分层嵌入大模型架构在多模态健康医疗数据治理中的应用

马 良[1] 马俊朋[1] 李向阳[1] 谢 超[1] 刘 超[1] 董尚文[2]
MA Liang MA Junpeng LI Xiangyang XIE Chao LIU Chao DONG Shangwen

摘 要 在人工智能时代的背景下，深入探索健康医疗数据治理领域对于推动医疗大数据应用和挖掘数据价值至关重要。文章提出了一种基于大模型技术的分层嵌入架构，通过构建分层次、模块化的嵌入表示智能体，在数据采集、标准化和集成等核心环节中实现高效协同，支持数据全生命周期的智能管理，从而推动多模态健康医疗数据治理的智能化。研究结果表明，该架构在电子病历解析、医学影像分析和专病数据库建设等应用中显著提升了数据治理的效率和准确性，为解决多模态健康医疗数据治理中的复杂问题提供了高效智能的解决方案，具有良好的扩展性和应用前景，未来有望在更多临床场景中发挥重要作用。

关键词 健康医疗大数据；分层嵌入大模型架构；人工智能；大模型；数据治理

doi：10.3969/j.issn.1672-9528.2025.01.032

0 引言

医疗行业数字化转型的快速发展催生了海量的多模态健康医疗数据，如电子病历、医学影像、实验室检验结果等。这些数据在疾病诊断、精准医疗及药物研发等方面发挥了巨大作用。但要实现其意义，需经历复杂且多环节的数据治理过程，而多模态健康医疗数据的异构性和复杂性，给数据治理带来了巨大挑战。在此背景下，大模型强大的语义理解和自然语言处理能力、多模态数据处理和特征提取能力、应对数据复杂性的泛化能力等，为有效解决这些挑战提供了新的机遇[1-2]。

本文提出了一种分层嵌入大模型（hierarchical embedding large model）架构，以大模型技术为核心引擎，深度融合数据治理各阶段的实际需求，通过构建分层次、模块化的嵌入表示智能体，将不同模态的数据映射到同一语义空间，实现数据治理任务的精细化分解与高效协同。

1 分层嵌入大模型架构

健康医疗数据治理包括数据采集、数据标准化和数据集成等多个环节[3]。数据采集汇聚海量且多样的健康医疗数据；数据标准化将采集到的异构、多模态数据进行清洗、转换和标准化，使其具备一致的格式和结构，便于后续处理与分析；数据集成将标准化后的多模态数据进行整合，构建统一的数据视图，使不同来源的数据能够协同工作。

现有数据治理框架和方法在处理多模态数据方面存在一定局限。首先，不同模态的数据存在不同的结构和语义特征，其异构性使得数据标准化和整合变得复杂；其次，多模态数据的复杂性增加了信息挖掘的难度，现有分析工具往往局限于单模态数据，无法利用全部信息；最后，健康医疗数据含有大量敏感信息，如何确保数据的机密性、完整性和可用性，防止数据泄露与滥用，仍是当前面临的一项严峻挑战[4]。

