Programmable Evidence-Grounded Clinical Copilot for Longitudinal EHRs

Boyang Gu
boyang.gu19@alumni.imperial.ac.uk

1 Introduction

Electronic Health Records (EHRs) contain rich longitudinal information about patients, but they are long, fragmented, and multi-modal. A single patient's record can span hundreds of pages of clinical notes, lab results, imaging reports, and other data over time [5, 10, 11, 13]. Clinicians face severe information overload in navigating these records, making it challenging to extract and analyze critical details for decision-making. Relevant facts are often scattered across different notes and data modalities, resulting in a cognitive burden [5]. This complexity can even lead large language model (LLM) systems to hallucinate or fabricate details when they cannot easily locate needed information in the input [17, 45]. If an LLM loses track of a detail in a lengthy history or encounters incomplete context, it may generate a plausible-sounding yet incorrect statement to fill the gap. In high-stakes clinical settings, such hallucinations pose serious safety concerns. Clearly, any clinical task solution must manage the length and heterogeneity of EHR data to prevent information overload and ensure factual accuracy.

Recent advances in LLMs have drawn interest in automating clinical text understanding for tasks like question answering (QA), summarization, and diagnostic reasoning [42, 46, 36, 20, 24, 19]. However, deploying LLM-based systems in healthcare requires far more than raw language generation capability. Factual consistency and explainability are crucial since clinicians need answers that are not only correct but also supported by evidence or reasoning that they can inspect [35, 6]. Similarly, models must integrate expert medical knowledge (such as clinical guidelines, best practices, and biomedical literature) rather than relying only on intrinsic patterns learned from data [47, 16, 42]. A purely data-driven model might not know the latest guidelines or could misinterpret rare conditions. Incorporating domain-expert knowledge helps ensure the results align with medical standards. In other words, clinical NLP tasks require trustworthy systems that can ground their outputs in verifiable evidence from both the patient's record and external medical sources, providing transparency into why an answer or recommendation was made.

Research Vision The long-term vision of this program is to create an **explainable, evidence-grounded framework for longitudinal EHRs** that allows clinicians to interact with complex patient records through high-level intents, while AI agents transparently handle the underlying reasoning, retrieval, and integration of multi-modal data. Instead of treating large language models as black boxes for single tasks, the goal is to move towards a programmable, auditable executor for longitudinal clinical reasoning.

This proposal will investigate the following research questions:

- **RQ1 (Evidence-grounded QA):** How can we design clinical question-answering systems that remain factually reliable on long, fragmented, multi-modal EHRs, while explicitly exposing their supporting evidence and uncertainty?
- **RQ2 (Longitudinal understanding):** How can we structure and summarize longitudinal patient trajectories so that downstream diagnostic reasoning is both accurate and inspectable by clinicians?
- **RQ3 (Clinical DSL and programmability):** Can a domain-specific language for clinical workflows provide a stable, interpretable interface between clinicians' intents and evolving LLM-based agents, enabling safe and reusable AI-assisted workflows?

If successful, this work would shift clinical AI from specific and single-task models to a unified, programmable backbone for longitudinal reasoning across the entire patient journey. Clinical experts would be able to specify high-level, auditable workflows that orchestrate patient data, medical knowledge, and AI agents, turning complex multi-step reasoning into a transparent and governable capability of the healthcare system.

2 Background and Motivation

2.1 Clinical QA, Summarization, and External Knowledge

Clinicians need to extract specific information from EHRs and to obtain concise overviews. Research on clinical QA and summarization shows that both are feasible. Datasets such as emrQA and related resources treat EHR QA as a text comprehension problem, while there are other work that studies extractive and abstractive summarization of clinical notes, discharge summaries, and progress reports [28, 38, 33]. More recently, LLM-based systems have demonstrated excellent fluency in producing patient summaries and answering clinical questions, including GPT models finetuned on clinical text.

