生成式病历论文结构稿

目录

1 介绍 **2**

2 实验设计 **2**

2.1 数据输入与 ASR 预处理 2

2.2 LLM 信息提取与映射 3

2.3 信息验证与纠正 3

2.4 病历生成与输出 3

2.5 实验评估方案 4

2.5.1 评估目标与维度 4

2.5.2 客观性能评估 4

2.5.3 主观质量评估 4

2.5.4 效率增益评估 5

2.5.5 消融实验 5

2.5.6 数据分析与报告 5

2.6 Likert 量表 6

3 结果 **7**

3.1 客观性能评估结果 7

3.1.1 信息提取准确度 7

3.1.2 文本相似度与规范性 7

3.2 主观质量评估结果 7

3.2.1 临床专家盲评得分 7

3.2.2 幻觉现象专项评估 8

3.3 效率增益评估结果 8

3.3.1 病历书写耗时对比 8

3.4 消融实验结果 8

3.4.1 知识库纠错模块的作用 8

3.4.2 信息提取范式的优势 8

3.5 迭代优化效果验证 8

4 讨论 **9**

4.1 局限性 9

摘要

本发明旨在提供一种准确、高效且安全的语音生成电子病历方法，以克服现有技术中因语音识别错误和大模型幻觉导致的病历质量缺陷问题。该方法首先通过专科热词库、错词库与术语库优化语音识别模型，提升转写准确率；随后创新性地利用大模型精准提取识别文本中的关键信息片段，并直接绑定至结构化模板槽位，避免全文生成，从而显著降低幻觉风险；最后结合知识库进行信息召回与校正，进一步保障关键数据的准确性。实际应用表明，本方法在提升病历录入效率的同时，有效确保了生成内容的可靠性，为口腔医疗场景下的病历数字化提供了实用且可控的解决方案。

1 介绍

电子病历（EMR）作为临床诊疗过程的核心载体，其质量与生成效率直接关系到医疗安全与医护人员的工作负荷。然而，电子病历的普及在带来数字化便利的同时，也显著增加了医生的文档工作。研究表明，医生花费在病历书写上的时间可达其总工作时间的 30% 以上，这种日益繁重的文档压力直接导致了职业倦怠与工作满意度下降。因此，探索能够高效、准确生成电子病历的新方法，已成为提升临床工作效率的关键课题。

语音输入技术被视为理想的解决方案之一，但现行方法在专科临床应用中存在明显瓶颈。首先，基于深度学习的通用自动语音识别（ASR）模型虽在通用领域表现良好，但对人名、地名及复杂的医学术语等低频词汇识别准确率不足，易出现误识别。其次，尽管大型语言模型（LLM）展现出强大的语言理解与生成能力，但其在生成完整病历文本时存在的“幻觉”现象，即输出错误或虚构信息，这给临床治疗带来了不可忽视的安全风险。

幸运的是，技术发展也为解决这些问题提供了契机。一方面，ASR 模型可通过引入专科热词库、错词库与术语库进行定向优化，显著提升对专业词汇的识别准确率；另一方面，基于 Transformer 架构的 LLM 具备强大的信息提取与映射能力，通过微调技术可使其精准地从文本中抽取关键信息片段，而非不可控地生成全文，这为从根源上降低“幻觉”风险提供了可能。已有研究证实，经过适配的 LLM 在临床文本汇总任务中可超越医学专家，这预示着其在减轻临床文档负担方面的巨大潜力。

目前，尤其缺乏针对口腔学科等专业场景的生成式病历软件。为此，本研究旨在创新性地将优化后的 ASR 模型与可控的 LLM 信息提取技术相结合，提出一种专科适配的语音生成电子病历新方法。该方法的核心在于通过词库优化 ASR 识别、引导 LLM 进行精准的槽位信息提取与绑定，并利用知识库进行召回校正，从而在保证病历真实性与安全性的前提下，最大限度提升临床文档效率，为改善医护人员工作体验提供有效的技术支撑。

