

基于思维链推理的 MAUDE 数据库自动化分析

华磊¹ 巩洋² 毋丽丽¹ 胡国华¹ 贺国平¹ 刘晋媛¹

(¹忻州师范学院 忻州 034000 ²美国得克萨斯大学休斯顿健康科学中心 休斯顿 77096)

〔摘要〕 **目的/意义** 针对医疗器械制造商与用户设施不良事件报告 (manufacturer and user facility device experience, MAUDE) 数据库分析中存在的技术复杂度高与效率低等问题, 设计并实现自动化分析框架, 降低数据库分析难度并加速研究产出。**方法/过程** 基于双链式结构设计 AutoQUEST 框架, 其具备研究问题自动生成、结构化 SQL 查询构建及数据分析能力。通过 5 组案例评估框架查询成功率、执行时间及报告质量。**结果/结论** 在最优案例中, 该框架查询成功率达 100%, 执行时间 158 秒, 综合质量评分 4.74 分 (满分 5 分), 既能实现 MAUDE 数据库高效自动化分析, 又能保持高质量的分析结果, 可为医疗机构、医疗器械制造商及监管机构提供便捷可靠的数据分析支持。

〔关键词〕 医疗器械不良事件; MAUDE 数据库; 大语言模型; 思维链推理

〔中图分类号〕 R-058 **〔文献标识码〕** A **〔DOI〕** 10.3969/j.issn.1673-6036.2025.08.014

Automated Analysis of MAUDE Database Based on Chain-of-Thought Reasoning

HUA Lei¹, GONG Yang², WU Lili¹, HU Guohua¹, HE Guoping¹, LIU Jinyuan¹

¹Xinzhou Normal University, Xinzhou 034000, China; ²University of Texas Health Science Center at Houston, Houston 77096, USA

〔Abstract〕 **Purpose/Significance** In view of the problems of high technical complexity and low efficiency existing in the database analysis of manufacturer and user facility device experience (MAUDE) of medical devices, to design and implement an automated analysis framework, so as to reduce the difficulty of database analysis and accelerate research output. **Method/Process** The AutoQUEST framework is designed based on a dual-chain, which has the capabilities of automatic generation of research questions, structured SQL query formulation, and data analysis. Five cases are used to evaluate the framework in terms of query success rates, execution duration, and report quality. **Result/Conclusion** In the best cases, the framework achieves a 100% query success rate, an execution time of 158 seconds, and an overall quality score of 4.74 out of 5. It can not only achieve efficient and automated analysis of the MAUDE database but also maintain high-quality analysis results, providing convenient and reliable data analysis support for medical institutions, medical device manufacturers, and regulatory authorities.

〔Keywords〕 medical device adverse event; MAUDE database; large language model; chain-of-thought reasoning

〔修回日期〕 2025-06-03

〔作者简介〕 华磊, 博士, 讲师, 发表论文 20 余篇。

〔基金项目〕 山西省留学回国人员科技活动择优资助项目 (项目编号: 20230039)。

1 引言

随着现代医疗技术迅速发展, 医疗器械在疾病诊断、治疗、康复等领域的作用日益重要。但不良事件时有发生, 给患者安全和医疗质量带来潜在威胁。构建科学有效的监测和研究体系, 以提升医疗器械的安全与有效性, 是行业监管的重要议题。

本研究基于大语言模型 (large language model, LLM), 结合思维链 (chain-of-thought, CoT) 推理技术在医疗器械制造商与用户设施不良事件报告 (manufacturer and user facility device experience, MAUDE) 数据库分析中的应用, 设计开发 AutoQUEST 数据库自动化智能分析工具, 并对其性能进行评估。结果表明, 该工具可以精准筛选数据, 并进行自动分析统计, 快速生成高质量结果, 有效降低 MAUDE 数据研究的技术门槛, 为医疗器械安全研究提供了自动化分析路径。

