融合图尔敏模型与大语言模型的社会科学循证知识库构建方法*

齐涵悦[1]　杨颜僖[2,3]　林泽斐[1]

（1. 福建师范大学社会历史学院，福州 350117；2. 中国科学院文献情报中心，北京 100049；
3. 中国科学院大学信息资源管理系，北京 100049）

摘要：社会科学领域的循证研究面临文献数量快速增长及论证情境复杂的双重挑战，亟需智能、高效的证据处理方法。本研究提出一种融合图尔敏模型和大语言模型的社会科学循证知识库构建方法，旨在提升社会科学证据的处理效率及证据分析的智能化水平。首先，设计以论证为中心的信息抽取流程；其次，构建面向语义检索的多维度知识组织模型；最后，设计基于链式推理的智能循证服务流程，并以图书情报学为例，开发原型系统ToulminQA。实验评测表明，ToulminQA在论证要素抽取方面具有较高的准确性与可靠性，同时系统能够有效整合大量证据，形成逻辑清晰、具备实用价值的循证结论。本研究为社会科学领域提供了一种新的循证研究路径及具有参考价值的实践范例。

关键词：图尔敏模型；大语言模型；社会科学；知识库构建；循证社会科学；证据；知识组织

中图分类号：C37; G356　DOI: 10.3772/j.issn.1673-2286.2025.10.001

引文格式：齐涵悦，杨颜僖，林泽斐. 融合图尔敏模型与大语言模型的社会科学循证知识库构建方法[J]. 数字图书馆论坛，2025，21（10）：1-11.

1992年，为响应流行病学家阿奇·科克伦（Archie Cochrane）[1]关于医疗决策应建立在可靠证据与系统检验基础之上的倡议，戈登·盖亚特（Gordon Guyatt）等[2]正式提出“循证医学”一词，标志着循证范式在医学领域兴起。此后，这一范式渗透至公共政策、教育学、心理学、社会工作等社会科学领域，催生出“循证社会科学”这一新兴的学术方向。循证社会科学致力于收集、综合、评估和利用社会科学证据，推动从理论研究到决策实践的有效转化，最终实现决策更科学、服务更有效、学科更专业的目标[3]。

循证范式从医学向社会科学移植，并非简单的概念平移。社会科学所研究的现象高度依赖情境，且人类行为和社会互动的因果链条往往不是孤立、静态的，会同时受时空背景、制度因素、文化语境的影响。因此，社会科学的证据形态更加多样和复杂，除了要回答“什么有效”以外，还要深入探究“为什么有效”“对谁有效”以及“在何种条件下有效”。这种复杂性对社会科学循证实践所依赖的知识基础设施建设提出了更高要求。

循证知识库是现有循证社会科学领域的关键知识基础设施[4]，其旨在通过收集、评估、整合高质量研究证据来缩小学术研究与社会实践之间的落差，为政策制定者、教育工作者及社会服务管理者等提供可靠的决策参考支持。尽管社会科学领域的循证知识库在近年来得到了快速发展，但文献数量的激增与论证结构的复杂性相互交织，成为知识库构建面临的两大挑战。现有循证知识库构建方法的核心局限在于难以在保障处

收稿日期：2025-08-20
*本研究得到福建省社会科学基金项目“基于大型语言模型的社会科学循证知识库构建”（编号：FJ2024BF034）资助。

理效率的同时，有效表征社会科学研究中复杂的、情境依赖的论证结构。

英国哲学家斯蒂芬·图尔敏（Stephen Toulmin）[5]提出的论证模型（图尔敏模型）与社会科学处理复杂论证的需求相契合，且大语言模型（LLM）在语义理解和逻辑推理方面取得了突破性进展，为实现论证要素的自动化抽取、评价和整合提供了可能。基于此，本研究提出一种融合二者的社会科学循证知识库构建方法，以期突破当前循证实践中处理规模与分析深度之间的固有矛盾，也为社会科学循证研究的智能化发展提供可资借鉴的方法论支撑。

1　相关研究与实践

1.1　社会科学循证知识库的构建

循证理念在社会科学领域的推广催生了多个专业化循证知识库的形成，根据构建策略可以分为两种典型模式。

（1）以Campbell协作网为代表的“深度综合”模式。这一模式的核心是严谨的系统评价，构建时高度依赖专家团队的人工操作，核心知识单元为经过同行评议的、长达数十乃至上百页的完整评估报告，最终产出为以文档为中心的报告库。Campbell协作网自2000年成立以来，通过其系统评价库持续汇总社会福利、犯罪、司法、教育、国际发展等领域的高质量系统评价报告，为全球社会科学研究以及政策决策提供证据支持[6]。为了构建高度的认知可信度，该模式通过详尽的定性叙述或定量元分析来保证综合分析的深度、透明度与方法论严谨性。然而，这种对严谨性的追求导致其很难在更新周期上应对学术文献数量指数级增长这一趋势。截至2025年8月，Campbell协作网系统评价库中已发布的报告数量为304篇[7]，这一数字表明Campbell协作网虽然产出高质量报告，但与当前海量的社会科学研究成果相比，其处理规模仍然有限。

（2）以有效实践信息库（What Works Clearinghouse，WWC）及Results First平台为代表的“有损压缩”模式。这一模式旨在提升证据结论的可及性和决策效用。美国教育科学研究所于2002年建立WWC，借助系统评价报告对教育干预措施的有效性进行分析[8]。WWC通过经认证的评审员和一套标准化的流程，将复杂的原始研究评估总结为对特定干预措施有效性的6个离散评级[9]，并提供名为“Find What Works”的在线查询系统，最终构建以干预措施为中心的、可进行公开检索的循证知识库。美国皮尤慈善信托基金会于2015年建立的Results First平台汇集了包含WWC在内的9个国家级知识库、逾3 100项社会政策项目信息，并通过统一的颜色编码进行效果评级，方便决策人员比较并获取不同社会项目的有效性证据[10]。虽然该模式在一定程度上提升了规模化处理能力，但为了实现跨研究的标准化评级，不得不剥离支撑结论的完整论证链条（如理论支撑、适用条件等），削弱了用户对“为什么有效”“对谁有效”以及“在何种条件下有效”等关键问题的深层理解。