1.1 整体架构

本文提出的分层嵌入大模型架构包含专能组件、智能体编排调度引擎和智能体三部分，整体架构如图 1 所示。

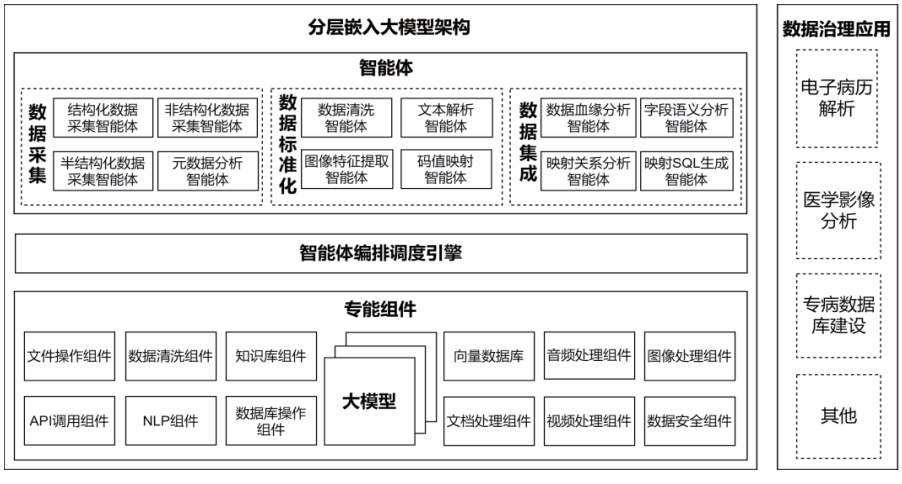

图 1 分层嵌入大模型架构示意图

通过编排调度引擎，基于大模型强大的数据分析和处理能力，构建针对数据采集、标准化、集成等环节的智能体，并调度协同多个智能体，共同完成具体复杂任务，实现数据

1. 山东浪潮智慧医疗科技有限公司 山东济南 250101
2. 河南科技大学 河南洛阳 471023

治理的自动化、智能化。具体而言，在数据采集层，智能体可以自动识别并提取结构化、非结构化和半结构化数据中的关键信息；在数据标准化层，智能体可以利用知识库和向量检索等技术对数据进行清洗和标准化处理；在数据集成层，智能体可以自动发现并建立数据之间的映射关系，实现数据的智能集成。

1.2 专能组件

在分层嵌入大模型架构中，专能组件（specialized component）是指为执行特定任务而设计的一段代码或模块。这些组件通常聚焦于处理特定的数据类型或执行特定的操作，是整个架构中的功能单元。可以是预训练的大模型、自定义的算法或第三方服务接口。专能组件清单如表 1 所示。

表 1 专能组件清单

组件名称	组件功能
大模型组件	负责处理自然语言理解、文本分析、SQL 生成、音视频解析、图像处理等复杂的多模态任务。
文件操作组件	负责对文件进行获取、读写、上传下载、格式转换等操作。
数据清洗组件	负责对数据进行过滤、去重、补全等预处理操作。
知识库组件	负责对医疗行业知识进行嵌入操作。
向量数据库组件	负责存储和检索高维向量数据等操作。
API 调用组件	负责调用外部的 API 接口，获取或传递数据，实现与其他系统或服务的交互。
NLP 组件	负责处理文本数据，包括分词、词性标注、命名实体识别等操作。
数据库操作组件	负责与数据库进行交互，实现数据的存储、查询、更新等操作。
文档处理组件	负责处理各种格式的文档，如 PDF、Word、Excel 等，包括文档转换、提取文本、表格数据等操作。
音频处理组件	负责处理音频数据，包括语音识别、音频分类等操作。
视频处理组件	负责处理视频数据，包括视频分类等操作。
图像处理组件	负责处理图像数据，包括图像分类、目标检测、图像分割等操作。
数据安全组件	负责提供数据加密解密、匿名化处理等操作。

专能组件的引入，使得分层嵌入大模型架构更加灵活、可扩展，能够更好地适应各种复杂的数据治理应用场景。通过合理地组合和配置这些组件，可以构建出高效的智能体应用。

1.3 智能体编排调度引擎

架构中的智能体编排调度引擎负责编排多个专能组件，从而构建一个智能体，并协调各专能组件与核心大模型之间的交互。

智能体是架构中的专能组件在健康医疗数据治理中某一特定场景下的动态组合应用，每个智能体都包含多个专能组件，应对复杂的多模态数据治理需求。具体来说，架构为智能体提供专能组件基础，智能体通过专能组件的编排实现具体任务，如电子病历解析任务、医学影像分析任务、多模态数据的联合分析等。通过这种方式，智能体可以灵活适应不同的医疗需求。

1.4 智能体嵌入

在数据治理的各环节嵌入智能体，可以充分利用智能体的自动化处理能力、数据安全和隐私保护能力以及协同工作能力等优势，解决多模态健康医疗数据治理中的复杂问题。

数据采集层是整个治理流程的基础环节，采集操作包括从多种数据源中获取原始数据、对原始数据进行元数据分析，发现数据之间的潜在关联。由于健康医疗数据通常表现为多种形式，包括结构化、半结构化和非结构化数据，因此，在数据采集层中嵌入多个智能体，用来处理不同类型的数据采集和元数据分析。数据采集层智能体清单如表 2 所示。

