Operation-Mechanism Alignment for Reliable Large Language Model Reasoning over Electronic Health Records

Guanyu Tao[1], **Siyao Wang**[1,2], **Yong Xue**[1], **Ashwani Tanwar**[1]
Yuting Ji[1], Kai Sun[1,2], Monica Mok[1], Marzana Chowdhury[1]
Deepa Gupta[1], Ashok Gupta[1], Jingqing Zhang[1*], Vibhor Gupta[1], Yike Guo[1,3]
[1]Pangaea Data Limited, UK and USA
[2]Data Science Institute, Imperial College London, London, UK
[3]Hong Kong University of Science and Technology, Hong Kong SAR, China

Abstract

Clinical reasoning over electronic health records (EHRs) involves heterogeneous operations, including text interpretation, numerical computation, temporal filtering, and guideline-based aggregation. However, many existing LLM-based approaches still cast these heterogeneous operations as a single end-to-end generation process, obscuring their different reliability requirements and making intermediate failures difficult to inspect. We therefore propose a framework based on operation-mechanism alignment that represents clinical reasoning as a directed acyclic graph of typed operations, where each node is assigned to the execution mechanism best suited to its reliability requirements. The framework also preserves structured evidence provenance for intermediate results. Across six clinician-annotated binary decision tasks, the framework outperforms direct prompting, single-step retrieval-augmented prompting, and chain-of-thought baselines, supporting operation-mechanism alignment as a practical design principle for reliable clinical reasoning over EHRs.

1 Introduction

Clinical reasoning over electronic health records (EHRs) often requires several different types of operations within the same task. For example, staging chronic kidney disease requires interpreting free-text nephrology notes to assess proteinuria, computing eGFR from a validated formula, selecting laboratory values within clinically relevant time windows, and combining the findings under guideline-defined logic. These steps have different reliability requirements. Numerical computation and temporal filtering require exact execution, because an arithmetic or date-filtering error can change the assigned disease stage. Proteinuria assessment, by comparison, depends on flexible interpretation of heterogeneous clinical language. Treating these steps as a single end-to-end generation problem obscures these operational differences and makes failures difficult to trace, inspect, or correct.

Common prompting-based approaches still cast clinical reasoning as a single inference pass. The model is prompted over either the full patient record or a retrieved subset of documents. Although retrieval-augmented generation can improve evidence relevance, the same generation step must still select evidence, perform calculations, apply clinical rules, and produce the final decision. This creates two practical problems. First, errors in intermediate operations, particularly numerical computation and temporal filtering, can propagate silently to the final answer. Second, the output is difficult to inspect because intermediate decisions and evidence assignments are not explicitly represented.

We argue that clinical reasoning should be decomposed into typed operations, including text interpretation, numerical computation, temporal filtering, and criteria aggregation. Different operation types should not be handled by the same generation step. Language interpretation can be handled by LLM-based evaluators, numerical and temporal operations by deterministic tools, and criteria aggregation by explicit logical operators. We call this principle *operation-mechanism alignment*. We implement this principle by representing clinical reasoning as a directed acyclic graph, where each node corresponds to a typed operation and each intermediate result carries structured evidence provenance. This makes the reasoning process inspectable at each step, rather than hidden behind a single generation call.

We evaluate the framework on clinician-annotated clinical reasoning tasks and compare it against direct prompting, single-step retrieval-augmented prompting, and chain-of-thought prompting baselines. Our contributions are:

*Correspondence to jzhang@pangaeadata.ai

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 605–616

(1) we formalize clinical reasoning over EHRs as a graph of typed operations with explicit operation-mechanism alignment, rather than single-pass generation; (2) we present a framework that routes evaluation, computation, and aggregation to distinct execution mechanisms while maintaining evidence provenance throughout; and (3) we provide an evaluation showing the benefits of this design over standard prompting and retrieval-based baselines.

We evaluate the framework on clinician-annotated clinical reasoning tasks and compare it with baselines including direct prompting, single-step retrieval-augmented prompting, and chain-of-thought prompting. Our contributions are: (1) we formalize EHR-based clinical reasoning as a graph of typed operations with explicit alignment between operation types and execution mechanisms; (2) we implement an executable framework that routes language evaluation, deterministic computation, temporal filtering, and logical aggregation to distinct mechanisms while preserving evidence provenance; and (3) we provide an evaluation showing the benefits of this structured design over standard prompting and retrieval-based baselines.

2 Related Work

Recent benchmarks study clinical reasoning with large language models in patient-specific EHR settings, requiring recovery of clinical state from records rather than decontextualized questions or pre-structured cases. EHRNoteQA and ER-REASON operationalize this setting through multi-note questions, emergency-department workflow tasks, longitudinal heterogeneous notes, and clinician- or physician-authored supervision (Kweon et al., 2024; Mehandru et al., 2025). DR.BENCH and MedHELM further show that clinical reasoning spans tasks with different operational demands, motivating methods beyond monolithic prompt-to-answer formulations, while related application studies apply LLMs directly to gold-labeled EHRs in largely one-pass prompting settings (Gao et al., 2023; Bedi et al., 2025; Zhang et al., 2024).