以上主流方案反映出当前社会科学循证知识库在分析深度与证据情境保全之间难以兼得的困境。“深度综合”模式提供了高保真、可信赖的论证，但生产速度限制了其规模和时效性；“有损压缩”模式可以实现规模化，但无法保留关键的论证情境和解释力。因此，需要探索既能以规模化的方式高效率处理证据，又能用情境化的方式保留必需的论证细节的新型知识库构建模式。

1.2　图尔敏模型在社会科学中的应用

图尔敏模型打破了传统形式逻辑中简单的“证据-结论”二元结构，把论证看作由6个关联要素构成的复杂过程[5]。这一模型在社会科学研究中具有广泛的适用性和应用价值。例如：在教育学领域，Osborne等[11]对课堂讨论与学生书面作业中的论证结构开展了分析工作，发现图尔敏模型有助于判断学生论证能力的发展水平和存在的问题；在犯罪心理学领域，Alison等[12]分析了21份犯罪心理画像报告，发现其中大量断言缺乏严谨的论证依据，建议引入图尔敏模型来构建和评估犯罪心理画像；在组织行为学研究中，Simosi[13]运用图尔敏模型开展了工作场合的员工论辩结构分析工作，进而探讨日常话语中的论证建构过程。

随着自然语言处理（NLP）技术的不断发展，有学者开始探索基于图尔敏模型的论证信息抽取及组织方法。Vitali等[14]提出了论证模型本体，对图尔敏模型要素开展了语义表示工作。Song等[15]借鉴图尔敏模型基本思想，构建了链接学术论文论证单元以及论证关系的知识图谱。Yang等[16]融合加权特征和BiLSTM-Attention模型，从学生写作中抽取图尔敏论证要素。

传统的NLP技术在自动抽取论证信息方面存在较大挑战，在处理社会科学复杂论证时具有局限性。比如德国达姆施塔特工业大学于2017年开发的ArgumenText系统利用基于注意力机制的神经网络和BiLSTM模型区分文本中的支持论点与反对论点以实现对论点句的自动识别与立场分类，但是这一系统的准确率不高，有47%的抽取结果被评估为不相关、缺乏逻辑或者立场错误[17]。为更新技术路径、突破该任务核心瓶颈，需要引入拥有强大语义理解和推理能力的大语言模型。

1.3　大语言模型在循证社会科学领域的应用

2022年以来，大语言模型在语义理解、逻辑推理和内容生成等方面获得了突破性进展，并开始被应用于科研辅助领域。例如：爱思唯尔的Scopus AI使用检索增强生成技术，依托AI综合分析Scopus数据库中海量同行评议期刊的摘要，为研究者提供科研文献检索、概念地图构建和领域专家推介等服务[18]；中国知网开发的知网研学AI利用大语言模型的文本分析、概括、总结能力，向科研人员提供课题选择、文献分析、学术综述等多项智能化服务[19]。

尽管大语言模型在科研辅助方向拥有广阔的应用前景，但在循证研究中系统性运用大语言模型的探索仍然十分有限。目前相关研究主要集中在医学与教育两个领域。在医学领域，Wang等[20]设计的TrialMind系统可以自动化地从非结构化的临床结果中提取证据，并处理成可以用于元分析的输入格式。在教育领域，褚乐阳等[21]设计的GPT4EBP系统旨在为教育工作者提供基于文献证据的交互式咨询服务。上述研究仅针对单一垂直领域，并且两者未能涉及论证要素的深度知识组织。

综合相关研究与实践可以看出，现有的社会科学循证知识库在处理规模与论证深度之间难以平衡，亟须探索智能高效的解决方案。图尔敏模型虽能刻画社会科学复杂的论证逻辑，但传统技术限制了精准的自动化抽取；大语言模型虽展现出强大的语义理解潜力，但目前仍缺乏在证据知识组织方面的系统性应用。因此，本研究拟将图尔敏模型的结构化分析框架与大语言模型的智能化处理能力相融合，提出一种新的社会科学循证知识库构建方法，以期在保障大规模文献处理效率的同时，有效保留循证决策所需的关键论证细节。

2　理论基础与协同机制

2.1　图尔敏模型

本研究以图尔敏模型为理论基础。图尔敏模型将完整的论证流程拆分为6个相互联系的核心要素，据此揭示从证据到结论的推理路径以及成立条件[5]。为便于理解，构建一个教育学领域示例予以说明。①主张，论证者所确立的核心观点或者结论，是进行论证的主要目标，例如“在小学阶段引入哲学教育，可以有效地提升学生的批判性思维能力”。②证据，为主张提供实质性支撑的事实性内容，包括数据、实验结果或者统计分析等，例如“一项涉及500名小学生的随机对照实验结果显示，经过一学期哲学教育，实验组学生在批判性思维测试中的平均得分为78.5分，显著高于对照组的65.2分（$p<0.01$）”。③保证，用于解释证据和主张之间的逻辑联系，即特定证据为何可以支持主张，例如“系统的哲学教育引导学生进行概念分析、逻辑推理和论证评估，从而使得实验组的学生在批判性思维测试中得分显著高于对照组，说明其批判性思维的核心认知能力得到了锻炼”。④支撑，为保证提供的外部理论或事实支撑，以此提高论证的可信度，例如“皮亚杰的认知发展理论表明，小学阶段正处于具体运算向形式运算过渡的关键时期，此时进行抽象思维训练效果较好”。⑤限定，主张的适用范围、程度或者成立条件，体现出论证者自身对于研究的审慎态度，比如“本研究结论只在教师接受过专业的儿童哲学教学培训的条件下成立”。⑥反驳，削弱或否定主张的特殊情况、反例或者质疑，呈现论证的批判性维度。反驳可由论证者自身预设性地指出，也可以由其他反对方提出，例如“该研究只进行了一学期的追踪，长期效果未经验证；实验组教师的教学热情或产生霍桑效应，从而影响结果的客观性”。

