基于聚类算法与 DeepSeek 协同的数据库结构降维研究

华磊[1]，巩洋[2]

1. 忻州师范学院，忻州市，034000
2. 德州大学休斯顿医学研究中心，休斯顿，77096

【摘要】目的/意义 医疗器械不良事件数据库（Manufacturer and User Facility Device Experience, MAUDE）存在模式复杂与命名格式不一致问题，严重阻碍研究人员对数据的有效获取与利用。本研究开发并评估一种混合聚类算法和以 DeepSeek 为代表的大语言模型（Large Language Model, LLM）的数据库表结构降维框架，实现美国食品药品监督管理局（Food and Drug Administration, FDA）的 MAUDE 数据库中相似表结构的匹配合并以降低数据分析难度。**方法/过程** 本研究共进行了 96 组实验，使用聚类算法与基于 DeepSeek 语义相似性评估组合的方法，识别相似的库表结构进行合并得到去冗后的表结构特征描述。特征提取使用了词频-逆文档频率向量化和句子转换器嵌入两种方式，相似性评估选取了三组阈值（0.7、0.8、0.9）来测量其对降维性能的影响。最后基于专家标准分组结果，采用调整兰德指数（Adjusted Rand Index, ARI）、归一化互信息（Normalized Mutual Information, NMI）、精确度、召回率和 F1 分数等指标对自动化分组结果进行评估。**结果/结论** 融合聚类与语义相似性评估的混合方法将 F1 分数从 0.51-0.95 显著提升至 0.95-1.00，成功将 113 个数据库表结构降维至 13-16 组，实现了 87.78%的词元（token）压缩率（从 111 787 减至 13 664），并减少了约 80%的 DeepSeek 调用。该混合方法结合大语言模型的使用可大幅度提升医疗设备不良事件数据的分析效率，此框架应用价值广泛，适用于各行业复杂数据结构分析，提升数据分析效率。

【关键词】 MAUDE 数据库、结构降维、大语言模型、患者安全

Database Structure Dimensionality Reduction Research Based on Collaborative Clustering Algorithms and DeepSeek

HUA Lei[1]，GONG Yang[2]

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

[Abstract] Purpose/Significance The Manufacturer and User Facility Device Experience (MAUDE) database suffers from complex schema patterns and inconsistent naming formats, severely hindering researchers' effective data access and utilization. This study aims to develop and evaluate a hybrid framework combining clustering algorithms with Large Language Model (LLM) represented by DeepSeek for database structure dimensionality reduction, enabling matching and merging of similar structures in the Food and Drug Administration (FDA) MAUDE database to reduce data analysis complexity. **Method/Process** This research conducted 96 experimental groups using a combination of clustering algorithms and DeepSeek-based semantic similarity assessment methods to identify similar database table structures for merging and obtaining redundancy-reduced structural feature representations. Feature extraction employed Term Frequency-Inverse Document Frequency vectorization and sentence transformer embeddings, while similarity assessment utilized three threshold

基金项目：山西省留学回国人员科技活动择优资助项目（20230039）.

作者简介：华磊，leihua@xztu.edu.cn

groups (0.7, 0.8, 0.9) to measure their impact on dimensionality reduction performance. Finally, based on expert standard grouping results, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), precision, recall, and F1-score metrics were used to evaluate automated grouping results. **Result/Conclusion** The hybrid method integrating clustering and semantic similarity assessment significantly improved F1-scores from 0.51-0.95 to 0.95-1.00, successfully reducing 113 database structures to 13-16 groups, achieving an 87.78% token compression rate (from 111 787 to 13 664), and reducing approximately 80% of DeepSeek API calls. This hybrid approach combined with LLM can dramatically improve the analysis efficiency of medical device adverse event data. The framework has broad application value, suitable for complex data structure analysis across various industries, enhancing data analysis efficiency.
[Keywords] MAUDE database; structure dimensionality reduction; large language model; patient safety

1. 引言

MAUDE 是 FDA 用于收集和监测医疗器械不良事件报告的数据库系统[1]。FDA 要求医疗器械制造商、进口商和使用机构在发现其产品可能导致或已导致死亡、严重伤害，或可能再次引起严重后果的故障时，必须提交医疗器械报告（Medical Device Reporting, MDR）。因此深入分析 MAUDE 数据库对于医疗器械在上市后的安全性和有效性监测至关重要，不仅能够为监管决策提供关键支持，提升患者安全，同时为医疗器械的改进与研发提供了宝贵的资源。通过分析大量真实发生的不良事件数据，研究人员和制造商可以更深入地了解医疗器械在实际应用中的表现，及时发现设计缺陷和潜在的安全风险，指导设备的设计优化、生产工艺改进和新产品开发，推动医疗器械的整体创新。

然而，MAUDE 数据库存在模式复杂与命名及格式不一致的问题。在官方网站提供的 113 个可下载数据库压缩文件中既同时包含 TXT 和 CSV 两种存储文件类型，又存在数据字段分隔符混用制表符和空格的混合情况，甚至部分库表完全缺失字段名定义。其严重阻碍了研究人员，尤其是领域新进者和计算机技能有限人员对数据的有效获取与利用[2]。在缺乏对数据结构和语义关系深入理解的情况下，有效的数据分析与关键信息提取变得几乎不可实现，最终导致研究效率与成果质量的双重受损[3]。

