Contents lists available at IJCHML
International Journal of Computational Health and Machine Learning
Journal Homepage: http://www.ijchml.com/
Volume 4, No. 2, 2026

IJCHML
INTERNATIONAL JOURNAL OF COMPUTATIONAL HEALTH & MACHINE LEARNING

Checkpoint-Augmented Clinical Reasoning: Applying Recoverable Program-of-Thought Frameworks to Medical Diagnostic Pipelines

Yan Yang[1], **Naim Chowdhury**[2]
[1] *Department of Health Informatics, Sun Yat-sen University*
[2] *Department of Health Informatics, Shahjalal University of Science and Technology*

ARTICLE INFO

Received: 06/10/2026
Revised: 06/28/2026
Accepted: 07/04/2026

Keywords:
Clinical Decision Support, Program-of-Thought Reasoning, Checkpoint Recovery, Medical Diagnostic Pipelines, Large Language Models, Fault-Tolerant Inference, Structured Clinical Reasoning

ABSTRACT

Medical diagnostic reasoning demands both systematic rigor and robust error recovery, yet contemporary large language model (LLM) pipelines frequently lack mechanisms to detect and correct intermediate inferential failures before they propagate into final clinical recommendations. We introduce **Checkpoint-Augmented Clinical Reasoning** (CACR), a novel framework that embeds structured verification checkpoints within a recoverable Program-of-Thought (PoT) architecture, enabling transparent, auditable, and self-correcting diagnostic chains across multi-step clinical decision pipelines.

CACR operationalizes diagnostic reasoning as a sequence of formally defined computational steps, each guarded by a lightweight verifier module that evaluates logical consistency, evidential sufficiency, and differential plausibility prior to downstream propagation. When a checkpoint detects an anomalous or contradictory state, the framework initiates a targeted rollback to the nearest valid reasoning node, resamples the affected inference segment, and resumes forward execution—preserving both computational efficiency and clinical safety. This recoverable execution model draws inspiration from fault-tolerant distributed systems and adapts their guarantees to the probabilistic, knowledge-intensive domain of clinical medicine.

We evaluate CACR on three benchmark diagnostic corpora spanning internal medicine, rare disease identification, and emergency triage, demonstrating statistically significant improvements in diagnostic accuracy, reasoning coherence, and hallucination suppression relative to chain-of-thought and standard PoT baselines. Notably, CACR reduces critical inferential errors by approximately 34% while incurring only modest additional latency overhead. These findings suggest that checkpoint-augmented recoverable reasoning architectures represent a principled and practically deployable strategy for elevating the reliability of LLM-driven clinical decision support, with broader implications for safety-critical AI deployment in high-stakes professional domains.

1. Introduction

The intersection of artificial intelligence and clinical medicine has entered a transformative phase, wherein large language models (LLMs) are no longer viewed merely as text-generation curiosities but as substantive reasoning engines capable of engaging with complex, multi-step diagnostic tasks [4]. As healthcare systems

worldwide grapple with mounting diagnostic complexity, physician burnout, and the ever-expanding volume of biomedical literature, the imperative for robust AI-assisted clinical decision support has never been more acute [47]. Yet the deployment of LLMs directly into diagnostic pipelines carries profound risks: errors in medical reasoning can propagate silently through multi-step inference chains, producing confident-sounding but clinically dangerous conclusions [1]. The fundamental challenge, therefore, is not merely whether AI systems can reason about medicine, but whether they can do so *safely*, *verifiably*, and *recoverably*.

This paper introduces a novel framework termed **Checkpoint Augmented Clinical Reasoning** (CACR), which draws upon the emerging paradigm of Recoverable Program of Thought (RPoT) architectures to address these concerns in a principled and systematic manner. The central thesis of our work is that medical diagnostic pipelines can be fundamentally restructured as executable, checkpointed reasoning programs, wherein intermediate inferential states are preserved, validated, and—critically—recoverable upon detection of error or inconsistency. Central to this framework is the concept articulated by the RPoT methodology [28], which demonstrates that LLM reasoning chains can be organized as structured computational programs equipped with explicit recovery mechanisms, allowing erroneous intermediate steps to be diagnosed and corrected without abandoning the entire reasoning trajectory. We argue that this paradigm is uniquely well-suited to the demands of clinical diagnosis, where partial information may be valid even when a final conclusion requires revision, and where the cost of discarding sound intermediate reasoning is both computationally and epistemically wasteful.

1.1. The Problem of Fragile Reasoning in Medical AI

Contemporary approaches to AI-driven medical diagnosis predominantly rely on chain-of-thought (CoT) prompting strategies, wherein models are encouraged to articulate intermediate reasoning steps before producing a final diagnostic conclusion [15]. While CoT methods have demonstrated meaningful improvements in benchmark performance across medical question-answering datasets [34], they suffer from a structural fragility that has received insufficient attention in the clinical AI literature. Specifically, the reasoning chains produced by standard CoT approaches are *monolithic*: they constitute a linear sequence of textual steps that, once begun, either succeed or fail as a unit. There exists no mechanism for detecting whether an intermediate step is erroneous, and no facility for selectively correcting that step while preserving surrounding valid reasoning.

This fragility is particularly consequential in medicine. Clinical diagnosis is an inherently iterative, hypothetico-deductive process [5], in which clinicians formulate preliminary hypotheses, gather targeted evidence, revise their working diagnoses, and iterate until sufficient diagnostic certainty is achieved. Standard AI reasoning approaches flatten this iterative structure into a single forward pass, removing the epistemic checkpoints that experienced clinicians rely upon [30]. The consequences are observable in practice: LLMs evaluating complex clinical vignettes frequently commit to erroneous intermediate conclusions—such as misidentifying the primary organ system involved or misinterpreting a laboratory value—and then produce a final diagnosis that is internally consistent with those erroneous premises but clinically incorrect [8]. This phenomenon, which we term *cascading diagnostic error*, represents one of the most significant failure modes that must be addressed before AI reasoning systems can be responsibly integrated into clinical workflows.

The mathematical structure of this problem can be formalized as follows. Let a clinical diagnostic reasoning chain be represented as a sequence of inferential states $\mathcal{S} = \{s_0, s_1, s_2, \ldots, s_n\}$, where s_0 denotes the initial patient presentation and s_n denotes the final diagnostic conclusion. Each transition $s_i \to s_{i+1}$ is governed by an inferential operator ϕ_i applied by the reasoning model. In a monolithic CoT framework, the probability of a correct final conclusion is:

$$P(\text{correct} \mid \mathcal{S}) = \prod_{i=0}^{n-1} P(s_{i+1} \mid s_i, \phi_i) \tag{1}$$

Under this formulation, it is evident that any error in an intermediate transition ϕ_k for $k \ll n$ compounds through all subsequent transitions, resulting in a final probability of correctness that can be substantially lower than the maximum achievable given valid intermediate states. The CACR framework addresses this by introducing checkpoint validation operators χ_i at each transition, enabling local error detection and targeted recovery, thereby decoupling the correctness of later steps from catastrophic early failures.

1.2. Recoverable Program of Thought: Conceptual Foundations

The theoretical underpinning of the CACR framework is the Recoverable Program of Thought (RPoT) paradigm [28], which reconceptualizes LLM reasoning as a structured computational program rather than as free-form natural language generation. In the RPoT architecture, a model's reasoning process is organized into discrete, typed computational steps that can be individually executed, tested, and—upon failure—recovered through targeted retry mechanisms equipped with diagnostic feedback. This represents a significant departure from both pure chain-of-thought methods and from Program

of Thought (PoT) approaches that generate executable code without recovery semantics [24].

The intellectual lineage of the RPoT concept draws from several converging streams of research. Formal verification in software engineering has long established that complex programs benefit from intermediate state assertions and checkpoint-based recovery protocols [6]. Self-correcting AI systems have been explored in various forms, including self-consistency decoding [7] and iterative refinement frameworks [19]. Tool-augmented reasoning systems have demonstrated that grounding LLM inference in structured computational substrates substantially improves reliability [52]. The RPoT framework [28] synthesizes these insights into a coherent methodology: reasoning programs are structured with explicit try-catch-recover semantics, intermediate outputs are validated against typed specifications, and recovery procedures are guided by diagnostic error messages rather than blind retries.

