石：《大数据技术与古代文学经典文本分析研究》，《数字人文》2020年第1期。

计了三项面向大语言模型的典故知识分析任务，分别为典故知识结构化整理、用典判断和典故识别。接下来，本节将介绍三种任务的设计思路和具体形式，并对大语言模型的选择和使用方法进行说明。

（一）任务形式

1. 典故知识结构化整理

任务一“典故知识结构化整理”旨在应用模型实现专家知识的结构化整理。在文史领域，为服务于文献阅读和研究，专家们编撰了大量的工具书。然而，这些工具书的词条往往采用自然语言的文本形式进行描述。如果能够对其进行结构化解析，进而构建一个系统的知识库，这将为大规模的知识查询、统计分析等研究提供重要支持。例如，表1示例出自《中国典故大辞典》^①（以下简称《辞典》）。该条目描述了典故“五马”的来源、释义及例句。表1右侧展示了该条目在知识库的存储格式，其中，典故“五马”的各类信息已整理归入对应的“键”，使计算机能够准确且高效地检索条目及相关信息。

表1 典故知识结构化整理示例

《辞典》原条目	知识库整理格式
五马《玉台新咏·日出东隅行》：“使君从南来，五马立踟蹰。”古乘驷马车，至汉时太守出则增一马，为五马。因以“五马”借指太守的车驾。唐钱起《送张中丞赴桂州》诗：“云衢降五马，林木引双旌。”亦为太守的代称。北周庾信《奉和永丰殿下言志》之四：“五马遥相问，双童来夹车。”唐白居易《西湖留别》诗：“翠黛不须留五马，皇恩只许住三年。”	<pre> { "典故名称": "五马", "典故来源": "《玉台新咏·日出东隅行》：“使君从南来，五马立踟蹰。”", "释义列表": [{ "释义": "古乘驷马车，至汉时太守出则增一马，为五马。因以“五马”借指太守的车驾。", "例句": ["唐钱起《送张中丞赴桂州》诗：“云衢降五马，林木引双旌。”"] }, { "释义": "亦为太守的代称。", "例句": ["北周庾信《奉和永丰殿下言志》之四：“五马遥相问，双童来夹车。”", "唐白居易《西湖留别》诗：“翠黛不须留五马，皇恩只许住三年。”"] }] }</pre>

① 赵应铎主编：《中国典故大辞典》，上海：上海辞书出版社，2012年。

但是,文史领域工具书的格式对结构化自动整理提出了挑战:首先,采用自然语言描述的工具书往往缺乏显式的切分标记,机器难以像人类一样辨识和理解每个知识实体的边界;其次,由不同专家合作编撰的工具书,语言表述格式可能具有不一致性,这会导致解析规则之间的冲突;再次,即使在同一领域内,不同的大型工具书也可能采用不同的编撰格式,这使得整理来自不同来源的内容时,需要耗费大量的时间和精力。

此前,结构化整理主要依赖于人工编写正则表达式,这一过程包含识别字符串中存在的模式并使用正则符号提取对应的文本内容。在使用正则表达式解析《辞典》时,边界误判问题尤为突出。以来源和释义的边界判断为例,如表1所示,《辞典》的标准格式通常是先引用诗文以说明来源,再解释典故的释义。为了区分来源与释义的边界,既可以识别来源的结束标记,也可以识别释义的开始标记。然而,如果以诗文引用结束的双引号作为来源的结束标记,则无法处理《辞典》中包含的还对诗文引用再做注解和补充笺注的条目。如果以释义中的术语,如“指”“谓”“因以”等作为释义的标记,则无法处理如表1中的例子,即先对典故做整体说明,再详细解释释义的情形。此外,《辞典》条目中还存在多个出处或释义、缺少释义或例句、直接描述典故来源等多种格式,每种情况都要编写对应的正则表达式。

大语言模型可以基于文本进行语义理解,同时长于形式语言,有望应对上述问题。为评测大模型进行结构化知识整理的能力,本文以《辞典》数据为对象,人工挑选了100条数据作为测试集A,这些数据均衡地涵盖了《辞典》中条目的所有编撰格式,具有多样的语言表述,能够较好地代表工具书数据的整体情况。

2. 用典判断

任务二“用典判断”属于文本分析层面任务,旨在要求模型判断给定句子是否用典,并输出理由。与任务一侧重整理已有工具书中的知识不同,任务二需要从未标注的语料中挖掘用典信息,属于知识加工型任务。需要指出的是,古代汉语中的语词表达具有高度歧义性,如例1、例2中都出现“桃李树”,但只有例2使用了“桃李不言,下自成蹊”的典故。此外,作者用典时并非简单使用某个固定词语或词组,而会对其表述进行变换,如例3、例4所示“宋玉悲秋”典故的不同典形。考虑到用典表达的歧义性和多样性,该任务能够有效评估大语言模型对深层语义信息的理解能力。

