Computer Science and Application 计算机科学与应用, 2026, 16(5), 58-65
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165164

Hans汉斯

基于大语言模型的非结构化流调数据治理与报告生成研究

蒋松冬

广西民族师范学院数学与计算机科学学院，广西 崇左

收稿日期：2026年4月11日；录用日期：2026年5月11日；发布日期：2026年5月22日

摘　要

目的：针对流调资料来源分散、表达口语化、时间信息隐含导致的信息整理效率低、报告编制负担重和审计追溯不足等问题，提出面向非结构化文本的数据治理与报告生成一体化方案。方法：以大语言模型为核心，建立数据接入登记、分级脱敏、个案主键关联、版本管理和质量标注机制；构建语义抽取、结构化校验和模板化成稿流程，并结合时间解析、术语映射和规则引擎，形成异常识别与人工复核闭环；通过签审流程、角色权限和审计日志实现分级人机协同。结果：研究形成了以个案为主线的分层架构和可落地流程，实现了原始叙述、结构化要素与规范文书的连续衔接。原型应用表明，该方案能够支持多源文本接入、隐私治理、冲突待办生成和可追溯签审管理，在流程规范性、结果一致性和责任追踪方面表现稳定。结论：该研究为疾控机构在私有化部署和强审计约束条件下推进非结构化流调数据治理与报告自动化提供了可实施的工程路径。

关键词

大语言模型，流行病学调查，非结构化数据，数据治理，信息抽取，报告生成，人机协同

Research on Governance of Unstructured Epidemiological Investigation Data and Report Generation Based on Large Language Models

Songdong Jiang

School of Mathematics and Computer Science, Guangxi Minzu Normal University, Chongzuo Guangxi

Received: April 11, 2026; accepted: May 11, 2026; published: May 22, 2026

文章引用: 蒋松冬. 基于大语言模型的非结构化流调数据治理与报告生成研究[J]. 计算机科学与应用, 2026, 16(5): 58-65.
DOI: 10.12677/csa.2026.165164

Abstract

Purpose: To address decentralized sources of epidemiological investigation materials, colloquial wording, implicitly expressed temporal information, and the resulting low efficiency in information collation, heavy workload in report drafting, and weak audit traceability, this study proposes an integrated approach to data governance and report generation for unstructured text. Methods: A large language model serves as the core engine. The approach establishes mechanisms for data intake registration, tiered de-identification, case-level primary-key linkage, version control, and quality annotation; builds a pipeline of semantic extraction, structured validation, and templated document generation; and integrates temporal parsing, terminology mapping, and a rules engine to form a closed loop of anomaly detection and human review. Tiered human-machine collaboration is implemented through approval workflows, role-based access control, and audit logs. Results: The work yields a case-centered layered architecture and practical workflows that connect raw narratives, structured elements, and standardized documents in a continuous chain. Prototype use indicates that the solution supports multi-source text intake, privacy governance, generation of conflict-related action items, and traceable approval management, with stable performance in procedural rigor, consistency of outputs, and accountability tracking. Conclusion: The study offers a feasible engineering path for public health agencies to advance governance of unstructured epidemiological investigation data and automation of reporting under on-premises deployment and strong audit requirements.

Keywords

Large Language Model, Epidemiological Investigation, Unstructured Data, Data Governance, Information Extraction, Report Generation, Human-Machine Collaboration

1. 引言

1.1. 研究背景与问题提出

流调业务的核心产出既包括用于决策与协查的结构化事实(时间、地点、人群、暴露和接触链等)，也包括符合规范的叙事型报告[1]。在实际工作中，大量信息最初以非结构化文本进入流程：一线人员通过不同渠道记录细节，同一要素可能反复出现且表述不一致，时间信息也常以相对说法或口语形式嵌入句段[2]。疾病文本挖掘与暴发情境分析研究表明，从叙述中恢复可计算要素是关键问题；互联网与机构信息化综述也显示，多源异构数据并存而缺乏统一元数据和质量约束时，下游分析往往难以复现。