Retrieval-augmented generation (RAG) is widely used to improve grounding in medical question answering and clinical reasoning over EHRs. MIRAGE/MedRAG shows that RAG can outperform chain-of-thought prompting while exposing retrieval-specific issues such as context position effects (Xiong et al., 2024b). i-MedRAG and Self-BioRAG explore multi-round evidence acquisition and self-critique for complex questions and retrieval noise (Xiong et al., 2024a). In EHR settings, ExpRAG uses case-based retrieval from other patients' discharge reports, while EHR-RAG combines temporally and event-aware retrieval with iterative refinement for long-horizon prediction (Ou et al., 2025; Cao et al., 2026). However, these approaches still typically execute clinical reasoning over EHRs within a single generative mechanism, even when it involves temporal filtering, numeric aggregation, and guideline-like decision logic.

Temporal and longitudinal reasoning remains challenging for LLMs in clinical reasoning over EHRs. TIMER introduces a temporal benchmark and instruction-tuning method grounded in time-indexed patient-record segments, suggesting that time-awareness is not captured by standard instruction tuning (Cui et al., 2025). Long-context clinical summarization and prediction studies similarly show that even with longer context windows, RAG variants, and chain-of-thought prompting, models struggle to maintain chronological coherence over extended patient trajectories (Kruse et al., 2025). These findings suggest that some EHR reasoning subproblems, especially temporal filtering and longitudinal aggregation, may require explicit operators such as time-window queries rather than free-form generation.

Graph-structured orchestration and agentic tool use provide another adjacent direction. MedAgentBench offers a Fast Healthcare Interoperability Resources (FHIR)-compliant environment for benchmarking medical agents on patient-specific record workflows (Jiang et al.). DSPy provides graph- or pipeline-based composition patterns, ReAct and Toolformer study tool invocation, and Graph-of-Thought extends linear chains to graph-structured intermediate thoughts (Khattab et al., 2023; Yao et al., 2022; Schick et al., 2023; Besta et al., 2024). However, these systems mainly use graphs to organize reasoning steps or tool calls, rather than define operation semantics or use operation types to determine execution mechanisms.

Another concern in clinical reasoning over EHRs with LLMs is how outputs are grounded, verified, and linked to patient-record evidence. VeriFact studies patient-record-grounded verification of generated clinical narratives, while FactEHR shows that clinical fact decomposition varies across LLMs

and note types (Chung et al., 2025; Munnangi et al., 2024). Related evaluation work validates LLM-as-a-judge assessment of multi-document EHR summaries aligned with clinician rubrics (Croxford et al., 2025). Standards such as CDS Hooks and CPG-on-FHIR show how guideline logic and evidence links can be made computable and integrated with EHR workflows. However, much of this literature attaches evidence to final outputs through verification, evaluation, or post hoc citation, rather than preserving provenance throughout intermediate reasoning steps.

In summary, existing work has improved clinical reasoning over EHRs through prompting, retrieval, tool use, and output verification. To the best of our knowledge, this is the first work to model clinical reasoning over EHRs as heterogeneous, typed operations routed to mechanisms aligned with reliability requirements, while preserving structured provenance across intermediate results.

3 Approach

Our central design principle is *operation-mechanism alignment*: clinical reasoning over EHRs should be modeled as a composition of heterogeneous operations with different reliability requirements, rather than as a monolithic generation problem. In our framework, the execution mechanism depends on the operation type. Language-mediated clinical judgment is handled by LLM-based modules; numeric, temporal, and unit-sensitive computation is handled by deterministic tools; and criteria aggregation is handled by explicit logical operators. This hybrid neural-symbolic design preserves interpretive flexibility while reducing avoidable errors in exactly computable operations.

Throughout this section, we use a **chronic kidney disease (CKD) assessment** task as a running example. The clinical question is binary: *does this patient have advanced-stage CKD with rapid kidney function decline?* Clinically, this decision is driven primarily by the estimated glomerular filtration rate (eGFR): advanced stage corresponds to eGFR below a threshold, and rapid decline to the eGFR decline rate over time. Answering it therefore requires interpreting clinical notes and laboratory reports to determine whether the patient has persistent albuminuria as a supporting criterion (evaluation), computing current and historical eGFR values from laboratory measurements and estimating the annual decline rate as the central quantitative driver (computation), and combining these results under clinical guideline logic to produce a yes-or-no decision (aggregation). Relevant clinical details are introduced as each operation type is discussed.

3.1 Problem Formulation

We define a clinical reasoning task template as $T = (Q, C)$, where Q is a task query describing the clinical question to be answered and C is a set of clinical criteria and decision constraints relevant to the task. Given a patient record D, the goal is to predict a binary label $\hat{y} \in \{0, 1\}$.

In the CKD example, Q asks whether the patient has advanced-stage CKD with rapid decline. C includes an eGFR threshold for advanced stage, an annual eGFR decline threshold for rapid progression, albuminuria severity rules, and a persistence requirement ($\geq$3 months). D comprises laboratory results, clinical notes, and demographics.

To solve (T, D), we instantiate a directed acyclic graph $G = (V, E, \tau)$, where each node $v_i \in V$ is associated with a typed operation $\tau(v_i) \in \{\text{EVALUATE}, \text{COMPUTE}, \text{AGGREGATE}\}$ and edges encode dependency constraints. Each node is executed over node-specific evidence e_i derived from D for that subproblem, together with any required upstream outputs $\{z_j : (v_j, v_i) \in E\}$. For some nodes—particularly aggregation nodes that only compose upstream results—e_i may be empty and execution depends entirely on upstream outputs. The node output $z_i = (r_i, \mathcal{E}_i)$ is a structured intermediate result consisting of an operation result r_i and supporting evidence references $\mathcal{E}_i$.