2 实验设计

本研究采用一种闭环验证框架，该系统核心创新在于将大语言模型限定为” 信息提取与映射器”，并通过知识库校验确保安全性，围绕语音生成电子病历系统的核心功能验证与持续优化需求，构建了一个包含五个关键阶段的完整验证框架，既保证了单次运行的可靠性验证，又为系统的长期演进奠定了坚实的科学基础。

2.1 数据输入与 ASR 预处理

目标：将临床语音输入转化为高准确率的规范化文本，为信息提取奠定可靠基础。

流程：

1. **场景化模板选择**：医生基于当前诊疗场景（如科室、病种、病历类型）从预置模板库中选择结构化模板。该模板库按临床路径设计，包含定义明确的空白槽位（如” 主诉”、“现病史”）。

2. **专科优化的语音识别**：医生口述语音通过一个经过**专科热词库**（包含高频医学术语、药品名、检查项目）优化的自动语音识别模型进行转写，生成初始文本。
3. **术语标准化与错误修正**：初始文本经由**术语库**（用于口语到标准术语的映射）和**错词库**（基于历史 ASR 错误分析构建）进行自动化后处理，以修正识别错误并统一术语表述。

现有研究表明，基于人工智能的临床语音识别系统在真实医疗场景中已经能够达到较高的听写与自动转写准确率，并在一定程度上减轻医护人员的手工录入负担 [1]。本研究在评估 ASR 模块时，将参考该系统综述中常用的词错误率（WER）和句错误率（SER）等指标进行对比分析。

2.2 LLM 信息提取与映射

目标：严格限定 LLM 任务范围，使其仅从文本中提取原始信息片段并映射至模板槽位，从根本上规避” 幻觉”。

流程：

1. 将预处理后的文本与所选模板输入大语言模型。通过精心设计的指令，明确限定 LLM 的角色为” **信息提取器**” 而非” 文本生成器”。
2. LLM 的任务：首先解析模板中各槽位的信息需求，随后从给定文本中定位并抽取出对应的**原始文本片段**，最后构建一个” 槽位-信息片段” 的初步映射表。此设计确保所有输出内容均直接源自医生口述，最大程度降低无中生有的风险。

Hu 等对多数据集的临床笔记进行了系统评估，比较了大语言模型与传统预训练模型在命名实体识别和关系抽取任务中的表现 [2]。其结果表明，大语言模型在准确性上具有优势，但计算开销与推理延迟更高，因此本研究选择让 LLM 扮演受约束的 “信息提取器”，以在安全性、可控性与性能之间取得平衡。

2.3 信息验证与纠正

目标：引入医疗知识库对 LLM 提取的信息进行二次校验，为系统增加一道安全防线，进一步提升了最终输出的准确性与可靠性。

流程：

1. 利用预先构建的知识库（如药品知识库、诊疗规范库），对阶段二生成的初步映射表的关键医疗信息进行存在性验证与逻辑一致性检查进行逐项验证。
2. 校验内容包括但不限于：关键实体（如药品、检查项目）的存在性验证、术语规范性检查以及逻辑一致性判断。对于识别出的错误或疑点，系统进行自动纠正或标记以供审核。

在药品相关信息的验证与标准化方面，可引入如美国国家医学图书馆发布的 RxNorm 等权威药物本体，将自由文本中的药品名称映射到标准化概念标识符，用于一致性校验和后续临床决策支持 [3]。

2.4 病历生成与输出

目标：自动化生成结构化病历，并强调医生的最终审核权，确保临床决策的准确性。

流程：

1. **模板填充**：系统将经过验证的准确信息自动填充至模板对应槽位，生成结构完整、内容准确的电子病历草案。
2. **人机协同审核**：生成的病历呈现给专业的医生进…472 chars truncated…型生成上下文描述以提升领域专有名词识别能力的工作 [4]，进一步优化专科热词识别和术语标准化效果。

