位置：使用大型语言模型对非结构化临床记录进行主题分析

Seungjun Yi[1], Joakim Nguyen[2], Terence Lim[3], Andrew Well[4], Joseph Skrovan[5], Mehak Beri[5], YongGeon Lee[5], Kavita Radhakrishnan[6], Liu Leqi[7], Mia Markey[1], Ying Ding[5,8]

[1]Department of Biomedical Engineering, The University of Texas at Austin

[2]Department of Computer Science, The University of Texas at Austin

[3]College of Natural Sciences, The University of Texas at Austin

[4]Department of Cardiac Surgery, Vanderbilt University School of Medicine

[5]School of Information, The University of Texas at Austin

[6]School of Nursing, The University of Texas at Austin

[7]IROM, McCombs School of Business, The University of Texas at Austin

[8]Dell Medical School, The University of Texas at Austin

{charlie.yi,jhn001, terence.lim}@utexas.edu, andrew.well@vumc.org, jskrovan@fastmail.fm, {mberi, yg910524}@utexas.edu, kradhakrishnan@mail.nur.utexas.edu, {leqiliu, mia.markey}@utexas.edu, ying.ding@ischool.utexas.edu

Abstract

这份立场文件探讨了大型语言模型（LLMs）如何支持对非结构化临床记录的主题分析，这是一种广泛使用但资源密集的方法，用于揭示患者和提供者叙述中的模式。我们系统地审查了最近将 LLMs 应用于主题分析的研究，并辅以与一名执业医师的访谈。我们的研究发现，当前方法在多个维度上仍然存在碎片化现象，包括主题分析类型、数据集、提示策略以及所使用模型，尤其是在评估方面最为明显。现有的评估方法差异很大（从定性的专家评审到自动相似度指标），阻碍了进展并妨碍了不同研究之间的有意义基准比较。我们主张建立标准化的评估实践对于推动该领域的发展至关重要。为此，我们提出了一种以三个维度为中心的评估框架：有效性、可靠性和可解释性。

1 介绍

主题分析（TA）是定性数据分析中最常用的方法之一，提供了识别、分析和报告文本数据中模式（主题[1]）的结构化但灵活框架 [1, 2]。在临床环境中，TA 经常应用于非结构化的患者访谈记录，以揭示潜在意义并生成可用于改善护理的见解 [3]，该过程通常由 Braun 和 Clarke 的六阶段框架指导 [1]。研究人员首先通过反复阅读和识别重要关键词（步骤 1-2）来熟悉数

[1]主题: 从将相关代码分组中浮现的更广泛的模式或类别，以捕捉参与者在与研究目标相关的经历或观点中的重要方面

Preprint. Under review.

据。然后使用这些关键词生成初始代码[2]（步骤 3），作为识别初步主题的基础（步骤 4）。主题经过迭代审查、完善和清晰定义和命名，通过解释性分析（第 5 步）。在最后阶段，开发了一个结构化的概念模型来表示主题关系和潜在结构（第 6 步）。为了确保可信度，至少两名具有领域专业知识的受过培训的研究人员必须独立地与数据互动并参与多轮讨论以达成共识。每年，超过 90 万美国医疗保健访谈在学术、临床和研究环境中进行主题分析 [4, 5]。大多数涉及非结构化数据，需要归纳主题分析 (ITA)，它直接从记录中提取主题，而无需预定义的代码 [6, 7, 8] 。手动 ITA 每年需要超过 610 万小时，相当于 3000 个全职工作和 3.054 亿美元的成本，这限制了可扩展性并延迟了在快速变化的医疗保健环境中的见解 [9, 10, 11]。因此，自动化 ITA 对于及时、可扩展和可重复的主题生成至关重要。

大型语言模型（LLMs）的最新进展为自动化六步手动归纳主题分析过程提供了有前景的选择 ~ [12, 13]。LLMs 可以在不到 10 分钟的时间内快速生成初始代码（第 2 步）和候选主题（第 3 – 5 步），与人类分析师所需的 5 – 8 小时相比，这是一个巨大的减少 ~ [14, 15]。