However, these systems often operate as black-box generators. Answers and summaries may be correct on average but can omit crucial details, over-emphasize recent notes, or hallucinate plausible-sounding facts that are not supported by the record. This is particularly problematic when questions require combining fragmented EHR evidence with general medical knowledge that is not explicitly documented in the notes. Retrieval-augmented generation (RAG) frameworks address part of this issue by explicitly retrieving relevant passages from EHRs, guidelines, and knowledge bases before generation, and have been shown to improve factual accuracy for both QA and summarization tasks [2, 25]. However, challenges remain in deciding what to retrieve from multi-modal, longitudinal records, how to balance patient-specific evidence with external guidance, and how to expose uncertainty when the evidence is incomplete or conflicting.

These limitations motivate **RQ1**: how to design clinical QA systems that remain factually reliable on long, fragmented, multi-modal EHRs while explicitly exposing their supporting evidence and uncertainty, and reinforce **RQ2**: how better longitudinal representations and summaries can serve as a stable component for answering complex, multi-step clinical queries.

2.2 Multi-modal Longitudinal EHR Modeling

Modern EHRs record diverse longitudinal information about each patient: diagnosis codes, lab results, medications, clinical notes, imaging reports, and time-series signals across many visits. This multi-modal, multi-visit structure creates challenges for processing. On the one hand, it provides a detailed view of disease trajectories, but on the other hand, it creates severe information overload for clinicians and models. Important facts are scattered across different documents and modalities, often with missing or inconsistent entries, making it difficult to reconstruct a coherent clinical story for a single patient.

A large amount of work has developed sequence and representation-learning models for EHRs, from early RNN-based and autoencoder-based approaches to transformer architectures that view a patient's history as a sequence of discrete events or tokens [8, 23, 18, 30]. More recent efforts extend this idea to multi-modal settings and even use LLMs to analyze patient timelines that mix structured data, notes, and other signals in a single sequence [29, 44, 41]. These models show that it is possible to learn powerful patient representations and to evaluate trajectories.

However, most of those works are optimized for prediction benchmarks (e.g. risk scores, readmission, mortality) rather than for the kinds of interactive tasks that clinicians actually face, such as quick understanding of a complex longitudinal course, free-form questions answering, and guideline consistency checking. Existing models rarely expose a human-interpretable view of the patient's trajectory or provide a stable abstraction that downstream agents can consume. This gap motivates **RQ2**: how to structure and summarize longitudinal patient trajectories so that downstream diagnostic reasoning is both accurate and inspectable by clinicians, and sets the stage for **RQ1** and **RQ3**, which depend on having usable representations of the record.

2.3 Explainable Clinical AI and the Case for a DSL

Since clinical decisions are highly regulated, AI systems must be not only accurate but also explainable and governable. Clinicians need to understand why a recommendation was made, what evidence supports it, and how it relates to established guidelines. Existing work on explainable AI in healthcare has explored post-hoc explanations and chain-of-thought (CoT) style rationales for LLMs, sometimes combined with explicit citation of guidelines or literature [1, 27].

These techniques can increase perceived transparency but still keep the full reasoning process buried inside a single model call. Hospitals and regulators have limited control over how the model decomposes tasks, which tools it uses, and how it behaves as models are updated over time.

In parallel, the medical informatics community has a long history of encoding clinical logic in domain-specific languages (DSLs) such as Arden Syntax and GLIF, which represent guidelines and rules as explicit, machine-interpretable programs [31, 26]. More recently, outside of biomedicine, several systems have combined DSLs with LLMs: the model proposes or edits programs in a human-readable DSL, while a separate interpreter executes them, providing a clear contract between human intent and model behavior [14, 21]. However, such DSL-based Copilots have not yet been explored for longitudinal clinical reasoning over EHRs. There is no established way for clinicians to express high-level intents in a form that can be executed, audited, and versioned independently of any particular model.