例1. 应似园中桃李树,花落随风子在枝。(未用典)

例2. 自是桃李树，何畏不成蹊。(用典)

例3. 当时宋玉悲感，向此临水与登山。(用典)

例4. 楚客忆江蓠，算宋玉未必为秋悲。(用典)

围绕该任务，我们通过人工标注的方式采集了100条数据，其中用典和未用典比例为1 : 1，称为测试集B。

3. 典故识别

任务三“典故识别”同样属于文本分析层面任务，除了要求模型判断是否用典外，还需给出具体使用何典，这对模型的文史知识面进行了深度考察。该任务同属于知识加工型任务，是计算机辅助构建文史知识库的重要环节。值得一提的是，古人表达高度凝练，可能在同一个句子中使用多个典故，如例5使用了“抱石疑玉”和“探骊得珠”这两个典故。因此，典故识别是三个任务中最具挑战性的任务。

例5. 抱石耻献玉，沉泉笑探珠。

同样地，我们通过人工标注采集了100条数据作为测试集C，其中用典数据为90条，包含40条使用单典数据和50条使用多典数据，另有10条未用典条目。

(二) 大语言模型选择

如表1所示，大语言模型通常有三种访问方式：通过网页或APP访问、使用API调用和下载开源模型使用，每种方式下可用的模型列表不同。考虑到本研究有数据批量处理和参数设置的需求，我们选择以API方式调用模型，并将Temperature参数设置为0，^①以确保实验结果的一致性和稳定性。

表2 大语言模型的三种访问方式

类别	示例	特点
网页或APP	ChatGPT、文心一言等	由用户通过网页或APP访问，研发团队在提示改写、联网检索、多轮会话等方面做了较多优化，以提升用户体验，但无法调整参数，不支持数据批量处理。

^①Temperature和Top_p参数影响模型输出结果的多样性和随机性，网页或APP端产品往往会设置较高的参数值，因此多次提问同一个问题会收到不同结果。

续表 2

类别	示例	特点
闭源模型 + 商用 API	GPT 系列、Qwen 系列、ERNIE-Bot 等	由用户参考 API 文档以网络请求方式调用模型，按 token 数计费，支持参数设置和批量调用，对设备无要求，仅涉及少量代码。
开源模型	Llama 系列、Qwen 系列、ChatGLM 系列等	由研发人员下载后在专用服务器上进行微调或二次开发，灵活性高，但对设备和代码有较高要求。

在选择具体的模型时，模型性能效果和使用成本是两项重要因素。首先，我们参考了多个大模型综合能力评测基准。其中，LMSYS Chatbot Arena^①国际评测采用大模型匿名对战的方式计算模型胜率，SuperCLUE^②和OpenCompass^③等专门针对模型中文能力开展了多种类型的测试。综合多个排行榜结果，选定目前表现最优的GPT-4o和Qwen-Max模型，此外，另选择了GPT-4o-mini和Qwen-Turbo这两个响应速度快、成本低的经济型模型作为参照。^④

（三）大语言模型提示与微调

在利用大语言模型完成特定任务时，可以通过提示设计和微调模型等方法对其进行引导。提示设计旨在为模型提供必要的背景知识和清晰的任务导引，提示词的质量对模型表现有重要影响。通常，可采用思维链、小样本学习、检索增强生成、系统指令等策略来丰富提示信息，以指导模型更好地回答问题。其中，思维链（Chain-of-Thought，简称CoT）指要求模型分步骤解决问题，较为适合复杂任务场景；小样本学习（Few-Shot Learning，简称FSL）指通过提供少量示例，令模型更好地理解任务目标，从而仿照示例回答问题；检索增强生成（Retrieval-Augmented Generation，简称RAG）则是利用外部专业数据库查询与问题有关的信息，将查询结果加入提示词，为模型作答提供参考；在调用API或本地部署模型时，系统指令（System Message）出现在首轮对话中，用于指定角色、领域、环境 and 能力等背景信息，能够增强对大模型的引导及限制。除了上述提示设计策略外，用户还可针对任务形式专门构造一批问答数据集，用于模型微调（Fine-tuning），使得模型较为系统地学习某类知识。在典故知识分析的三个任务中，本研究综合运用了多种提示策略和微调方法。附录展示了各任务所使用的提示词。