然而，将大语言模型完全自动化地整合到助教过程中仍处于初期阶段。现有研究主要集中在主题分析的具体阶段上，例如编码 [16]，以及主题生成 [17, 18]。它们在方法选择上的差异也很大，包括所采用的主题分析类型（即归纳性与演绎性）、使用的模型（如 `gpt-4o`、`llama`）和采取的提示策略（例如少样本提示、思维链），这使得难以建立起一个连贯的进步图景。这一挑战因大语言模型开发速度迅猛而进一步加剧，这不断重塑了方法论景观。更重要的是，评估主题分析的结果仍然是一个主要难题。评估实践仍不一致，主要依赖专家的定性评估和大语言模型生成与人类生成的主题之间的比较，这限制了研究结果的可重复性和可比性。

在这项工作中，我们的目标是：

1. 提供大型语言模型（LLMs）支持的主题分析方面的最新发展和研究进展概述。
2. **强调基于大语言模型的主题分析中标准化评估的必要性，并提出一个基于有效性、可靠性和可解释性的评估框架**。
3. 强调需要端到端框架以确保主题分析在临床环境中的实际应用价值。

2 方法

文献搜索与选择 我们的数据收集针对的是与大型语言模型（LLMs）相关的主题分析（TA）的研究论文。通过使用 arXiv 元数据（截至：2025 – 08 – 15），我们通过对标题和摘要进行基于关键词的搜索来识别相关论文。总结统计信息见表 1，完整搜索详情请参阅附录 C。为解决一些同行评审论文可能未在 arXiv 中收录的问题，我们在 Elicit[19] 和主要索引平台的手动查询中进行了补充（详见附录 E）。完整的程序、验证步骤、调查提示和关键词列表见附录 D 表 3。从收集的文献中，我们回顾了最近三年（截至日期前三年）应用 LLMs 进行主题分析的 56 项研究，重点关注五个维度：分析类型、模型、数据集、提示策略和评估方法。我们包括同行评审出版物、预印本，但排除论文、学位论文和未完成的工作。

专家访谈 除了系统的文献综述外，我们还对一位拥有八年临床经验的心脏外科医生进行了深入访谈，该医生之前曾对患者记录进行手动主题分析 [20]。访谈通过为期两小时的 Zoom 会议进行，在此期间，医生获得了文献综述结果的总结，并被邀请讨论其影响以及 LLM 在增强主题分析工作流程中的更广泛作用。本次讨论的详细见解在第 4 节中呈现。

[2] 代码：简洁的标签，用于识别与研究问题相关的文本数据特征

3 结果

3.1 大型语言模型支持的主题分析（TA）的趋势和数据集特征

从表 1 中，我们观察到利用大语言模型的 TA 是一个相对较新的现象，第一项工作出现在 2022 年 11 月 GPT-3.5 发布之后 [21]。自那时以来，数量呈指数级增长（图 1）。截至 2025 年 8 月，在 arXiv 元数据中总共识别出 46 篇论文，其中只有三篇与医疗保健相关。

表 1: 包含与主题分析和大型语言模型相关的关键词的检索论文总结。额外的 “(+11)” 表示在 arXiv 之外识别出的论文。关键词列表定义见附录表 3。

关键词分类	计数
Thematic Analysis (核心_ta)	381
Inductive Thematic Analysis (归纳_塔)	14
Deductive Thematic Analysis (演绎_ta)	3
Thematic Analysis (核心_ta) + LLMs (通用 llm)	45 (+11)
Thematic Analysis (核心_ta) + LLMs (通用 (llm_generic 不做翻译)) + Healthcare (医疗卫生)	3
Total TA (any category above)	428