表 2 数据采集层智能体清单

智能体名称	包含的组件及功能
结构化数据采集智能体	包含数据库操作组件、API 调用组件。负责关系型数据库数据的采集。
半结构化数据采集智能体	包含文件操作组件、API 调用组件。负责 TXT 文件、XML 文件、JSON 文件等文件的采集。
非结构化数据采集智能体	包含文件操作组件、API 调用组件。负责医学影像、音视频文件等文件的采集。
元数据分析智能体	包含大模型组件、文件操作组件、API 调用组件、知识库组件。负责通过数据上下文及知识库中预置的领域知识自动识别它们之间的等价关系，基于元数据推断数据间的潜在关系。

数据标准化层是数据治理流程中承上启下的关键一环，通过一系列的规则和规范，对采集到的数据进行处理，使其符合预定的标准。标准化层的操作包括数据清洗、文本解析、图像特征提取、将原始数据码值转换成标准码值等操作。由于健康医疗数据来源多样，数据格式、编码标准和表示方式可能存在较大差异，为解决这一层面问题，在数据标准化层中嵌入多个智能体，分别处理数据清洗、文本解析、图像特征提取和码值映射等任务。数据标准化层智能体清单如表 3 所示。

表 3 数据标准化层智能体清单

智能体名称	包含的组件及功能
数据清洗智能体	包含文档处理组件、数据清洗组件。负责处理数据中的异常、缺失值和重复数据，为后续的标准化和分析过程提供高质量的数据输入。
文本解析智能体	包含 NLP 组件、知识库组件。负责通过自然语言处理技术，对医疗文本进行语义分析和结构化处理。
图像特征提取智能体	包含图像处理组件、向量数据库组件。负责处理医学影像数据（如 CT、MRI、X 光片等），提取图像中的关键特征并向量化。
码值映射智能体	包含 NLP 组件、向量数据库组件、知识库组件、大模型组件。负责将不同来源、不同格式的医疗数据码值（如疾病编码、药品编码等）进行标准化映射，将不同编码系统之间的码值转换为标准码值。

数据集成层的任务是将来自不同来源的、标准化后的数据整合在一起，形成一个统一的数据视图，为后续数据分析和应用提供基础。集成层的操作包括数据血缘分析、数据映射关系分析、数据映射 SQL 生成等操作。由于数据来自多个异构系统和来源，具有复杂的格式和结构，因此在集成过程中需要确保数据的可追溯性、准确性和一致性。为解决这一层面的问题，在数据集成层中嵌入多个智能体，分别处理数据血缘分析、字段语义分析、映射关系分析和映射 SQL 生成等任务。数据集成层智能体清单如表 4 所示。

表 4 数据集成层智能体清单

智能体名称	包含的组件及功能
数据血缘分析智能体	包含数据库操作组件、大模型组件。负责追踪数据的来源、演变和加工流程，通过大模型生成数据血缘关系图，辅助理解数据的流动和加工过程。
字段语义分析智能体	包含数据库操作组件、向量数据库组件、知识库组件、大模型组件。负责对不同来源数据表中的字段进行语义分析，识别字段的实际含义，并将不同系统中含义相同但命名或格式不同的字段进行统一映射。
映射关系分析智能体	包含数据库操作组件、大模型组件。负责分析和构建不同数据源之间的映射关系，包括来源表、来源字段、目标表、目标字段等。
映射 SQL 生成智能体	包含数据库操作组件、大模型组件。根据字段映射关系和语义分析结果，由大模型组件生成用于数据集成的 SQL 转换语句。

本文提出的分层嵌入大模型架构为多模态健康医疗数据治理提供了新的思路，其丰富的专能组件体系、精细化的数据处理能力以及统一的编排策略，为构建智能化、高效化、安全化的多模态健康医疗数据治理体系提供了重要的技术支撑与实践路径。

2 多模态健康医疗数据治理应用

2.1 电子病历解析

电子病历作为患者诊疗过程的数字化记录，包含了患者基本信息、疾病诊断内容等，是典型的半结构化数据，电子病历中不仅有结构化数据，也有大量的自由文本记录，如口语化的病情描述、病史、症状，格式和编码各异，难以进行准确地数据分析和处理。