This gap between the need for explainability and governance on one side, and the flexibility of LLM-based agents on the other, motivates **RQ3**: whether a clinical DSL can provide a stable, interpretable interface between clinicians' intents and evolving LLM-based agents, enabling safe, reusable workflows. Together with the motivations for **RQ1** and **RQ2** above, this forms the motivation for the proposed explainable, evidence-grounded framework for longitudinal EHRs.

3 Research Plan and Methods

To address the above challenges, we propose a clinical domain-specific language (DSL) with its AI Copilot that explicitly separates the high-level intent of a task (what) from the procedural details of implementation (how). In detail, we will separate the research into three aims. Each aim produces a set of tools or methodologies that feed into the next. Below, we detail the plan for each aim, including approach and evaluation strategies.

3.1 Aim 1: Evidence-Centric Question Answering with RAG and Micro-Agents

Objective: Develop a QA system that accurately answers clinicians' questions about a patient's case by drawing evidence from both the patient's EHR and external knowledge sources, with answers explicitly supported by references or explanations.

This aim establishes a trustworthy QA layer by combining multi-source retrieval with an agent-based orchestration. We plan to:

- **Characterize Question Types and Sources:** We will analyze typical clinician queries (e.g., patient-specific questions such as "Has the patient ever had condition X?" and knowledge-oriented questions such as "What is the recommended dosing of Drug Z in this context?"). This taxonomy will drive which sources are consulted: deep EHR retrieval for patient-specific questions, and external resources (guidelines, drug references, literature) for knowledge questions, with many complicated questions requiring both.
- **Build Retrieval Tools as Micro-Agents:** We will implement retrieval modules over unstructured notes, structured EHR fields, and external expert sources. Notes will be indexed by clinical concepts and time. Structured data (e.g., labs, vitals, medications) will be exposed through focused queries (e.g., trends in renal function). External knowledge (guidelines, textbooks, literature) will be stored and searched with dense or sparse retrieval. Each module will behave as a micro-agent that, given a query, returns targeted snippets or data.
- **Design a Central QA Agent:** A central LLM-based QA agent will orchestrate these micro-agents. For a given question, it will decide which retrieval tools to call (e.g., EHR search first, guideline search next), accumulate a working set of evidence, and then generate an answer constrained to that evidence, encouraged not to introduce unsupported information. Conceptually, it will behave as an open-book model that composes answers by quoting and understanding retrieved content rather than hallucinating new facts.
- **Produce Explainable Outputs and Refine with Clinician Feedback:** Answers will be formatted for transparency, using bullet points with explicit citations to notes, lab values, or guideline passages, and structured criteria where appropriate. Early prototypes will be evaluated on de-identified EHR cases (e.g., public datasets or synthetic records), and iteratively refined with feedback from clinician collaborators on relevance, completeness, format, and handling of uncertainty.

Evaluation (Aim 1): We will evaluate the QA system on both accuracy and its handling of evidence and uncertainty. For accuracy, we will construct or reuse QA pairs derived from clinical datasets or case reports, measuring whether the system's answers match expert references and capture key facts. For factual grounding, medical experts will review a sample of answers to assess whether all statements are supported by cited evidence and to estimate hallucination rates (unsupported statements). For uncertainty and trust, we will examine whether the system appropriately abstains or flags low-evidence situations, and clinicians will rate how well they understand the system's reasoning and whether the evidence presentation and uncertainty handling increase their confidence. We will compare against a baseline LLM (e.g., ChatGPT) answering without explicit retrieval or citations. By the end of Aim 1, we expect a robust QA module that reliably answers patient-related questions using a combination of record retrieval and external knowledge, while making evidence usage and failures visible, forming a core building block for Aims 2 and 3.

3.2 Aim 2: Longitudinal Summarization and Diagnosis Support Agents

Objective: Develop two complementary agents: (i) a summarization agent that produces concise longitudinal overviews of a patient‘s record; and (ii) a diagnosis support agent that suggests likely diagnoses or explanations for a patient's findings, grounded in both patient data and medical knowledge. This aim targets two core needs in clinical practice, which both require a strong grounding in the record and transparent reasoning.