2 MAUDE 数据库

2.1 研究意义

美国食品药品监督管理局 (Food and Drug Administration, FDA) 于 1991 年推出的 MAUDE 系统是目前全球范围内使用最广泛的医疗器械不良事件监测平台之一。该系统积累了海量医疗器械安全性数据, 几乎完整覆盖医疗器械风险监测生命周期, 并对敏感信息进行脱敏处理, 可为监管机构、医疗器械制造商及医疗机构提供公开透明的信息支持^[1]。研究人员可以利用 MAUDE 数据库精准定位医疗器械的故障模式、发生频率及其成因, 为产品改进和监管部门管理提供科学依据。同时, 该系统的数据库反馈机制打通了医疗机构、监管部门和供应商等之间的信息壁垒, 实现了对潜在安全隐患早发现、早干预的闭环管理。因此, 利用 MAUDE 数据库进行深入研究对于提高医疗器械监管效率和保障患者安全具有极其重要的意义。

2.2 技术挑战

科研人员在使用 MAUDE 数据库时面临众多技

术挑战。该数据库包含多个紧密关联的数据表和大量非结构化文本数据, 导致数据预处理与信息提取流程复杂而耗时, 精准定位关键信息难度高、效率低, 且易引起人为疏漏, 可能造成重要安全信号遗漏, 直接给患者安全带来负面影响。此外, 随着研究的深入, 分析需求不断变化, 研究人员须频繁调整查询策略和数据处理流程。过度依赖人工操作会耗费大量时间, 增加人为误差风险, 从而影响研究结果的准确性与可重复性。因此, 实现分析过程的计算机辅助不仅能显著提升效率, 更是确保分析精准性和患者安全的关键。

FDA 官方提供在线查询页面^[2]与应用程序编程接口 (application programming interface, API) 服务^[3]两种数据访问途径, 研究者可通过关键词检索和编程实现批量调用与处理。但前者功能单一, 后者依赖编程实现且受限于调用频次及系统功能, 难以兼顾易用性与灵活性。近年来, 学界虽在不良事件文本分类模型^[4]、自动结构化查询语句 (structured query language, SQL) 生成^[5]、数据集成框架^[6]及交互式分析报告方面取得一定成果^[7], 但仍缺乏贯穿数据采集分析到结果呈现的端到端自动化解决方案。

2.3 研究目标

本研究旨在提出一种基于 LLM 与 CoT 推理^[8]的 AutoQUEST 分析框架, 构建具备自纠错能力的 SQL 生成机制 (支持跨表复杂查询), 并通过提示词优化技术, 即词元 (token) 的压缩与少样本学习, 提升文本解析能力。研究目标包括: 构建端到端的自动化研究分析框架, 实现研究流程的高度自动化; 应用 CoT 推理与提示词优化方法, 提高复杂任务的自动完成能力; 通过多维度测试对框架性能和结果质量进行系统评估, 以验证其可靠性与有效性。

3 AutoQUEST 框架及其评估方法

3.1 框架介绍

AutoQUEST 框架采取“认知-执行”双链结构, 模拟人类活动中“先思考、后行动”的闭环逻辑, 见图 1。采用 Gemini 2.0 Flash Thinking 为基础

模型，因其具备 1.8 % 的低“幻觉”率（DeepSeek R1 对应数值为 14.8 %）^[9]、每秒 200 token 以上的输出速度和可免费调用的 API。框架遵循“提示词优化 + 双链结构”的轻量化思路：每次调用前，将 MAUDE 数据库的表结构、示例数据和专业术语等关键信息按任务指令的要求进行筛选，确保提示词上下文内容完整、精确。解耦式设计使框架无须修改代码即可即时继承模型升级带来的性能红利。