任务一“典故知识结构化整理”旨在生成符合规范的解析字符串。为帮助模

①LMSYS Chatbot Arena 网址：<https://chatbot-arena.web.app/>。

②SuperCLUE 网址：<https://www.superclueai.com>。

③OpenCompass 网址：<https://opencompass.org.cn/home>。

④ 本文所使用的具体模型版本包括：GPT-4o-2024-08-06、GPT-4o-mini-2024-07-18（还用于微调任务）、Qwen-Max-0428、Qwen-Turbo-0624。

型理解《辞典》条目的组织形式及各实体的核心内容，我们首先阐述了这些条目的整体结构。对于超出普遍格式的条目，如包含多个来源或多个释义的情况，我们进一步明确了模型应遵循的整理规则，并规定了统一的输出格式。最后，我们通过一个实例示范了理想的输出格式。

在任务二“用典判断”中，我们概述了“用典”的特征，要求模型针对“是否用典”(Label)进行判断，并要求模型输出“用典位置”(Span)和“用典依据”(Evidence)信息，以便检查模型判断的合理性。此外，本研究还对微调模型的效果进行了对比分析。我们从人工标注数据集中按用典和未用典1:1的比例，随机采样了200条数据，构建了问答格式的微调数据集，并使用GPT-4o-mini模型进行了微调。为保持微调和输出的格式的一致性，微调模型仅生成“是否用典”(Label)和“用典位置”(Span)结果。

任务三“典故识别”引入了基于外部知识库的检索增强生成方法，并对比了不同背景信息对检索增强生成的效果。具体来说，对于每一条待判断的语料，我们使用了古汉语预训练语言模型BERT模型^①为语料生成句向量，并在外部典故库中进行句子检索匹配和排序，得到与该语料相似度最高的十条候选典故数据，^②每条数据包含典故名称、释义与例句信息。我们对这些信息的不同组合方式及其在检索增强生成中的效果进行了横向对比。附录中展示了典故名称作为检索增强信息的提示词。

二、实验结果与分析

在实验阶段，我们采用前述方法调用模型，在测试集A、B、C上进行了实验。针对“典故知识结构化整理”任务，我们逐一校验了模型的生成结果，统计了模型整理正确率及各类错误的数量；在用典判断和典故识别任务中，我们以人工标注的结果为基准，采取自动评估的方式，分别计算了这两个分类任务的精确率、召回率、F1值等评估指标。其中，典故识别的多分类结果通过微平均法计算，“未使用典故”也被视为一个独立的分类标签。下文将依次介绍每项任务的评测结果。

①胡韧奋、李绅、诸雨辰：《基于深层语言模型的古汉语知识表示及自动断句研究》，《中文信息学报》2021年第4期。

②外部典故库参考见莫凯洁、丘子靓、胡韧奋等：《古汉语典故资源库的构建及应用研究》，《中文信息学报》2024年第11期。该外部典故知识库为《中国典故大辞典》的知识结构化整理结果，每个典故有若干条例句，其向量由下属所有例句的句向量取平均获得。例句和待判断语料的句向量均由所有Token取平均获得。

（一）典故知识结构化整理

表3展示了各模型在执行任务一时的表现。首先，GPT-4o和Qwen-Max的解析准确率均达到90%及以上，显示出大模型在进行文本结构整理和知识库构建方面的能力。在某些实例中，大模型甚至表现出超越人类判断的潜力。以例6中的“一字褒贬”条目为例，模型结果与人工整理有显著差异。《辞典》中先说明了该典故的本义，随后指出其引申义。人工整理时，倾向于将本义与引申义合并，而大模型则严格遵循指令，将多个释义分别存储为独立的键值。尽管前者更符合人类的阅读习惯，后者则更契合知识库整理的规范。此外，大模型还依据语义，将“一字褒贬”的两条例句正确归类至对应的释义下，这是基于字符串表层形式的正则表达切分所难以实现的功能。

表3 典故知识结构化整理结果

类型 \ 模型		GPT-4o	GPT-4o-mini	Qwen-Max	Qwen-Turbo
切分正确率		96%	71%	90%	70%
键值缺失	典源缺失	1	2		4
	释义缺失		3	2	5
	注释缺失	1	1	1	6
	例句缺失		1		3
键识别错误	例句识别为释义		10	3	5
	例句识别为典源	2	6	3	4
	释义识别为典源		4	1	1
	注释识别为释义		3		2
	典源识别为释义				2
格式错误	生成超出格式的键				2
	*JSON 格式解析错误 ^①			73	30

例6.