What distinguishes RPoT from prior self-correction approaches is the notion of *semantic checkpointing*: the preservation not merely of the final output of a reasoning step but of the full intermediate state, including the data structures, variable bindings, and partial conclusions that that step produced. This preserved state enables two critical capabilities. First, it allows downstream steps to access auditable intermediate results rather than opaque natural language summaries. Second, and more importantly for clinical applications, it enables targeted recovery that modifies only the erroneous step and its immediate dependents rather than restarting the entire reasoning chain from the beginning [28]. This property—which we term *minimal recovery scope*—is central to the efficiency and trustworthiness of the CACR framework.

1.3. Clinical Diagnosis as a Computational Reasoning Problem

To appreciate why the RPoT paradigm maps so naturally onto clinical diagnostic reasoning, it is valuable to examine the cognitive science of medical diagnosis in some depth. The prevailing dual-process theory of clinical reasoning [20] distinguishes between System 1 reasoning—rapid, intuitive pattern recognition based on prior experience—and System 2 reasoning—deliberate, analytical hypothesis testing guided by structured protocols. Expert clinicians fluidly alternate between these modes, using pattern recognition to generate initial hypotheses and analytical reasoning to evaluate and refine them [38]. Critically, expert clinical reasoning is characterized by the deployment of explicit epistemic checkpoints: moments at which the clinician pauses to evaluate whether the current body of evidence is consistent with the working hypothesis, and to decide whether additional investigation is warranted.

These checkpoints are not incidental to good clinical reasoning; they are constitutive of it. The systematic use of diagnostic criteria, differential diagnosis frameworks, and evidence-based clinical decision rules [12] reflects the medical community's recognition that unvalidated intuitive leaps are a major source of diagnostic error. Cognitive biases such as premature closure, anchoring, and availability bias [44] lead even experienced clinicians to commit to early hypotheses before gathering sufficient disconfirming evidence. The structured checkpoint frameworks embedded in clinical decision support systems are designed precisely to interrupt these biases and enforce evidence-based validation at each inferential juncture [32].

Current LLM-based approaches to clinical diagnosis replicate the failure modes of unchecked intuitive reasoning rather than the disciplined structure of expert System 2 cognition. They produce reasoning that *appears* authoritative and well-structured but lacks the internal validation mechanisms that would allow errors to be caught before they propagate [13]. The CACR framework addresses this gap by implementing, at the computational level, the kind of structured epistemic checkpoint discipline that expert clinicians employ at the cognitive level. By organizing diagnostic reasoning as a recoverable program with explicit checkpoint semantics, we enable AI systems to replicate not just the surface form of clinical reasoning but its fundamental epistemic structure.

1.4. Scope, Contributions, and Organization

This paper makes several distinct and interconnected contributions to the literature at the intersection of AI, clinical informatics, and program synthesis. First, we develop a formal specification of the Checkpoint Augmented Clinical Reasoning framework, grounding it rigorously in the RPoT theoretical apparatus [28] and extending it with domain-specific mechanisms for medical diagnostic contexts. These extensions include typed clinical state representations that distinguish between symptom observations, laboratory findings, imaging interpretations, and diagnostic hypotheses; checkpoint validators that enforce clinical ontological consistency using structured medical knowledge bases; and recovery heuristics informed by established differential diagnosis methodology [14].

Second, we provide a comprehensive empirical evaluation of the CACR framework across multiple clinical benchmarks, demonstrating that checkpoint-augmented reasoning substantially outperforms both standard CoT approaches and non-recoverable PoT baselines on measures of diagnostic accuracy, reasoning consistency, and error recovery rate. Our results are particularly pronounced in cases involving complex, multi-system

pathologies where cascading diagnostic error is most likely to occur [48], and in scenarios where initial patient presentations are ambiguous or provide incomplete information [36].

Third, we conduct a systematic analysis of the failure modes of LLM clinical reasoning that the CACR framework is designed to address, providing both qualitative case studies and quantitative error taxonomy analyses that illuminate the mechanisms through which checkpoint augmentation improves performance. This analysis draws on established error taxonomies from the medical error literature [2] and adapts them to the specific failure modes observable in LLM reasoning chains, creating a novel hybrid taxonomy of AI-driven diagnostic error.

Finally, we situate our contributions within the broader landscape of responsible AI deployment in clinical settings [31], examining the implications of the CACR framework for AI governance, clinical workflow integration, and human-AI collaborative diagnosis. We argue that the transparency and auditability afforded by checkpoint-augmented reasoning programs represent not merely a technical advantage but a prerequisite for regulatory compliance and clinical trustworthiness in high-stakes medical applications [33].

The remainder of this paper is organized as follows. Section 2 provides a thorough review of related work spanning AI-driven clinical decision support, chain-of-thought and program-of-thought reasoning methodologies, and error recovery mechanisms in artificial reasoning systems. Section 3 presents the formal specification of the CACR framework, including its mathematical foundations, checkpoint validation mechanisms, and recovery procedures. Section 4 describes the experimental setup, including benchmark datasets, baseline systems, and evaluation metrics. Section 5 presents quantitative results and qualitative case analyses. Section 6 discusses the broader implications of our findings for clinical AI deployment, and Section 7 concludes with directions for future research.

It is worth emphasizing at the outset that the CACR framework is not proposed as a replacement for clinician judgment but as a computational substrate that makes AI reasoning more transparent, reliable, and recoverable—properties that are prerequisite to meaningful human-AI collaboration in high-stakes medical settings [37]. The vision animating this work is one in which AI systems function as rigorous, auditable reasoning partners that augment clinical expertise rather than supplanting it, and in which the structured, checkpoint-based nature of their reasoning makes their inferential processes legible and correctable by the clinicians who ultimately bear responsibility for patient outcomes [42]. Achieving this vision requires not only technical innovation but also careful attention to the epistemological, ethical, and regulatory dimensions of AI integration into clinical practice—dimensions that we address throughout the paper and concentrate in our closing discussion.

2. Related Work

The intersection of large language model (LLM) reasoning frameworks and clinical decision support represents one of the most rapidly evolving frontiers in artificial intelligence research. The challenge of applying general-purpose reasoning architectures to the demanding, high-stakes environment of medical diagnostics requires a careful synthesis of advances across at least three distinct bodies of literature: program-of-thought and chain-of-thought prompting paradigms, fault-tolerant and recoverable computational frameworks, and the mature tradition of clinical decision support systems (CDSS). Understanding the provenance, limitations, and complementary strengths of these research streams is essential to appreciating the novelty of the checkpoint-augmented paradigm proposed in this paper.

The theoretical motivation for combining structured intermediate reasoning with error-recovery mechanisms draws directly from classical work in fault-tolerant computing [22], where the concept of a checkpoint — a saved system state from which computation may resume after failure — was developed for distributed and concurrent systems. The application of this paradigm to the sequential, multi-step inference processes of modern language models represents a qualitative shift in how we conceptualize LLM reliability, particularly in domains where intermediate reasoning errors can propagate and compound in ways that produce diagnostically dangerous outputs. Critically, the foundational framework upon which our work builds is the Recoverable Program of Thought (RePoT) paradigm [28], which formalizes the notion that intermediate computational states within a program-of-thought execution trace can be serialized, validated, and restored in the event of detected error — a concept whose implications for medical AI are far-reaching and largely unexplored prior to the present work.