For Aim 2i (Summarization), we plan to:

- **Extract-then-Abstract Pipeline:** We will first extract key elements from the longitudinal record, such as problem threads (diabetes, kidney disease), major events (admissions, surgeries, critical results), and trends in numeric data. Existing clinical tools, especially the QA agent from Aim1, will be used to detect important events and assemble a structured intermediate representation (problem clusters, timelines, and numeric summaries).
- **Configurable Summary Generation and Faithfulness:** An LLM-based agent will then generate summaries from this representation, supporting different modes (problem-oriented vs chronological), focus (specific condition or whole history), and time ranges. Prompts will explicitly restrict the agent to extracted facts and encourage citation of source snippets. We will experiment with sentence-level alignment or a secondary verification agent to flag or remove statements that cannot be matched to underlying notes or labs, prioritizing factual faithfulness over stylistic fluency.

For Aim 2ii (Diagnosis Support), we plan to:

- **Evidence-Backed Differential Suggestions:** The diagnosis support agent will have a current presentation (structured data plus a summary from Aim 2a) as the starting point and output a small set of candidate diagnoses or issues, each with supporting findings and potential next steps. The goal is not to replace clinical judgment but to surface plausible hypotheses with explicit rationales.
- **Knowledge and Case-Based Reasoning:** We will combine two information sources: a disease knowledge base encoding typical presentations and diagnostic criteria, and a similar-patient retrieval component that finds past cases with comparable patterns and their final diagnoses. The agent will parse key findings, query these sources, and generate a differential list with brief reasoning and optional "what to check next" suggestions. Constraints on list length and frequency priors (common before rare) will aim to avoid overlong or unrealistic differentials.
- **Integration with QA and Summarization:** Diagnostic support will be tightly integrated with other tools. Clinicians can use the QA agent from Aim 1 to probe hypotheses, while the Aim 3 DSL will support this interactive exploration. The diagnosis agent may also invoke summarization from Aim 2a, thus acting as an internal downstream task of both QA and summarization capabilities.

Evaluation (Aim 2): Summarization will be evaluated primarily via clinician review rather than purely automatic metrics. Physicians will rate summaries on completeness, conciseness, correctness, and usefulness, and we will compare against baselines such as last-note-only or generic LLM summarization. We will also assess whether access to our summaries improves downstream clinicians' task performance. For diagnosis support, we will test on retrospective cases with known outcomes, measuring whether the true diagnosis appears among the top suggestions and whether the reasoning aligns with the documented explanation. Clinicians will qualitatively rate the plausibility and clarity of the differential lists. By completing Aim 2, we will obtain robust summarization and diagnostic-reasoning modules that, together with the QA agent from Aim 1, will provide a set of reusable clinical components for the orchestration in Aim 3.

3.3 Aim 3: Design and Implementation of a Clinical Domain-Specific Language (DSL)

Objective: Create a high-level DSL that allows users (or systems) to specify complex workflows involving the tools from Aim 1 and Aim 2. The DSL will encode a clinician's intent as a simple script or command sequence, which the system executes by calling the appropriate agents at the backend.

- **DSL Design:** We will first define the syntax and core operations of the DSL, driven by representative use cases. Likely primitives include data retrieval (`GetNotes`, `GetLabs`), analysis (`Summarize`, `Diagnose`, `AnswerQuestion`), and lightweight control flow (sequencing, simple conditions). The language should be simple yet expressive, perhaps in a line-based or JSON-like form rather than fully free text. We will iterate on the command set with clinical experts to ensure it aligns with how they naturally think about tasks, even if they do not write DSL scripts directly.

- **Interpreter and Execution:** We will implement an interpreter that parses DSL scripts and dispatches them to the underlying agents. For example, `AnswerQuestion(record, "Does the patient have X?")` will call the Aim 1 QA agent. Variables and intermediate results will be passed between steps as the script executes. The interpreter will be implemented in Python and treat agents as callable functions or services, handling errors (e.g., missing data or failed calls) in a controlled way.