大语言模型在语义解析、指代消解和篇章组织方面具备较强通用能力，其技术基础可追溯至基于注意力机制的序列建模范式[3]。规模化预训练形成的迁移能力，为复杂文本处理提供了可用基础。临床笔记等非结构化文本研究也提示，面向领域的表示学习有助于提升抽取和填充任务稳定性。医学场景研究进一步强调，模型建议与人工决策之间应保持清晰分工和责任边界[4]。检索增强生成通过引入外部片段约束输出，可在一定程度上降低幻觉风险；在需要引用制度条文或个案既往材料时，可作为辅助模块按需部署。本文将 LLM 放在流调数据治理、信息抽取、报告成文和复核闭环中的关键位置，而不将其扩展

为泛化的一体化平台。

1.2. 研究不足

现有文献中，面向医疗工作流的大模型评测与系统综述多集中在对话、摘要或单一任务原型，对长周期、强合规业务链条中的工程约束讨论仍较分散[5]，例如字段口径归属、结构化结果与个案主键绑定、报告版本与审计日志对齐等问题。相关研究虽扩大了应用范围，但在可问责流程设计方面尚未形成统一做法[6]。临床工作流可用性研究提示，若缺少权限嵌入、过程记录和复核机制，对话式能力难以转化为可问责环节[7]。在流调场景中，尤其需要区分抽取准确率与法定或业务审签质量，避免以演示级准确率替代发布质量[8]。

1.3. 本文贡献

本文贡献可概括为四点。第一，提出面向非结构化流调文本的数据治理框架，覆盖接入登记、敏感分级、脱敏策略、个案主键对齐、版本血缘和质量标签，使文本处理结果可追溯、可问责。第二，给出信息抽取环节的 Schema 设计、提示与后处理、冲突检测和人工待办接口，强调规则与模型协同。第三，阐述报告生成与机构模板、段落槽位、证据片段绑定的工程方法，并通过状态管理区分草稿与可提交版本。第四，系统化描述人机协同机制，包括分级发布、双重复核、模型版本与提示版本留痕，并说明其与监测和信息流行病治理体系的对接关系。

2. 相关工作

2.1. 医学文本中的知识增强生成与结构化应用

在医学问答与指南场景中，检索增强生成(RAG)将检索片段嵌入提示以约束输出，相关研究在提示构造、医学事实对齐与知识更新节奏上积累了可迁移经验[9]。后续综述进一步说明，该路线在医学生成任务中的核心价值在于以外部证据约束模型自由生成[10]。面向指南与病例材料的实践强调证据可引用性与临床流程耦合[11]。临床场景大规模专家评测表明，证据筛选、冲突处置与机构口径一致往往依赖制度流程而非单次模型打分。其典型落点仍在问答、摘要与辅助阅读，对连续法定文书写作、模板嵌套及逐级签审链讨论有限。医学 RAG 多假设查询短、证据块相对独立，与流调跨段指代及时间线嵌套的张力须在架构上单列处理。对流调而言，RAG 适合在报告阶段按需启用，用于制度或类案的可追溯摘录，与基于个案主键与原文位置的抽取结果互补，而不宜被当作事实推断主体。

2.2. 大语言模型与非结构化流调数据治理

流行病学调查所形成的非结构化数据，常见形式包括访谈记录、通讯转写、现场工作笔记及多媒体配文等，多呈现来源多元、口语化程度高、时间信息隐含与指代跨段等特征[12]。流调数据治理强调在可存储、可共享之外，将文本与个案或事件标识、采集责任、敏感信息分级、脱敏规则、版本血缘与数据质量标签一并管理，以保障后续信息抽取与报告编制的溯源与审计。大语言模型可辅助完成敏感片段初筛、表述归一、缺项与冲突提示、低质文本预警等任务，其输出宜作为待复核建议进入人机协同流程；对法定要素认定、涉他人信息及跨机构共享范围等高风险事项，仍应以制度规则与人工确认为准。上述技术路线在医疗文本场景的系统评测中也体现出共同结论：规则校验与人工审阅是确保可发布质量的必要环节。需要指出的是，流调材料在协查对象、舆情要素与执法表述等方面与住院病历并不等同，借鉴相关结论时应同步调整规则库、提示模板与抽检方案，避免不加论证地套用临床场景的缺省配置。