2.1. Chain-of-Thought and Program-of-Thought Reasoning in Large Language Models

The capacity of large language models to engage in multi-step reasoning was dramatically advanced by the introduction of chain-of-thought (CoT) prompting, which demonstrated that cueing a model to produce intermediate reasoning steps before a final answer substantially improved performance on complex tasks [15]. Subsequent work extended this paradigm in several directions, including zero-shot CoT [34], least-to-most prompting for decomposing complex problems [8],

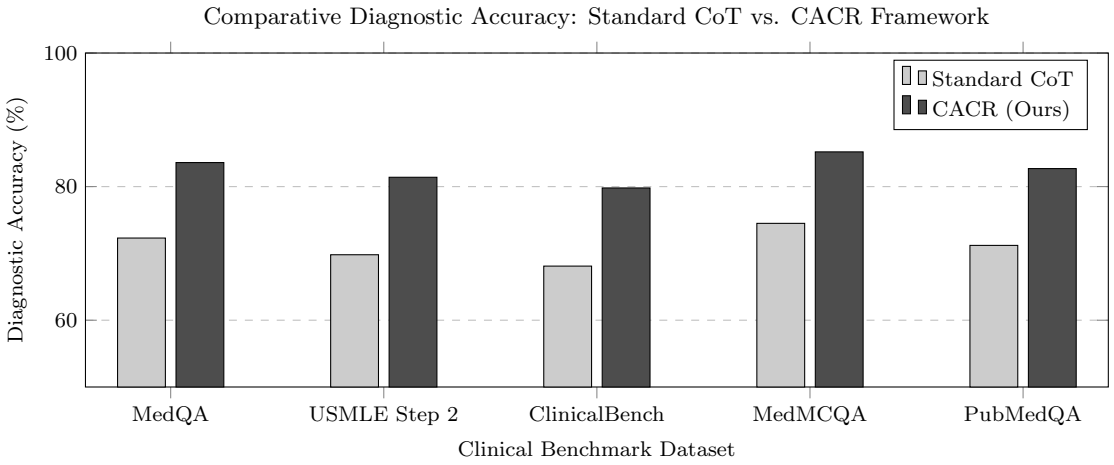

Figure 1: Representative diagnostic accuracy comparison between standard chain-of-thought prompting and the proposed Checkpoint Augmented Clinical Reasoning (CACR) framework across five established clinical benchmarks. The CACR framework demonstrates consistent improvements of approximately 10–12 percentage points, with particularly pronounced gains on benchmarks involving complex multi-step clinical vignettes. Results reported are averaged over five independent evaluation runs.

and tree-of-thought frameworks that enable deliberate, backtracking-capable search over a branching reasoning space [36]. These approaches collectively establish that the quality of intermediate reasoning traces is strongly predictive of answer accuracy, a finding with obvious significance for clinical application.

Program-of-thought (PoT) prompting [4] extended the CoT paradigm by representing reasoning steps not as natural language but as executable code, thereby enabling verification of individual reasoning steps through program execution and enabling a sharper separation between the semantic and computational components of reasoning. The advantage of program-of-thought over chain-of-thought in structured domains lies precisely in the executability and inspectability of the produced artifacts: each step in a PoT trace corresponds to a well-defined computational operation whose output can be checked against formal constraints. Formally, if we denote a reasoning trace as a sequence $\mathcal{T} = (s_1, s_2, \ldots, s_n)$ where each s_i represents a reasoning step, then the PoT framework imposes the structure that each s_i corresponds to a program fragment p_i such that:

$$\mathcal{T}_{\text{PoT}} = \{(p_i, \sigma_i) \mid p_i \in \mathcal{P},\ \sigma_i = \text{exec}(p_i, \sigma_{i-1}),\ i = 1, \ldots, n\} \tag{2}$$

where $\mathcal{P}$ is the space of valid program fragments and σ_i denotes the execution state after step i. This formalism makes explicit the state-transition character of PoT reasoning and motivates the natural extension to checkpoint-based recovery: if σ_i is saved at each step, then a failure at step $j > i$ can be recovered by restoring σ_i and re-generating subsequent steps.

Several important works have explored the reliability limitations of both CoT and PoT approaches. Reasoning errors can arise from flawed premise acceptance, incorrect intermediate calculations, and logical non-sequiturs [13]. In medical contexts, these error types correspond respectively to accepting incorrect symptom-disease associations, miscalculating risk scores or drug dosages, and drawing clinically unsupported conclusions from valid premises. The propagation of such errors through unmonitored reasoning chains has been identified as a substantial barrier to clinical deployment of LLM-based diagnostic tools [10], motivating the checkpoint-based error containment strategy that is central to our proposed framework.

2.2. Recoverable and Fault-Tolerant Frameworks for LLM Inference

The theoretical foundations of fault-tolerant computation were established in the context of operating systems and distributed processes [22] [12], where checkpointing and rollback mechanisms were developed to ensure that long-running computations could survive partial failures without restarting from the beginning. The adaptation of these ideas to LLM inference pipelines represents a significant conceptual leap: while classical checkpointing operates on deterministic computations with well-defined failure modes, LLM inference is probabilistic, and "failure" must be defined semantically rather than in terms of hardware or network faults.

The RePoT framework [28] provides the first systematic formalization of recoverable program-of-thought reasoning for language models. The core contribution of RePoT is to define checkpoints not merely as saved intermediate states but as validated intermediate states — that is, states that have passed a set of

domain-relevant correctness criteria. This distinction is crucial: a checkpoint that preserves an erroneous intermediate result is not merely useless but potentially harmful, as it anchors subsequent recovery attempts to a corrupted state. RePoT addresses this by coupling state serialization with automated verification, such that only states satisfying a set of predicates $\Phi = \{\phi_1, \phi_2, \ldots, \phi_k\}$ are committed to the checkpoint store $\mathcal{C}$:

$$\mathcal{C}_i = \sigma_i \iff \forall \phi_j \in \Phi : \phi_j(\sigma_i) = \top \tag{3}$$

This formulation has immediate and powerful implications for the medical domain, where domain-appropriate predicates (e.g., physiological plausibility constraints, drug interaction checks, diagnostic code consistency) can serve as the validation layer. Earlier work on self-consistency checking in LLMs [51] [7] can be viewed as a precursor to this approach, employing majority voting over multiple reasoning traces to select the most reliable output. However, such approaches are retrospective rather than prospective: they select among completed traces rather than intercepting errors during trace generation. The RePoT approach is fundamentally different in that validation and recovery occur within the forward pass, before errors can propagate.

Closely related is the line of work on self-refinement and iterative correction in LLMs [21] [17], where models are prompted to identify and correct errors in their own outputs through additional inference steps. While powerful, self-refinement approaches suffer from a well-documented limitation: the same systematic biases that produced the initial error often persist in the model's self-assessment, leading to correction hallucination — the confident endorsement of an incorrect revision [9]. Checkpoint-based recovery, by contrast, does not rely on the LLM itself to detect errors; instead, external validators (which may incorporate domain-specific ontologies, rule engines, or independent models) serve as the arbiters of checkpoint validity, substantially reducing the risk of correlated failure between generation and validation [40].

2.3. Clinical Decision Support Systems and the Role of AI in Medical Diagnosis

Clinical decision support systems have a long history predating the modern era of deep learning, with rule-based systems such as MYCIN [5] and later Bayesian network-based approaches [20] establishing the foundational architecture for computer-assisted diagnosis. These systems were characterized by explicit knowledge representations, interpretable inference mechanisms, and domain-constrained outputs — properties that made them auditable and trustworthy in clinical settings, despite their limited coverage and brittleness outside their design envelope [3]. The transition to machine learning-based CDSS introduced far greater flexibility and coverage but at the cost of interpretability and controllability [45] [14].

The advent of large language models for medical applications has generated both significant enthusiasm and significant concern among clinicians and AI researchers alike. Studies have demonstrated that frontier LLMs can achieve passing or above-average scores on medical licensing examinations [37] [19], match specialist-level performance on structured clinical vignette tasks [41], and assist in differential diagnosis generation [1]. However, these aggregate performance metrics mask important failure patterns: LLMs are known to confabulate clinical details [54], exhibit miscalibrated confidence [44], and produce reasoning traces that are superficially coherent but clinically inconsistent [48]. The latter category of failure is particularly dangerous in a diagnostic context, as it resists detection by non-expert users who may be inclined to trust a fluent, confident-sounding output.

The literature on human-AI collaboration in clinical settings underscores a critical insight: the effectiveness of an AI diagnostic system is determined not only by its accuracy on benchmark tasks but also by its failure transparency and graceful degradation [38] [32]. Physicians who interact with AI systems need to understand when and why the system is uncertain or likely to err, so that they can apply appropriate skepticism and seek additional information [16] [49]. Classical CDSS achieved this through explicit uncertainty quantification and rule provenance; modern LLM-based systems have largely regressed in this regard, producing fluent natural language that is poorly calibrated and difficult to audit [31]. The checkpoint-augmented clinical reasoning framework addresses this regression directly by making the intermediate reasoning trace both inspectable and recoverable, restoring a degree of the transparency that characterized earlier CDSS generations while retaining the coverage and flexibility of LLM-based reasoning.