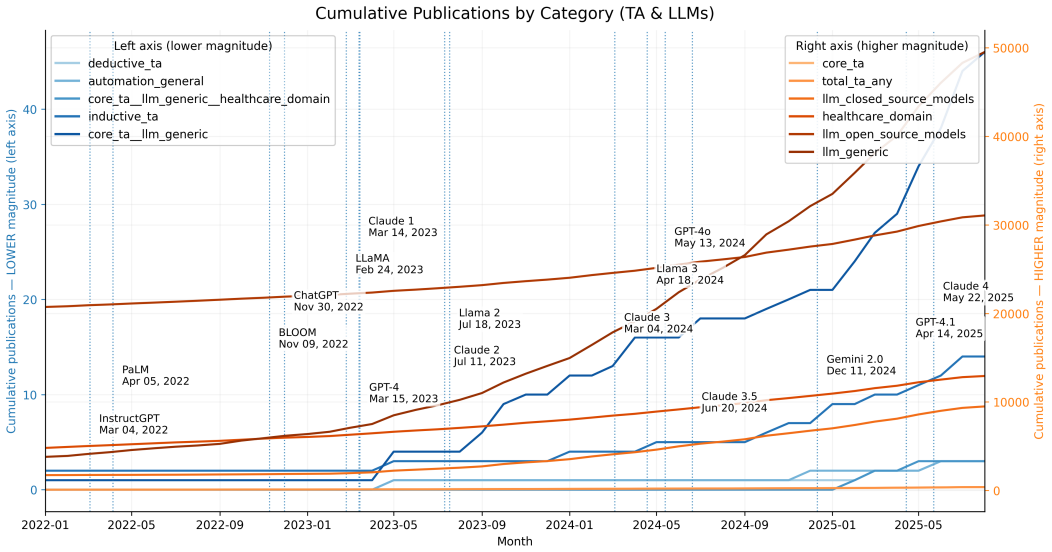

图 1: 自 2022 年 1 月以来按类别累计提及主题分析（TA）和大语言模型（LLMs）的出版物数量。明确结合 TA 与 LLMs 的研究在 GPT-4 发布后开始出现。垂直线表示 LLM 开发的主要里程碑（见附录 B）。所示类别对应表 1 中总结的类别。

3.2 LLM 辅助主题分析（TA）的调查

我们探讨了 LLM 辅助 TA 的五个关键维度，包括采用的 TA 类型、利用的模型、分析的数据集、所用的提示策略和所应用的评估方法（表 2）。

主题分析类型：归纳法与演绎法。 演绎以先验代码本为基础，提供了更清晰的定义和可比性，而归纳的 TA 则直接从记录中生成主题，从而能够得出新颖见解。在 56 项研究中，归纳 TA 最为普遍，占 36 篇论文（64%）。结合归纳和演绎元素的混合方法是第二常见的（22%），

表 2: 跨越维度的 56 篇论文（2022 年 8 月-2025 年 8 月）频率总结。由于单篇论文可能涵盖多个类别，计数可能会超过 56。被认为不适用的类别被排除在表格之外。

Dimension	**Attributes (# out of 56 / %)**
Types of TA	Inductive (36/64%), Few-shot (9/16%), Hybrid (12/22%), Other (3/5%)
Models (by family)	GPT (32/58%), Claude (7/13%), LLaMA (6/11%), Gemini (6/11%), Mistral/Mixtral (3/5%), DeepSeek (4/7%), Qwen (2/4%), Other/Custom (7/13%)
Datasets by Domain	Social Media/Online Communities (14/25%), Education (12/21%), Software Engineering/Programming (11/20%), Arts/Humanities/Other (10/18%), Healthcare/Clinical (9/16%)
Prompting	Zero-shot (19/35%), Few-shot (9/16%), Chain-of-thought (7/13%), Self-consistency/Reflexion (8/15%), Tool/Agent use (7/11%)
Evaluation	Human qualitative review (22/40%), Automatic text metrics (16/27%), Task-based/utility eval (7/13%), Hybrid (11/20%)

而纯粹的演绎分析相对罕见（9%）。一小部分工作（5%）采用了其他或不太明确的策略。*演绎和归纳的* TA 需要不同的评估方法。因此，为一种方法开发的评估方法很难转移到另一种上，这导致了评估实践中的持续不一致性。

所用模型。 在调查的文献中，大多数研究依赖于 OpenAI GPT 系列，GPT-3.5 [22] 和 GPT-4 [23] 变体作为主要分析引擎。较少的研究纳入了 Claude [24] 或 Gemini [25] 模型，通常作为比较基准而非中心分析工具。开源系列如 LLaMA [26]、Mistral [27] 和 DeepSeek [28] 出现在探索性或集成配置中，反映了对透明度和成本效益替代方案日益增长的兴趣。仅有少数论文研究了微调或专门化的部署（例如，InCoder [29], CodeGen [30], ProgGP [31]），通常用于特定的用例。按模型系列的详细分解见附录 F。