值得注意的是，MAUDE 数据库中面临的挑战在私有数据分析领域也同样存在。许多医疗机构在处理历史遗留或孤岛整合后的数据时，仍需投入大量时间理解和解析不同数据库、表格与字段之间的隐含关系[4]，即使经验丰富的数据提取转换加载（Extract-Transform-Load, ETL）专业人员也难以避免。通过表结构特征描述的匹配合并，可以在保障结构完整与一致性前提下降低分析复杂度提升集成效率[5,6]。这类方法通过精准分组相似结构消除冗余描述，简化并加速不同数据源的对齐，为实现高效、可靠的数据分析奠定坚实基础[7]。

另一方面，LLM 的崛起极大降低了深度语义分析的技术门槛，为数据理解与应用创造了前所未有的可能性[8,9]。然而，现有 LLM 的上下文处理能力仍面临 token 长度限制的挑战——主流模型线上接口调用支持的最大输入长度多为 128k，网页应用端通常仅为 32k 至 64k。因此借助 LLM

对数据库表结构及案例数据进行整体分析时，必须在保持语义完整性的同时有效降低输入内容的长度。

本研究提出了一个创新性的数据库表结构降维框架，旨在高效理解与简化复杂数据库的表结构描述，同时解答以下三个关键研究问题：(1) 如何实现数据库表结构的自动高效识别与匹配？(2) 聚类算法与语义相似性评估的结合能在多大程度上提升模式匹配准确性并降低计算成本？(3) 这种降维方法能否有效满足大型语言模型的上下文窗口限制？该方法融合了聚类算法与语义相似性评估技术的优势，通过两阶段处理实现了表结构匹配的高效自动化与冗余特征消除。在第一阶段，系统采用 K 均值聚类（K-Means）[10]、层次聚类[11]和基于密度的空间聚类（Density-Based Spatial Clustering of Applications with Noise, DBSCAN）[12]算法对表结构进行初步快速分组；随后在第二阶段，利用预训练大型语言模型 DeepSeek V2.5[13]进行深度语义关系评估，进一步优化初始分组结果。实验结果表明该框架在维持较低计算开销的前提下，显著提高了模式分组的精确度，不仅简化并自动化了数据处理流程，还确保了后续分析缓解与大型语言模型等高级分析工具的兼容。这一技术突破为多领域复杂数据结构分析提供了新的可行路径，有望加速数据驱动决策的实现过程。

2. 相关技术背景

2.1 数据库模式匹配和映射

模式匹配作为异构数据库集成的基础环节，专注于识别不同数据库模式元素间的语义对应关系[14]。其匹配方法主要分为两个互补维度：元素级匹配与结构级匹配。元素级匹配通过评估名称、数据类型、约束和值分布等属性对单个模式元素进行比较；而结构级匹配则分析模式元素间的关系拓扑，包括层次架构和引用完整性约束。

为提升匹配精度，模式匹配常使用多维度的相似性评估策略，如通过语言相似性分析评估术语对应关系，用结构相似性量化关系模式一致性，以及基于数据库约束规则匹配识别表结构功能等价性，通过混合以上多维度策略及机器学习方法，模式匹配的性能够得到显著提升[5,6,15]。同时通过综合模式元数据和数据实例信息，派生出精确的匹配规则，可实现匹配过程的精细化和自动化，提升大规模数据环境中模式对齐的准确性和效率。

2.2 聚类和语义相似性增强数据集成

聚类算法是该降纬框架实现高效运行的重要基础，它简单来说就是把相似的数据元素或表格归类到一起，从而让后续的匹配和合并工作变得更加容易[16,17]。通过在匹配前先按照相似度将模式元素分组，聚类大大减少了需要比较的数据对数量，提高了处理效率。这一点在处理大型复杂数据时尤为重要，比如 COMA++[17]等工具就证明了聚类技术能有效地将复杂模式组织成易于管理的群组，从而优化整个匹配过程。

在聚类基础上，语义相似性度量进一步深化了数据集成能力，精确捕获数据元素的上下文含义[18,19]。借助自然语言处理和本体技术，系统能够深入理解模式标签和数据值的语义内涵。集成词汇和结构相似性的方法，结合 WordNet 等语义资源，使得精确计算模式元素间的语义距离成为可能。这种语义丰富化显著提升了匹配质量，通过考量术语超越其语法表面的深层含义与关系。近年来，深度学习的快速发展进一步优化了语义相似性计算，通过词嵌入和大参数预训练语言模型，在承载丰富语义信息的连续向量空间中表征词汇与短语。这些基于深度学习的先进技术能够精准捕获上下文细微差别，通过解析复杂语义关系，大幅提高模式匹配和实体解析的准确性。

聚类与语义相似性技术的有机整合为数据模式匹配提供了高效协同的方法论支持。聚类技术通过合理限制潜在匹配范围显著降低计算复杂度，而语义相似性技术则确保精准化匹配。这种优势互补的组合充分发挥了聚类应对大规模数据集的高效性和语义相似性提供深度上下文理解的精确性。两种技术的协同应用显著增强了模式匹配和实体解析的效率与准确度，促进了更为有效的数据整合与系统互操作性，为本文降维框的提出奠定了坚实基础。