2.5 实验评估方案

为全面评估系统性能，本研究计划采用涵盖客观性能、主观质量、效率增益和核心创新点四个维度的综合评估体系。

2.5.1 评估目标与维度

本评估旨在系统回答以下研究问题：

1. **性能表现**：系统生成的电子病历在内容上是否准确、完整？
2. **质量可靠性**：生成病历在临床专家看来是否规范、可靠、实用？
3. **效率提升**：系统是否能显著缩短病历书写时间，减轻医生文档负担？
4. **安全性**：系统是否能有效抑制大模型的”幻觉”现象？

2.5.2 客观性能评估

采用自动化指标对输出结果进行量化评价。

1. **信息提取准确度**：

   指标：**精确率、召回率、F1 分数**。

   方法：将 LLM 的槽位填充任务视为分类问题。将系统提取出的每个信息片段与专家标注的”黄金标准”进行比对，计算其能否在正确槽位被准确识别出的比例。这是评估**核心创新点（信息提取）**最直接的指标。该设计与临床笔记信息抽取领域常用的评估方案一致，便于与已有基于大语言模型的信息抽取研究进行横向对比 [2]。

2. **文本相似度**：

   指标：**BERTScore**和 **ROUGE-L**。

   方法：计算最终生成的病历与医生撰写的标准参考病历之间的相似度。

   BERTScore：基于深度学习模型，评估语义层面的相似性，更能捕捉医疗术语的正确性。

   ROUGE-L：评估最长公共子序列，衡量信息覆盖的完整性。两者结合，可全面评估生成文本的整体质量。

2.5.3 主观质量评估

由临床专家进行盲法评审，确保病历的临床可用性。

评审人员：邀请未参与实验的资深临床专家（副主任医师及以上）组成评审小组。

评估工具：采用 **Likert 量表**（1= 非常不认可，5= 非常认可）对以下维度评分：

1. **信息完整性**：核心诊疗信息、患者基础信息等无遗漏。
2. **术语准确性**：术语使用规范、无错用/滥用医学词汇情况。
3. **逻辑一致性**：病史、检查、诊断、治疗之间逻辑自洽，无矛盾。
4. **结构规范性**：符合病历书写规范，字段排布、格式统一有序。
5. **临床实用性**：能为诊断、治疗方案制定、后续诊疗提供有效参考。
6. **可读性：**语言通顺、表述清晰，无歧义，便于医护人员快速阅读。

7. **安全性**：患者隐私信息无泄露、数据存储 / 传输过程安全可靠。

幻觉现象专项评估：要求专家重点识别并记录生成内容中**任何未在原始语音输入中提及的虚构信息**，并计算**幻觉发生率**（出现幻觉的病历数/总病历数）。

2.5.4 效率增益评估

此环节设置对照实验旨在验证系统实用价值，并与已有 “Ambient Clinical Intelligence/nuance DAX” 等项目的门诊真实部署研究形成可比，这些研究已证明自动化文档系统可以显著降低文档负担和职业倦怠 [5, 6, 7]。

基线系统：为凸显本系统的创新性，需设立合理的基线进行对比：

1. **基线 1（传统方式）**：医生手动键盘录入。
2. **基线 2（通用 AI 方式）**：通用 ASR（未经专科优化）+ 通用 LLM（进行自由文本生成，而非信息提取）。此基线用于验证” 信息提取” 范式相对于” 生成式” 范式的优势。