按领域划分的数据集。 被调查的研究涉及多个领域，其中社交媒体和在线社区 [32, 33, 34] 是最常见的（25%），其次是教育 [35, 36]（21%），软件工程和编程 [35, 37]（20%），以及医疗或临床环境 [13, 15, 38]（16%）。其余的 18%是艺术、人文学科和其他领域，如法庭案例、纪律处分决定和设计工作坊。

提示策略。 研究最常依赖零样本提示（即，没有具体示例给出），通常通过多阶段管道实现，较小的比例使用少量样本模板、链式思维推理或迭代改进方法，如自洽性和反射。少数探索了多代理或多工具的提示方式，而有些则完全不使用提示，而是专注于微调或人工编码的方法。总体而言，零样本提示因其可访问性以及与归纳主题分析的契合度而占主导地位，而更复杂的提示设计反映了试图稳定输出或近似协作编码实践的努力。在附录 G 中提供了更详细的分解。此外，大多数现有方法仍然依赖于需要全面审查转录的人类参与的工作流，这严重限制了可扩展性。由于熟悉（Braun & Clarke 的六步主题分析过程的第一步 [1]）转录是最耗时的步骤，因此保留这一要求对于临床环境中 LLM 辅助的主题分析的实际效用至关重要。

评估。 LLM 辅助的主题分析评估仍处于分散状态，没有普遍接受的黄金标准。人类定性审查是最常见的方法（$\approx 40\%$），包括编码者一致性的检查、反思讨论以及专家对合理性和主题覆盖率的评价，尽管再现性往往有限。通常使用克利彭多夫阿尔法系数 [39] 来评估编码者一致性，这是一种考虑偶然一致性的可靠性系数，支持不同测量水平，并处理不完整数据，或者使用科恩卡帕系数 [40]，后者在假设完全数据和相等类别分布的情况下衡量两个评价者之间的协议。除了编码者一致性之外的其他标准是通过调查来衡量的，在这些调查中，专家们在一个 1-5 李克特量表 [15, 16] 上进行评分。大约四分之一的研究（27%）使用自动相似度

指标将 LLM 生成的主题与人类基准进行了对比评估。词汇重叠度量，如 Jaccard 指数、模糊匹配、ROUGE [41]、BLEU [42]、GLUE [43] 和 METEOR [44] 基于共享的单词或 n-grams[3] 来评估相似度。语义相似度量包括余弦相似度、BERTScore [45]，比较向量嵌入 [4]。分类器基础的准确率和 F1 分数偶尔被应用。一个较小的子集（13%）通过基于任务的措施（例如，学习成果）间接评估输出，而大约五分之一（20%）采用了结合上述人类判断与计算指标的混合设计 [16]。

然而，评估方面仍存在主要挑战。即使报告了相同的指标，基础嵌入的差异也阻碍了研究之间的可比性和综合分析。透明度问题进一步加剧了这些挑战：被调查的工作没有披露完整的 ground truth，常常以隐私或 IRB 限制为由，而其他一些工作则未明确解释为何不提供此信息。对于自动评估指标的使用，直接评估人类与大语言模型生成主题之间相似性的主要局限性在于缺乏一对一映射。主题分析本质上会产生部分重叠的主题集：一个由大语言模型生成的主题可能对应多个人类主题，而某些人类主题则可能没有直接对应的。