- **Evaluation nodes** determine whether a clinical condition is satisfied using language-mediated reasoning over retrieved evidence; in the running example, this includes assessing persistent albuminuria from notes and laboratory reports.
- **Computation nodes** invoke deterministic operators for numeric, temporal, or unit-sensitive operations; in the running example, these compute current and historical eGFR values and the annualized decline rate.
- **Aggregation nodes** combine upstream outputs into higher-level criterion decisions or the final prediction $\hat{y}$; in the running example, eGFR-based staging serves as the pri-

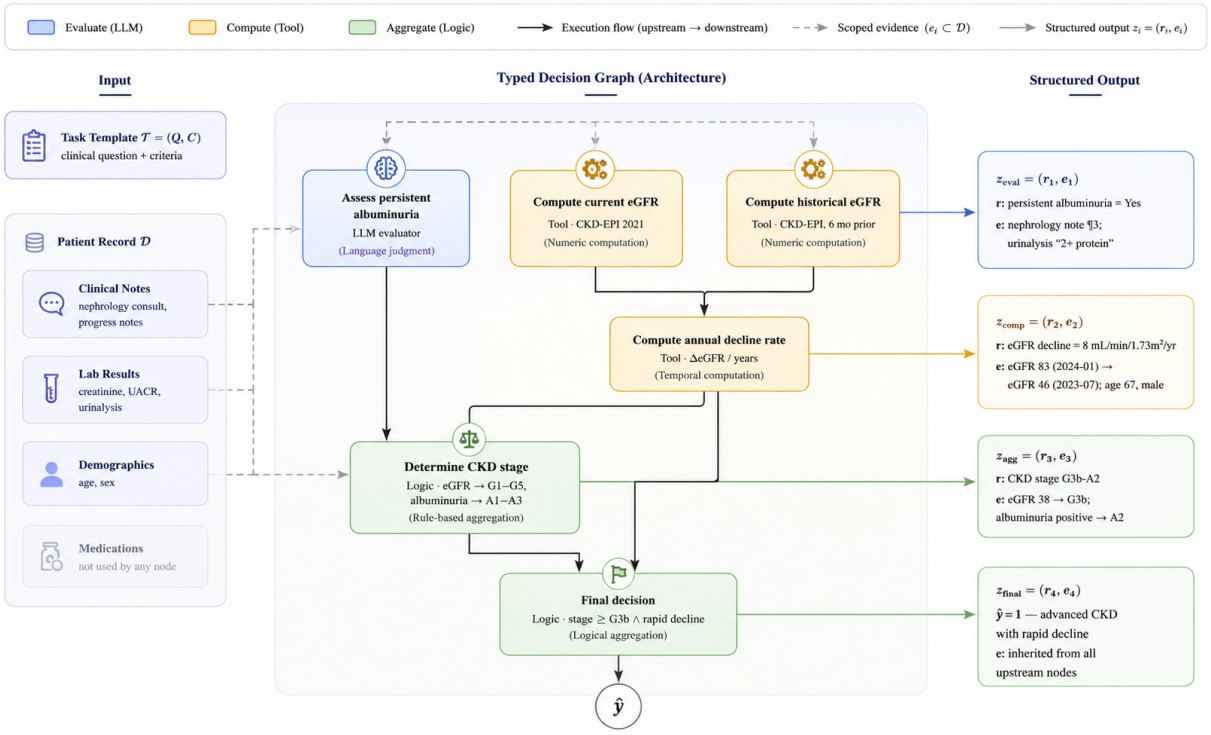

Figure 1: Overview of our framework. A clinical reasoning task template and patient record are instantiated as a typed decision graph. Each node is routed to the execution mechanism aligned with its operation type: LLM-based evaluators for language-mediated judgment, deterministic tools for numeric and temporal computation, and explicit logical operators for aggregation. Node outputs are structured and evidence-grounded, enabling provenance-preserving execution traces. Dashed arrows indicate scoped evidence: different nodes operate over different subsets of the patient record, rather than a single shared context.

mary backbone, combined with the albuminuria category and the progression criterion under guideline logic.

Figure 1 illustrates this decomposition for the CKD running example, showing how different patient record sources provide scoped evidence to specific nodes, how nodes are routed to distinct execution mechanisms, and how each node output carries structured evidence references.

3.2 Task Decomposition and Execution

A key property of the graph is that each node operates on a *scoped* subproblem with its own evidence requirements, rather than on a single shared context. This contrasts with retrieval-augmented approaches, where one context bundle must support evidence identification, calculation, temporal reasoning, and final decision making at once. In our framework, evidence is constructed locally: proteinuria evaluation uses clinical notes and urinalysis reports, whereas eGFR computation uses structured laboratory observations within defined temporal windows.

Edges in G encode explicit data dependencies. The eGFR decline computation depends on both the current and historical eGFR computations; the final aggregation depends on the eGFR-based staging, the decline rate, and the albuminuria evaluation. This dependency structure determines a partial execution order: independent nodes—such as the albuminuria evaluation and the two eGFR computations—can execute in parallel, while dependent nodes wait for upstream results.

The graph G is instantiated from the task template T before execution. Each criterion in C maps to one or more nodes in V, and dependencies are derived from the logical relationships among criteria. This turns a declarative task specification into an executable plan, separating *what* must be decided from *how* each subproblem is resolved.

3.3 Operation-Mechanism Alignment

The three node types are not merely a descriptive taxonomy; they determine which execution mechanism is invoked. This section describes how each mechanism operates and why the routing matters.