- **Combining Patient-Level and Expert-Level Evidence:** The DSL will support workflows that mix EHR retrieval with external knowledge. A typical script may filter the record, summarize it, retrieve guidelines, and pose a question combining both:

  ```
  notes   = GetNotes(record, filter="oncology", timeframe="last 2 years");
  summary = Summarize(notes, focus="cancer treatment response");
  guideline = GetGuideline(topic="cancer follow-up");
  answer  = AnswerQuestion(record,
      "Next follow-up recommendation", context=[summary, guideline]);
  ```

  This illustrates how patient-level summaries and expert-level guidance can be composed to answer higher-level clinical questions.

- **Natural Language to DSL:** To avoid expecting clinicians to write DSL code, we will explore an LLM-based translator that converts natural language requests into DSL scripts for common patterns. For example, a request like "Summarize the patient's cardiac history and relevant guidelines" could be mapped to a script similar to the example above. Here, the LLM acts as a Copilot that chooses which DSL commands to invoke and in what order.

- **Workflow Prototypes and User Feedback:** We will design several canonical DSL workflows (e.g., comprehensive case review, diagnostic drill-down, guideline adherence check) and test them end-to-end, verifying that each step's output is reasonable and that the overall workflow matches clinical expectations. We will present the DSL concept and sample outputs to clinicians or informaticians for qualitative feedback on usefulness and coverage, refining commands and abstractions if important tasks are hard to express.

Evaluation (Aim 3): We will evaluate the DSL primarily through realistic use-case demonstrations and user feedback rather than a single metric. For a set of multi-step clinical queries (e.g., "For this complex patient, what should I focus on today?"), we will show how the DSL-based system decomposes the request into summaries, checks, and questions, and ask clinicians whether the outputs capture the key issues they would consider. We will also estimate effort savings compared to using the EHR alone. The ultimate success criterion is improved usability. If clinicians can, via a UI of the DSL, obtain comprehensive analyses that previously required substantial manual effort or data science support, then the DSL will have achieved its goal of making advanced AI tools accessible, transparent, and clinically actionable.

Timeline We expect to spend about one year on each aim, with some overlap. Aim 1 will begin in Year 1, focusing on building the question taxonomy, micro-agent retrieval layer, and the first evidence-centric QA prototype evaluated on de-identified cases. Aim 2 will start in late Year 1 and run through Year 2, first developing the structured longitudinal representation and summarization agent, then adding the limited-scope diagnosis support functions and testing them on retrospective cohorts. Aim 3 will begin in Year 2 and continue into Year 3, during which we will design the DSL, implement the interpreter and safety checks, and integrate the QA, summarization, and diagnosis agents into end-to-end workflows that can be piloted with clinician users.

Novelty and Key Challenges This project is novel in both its technical approach and its view of how clinical AI should be used. Rather than building yet another single-task model, it proposes a unified, programmable framework in which evidence-centric QA, longitudinal summarization, and diagnostic reasoning are exposed as modular agents and composed via a clinical DSL. This explicit separation of intent (what) from implementation (how) is, to our knowledge, largely unexplored for longitudinal EHR reasoning. At the same time, the central challenges lie less in creating individual modules but more in making the DSL and its Copilot safe, reliable, and clinically adoptable. The first challenge is how to give the DSL a clear semantics so that every command has a predictable effect and how to guarantee that failures are visible (e.g., missing data, low-confidence outputs). We need to ensure that complex workflows remain debuggable rather than becoming another black box. The second challenge is how to ensure safe composition of agents. Once QA, summarization, and diagnostic modules can be chained into longer workflows, we need explicit rules and safeguards so that the Copilot respects model limits, exposes uncertainty instead of hiding it, and does not propose workflows that conflict with local policies or clinical standards. The third challenge is about governance and evolution. The DSL and its libraries must be easy to version, extend, and adapt to changing guidelines or site-specific practices, while still allowing clinicians and informatics teams to inspect and audit what the system is actually doing on their patients' data.