《辞典》条目：

一字褒贬 晋杜预《春秋经传集解序》：“《春秋》虽以一字为褒贬，然皆须数字以成言。”孔颖达疏：“《庄二十五年》：‘陈侯使女叔来聘。’传曰：‘嘉

①JSON 格式解析错误是指大模型返回的结果未严格遵循 JSON 格式，或存在部分格式不完整的情况，例如缺少引号、花括号等标识符。由于某些模型在输出中频繁出现此类错误，因此在人工校对时，仅关注了模型返回的键值是否正确，而没有考虑 JSON 格式解析的影响。

之，故不名。’《僖二十五年》：‘卫侯毁灭邢。’传曰：‘同姓也，故名。’褒则书字，贬则称名，褒贬在于一字。”原谓《春秋》笔法严谨，一字即寓褒贬之意。后亦泛指论人议事用词严谨而有分寸。南朝梁刘勰《文心雕龙·征圣》：“故《春秋》以一字为褒贬。”唐柳宗元《上扬州李吉甫相公献所著文启》：“又不得献薄技于前，以希一字之褒贬。”

大模型输出结果（Qwen-Turbo）：

```
{ "典故名称": "一字褒贬", "典故来源": "晋杜预《春秋经传集解序》：
“《春秋》虽以一字为褒贬，然皆须数字以成言。”孔颖达疏：“《庄二十五年》：‘陈侯使女叔来聘。’传曰：‘嘉之，故不名。’《僖二十五年》：‘卫侯毁灭邢。’传曰：‘同姓也，故名。’褒则书字，贬则称名，褒贬在于一字。”",
"释义列表": [ { "释义": "原谓《春秋》笔法严谨，一字即寓褒贬之意。",
"例句": [ "南朝梁刘勰《文心雕龙·征圣》：“故《春秋》以一字为褒贬。”" ] },
{ "释义": "后亦泛指论人议事用词严谨而有分寸。", "例句": [ "唐柳宗元《上扬州李吉甫相公献所著文启》：“又不得献薄技于前，以希一字之褒贬。”" ] } ] }
```

进一步分析发现，大模型执行结构化整理任务时，主要出现了三类错误：键值缺失、键识别错误和格式错误。“键值缺失”指某个键应有的值未能完整显示，且该信息未出现在其他键；如果该信息出现在了其他键下，则归为“键识别错误”。其中，大模型典型的错误表现为“将例句误识别为释义”以及“将例句/释义误识别为典源”。提示词规则要求将例句与对应的释义一起保存，且例句键的值可以为空。如果模型未能找到例句的对应释义，可能会将例句误作为释义或保留例句键为空值。为避免此类错误，可在提示词中进一步阐明规则。典源信息的识别错误主要发生在以“同”格式标识典故出处的条目中，如附录任务一的提示词中规则1和5所示。这类条目通常采取“同某典故”的方式，省略了典故来源和释义的详细说明。如果模型未能正确理解该格式，则可能将其他键的信息误识别为典故来源。针对这些复杂且对理解要求较高的任务，两个更高级的大模型（GPT-4o和Qwen-Max）表现出更好的适应性。因此，为提高准确率，可考虑使用这些更高级的模型，或通过提供更多示例来加强规则说明。

（二）用典判断

表4呈现了各模型在用典判断任务上的结果，其中精确率（Precision）为“模型正确识别用典数/模型识别用典数”，召回率（Recall）为“模型正确识别用典数/数据集中的真实用典数”，F1值通过公式 $F1=2 \times P \times R / (P+R)$ 计算，准确率（Accuracy）为“模型正确识别数/数据总条数”。结果显示，在所有未经微调的模型中，GPT-4o综合表现最优，Qwen-Turbo在各指标上较其他模型均有一定差距。GPT-4o和Qwen-Max在召回率上均达100%，即没有遗漏地识别出了所有用典条目，但其精确率较低，存在较多将未用典错误识别为用典的情况。经过微调后，GPT-4o-mini的精确率显著提升，召回率也达到90%，综合表现最佳，这充分证明了微调方法在该任务上的有效性。

表4 用典判断实验结果

模型	统计类型	精确率 %	召回率 %	F1 值 %	准确率 %
GPT-4o		76.92	100.00	86.96	85.00
GPT-4o-mini		69.57	96.00	80.67	77.00
GPT-4o-mini（微调）		90.00	90.00	90.00	90.00
Qwen-Max		60.24	100.00	75.19	67.00
Qwen-Turbo		71.79	56.00	62.92	67.00

我们筛选出GPT-4o和Qwen-Max正确识别用典的条目，人工核验了两个模型划分出的用典位置和用典依据。在50个用典条目中，GPT-4o出现了3处用典位置错误和35处用典依据错误，而Qwen-Max有1处用典位置错误和33处用典依据错误。用典依据错误集中在“错误解释典故含义”和“错误解释典故出处”两处。这指向了大模型的幻觉问题，即在训练数据相对缺乏的领域，大模型会生成与事实不一致的回答。在例7中，大模型正确地识别出该句使用了典故的两处表达，并准确解释了“五斗禄”的含义和来源。然而，“七不堪”实际典出嵇康《与山巨源绝交书》，而非《庄子》。这表明，通用大语言模型在处理中国传统文化知识，尤其是典故方面，仍有所欠缺。不过，大模型已经掌握了一些典故的知识，如例8中，大模型正确地输出了典故“卖剑买牛”的相关信息。