图尔敏模型的独特价值在于超越了简单的“证据-结论”二元论证结构，利用保证和支撑给出论证逻辑，依靠限定与反驳明确了结论的适用边界以及局限，这一特性使它特别适合处理社会科学中普遍存在的情境依赖性及条件限制性论证。

2.2　图尔敏模型与大语言模型的协同机制

在处理社会科学的复杂论证中图尔敏模型与大语

言模型具有显著的互补性。前者可以提供结构化的解析框架，捕获“什么有效”“为什么有效”以及适用边界，但是在学术文献中保证与支撑等关键要素往往缺乏显著的形式特征，因而通过规则或传统NLP技术难以有效抽取；后者基于强大的语义理解与逻辑推理能力可以捕获学术文献里的隐性论证，但如果缺乏明确的抽取框架，则很难系统化地抽取证据知识。因此，本研究设计了“结构化约束+语义化推理”的融合机制，将图尔敏模型的六要素结构作为提示语，为大语言模型确立明确的抽取目标与逻辑边界；同时利用大语言模型的深度语义推理能力，从非结构化文本中识别显性证据与隐性论证逻辑，这既提高了大规模处理的效率，又有利于实现对复杂情境依赖性论证的深度处理。

3 循证知识库构建方法

本研究依据信息抽取、知识组织、智能服务的路径构建融合图尔敏模型与大语言模型的社会科学循证知识库（见图1）。通过3个阶段，将非结构化的学术文献转化为支持智能化循证服务的知识体系，以期在保证处理效率的同时相对完整地捕获和利用社会科学研究所需的论证细节。

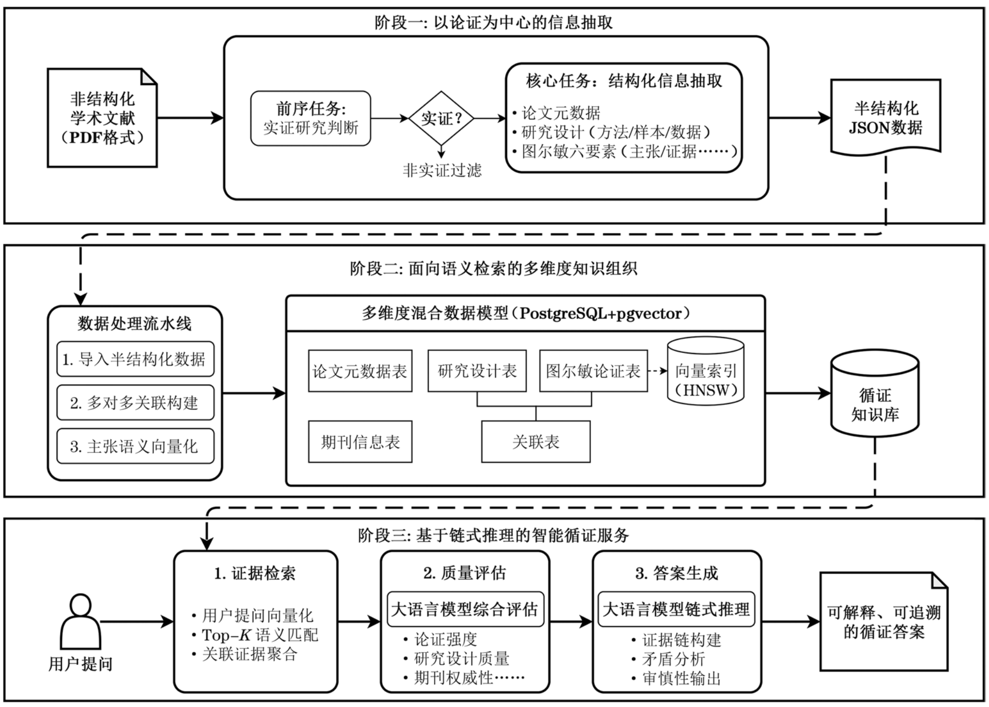

图1　融合图尔敏模型与大语言模型的社会科学循证知识库构建框架

3.1 以论证为中心的信息抽取

本阶段旨在从非结构化的学术文献中提取半结构化的论证信息。由于社会科学文献版面复杂、论证逻辑嵌套，本研究采用“端到端”的抽取策略。将PDF文件的二进制流与结构化指令输入具备长上下文理解能力的多模态大语言模型，避免了传统光学字符识别（OCR）方法造成的内容损失与误差累积。

为高效处理大规模文献并确保输出结果的一致性，设计了基于并行架构的抽取流程，并将模型生成的温度系数设置为0以降低随机性。信息抽取处理流程如算法1所示。

算法1：信息抽取处理流程

输入：学术文献PDF集合D，提示语模板P

输出：结构化JSON数据集合J

参数：并发线程数N，温度系数T=0

初始化线程池Executor(max_workers=N)

procedure ExtractProcess(pdf_file)

　binary_stream ← ReadFile(pdf_file)

　Response ← LLM.generate_content

　　(contents=[binary_stream, P],

　　config={temperature: T, format: “application/

　　　json”})

　Data ← ParseJSON(Response.text)

```
        if Data is Valid then
            Save Data to J
        end if
    end procedure
    for each file in D do
        Submit ExtractProcess(file) to Executor
    end for
```

本阶段使用的信息抽取提示语如提示语1所示。该提示语首先定义模型角色，并要求其执行实证性判断这一前置任务。这是为了过滤社会科学领域中的纯理论探讨或文献综述类文章，确保知识库的数据源均为基于数据的实证研究。若判断为实证研究，要求模型严格遵循预定义的JSON Schema进行信息抽取，涵盖了论文元数据、研究设计细节及图尔敏论证六要素。

提示语1：信息抽取提示语

上下文信息：单篇PDF格式论文全文

角色设定：你是一位专业的学术信息抽取助手

前序任务：判断输入的论文是否属于实证类研究。若论文未展现任何针对研究问题的数据收集或数据分析过程（如问卷、访谈、实验、文本挖掘等），或者仅为纯理论推演、文献综述等，视为非实证类直接输出{“error”: “non_empirical”}