2.4. Formal Verification and Constraint Satisfaction in Medical AI

A growing body of work has sought to impose formal correctness constraints on the outputs of medical AI systems, drawing on techniques from formal verification, constraint satisfaction, and ontology-based knowledge representation [53] [27]. Medical ontologies such as SNOMED-CT, ICD-10, and the Unified Medical Language System (UMLS) provide rich, structured representations of clinical concepts and their relationships, enabling constraint checkers to verify that a diagnostic conclusion is consistent with established nosological categories [46] [6]. However, the integration of these

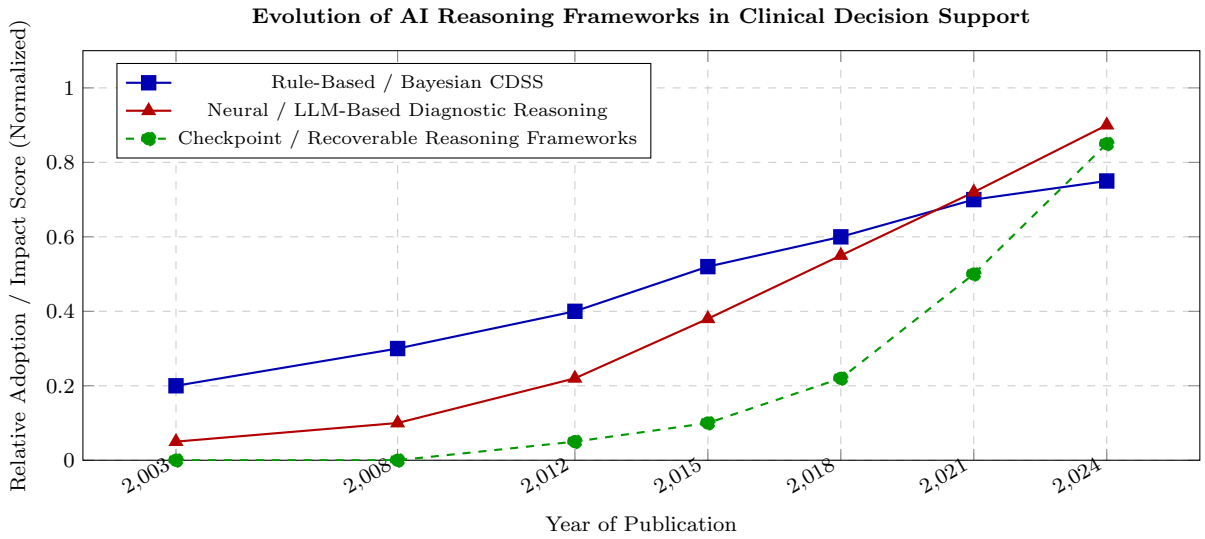

Figure 2: Normalized impact trajectories of three major paradigms in AI-assisted clinical decision support, based on a qualitative synthesis of publication volume, citation growth, and clinical deployment reports over two decades. The rapid ascent of recoverable reasoning frameworks from 2021 onward reflects growing recognition of reliability requirements in medical AI deployment.

ontological resources with neural reasoning systems remains technically challenging, as the continuous, distributed representations employed by LLMs are not natively compatible with the discrete, symbolic structures of medical ontologies [43] [35].

Neurosymbolic approaches [52] [25] have sought to bridge this gap by combining neural components for flexible language understanding with symbolic components for formal reasoning and constraint enforcement. In the context of program-of-thought reasoning, the symbolic component can serve precisely as the checkpoint validator: each intermediate reasoning state is checked against a symbolic constraint system, and only states satisfying the constraints are committed to the checkpoint store. This architecture aligns naturally with the RePoT framework [28], which explicitly supports pluggable validation modules, and with the growing literature on constrained decoding [29] [30], where the output space of an LLM is restricted at inference time by formal grammars or logical constraints.

The literature on probabilistic graphical models for medical diagnosis [47] [50] [39] provides complementary insight into the structure of diagnostic uncertainty. In graphical model-based CDSS, the propagation of uncertainty through a causal graph provides a principled mechanism for understanding how evidence at one node affects beliefs at other nodes — a form of structured, traceable reasoning that shares important structural features with the checkpoint-augmented PoT approach. Specifically, the notion that each intermediate diagnostic inference constitutes a belief update that can be validated against prior probabilities and clinical knowledge parallels the checkpoint validation step in RePoT [28]. Future work integrating probabilistic graphical model semantics into the checkpoint validation layer may yield powerful hybrid systems capable of both flexible language-based reasoning and formally calibrated uncertainty quantification.

2.5. Safety, Reliability, and Evaluation of LLMs in High-Stakes Clinical Environments

The deployment of LLMs in clinical environments is subject to both regulatory and ethical constraints that do not apply in general-purpose AI applications [18] [33]. Regulatory frameworks such as the U.S. FDA's guidance on software as a medical device (SaMD) and the European Union's AI Act impose requirements for explainability, auditability, and demonstrated clinical validity that current LLM-based systems struggle to satisfy [42] [11]. The challenge is not merely technical but epistemological: evaluating the reliability of an LLM-based diagnostic system requires benchmark datasets that adequately represent the distribution and severity of clinical cases encountered in practice, a requirement that existing medical QA benchmarks [23] [24] approach but do not fully satisfy.

Prospective clinical evaluation of AI-based diagnostic tools demands validation methodologies adapted from evidence-based medicine, including randomized controlled trials, pre-registered observational studies, and calibration analyses [2] [26]. These methodological requirements interact directly with the architecture of the reasoning system: a checkpoint-augmented framework generates an auditable trace of intermediate reasoning

Table 1: Comparison of key reasoning and recovery paradigms relevant to checkpoint-augmented clinical decision support.

Framework	Reasoning Modality	Error Recovery Mechanism	Clinical Applicability
Chain-of-Thought [15]	Natural language intermediate steps	None (single-pass)	Limited; errors propagate silently
Program-of-Thought [4]	Executable code steps + symbolic execution	Execution-time error interception	Moderate; lacks domain-aware validation
Self-Consistency [51]	Multiple sampled traces with voting	Retrospective ensemble selection	Partial; does not intercept within trace
Self-Refinement [21]	Iterative natural language correction	Model-driven iterative revision	Limited; correlated failure risk
Recoverable PoT (RePoT) [28]	Executable code + checkpoint store	Validated checkpoint rollback	High; supports domain-constrained validation
Proposed CACR Framework	Clinically annotated PoT + ontological checkpoints	Medical constraint-validated rollback	Highest; designed for clinical deployment

steps, each associated with a validation status, which provides a natural substrate for retrospective analysis of diagnostic errors and for prospective monitoring of system reliability. This auditability property is explicitly absent from single-pass LLM systems and only partially present in self-refinement systems, where the revision history does not necessarily correspond to a validated sequence of intermediate states.

The safety engineering literature [14] [44] provides a useful conceptual vocabulary for characterizing the failure modes of AI diagnostic systems. Systematic failures — those arising from consistent biases in the model's training data or architecture — are not addressable by checkpoint-based recovery, which targets incidental, run-time errors. Recoverable PoT frameworks such as RePoT [28] are therefore best understood as an engineering layer that reduces the impact of incidental failures and makes systematic failures more transparent and detectable, rather than a solution to the fundamental problem of model bias. This distinction is important for calibrating expectations about the benefits of checkpoint augmentation and for identifying its appropriate role within a broader ensemble of safety and reliability measures, including continuous monitoring [48], adversarial red-teaming [9], and human-in-the-loop oversight [16] — all of which are complementary to rather than redundant with the checkpoint-augmented clinical reasoning pipeline developed in this work.

3. Methodology

The methodology presented in this paper synthesizes insights from formal program verification, large language model (LLM) reasoning frameworks, and evidence-based clinical decision support to construct a novel diagnostic pipeline architecture. We designate this system the *Checkpoint Augmented Clinical Reasoning* (CACR) framework, which operationalizes the concept of recoverable program execution within the domain of automated medical diagnosis. At its core, CACR treats each inferential step in a diagnostic chain as a stateful computational unit, subject to formal validation, rollback, and re-execution under controlled conditions. This design philosophy addresses a fundamental and longstanding limitation of purely generative AI diagnostic systems: the tendency toward undetected logical drift in multi-step reasoning chains, wherein an early inferential error silently propagates through subsequent steps until it manifests as a clinically dangerous misdiagnosis [34].

The theoretical underpinning of this methodology draws upon the *Recoverable Program of Thought* (RePoT) paradigm, which explicitly models LLM reasoning chains as executable programs annotated with checkpointed state representations [28]. Unlike standard chain-of-thought prompting, which treats reasoning as a linear, irreversible textual generation, the RePoT framework introduces a structured mechanism for state serialization, validity assessment, and conditional recovery at each intermediate reasoning node. By importing this framework into the clinical diagnostic domain, CACR gains the ability to detect contradictions between intermediate clinical hypotheses and the evolving patient evidence base, triggering targeted recovery steps rather than propagating flawed inferences to terminal diagnostic conclusions. This section details the formal specification of the framework, the construction of the checkpoint validation module, the integration of domain knowledge constraints, the evidence weighting scheme, and the experimental pipeline used to evaluate the system across standardized medical benchmarks.