为了解决这些限制，我们建议从三个关键维度评估自动化归纳主题分析：**有效性**、**可靠性**和**可解释性**。(1) **有效性**。任何一组生成的主题可能无法在人类和 LLMs 之间保持 1:1 的映射，因此我们在相似性矩阵 $\boldsymbol{S}$ 上计算**最大权二分匹配**，其中 S_{ij} 表示人类主题 i 与模型主题 j 之间的相似度。使用以下方法：(a) 词汇重叠：Jaccard；可选地使用 ROUGE 进行 n-gram 重叠。(b) 语义相似度：句子嵌入或 BERTScore / MoverScore / BLEURT [45] 的余弦值，以及报告：Precision/Recall/F1@match（匹配精度/召回率/F1），Coverage@τ（超过阈值的人类主题匹配比例），冗余（平均内部 LLM 主题相似度）和新颖率（没有对应人类主题的 LLM 主题）。(2) **可靠性**。Krippendorff's α 或 Cohen/Fleissκ 可以用作诊断模糊定义或不清晰编码边界的指标。为了测试稳定性，重新运行管道的不同部分并使用不同的随机种子或引导样本，并计算代码分配的调整兰德指数 (ARI) 或信息变异 (VI)。在主题层面上，可以通过比较支持摘录中的重叠来对系统生成的主题与人类主题进行对齐。当将大型语言模型应用于管道时，可确认性评估主题是数据驱动还是由不同大型语言模型中固有的内部偏见所驱动的。(3) **可解释性**。分析深度和细微差别仍然具有挑战性 [46]。缓解措施包括生成理由 [47]，专家裁定 [15] 和自适应代码本 [33]。同一主题内的摘录应具有一致的意义，而不同的主题应该是可区分的。嵌入相似性可以检查定义为主题内引用之间的平均相似性的一致性以及作为主题质心之间距离的区分度 [48]。报告（i）分配给任何主题的段落和（ii）每个主题所代表的参与者的覆盖范围。主题的可信度评估生成的主题是否忠实于数据的表示。对于特定领域的上下文，将自动化指标与人工验证结合使用 [49, 34]。

其他考虑因素：主题分析的成本。 使用 LLMs 的成本已经变得显著更具可行性。最近的分析估计推理价格约为每百万输入标记 0.15 至 2.50 美元，每百万输出标记 0.60 至 10 美元，相比之下，手动主题分析通常需要 200,000+美元。[50, 51]

4 讨论与结论

我们的研究结果表明，当前基于大语言模型的主题分析方法仍然支离破碎，特别是在不同研究中使用的评估方法方面。正如我们研究访谈中的临床医生所观察到的，在缺乏明确的基本事实或普遍接受的标准的情况下，对归纳主题分析进行评估可能会觉得像“盲人摸象”一样。当前不同的评估实践阻碍了跨研究的直接比较。没有这样的标准化，研究工作仍然支离破碎，并且使用大语言模型开发端到端自动化的主题分析框架受到了阻碍。

[3] 一个或多个连续单词的重叠序列。

[4] 捕捉语义意义的文本数值表示，由诸如 `all-mpnet-base-v2` 和文本嵌入-3-大型等模型生成。

总结而言，虽然 LLMs 在变革主题分析方面具有巨大潜力，但如果没有统一的评估标准，进展仍将支离破碎。通过推进有效性、可靠性和可解释性作为共享维度，该领域可以朝着更严格、可扩展和临床意义重大的自动化定性研究应用迈进。

参考文献

[1] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[2] Virginia Clarke and Virginia Braun. Teaching thematic analysis: Overcoming challenges and developing strategies for effective learning. *The Psychologist*, 26(2):120–123, 2013.

[3] Sirwan Khalid Ahmed, Ribwar Arsalan Mohammed, Abdulqadir J. Nashwan, Radhwan Hussein Ibrahim, Araz Qadir Abdalla, Barzan Mohammed M. Ameen, and Renas Mohammed Khdhir. Using thematic analysis in qualitative research. *Journal of Medicine, Surgery, and Public Health*, 6:100198, 2025.

[4] Melissa DeJonckheere and Lisa M Vaughn. Semistructured interviewing in primary care research: a balance of relationship and rigour. *Family Medicine and Community Health*, 7:e000057, 2019.

[5] Konstantina Vasileiou, Julie Barnett, Susan Thorpe, and Terry Young. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC Medical Research Methodology*, 18(1):148, Nov 21 2018.

[6] Kaitlyn A. Campbell, Elizabeth Orr, Pauline Durepos, Linh Nguyen, Lanting Li, Caitlin Whitmore, Paula Gehrke, Lorraine Graham, and Susan M. Jack. Reflexive thematic analysis for applied qualitative health research. *The Qualitative Report*, 26(6):2011–2028, 2021.

[7] Bethan Taylor, Catherine Henshall, Sara Kenyon, Ian Litchfield, and Sheila Greenfield. Can rapid approaches to qualitative analysis deliver timely, valid findings to clinical leaders? a mixed methods study comparing rapid and thematic analysis. *BMJ Open*, 8(10):e019993, 2018. Published 2018 Oct 8.