核心任务：若属于实证类研究，按以下Schema抽取

{“paper”: {“title”: “标题”, “authors”: [“作者”, ...], “abstract”: “摘要”, “keywords”: [“关键词”, ...], “publish_year”: int, “funding”: [“资助基金”, ...]}, “study_designs”: [{“study_design_id”: int, “study_type”: [“研究类型”, ...], “sample_size”: int|null, “data_source”: [“数据来源”, ...], “analysis_techniques”: [“分析方法”, ...], “study_description”: “研究目的和过程”, “toulmin_argument_ids”: [int, ...]}, ...], “toulmin_arguments”: [{“toulmin_argument_id”: int, “claim”: “主张”, “evidence”: “证据”, “warrant”: “保证”, “backing”: “支撑|null”, “qualifier”: “限定|null”, “rebuttal”: “反驳|null”, “study_design_ids”: [int, ...]}, ...]}

示例：……

研究设计与图尔敏论证之间存在多对多的对应关系。例如一项研究可能同时采用问卷调查和深度访谈两种研究设计来支撑“社交媒体使用对青少年阅读习惯有负面影响”的论证，而问卷调查的数据还可以支撑“不同社交媒体类型对阅读偏好有不同影响”的论证。如提示语1所示，为了准确地表示这种复杂关系，提示语要求大语言模型通过ID数组建立双向映射，即为每个研究设计和论证单元分配唯一的ID标识符，通过study_design_ids和toulmin_argument_ids数组记录关联关系。研究设计与图尔敏要素之间的具体对应关系由大语言模型基于对文献内容的理解自动判断并生成，可以避免为每种研究范式预设固定的映射规则，保证系统处理不同类型和结构的社会科学研究的灵活性。

这种结构化的提示设计和智能化的关系识别有利于系统捕获社会科学研究中的复杂论证结构，在保留了研究关键发现的同时，还记录了支撑这些发现的方法学细节、推理逻辑、理论基础、适用边界以及潜在局限，为后续的知识组织和智能服务阶段奠定了基础。

3.2　面向语义检索的多维度知识组织

本阶段的任务是将阶段一抽取的半结构化信息进行有效存储与组织。为此，设计能够保持论证结构完整性并支持高效语义检索的多维度数据模型，进而通过系统化的处理流程完成知识组织。

本研究设计的社会科学循证知识库的数据模型（见图2）由5个核心数据库表构成，采用关系数据模型以及向量存储的混合架构：一方面存储文献元数据、研究设计及图尔敏论证要素等信息；另一方面对论证主张文本进行向量化处理，以支持后续的语义检索。①论文表（Paper）记录论文的基本情况。其中，作者、关键词和基金项目等信息采用半结构化的JSON数组存储，以增加扩展性。②研究设计表（StudyDesign）将研究方法信息作为单独实体建模。一篇论文可同时包含多个研究设计，对每个设计记录其样本数量、数据来源和数据分析方法等。③图尔敏论证表（ToulminArgument）构成知识库的核心，映射图尔敏模型的六要素。特别是该表中claim_vector字段存储了主张文本的高维语义向量表示，以支持基于语义相似度的智能检索。一篇论文允许含有多个图尔敏论证。④关联表（ToulminArgumentToStudyDesign）实现了论证与研究设计之间的多对多关系映射。本设计有利于系统灵活地处理复杂的证据关系。例如，追踪某个论证依赖的所有研究设计，或查询某个研究设计所支持的所有论证。⑤期刊表（Journal）存储期刊的评价指标，包括被权威索引收录情况、受社科基金资助的情况及期刊影响因

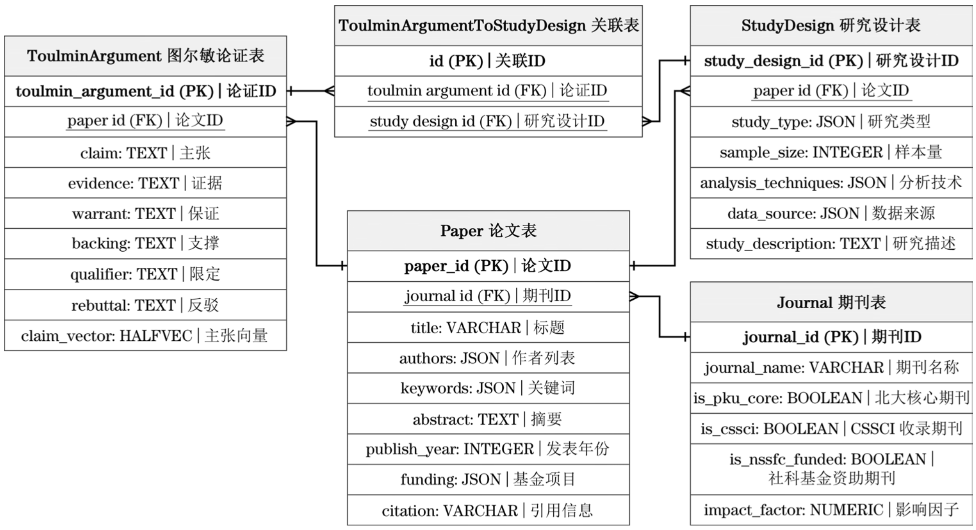

图2 社会科学循证知识库的数据模型

子等。这些指标为证据质量评估提供了辅助依据。

本阶段的知识组织处理流程如算法2所示，包括结构化数据导入和主张向量化两个环节。在结构化数据导入环节，论文元数据、研究设计和图尔敏论证被依次存入数据库，同时通过ID映射表维护JSON文件中临时ID与数据库实际ID间的对应关系，从而正确建立论证与研究设计之间的多对多关联。主张承载了论证的核心观点，是用户查询与证据匹配的关键语义锚点，因此选择主张作为向量化对象。在主张向量化环节，针对所有未向量化的主张文本，程序调用嵌入模型接口生成语义向量并更新至claim_vector字段。