2.3. 大语言模型与疾控报告生成

在传染病监测预警和突发公共卫生事件处置链条中，流行病学个案材料、事件通报和互联网讨论常

同时出现，文本类型复杂且更新频繁。针对事件监测非结构化数据源(如 ProMED、WHO Disease Outbreak News)的研究，已系统比较多种大语言模型在病例要素和地理实体抽取任务中的表现，并尝试通过多模型集成提升稳定性，为从开放叙述恢复报送字段提供了方法参考。英国卫生安全机构等团队评测显示，在疾病负担、危险因素与干预相关任务中，先进开放权重模型可接近主流闭源模型，但任务间性能差异明显，因此提示落地时需配套提示设计、规则校验和机构内再校准。针对疫苗接种、个人防护和心理健康等议题的舆情监测研究也提示，模型应用应与业务口径和隐私合规同步设计，并纳入信息流行病治理视角。综合现有研究，对流调与疾控报告生成更具操作性的做法是：以模板槽位绑定文书结构，以已校验结构化字段约束叙述段落，区分自动生成指标和可签审业务质量，并在发布前落实人机协同复核和版本留痕。

2.4. 小结与本文定位

检索增强生成多用于短问答和片段引用，在流调成文阶段可用于补充制度或类案依据，但难以单独从长篇叙述中还原事实主线。流调文本与病历在来源和口语表达上并不一致，因此仍需将个案标识、脱敏处理、版本管理和质量标注与人工复核稳定衔接。公卫与疾控文本的信息抽取及舆情研究也表明，指标提升并不等于可签审成品，仍离不开模板字段约束和规则校验。本文围绕流调零散文本提出分步实施路径，依次为整理脱敏、信息抽取、报告起草和人工定稿，并补充个案绑定、变更留痕和模型及提示词版本登记要求，以支撑复核与签审。

3. 非结构化流调数据治理与报告生成系统设计

3.1. 整体架构

在突发公共卫生事件与常态化传染病防控并行的业务环境下，流调资料普遍存在来源分散、文本口语化、时间线隐含和跨部门口径不一致等特点，进而带来信息整理慢、报告产出慢、复核压力大和责任追溯难等问题。不同岗位对系统能力的关注点也存在差异：采集人员关注录入和脱敏效率，审核人员关注字段冲突和证据定位，撰写人员关注模板一致性和成文速度，签审人员关注可追溯和可问责。为同时覆盖这些需求，本研究采用以个案为主线的分层架构，并遵循先治理、后抽取、再成文、全程留痕的组织原则。

具体而言，系统由五个协同层构成：第一层为数据源与接入层。第二层为数据治理层，负责非结构化流调文本的接入登记、敏感信息分级脱敏、质量标注与版本管理。第三层为信息抽取层，由大语言模型与规则引擎联合完成结构化字段抽取、冲突检测与缺失提示。第四层为报告生成层，在机构模板约束下将已校验字段与证据片段组织为报告草稿，并执行格式与一致性检查。第五层为人机协同与审计层，通过角色权限、任务队列、签审状态机与全流程日志，保证关键操作可复核、可回溯、可问责(见图 1)。

3.2. 技术实现

在技术实现上，系统采用语言模型能力、结构化规则约束与人工审核闭环协同的实现路线，并在数据、模型、知识、服务和运维五个层面联动。整体技术流程覆盖数据接入与治理、信息抽取与校验、报告生成、流程控制与审计等主链路环节。

首先，在多源数据接入与预处理环节，系统支持 API 推送、批量文件导入和消息队列三种接入方式，统一接收访谈纪要、电话录音转写、即时通讯记录、现场图片文字等材料。对非文本输入分别采用 OCR 与 ASR 转写为标准文本，随后完成编码统一、句段切分、哈希去重、时间标准化和地名规范化。数据入库前写入来源、采集时间、采集人员、区域、版本号等元数据并绑定个案主键，保证跨轮次信息可追踪(见图 2)。