3. 材料与方法

为了有效的分析与利用 MAUDE 数据库，本研究开发了语义聚类与合并框架（Semantic Clustering and Merging Framework, SCMF）进行数据库结构降维，如图 1 所示。这里的"降维"特指压缩数据库结构描述文本的长度，以适应大语言模型提示词的输入限制，而非对实际数据表或数据内容进行物理合并或标准化处理。该框架包含四个关键环节：(1)数据库结构特征和样例提取；(2)基于聚类算法的初步分组；(3)基于 DeepSeek 的语义相似性评估和组间优化；(4)与专家标准分组的对比评估。通过识别语义相似的数据结构并在描述层面进行归类整合，该方法能够显著减少提示词中的冗余信息，确保在保持语义完整性的前提下满足 LLM 的上下文窗口限制。

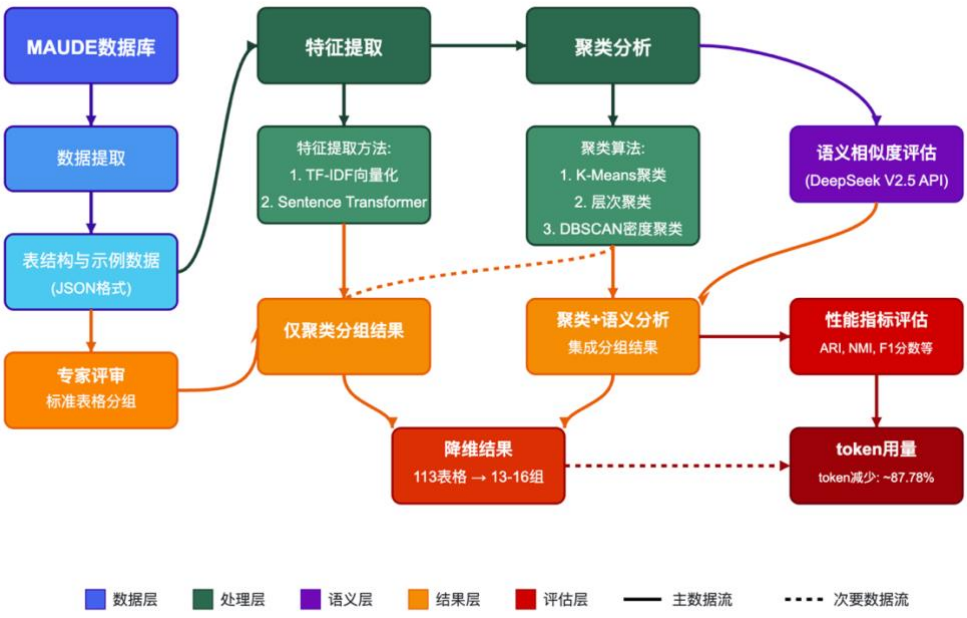

图 1 SCMF 框架下进行数据库表结构分组及性能评估的工作流程

3.1 系统配置

本框架的开发基于 Ubuntu 20.04 操作系统，并使用 PostgreSQL 12.18 管理 MAUDE 数据，确保大规模关系数据的稳健处理和高效查询。对于特征提取和聚类等计算密集型任务，采用了配备 24GB 内存的 Nvidia RTX 3090 GPU 进行加速。语义相似性计算使用 DeepSeek V2.5 应用接口服务（Application Programming Interface，API），一个开源、高性能的生成式预训练模型，语义理解能力优秀，使用成本低且 token 输出稳定速度极快（实验执行时间为 2024 年 10 月）。所有实验代码、数据分析和图形生成均使用 Python 3.9 实现。本研究涉及的源代码（包含所有提示词）、程序执行过程数据和结果报告均遵循 MIT 开源协议，发布于作者的 GitHub 页面，便于研究结果复现和进行二次开发[20]

3.2 数据集

本研究的数据集来源于 FDA 的 MAUDE 数据库，包含 1991 年至 2023 年 10 月的超过 4 000 万条记录。这些记录分布在 113 个文本文件中，包括 111 个可直接下载的 MAUDE 原生数据表格和 2 个在研究过程中衍生出的自定义表。为评估降维性能，领域专家将这 113 个表格根据其结构及内容进行分组形成最终的评价标准。完整的数据表名称及专家分组结果请详见附录 A。

3.3 数据和特征提取

数据提取涉及从指定网页下载所有压缩的 MAUDE 数据库文件，解压为 TXT 和 CSV 格式，并导入 PostgreSQL 12.18。为确保数据完整性，每个导入表格的行和列计数与 MAUDE 规范进行了验证。然后将每个表格转换为 JSON 文件，以封装其结构和样本数据，便于后续特征提取和相似性分析。特征提取旨在捕获结构和语义特征，以增强聚类性能。本研究采用词频-逆文档频率（Term Frequency-Inverse Document Frequency, TF-IDF）[21]向量化和句子转换器（Sentence Transformer）嵌入模型 all-MiniLM-L6-v2[22]进行语义特征提取。通过测试各种特征组合，确定了最有效的组合以提高聚类准确性。

3.4 表格聚类和相似性计算

表格聚类旨在通过提取的特征将 MAUDE 数据库中相似表格有效分组，从而揭示数据集内在关系与模式。该实验构建了一套较为简洁的特征集，仅通过拼接表结构描述与样本数据进行文本特征提取。通过应用多种聚类算法，包括 K-Means（涵盖手动指定和自动确定聚类数的方法）、层次聚类和 DBSCAN，高效实现了初步分组。