Evaluation: LLM-based clinical judgment. Evaluation nodes handle operations that require interpreting diverse and often ambiguous clinical language. In the running example, determining whether a patient has persistent albuminuria may involve reasoning over a urinalysis showing "2+ protein," a nephrology consult noting "persistent albuminuria documented over the past 4 months, now in the moderately increased range," and structured uACR (urine albumin-to-creatinine ratio) measurements—all expressed differently and scattered across the record. These judgments resist rule-based encoding because the space of possible phrasings is open-ended. LLM-based reasoning is therefore the appropriate mechanism for evaluation nodes: it provides the flexibility to interpret heterogeneous clinical text, at the cost of being non-deterministic.

Computation: deterministic tools. Computation nodes handle operations where correctness is exactly verifiable and LLM generation introduces unnecessary risk. In the running example, computing the patient's estimated glomerular filtration rate (eGFR) requires applying a validated clinical formula (CKD-EPI 2021) to serum creatinine, age, and sex. A small numeric error—an eGFR of 46 versus 38—can shift the patient between adjacent disease stages, directly changing the clinical decision. The same concern applies to unit conversion (creatinine reported in mg/dL versus μmol/L), temporal filtering (selecting laboratory values within a specific time window), and rate-of-change computation (pairing values across time points and annualizing the difference). These operations are delegated to validated deterministic tools: clinical calculators for formula evaluation, unit-aware conversion routines, temporal window filters, and arithmetic operators for derived quantities.

Aggregation: explicit logical composition. Aggregation nodes combine upstream results under predefined clinical logic rather than open-ended synthesis. In the CKD example, the final staging decision follows clinical guidelines: map eGFR to a disease stage, combine it with the albuminuria category, and check the progression criterion. This composition is a deterministic function of its inputs and should not be subject to LLM variability. Aggregation therefore uses explicit logical operators—boolean combinations, threshold comparisons, and lookup tables—to produce the final prediction $\hat{y}$.

Why alignment matters. The key claim is not that any one mechanism is universally superior, but that misalignment between operation type and execution mechanism introduces avoidable errors. Delegating eGFR computation to an LLM risks arithmetic mistakes in a context where precision determines the clinical outcome. Conversely, attempting to encode proteinuria assessment as a rule-based system would require enumerating an impractical number of textual patterns. Operation-mechanism alignment is the principle that each node should be executed by the mechanism whose failure modes are least harmful for that operation's reliability requirements.

3.4 Evidence-Grounded Output

Every node output $z_i = (r_i, \mathcal{E}_i)$ includes structured evidence references $\mathcal{E}_i$ that link the result back to specific records in the patient record D. These references are not post-hoc citations but are constructed during execution: evaluation nodes record which passages informed their judgment, computation nodes record which laboratory values and timestamps were used, and aggregation nodes inherit the evidence from all upstream nodes they compose.

This design serves two purposes. First, it supports clinical auditability: a reviewer can trace any component of the final decision back to its source evidence without re-running the system. Second, it makes unsupported conclusions structurally detectable—if a node produces a result without corresponding evidence references, this is visible in the output structure rather than hidden in a free-text rationale.

In the running example, the final decision that a patient has advanced-stage CKD with rapid decline is accompanied by the specific laboratory values and their timestamps used to compute both eGFR values, the clinical note passages that supported the albuminuria assessment, and the guideline rule that combined these into the staging decision. Each component is individually inspectable.

4 Experiments

4.1 Data and Tasks

We evaluate the framework on de-identified EHR data from selected subsets of MIMIC-III and MIMIC-IV, together with a de-identified partner-hospital dataset (from Cone Health). The evaluation covers six binary clinical decision tasks under

a shared annotation protocol.

COPD exacerbation history. The first task asks whether a patient has a confirmed recent COPD exacerbation history, defined by recurrent outpatient exacerbation treatment, at least one inpatient hospitalization for acute COPD exacerbation, or repeated prescription courses of systemic corticosteroids or antibiotics for exacerbation within the past year. The task is evaluated on two cohorts under the same task definition: 151 MIMIC-III patients (108 positive, 43 negative), and 100 patients from the de-identified Cone Health dataset (18 positive, 82 negative).

Asthma exacerbation risk. The second task determines whether a patient with confirmed asthma is at risk of exacerbation or asthma-related death under the Global Initiative for Asthma (GINA) 2024 criteria (Global Initiative for Asthma, 2024), including severe exacerbation history, poor symptom control, inadequate ICS treatment, high SABA use, low FEV_1, relevant comorbidities, or prior near-fatal events. The cohort comprises 50 MIMIC-IV patients (46 positive, 4 negative).

Cancer cachexia screening (Fearon). The third task applies the cachexia criteria of Fearon et al. (Fearon et al., 2011), requiring an underlying cancer diagnosis together with weight loss $>5\%$ over 6 months, or BMI <20 kg/m^2 with weight loss $>2\%$, or sarcopenia with weight loss $>2\%$. The cohort comprises 50 patients (15 positive, 35 negative) from MIMIC-III and IV.

Cancer cachexia screening (Evans). The fourth task applies the cachexia criteria of Evans et al. (Evans et al., 2008), requiring an underlying chronic disease including cancer, involuntary weight loss $\geq 5\%$ in the past 12 months, and at least three of five additional features: decreased muscle strength, anorexia, low fat-free mass index, abnormal biochemistry, or decreased functional status. The cohort comprises 50 patients (8 positive, 42 negative) from MIMIC-III and IV.