例7. 输入语料：欲徇五斗禄，其如七不堪。

大模型输出结果（GPT-4o）："五斗禄"出自陶渊明的典故，陶渊明辞官时说“不为五斗米折腰”，指不愿为微薄的俸禄屈身。“七不堪”出自《庄子·人间世》，指七种不能忍受的情况。这句话的意思是想要追求微薄的俸禄，但又难

以忍受多种困境，引用了古籍中的故事和言论，传达了丰富的意蕴，因此判断使用了典故。", "allusion_span": ["五斗禄", "七不堪"]（正确划分用典位置，错误解释典故出处。）

例8. 输入语料：薄有田园能自适，不妨卖剑买黄牛。

大模型输出结果（Qwen-Max）："“卖剑买牛”是一个典故，源自《汉书·龚遂传》，讲述了龚遂劝导农民放弃争斗回归农耕生活的故事，后用以比喻放弃武力、从事生产或改业归农。此诗句中“卖剑买黄牛”即化用了这一典故，表达了一种由征战或武事转而追求田园自适生活的意向。", "allusion_span": ["卖剑买黄牛"]（正确解释典故出处。）

（三）典故识别

任务一和任务二实验结果显示，GPT-4o模型在处理典故相关任务时表现更为优异。因此，在任务三中，我们选用GPT-4o模型来测试不同的RAG（即检索增强生成）策略，以进一步评估大模型在复杂任务中的表现。我们将未使用RAG的提示词作为基线，与各类RAG策略进行了对比，相关结果见表5。数据显示，所有RAG策略的效果均显著优于无检索增强的基线，这说明了在知识密集型任务中引入外部知识库的必要性。在引入RAG策略后，各组合识别的精确率都有显著提升，但召回率均略有下降。提供检索信息时，并非模型呈现越多信息越好，典故释义信息对大模型的理解帮助较为明显。

表5 典故识别实验结果

语料类型		统计类型	精确率 %	召回率 %	F1%
无检索增强			23.71	16.08	19.17
有检索增强	零样本学习 (无示例)	名称	79.41	55.86	65.59
		名称 + 释义	81.73	58.62	68.27
		名称 + 例句	78.10	57.34	66.13
		名称 + 释义 + 例句	82.65	55.86	66.67
	小样本学习 (一则示例)	名称	84.62	53.85	65.81
		名称 + 释义	89.25	57.24	69.75
		名称 + 例句	86.17	56.25	68.07
		名称 + 释义 + 例句	89.29	52.08	65.79

此外，RAG策略中的候选典故列表是通过句子向量相似度匹配得到，可能未包含或仅部分包含正确答案。对此，在构建数据集C时，我们特别标注了这些列

表是否包含正确答案。表6展示了不同类型数据上RAG策略的表现，对于用典情形来说，检索结果对模型识别能力影响巨大，当用典信息在候选列表中时，模型能够较好地参考检索结果作答，平均正确识别典故数高；而当用典信息未被检索到时，模型几乎难以识别具体所用何典。对于未用典的情形，丰富的典故候选信息对模型的干扰较小，可见模型能够从语义层面分析当前语料是否与某一候选典故相关。

表 6 检索增强类型及统计结果

语料类型 \ 统计类型		语料数	典故数	平均正确识别典故数	平均漏识别典故数	平均错识别典故数
答案在检索增强列表	多典全部答案	20	40	34.75	5.25	1.63
	多典部分答案	20	41	18.63	22.38	4.88
	单典含答案	30	30	27.00	3.00	0.63
答案完全不在检索增强列表	多典无答案	10	22	0.13	21.88	6.88
	单典无答案	10	10	0.13	9.88	1.75
未用典		10	0			1.25

三、大语言模型应用于文史知识整理的启示

由上述实验结果可见，大语言模型在执行结构化整理任务中的指令时表现出较高的准确率，而经过微调和检索增强后，大模型在典故判断和典故识别两个任务上的效果也取得了显著提升。基于这些实验，我们可以得到以下三点主要启示。