传统解析方法大多依赖于手工规则和特征工程，灵活性较差，开发和维护成本高，难以适应复杂的医疗文本，而大模型在自然语言处理和语义理解方面具有显著优势，特别是在处理复杂和非结构化文本时，能够有效捕捉上下文信息和语义关系。通过编排多个智能体和组件来实现电子病历解析。首先，通过半结构化数据采集智能体，从数据库或磁盘中获取电子病历文件，由文档处理组件读取病历文件，进行格式处理、文本分段操作，完成这些处理后，调用数据安全组件对病历中的敏感信息进行脱敏处理，脱敏后的内容传送至数据清洗智能体，进行文本过滤、去重等操作。数据清洗后，结合电子病历标准标签知识库，解析出病历中包含的标准标签。最后，通过大模型组件进行语义解析，提取关键信息，如患者的症状、诊断结果和治疗方案，并结合知识库中的医学术语，对文本内容进一步进行语义分析，实现标准化的医学知识对齐。

电子病历解析是一个复杂任务，通过在解析环节嵌入灵活的智能体，本文构建了一个可配置、可扩展的电子病历解析框架。当治理需求发生变化时，只需对相关智能体或组件的配置进行调整，即可快速适应新的要求，而无需进行大规模改动，从而保持解析的高效性和准确性。使用智能体方式解析 500 万份电子病历，相比于传统的深度模型解析方法，时间缩短了 37%。

2.2 医学影像分析

医学影像分析是现代医学诊断的重要组成部分，包括病灶检测、图像分割、图像分类等多个方面，传统的手工影像分析耗时费力，容易受到主观因素的影响，出现误诊，而基于特定数据集训练模型需要专业医生的参与，数据特征提取复杂、标注成本高，模型在面对不同机构或设备生成的数据时泛化能力不足。

分层嵌入大模型架构在进行医学影像分析时，首先调用非结构数据采集智能体获取医学影像文件，后调用图像处理组件进行旋转、缩放、去噪、对比度调整等操作，由大模型组件提取影像特征，识别出可能存在的异常区域，图像特征会被向量化并存储在向量数据库中，再次调用大模型组件，结合异常区域的特征向量及知识库，对异常区域进行进一步的分类与分割。

分层嵌入大模型架构可以动态地调整分析策略，适应多种医学影像数据源、模态和应用场景，使得智能体能够快速适应新的医疗场景，从而保持较强的泛化能力和准确性。同样，相比于深度模型分析方式，智能体方式将医学影像分析的时间缩短了 25%，病灶分割准确率提升了 30%。

2.3 专病数据库建设

专病数据库（专病库）是专门针对单病种建立的包含该病种患者临床资料和随访信息的数据库，如肺结节专病数据库、糖尿病专病数据库、高血压专病数据库等。专病库汇集了大量特定疾病的临床数据，可为临床科研、辅助诊疗、医疗资源优化提供宝贵的数据资源。

在专病库的建设过程中，往往存在数据不充分、缺乏关键数据、存在逻辑错误、录入误差、数据不标准等问题，导致专病库质量下降，进而影响临床研究结果的产出和质量，使用传统手段很难解决数据质量和数据标准化等问题。

大模型在处理大规模医疗数据，如数据采集、清洗、标准化等方面具有优势。在临床科研中，数据获取的过程往往占据了临床科研整个流程时间的 1/3 以上，因此，通过分层嵌入大模型架构，在数据采集层，调用多模态数据采集智能体，对结构化、非结构化、半结构化数据进行高效地采集，解决数据不充分的问题，调用数据安全组件对采集的医疗数据进行脱敏，保护数据隐私；在数据标准化层，调用数据清洗智能体来识别和纠正数据中的错误和异常，调用码值映射智能体对医疗机构中采用不同标准的数据进行标准化。通过在专病库建设的关键环节嵌入智能体或专能组件，有效提升了医疗数据的质量，为后续临床科研提供了有力的数据支持，促进了临床科研的发展。

分层嵌入大模型架构内置了多个智能体和专能组件，通过调度这些智能体和专能组件，使得数据采集、处理效率、处理准确度大幅提升，为数据治理应用提供了可靠、高质量的数据资源。

（下转第 146 页）

strategy[DB/OL].(2019-09-10)[2024-07-25].https://doi.org/10.48550/arXiv.1909.04273.