[8] Daphne C Watkins. Qualitative research: The importance of conducting research that doesn't "count". *Health Promotion Practice*, 18(5):601–602, 2017.

[9] Emily Namey, Greg Guest, Lucy Thairu, and Lauri Johnson. Data reduction techniques for large qualitative data sets. *Handbook for Team-Based Qualitative Research*, pages 137–161, 2008.

[10] Hamza Alshenqeeti. Interviewing as a data collection method: A critical review. *English Linguistics Research*, 3, 01 2014.

[11] Monique M. Hennink, Bonnie N. Kaiser, and Vincent C. Marconi. Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research*, 27(4):591–608, March 2017. Epub 2016 Sep 26.

[12] Seungjun Yi, Jaeyoung Lim, and Juyong Yoon. Protomed-llm: An automatic evaluation framework for large language models in medical protocol formulation, 2025.

[13] Seungjun Yi, Joakim Nguyen, Huimin Xu, Terence Lim, Andrew Well, Mia Markey, and Ying Ding. Auto-TA: Towards scalable automated thematic analysis (TA) via multi-agent large language models with reinforcement learning. In *ACL 2025 Student Research Workshop*, 2025.

[14] Huimin Xu, Seungjun Yi, Terence Lim, Jiawei Xu, Andrew Well, Carlos Mery, Aidong Zhang, Yuji Zhang, Heng Ji, Keshav Pingali, Yan Leng, and Ying Ding. Tama: A human-ai collaborative thematic analysis framework using multi-agent llms for clinical interviews, 2025.

[15] Muhammad Zain Raza, Jiawei Xu, Terence Lim, Lily Boddy, Carlos M. Mery, Andrew Well, and Ying Ding. Llm-ta: An llm-enhanced thematic analysis pipeline for transcripts from parents of children with congenital heart disease, 2025.

[16] Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6456–6469, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[17] Insa Mannstadt, Susan M. Goodman, Mangala Rajan, Sarah R. Young, Fei Wang, Iris Navarro-Millán, and Bella Mehta. A novel approach for mixed-methods research using large language models: A report using patients' perspectives on barriers to arthroplasty. *ACR Open Rheumatology*, 6(6):375–379, 2024.

[18] M Deiner, V Honcharov, J Li, T Mackey, T Porco, and U Sarkar. Large language models can enable inductive thematic analysis of a social media corpus in a single prompt: Human validation study. *JMIR Infodemiology*, 4:e59641, 2024.

[19] Janice Y. Kung. Elicit. *Journal of the Canadian Health Libraries Association*, 44(1):15–18, April 2023.

[20] C.M. Mery, A. Well, K. Taylor, K. Carberry, J. Colucci, C. Ulack, A. Zeiner, M. Mizrahi, E. Stewart, C. Dillingham, T. Cook, A. Hartounian, E. McCullum, J.T. Affolter, H. Van Diest, A. Lamari-Fisher, S. Chang, S. Wallace, E. Teisberg, and C.D. Jr Fraser. Examining the real-life journey of individuals and families affected by single-ventricle congenital heart disease. *Journal of the American Heart Association*, 12(5):e027556, March 2023. Epub 2023 Feb 21.

[21] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis, 2023.

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[23] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report, 2024.

[24] Anthropic. Introducing claude. `https://www.anthropic.com/index/introducing-claude`, 2023.

[25] Google DeepMind. Gemini: A family of highly capable multimodal models. `https://deepmind.google/technologies/gemini/`, 2023.

[26] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[27] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[28] DeepSeek AI. Deepseek llms. `https://github.com/deepseek-ai`, 2025.

[29] Daniel Fried, Armen Aghajanyan, Jessy Lin, Sida Wang, Eric Wallace, Freda Shi, Ruiqi Zhong, Wen tau Yih, Luke Zettlemoyer, and Mike Lewis. Incoder: A generative model for code infilling and synthesis, 2023.

[30] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis, 2023.

[31] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models, 2021.

[32] Mohammed-Khalil Ghali, Abdelrahman Farrag, Sarah Lam, and Daehan Won. Beyondwords is all you need: Agentic generative ai based social media themes extractor, 2025.