算法2：知识组织处理流程

输入：JSON文件集合*J*，嵌入模型EmbedModel

输出：结构化存储的循证知识库

```
//第一部分：结构化数据导入
for each json_file in J do
  Data ← LoadJSON(json_file)
  paper_id ← DB.insert(Paper, Data.paper)
  for each design in Data.study_designs do
    DB.insert(StudyDesign, design, paper_id)
  end for
  for each arg in Data.toulmin_arguments do
    toulmin_id ← DB.insert(ToulminArgument, arg, paper_id)
    DB.bindRelation(toulmin_id, arg.study_design_ids)
  end for
end for
//第二部分：主张向量化
claims ← DB.query(ToulminArgument, where: claim_vector IS NULL)
for each (id, text) in claims do
  vector ← EmbedModel.embed(text)
  DB.update(ToulminArgument, id, claim_vector=vector)
end for
```

选用PostgreSQL作为数据库管理系统，利用其pgvector扩展支持高维语义向量的存储与查询，并为claim_vector字段构建基于层次导航小世界（HNSW）算法[22]的向量索引，从而能够在亚线性时间复杂度内完成大规模向量的相似度检索。此外，期刊表的数据通过调用基于Dify平台的自建接口服务，根据论文标题和作者信息自动检索并补充。

本阶段设计的多维度数据模型同时表现出完整性和灵活性。完整性体现在对研究设计、研究质量等支撑性要素的详细保留，使证据可追溯、可评估；灵活性体现在关联表的多对多映射机制可以适配社会科学研究中的多重研究设计与复杂论证关系。结构化存储与向量索引的混合架构进一步拓展了循证知识库的检索能力，为后续智能循证服务阶段提供数据结构支撑。

3.3 基于链式推理的智能循证服务

经过前两个阶段，循证知识库已具备提供智能服务的数据基础。本阶段旨在为用户提供可解释、可追溯

的智能问答服务，将用户的自然语言提问，按照证据检索、质量评估、答案生成3个步骤转换为有理有据、逻辑清晰的循证答案，实现知识库的最终价值。智能循证服务处理流程如算法3所示。

算法3：智能循证服务处理流程

输入：用户查询Q，检索数量K，嵌入模型EmbedModel，评估模型RefineModel，推理模型ReasonModel

输出：可解释的循证答案Answer

```
//第一部分：证据检索
query_vector ← EmbedModel.embed(Q)
candidates ← DB.vectorSearch(ToulminArgument.claim_vector, query_vector, limit=K)
evidence_list ← []
for each claim in candidates do
    paper ← DB.query(Paper, claim.paper_id)
    journal ← DB.query(Journal, paper.journal_id)
    designs ← DB.query(StudyDesign, claim.id)
    evidence_list.append(Package(claim, paper, journal, designs))
end for
//第二部分：质量评估
prompt_refine ← BuildRefinePrompt(Q, evidence_list)
response ← RefineModel.generate(prompt_refine, format=JSON)
refined_list ← FilterAndReorder(evidence_list, response.ranked_relevant_ids)
//第三部分：答案生成
prompt_summary ← BuildSummaryPrompt(Q, refined_list)
Answer ← ReasonModel.stream_generate(prompt_summary)
return Answer
```

在证据检索环节，首先，借助嵌入模型把用户的自然语言问题转化为机器可理解的语义表示向量。此阶段，为确保用户问题向量与数据库中存储的主张向量位于同一语义空间，采用与3.2节中生成claim_vector字段相同的嵌入模型。其次，采用以主张为核心的语义匹配策略，通过计算用户查询向量与数据库中所有主张向量的余弦相似度来提取语义相关的Top-K个主张。这种基于语义的检索方法可以处理同义词、近义词等语言变化，即使用户查询与主张的表述方式不完全一致，也能保障相关语义主张的检出。最后，利用数据模型中的外键关系，自动聚合主张相关的图尔敏论证、研究设计、论文、期刊等信息，包装为半结构化JSON数据格式，形成候选证据集合。

候选证据集合与用户提问有语义上的相关关系，但并不代表相关证据可直接用于回答用户提问。因此，在质量评估环节，需要对候选证据进行深度的语义筛选与排序。为了解决传统检索系统无法理解证据逻辑强度的问题，构建了专门的质量评估提示语（见提示语2），指示大语言模型从内容相关性、论证逻辑、方法学质量等维度对证据进行综合评估。

提示语2：质量评估提示语

上下文信息：用户提问，候选证据集合

任务：根据用户提问和候选证据集合，对证据进行过滤和排序

评估标准：①与问题的相关性。evidence是否适合作为回答用户问题的证据。②论证强度。evidence的充分性、warrant的逻辑性、backing的可靠性。③研究设计质量。研究类型是否恰当、样本量是否充足、分析方法是否严谨、数据来源是否可靠。④期刊质量。期刊的核心地位、影响因子等（作为辅助参考）。⑤反驳与限定。qualifier和rebuttal的存在可能影响结论的普适性，须纳入考量。

输出格式：最终输出结果为JSON格式，包含irrelevant_ids和ranked_relevant_ids两个数组。①识别与query相关度低（难以直接回答）的证据，将其ID归入irrelevant_ids。②对保留的evidence按照上述评估标准进行综合质量排序，从高到低将ID按顺序归入ranked_relevant_ids。

通过上述提示语，系统利用大语言模型的语义理解能力，不仅自动剔除了不相关的噪声数据，还实现了基于证据质量而非单纯文本相似度的排序，确保了后续生成的答案建立在高质量的实证研究基础之上。

在答案生成环节，使用具备链式推理能力的大语言模型（如DeepSeek-R1）生成基于质量评估环节所得证据及其相关信息的综合回答。通过强化学习技术，此类大语言模型获得了先思考后回答的能力，可以进行多步骤的逻辑推理。本研究设计了专门的答案生成提示语（见提示语3），要求模型在生成过程中显式引用证据来源，并运用图尔敏思维框架进行辩证分析。

提示语3：答案生成提示语

上下文信息：用户提问，相关证据列表

任务：基于用户提问和相关证据列表，给出一个系统性的综合回答

核心指令要求：①事实支撑。回答须有实证研究过程和证据支撑。②引用规范。在证据描述后用括号注明证据ID，如（证据12）。③审慎性。结合证据中的图尔敏要素（如限定、反驳），对结论的适用边界进行说明。④质量优先。优先使用高质量证据。⑤证据链构建。尝试串联多条证据，形成逻辑链条。⑥分歧处理。证据之间存在矛盾或分歧时，分析导致差异的潜在原因，呈现辩证视角。