为精确细化分组并识别相似的数据表，实验设计了多层次相似性计算方案。首先，基于字段名称集合计算 Jaccard [23]相似系数，量化表格间的结构重叠度，并应用 0.1 的阈值识别结构相似性极低的数据表分组，提示异常情况及人工干预。其次，利用 DeepSeek 计算表格描述之间的语义相似性得分，这使得系统能够捕捉并量化细微的语义差异。然后再基于这些相似性得分，以 0.7、

0.8 和 0.9 为阈值确定表格合并的适用性。合并过程会优先保留数据量最大的数据表结构作为指代，同时记录被合并的所有数据表名称。

为评估聚类和相似性计算的有效性，实验将自动生成的分组结果与领域专家提供的手动分类进行了系统比较。并采用多种评估指标，包括调整兰德指数 ARI[24]、归一化互信息 NMI[25]、精确度、召回率和 F1 分数，以量化自动聚类与专家分类之间的一致性程度，从而确保 SCMF 框架的可靠性和准确性。

具体来说，ARI 通过考量所有样本对，计算在预测聚类和真实聚类中被分配到相同或不同集群的对数，并对随机分组进行校正，从而精确测量两个聚类结果之间的相似度。ARI 的公式为：

$$ARI = \frac{Index - Expected\ Index}{\mathrm{Max}\ Index - Expected\ Index} \quad (1)$$

其中$Index$表示两个聚类之间的一致数，$Expected\ Index$是由于随机机会而预期的一致数，$\mathrm{Max}\ Index$是指数的最大可能值。ARI 范围从-1 到 1，其中 1 表示聚类结果之间的完美一致，0 表示随机一致，负值表示比随机预期的一致性更少。

NMI 是另一种强大的聚类评估指标，它精确量化两个聚类分配之间的相互依赖程度，有效测量它们所共享的信息量。NMI 的标准化公式表示为：NMI 的公式为：

$$NMI = \frac{2 \times I(U;V)}{H(U) + H(V)} \quad (2)$$

其中$(U;V)$表示聚类U和V之间的互信息，量化两种分组方案之间共享的信息内容。术语$H(U)$和$H(V)$表示聚类U和V的熵，用于测量各自聚类内的不确定性或信息复杂度。通过将互信息与熵的几何平均值相结合，NMI 指标有效调整了不同聚类复杂度的影响，提供了一种平衡而稳健的相似性度量。NMI 值范围从 0 到 1，其中 0 表示两个聚类完全独立，而 1 表示完全一致的聚类结果。

为全面评估聚类性能，实验还使用了基于数据表分组的精确度、召回率和 F1 分数指标对分组结果进行评估。这些指标将聚类任务视为二元分类问题，即判断样本对是否应归入同一集群。精确度衡量模型预测为同类的样本对中实际确实属于同类的比例，其计算公式为：

$$Precision = \frac{TP}{TP + FP} \quad (3)$$

召回率则测量在所有实际应归为同类的样本对中，被模型正确识别的比例，表示为：

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

F1 分数是精确度和召回率的调和平均值，在两者之间提供平衡。

$$F1\text{-}Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (5)$$

在表格聚类评估中，这些指标可具体理解为：真阳性(TP)代表自动算法与专家标注均将其归为同组的数据表对数量；假阳性(FP)表示自动方法归为同组但专家标注为不同组的表对；假阴性(FN)则是专家标注为同组但自动方法未能正确分组的表对。通过此套评估体系，本研究能够全面客观地衡量聚类算法的有效性，为后续优化提供可靠依据。

3.5 token 用量变动评估

在利用预训练语言模型对数据库表结构进行分析和挖掘过程中，数据模式的描述信息要以提示词形式进行输入，所以实验创建了两组（SCMF 降维前后）原始提示词方便 MAUDE 研究人员直接使用与效果对比。同时本研究为了精确量化 SCMF 在降维 MAUDE 数据库结构表达方面的实际影响，对两组提示词 token 用量进行了比较，通过计算降维前后的 token 数量差异，量化 SCMF 对提示词文本长度的影响。

两组提示器文本由官方描述和数据库表结构相关信息两类内容组成，即 FDA 官方网站对 MAUDE 数据库结构和关键字段的详细描述，以及 SCMF 框架从数据库中提取的数据模式信息。通过两者进行叠加得到降维前后两组提示词。前者包含官方描述与 SCMF 处理前 113 个原始 MAUDE 数据表内容，后者是官方描述与 SCMF 处理后生成的描述。最后，基于 tiktoken 库[26]和 gpt2 编码[27]标准对以上两组提示内容进行 token 计数分析，得到提示词在降维前后 token 总数的绝对变化值和相对百分比变化，以客观量化该框架对 token 用量的影响。为确保评估结果的精确，即仅反映数据模式描述合并带来的 token 变化，该评估从提示文件的总 token 数中扣除了官方描述所占的固定 token 数量，确保对比的对象仅为 SCMF 作用的部分。

4. 结果

通过 96 组实验进行的综合分析成功地将 MAUDE 表结构数量从 113 个降维到 13，14 和 16 个（分别对应相似度阈值 0.7、0.8 和 0.9）。降维效果验证了本方法在结构简化方面的有效性，不仅能显著提高后续数据分析的效率，同时也确保了处理后的数据结构能够适配主流 LLM 的输入长度限制（从 111 787 减少至 13 664 个 token，压缩率达 87.78%）。实验结果表明，即使在三组中最保守的相似度阈值设置下，该方法仍能实现减少 85%的表结构的描述长度，为复杂数据模式的简化表达提供了有力支持。