[33] Tingrui Qiao, Caroline Walker, Chris W Cunningham, and Yun Sing Koh. Thematic-LM: a LLM-based multi-agent system for large-scale thematic analysis. In *THE WEB CONFERENCE 2025*, 2025.

[34] JaMor Hairston, Ritvik Ranjan, Sahithi Lakamana, Anthony Spadaro, Selen Bozkurt, Jeanmarie Perrone, and Abeed Sarker. Automated thematic analyses using llms: Xylazine wound management social media chatter use case, 2025.

[35] Majeed Kazemitabaar, Xinying Hou, Austin Henley, Barbara J. Ericson, David Weintrop, and Tovi Grossman. How novices use llm-based code generators to solve cs1 coding tasks in a self-paced learning environment, 2023.

[36] Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. "don't forget the teachers": Towards an educator-centered understanding of harms from large language models in education, 2025.

[37] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI ' 24, page 1 – 20. ACM, May 2024.

[38] Walter S. Mathis, Sophia Zhao, Nicholas Pratt, Jeremy Weleff, and Stefano De Paoli. Inductive thematic analysis of healthcare qualitative interviews using open-source large language models: How does it compare to traditional methods? *Computer Methods and Programs in Biomedicine*, 255:108356, 2024.

[39] Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, 30(3):411–433, 2004.

[40] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[41] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, pages 74–81. Barcelona, Spain, 2004.

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[43] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[44] Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors, *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[46] Karisa Parkington, Bazen G. Teferra, Marianne Rouleau-Tang, Argyrios Perivolaris, Alice Rueda, Adam Dubrowski, Bill Kapralos, Reza Samavi, Andrew Greenshaw, Yanbo Zhang, Bo Cao, Yuqi Wu, Sirisha Rambhatla, Sridhar Krishnan, and Venkat Bhat. Human vs. llm-based thematic analysis for digital mental health research: Proof-of-concept comparative study, 2025.

[47] Zackary Okun Dunivin. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks, 2024.

[48] Tadeusz Calinski and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. Variance Ratio Criterion (VRC).

[49] Rewina Bedemariam, Natalie Perez, Sreyoshi Bhaduri, Satya Kapoor, Alex Gil, Elizabeth Conjar, Ikkei Itoku, David Theil, Aman Chadha, and Naumaan Nayyar. Potential and perils of large language models as judges of unstructured textual data, 2025.

[50] Mehmet Hamza Erol, Batu El, Mirac Suzgun, Mert Yuksekgonul, and James Zou. Cost-of-pass: An economic framework for evaluating language models, 2025.

[51] Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The economics of large language models: Token allocation, fine-tuning, and optimal pricing. Technical report, Cowles Foundation Discussion Paper No. 2425, Yale University; also available on arXiv, 2025. Online; February 11, 2025.

表 3: 关键词词典（关键词_字典）用于主题分析文献检索。

字典键	关键词
核心_ta	thematic analysis; qualitative thematic analysis; theme analysis; thematic coding; theme coding; identify themes; theme identification; Braun and Clarke; Braun & Clarke; six-phase framework; six phase framework
归纳型_tA	inductive thematic analysis; data-driven thematic analysis; bottom-up thematic analysis; open coding (qualitative)
演绎_ta	deductive thematic analysis; theory-driven thematic analysis; top-down thematic analysis; a priori codes
通用 llm	LLM; large language model; transformer-based model; foundation model; prompting; chain-of-thought; self-reflection (LLM); RAG; few-shot; zero-shot
医疗卫生保健	clinical interview; patient interview; healthcare; medicine; clinical transcript; medical transcript; qualitative health research; nursing