Confirmed CKD diagnosis. The fifth task asks whether the patient has an explicit documented diagnosis of chronic kidney disease (CKD) in the clinical record. The cohort comprises 100 patients from the de-identified Cone Health dataset (25 positive, 75 negative).

At-risk CKD identification. The sixth task asks whether a patient is at risk of CKD but does not yet have a confirmed diagnosis. Positive cases require at least one recognized CKD risk factor or non-chronic kidney abnormality, including hypertension, diabetes mellitus, cardiovascular disease, prior AKI/AKD, reduced eGFR, elevated albuminuria, urine sediment abnormalities, hematuria, histology abnormalities, tubular or electrolyte disorders, or structural kidney abnormalities on imaging, while excluding patients who meet confirmed CKD criteria. The cohort comprises 100 patients from the same Cone Health dataset (42 positive, 58 negative).

Gold labels are created by clinician annotators on curated subsets using task-specific criteria. For each case, annotators provide the expected binary label and, when applicable, supporting evidence from the record. Each case is reviewed by two clinician annotators independently and any disagreement is resolved by a third clinician annotator or with consensus.

All methods are evaluated using GPT-5.4 as the primary model; direct-prompting sensitivity to model choice is additionally assessed with GPT-5.2 and Claude Opus 4.6 (see Appendix A).

4.2 Baselines

We compare the proposed framework against three baseline settings that use the same underlying patient records and task definitions, but differ in how evidence is constructed and how reasoning is organized:

- **Direct prompting**: The patient record is provided to the model in a single pass with a task instruction, without explicit task decomposition, intermediate structure, or tool use. This baseline tests whether a general-purpose prompt over the available patient context is sufficient for the binary decision.

- **Single-step RAG**: Task-relevant evidence is retrieved once using the task query, and the retrieved context is then provided to the model for end-to-end reasoning in a single generation step. This baseline isolates the value of retrieval while still requiring one model call to perform evidence interpretation, calculation, temporal reasoning, and final decision making jointly.

- **Chain-of-thought (CoT) prompting**: The model receives the same retrieved context as

in the single-step RAG baseline, but is explicitly prompted to reason step-by-step before producing the final label. This baseline tests whether prompting for intermediate natural-language reasoning alone is sufficient, without explicit task-scoped execution or deterministic tool support.

For fairness, all methods use the same underlying task definitions, model set, and binary output schema. When the full patient record exceeds the model context budget, the direct-prompting baseline uses a shared record-linearization pipeline with budget-matched truncation, while retrieval-based baselines and our framework use the same task description and retriever configuration wherever applicable. These baselines are therefore designed to separate the effects of retrieval, explicit reasoning structure, and tool-supported execution rather than by differences in model access or output format.

4.3 Metrics

Because our gold labels are binary clinical decisions, we use sensitivity, specificity, precision, and F1 as the primary task metrics. Accuracy is included as a supplementary metric.

4.4 Results

Table 1 reports the main comparison across reasoning strategies using GPT-5.4 on all six tasks.

On COPD exacerbation, the framework achieves the best F1 and accuracy on both cohorts. On MIMIC-III, it reaches F1 = .893 and accuracy = .861, with a marked increase in specificity (.977 versus .349 for direct prompting and .186 for both retrieval-based baselines). The same pattern holds on Cone Health, where the framework attains F1 = .882 and specificity = .988, while baseline specificities range from .780 to .866; the four methods identify the same set of 15 true positives but the framework yields only 1 false positive versus 11–18 for the baselines. The consistent specificity advantage across two independent datasets is consistent with baseline methods over-predicting positive cases when temporal and diagnostic attribution must be resolved precisely, whereas deterministic temporal filtering and explicit diagnostic-attribution logic are associated with fewer false positives. On asthma exacerbation risk, the framework again achieves the best F1 (.957) and accuracy (.920) and is the only method with nonzero specificity (.500, 2/4). This specificity estimate should be interpreted cautiously because the task contains only four negative cases.

On cancer cachexia screening, the framework achieves the highest F1 and accuracy for both Fearon and Evans criteria. For the Fearon criteria, it attains F1 = .839 and accuracy = .900, while chain-of-thought prompting performs worst (F1 = .581). This result is consistent with the difficulty of applying quantitative thresholds, such as weight loss, BMI, and sarcopenia indices, through free-text step-by-step reasoning alone. For the more stringent Evans criteria, the framework reaches sensitivity = 1.00 and F1 = .727, compared with the best baseline F1 of .480, consistent with the difficulty of aggregating multiple concurrent conditions in a single-pass setup.

On the CKD tasks evaluated on Cone Health data, the framework achieves the highest F1 and accuracy on both tasks. For confirmed CKD diagnosis, it reaches F1 = .909 with perfect sensitivity (1.000), recovering all 25 confirmed cases; chain-of-thought and single-step RAG both reach .960 sensitivity but with one missed case each, while direct prompting lags further (F1 = .792, sensitivity = .760). Specificity is comparable across the four methods (.920–.947) on this task, so the framework's advantage concentrates on sensitivity rather than specificity. For at-risk CKD identification, the framework attains F1 = .900 and accuracy = .920, with the highest precision (.947) and specificity (.966); chain-of-thought is competitive on sensitivity (.881 versus .857 for the framework), but produces more false positives. Together, these results suggest that detailed multi-criteria CKD definitions can challenge single-pass reasoning, while decomposition into independent criteria evaluations is associated with stronger performance in these tasks.