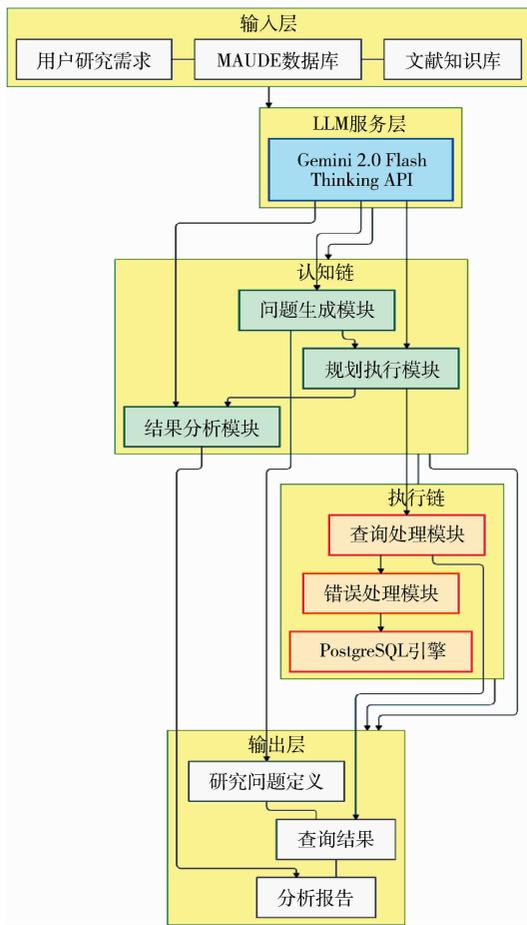


图 1 框架流程及“认知-执行”双链式结构

认知链是以“想”的模式提供服务，依次完成问题生成、规划执行和结果分析：先将用户需求转换为可映射至 MAUDE 数据库的具体研究问题，再自动制定分析策略，最终生成分析结果和报告。执行链是以“做”的模式提供服务，将认知链的规划转换为 SQL，完成数据查询分析和结果格式化。整个处理通过输入层、LLM 服务层和输出层的结构和顺序完成；统一的数据接口与实时监控保障信息流畅和系统稳定。用户提出问题后，即可启动从数据

抽取到报告生成的全自动处理流程。

3.2 核心功能

AutoQUEST 框架的核心是基于提示词优化的链式思维推理。先自动整合用户需求与数据库元信息，将研究问题细化和分解为可执行的子任务。再结合数据库结构和样例数据对前述输出进行二次校验，以有效抑制“幻觉”和错误生成。SQL 查询采用渐进式生成策略，设定 60 秒数据库查询执行超时及最多 32 条查询结果返回的数量限制（达到或超过时自动生成统计类子查询），在控制链路上下文提示词长度的前提下，最大程度减少信息丢失。系统内置 3 重纠错机制：预验证、运行监控及指数退避重试（初始间隔 1 秒，递增因子为 2，最多重试 5 次）。最终查询结果均标准化为 JSON 格式，实现面向研究问题的多维统计与语义分析，输出包括趋势、成因及局限性的综合报告。

3.3 运行环境及使用方法

AutoQUEST 框架支持在 Linux、MacOS 或 Windows 操作系统的 Python 环境下运行，默认以 PostgreSQL 数据库加载 MAUDE 数据。仅需双核 CPU、16 GB 内存和 250 GB 硬盘空间即可完成部署。

使用方法简单，具体步骤：一是在配置文件中填写数据库连接信息、推理模型 API 地址与密钥、输出路径等关键参数；二是使用命令行模式执行语句（python main.py -question “你的研究问题或偏好”）。程序通常会在 3 ~ 10 分钟自动完成全过程，输出数据报告。报告内容涵盖研究问题、执行计划、生成的 SQL 语句、查询次数、执行耗时、token 数量消耗、查询失败次数、最终查询结果与综合分析报告。

本研究相关完整源代码（含所有提示词）、程序运行数据及结果报告已发布于 GitHub（<https://github.com/leiMizzou/AutoQUEST>），便于研究结果复现与进行二次开发^[10]。

3.4 评估框架及标准

围绕技术执行和内容质量两个维度对 AutoQUEST 框架进行评估，见表 1。其中，系统运行性

能通过查询成功率、查询数量、重试次数以及执行时间来衡量，见表 2；分析复杂度评分标准，见表 3；内容质量评分标准，见表 4。

表 1 综合评分权重分布

| 一级维度 | 二级维度 | 三级维度 |
|--------------|--------------|--------------|
| 技术执行维度 (40%) | 系统运行性能 (25%) | 查询成功率 (7.5%) |
| | | 查询数量 (6.25%) |
| | | 重试次数 (6.25%) |
| | | 执行时间 (5.0%) |
| | | 分析复杂度 (15%) |
| 内容质量维度 (60%) | 内容质量 (60%) | 问题定义 (6.0%) |
| | | 数据关联 (4.5%) |
| | | 分析方法 (4.5%) |
| | | 问题阐述 (12.0%) |
| | | 分析框架 (15.0%) |
| | | 数据展示 (12.0%) |
| | | 结论深度 (12.0%) |
| | | 实用价值 (9.0%) |