4.1 SCMF 性能评估

性能评估显示，仅应用聚类时 ARI 值在 0.31 至 0.93 之间， NMI 值在 0.50 至 0.93 之间，F1 分数在 0.51 至 0.95 之间（详见表 1 和图 2A）。当混合使用聚类与 DeepSeek 时，ARI 提高至 0.85 至 1.00 之间，NMI 提高至 0.93 至 1.00 之间，F1 提高至 0.88 至 1.00 之间（详见表 1 和图 2B）。

表 1 综合性能指标和特征提取

类别	特征提取方法	ARI	NMI	F1 分数	图例
聚类	无	0.31 – 0.93	0.50 – 0.93	0.51 – 0.95	2A
聚类+DeepSeek	无	0.85 – 1.00	0.93 – 1.00	0.88 – 1.00	2B
聚类+特征提取	TF-IDF	0.31 – 0.84	0.50 – 0.88	0.51 – 0.87	2C
聚类+特征提取	Sentence Transformer	0.31 – 0.93	0.50 – 0.93	0.51 – 0.95	2C
聚类+特征提取+ DeepSeek	TF-IDF	0.94 – 1.00	0.97 – 1.00	0.95 – 1.00	2D
聚类+特征提取+ DeepSeek	Sentence Transformer	0.85 – 1.00	0.93 – 1.00	0.88 – 1.00	2D

关于特征提取方法，TF-IDF 向量化实现了 0.31 至 0.84 的 ARI 值，0.50 至 0.88 的 NMI 值，0.51 至 0.87 的 F1 分数，而 Sentence Transformer 嵌入实现了 0.31 至 0.93 的 ARI 值，0.50 至 0.93 的 NMI 值，0.51 至 0.95 的 F1 分数（详见表 1 和图 2C）。与 API 匹配结合时，TF-IDF 向量化产生了 0.94 至 1.00 的 ARI 值，0.97 至 1.00 的 NMI 值，0.95 至 1.00 的 F1 分数，而 Sentence Transformer 嵌入实现了 0.85 至 1.00 的 ARI 值，0.93 至 1.00 的 NMI 值，0.88 至 1.00 的 F1 分数（详见表 1 和图 2D）。

表 2 相似性阈值效果

相似性阈值	对分组的影响	组数量	图例
0.7	更多合并组	13	2E
0.8	平衡合并和精细化组	14	2E
0.9	更严格的相似性要求，更精细的分组	16	2E

相似性阈值分析显示，将阈值从 0.7 提高到 0.9 分别导致更多合并组和更精细的分组（详见表 2 和图 2E）。

表 3 按聚类方法的减少比率

聚类方法	减少比率(%)	图例
K-Means	高达 83%	2F
层次聚类	高达 80%	2F
DBSCAN	高达 77%	2F

DeepSeek 调用减少比例分析表明，通过使用 K-Means 进行预分组结合 DeepSeek 细分组的方法与仅使用 DeepSeek 相似度对比进行直接分组对比，可减少最高 83%的 DeepSeek 调用数量，使用层次聚类时上限为 80%，使用 DBSCAN 时为 77%（详见表 3 和图 2F）。因参数配置有差异，该比例会存在波动，但总体而言，集成使用聚类和语义相似性技术增强了 MAUDE 数据库内模式匹配和实体解析的效率和准确性。

4.2 token 用量变化

在评估过程中，当选取的相似性阈值为 0.8，聚类算法为 K-Means（K=3）及使用 TF-IDF 进行特征提取，得到了 F1 分数为 1.0 的最优解决方案。该参数配置将 113 个表结构描述降维为 14 个组。降维前提示文本共包含 113 954 个 token，其中 2 167 个来自官方解释，111 787 个来自数据

库模式描述。而降维后提示词仅包含 15 831 个 token，剔除官方解释文本长度，数据库描述内容长度减少至 13 664 个 token，其绝对值减少 98 123 个，百分比减少为 87.78%。降维后提示长度满足已知全部 LLM 的上下文窗口要求。