4 Related Work

4.1 Sequence and Representation Learning for EHRs

There are a lot of works studying how to represent longitudinal EHR data for prediction tasks. Doctor AI employed recurrent neural networks (RNNs) to predict future clinical events from past visits [8], while Deep Patient used autoencoders to learn general-purpose patient representations from heterogeneous EHR data and improved performance on multiple downstream tasks [23]. Transformer-based architectures further advanced this area. BEHRT treats visit sequences analogously to sentences and improves multi-disease prediction [18], and Med-BERT adapts BERT-based pretraining to structured EHR sequences, yielding gains on readmission and mortality prediction [30]. More recently, approaches such as EHR2Path serialize a patient's entire hospital course into long textual sequences so that an LLM can summarize context and predict future events [29].

These models demonstrate that powerful representations over complex, longitudinal records are possible, but they are typically evaluated on aggregate prediction benchmarks rather than on interactive tasks like free-form QA, longitudinal case review, or guideline checks. They also rarely expose structured, human-interpretable abstractions (e.g., problem threads, episodes) that can be inspected by clinicians or consumed by higher-level agents, motivating the need for longitudinal representations designed for interactive reasoning.

4.2 Clinical Question Answering and Summarization over EHRs

There is extensive work on extracting or generating information from clinical texts, particularly QA and summarization over patient records. As discussed in Section 2.1, resources such as emrQA provide large corpora of question–answer pairs grounded in EHRs [28], and neural architectures including BiDAF and BERT-based models have proved effective at clinical QA by treating it as text comprehension over notes [34, 3]. To answer questions requiring general medical knowledge, some systems augment EHR text with external knowledge bases, guidelines, or drug references [15, 7].

Clinical summarization has evolved from extractive and pattern-based methods to abstractive sequence-to-sequence models that generate more fluent summaries [33, 39]. LLMs, sometimes finetuned on clinical data, can distill key diagnoses, procedures, and medications into concise narratives [4]. However, faithfulness to the underlying record remains a major concern. LLM-generated summaries may omit subtle but important details or hallucinate unsupported content [9]. Recent work therefore introduces constraints and verification mechanisms to reduce hallucinations [37]. Retrieval-augmented generation (RAG) is one such strategy. By explicitly fetching relevant passages or similar prior cases, systems can improve factual accuracy and reduce hallucinations in both QA and summarization tasks [2, 25].

Overall, existing QA and summarization systems show that targeted information access and condensed overviews are achievable, but they are typically built as single-task models. They seldom unify patient-specific evidence and external knowledge into a coherent, inspectable reasoning process, nor do they provide a programmable interface for composing QA, summarization, and retrieval into multi-step workflows over longitudinal records.

4.3 Explainability, Evidence Grounding, and Domain-Specific Languages

Since clinical applications are safety-critical and regulated, AI systems must be explainable and governable. Work on explainable clinical AI has explored post-hoc explanations and chain-of-thought-style rationales for LLMs [1, 40, 22, 32], as well as explicit grounding in external medical knowledge, for example through citation of guidelines or literature and platforms such as OpenEvidence that synthesise evidence-based recommendations from trusted sources [27, 43, 12]. These methods go beyond opaque predictions by exposing rationales and references, but the underlying computation usually remains a single, monolithic model call, with limited control over how tasks are decomposed or how behaviour changes as models evolve.

Domain-specific languages (DSLs) offer a complementary route to transparency and control. In medical informatics, languages such as Arden Syntax and GLIF encode clinical rules and guidelines as machine-interpretable logic, enabling decision support systems that follow guideline-like steps [31, 26]. More recently, outside biomedicine, systems like AIDL and MetaGen combine DSLs with LLMs. Models propose or edit programs in a human-readable DSL, while a separate interpreter executes them, providing a clear contract between human intent and model behaviour [14, 21].