输出格式：按照“简明引言-分点论述-简明结语”的结构输出自然语言回答。

在该提示语的约束下，大语言模型能够识别不同证据之间的逻辑关系，自动构建证据链并生成兼具逻辑性和可读性的循证答案。当证据之间存在矛盾时，模型依据指令分析矛盾产生的可能原因（如样本差异、研究情境不同等），而不是简单选择一方，从而帮助用户形成全面、辩证的认识。

4　实证研究

4.1　领域选择与数据准备

为了检验上述构建方法的可行性和有效性，本研究以图书情报学为实证领域，开发名为ToulminQA的原型系统。选择图书情报学作为实证学科有两个方面的考量。一方面，该学科有开展循证研究与实践的历史。21世纪初期兴起的循证图书馆与情报实践即倡导在图情实践中应用最佳的研究证据。另一方面，该领域研究方法定量与质性并存，具备多样化的特性，能够充分验证系统对不同研究范式的适应性，因而该学科循证知识库的构建经验对其他社会科学领域有重要的示范价值。

具体而言，实证研究主要聚焦于阅读推广和信息行为两个领域。这两个领域分别反映了图书馆学和情报学研究的重要实证方向。本研究借助中国知网数据库，获取2015—2025年发表的《中文核心期刊要目总览》及CSSCI来源期刊论文全文。在具体检索策略上，针对阅读推广领域，以“阅读推广”为主题词进行检索；针对信息行为领域，考虑到该主题下位概念丰富，为保证查全率，采用了扩展检索策略，选取“信息行为”“信息搜寻”“信息分享”“信息规避”“信息偶遇”“信息采纳”作为主题词联合检索。最终共获得阅读推广领域相关论文3 211篇，信息行为领域相关论文2 308篇，总计5 519篇论文作为知识库的数据基础。

需要说明的是，本研究以学术期刊论文为证据来源文献，主要基于3个方面的考虑：①经过严格的同行评议，质量较为可靠；②包含完整的研究设计描述和论证过程，信息密度高；③是社会科学研究成果的主要载体，具有良好的代表性。尽管本研究以期刊论文为文献来源，但是所提出的方法框架具有扩展性，未来可以扩展至包括学位论文、研究报告、会议论文在内的其他类型学术文献。

4.2　循证知识库构建

知识库构建遵循前述方法，通过论证信息的抽取、存储与组织，将非结构化学术文献转化为可查询的循证知识库。

在信息抽取的具体实现上，选用Gemini 2.5 Pro多模态大语言模型作为提取工具。与其他大语言模型相比，Gemini 2.5 Pro具备处理百万级词元超长上下文的能力，能够一次性处理完整的学术论文；其多模态理解能力使系统可直接读取PDF文件，解析其中的文本、表格、图片等多种形式的信息，克服了传统OCR转换造成的信息损失和误差。本研究借助多线程处理方式，对5 519篇论文进行并发抽取，要求模型先判断每篇论文是否为实证类，若为实证类论文则根据图尔敏模型框架，系统化抽取论文元数据、研究设计信息和6类论证要素，借助ID标识建立论证单元与研究设计之间的关系。在5 519篇论文中，有3 218篇（58.3%）被判定为实证类并进行完整的信息抽取，另有2 301篇论文被识别为非实证类而未进行抽取。在数据存储技术方面，选用开源数据库管理系统PostgreSQL作为解决方案，使用Gemini Embedding嵌入模型，为claim字段的每个主张文本生成3 072维的语义向量。

经过系统化的信息抽取与组织，本研究构建的循证知识库数据规模如表1所示，共包含3 218篇实证类论文

表1　循证知识库的数据规模

类别	实证类论文数/篇	研究设计数/个	图尔敏论证单元数/个
阅读推广领域	1 608	2 007	5 475
信息行为领域	1 610	1 862	6 828
合计	3 218	3 869	12 303

文、3 869个研究设计和12 303个图尔敏论证单元（每个论证单元含6个要素）。

在研究方法分布上，两个领域呈现出明显的特性差异。统计每篇论文所运用的研究方法（一篇论文可运用多种方法）可以发现：阅读推广领域以案例研究（51.45%）为主，其余依次为质性研究（28.25%）、文献计量（21.50%）、问卷调查（19.25%）以及行动研究（11.75%）等，体现了该领域侧重定性研究以及经验总结的特性；信息行为领域则以问卷调查（39.86%）最为常见，其余依次为质性研究（25.12%）、访谈研究（17.81%）、案例研究（13.77%）和文献计量（12.05%）等，反映了这一领域更重视对用户行为规律的定量分析。

为防范大语言模型的幻觉问题，采用人工方法对信息抽取准确性进行评估，具体方法为：从两个数据集的实证论文中各随机抽取15篇，共计30篇论文，每篇论文随机选择一个完整的六要素图尔敏论证单元，共计180个要素样本。对于主张、证据、支撑、限定4类要素，核查所抽取内容是否在原文中有明确对应的表述；对于保证要素，由于其旨在解释证据和主张间的逻辑联系，核查是否正确反映了原文作者建立证据与主张之间逻辑联系的推理过程；对于反驳要素，核查是否针对原文论证进行了科学合理的质疑。考虑到社会科学文献的特性，支撑、限定和反驳3类要素允许为null值，分别表示文献中未明确提及理论支撑、未说明适用限制，以及研究设计严谨无明显可反驳之处。评分采用二分类标准，将结果判定为“未发现明显幻觉”与“存在明显幻觉”（明显的数据虚构、事实扭曲、逻辑错误等）两类之一。

评估由2名图书情报专业研究生背对背独立完成。评估前，2名评估者接受了针对图尔敏模型和评分标准的培训。评估结果显示，2名评估者的独立评估结果未出现分歧，表明该评估结果具有较高的可靠性。主张、保证、支撑、限定、反驳5类要素的幻觉率均为0，证据要素中出现了1例明显幻觉，总体幻觉率为0.55%（1/180）。整体而言，系统在论证信息抽取方面表现良好，但在个别情况下存在数据虚构，这提示在较严谨的应用场景中，仍有必要对证据信息作额外审核。