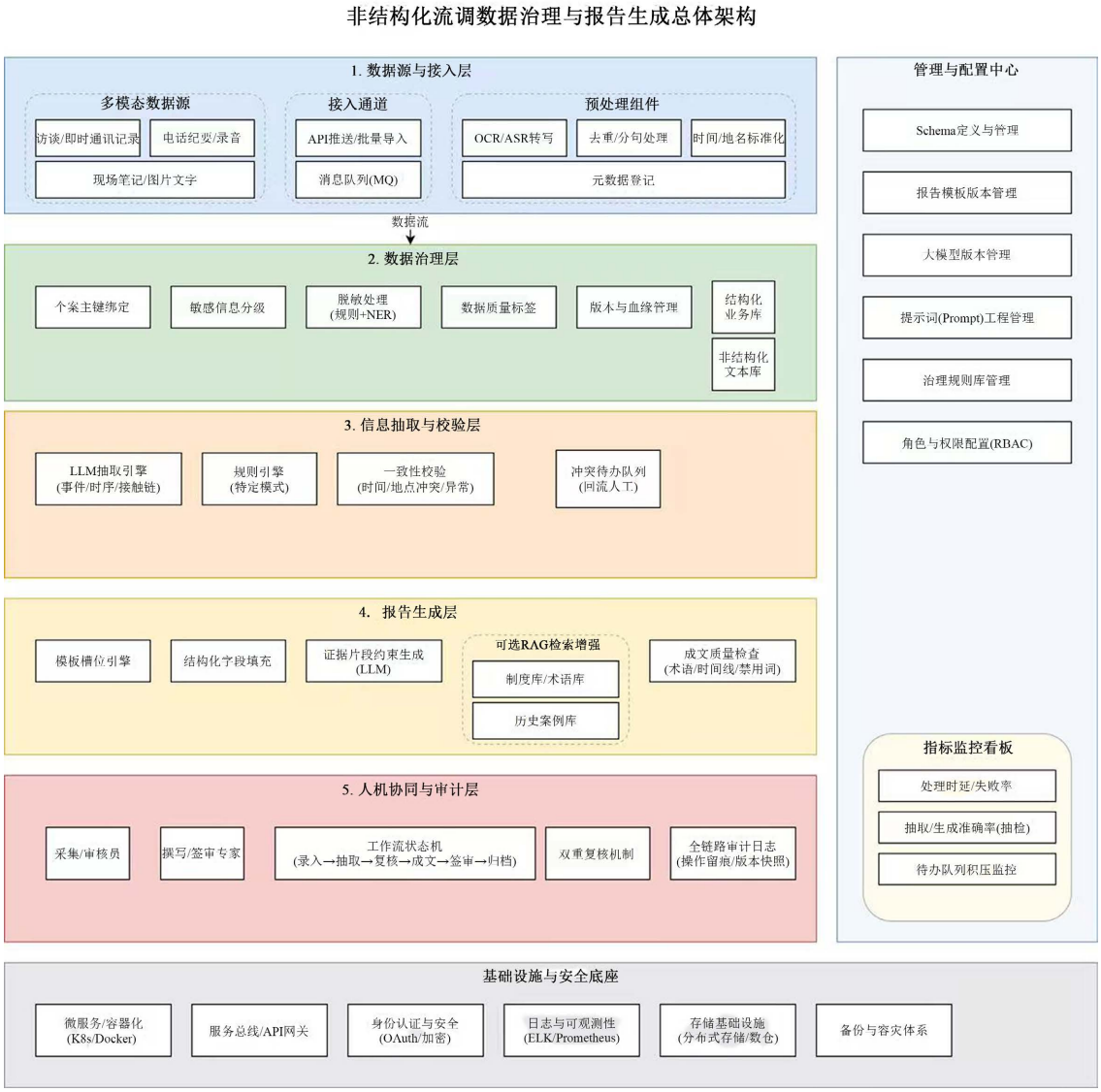

Figure 1. Overall architecture of unstructured epidemiological investigation data governance and report generation
图 1. 非结构化流调数据治理与报告生成总体架构

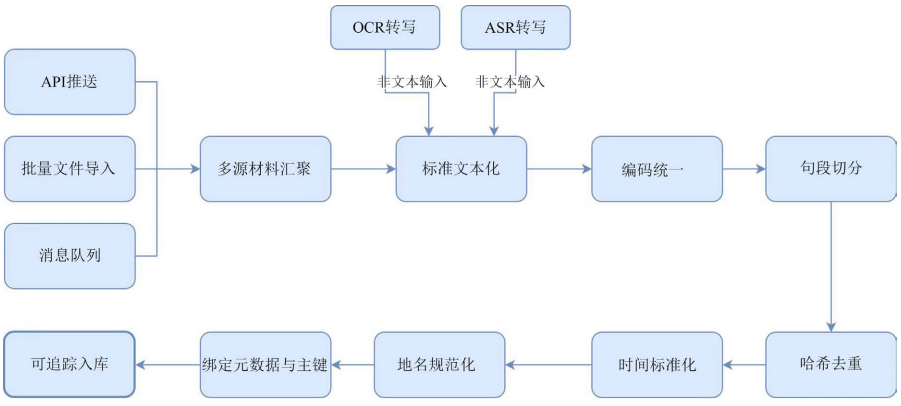

Figure 2. Data flow of multi-source data ingestion and preprocessing
图 2. 多源数据接入与预处理数据流程图