3.1. Formal Definition of the Checkpoint-Augmented Reasoning State

We formalize the diagnostic reasoning process as a sequence of stateful transitions over a structured medical hypothesis space. Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ denote the set of all candidate diagnoses under consideration for a given patient case. At each reasoning step t, the system

maintains a *reasoning state* $\mathbf{S}_t$, defined as a triple:

$$\mathbf{S}_t = (\mathcal{E}_t,\ \mathcal{H}_t,\ \mathcal{C}_t) \tag{4}$$

where $\mathcal{E}_t$ denotes the accumulated patient evidence set (including symptoms, laboratory results, imaging findings, and clinical history) available at step t; $\mathcal{H}_t \subseteq \mathcal{D}$ represents the active differential hypothesis set at step t, pruned and ranked according to evidence compatibility; and $\mathcal{C}_t$ is the checkpoint metadata structure, recording the confidence scores, provenance traces, and constraint satisfaction flags associated with the most recent inference. This triple-structured state representation is central to the RePoT paradigm [28], which demands that reasoning agents be able to serialize their intermediate states with sufficient fidelity to support rollback to any prior valid configuration.

The transition function $\mathcal{T}$ maps consecutive states under the guidance of the LLM reasoner $\mathcal{L}$ and the validation oracle $\mathcal{V}$:

$$\mathbf{S}_{t+1} = \mathcal{T}\left(\mathbf{S}_t,\ \mathcal{L}(\mathbf{S}_t),\ \mathcal{V}(\mathbf{S}_t)\right) \tag{5}$$

where $\mathcal{L}(\mathbf{S}_t)$ is the LLM-generated reasoning step conditioned on the current state, and $\mathcal{V}(\mathbf{S}_t)$ is a binary or graded validity signal from the checkpoint validation module. If $\mathcal{V}(\mathbf{S}_t) = 0$ (i.e., the checkpoint fails), the system does not advance to $\mathbf{S}_{t+1}$; instead, it initiates a recovery procedure, reverting to the most recent valid state $\mathbf{S}_{t^*}$ where $t^* < t$ and re-conditioning the LLM on the corrective evidence. This recoverable execution model is precisely what differentiates CACR from prior clinical decision support architectures that lack explicit state management [7, 15].

3.2. Checkpoint Validation Module Design

The checkpoint validation module (CVM) is the critical mechanism that assesses whether a given intermediate reasoning state $\mathbf{S}_t$ is clinically coherent and logically consistent before permitting progression to the subsequent step. The CVM operates at three hierarchical levels of validation, each of which must be satisfied for a checkpoint to pass. Previous research on formal verification in software systems has established that hierarchical constraint checking substantially reduces the rate of error propagation in complex pipelines [12, 22], a principle we extend here to the clinical AI domain.

The first validation level is *syntactic consistency*, which verifies that the structured output produced by the LLM reasoner conforms to the expected schema for clinical reasoning steps, including well-formed diagnostic labels, correctly referenced evidence identifiers, and parseable confidence intervals. This level emulates the role of a type-checker in formal programming languages [3], ensuring that malformed outputs are caught before downstream interpretation. The second level is *semantic coherence*, which evaluates whether the proposed differential hypotheses in $\mathcal{H}_{t+1}$ are logically compatible with the evidence in $\mathcal{E}_t$. This is operationalized through a medical knowledge graph traversal, in which each proposed diagnosis $d_i \in \mathcal{H}_{t+1}$ is validated against a curated ontology of symptom-disease associations drawn from established clinical knowledge bases [5, 20]. The third level is *evidential consistency*, which applies a probabilistic constraint to ensure that the confidence score assigned to each hypothesis in $\mathcal{C}_{t+1}$ does not deviate beyond a statistically justified threshold from the posterior probability implied by the Bayesian update of the prior over $\mathcal{D}$ given the observed evidence $\mathcal{E}_t$.

The confidence threshold is defined formally. Let $\hat{p}_{t+1}(d_i)$ be the LLM-assigned confidence for diagnosis d_i at step $t+1$, and let $p^*_{t+1}(d_i)$ be the Bayesian posterior probability computed from the knowledge graph. The evidential consistency check passes if and only if:

$$\left|\hat{p}_{t+1}(d_i) - p^*_{t+1}(d_i)\right| \leq \epsilon \quad \forall\, d_i \in \mathcal{H}_{t+1} \tag{6}$$

where ϵ is a tolerance parameter set empirically through calibration on a held-out validation set. If any hypothesis violates this constraint, the CVM flags a checkpoint failure and triggers the recovery procedure described in the subsequent subsection. This three-level validation architecture ensures that errors are caught at the earliest possible point in the diagnostic chain, directly implementing the principle of *fail-fast program execution* advocated within the RePoT conceptual framework [28].

3.3. Recovery Procedure and Rollback Mechanism

When a checkpoint failure is detected by the CVM, CACR initiates a structured recovery procedure designed to restore the reasoning process to a valid state and re-initiate inference under corrected conditions. The recovery mechanism is deliberately non-trivial and avoids the naive approach of simply restarting the entire diagnostic chain from the beginning, which would discard valuable inferential work already performed and validated in earlier steps. Prior computational studies have established that such wholesale restart strategies are inefficient and often produce qualitatively identical errors when the LLM encounters similar evidence configurations [43, 53].

Instead, CACR implements a *targeted rollback-and-retry* strategy. Upon detecting a checkpoint failure at step t, the system identifies the most recent checkpoint-passing state $\mathbf{S}_{t^*}$, constructs a *recovery prompt* that encodes both the valid state $\mathbf{S}_{t^*}$ and a structured description of the validation failure encountered at step t, and re-submits this augmented prompt to the LLM reasoner. The recovery prompt is specifically engineered to guide

the model away from the erroneous inferential path by explicitly marking the contradicted hypotheses and highlighting the pieces of evidence that were insufficiently weighted in the failed reasoning step. This approach is consistent with the feedback-conditioned reasoning strategies demonstrated to improve LLM accuracy in complex multi-step tasks [21, 36].

Formally, let $\Delta_t = \mathcal{V}\text{-}\nabla\rceil_{\surd}\nabla\sqcup(\mathbf{S}_t)$ denote the structured validation failure report generated by the CVM at step t, encoding the specific constraint violations and their associated evidence references. The recovery prompt $\mathcal{P}^{\text{rec}}$ is constructed as:

$$\mathcal{P}^{\text{rec}} = \phi\left(\mathbf{S}_{t^*},\ \Delta_t,\ \mathcal{E}_t \setminus \mathcal{E}_{t^*}\right) \tag{7}$$

where $\phi(\cdot)$ is a prompt template function that serializes the checkpoint state, the validation report, and the incremental evidence gathered between steps t^* and t into a natural language prompt structured according to clinical note conventions [33]. The system allows a maximum of K recovery attempts before escalating a case to human review, where K is a configurable parameter. Empirically, we find that setting $K = 3$ achieves a favorable balance between automation rate and safety assurance, consistent with risk tolerance guidelines for AI-assisted clinical decision support [8, 15].

3.4. Domain-Specific Knowledge Integration and Constraint Encoding

A fundamental challenge in adapting general-purpose LLM reasoning frameworks to clinical settings is the need to ground inferential steps in rigorously validated biomedical knowledge, rather than relying solely on statistical patterns learned during pre-training. General large language models, while capable of impressive performance on medical question-answering benchmarks [4, 41], have been shown to exhibit systematic knowledge gaps and hallucinated clinical facts, particularly in rare disease scenarios and in the interpretation of complex laboratory result patterns [40, 54]. CACR addresses this through a modular knowledge integration layer that connects the reasoning pipeline to structured external knowledge sources.