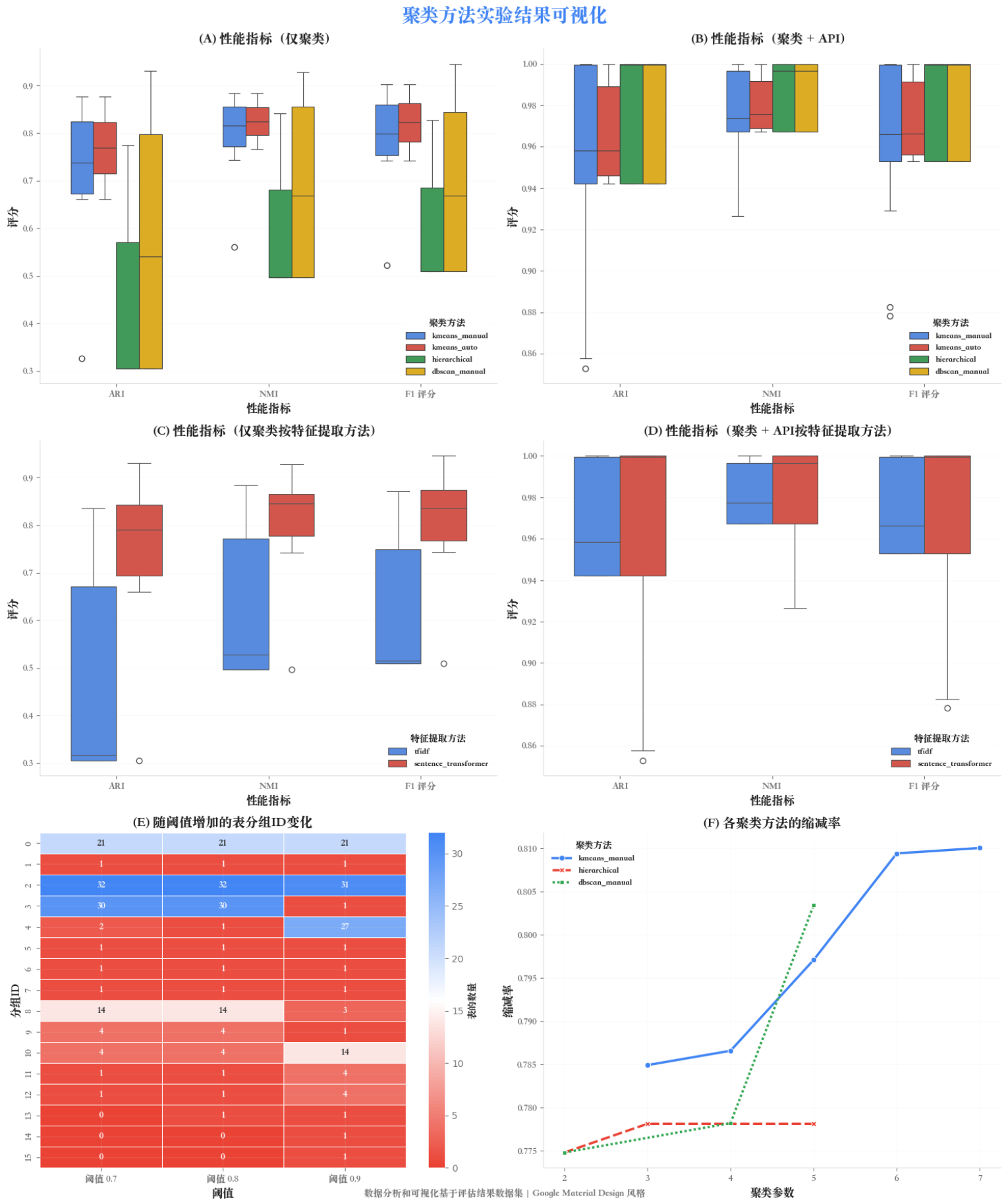

图 2 SCMF 框架多维度性能评估与方法对比分析

5. 讨论

5.1 结合聚类与语义相似性评估的有效性

聚类与语义相似性评估算法的结合显著提高了 MAUDE 数据库降维效果。F1 分数从仅使用聚类的 0.51 提高到与语义评估结合时的 1.00，强调语义理解在准确识别和合并相似表格中的关键作用。这种组合方法不仅提高了表格合并的精度，还确保了全面的召回，有效最小化假阳和假阴的案例。

5.2 计算效率和成本分析

初始聚类阶段，辅以基于 Jaccard 相似性的预过滤，将需要语义相似性计算的表格对数量减少了 77%至 83%。这种减少对管理计算资源至关重要，特别是处理具有众多潜在表格对的大规模数据库时。尽管使用 DeepSeek 会引入大规模参数模型的计算开销，但聚类精度的显著提高证明了这种投资的合理性。使用 1,425 次 API 调用实现完美的 F1 分数证明了所提出方法的成本效益，特别是在由于数据集规模或复杂性而无法进行手动表格合并的情况下。

5.3 SCMF 框架优势

SCMF 框架整合了结构相似性评估与深度语义相似性分析，实现了对跨数据库表结构和样例数据的多维度语义关联评估。实验结果表明，这种双层评估机制能够精确捕获数据表之间的结构重叠度及语义差异的细微变化，从而在数据模式匹配领域实现了可量化的精度提升。

值得强调的是，SCMF 框架的设计原理不依赖于 MAUDE 数据库的任何特定特征或独有属性，这保证了其在面对复杂度相当的其他领域数据库系统时依然保持良好的适应性和泛化能力。本研究对方法和参数配置的全面而详尽的描述，在确保实验结果可复现的同时也为其他研究团队提供了执行路径参考，便于在不同数据环境中进行验证测试和应用迁移。

5.4 兼容大语言模型上下文窗口限制的价值

SCMF 框架实现了 87.78%的 token 压缩率，有效解决了大型语言模型在处理复杂数据库结构时的三大挑战：(1)处理能力限制，将原本超出大多数 LLM 上下文窗口的数据表结构描述长度降至可处理范围；(2)计算资源优化：显著降低 API 调用的 token 用量及相关成本；(3)响应质量提升：通过消除冗余信息优化模型对数据结构的理解效率与输出质量。这一突破对医疗机构等数据密集型环境尤为有价值，为基于 LLM 的智能数据库分析提供了实用可行的技术路径。

5.5 局限性和未来方向

尽管 DeepSeek 为 SCMF 框架提供了强大的语义相似性评估能力，但对特定工具的依赖不可避免地引入了长期维护和版本更新方面的潜在风险。针对这一局限性，未来研究可着力于构建模块化的语义评估架构，通过整合多种开源语义相似性评估工具或开发垂类模型，以增强框架的技术灵活性和适应能力。