表 2 系统运行性能评分标准

| 分数 | 查询成功率 (%) | 查询数量 (个) | 重试次数 (次) | 执行时间 (秒) |
|----|-----------|------------------|----------|----------|
| 5 | 100 | 最优 (8~12) | <5 | <200 |
| 4 | 90~99 | 适中 (6~7 或 13~15) | 5~10 | 200~400 |
| 3 | 80~89 | 稍弱 (4~5 或 16~18) | 11~15 | 400~600 |
| 2 | 70~79 | 较弱 (2~3 或 19~20) | 16~20 | 600~800 |
| 1 | <70 | 过多或过少 (<2 或 >20) | >20 | >800 |

表 3 分析复杂度评分标准

| 分数 | 问题定义 | 数据关联 | 分析方法 |
|----|------|----------|--------|
| 5 | 多维度 | >3 个表关联 | 复杂统计分析 |
| 4 | 双维度 | 2~3 个表关联 | 多维度分析 |
| 3 | 单维度 | 2 个表关联 | 基础统计分析 |
| 2 | 一般描述 | 单表分析 | 简单统计 |
| 1 | 简单描述 | 无关联分析 | 基础计数 |

表 4 内容质量评分标准

| 分数 | 问题阐述 | 分析框架 | 数据展示 | 结论深度 | 实用价值 |
|----|------------------|----------------|-----------------|----------------|-------------|
| 5 | 问题定义清晰，研究目标和边界明确 | 框架完整，层次分明，逻辑严密 | 数据全面，可视化专业，解释充分 | 见解独到，论证充分，启发性强 | 直接可用，建议具体可行 |
| 4 | 问题定义基本清晰，目标较明确 | 框架较完整，逻辑较清晰 | 数据较全面，展示合理 | 见解较好，论证较充分 | 较易应用，建议具体 |
| 3 | 问题描述一般，目标不够具体 | 框架基本完整，逻辑一般 | 数据基本完整，展示一般 | 见解一般，论证基本可信 | 可以参考，建议一般 |
| 2 | 问题描述模糊，目标不明确 | 框架不完整，逻辑欠缺 | 数据不足，展示简单 | 见解浅显，论证不足 | 应用困难，建议模糊 |
| 1 | 问题描述混乱或缺失 | 框架混乱或缺失 | 数据混乱或缺失 | 见解混乱或缺失 | 难以应用或无建议 |

3.5 评估流程

以“请提出一个与医疗器械相关的研究问题”为示例需求，由 AutoQUEST 系统自动生成 5 组待评估数据（分别记为报告 1—5）。系统在每次生成研究问题后，自动结合 MAUDE 数据库的表结构和案例数据完成研究问题拆解、SQL 查询及统计分析等操作，最终产出对应的分析报告。在此过程中，系统自动记录每个报告对应的执行时间、模型 token 消耗总量、SQL 语句及其数量、查询成功率、重试次数等指标。在主观评分环节，邀请两位计算机专

家，依据内容质量评分标准，对报告进行评审。如果两位专家意见有差异，则引入团队讨论协商达成一致。最后，基于加权评分法，计算综合得分，并对 5 份报告进行整体排名。

4 实验结果

4.1 系统性能详细评估

系统性能详细评估，见表 5。报告 3 和报告 4 的查询成功率均达到 100%，耗时分别为 158 秒和 290 秒，表现出较高执行效率和稳定性。报告 1 结

果相对良好。报告 2 重试次数高达 23 次，表明在处理查询时存在大量错误重试。报告 5 由于总查询数

过多，尽管查询成功率达到 100%，但执行时间显著增加。

表 5 系统性能详细评估

| 报告 | 执行时间 (秒) | token 数 (千) | 总查询数 (个) | 成功查询 (次) | 失败查询 (次) | 重试次数 (次) | 查询成功率 (%) |
|----|----------|-------------|----------|----------|----------|----------|-----------|
| 1 | 178 | 173 | 9 | 8 | 1 | 5 | 89 |
| 2 | 669 | 341 | 10 | 8 | 2 | 23 | 80 |
| 3 | 158 | 120 | 7 | 7 | 0 | 4 | 100 |
| 4 | 290 | 141 | 7 | 7 | 0 | 5 | 100 |
| 5 | 877 | 140 | 32 | 32 | 0 | 2 | 100 |