其次，在隐私保护与数据治理技术环节，脱敏采用三级处理策略，包括规则识别、命名实体识别和人工抽检。系统先识别证件号、手机号等确定性较高的字段，再补充识别姓名、地址、机构等依赖上下文判断的敏感实体，最后由审核端进行抽检确认。数据存储阶段采用字段级加密与令牌化处理，模型输入阶段采用按需掩码和最小必要字段输入，在保障安全性的同时兼顾可用性。

再次，在信息抽取与结构化校验环节，抽取模块通过提示设计、函数调用和 JSON Schema 约束输出结构化结果，核心任务包括事件要素抽取、时序关系识别、接触链关系识别和风险标签判定。后处理阶段引入时间解析器、术语码表映射、规则引擎和一致性校验器，对时间顺序矛盾、地点冲突、人数不一致、字段缺失等问题自动标记并生成待办任务。对于表达复杂的文本场景，系统采用两阶段抽取方式，即先粗抽取再精修订，以提升结果稳定性。

随后，在知识增强与报告生成环节，报告模块由模板槽位结构驱动，确定性字段优先使用已校验的结构化数据进行填充，叙事段落在证据片段约束下由模型生成。为降低幻觉风险，系统可选用 RAG 流程，先在制度库、术语库和历史案例库中检索相关内容，再由生成模块综合输出。生成完成后，系统继续执行占位符检查、术语规范检查、时间线一致性检查和禁用词检查，未通过项将回流到复核队列。

最后，在流程编排与人机协同环节，系统采用工作流引擎和状态机管理录入、抽取、复核、成文、签审和归档全流程，并配套 RBAC 权限模型、任务队列、超时升级和回退机制。制度层面明确，大语言模型仅承担证据整理、候选信息抽取和文稿起草等辅助职责，不承担法定要素认定和处置决策职责。每次人工订正都会记录差异变更，并回写至提示样例库或规则库，形成模型输出、人工修正与策略更新的持续迭代闭环。

4. 非结构化流调数据治理与报告生成系统的应用案例

本研究以非结构化流调文本入库、结构化抽取、报告草稿生成和签审归档为主流程，构建原型系统并开展案例应用。考虑到疾控场景对权限、审计和数据安全的要求，系统采用私有化部署，模型服务、规则引擎、模板服务和审计模块通过接口解耦，便于在不改变主流程的前提下替换模型或增减规则。系统前端按岗位划分为采集端、审核端、撰写端和签审端，后端统一记录模型版本、提示词版本、字段订正记录和签审日志，以支持事后复盘和责任追踪。

在应用流程上，系统首先接收访谈纪要、电话记录和通讯转写等文本，完成登记、脱敏和主键绑定；随后调用抽取模块生成结构化字段，并通过规则校验输出冲突待办；审核人员完成字段订正后，报告模块按模板槽位生成草稿，最终进入签审与归档环节。该流程覆盖个案流调报告的主要生产步骤，可将原始叙述、结构化要素和规范文书串联为连续作业链。

5. 讨论：价值、局限与扩展

5.1. 预期价值

系统的首要价值体现在效率提升和质量可控两个方面。其一，系统将分散文本的接入、抽取、成文和签审衔接为连续流程，可减少跨系统搬运、重复录入和反复校对的时间消耗。其二，结构化字段与模板槽位的协同约束有助于降低文书口径波动，提升报告一致性和可读性。其三，版本管理与审计日志可完整保留原始文本、抽取结果、报告草稿和签审定稿等关键节点，使结果具备生成能力、复盘能力和问责能力。

5.2. 局限性

当前方案仍存在四类局限。第一，语言理解边界：面对罕见表达、多方言口语、跨段省略与跨文档

指代时，模型仍可能出现要素遗漏或关系误判。第二，规则与模板维护成本：机构模板、码表与签审口径若频繁调整，会同步带来槽位配置、规则更新与回归测试压力。第三，工程与算力成本：在高峰期并发调用、长文本处理和审计留痕并行运行时，对算力、存储与网络稳定性提出更高要求。第四，数据与治理风险：即使完成脱敏，外部数据接入、跨系统对接和人员操作仍可能引入权限越界、误用误传和责任边界不清等问题。