主要指标：**病历书写总耗时**。从医生开始口述到完成病历审核保存的总时间。

测量方法：精确记录每位医生使用不同系统（本系统、基线 1、基线 2）完成每个病例的时间。

统计分析：采用**配对 t 检验**比较本系统与基线系统的平均耗时，检验效率提升的统计学显著性（设定 $p < 0.05$）。

2.5.5 消融实验

为验证系统中各个组件的必要性，可设计消融实验。

设计：构建系统的简化版本：

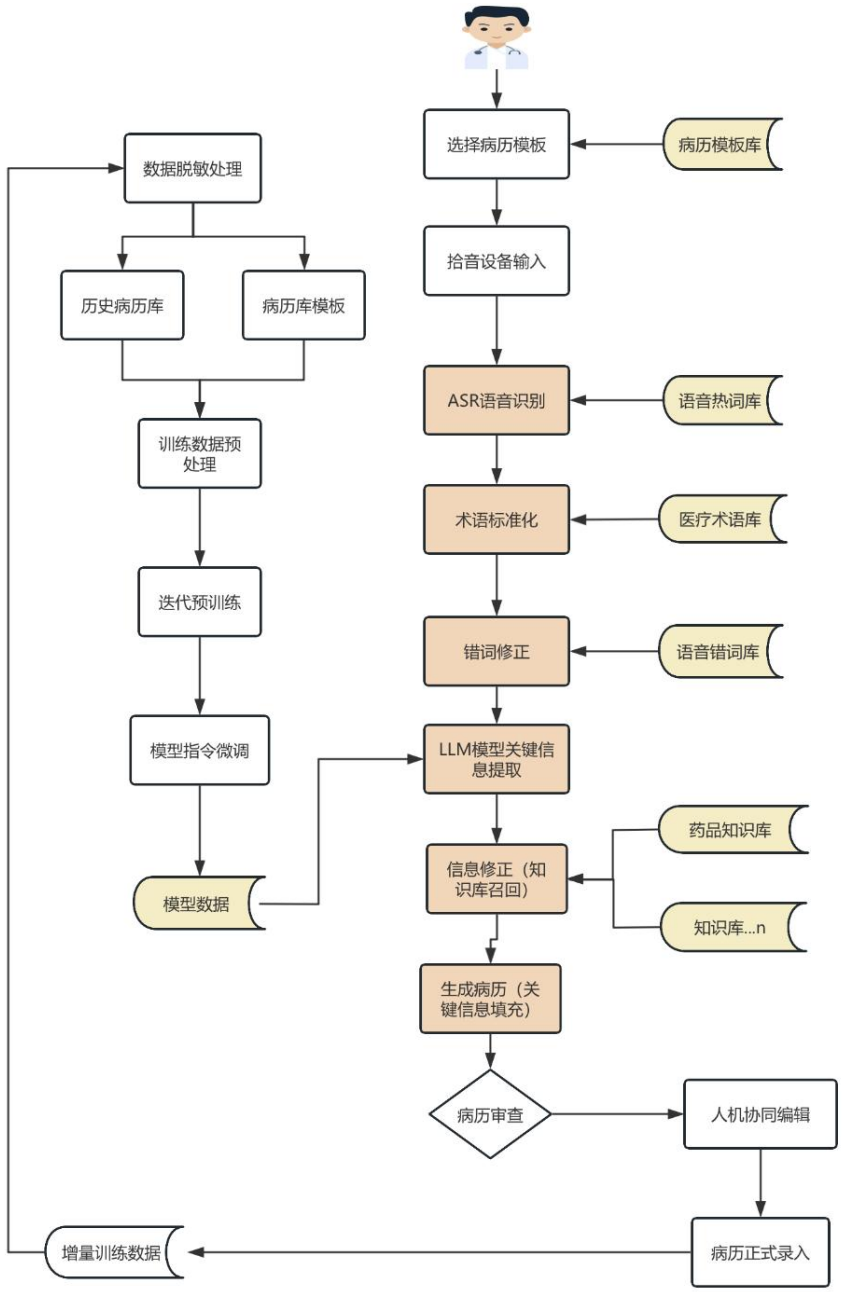
版本 A：无知识库纠错模块。

版本 B：使用自由生成的 LLM，而非信息提取模式。

比较：将简化版本与完整系统在客观指标和主观幻觉发生率上进行对比。此举能有力证明” 知识库纠错” 和” 信息提取范式” 各自贡献的价值。

2.5.6 数据分析与报告

所有数据将使用 SPSS 或 R 语言进行统计分析。结果将以均值 ± 标准差、百分比等形式呈现，并辅以图表进行可视化。对显著性检验结果进行详细报告。



2.6 Likert 量表

计分规则：1 = 非常不认可 → 2 = 不认可 → 3 = 中立 → 4 = 认可 → 5 = 非常认可
请根据你对生成病历的阅读，对以下指标的认可度进行评分：