4.2 系统运行性能评分结果

基于系统运行性能评分标准，结合执行时间及查询成功率，计算 5 份报告的 SQL 执行性能评分，见表 6。综合来看，报告 3 的表现最佳，报告 2 在执行时间和重试次数上失分较多。

表 6 系统运行性能评分 (分)

| 报告 | 查询成功率 | 查询数量 | 重试次数 | 执行时间 | 加权总分 |
|----|-------|------|------|------|------|
| 1 | 3.0 | 5.0 | 4.0 | 5.0 | 4.15 |
| 2 | 3.0 | 5.0 | 1.0 | 2.0 | 2.85 |
| 3 | 5.0 | 4.0 | 5.0 | 5.0 | 4.75 |
| 4 | 5.0 | 4.0 | 4.0 | 4.0 | 4.30 |
| 5 | 5.0 | 1.0 | 5.0 | 1.0 | 3.10 |

4.3 分析复杂度评分结果

5 份报告的分析复杂度评分，见表 7。报告 3 在问题定义、数据关联、分析方法 3 个方面均较为突出。

表 7 分析复杂度评分 (分)

| 报告 | 问题定义 | 数据关联 | 分析方法 | 加权总分 |
|----|------|------|------|------|
| 1 | 4.0 | 4.0 | 4.5 | 4.15 |
| 2 | 3.5 | 3.5 | 4.0 | 3.65 |
| 3 | 4.5 | 4.5 | 4.5 | 4.50 |
| 4 | 4.0 | 4.0 | 4.0 | 4.00 |
| 5 | 3.0 | 3.5 | 3.0 | 3.15 |

4.4 内容质量评分结果

5 份报告的内容质量评分，见表 8，报告 3 和报告 1 得分较高。

表 8 内容质量综合评分 (分)

| 报告 | 问题阐述 | 分析框架 | 数据展示 | 结论深度 | 实用价值 | 加权总分 |
|----|------|------|------|------|------|------|
| 1 | 5.0 | 5.0 | 4.5 | 4.5 | 4.5 | 4.75 |
| 2 | 4.0 | 4.0 | 4.0 | 3.5 | 4.0 | 3.90 |
| 3 | 5.0 | 5.0 | 4.5 | 4.5 | 5.0 | 4.80 |
| 4 | 4.5 | 4.5 | 4.5 | 4.5 | 4.0 | 4.45 |
| 5 | 4.0 | 4.0 | 4.0 | 3.5 | 3.5 | 3.85 |

4.5 综合评分与排名

5 份报告综合得分，见表 9。其中，报告 3 综合得分最高，为 4.74。

表 9 综合评分与排名

| 报告 | 技术维度 (分) | 内容维度 (分) | 综合总分 (分) | 排名 |
|----|----------|----------|----------|----|
| 1 | 1.67 | 2.85 | 4.52 | 2 |
| 2 | 1.27 | 2.34 | 3.61 | 4 |
| 3 | 1.86 | 2.88 | 4.74 | 1 |
| 4 | 1.68 | 2.67 | 4.35 | 3 |
| 5 | 1.26 | 2.31 | 3.57 | 5 |

5 讨论

5.1 性能表现

AutoQUEST 能够持续生成高质量分析结果并在大多数实验场景中保持一致，证明其基础设计的稳

健性,表明该系统不仅是一个概念验证,更是一个切实可行的工具,能够真正增强人类研究能力,特别是在减少数据分析的人工工作量的同时,确保了分析的严谨性。AutoQUEST 采用的自动化流程显著简化了多次迭代查询与分析的操作过程,不仅大幅提升了研究效率,还能提供高质量分析结果。以报告 1、报告 3、报告 4 为例,其在问题阐述、分析框架、数据展示等关键维度均获得较高评分,充分体现了 AutoQUEST 在自动化效率与结果质量方面的双重优势,为研究者提供了便捷而可靠的研究辅助工具。然而,系统在处理复杂查询需求时仍有改进空间。报告 2 和报告 5 的测试结果显示,面对大规模数据或复杂分析任务时,系统出现了执行时间较长和重试次数增多的问题,这表明 AutoQUEST 在性能方面有待进一步提升。