5 Conclusion and Impact

In summary, this research will develop a novel, explainable clinical framework to help clinicians make sense of longitudinal patient records. We will tackle the problems of information overload and AI hallucinations by grounding all outputs in evidence from both the patient's data and the medical knowledge base. Through Aim 1, we ensure that any question the clinician asks is answered factually with citations, thereby building trust. Aim 2 provides higher-level insights. Distilled patient-course summaries and diagnostic-reasoning support will mitigate frequent challenges clinicians face in routine practice. Aim 3 combines everything together with a DSL that empowers flexible, complex queries without exposing the user to the complexity behind them. If successful, this project will produce a working prototype of an AI clinical assistant that is far more transparent and reliable than current black-box medical models. The system will not only answer questions and generate narratives, but it will also show its work like a careful medical expert.

References

[1] Gwénolé Abgrall, Andre L Holder, Zaineb Chelly Dagdia, Karine Zeitouni, and Xavier Monnet. Should ai models be explainable to clinicians? *Critical Care*, 28(1):301, 2024.

[2] Mohammad Alkhalaf, Ping Yu, Mengyang Yin, and Chao Deng. Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *Journal of biomedical informatics*, 156:104662, 2024.

[3] Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. Question answering for electronic health records: Scoping review of datasets and models. *Journal of Medical Internet Research*, 26:e53636, 2024.

[4] Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, and Christian Lovis. Scientific evidence for clinical text summarization using large language models: scoping review. *Journal of Medical Internet Research*, 27:e68998, 2025.

[5] Guthrie S Birkhead, Michael Klompas, and Nirav R Shah. Uses of electronic health records for public health surveillance to advance public health. *Annual review of public health*, 36(1):345–359, 2015.

[6] Nathan Brake and Thomas Schaaf. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? *arXiv preprint arXiv:2404.06503*, 2024.

[7] Laura Cabello, Carmen Martin-Turrero, Uchenna Akujuobi, Anders Søgaard, and Carlos Bobed. Meg: Medical knowledge-augmented large language models for question answering. *arXiv preprint arXiv:2411.03883*, 2024.

[8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR, 2016.

[9] Xinsong Du, Zhengyang Zhou, Yifei Wang, Ya-Wen Chuang, Yiming Li, Richard Yang, Wenyu Zhang, Xinyi Wang, Xinyu Chen, John Lian, et al. Adapting generative large language models for electronic health record applications: A systematic review of methodologies, evaluation, and hallucinations. *medRxiv*, pages 2024–08, 2024.

[10] JaWanna Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008-2015. *ASTP Health IT Data Brief [Internet]*, 2016.

[11] Chun-Ju Hsiao, Esther Hing, and Jill Ashman. *Trends in electronic health record system use among office-based physicians, United States, 2007-2012*. Number 75. US Department of Health and Human Services, Centers for Disease Control and . . . , 2014.

[12] Ryan T Hurt, Christopher R Stephenson, Elizabeth A Gilman, Christopher A Aakre, Ivana T Croghan, Manpreet S Mundi, Karthik Ghosh, and Jithinraj Edakkanambeth Varayil. The use of an artificial intelligence platform openevidence to augment clinical decision-making for primary care physicians. *Journal of Primary Care & Community Health*, 16:21501319251332215, 2025.

[13] E Jamoom and N Yang. Table of electronic health record adoption and use among office-based physicians in the us, by state: 2015 national electronic health records survey. *Hyattsville, MD: National Center for Health Statistics*, 2016.

[14] Benjamin T Jones, Felix Hähnlein, Zihan Zhang, Maaz Ahmad, Vladimir Kim, and Adriana Schulz. A solver-aided hierarchical language for llm-driven cad design. *arXiv preprint arXiv:2502.09819*, 2025.

[15] Julien Khlaut, Corentin Dancette, Elodie Ferreres, Benani Alaedine, Herent Herent, and Pierre Manceron. Efficient medical question answering with knowledge-augmented question generation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 10–20, 2024.