MAUDE 等复杂医疗数据集中固有的语义歧义和术语变异仍然构成重要挑战。为解决这一问题，后续工作可考虑引入医疗领域特定的本体或知识图谱，从而进一步提升语义消歧能力和相似性

评估的精准度。同时，拓展方法包含更多的降维技术如自动编码器[28]和流形学习方法[29]等有望进一步增强框架在处理超大规模异构数据时的鲁棒性和可扩展性。

5.6 对医学信息学及其他领域的影响

SCMF 自动降维方法在 MAUDE 数据库中的成功应用可显著提升数据分析效率，间接促进了患者安全事件的快速和精准洞察，有助于行业监管和改善临床结果。此外其应用价值并不局限于医疗健康领域，金融、政务、制造等同样面临复杂异构数据挑战的行业也可以采纳应用。特别值得注意的是，该方法能够在 LLM 上下文窗口限制下有效降维大规模医疗数据集，为高级数据分析和智能解释开辟了全新可能性，并使研究人员和实践者能够将更多精力投入到高价值的分析任务和知识发现中。这种创新性的数据简化方法不仅解决了当前大规模数据处理中的技术瓶颈，更为医学信息学及相近领域的数据驱动决策提供了坚实基础，有望加速从数据到洞察、从洞察到行动的转化进程。

6. 结论

本研究提出了一种创新的自动化降维方法，通过 FDA MAUDE 数据库的实证研究展示了其在处理复杂数据结构方面的有效性。实验整合了聚类算法与 DeepSeek V2.5 的语义相似性评估能力，成功将原始 113 个表结构降维至 13-16 个语义一致组，在数据模式表达显著简化的基础上达到了与专家分组结果的高度一致，验证了其可靠性。更重要的是，通过将表结构描述长度降至适合 LLM 输入限制的规模同时保留上下文完整性，有效解决了利用 LLM 分析复杂数据库的关键障碍。虽然本研究以医疗数据为实证基础，但所提出方法的灵活性使其具备广泛的跨领域应用潜力，为数据处理工作流程简化、分析效率提升和监管决策优化提供了有力工具，有望在医疗保健及其他数据密集型领域产生深远影响。

作者贡献： 华磊负责研究设计、数据分析和论文撰写；巩洋负责研究指导和论文修改。

利益声明： 所有作者均声明不存在利益冲突。

7. 参考文献

[1] U.S. FOOD AND DRUG ADMINISTRATION. Manufacturer and User Facility Device Experience (MAUDE)[EB/OL]. (2024)[2024-11-01]. https://www.fda.gov/medical-devices/mandatory-reporting-requirements-manufacturers-importers-and-device-user-facilities/about-manufacturer-and-user-facility-device-experience-maude-database.

[2] LI X, FENG Y, GONG Y, 等. Assessing the Reproducibility of Research Based on the Food and Drug Administration Manufacturer and User Facility Device Experience Data[J/OL]. Journal of Patient Safety, 2024, 20(5). DOI:10.1097/PTS.0000000000001220.

[3] BATINI C, CAPPIELLO C, FRANCALANCI C, 等. Methodologies for data quality assessment and improvement[J/OL]. ACM Computing Surveys, 2009, 41(3). DOI:10.1145/1541880.1541883.

[4] SHI Y, YU Y, FENG Y, 等. A Data Pipeline for Enhancing Quality of MAUDE-Based

Studies.[J/OL]. Studies in health technology and informatics, 2024, 316: 1214-1218. DOI:10.3233/SHTI240629.

[5] SHETH A P, LARSON J A. Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases[J/OL]. ACM Computing Surveys (CSUR), 1990, 22(3). DOI:10.1145/96602.96604.

[6] DOAN A H, HALEVY A, IVES Z. Principles of Data Integration[M/OL]//Principles of Data Integration. 2012. DOI:10.1016/C2011-0-06130-6.

[7] VASWANI A, SHAZEER N, PARMAR N, 等. Attention is all you need[C]//Advances in Neural Information Processing Systems: 卷 2017-December. 2017.

[8] BROWN T B, MANN B, RYDER N, 等. Language models are few-shot learners[C]//Advances in Neural Information Processing Systems: 卷 2020-December. 2020.

[9] YU Y, SHI Y, FENG Y, 等. Developing a Generative AI-Powered Chatbot for Analyzing MAUDE Database.[J/OL]. Studies in health technology and informatics, 2024, 316: 1255-1259. DOI:10.3233/SHTI240639.

[10] LLOYD S P. Least Squares Quantization in PCM[J/OL]. IEEE Transactions on Information Theory, 1982, 28(2). DOI:10.1109/TIT.1982.1056489.

[11] MURTAGH F, CONTRERAS P. Algorithms for hierarchical clustering: An overview[J/OL]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2(1). DOI:10.1002/widm.53.

[12] ESTER M, KRIEGEL H P, SANDER J, 等. A Density-Based Algorithm for Discovering Clusters A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise[C]//Proceedings - 2nd International Conference on Knowledge Discovery and Data Mining, KDD 1996. 1996.

[13] LIU A, FENG B, WANG B, 等. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model[J/OL]. arxiv.org, 2024[2024-10-22]. https://arxiv.org/abs/2405.04434v2

[14] RAHM E, BERNSTEIN P A. A survey of approaches to automatic schema matching[A/OL]//VLDB Journal. (2001). DOI:10.1007/s007780100057.