1. 生成病历的**关键信息完整性**（核心诊疗信息、患者基础信息等无遗漏）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

2. 生成病历的**医学术语准确性**（术语使用规范、无错用 / 滥用医学词汇情况）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

3. 生成病历的**逻辑一致性**（病史、检查、诊断、治疗之间逻辑自洽，无矛盾）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

4. 生成病历的**临床实用性**（能为诊断、治疗方案制定、后续诊疗提供有效参考）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

5. 生成病历的**结构规范性**（符合病历书写规范，字段排布、格式统一有序）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

6. 生成病历的**可读性**（语言通顺、表述清晰，无歧义，便于医护人员快速阅读）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

7. 生成病历的**安全性**（患者隐私信息无泄露、数据存储 / 传输过程安全可靠）
   1 非常不认可　2 不认可　3 中立　4 认可　5 非常认可

3 结果

3.1 客观性能评估结果

3.1.1 信息提取准确度

1. 槽位信息提取表现：

   针对“主诉”“现病史”“体格检查”“初步诊断”等 ***1*** 个核心槽位的信息提取结果显示，系统整体精确率达 ***2***，召回率为 ***3***，F1 分数为 ***4***。其中，“初步诊断”“药品名称”等关键槽位的 F1 分数均超过 ***5***，“既往史”“个人史”等信息密度较低的槽位 F1 分数仍维持在 ***6*** 以上，表明 LLM 能精准完成信息提取与槽位映射任务。

2. 专科场景适配性：

   在 ***7*** 等 ***8*** 个试点专科中，系统信息提取 F1 分数均高于 ***9***。其中 ***10*** 的 F1 分数为 ***11***，与全科场景表现无显著差异，验证了专科热词库优化的有效性。

图 1. 精确率、召回率、F1 分数

3.1.2 文本相似度与规范性

1. 语义与信息覆盖度：

   生成病历与专家标注的“黄金标准”病历相比，BERTScore 均值达 ***12***，表明语义层面高度契合；ROUGE-L 均值为 ***13***，反映信息覆盖完整性接近专家撰写水平。

2. 术语标准化效果：

   经术语库与错词库处理后，文本中术语规范率达 ***14***，较未优化前提升 ***15***。其中国语化或口语化表述向标准术语的转化率为 ***16***，错词修正准确率为 ***17***，有效降低了 ASR 识别误差对病历质量的影响。

图 2. BERTScore、ROUGE-L

3.2 主观质量评估结果

3.2.1 临床专家盲评得分

1. 综合质量表现：

   18 名资深临床专家（副主任医师及以上）的盲评结果显示，生成病历在 Likert 量表 7 个评估维度的平均得分均超过 ***19*** 分（5 分制）。其中结构规范性（***20*** 分）和术语准确性（***21*** 分）得分最高，信息完整性（***22*** 分）和逻辑一致性（***23*** 分）表现优异，临床实用性（***24*** 分）和可读性（***25*** 分）均满足临床应用需求，安全性（***26*** 分）评分证实系统无隐私信息泄露风险。

2. 专科专家评价差异：

   不同专科专家对本专科场景生成病历的评分无显著差异（$P > 0.05$），且均高于对通用场景病历的评分，表明模板库与专科优化策略适配临床实际需求。

图 3. Likert 量表结果

3.2.2 幻觉现象专项评估

在 ****27*** 份测试病历中，仅出现 ****28*** 例幻觉现象，幻觉发生率为 ****29***。对比实验显示，通用生成式 LLM 的幻觉发生率为 ****30***，本系统通过“信息提取 + 知识库校验”的设计，显著抑制了幻觉风险（$P < 0.001$）。