4.3　原型系统的开发与评估

本研究在循证知识库的基础之上，运用3.3节所描述的方法，开发了名为ToulminQA的循证问答原型系统（https://historylab.cn/qa），从而验证所提方法在实际应用中的有效性。

在进行证据检索时，系统首先使用Gemini Embedding嵌入模型生成用户查询向量，然后基于claim字段的向量索引，以Top-K阈值初步筛选出相关主张及其论证和质量信息，从而获得候选证据集。为了应对多种场景的需要，系统预设了3种Top-K阈值：快速查询（K=10）适用于快速了解研究概况、把握核心证据；普通查询（K=30）是默认选项，有效平衡了查询速度和结果覆盖率，可以满足大多数研究者的查询需求；深度查询（K=50）则为希望详尽了解研究情况的用户提供更多证据。得到候选证据集后，调用DeepSeek-V3模型，通过提示语指示模型全面考虑证据强度、研究设计质量、期刊质量以及反驳与限制等要素，对候选证据集进行实时的过滤与排序。

质量评估环节可以自动过滤质量不高、关联性弱的证据，并将重要证据的排名提前。例如，针对信息行为领域“生成式AI对用户信息搜寻行为产生何种影响”这一查询，ToulminQA在深度查询模式下经DeepSeek-V3智能过滤后，返回了20条高品质证据。其中，排名首位的是王若佳等[23]的准实验研究：该研究招募40名大学生分别使用传统全文搜索引擎Bing和集成GPT-4的New Bing完成5种检索任务，通过屏幕录制软件记录用户行为数据，并利用Wilcoxon非参数检验等方法进行分析，具有较高的方法学严谨性。紧随其后的是董焕晴等[24]的质性研究：该研究通过对24名大语言模型用户的深入访谈构建了信息检索行为的理论模型，与用户提问亦有密切关联。系统不仅展示了所有证据的研究设计和证据描述，还允许用户进一步查看论证要素、期刊质量指标等细节信息。

系统的另一创新之处体现在将具备链式推理能力的大语言模型引入社会科学的证据综合处理。系统运用DeepSeek-R1模型作为推理引擎，该模型凭借强化学习训练增强了链式推理能力，适用于本研究中串联多证据来开展综合推理的需求。在证据综合阶段，以提示语指示模型关注不同证据之间的逻辑关联，同时考虑研究设计局限性对结论的影响，构建一个能够贯穿多研究的证据链，从而对用户提问给出具有针对性的回复。系统会依次展示思考过程和思考结论。其中：思考过程完整地保留DeepSeek-R1的推理过程，这有助于用户跟踪AI处理证据的内部思路，以增强系统的可解释性；思考结论则会给出证据综合的最终结果。

为综合评价ToulminQA循证问答原型系统的应用价值，邀请了4位图书情报领域专家（含2位阅读推广专家和2位信息行为专家）参与测评，4位专家均具备博士学位、中级及以上职称。测评以独立计分形式开展，每位专家向系统提出5个以上与其研究领域相关的问题，其后从相关性（所生成的答案是否能清晰回应所提问题）、可靠性（所生成的答案是否有可信的依据）、逻辑性（所生成答案的论证逻辑是否清晰合理）、全面性（所生成的答案是否能充分覆盖问题相关的多个视角）以及实用性（所生成的答案是否能为实际研究或决策提供帮助）5个方面进行评分，评分采用李克特5级量表（1=非常不同意，5=非常同意）。

评测结果显示，系统在所有维度的平均得分均在4分以上。其中，相关性4.25分，可靠性4.50分，逻辑性4.25分，全面性4.25分，实用性4.75分。结果表明专家在整体上对系统在科研辅助方面的综合能力给予认可。其中，实用性维度分值最高，说明该系统能够为研究工作提供高效的辅助。在开放式反馈中，专家认为系统所呈现的证据内容不仅覆盖面广，还具备多元分析视角，能够显著节约人力投入；同时，系统在推理过程上保持了严谨性，所得结论较为可靠。整体来看，ToulminQA能够对多来源证据进行有机整合，为用户提供智能化的科研辅助服务，达到了既定目标。

5 总结与展望

为提升社会科学领域循证研究效率及智能化水平，本研究提出并探索了一种将图尔敏模型和大语言模型技术相融合的社会科学循证知识库构建方法，同时选取图书情报学作为实证应用领域开展分析。通过开发和评测ToulminQA原型系统，进一步验证了该方法在自动抽取论证要素和整合多来源证据进而提供智能循证服务等方面具备有效性。

本研究的创新主要体现在3个方面：①提出并实践了将图尔敏模型与生成式AI相结合支撑社会科学循证研究的技术路径；②设计了整合研究设计信息、研究质量信息及图尔敏论证要素的多维度数据模型；③设计了面向社会科学循证研究的智能问答原型系统，将具有链式推理能力的大语言模型引入循证研究，实现先思考后回答，使结果具备良好的可解释性和可追溯性。

有必要指出的是，本研究所提出的方法是经典循证研究方法的一种有益补充，而非替代方案。系统评价以及元分析等方法凭借其显著的严谨性，依然是确保研究结论可靠性的核心手段。本方法的独特应用价值在于，作为系统评价前的有效探索工具，或在时间、资源有限的情况下，帮助研究者高效地综合特定领域的核心证据，从而在现有循证研究模式之外提供一条更为便捷的路径。

本研究在以下方面还存在一定局限性：①实证分析仅聚焦于图书情报学科的两个具体领域，因此所提方法在更多社会科学领域中的普适性仍须进一步验证；②知识库的证据主要来源于学术期刊论文，暂未覆盖其他文献类型。展望未来，该方法框架可扩展到更广泛的社会科学领域，还可纳入学位论文、会议文献、研究报告等其他类型的学术文献。此外，进一步的研究方向包括：将循证社会科学中的系统评价和元分析方法与大语言模型深度结合；借助多个大语言模型实现结论的交叉验证；与现有的循证社会科学平台整合，构建智能化的循证生态系统。

参考文献

[1] COCHRANE A L. Effectiveness and efficiency: random reflections on health services[M]. London: Nuffield Provincial Hospitals Trust, 1972: 22-25.