Model sensitivity. As a secondary check, we repeated direct prompting with GPT-5.2 and Claude Opus 4.6. Performance varied by task and model, but model substitution did not eliminate the limitations of single-pass prompting; full results are provided in Appendix A. We therefore treat model choice as an important but orthogonal factor.

4.5 Discussion

The results reveal several consistent patterns across tasks.

Specificity as the differentiator. Across the six tasks, the proposed framework consistently achieves the highest or near-highest specificity.

Task	Method	Sens.	Spec.	Prec.	F1	Acc.
COPD Exacerbation (MIMIC-III)	Direct prompting	**.843**	.349	.765	.802	.702
	Single-step RAG	**.843**	.186	.722	.778	.656
	Chain-of-thought	.806	.186	.713	.757	.629
	Ours	.815	**.977**	**.989**	**.893**	**.861**
COPD Exacerbation (Cone Health)	Direct prompting	.778	.780	.438	.560	.780
	Single-step RAG	**.833**	.866	.577	.682	.860
	Chain-of-thought	**.833**	.829	.517	.638	.830
	Ours	**.833**	**.988**	**.938**	**.882**	**.960**
Asthma Risk†	Direct prompting	.913	.000	.913	.913	.840
	Single-step RAG	**.935**	.000	.915	.925	.860
	Chain-of-thought	.913	.000	.913	.913	.840
	Ours	.957	**.500**	**.957**	**.957**	**.920**
Cachexia (Fearon)	Direct prompting	.800	.743	.571	.667	.760
	Single-step RAG	.733	.829	.647	.688	.800
	Chain-of-thought	.600	.800	.563	.581	.740
	Ours	**.867**	**.914**	**.813**	**.839**	**.900**
Cachexia (Evans)	Direct prompting	.875	.452	.233	.368	.520
	Single-step RAG	.875	.619	.304	.452	.660
	Chain-of-thought	.750	.738	.353	.480	.740
	Ours	**1.00**	**.857**	**.571**	**.727**	**.880**
Confirmed CKD	Direct prompting	.760	**.947**	.826	.792	.900
	Single-step RAG	.960	.933	.828	.889	.940
	Chain-of-thought	.960	.920	.800	.873	.930
	Ours	**1.000**	.933	**.833**	**.909**	**.950**
At-Risk CKD	Direct prompting	.833	.862	.814	.824	.850
	Single-step RAG	.833	.948	.921	.875	.900
	Chain-of-thought	**.881**	.931	.902	.892	.910
	Ours	.857	**.966**	**.947**	**.900**	**.920**

Table 1: Main results across reasoning strategies (GPT-5.4). Bold indicates the best result per metric per task. Sens. = sensitivity, Spec. = specificity, Prec. = precision, Acc. = accuracy. †High class imbalance (46/4); specificity estimates are unstable due to very few negative cases.

This is most pronounced in COPD exacerbation across both cohorts (.977 vs. .349 for direct prompting on MIMIC-III; .988 vs. .780–.866 on Cone Health), cachexia under Evans criteria (.857 vs. .452), and asthma exacerbation risk, where all three baselines achieve zero specificity while the framework is the only method that correctly rejects any negative cases. On at-risk CKD, the framework reaches .966 specificity versus .862–.948 for baselines, with .947 precision versus .814–.921, suggesting it is more conservative in rejecting borderline cases that satisfy only one or two soft risk factors. On confirmed CKD, specificity differences across methods are small (.920–.947); the framework's lead instead concentrates on sensitivity (1.000 versus .760–.960), recovering all confirmed cases. The framework's explicit criteria decomposition and deterministic aggregation logic enforce the full set of diagnostic requirements (e.g., temporal windows, primary-diagnosis attribution, multi-condition conjunctions), preventing the model from short-circuiting to an affirmative answer when only partial evidence is present.

Structured execution outperforms unstructured reasoning. Chain-of-thought (CoT) prompting, which encourages step-by-step reasoning in natural language, does not consistently improve over direct prompting and in some cases degrades performance. On cachexia (Fearon), CoT achieves the lowest F1 of any method (.581 vs. .667 for direct prompting). This suggests that prompting the model to "think step by step" does not substitute for routing quantitative operations to deterministic tools. When intermediate steps involve numeric thresholds, weight-loss calculations, or temporal window enforcement, natural-language reasoning introduces opportunities for error that structured execution avoids.

Task complexity modulates framework advantage. The framework's advantage is largest on tasks requiring multi-step quantitative criteria and smallest on simple diagnostic lookup. These tasks share a common structure: evaluating multiple in-

dependent clinical criteria with their own evidence, chronicity, and exclusion rules. Providing such detailed criteria as a single instruction degrades baseline performance, which the framework avoids by decomposing the specification into independent, single-feature evaluations.

Relation to tool-using agents. Autonomous tool-using agents (e.g., ReAct-style systems (Yao et al., 2022) or Toolformer (Schick et al., 2023)) address a distinct problem: discovering task decomposition at inference time when the decomposition is unknown. In our setting the decomposition is known—each task graph is derived from a validated clinical guideline (e.g., GOLD 2025, KDIGO 2024). Letting an agent re-discover this structure at inference replaces a vetted plan with one assembled on the fly, sacrificing auditability without addressing how each step should be executed. The latter question—operation-mechanism alignment—is orthogonal to tool use: even a ReAct agent with calculator access must still decide when to invoke the calculator versus reason in natural language, which is the same alignment decision our framework resolves ahead of time. Our contribution therefore complements rather than substitutes for autonomous tool use.