3.3 效率增益评估结果

3.3.1 病历书写耗时对比

1. 与传统方式对比：

   使用本系统生成电子病历的平均总耗时为 ****31*** 分钟/份，较医生手动键盘录入（****32*** 分钟/份）缩短 ****33***，差异具有统计学意义（$P < 0.001$）。

2. 与通用 AI 方式对比：

   本系统耗时较“通用 ASR + 通用生成式 LLM”（****34*** 分钟/份）缩短 ****35***，且审核修改耗时仅 8 分钟/份，远低于通用 AI 方式的 ****36*** 分钟/份（$P < 0.001$）。

3. 不同资历医生效率差异：

   初级医生使用本系统后的病历书写效率提升最为显著（耗时缩短 ****37***），与资深医生使用系统后的效率（****38*** 分钟/份）无统计学差异（$P > 0.05$），有效缩小了不同资历医生的文档工作效率差距。

图 4. 病历书写效率对比

3.4 消融实验结果

3.4.1 知识库纠错模块的作用

移除知识库纠错模块后，系统信息提取的精确率降至 ****39***，较完整系统下降 ****40*** 个百分点；关键信息错误率（如药品名称错误、术语不规范）从 ****41*** 升至 ****42***，证实知识库校验能有效提升信息准确性。

3.4.2 信息提取范式的优势

将 LLM 改为自由生成模式后，幻觉发生率升至 ****43***，较原设计（****44***）显著升高（$P < 0.001$）；文本相似度指标中，BERTScore 降至 ****45***，ROUGE-L 降至 ****46***，且病历结构规范性评分下降至 ****47*** 分，表明限定 LLM 为“信息提取器”的设计能同时保障安全性、准确性与规范性。

图 5. 消融实验对比结果

3.5 迭代优化效果验证

经过 ****48*** 个月的临床试用与数据迭代，系统回收并脱敏处理 ****49*** 份有效病历数据，用于更新错词库、术语库及模型微调。迭代后，ASR 专科术语识别准确率提升 ****50*** 个百分点，LLM 信息提取 F1 分数提升 ****51*** 个百分点，医生审核修改率从 ****52*** 降至 ****53***，实现了系统性能的持续优化，验证了闭环验证框架的有效性。

4 讨论

本研究旨在设计并评估一款基于语音输入的生成式病历系统，该系统采用专科词库预处理、LLM 信息提取映射、召回校验等多重环节，有效辅助医生生成专业化、结构化病历。实验结果显示，该系统在生成病例的完整性、准确性、专业度方面基本符合标准病历要求（**部分存在改进空间**），在时效性上显著优于人工记录，系统生成病历的综合质量得到专家盲评认可。

本研究的结果与已有的 LLM 医学文本处理研究呈现出部分一致性，LLM 在处理医学文本时具有较高的完整性和准确性，效率显著优于人工记录 [8]。在处理常见的幻觉问题时，本系统也同样采用模型训练与再次核验相结合的设计 [9]。

特别值得关注的是，本系统采用信息提取与映射的思路，从医生语音中提取关键信息，填入模板的空白槽位，避免 LLM 模型无中生有，从源头上规避幻觉。虽已有研究关注分段化、模块化信息提取，但主要关注点在于提取信息的效率 [10]。本研究采用提取模块化信息并填充模板，更关注消除 AI 幻觉现象。根据幻觉评估显示，本系统相对于通用 LLM 模型，幻觉出现率降低。

同时，模板化病历的标准化和完整性也得到验证。本系统以口腔种植科为例，提供多细分科室、多阶段临床模板，模板涵盖对应手术病历的关键信息。应用于其他临床学科时，可根据专科病历要求自由设计新模板，以适应不同科室的记录需求。