5.2 优势与特点

AutoQUEST 开创了医疗器械不良事件研究领域智能分析新范式,创新性地将 LLM 与 CoT 推理技术结合,实现了从问题生成到数据分析的全流程自动化。实验结果表明,在 CoT 推理技术加持下,AutoQUEST 不仅能够模拟专家思维,根据用户研究意图与数据特点自动构建并执行高质量查询,还能生成在问题阐述、分析框架、结论深度等方面均较出色的研究报告,确保每份报告具有逻辑闭环。

该系统的核心优势在于其模块化的双链式结构,通过认知链完成问题分析、执行策略优化等关键环节,再由执行链负责数据检索、分析及结果呈现。这一系列操作形成了端到端的智能化闭环,充分体现了 CoT 将复杂任务分解为更易管理的子任务的内在逻辑,为输出结果连贯性和上下文关联性提供了关键保障。

AutoQUEST 系统配备独特的提示词压缩技术,通过聚类和 LLM 对数据表进行语义相似度计算,将 MAUDE 原始 113 张数据表合并为 13 ~ 16 个表组,再按需还原使用。该方法显著简化了数据库模式的信息表达,提示词长度减少了 85% ~ 88%^[11]。既满足了模型输入长度方面的限制(通常为 32K ~ 128K),又显著提高了 AutoQUEST 在研究问题生成

阶段的创造性与准确性。双链架构与提示词压缩、少样本学习的结合,形成了强大的协同效应,使 AutoQUEST 能够超越简单的自动化,实现智能、专家级辅助。

5.3 研究局限

尽管 AutoQUEST 在实验中综合表现灵活出色,但在复杂任务及大规模数据测试场景中,仍存在 3 方面问题有待优化。首先,AutoQUEST 对数据规范性和完整度要求较高。一旦出现文本标准化缺失或数据不完整,可能对分析可靠性产生不利影响。后续可通过引入数据清洗、缺失值补全等预处理技术,增强系统鲁棒性。其次,AutoQUEST 对 MAUDE 数据库中的海量非结构化文本信息主要采用规则匹配及少样本学习算法,在深层语义理解和自动知识抽取方面仍存在明显局限。将自然语言处理、机器学习等前沿技术融入分析过程,是未来拓展系统功能的重要方向。最后,从实验结果来看,报告 2 和报告 5 执行时间明显增长、重试次数增多,说明系统在面对多表复杂连接或大规模数据时存在性能瓶颈。为实现跨领域、更大规模的通用应用,应在算法优化、资源弹性调度、查询缓存和索引机制等方面进行深入研究。

6 结语

AutoQUEST 框架通过 LLM 结合 CoT 的“认知-执行”双链,实现了 MAUDE 数据库从问题生成到结果报告的全流程自动化分析,不仅大幅提升了分析效率,还保证了高质量的结果输出,降低了 MAUDE 数据研究的技术门槛。尽管在深层语义理解与大规模数据处理方面仍存在改进空间,但实验结果已充分验证了系统应用于监管、研发与临床实践等领域的现实意义。随着因果推理、多模态输入及可视化交互的进一步融合,以及基础模型的持续进步,AutoQUEST 有望成为医疗器械安全研究领域的重要基础设施,为临床、产业和监管部门持续赋能。

(下转第 102 页)