5 Conclusion

We presented a framework that structures clinical reasoning over EHRs as a directed acyclic graph of typed operations, where each node is routed to the execution mechanism—LLM-based evaluation, deterministic computation, or explicit logical aggregation—best suited to its reliability requirements. Every intermediate result carries structured evidence provenance, making the reasoning process inspectable at each step.

Experiments on six clinician-annotated tasks across four disease areas show that this design consistently improves specificity over direct prompting, retrieval-augmented, and chain-of-thought baselines, with the largest gains on tasks involving multi-step quantitative criteria such as temporal filtering and threshold evaluation. The results also show that the framework's advantage diminishes on simpler diagnostic-lookup tasks, suggesting that operation-mechanism alignment is most valuable when clinical criteria demand heterogeneous reasoning operations that single-pass generation handles unreliably.

These findings support a general design principle: clinical decision support systems should match each reasoning operation to the mechanism whose failure modes are least harmful for that operation, rather than delegating all operations to a single generative model.

Limitations

Evaluation scope. Our evaluation covers six clinician-annotated binary decision tasks drawn from MIMIC-III, MIMIC-IV, and a de-identified partner-hospital dataset across four disease areas. Cohort sizes (50–151 patients per task) and the class imbalance in some tasks (e.g., 46/4 for asthma exacerbation risk) limit the precision of absolute per-task estimates; we therefore focus claims on relative comparisons across methods. Although clinicians were involved in task design and annotation, the auditability benefits of provenance traces remain to be evaluated systematically, including their effect on review efficiency, error detection, and decision review. Broader validation across additional institutions, languages, disease areas, and non-binary decision formats is a natural next step.

Manual task template construction. Task templates—including clinical criteria, node types, and dependency structure—are authored by domain experts for each task. This is a deliberate design choice: explicit expert-authored templates prioritize auditability, clinical face validity, and adherence to established guideline logic. At the same time, extending the framework to a new clinical task currently requires human authoring effort. Automating template construction from guideline text, clinical pathways, or existing decision-support artifacts is a promising extension that would improve scalability.

Mechanism-level ablation and error decomposition. Our experiments evaluate the framework holistically against end-to-end baselines, rather than isolating the contribution of each execution mechanism. A finer-grained ablation—holding the task graph fixed while varying which mechanism handles each node type, for instance routing computation nodes through an LLM rather than a deterministic tool—would more directly test the operation-mechanism alignment claim. We also do not report a per-node-type decomposition of residual errors, which would help localize failures to evaluation ambiguity, computation edge cases, or upstream error propagation. Both are natural follow-ups that the typed graph structure is de-

signed to support.

Empirical comparison with tool-using agents. We do not include a direct empirical comparison with autonomous tool-using agents on our task suite. As discussed above, we view such systems as complementary to operation-mechanism alignment rather than as substitutable baselines. We also note that "tool-using agents" is not a single configuration but a design space spanning tool selection, agent loop (e.g., ReAct, Toolformer, function-calling, code interpreter), planning strategy, error handling, and stopping criteria, each of which can substantially affect performance. Designing a comparison that controls these factors fairly is a non-trivial undertaking that we leave to future work.

Computational cost and latency. Because the framework decomposes a single decision into multiple typed nodes executed across heterogeneous mechanisms, end-to-end inference incurs higher computational cost and latency than single-pass prompting, which we do not quantify in this work. Systematic characterization of the accuracy–cost trade-off—including strategies for adaptive depth or selective decomposition—is an important practical follow-up for deployment settings where throughput and API cost are binding constraints.

Acknowledgments

We gratefully acknowledge Cone Health for providing the de-identified, real-world electronic health record (EHR) data used in this study.

References

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, and 1 others. 2025. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 17682–17690.

Lang Cao, Qingyu Chen, and Yue Guo. 2026. Ehr-rag: Bridging long-horizon structured electronic health records and large language models via enhanced retrieval-augmented generation. *arXiv preprint arXiv:2601.21340*.

Philip Chung, Akshay Swaminathan, Alex J Goodell, Yeasul Kim, S Momsen Reincke, Lichy Han, Ben Deverett, Mohammad Amin Sadeghi, Abdel-Badih Ariss, Marc Ghanem, and 1 others. 2025. Verifying facts in patient care documents generated by large language models using electronic health records. *NEJM AI*, 3(1):AIdbp2500418.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Hejie Cui, Alyssa Unell, Bowen Chen, Jason Alan Fries, Emily Alsentzer, Sanmi Koyejo, and Nigam H Shah. 2025. Timer: Temporal instruction modeling and evaluation for longitudinal clinical records. *npj Digital Medicine*, 8(1):577.

William J Evans, John E Morley, Josep Argilés, Connie Bales, Vickie Baracos, Denis Guttridge, Aminah Jatoi, Kamyar Kalantar-Zadeh, Herbert Lochs, Giovanni Mantovani, and 1 others. 2008. Cachexia: a new definition. *Clinical nutrition*, 27(6):793–799.

Kenneth Fearon, Florian Strasser, Stefan D Anker, Ingvar Bosaeus, Eduardo Bruera, Robin L Fainsinger, Aminah Jatoi, Charles Loprinzi, Neil MacDonald, Giovanni Mantovani, and 1 others. 2011. Definition and classification of cancer cachexia: an international consensus. *The lancet oncology*, 12(5):489–495.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M Churpek, and Majid Afshar. 2023. Dr. bench: Diagnostic reasoning benchmark for clinical natural language processing. *Journal of biomedical informatics*, 138:104286.