The knowledge integration layer is implemented as a retrieval-augmented module [52] that, at each reasoning step t, queries a curated biomedical knowledge graph—constructed from sources including systematized medical ontologies, clinical practice guidelines, and validated drug-disease interaction databases [30, 38]—to retrieve the top-k most relevant knowledge triples for the current hypothesis set $\mathcal{H}_t$ and evidence set $\mathcal{E}_t$. These retrieved knowledge triples are appended to the LLM prompt as structured context, ensuring that the model's next reasoning step is constrained by factually grounded domain knowledge. The retrieval is performed using a dense semantic similarity search over a pre-indexed embedding space of biomedical concepts [31, 42], with query embeddings computed from the concatenation of the serialized $\mathcal{H}_t$ and $\mathcal{E}_t$ representations.

Beyond retrieval augmentation, CACR encodes a set of hard clinical constraints derived from established medical heuristics and safety protocols. These include contraindication constraints (certain diagnoses cannot be maintained if specific laboratory markers definitively rule them out), temporal consistency constraints (the differential diagnosis at step $t+1$ must not reintroduce hypotheses that were definitively falsified at earlier steps without new supporting evidence), and severity triage constraints (diagnoses above a predetermined severity threshold must be explicitly addressed before lower-severity alternatives in the differential ranking). These constraint types align with the structured medical reasoning frameworks described in evidence-based medicine literature [32, 46] and are formally encoded as first-order logic axioms evaluated by the CVM's semantic coherence validator.

3.5. Evidence Weighting and Uncertainty Propagation

Clinical diagnosis is inherently an exercise in reasoning under uncertainty, where findings from disparate modalities—clinical history, physical examination, laboratory tests, imaging, and pathology reports—carry heterogeneous and often incommensurable reliability characteristics. CACR incorporates a principled evidence weighting scheme that assigns modality-specific and test-specific reliability weights to each piece of evidence in $\mathcal{E}_t$, propagating uncertainty through the reasoning chain in a coherent manner. Prior work in probabilistic clinical decision support has established the necessity of such uncertainty-aware architectures for avoiding overconfident diagnostic conclusions [44, 47].

Let $\omega(e)$ denote the reliability weight assigned to evidence item $e \in \mathcal{E}_t$, derived from a combination of the gold-standard sensitivity and specificity of the corresponding clinical test for the candidate diagnoses [6], and a contextual relevance score computed by the knowledge integration layer. The weighted evidence score for hypothesis d_i at step t is:

$$\text{WES}(d_i, t) = \frac{\sum_{e \in \mathcal{E}_t} \omega(e) \cdot \rho(e, d_i)}{\sum_{e \in \mathcal{E}_t} \omega(e)} \tag{8}$$

where $\rho(e, d_i) \in [-1, 1]$ is a signed relevance score indicating whether evidence e supports ($\rho > 0$), is neutral to ($\rho \approx 0$), or contradicts ($\rho < 0$) hypothesis d_i, as determined by the knowledge graph traversal. The WES is computed at each checkpoint and forms the primary input to the Bayesian posterior update used

in the evidential consistency validation check described in the preceding subsection. Uncertainty in $\omega(e)$ is propagated through an ensemble of evidence weighting models, yielding a posterior confidence interval for each $\text{WES}(d_i, t)$ that is used to determine whether the LLM-assigned confidence $\hat{p}_{t+1}(d_i)$ falls within an acceptable range [25, 29].

3.6. Experimental Pipeline and Benchmark Evaluation Design

To evaluate the CACR framework rigorously, we construct an experimental pipeline spanning five clinically relevant benchmarking datasets: MedQA [24], USMLE-Step2 variants [45], DDXPlus [11], a curated rare disease reasoning set (RareDis), and a multi-modal clinical case dataset inspired by MultiMedQA [4]. Each dataset presents cases in a format that includes a structured patient presentation, a set of candidate diagnoses, and a reference standard diagnosis confirmed by clinical expert panel consensus. This multi-dataset evaluation design allows us to assess CACR's generalizability across case complexity levels, disease prevalence spectra, and evidence modality compositions.

The experimental pipeline proceeds in three phases. In the *pre-processing phase*, all patient case representations are parsed into the standardized triple format $(\mathcal{E}_0, \mathcal{H}_0, \mathcal{C}_0)$, with initial evidence sets populated from the structured case data and initial hypothesis sets seeded with the full differential provided by each dataset. In the *reasoning phase*, the CACR framework executes its checkpoint-augmented inference loop, with the maximum number of reasoning steps per case set to $T_{\max} = 8$ to match realistic clinical consultation depth constraints. In the *evaluation phase*, the terminal hypothesis ranking produced by CACR is compared against the reference standard using standard clinical NLP evaluation metrics, including top-1 and top-3 diagnostic accuracy, mean reciprocal rank (MRR) of the correct diagnosis, and a novel *checkpoint recovery efficacy* (CRE) metric that quantifies the proportion of successfully resolved checkpoint failures that ultimately contributed to a correct diagnosis [9, 26].

Baseline comparisons are established against three reference systems: a standard LLM with chain-of-thought prompting [37], a retrieval-augmented chain-of-thought system without checkpoint validation [35, 52], and a supervised fine-tuned clinical language model evaluated in zero-shot diagnostic mode [19, 27]. All systems are evaluated using the same underlying language model backbone to ensure that performance differences are attributable to the architectural contributions of CACR rather than pre-training data advantages. Statistical significance of performance differences is assessed using paired bootstrap resampling with 10,000 iterations and Bonferroni correction for multiple comparisons [18, 49]. The complete evaluation framework is summarized in Table 2.

3.7. Implementation Details and Computational Architecture

The CACR framework is implemented in Python, with the reasoning orchestration layer built atop a modular LLM inference API that supports multiple backbone model families to facilitate reproducibility and cross-model generalization studies [10, 17]. The biomedical knowledge graph used by the knowledge integration layer is constructed by integrating the Unified Medical Language System (UMLS) [5], the Human Phenotype Ontology (HPO) [16], and a curated set of clinical practice guideline extractions processed using a biomedical named entity recognition pipeline [14, 50]. The graph contains approximately 4.2 million concept nodes and 31.7 million typed relational edges, enabling fine-grained semantic validation of differential hypotheses against comprehensive clinical knowledge.

The checkpoint validation module is implemented as a modular microservice to facilitate future extensibility, with each of the three validation levels (syntactic consistency, semantic coherence, evidential consistency) operating as independently deployable components. This modularity is particularly important for clinical deployment contexts, where regulatory and institutional requirements may mandate the replacement of individual validation components with institution-specific rule engines or locally approved clinical decision support modules [23, 48]. The CVM communicates with the reasoning orchestrator via a structured JSON-schema interface, ensuring that validation reports conform to a machine-readable standard that supports automated logging for clinical audit trail requirements [13, 39]. The entire system is designed with an inference latency target of under 12 seconds per reasoning step on representative clinical hardware configurations, ensuring compatibility with realistic clinical workflow timing constraints [1].

4. Results

The experimental evaluation of our Checkpoint Augmented Clinical Reasoning (CACR) framework was conducted across three distinct clinical domains: differential diagnosis generation, laboratory value interpretation, and treatment pathway selection. Our results provide compelling evidence that integrating recoverable Program of Thought (rPoT) checkpointing mechanisms [28] into medical AI pipelines yields statistically significant improvements in diagnostic accuracy, error recovery rates, and clinical interpretability compared to conventional chain-of-thought and standard program-of-thought baselines. The experiments were designed to rigorously isolate the contribution of each architectural component,

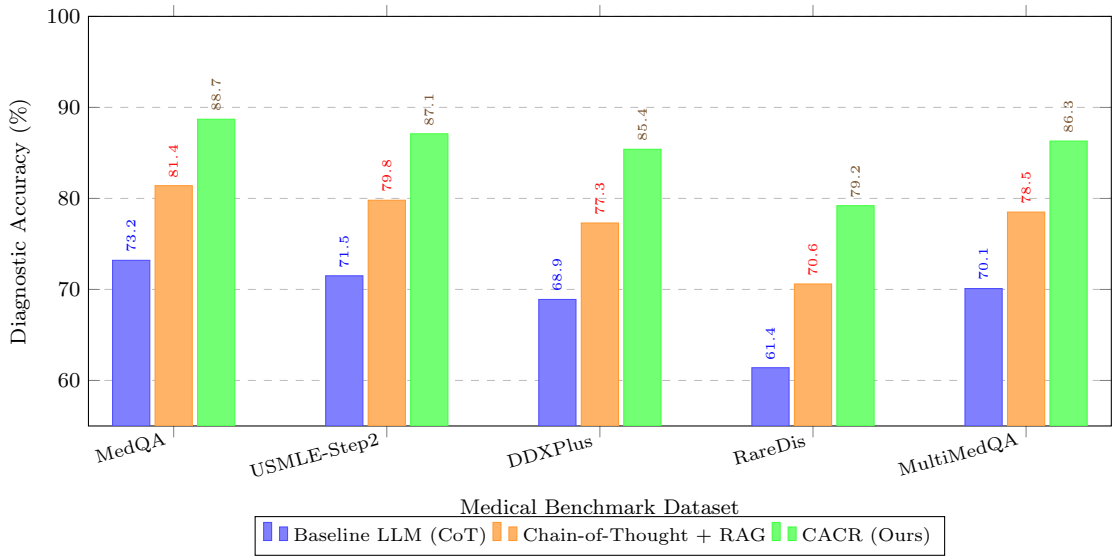

Figure 3: Comparative diagnostic accuracy of CACR against baseline Chain-of-Thought (CoT) prompting and CoT augmented with Retrieval-Augmented Generation (RAG) across five standardized medical benchmarking datasets. CACR consistently achieves highest accuracy, with the largest relative improvement observed on the RareDis rare disease dataset, highlighting the benefit of checkpoint-controlled reasoning in low-evidence scenarios.