- [J]. 医学信息, 2018, 31 (12): 23 - 25.
- 13 罗晓兰, 李明. 国外医学图书馆员的角色定位与职业能力需求 [J]. 中华医学图书情报杂志, 2018, 27 (1): 62 - 67.
 - 14 黄成. 古籍文献资源整理保护与开发利用简论 [J]. 兰台世界, 2016 (12): 75 - 77.
 - 15 海佳丽, 汪润, 袁良志, 等. 基于检索增强的中医药标准知识问答系统构建探索与实践 [J/OL]. 数据分析与知识发现, 1 - 13 [2025 - 02 - 18]. <http://kns.cnki.net/kcms/detail/10.1478.G2.20250120.1024.004.html>.
 - 16 曾刚. 数字人文图像资源语义组织研究 [D]. 武汉: 武汉大学, 2021.
 - 17 RAMEZANI M, TAKIAN A, BAKHTIARI A, et al. Research agenda for using artificial intelligence in health governance: interpretive scoping review and framework [J]. *BioData mining*, 2023, 16 (1): 31.
 - 18 JUNGWIRTH D, HALUZA D. Artificial intelligence and public health: an exploratory study [J]. *International journal of environmental research and public health*, 2023, 20 (5): 4541.
 - 19 陈淦涛. AI 时代人工智能在健康教育领域的应用探索与思考 [C]. 广州: 广州市第十四届健康教育与健康促进学术交流会议, 2024.
 - 20 郑琰莉, 韩福海, 李舒玉, 等. 人工智能大模型在医疗领域的应用现状与前景展望 [J]. 医学信息学杂志, 2024, 45 (6): 24 - 29.
 - 21 刘北平. 健康医疗大数据的发展现状与应用模式研究 [J]. 中国高科技, 2022 (6): 8 - 9.
 - 22 付晓燕. 大数据视角下医疗档案信息交互共享平台构建研究 [J]. 办公室业务, 2024 (13): 62 - 64.
 - 23 郑荣, 雷亚欣, 张默涵, 等. 基于联盟区块链的多源个人健康信息协同共享模式研究 [J]. 图书情报工作, 2023, 67 (20): 79 - 92.
 - 24 齐娜, 宋立荣. 医疗健康领域微博信息传播中的信息质量问题 [J]. 科技导报, 2012, 30 (17): 60 - 65.
 - 25 毛丽, 张文一, 马德好, 等. 某三甲医院基于 AI 的病历质量管理体系应用 [J]. 中国病案, 2022, 23 (11): 26 - 29.
 - 26 张瑜, 徐可, 陈俊. 基于大数据与人工智能的全程病历质控系统构建 [J]. 产业与科技论坛, 2022, 21 (24): 46 - 48.

(上接第 90 页)

作者贡献: 华磊负责研究设计与实施、系统开发、论文撰写; 巩洋负责提供指导、论文审核; 毋丽丽负责组织主观评分及结果分析; 胡国华、贺国平负责主观评分; 刘晋媛负责评估数据收集与分析。

利益声明: 所有作者均声明不存在利益冲突。

参考文献

- 1 吴静, 金乐. 美国医疗器械不良事件监测体系及 MAUDE 数据库在医疗器械注册申报中的应用 [J]. 中国医疗器械杂志, 2024, 48 (4): 451 - 456.
- 2 U. S. Food and Drug Administration. Manufacturer and user facility device experience (MAUDE) database [EB/OL]. [2025 - 02 - 22]. <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm>.
- 3 U. S. Food and Drug Administration. OpenFDA: open - source APIs [EB/OL]. [2025 - 02 - 22]. <https://open.fda.gov/>.
- 4 WANG E, KANG H, GONG Y. Generating a health information technology event database from FDA MAUDE reports [J]. *Studies in health technology and informatics*, 2019, 264 (8): 883 - 887.
- 5 YU Y, SHI Y, FENG Y, et al. Developing a generative AI - powered chatbot for analyzing MAUDE database [J]. *Studies in health technology and informatics*, 2024, 316 (8): 1255 - 1259.
- 6 SHI Y, YU Y, FENG Y, et al. A data pipeline for enhancing quality of MAUDE - based studies [J]. *Studies in health technology and informatics*, 2024, 316 (8): 1214 - 1218.
- 7 YU Y, SHI Y, FENG Y, et al. Enhancing MAUDE database utility by GPT - 4 and cause - effect visualization [J]. *Studies in health technology and informatics*, 2024, 315 (7): 290 - 294.
- 8 WEI J, WANG X, SCHUURMANS D, et al. Chain - of - thought prompting elicits reasoning in large language models [J]. *Advances in neural information processing systems*, 2022, 35 (11): 24824 - 24837.
- 9 VECTARA. Hallucination Leaderboard [EB/OL]. [2025 - 05 - 19]. <https://github.com/vectara/hallucination-leaderboard>.
- 10 HUA L. AutoQUEST [EB/OL]. [2025 - 02 - 22]. <https://github.com/leiMizzou/AutoQUEST>.
- 11 HUA L. MAUDE - schema - compressor [EB/OL]. [2025 - 01 - 09]. <https://github.com/leiMizzou/MAUDE-Schema-Compressor>.