Global Initiative for Asthma. 2024. Global strategy for asthma management and prevention. 2024 update. Accessed: 2026-04-17.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: A realistic virtual ehr environment to benchmark medical llm agents, 2025. *URL https://arxiv. org/abs/2501.14654*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025. Large language models with temporal reasoning for longitudinal clinical summarization and prediction. In *Findings of ACL. EMNLP. Conference on Empirical Methods in Natural Language Processing*, volume 2025, pages 20715–20735.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jeewon Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.

Nikita Mehandru, Niloufar Golchini, David Bamman, Travis Zack, Melanie F Molina, and Ahmed Alaa. 2025. Er-reason: A benchmark dataset for llm-based clinical reasoning in the emergency room. *arXiv preprint arXiv:2505.22919*.

Monica Munnangi, Akshay Swaminathan, Jason Alan Fries, Jenelle Jindal, Sanjana Narayanan, Ivan Lopez, Lucia Tu, Philip Chung, Jesutofunmi A Omiye, Mehr Kashyap, and 1 others. 2024. Factehr: A dataset for evaluating factuality in clinical notes using llms. *arXiv preprint arXiv:2412.12422*.

Justice Ou, Tinglin Huang, Yilun Zhao, Ziyang Yu, Peiqing Lu, and Rex Ying. 2025. Experience retrieval-augmentation with electronic health records enables accurate discharge qa. *arXiv preprint arXiv:2503.17933*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.

G Xiong, Q Jin, X Wang, M Zhang, Z Lu, and A Zhang. 2024a. Improving retrieval-augmented generation in medicine with iterative follow-up questions. arxiv; 2024. *arXiv preprint arXiv:2310.19988*.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024b. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Jingqing Zhang, Kai Sun, Akshay Jagadeesh, Parastoo Falakaflaki, Elena Kayayan, Guanyu Tao, Mahta Haghighat Ghahfarokhi, Deepa Gupta, Ashok Gupta, Vibhor Gupta, and 1 others. 2024. The potential and pitfalls of using a large language model such as chatgpt, gpt-4, or llama as a clinical assistant. *Journal of the American Medical Informatics Association*, 31(9):1884–1891.

A Model Sensitivity Under Direct Prompting

Table 2 examines sensitivity to model choice under the direct-prompting setting. Unlike Table 1, which fixes GPT-5.4 and compares reasoning strategies under a controlled setup, this analysis varies the underlying model while keeping the prompting paradigm unchanged.

Performance under direct prompting varies across models and tasks. On COPD exacerbation evaluated on MIMIC-III, all three models have identical specificity (.349), suggesting that the false-positive pattern is a limitation of direct prompting rather than a model-specific weakness. On the same task evaluated on Cone Health, by contrast, the three models occupy different points on the sensitivity-specificity trade-off (GPT-5.2 conservative at .556/.951, GPT-5.4 balanced at .778/.780, Claude Opus 4.6 over-predicting at .722/.756); the framework in Table 1 (.833/.988) dominates all three on both dimensions, indicating that the gains are not attributable to a particular model's operating point. On the cachexia tasks, between-model differences are smaller than the improvement from replacing direct prompting with our structured framework. On the CKD tasks, direct-prompting performance is more consistent across models, but no single model reaches the precision or F1 of the structured framework.

Overall, the effect of model choice is task-dependent. We therefore interpret Table 2 as a sensitivity analysis showing that changing the underlying model does not by itself resolve the limitations of single-pass prompting, and that the gains in Table 1 cannot be attributed to model substitution alone.

Task	Model	Sens.	Spec.	Prec.	F1	Acc.
COPD Exacerbation (MIMIC-III)	GPT-5.2	.843	.349	.765	.802	.702
	GPT-5.4	.843	.349	.765	.802	.702
	Claude Opus 4.6	.870	.349	.771	.817	.722
COPD Exacerbation (Cone Health)	GPT-5.2	.556	.951	.714	.625	.870
	GPT-5.4	.778	.780	.438	.560	.780
	Claude Opus 4.6	.722	.756	.394	.510	.750
Asthma Risk[†]	GPT-5.2	.913	.250	.933	.923	.860
	GPT-5.4	.913	.000	.913	.913	.840
	Claude Opus 4.6	1.00	.000	.920	.958	.920
Cachexia (Fearon)	GPT-5.2	.800	.800	.632	.706	.800
	GPT-5.4	.800	.743	.571	.667	.760
	Claude Opus 4.6	.733	.743	.550	.629	.740
Cachexia (Evans)	GPT-5.2	.750	.595	.261	.387	.620
	GPT-5.4	.875	.452	.233	.368	.520
	Claude Opus 4.6	.625	.762	.333	.435	.740
Confirmed CKD	GPT-5.2	.720	.946	.818	.766	.880
	GPT-5.4	.760	.947	.826	.792	.900
	Claude Opus 4.6	.760	.960	.864	.809	.910
At-Risk CKD	GPT-5.2	.805	.862	.805	.805	.830
	GPT-5.4	.833	.862	.814	.824	.850
	Claude Opus 4.6	.857	.897	.857	.857	.880

Table 2: Model sensitivity under direct prompting. Performance varies across models and tasks, indicating that model choice matters, but this variation does not by itself resolve the limitations of single-pass prompting observed in Table 1. [†]High class imbalance (46/4).