A 关键词词典

B 模型发布里程碑

我们在LLMs的发展过程中确定了以下里程碑；然而，根据文献综述中突出的模型，只选择了这些关键里程碑的一部分纳入我们的图表。InstructGPT 于 2022 年 3 月 4 日推出，随后 PaLM 于 2022 年 4 月 5 日推出，BLOOM 于 2022 年 11 月 9 日推出。ChatGPT 于 2022 年 11 月 30 日发布。在 2023 年，LLaMA 于 2 月 24 日出现，Claude 1 于 3 月 14 日推出，GPT-4 于 3 月 15 日推出，Claude 2 于 7 月 11 日推出，Llama 2 于 7 月 18 日推出。2024 年带来了 Claude 3 于 3 月 4 日发布，Llama 3 于 4 月 18 日发布，GPT-4o 于 5 月 13 日发布，Claude 3.5 于 6 月 20 日发布，Gemini 2.0 于 12 月 11 日发布。在 2025 年，GPT-4.1 于 4 月 14 日推出，随后是 Claude 4 于 5 月 22 日推出，Claude 4.1 于 8 月 5 日推出，以及 GPT-5 于 8 月 7 日推出。

C 数据收集和关键词搜索程序

我们通过 Kaggle API 访问了官方的 arXiv 元数据（截至日期：2025-08-15）。每条记录都包括标题、摘要和主题类别等字段。

- **关键词构建：**我们为主题分析（核心 TA、归纳 TA、演绎 TA）、大型语言模型（LLMs）和医疗保健领域编译了关键词列表（附录表 3）。
- **搜索方法：**在标题和摘要字段中应用了不区分大小写的字符串搜索。如果一篇论文包含列表中的至少一个关键词，则将其计入该类别。在一个类别内多次匹配只计一次。例如，含有多个 TA 条目的论文在核心_ta 下仅记录一次。
- **类别交集：**同时包含 TA 和 LLM 术语的论文被归类为核心_ta+通用（llm_）。那些还包含医疗保健术语的论文则计入三重交集 (核心_ta+通用_llm+医疗卫生)。

结果计数汇总在表 1 中。

D 调查提示

用于文献检索的提示在诱发 [19] 如下：*如何在比较直接的 LLM 编码和混合的人工-AI 工作流时，使自动归纳主题分析的评估框架平衡有效性、可靠性和可解释性？*

E 补充文献检索策略

为了补充基于 arXiv 的搜索并承认一些同行评审论文可能未包含在 arXiv 中（因为它是一个非存档平台），我们实施了两种额外策略：

- **诱饵辅助搜索：** 我们使用了 Elicit [19]，这是一款旨在简化文献回顾过程的人工智能研究助手。通过 Elicit 识别的所有参考文献均由作者手动验证和审查。调查提示的具体示例见附录 D。
- **手动数据库搜索：** 我们在主要的索引平台 [5] 上进行了手动搜索，使用查询词主题分析结合特定领域的术语。确切的关键字列表见附录表 3。

F 使用的模型

OpenAI GPT 家族（41 篇论文，∼74%）: GPT-3.5 和 GPT-4 变体（Turbo，16k/32k，GPT-4o，GPT-4V）。大多数研究的基础。

克劳德（8 篇论文，∼15%）: Claude-instant, Claude-2/3.5/3.7 Sonnet。主要用于比较。

LLaMA & 开源（7 篇论文，∼13%）：LLaMA-2/3, CodeLlama, Mistral, Mixtral, Gemma, DeepSeek。与 GPT 对比基准；因其透明性而受到重视。**谷歌 Gemini（6 篇论文，∼11%）**：双子座专业/闪存/超极/2.5 预览。通常是基线；在归纳任务上表现混合。**其他（5 篇论文，∼9%）**: InCoder, CodeGen, ProgGP, genre-CTRL, Whisper, Sentence-T5。特定或辅助角色。

G 提示策略

零样本（35%）：最常见；用于代码提取、主题整合和命名的单阶段或多阶段流水线。高效但对措辞敏感，引发可重复性担忧。**少样本 (16%)**: 嵌入提示中的示例、模板或伪代码框架；提高了可控性但引入了示例选择偏差。**思维链（13%）**: 鼓励中间推理步骤；支持透明度，尽管输出通常冗长或不一致。**自洽性/反射（15%）**：迭代或递归重新提示；旨在稳定，但与单次通过输出相比效果不一。**基于工具/代理的（11%）**：多代理角色（编码者、评估者、裁判）或辩论框架；创新但资源密集型，收益变化不定。**不适用（11%）**：未使用提示（例如，微调模型、人工编写的数据库、概念性论文）。

[5] 谷歌学术, PubMed, 斯库伯斯, 科学网