[2] GUYATT G, CAIRNS J, CHURCHILL D, et al. Evidence-based medicine: a new approach to teaching the practice of medicine[J]. JAMA, 1992, 268 (17): 2420-2425.

[3] 拜争刚. 循证社会科学[M]. 上海：华东理工大学出版社，2019：11-18.

[4] 王思丽，祝忠明，李慧佳. 国外循证社会科学证据知识库的证据类型与标准探析及其启示[J]. 现代情报，2023，43（2）：105-116.

[5] TOULMIN S E. The uses of argument[M]. Updated ed. Cambridge: Cambridge University Press, 2003.

[6] 拜争刚，赵坤，刘丹，等. 循证社会科学的推动者：Campbell协作网[J]. 中国循证医学杂志，2018，18（12）：1380-1385.

[7] Campbell Collaboration. Evidence[EB/OL]. [2025-09-02]. https://www.campbellcollaboration.org/evidence.

[8] LORY C, GREGORI E. Comparison of the What Works Clearinghouse standards for single-case research: applications for systematic reviews[J]. Behavioral Disorders, 2024, 50 (1): 34-45.

[9] What Works Clearinghouse. Procedures and standards handbook[EB/OL]. [2025-09-02]. https://ies.ed.gov/ncee/wwc/

docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf.

[10] Pew Charitable Trusts. Results First Clearinghouse Database[EB/OL]. [2025-05-25]. https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2015/results-first-clearinghouse-database.

[11] OSBORNE J，ERDURAN S，SIMON S. Enhancing the quality of argumentation in school science[J]. Journal of Research in Science Teaching，2004，41（10）：994-1020.

[12] ALISON L，SMITH M D，EASTMAN O，et al. Toulmin's philosophy of argument and its relevance to offender profiling[J]. Psychology，Crime & Law，2003，9（2）：173-183.

[13] SIMOSI M. Using Toulmin's framework for the analysis of everyday argumentation：some methodological considerations[J]. Argumentation，2003，17（2）：185-202.

[14] VITALI F，PERONI S. The argument model ontology[EB/OL]. [2025-05-15]. http://www.essepuntato.it/2011/02/argumentmodel.

[15] SONG N Y，CHENG H H，ZHOU H M，et al. Linking scholarly contents：the design and construction of an argumentation graph[J]. Knowledge Organization，2022，49（4）：213-235.

[16] YANG J C，ZHENG M，LIU Y L. Fusion weighted features and BiLSTM-attention model for argument mining of EFL writing[J]. Frontiers in Psychology，2023，14：1049266.

[17] STAB C，DAXENBERGER J，STAHLHUT C，et al. ArgumenText：searching for arguments in heterogeneous sources[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Demonstrations. Stroudsburg：ACL，2018：21-25.

[18] Elsevier. Learn about the Scopus AI difference[EB/OL]. [2025-06-26]. https://blog.scopus.com/posts/learn-about-the-scopus-ai-difference.

[19] 中国知网. 知网研学AI[EB/OL]. [2025-05-28]. https://xai.cnki.net/welcome.

[20] WANG Z F，CAO L，DANEK B，et al. Accelerating clinical evidence synthesis with large language models[EB/OL]. [2025-05-26]. https://arxiv.org/abs/2406.17755.

[21] 褚乐阳，刘泽民，王浩，等. 大模型支持的教师循证实践：行动框架与案例应用[J]. 开放教育研究，2024，30（4）：91-103.

[22] MALKOV Y A，YASHUNIN D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2020，42（4）：824-836.

[23] 王若佳，范科鸣，刘智锋，等. 生成式人工智能环境下用户信息检索式行为研究[J]. 数据分析与知识发现，2024，8（Z1）：20-30.

[24] 董焕晴，曹高辉，庹沛. 生成式AI工具用户信息搜寻行为形成路径：以数字素养与人工智能素养为视角[J]. 图书馆论坛，2025，45（10）：152-160.

作者简介

齐涵悦，女，硕士研究生，研究方向：知识组织。
杨颜僖，女，硕士研究生，研究方向：知识组织。
林泽斐，男，博士，副教授，通信作者，研究方向：知识组织、数字人文，E-mail：linzf@fjnu.edu.cn。

Evidence-Based Knowledge Base Construction Method for Social Sciences Integrating Toulmin Model and LLM

QI HanYue[1] YANG YanXi[2,3] LIN ZeFei[1]
(1. College of Sociology and History, Fujian Normal University, Fuzhou 350117, P. R. China; 2. National Science Library, Chinese Academy of Sciences, Beijing 100049, P. R. China; 3. Department of Information Resources Management, University of Chinese Academy of Sciences, Beijing 100049, P. R. China)

Abstract: Evidence-based research in the social sciences faces dual challenges: the rapid growth of literature and complex argumentation scenarios, necessitating intelligent and efficient evidence processing methods. This study proposes a method for constructing a social sciences evidence-based knowledge base that integrates the Toulmin argumentation model with large language model (LLM), aiming to improve the efficiency of processing social sciences evidence and the level of intelligence in evidence analysis. First, we design an information extraction process centered on argumentation. Second, we construct a multidimensional knowledge organization model for semantic retrieval. Finally, we design an intelligent evidence-based service process based on chain reasoning, and take library and information science as an example to develop the prototype system ToulminQA. Experimental evaluations demonstrate that ToulminQA exhibits high accuracy and reliability in argumentative element extraction. The system effectively integrates extensive evidence to generate logically coherent, practically valuable evidence-based conclusions. This research provides a novel reference and practical example for evidence-based research pathways in the social sciences.

Keywords: Toulmin Model; Large Language Model; Social Sciences; Knowledge Base Construction; Evidence-Based Social Sciences; Evidence; Knowledge Organization

（责任编辑：王玮）