[12] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[DB/OL].(2019-09-07)[2024-07-19].https://doi.org/10.48550/arXiv.1909.03227.

[13] ZHENG H Y, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[DB/OL].(2021-06-18)[2024-06-19].https://doi.org/10.48550/arXiv.2106.09895.

[14] REN F L, ZHANG L H, ZHAO X F, et al. A simple but effective bidirectional framework for relational triple extraction[C]//Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. NewYork:ACM, 2022: 824-832.

[15] ZHANG X L, ZHANG T, YAN R. Chinese medical intent recognition based on multi-feature fusion[J]. Neural information processing, 2023, 1962(11): 504-515.

[16] SUN C A, LI T J, LIU T L, et al. A joint model based on interactive gate mechanism for spoken language understanding[J]. Applied intelligence, 2021, 52(8): 6057-6064.

[17] LI S Y, WANG T T, LI G Y, et al. Short-term ship roll motion prediction using the encoder–decoder Bi-LSTM with teacher forcing[J]. Ocean engineering, 2024,295(3): 116917.

[18] 甘子发，昝红英，关同峰，等 .CHIP 2020 评测任务 2 概述：中文医学文本实体关系抽取 [J]. 中文信息学报，2022, 36(6): 101-108.

[19] SUI D B, ZENG X R, CHEN Y B, et al. Joint entity and relation extraction with set prediction networks[J]. IEEE transactions on neuralnetworks and learning systems, 2023, 35(9): 12784-12795.

[20] REN F L, ZHANG L H, YIN S J, et al. A novel global feature-oriented relational triple extraction model based on table filling[DB/OL]. (2021-09-14)[2024-05-25].https://doi.org/10.48550/arXiv.2109.06705.

[21] 余肖生，李琳宇，周佳伦，等 . AMFRel: 一种中文电子病历实体关系联合抽取方法 [J]. 重庆理工大学学报 (自然科学), 2024,38(2):189-197.

【作者简介】

张云涛（1999—），通信作者（email: 964447582@qq.com），男，湖北武汉人，硕士研究生，研究方向：大数据分析技术研究。

黄莺（2000—），女，湖北武汉人，硕士研究生，研究方向：大数据分析技术研究，email: 156376342@qq.com。

（收稿日期：2024-09-02）

（上接第 139 页）

3 结语

本文提出了一种分层嵌入大模型架构，该架构以大模型技术为基础，提供了一个高效、智能的解决方案，以应对当前多模态健康医疗数据治理中所面临的复杂挑战。通过编排调度引擎将专能组件进行编排组合，形成数据治理各环节中的智能体，实现了对多模态数据的智能化处理，并成功应用于电子病历解析和医学影像分析。事实证明，该框架可为数据治理中的复杂任务提供更加精确、智能、高效的解决方案。

未来，随着大模型技术的进一步发展和医疗数据的不断积累，基于分层嵌入大模型的智能治理架构有望在更多的临床场景中发挥作用，为精准医疗、个性化健康管理等领域提供重要的技术支撑。

参考文献：

[1] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: transformers for image recognition at scale[DB/OL].(2021-06-03)[2024-02-18].https://doi.org/10.48550/arXiv.2010.11929.

[2] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[DB/OL].(2019-05-24)[2024-02-06].https://doi.org/10.48550/arXiv.1810.04805.

[3] 阮彤，邱加辉，张知行，等 . 医疗数据治理 : 构建高质量医疗大数据智能分析数据基础 [J]. 大数据，2019(1):12-24.

[4] 刘志红 . 人工智能大模型的隐私保护与数据安全技术研究 [J]. 软件，2024, 45(2): 143-145+151.

【作者简介】

马良（1989—），男，山东泰安人，硕士研究生，高级工程师，研究方向：大数据、人工智能、数据治理。

马俊朋（1979—），男，山东泰安人，本科，工程师，研究方向：大数据、人工智能。

李向阳（1981—），男，山东济南人，本科，高级工程师，研究方向：大数据、物联网、元宇宙、人工智能。

谢超（1983—），男，山东枣庄人，本科，工程师，研究方向：大数据、人工智能、数据治理。

刘超（1993—），男，山东临沂人，博士，工程师，研究方向：生物医学图像处理、生成式大模型应用研究。

（收稿日期：2024-09-10）