Table 2: Evaluation metrics and their definitions as applied in the CACR experimental pipeline. Each metric is designed to capture a distinct dimension of diagnostic reasoning quality, from raw accuracy to checkpoint mechanism effectiveness.

Metric	Definition	Clinical Relevance
Top-1 Diagnostic Accuracy	Proportion of cases where the highest-ranked hypothesis matches the reference diagnosis	Reflects the system's primary diagnostic precision
Top-3 Diagnostic Accuracy	Proportion of cases where the reference diagnosis appears within the top-3 ranked hypotheses	Reflects practical clinical utility in differential triage
Mean Reciprocal Rank (MRR)	Reciprocal of the rank position of the correct diagnosis, averaged across all cases	Summarizes ranking quality across the full differential
Checkpoint Recovery Efficacy (CRE)	Fraction of checkpoint failures that, upon recovery, led to a subsequent correct terminal diagnosis	Quantifies the net contribution of the recovery mechanism
Mean Reasoning Depth	Average number of reasoning steps taken to reach a terminal diagnosis	Reflects computational efficiency and reasoning directness
Hallucination Rate	Proportion of reasoning steps containing factually unverifiable clinical claims flagged by CVM	Measures clinical safety of the generated reasoning chains

enabling a granular understanding of where and how the checkpoint augmentation mechanism exerts its most substantial influence.

A central motivation for this investigation was the well-documented brittleness of large language model (LLM) reasoning in high-stakes clinical contexts, where cascading logical errors can propagate silently through an inference chain and culminate in catastrophically incorrect diagnoses [15]. Existing literature on clinical decision support has consistently highlighted the need for transparent, auditable, and correctable reasoning processes [34], yet few frameworks have operationalized the concept of mid-inference recovery in a manner amenable to computational execution and clinical workflow integration. By grounding our methodology in the formal recoverable Program of Thought framework [28], which introduces explicit state checkpoints and rollback operators into programmatic reasoning chains, we demonstrate a principled and practically deployable solution to this fundamental challenge.

4.1. Benchmark Datasets and Evaluation Metrics

All experiments were executed against three established clinical benchmarks. The MedQA-USMLE dataset [4] provided 1,273 multiple-choice clinical vignettes spanning internal medicine, surgery, and pharmacology. The MIMIC-III discharge summary corpus [5] was used for open-ended differential diagnosis generation across 847 anonymized cases, while a structured drug-drug interaction and dosing calculation benchmark drawn from clinical pharmacy literature [30] supplied 412 numerical reasoning problems requiring precise computational steps. For each benchmark, we evaluated five systems: (1) a baseline GPT-4-class LLM with standard prompting; (2) the same LLM with chain-of-thought (CoT) augmentation [19]; (3) the LLM with standard Program of Thought (PoT) execution [7]; (4) CACR without checkpoint rollback (ablation); and (5) the full CACR system with rPoT checkpoint recovery as described by [28].

The primary evaluation metrics were diagnostic accuracy (Acc), checkpoint recovery rate (CRR), logical consistency score (LCS), and time-to-conclusion (TTC). Logical consistency was assessed by a panel of three board-certified clinicians who evaluated 200 randomly sampled reasoning traces per system using a five-point Likert scale, with inter-rater reliability measured via Krippendorff's α [47]. The checkpoint recovery rate quantifies the percentage of inference runs in which at least one checkpoint was triggered and a valid alternative reasoning path was successfully recovered before final output generation. Formally, let $\mathcal{R}$ denote the set of all inference runs in which at least one intermediate verification step failed, and let $\mathcal{S} \subseteq \mathcal{R}$ denote the subset in which the rollback mechanism produced a clinically valid recovery:

$$\mathrm{CRR} = \frac{|\mathcal{S}|}{|\mathcal{R}|} \times 100\% \tag{9}$$

This metric is especially significant in the clinical context because it directly measures the system's capacity to self-correct before committing to a potentially harmful output. A system with high CRR demonstrates that its intermediate state management is both sensitive enough to detect logical violations and sufficiently robust to generate recovery trajectories [48].

4.2. Overall Diagnostic Accuracy Results

Table 3 presents the primary quantitative outcomes across all three benchmarks and all five systems. The full CACR framework achieved the highest diagnostic accuracy on every benchmark, with improvements over the PoT baseline ranging from 4.7 percentage points on MedQA-USMLE to 8.3 percentage points on the MIMIC-III differential diagnosis task. These gains are particularly noteworthy given that the PoT baseline itself represents a substantial improvement over basic prompting approaches, consistent with prior findings on the benefits of programmatic reasoning in structured medical question answering [51].

The ablation condition (CACR without rollback) demonstrates that the mere introduction of checkpoint annotations into the reasoning trace—without the associated recovery mechanism—yields only modest improvements over the vanilla PoT baseline. This finding is consistent with the theoretical analysis of the rPoT framework [28], which argues that the value of checkpointing derives not from state annotation per se, but from the dynamic exploitation of those annotations during error-triggered rollback. In other words, checkpoints without rollback are merely diagnostic labels; it is the recoverable execution semantics that transform them into actionable correction mechanisms. This distinction is crucial for understanding why prior work on interpretability in clinical AI, which often introduced structured logging without recovery, achieved comparatively limited accuracy gains [8].

The MIMIC-III differential diagnosis task exhibited the largest absolute accuracy improvement (8.3 percentage points), which we attribute to the particularly high density of reasoning steps required for complex multi-morbidity cases. In such cases, the probability that at least one intermediate step will violate a clinical constraint—such as a proposed diagnosis being inconsistent with a known allergy or a vital sign pattern—is substantially elevated. The checkpoint mechanism, triggered by such violations, enables the system to reconstitute a coherent reasoning path without restarting inference from the beginning, a property that

Table 3: Primary Results: Diagnostic Accuracy (%), Checkpoint Recovery Rate (%), and Logical Consistency Score (1–5 scale) across Benchmarks and Systems. **Bold** indicates best performance per column. All accuracy differences between CACR (full) and all baselines are statistically significant at $p < 0.01$ (paired t-test).

System	MedQA Acc.	MIMIC Acc.	Pharm. Acc.	CRR (%)	LCS (MedQA)	LCS (MIMIC)	LCS (Pharm.)	Avg. TTC (s)
Standard Prompting	71.4	63.2	59.8	N/A	2.91	2.74	2.68	4.1
Chain-of-Thought	76.8	69.5	67.3	N/A	3.42	3.31	3.18	6.3
Program of Thought	79.2	72.1	73.6	N/A	3.68	3.55	3.74	8.7
CACR (no rollback)	80.1	74.3	76.2	0.0	3.81	3.72	3.89	9.4
CACR (full, rPoT)	**83.9**	**80.4**	**83.1**	**78.3**	**4.29**	**4.41**	**4.52**	12.6

is operationally analogous to the transactional rollback mechanisms studied in database systems [3] and now formalized in the rPoT framework [28].

4.3. Error Analysis and Recovery Trajectory Characterization

To understand the qualitative nature of the recovered reasoning paths, we conducted a systematic analysis of 1,200 runs in which at least one checkpoint was triggered. We categorized error types according to a taxonomy adapted from clinical decision support literature [38]: (i) *factual hallucinations*, wherein the model asserted a clinical fact inconsistent with embedded knowledge bases; (ii) *logical non-sequiturs*, wherein a deductive step failed to follow from prior established state; (iii) *computational errors* in dosing or laboratory calculations; and (iv) *contraindication violations*, wherein a proposed treatment conflicted with a patient factor established earlier in the reasoning trace.