[16] Linfeng Li, Peng Wang, Jun Yan, Yao Wang, Simin Li, Jinpeng Jiang, Zhe Sun, Buzhou Tang, Tsung-Hui Chang, Shenghui Wang, et al. Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine*, 103:101817, 2020.

[17] Lingyao Li, Jiayan Zhou, Zhenxiang Gao, Wenyue Hua, Lizhou Fan, Huizi Yu, Loni Hagen, Yongfeng Zhang, Themistocles L Assimes, Libby Hemphill, et al. A scoping review of using large language models (llms) to investigate electronic health records (ehrs). *arXiv preprint arXiv:2405.03066*, 2024.

[18] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

[19] Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, et al. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20, 2025.

[20] Siru Liu, Allison B McCoy, and Adam Wright. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4):605–615, 2025.

[21] Liane Makatura, Benjamin Jones, Siyuan Bian, and Wojciech Matusik. Metagen: A dsl, database, and benchmark for vlm-assisted metamaterial generation. *arXiv preprint arXiv:2508.17568*, 2025.

[22] Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1):148, 2024.

[23] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):26094, 2016.

[24] Zabir Al Nazi and Wei Peng. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI, 2024.

[25] Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. Retrieval-augmented generation (rag) in healthcare: A comprehensive review. *AI*, 6(9):226, 2025.

[26] Lucila Ohno-Machado, John H Gennari, Shawn N Murphy, Nilesh L Jain, Samson W Tu, Diane E Oliver, Edward Pattison-Gordon, Robert A Greenes, Edward H Shortliffe, and G Octo Barnett. The guideline interchange format: a model for representing guidelines. *Journal of the American Medical Informatics Association*, 5(4):357–372, 1998.

[27] David Oniani, Xizhi Wu, Shyam Visweswaran, Sumit Kapoor, Shravan Kooragayalu, Katelyn Polanska, and Yanshan Wang. Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 694–702. IEEE, 2024.

[28] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*, 2018.

[29] Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Matthias Keicher, and Nassir Navab. From ehrs to patient pathways: Scalable modeling of longitudinal health trajectories with llms. *arXiv preprint arXiv:2506.04831*, 2025.

[30] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

[31] Matthias Samwald, Karsten Fehre, Jeroen De Bruin, and Klaus-Peter Adlassnig. The arden syntax standard for clinical decision support: Experiences and directions. *Journal of biomedical informatics*, 45(4):711–718, 2012.

[32] Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20, 2024.

[33] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[34] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[35] Krishna Subedi. The reliability of llms for medical diagnosis: An examination of consistency, manipulation, and contextual awareness. *arXiv preprint arXiv:2503.10647*, 2025.

[36] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158, 2023.

[37] Ritchie Verma, Emily Alsentzer, Zachary Strasser, Leslie Chang, Kirollos Roman, Esteban Gershanik, Camellia Hernandez, Miguel Linares, Jorge Rodriguez, Durga Thakral, et al. Verifiable summarization of electronic health records using large language models to support chart review. *medRxiv*, pages 2025–06, 2025.

[38] Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*, 2020.

[39] Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 485–497, 2020.

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[41] Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. The shaky foundations of large language models and foundation models for electronic health records. *npj digital medicine*, 6(1):135, 2023.

[42] Chengyan Wu, Zehong Lin, Wenlong Fang, and Yuyan Huang. A medical diagnostic assistant based on llm. In *China Health Information Processing Conference*, pages 135–147. Springer, 2023.

[43] Velyn Wu and Jed Casauay. Openevidence. *Family Medicine*, 57(3):232, 2024.

[44] Zhenbang Wu, Anant Dadu, Mike Nalls, Faraz Faghri, and Jimeng Sun. Instruction tuning large language models to understand electronic health records. *Advances in Neural Information Processing Systems*, 37:54772–54786, 2024.

[45] Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. One llm is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*, 2023.

[46] Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. Llm-medqa: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*, 2024.

[47] Jinzhu Yang et al. Integrated application of llm model and knowledge graph in medical text mining and knowledge extraction. *Soc. Med. Health Manag*, 5:56–62, 2024.