第一，大语言模型能够理解和处理文史知识，尤其在解析和整理已有知识方面表现突出。然而，在自主加工领域特定知识时，大模型仍存在一定的局限性。在结构化整理任务中，相较正则表达式等传统方法，大语言模型展示出明显优势。大模型不仅能够通过一套统一的规则处理工具书中的多种条目格式，还能凭借其语言理解能力，识别并处理缺乏显式标记的自然语言条目，从而应对超出格式规范的情况。但是，当涉及对典故的解释性任务，或在检索增强信息缺乏相关知识时，模型的表现仍然不够理想。面对这些对领域知识要求较高的任务时，人机协同至关重要。以“用典判断”任务为例，大语言模型在判断准确度上已经达到较高水平，能够在大规模数据处理中承担基础标注工作，专业人员则可以专注于深度加工，完成解释性等大模型尚不擅长的任务。同时，人工总结和校验的结果也可以用于模型调优，进而提升模型处理复杂任务的能力。

第二，选择合适的模型及调用方式对于高效的数据处理来说至关重要。在模型的选择过程中，需要综合考虑任务执行难度、模型处理效果和模型调用成本等多重因素。例如，在调用价格方面，GPT-4o与Qwen-Max相近，但均明显高于两个更小的模型。而在结构化整理任务中，GPT-4o的表现较其他模型显著更优，尤其是在处理大规模数据时，选择GPT-4o模型将大幅节约人工成本。此外，在确定了模型以后，参数设置同样对任务效果造成显著影响。因此，应当参考文档，设置合理的参数值。例如，为了追求模型反馈的一致性和稳定性，可以将Temperature或Top_p参数设为接近0的数值。而在执行创意度要求更高的任务时，适当提高Temperature参数，有助于模型生成丰富多样的回答。

第三，大语言模型在数字人文领域的应用潜力需要使用者的悉心引导。在提示设计中，可以通过以下几种策略有效激发大模型的相关领域知识：（1）在系统指令中为模型设定角色，帮助模型熟悉任务背景；（2）提供清晰具体的任务说明；（3）给出问题和答案示例；（4）对于复杂任务，可为模型设定分步骤解决思路；（5）可通过外部知识库或工具检索必要的专业知识作为参考，检索结果的相关性对模型分析能力有极大影响，因此RAG策略需选择精准、高效的检索匹配算法；（6）设置限制和要求，或要求模型输出后进行自查和纠错。此外，如果对模型输出有更强的控制需求，还可以通过构造数据集微调模型等方式，对模型进行定向、系统地引导。

四、总结与展望

本文立足于文史知识库构建的实际需求，以典故知识分析为案例，设计了三项面向大语言模型的评测任务（典故知识结构化整理、用典判断和典故识别），并系统评测了多种模型在该系列任务上的表现。实验结果表明，大语言模型能够较为高效地完成文本结构化处理类的工作，并表现出一定的文史知识判断能力，而检索增强生成、微调模型等策略均能有效提升大模型的文史知识。进一步地，本文在大语言模型能力、调用方式、引导策略等方面获得了有益启示。

本研究发现，经恰当的提示设计或微调，大语言模型能够有效促进数字人文领域的数据整理和语言文化信息挖掘工作，为文史知识库构建提供了新思路。一方面，可利用其强大的语言理解和生成能力，自动提取专家知识并组织为结构化形式，提升知识库构建效率。另一方面，可引导大模型从海量文献中挖掘隐性知识，拓宽和加深知识加工的广度和深度。在上述过程中，专家可审核模型提取的信息，并反馈领域知识指导模型优化。这种人机协同模式可兼顾效率和准确性，

构建大规模、高质量的文史知识库，为数据驱动的人文研究提供支撑。

为了更好地利用大语言模型辅助人文研究，我们需要在以下两个方面继续努力：

（一）构建高质量的数据集以增强知识。在大语言模型的应用中，对大规模高质量数据集的需求尤为突出。首先，我们需要将稀缺的文史资源数字化，转换为计算机可识别、可处理的形式。其次，我们需要关注使用数据集的平衡性，尽可能提供全面丰富的数据，以帮助提升模型在文史知识方面的生成和加工能力。

（二）综合计算与人文视野分析结果。大数据与人工智能技术正在逐渐影响人文学科的研究范式。开展交叉方向课题研究时，需要首先正确定位技术在人文研究中的角色，从研究问题出发统摄全局；其次，需拓展传统研究的思路，理解并适应技术的逻辑，选择适合的技术手段；最后，需发挥人文研究者的批判思维和理论总结能力，以理论指导实验，分析错例并提出改进方案。

展望未来，大语言模型在数字人文领域具有广阔的应用前景。当下，多模态大语言模型、多智能体和数字人等前沿技术的发展，将会带来更多的应用契机。例如，多模态大语言模型可以更全面地分析历史文献、艺术作品和文化遗产；多智能体系统可以模拟复杂的人类社会互动和文化现象；数字人技术则为文化遗产的保护和传承提供了新的可能。这些创新应用不仅能加深和拓宽数字人文研究的深度和广度，也为文化遗产和教育提供了新的路径。随着人工智能技术的不断进步，传统人文学科有望迎来新的发展机遇和挑战。正如袁毓林所呼吁的，“应该拥抱和投身于数据/计算密集型的第四/五范式”，因为“这是技术的催促，也是时代的召唤”。^①