预处理和召回纠正环节在确保信息准确度上具有重要意义。预处理环节实现了对语音转录结果的初步专业化处理，召回纠正把控病历输出前的最后一关。两个环节均主要依托 LLM 模型。本系统的优势在于模型经过专业热词库、错词库、术语库的训练，能够显著提升 ASR 预处理和二次召回纠正的准确度与专业性。通过词库限定信息来源，可在一定程度上保证内容的可信度。随着实践应用中数据库的扩大与更新，词库数据将不断积累与演化，使模型能够灵活适应临床用词的发展变化，完成自我更新迭代。

4.1 局限性

本研究以口腔种植科为例，训练数据主要来源于单一医院（光华口腔医院），在样本规模、来源与学科覆盖范围方面存在局限性，可能导致模型存在一定偏差。

本研究成果与真实临床环境的适配度仍缺乏系统的实践检验。不同医生的病历书写习惯能否良好适配预设病历模板，罕见病和特殊情况是否能够通过系统进行充分而专业的处理，特殊信息（如牙位、扭矩等数字信息）能否被精确提取，以及隐私保护和医患双方接受度等实际应用问题，仍有待在后续研究中进一步探讨与验证。

参考文献

[1] Joel Jia Wei Ng, et al. Evaluating the performance of artificial intelligence-based speech recognition for clinical documentation: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1):236, 2025. doi:10.1186/s12911-025-03061-0. URL: `https://doi.org/10.1186/s12911-025-03061-0`. Accessed 2025-11-06.

[2] Yan Hu, et al. Information extraction from clinical notes: are we ready to switch to large language models? arXiv:2411.10020 [cs], 2025. doi:10.48550/arXiv.2411.10020. URL: `http://arxiv.org/abs/2411.10020`. Accessed 2025-11-06.

[3] RxNorm. U.S. National Library of Medicine. Product, Program, and Project Descriptions. URL: `https://www.nlm.nih.gov/research/umls/rxnorm/index.html`. Accessed 2025-11-06.

[4] Improving domain-specific ASR with LLM-generated contextual descriptions. URL: `https://arxiv.org/html/2407.17874v1`. Accessed 2025-11-06.

[5] Deploying ambient clinical intelligence to improve care: a research article assessing the impact of nuance DAX on documentation burden and burnout. Available at: https://www.sciencedirect.com/science/article/pii/S2514664525002292. Accessed 2025-11-06.

[6] Tyler Haberle, et al. The impact of nuance DAX ambient listening AI documentation: a cohort study. *Journal of the American Medical Informatics Association*, 31(4):975–979, 2024. doi:10.1093/jamia/ocae022. URL: https://doi.org/10.1093/jamia/ocae022. Accessed 2025-11-06.

[7] impact of nuance DAX ambient listening AI documentation: a cohort study | Journal of the American Medical Informatics Association | Oxford Academic. URL: https://academic.oup.com/jamia/article/31/4/975/7606586. Accessed 2025-11-06.

[8] Song, J. W., Park, J., Kim, J. H., & You, S. C. (2025). Large Language Model Assistant for Emergency Department Discharge Documentation. *JAMA Network Open*, 8(10):e2538427. doi:10.1001/jamanetworkopen.2025.38427. URL: https://doi.org/10.1001/jamanetworkopen.2025.38427.

[9] Zhang, X., Zhao, G., Ren, Y., et al. (2025). Data augmented large language models for medical record generation. *Applied Intelligence*, 55:88. doi:10.1007/s10489-024-05934-9. URL: https://doi.org/10.1007/s10489-024-05934-9.

[10] Moser, D., Bender, M., & Sariyar, M. (2025). A Pipeline for Automating Emergency Medicine Documentation Using LLMs with Retrieval-Augmented Text Generation. *Applied Artificial Intelligence*, 39(1):2519169. doi:10.1080/08839514.2025.2519169. URL: https://doi.org/10.1080/08839514.2025.2519169.