5.3. 关联系统与未来工作

围绕本文主链路，后续扩展可从能力外延、评估深化和治理完善三个方面推进。能力外延方面，监测预警、疾控知识库 RAG、多语言协作、对话式指标查询和外网情报辅助可作为独立模块接入。评估深化方面，建议在更多地区、病种和文本类型上开展分层验证，形成跨场景可比的评测基线，并持续跟踪字段级准确率、签审通过率和人工修订负荷等核心指标。治理完善方面，可进一步探索多模态数据统一治理、跨区域平台互联标准、联邦审计和最小必要共享机制，在确保隐私与合规前提下提升系统协同能力和可推广性。

6. 结论与展望

本文针对非结构化流调数据，提出以 LLM 为核心的数据治理、信息抽取、报告生成与人机协同一体化方案，细化各环节可操作要素与状态划分，并明确监测、问答、可视化等能力的关联定位，突出系统落地的可实施性。后续工作将结合具体机构模板与真实脱敏语料，在严格伦理审查下开展迭代评测与提示治理。

基金项目

广西民族师范学院校级科研项目《基于大语言模型的多模态数据融合研究》(项目编号：2024QN055)。

参考文献

[1] Gautam, A.S. and Raza, Z. (2024) Disease Outbreak Prediction Using Natural Language Processing: A Review. *Knowledge and Information Systems*, **66**, 6561-6595.

[2] McClymont, H., Lambert, S.B., Barr, I., Vardoulakis, S., Bambrick, H. and Hu, W. (2024) Internet-Based Surveillance Systems and Infectious Diseases Prediction: An Updated Review of the Last 10 Years and Lessons from the COVID-19 Pandemic. *Journal of Epidemiology and Global Health*, **14**, 645-657. https://doi.org/10.1007/s44197-024-00272-y

[3] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. https://arxiv.org/abs/1706.03762

[4] Lee, P., Bubeck, S. and Petro, J. (2023) Benefits, Limits, and Risks of GPT-4 as an AI Chatbot for Medicine. *New England Journal of Medicine*, **388**, 1233-1239. https://doi.org/10.1056/nejmsr2214184

[5] Bedi, A., Nadkarni, P.M., Somai, M., *et al*. (2024) Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA Network Open*, **7**, e2440819.

[6] Busch, F., Hoffmann, L., Rueger, C., van Dijk, E.H., Kader, R., Ortiz-Prado, E., *et al*. (2025) Current Applications and Challenges in Large Language Models for Patient Care: A Systematic Review. *Communications Medicine*, **5**, Article No. 26. https://doi.org/10.1038/s43856-024-00717-2

[7] Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A.K., *et al*. (2023) Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *Journal of Medical Internet Research*, **25**, e48659. https://doi.org/10.2196/48659

[8] 世界卫生组织. 信息流行病[EB/OL]. https://www.who.int/health-topics/infodemic, 2026-03-28.

[9] He, Z., Wu, J., Peng, A., *et al*. (2023) MKRAG: Medical Knowledge Retrieval Augmented Generation for Medical Question Answering. https://arxiv.org/abs/2309.16035

[10] Xing, Z., Ye, C., Han, T., *et al*. (2024) Retrieval-Augmented Generation for Generative Artificial Intelligence in Medicine. https://arxiv.org/abs/2406.12449

[11] Zakka, C., Chaurasia, A., Shad, R., Dalal, A.R., Kim, J.L., Moor, M., *et al*. (2023) Almanac: Retrieval-Augmented Language

Models for Clinical Medicine. arXiv:2303.01229.

[12] Consoli, S., Markov, P., Stilianakis, N.I., Bertolini, L., Gallardo, A.P. and Ceresa, M. (2024) Epidemic Information Extraction for Event-Based Surveillance Using Large Language Models. In: Yang, X.S., Sherratt, S., Dey, N. and, Joshi, A., Eds., *Lecture Notes in Networks and Systems*, Springer, 241-252. https://doi.org/10.1007/978-981-97-4581-4_17