^①袁毓林：《人工智能大飞跃背景下的语言学理论思考》，《语言战略研究》2023年第4期。

Leveraging Large Language Models for Building a Cultural and Historical Knowledge Base

Qiu Ziliang, Hu Renfen, Mo Kaijie, Wang Yupei, Liu Zhiying

Abstract: High-quality Cultural and Historical Knowledge Bases are crucial foundations for advancing digital humanities research. In recent years, Large Language Models (LLMs), with their outstanding language understanding and generation capabilities, have presented new opportunities for the deep processing of humanities knowledge. This study, using the analysis of classical Chinese allusions as a case study, evaluates the capabilities of LLMs in the extraction and processing of cultural and historical knowledge through three tasks: structuring allusions knowledge, determining the presence of allusions, and identifying specific allusions used. The experimental results demonstrate that, with well-designed prompts, LLMs can effectively accomplish the organization of cultural and historical knowledge and exhibit a certain degree of judgment in this domain. Additionally, strategies such as Few-Shot Learning, Retrieval-augmented Generation, and Fine-tuning significantly enhance the model's ability to handle issues related to cultural and historical knowledge. Based on these findings, this paper discusses the application strategies for LLMs in the processing of cultural and historical knowledge, including model selection, deployment methods, prompting, and fine-tuning mechanisms. Furthermore, the paper outlines a fundamental approach for constructing a Cultural and Historical Knowledge Base using LLMs and explores potential future directions for the application of LLMs in digital humanities.

Keywords: Large Language Models; Knowledge Base; Classical Chinese Allusions; Digital Humanities; Cultural and Historical Knowledge

附录：典故知识分析三种任务采用的提示词

任务一：典故知识结构化整理	
System	你是一个语料整理助手，请根据任务要求，完成辞典语料格式化整理任务。
User	<p>## 任务说明 ##</p> <p>你会得到一条辞典条目，条目包含典故名称、典故来源、典故释义和例句信息。条目以典故名称开始，后附典故来源。典故来源通常为中国古代历史故事、传说或前人文献中的语词，也可能指向其他辞典条目。典故释义是对典故内容的完整解释，一条完整解释可能包含典故的修辞用法和对单个字词的注解。每个典故可包含一条或多条释义，每条释义后附一条或多条例句。遵循以下规则，完成条目整理任务：1. 典故来源有两种格式，一种是通过诗文直接阐释其来源，另一种是通过“源自”“源见”“同”等标识引述其他辞典条目而不直接包含诗文内容。例如，“一丸同‘一丸泥’”，整理为：{"典故名称": "一丸", "典故来源": "一丸泥"}。每个典故可以有多个来源，将不同来源存储在同一个“典故来源”键下。例如，“营蝇斐锦，源见‘营营青蝇’。又《诗·小雅·巷伯》：‘斐兮斐兮，成是贝锦。’”条目中，“又”标明了典故的第二个来源，整理为：{"典故名称": "营蝇斐锦", "典故来源": "源见‘营营青蝇’。又《诗·小雅·巷伯》：‘斐兮斐兮，成是贝锦。’"}。2. 每个典故可以包含多个释义，每个释义应单独存储，并与对应的例句一起保存。注解单个字词的情况应合并到对应的释义中。例如，“量体裁衣《张融传》：手诏赐融衣曰：‘是吾所着，已令裁剪称卿之体。’后因以‘量体裁衣’谓比照身材裁制衣裳。量，裁度。”条目中，“量，裁度”是对释义中“量”的详细注解，应视为释义的一部分，整理为：{"典故名称": "量体裁衣", "典故来源": "《张融传》：手诏赐融衣曰：‘是吾所着，已令裁剪称卿之体。’", "释义列表": [{"释义": "后因以‘量体裁衣’谓比照身材裁制衣裳。量，裁度。"}], "例句": []} 3. 只提取原文表述，不要修改原文表述。所有键的值能够组合为完整的原输入条目文本。4. 不同键的值不应重复。例句键的值可以为空，其他键的值不可为空。5. 典故来源以“同”格式说明且没有释义信息时，将“典故来源”填入“释义”键的值。如“鱼信同‘鱼书’”，唐胡曾《车遥遥》诗：‘玉枕夜残鱼信绝，金钿秋尽雁书遥。’”整理为：{"典故名称": "鱼信", "典故来源": "同‘鱼书’", "释义列表": [{"释义": "同‘鱼书’"}], "例句": "唐胡曾《车遥遥》诗：‘玉枕夜残鱼信绝，金钿秋尽雁书遥。’"}。6. 输出要求 ##</p> <p>输出为 JSON 格式，结构如：{"典故名称": "...", "典故来源": "...", "释义列表": [{"释义": "..."}], "例句": "..."}。仅返回 JSON 格式。</p> <p>## 参考示例 ##</p> <p>example input: 犄角《左传·襄公十四年》：“譬如捕鹿，晋人角之，诸戎犄之，与晋掎之。”孔颖达疏：“角之谓执其角也；犄之言戾其足也。”后以“犄角”谓分兵牵制或夹击敌人。《三国志·陆逊传》：“犄角此寇，正在今日。”亦谓分兵互相呼应。《旧唐书·德宗纪下》：“宜令诸道各出师徒，犄角齐进。”</p> <p>example output: {"典故名": "犄角", "典源内容": "《左传·襄公十四年》：“譬如捕鹿，晋人角之，诸戎犄之，与晋掎之。”孔颖达疏：“角之谓执其角也；犄之言戾其足也。”", "释义列表": [{"释义": "后以“犄角”谓分兵牵制或夹击敌人。"}, {"释义": "亦谓分兵互相呼应。"}], "例句": "《三国志·陆逊传》：“犄角此寇，正在今日。” 《旧唐书·德宗纪下》：“宜令诸道各出师徒，犄角齐进。”"}。7. 输入文本 ##</p> <p>以下是需要整理的辞典条目： ```\${text}```</p>