[15] BERLIN J, MOTRO A. Database schema matching using machine learning with feature selection[C/OL]//Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): 卷 2348. 2002. DOI:10.1007/978-3-642-36926-1_25.

[16] GAL A. Uncertain Schema Matching[J/OL]. Synthesis Lectures on Data Management, 2011, 3(1). DOI:10.2200/s00337ed1v01y201102dtm013.

[17] AUMUELLER D, DO H H, MASSMANN S, 等. Schema and ontology matching with COMA++[C/OL]//Proceedings of the ACM SIGMOD International Conference on Management of Data. 2005. DOI:10.1145/1066157.1066283.

[18] CASTANO S, FERRARA A, MONTANELLI S. Matching ontologies in open networked systems: Techniques and applications[C/OL]//Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): 3870 LNCS. 2006. DOI:10.1007/11617808_2.

[19] WANG Z, LI J, WANG Z, 等. Cross-lingual knowledge linking across wiki knowledge bases[C/OL]//WWW'12 - Proceedings of the 21st Annual Conference on World Wide

Web. 2012. DOI:10.1145/2187836.2187899.
[20] HUA L. MAUDE-Schema-Compressor[A/OL]//Github.com. github.com, 2024[2025-01-09]. https://github.com/leiMizzou/MAUDE-Schema-Compressor.
[21] JONES K S. A statistical interpretation of term specificity and its application in retrieval[A/OL]//Journal of Documentation. (1972). DOI:10.1108/eb026526.
[22] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence embeddings using siamese BERT-networks[C/OL]//EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. 2019. DOI:10.18653/v1/d19-1410.
[23] JACCARD P. The Distribution of the Flora in the Alpine Zone[J/OL]. New Phytologist, 1912, 11(2). DOI:10.1111/j.1469-8137.1912.tb05611.x.
[24] RAND W M. Objective criteria for the evaluation of clustering methods[J/OL]. Journal of the American Statistical Association, 1971, 66(336). DOI:10.1080/01621459.1971.10482356.
[25] VINH N X, EPPS J, BAILEY J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance[J]. Journal of Machine Learning Research, 2010, 11.
[26] OPENAI. tiktoken: An OpenAI library for tokenization[A/OL]. Github, 2023[2024-11-01]. https://github.com/openai/tiktoken.
[27] ALEC R, JEFFREY W, REWON C, 等. Language Models are Unsupervised Multitask Learners | Enhanced Reader[J]. OpenAI Blog, 2019, 1(8).
[28] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J/OL]. Science, 2006, 313(5786). DOI:10.1126/science.1127647.
[29] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J/OL]. Science, 2000, 290(5500). DOI:10.1126/science.290.5500.2323.

8. 资金支持

本研究得到了山西省留学回国人员科技活动择优资助项目（项目编号：20230039）的支持，但资助机构对研究设计、数据收集与分析、发表决策以及手稿撰写均未施加任何影响。

---

9. 数据可用性

MAUDE 开源数据库文件可通过以下链接从 FDA 的 MAUDE 数据库获取，访问地址为：https://www.fda.gov/medical-devices/maude-database。本研究相关的所有数据，包含程序完整 python 代码，原始提示词内容，测试过程及结果数据可以通过以下链接进行访问下载，地址为：https://github.com/leiMizzou/MAUDE-Schema-Compressor

---

10. 补充材料

附录 A：MAUDE 及自建统计表格的分组情况

Group	Table Names
0	ASR_2006, ASR_2016, ASR_2010, ASR_2009, ASR_2008, ASR_2013, ASR_2015, ASR_2002, ASR_2000, ASR_2012, ASR_2004, ASR_1999, ASR_2007, ASR_2011, ASR_2003, ASR_2017, ASR_2005, ASR_2014, ASR_2001, ASR_2018, ASR_2019
1	DISCLAIM
2	ASR_PPC
3	foitext, foitext2012, foitext2005, foitext2002, foitext1998, foitext2010, foitext2014, foitext2001, foitext2008, foitext1999, foitext1997, foitext2011, foitext2017, foitextthru1995, foitext2009, foitext2013, foitextChange, foitext2016, foitext2023, foitextAdd, foitext2003, foitext2021, foitext2006, foitext2004, foitext2018, foitext2015, foitext2020, foitext2019, foitext1996, foitext2000, foitext2007, foitext2022
4	mdr96, mdr97, mdr84, mdr88, mdr91, mdr90, mdr93, mdr87, mdr92, mdr85, mdr95, mdr86, mdr94, mdr89
5	mdrfoiChange, mdrfoi, mdrfoiAdd, mdrfoiThru2023
6	patientThru2023, patientAdd, patient, patientChange
7	patientproblemcode
8	deviceproblemcodes
9	table_statistics_selfdefined
10	DEVICE2018, DEVICE2022, DEVICE2013, DEVICE2015, DEVICE2017, DEVICE2014, DEVICE2008, DEVICE2003, DEVICE2016, DEVICEAdd, DEVICE2021, DEVICE2005, DEVICE2007, DEVICE2023, DEVICE2006, DEVICE2000, DEVICE2020, DEVICEChange, DEVICE2002, DEVICE2001, DEVICE, DEVICE2011, DEVICE2009, DEVICE2004, DEVICE2019, DEVICE2010, DEVICE2012, foidevthru1997, foidev1999, foidev1998, foidevproblem
11	foidevproblem
12	foiclass_selfdefined
13	patientproblemdata