续附表

任务二：用典判断	
System	你是一位精通中国古代文化知识的专家，请判断语料是否使用了典故。
User	<p>## 任务说明 ##</p> <p>用典指的是通过引用古籍中的故事或词句，使文本具有超出字面意思的深层文化含义。给定一条语料，请判断其是否用典（label），并说明你的判断依据（allusion_evidence）。如果使用了典故，指出文本中的用典片段（allusion_span）。</p> <p>## 输出要求 ##</p> <p>输出为 JSON 格式，结构如：{"label":int,"allusion_evidence":"...", "allusion_span":["片段 1","片段 2"]}。"label" 是一个整数，使用典故则为 1，否则为 0。"allusion_evidence" 是一个字符串，描述你的判断依据。"allusion_span" 是用典的文本片段的字符串列表，请保留原文。仅返回 JSON 结果。</p> <p>## 参考示例 ##</p> <p>example_input: 愿天下人少察狂瞽之言，毋蹈前车之覆。example_output:{"label":1,"allusion_evidence": "“狂瞽之言”出自《十渐不克终疏》，指愚妄无知的言论，也用于自谦。“前车之覆”，也作“前车之鉴”，《晋书》中说“前车之覆轨，后车之明鉴”，比喻以往的失败，可以当作后来的教训。这句话的意思是说，希望天下人不被愚昧言论蒙蔽，不要重复失败的教训。“狂瞽之言”和“前车之覆”出自古籍中的故事和言论，简练地传达了丰富的意蕴，因此判断使用了典故。", "allusion_span": ["狂瞽之言", "前车之覆"]}</p> <p>## 输入文本 ##</p> <p>以下是需要判断的文本：``{text}``</p>
任务三：典故识别（以典故名称作为 RAG 信息）	
System	你是一位精通中国古代文化知识的专家，请帮助识别语料中的用典现象。
User	<p>## 任务说明 ##</p> <p>用典指的是通过引用古籍中的故事或词句，使文本具有超出字面意思的深层文化含义。给定一条语料，请判断其中是否用典，如果用典，请指出其所用典故。你会得到一个从外部知识库匹配的候选列表作为参考，列表包含十组候选典故，格式为：["候选典故 1","候选典故 2","候选典故 3"]，但注意语料所用典故不局限于此表，语料也可能没有使用典故。请逐一查看后再做出判断。</p> <p>## 输出要求 ##</p> <p>输出为 JSON 格式，结构如：{"result":["典故 1","典故 2"]}。"result" 的值是一个 list 字段，包含识别出的典故名称（str）。如果未识别到典故，返回空列表。仅返回 JSON 格式。</p> <p>## 参考示例 ##</p> <p>example_input: "莼鲈归思满吴淞，渺渺江云隔几重。" example_candidate_list:["云树之思","白云亲舍","莼羹鲈脍","范蠡扁舟","雁足书","羊裘钓","华亭鹤唳","日近长安远","一苇","罗浮梦"] example_output: {"result":["莼羹鲈脍"]}</p> <p>## 输入文本 ##</p> <p>以下是需要判断的文本：``{sentence}`` 以下是候选列表：``{waitinglist}``</p>

（编辑：饶高琦）