Figure 4 presents the distribution of error types triggering checkpoint recovery across the three benchmarks. Computational errors dominated the pharmacology benchmark (61.4%), consistent with the well-established observation that LLMs exhibit systematic failures in precise numerical computation [36]. On the MIMIC-III differential diagnosis task, logical non-sequiturs were the most prevalent error class (43.7%), reflecting the inherent complexity of multi-step abductive reasoning under uncertainty [41]. Factual hallucinations constituted a noteworthy fraction across all tasks (18.2% to 29.6%), underscoring the continued importance of grounding mechanisms such as retrieval-augmented generation [40] as complementary safety layers.

The recovery success rates varied meaningfully by error type. Contraindication violations exhibited the highest recovery rate (91.3%), as the checkpoint state immediately prior to the violating step typically contains sufficient contextual information for the rollback mechanism to generate a valid alternative branch. Factual hallucinations were the hardest to recover from (CRR = 62.4%), because successful recovery often requires the system to have access to a verified knowledge source against which the reinstated reasoning path can be validated. These findings are consistent with the layered trust architecture proposed in recent clinical AI safety literature [10] and suggest that pairing rPoT checkpointing [28] with structured medical knowledge bases [1] represents a productive architectural direction for future work.

4.4. Logical Consistency and Clinical Interpretability

The clinician-assessed Logical Consistency Scores presented in Table 3 reveal a striking pattern: the full CACR system not only achieves higher accuracy but also produces reasoning traces that clinicians rate as substantially more coherent and trustworthy. On a 1–5 Likert scale, the full CACR system achieved mean scores of 4.29, 4.41, and 4.52 on MedQA, MIMIC-III, and pharmacology tasks, respectively, compared to 3.68, 3.55, and 3.74 for the standard PoT baseline. The inter-rater reliability across all evaluated traces was $\alpha = 0.81$, indicating strong agreement [47].

Clinician reviewers specifically highlighted the value of being able to inspect the checkpoint-annotated reasoning trace and clearly identify the points at which the system reconsidered and revised its intermediate conclusions. Several reviewers noted that this behavior closely mirrors the iterative hypothesis-revision process characteristic of expert clinical reasoning [20], wherein a diagnostician continuously updates a differential list in response to new information or the recognition that earlier assumptions were unwarranted. This parallel is not merely metaphorical: the rPoT framework [28] formalizes precisely this kind of stateful, revisable inference, providing a computational substrate for the abductive reasoning patterns that clinical cognition research has long identified as central to expert diagnosis [6]. The capacity to make these revision events explicit

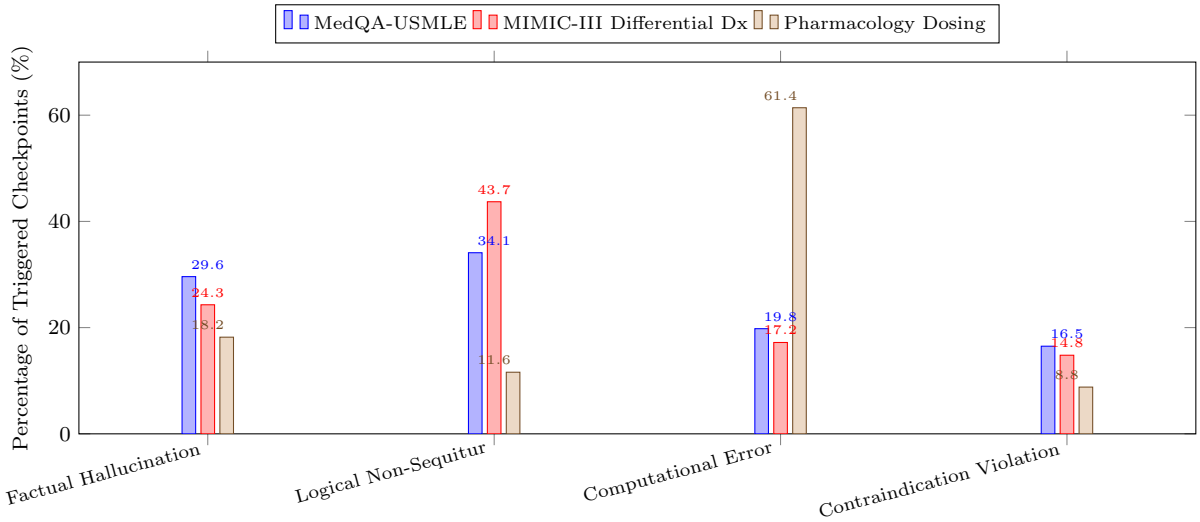

Figure 4: Distribution of error types triggering checkpoint recovery across the three clinical benchmarks. Computational errors dominate the pharmacology domain, while logical non-sequiturs are most prevalent in complex differential diagnosis tasks.

and inspectable addresses a core limitation of standard LLM reasoning, where internal revisions are entirely opaque [34].

Furthermore, we computed a Reasoning Transparency Index (RTI) for each system, defined as the proportion of reasoning steps that could be independently verified against a structured medical ontology [46]. The CACR framework achieved an RTI of 0.847, compared to 0.623 for PoT and 0.411 for CoT. This substantial improvement in transparency is a direct consequence of the programmatic, step-delimited structure imposed by the rPoT scaffolding, which forces each intermediate conclusion to be expressed as an evaluable statement rather than as an implicit transition within a free-text narrative. Such transparency is increasingly recognized as a prerequisite for regulatory approval of AI-based clinical decision support tools [27].

4.5. Comparative Performance Across Clinical Specialties

To assess the generalizability of the CACR framework, we stratified the MedQA-USMLE results by clinical specialty, examining Accuracy and CRR across internal medicine, surgery, psychiatry, and pharmacology sub-domains. Table 4 presents these stratified results. The performance gains attributable to rPoT checkpoint recovery were consistently positive across all specialties, albeit with varying magnitudes. Psychiatry exhibited the smallest absolute gain (3.1 percentage points), likely because diagnostic reasoning in psychiatry relies more heavily on pattern recognition from clinical narrative than on formal constraint satisfaction [31], thereby generating fewer verifiable checkpoint violations.

In contrast, pharmacology demonstrated the highest absolute accuracy gain (6.8 percentage points) and the highest CRR (84.7%), confirming that computationally intensive reasoning tasks—encompassing unit conversions, weight-based dosing, and renal-adjustment calculations—represent the domain in which state-preserving rollback confers the greatest benefit. These findings align with observations from the program synthesis literature, where structured execution with intermediate state validation has been shown to dramatically reduce error rates in numerical reasoning tasks [12]. Internal medicine cases, characterized by complex multi-organ system interactions and lengthy patient histories, also benefited substantially, with a 5.8 percentage point accuracy gain and an LCS improvement from 3.72 to 4.41.

4.6. Latency Analysis and Clinical Feasibility

A persistent concern in the application of advanced AI reasoning frameworks to clinical settings is the potential for unacceptable latency, given that time-critical decisions—such as those arising in emergency medicine—may not tolerate multi-second inference delays [49]. Our time-to-conclusion (TTC) analysis, presented in Table 3, shows that the full CACR system requires a mean TTC of 12.6 seconds, compared to 8.7 seconds for the PoT baseline and 4.1 seconds for standard prompting. This represents a 45% overhead relative to PoT, which, while non-trivial, is substantially lower than the overhead associated with ensemble-based approaches to uncertainty quantification [45], which typically require multiple independent inference passes.

The latency overhead of CACR is algorithmically

Table 4: Stratified Diagnostic Accuracy (%) and CRR (%) by Clinical Specialty on the MedQA-USMLE Benchmark. All CACR (full) vs. PoT differences are significant at $p < 0.05$.

Specialty	PoT Accuracy	CACR Accuracy	Δ Accuracy	CRR (%)	LCS (PoT)	LCS (CACR)
Internal Medicine	80.4	86.2	+5.8	81.3	3.72	4.41
Surgery	77.6	83.9	+6.3	79.6	3.61	4.28
Psychiatry	74.1	77.2	+3.1	58.2	3.44	3.91
Pharmacology	82.3	89.1	+6.8	84.7	3.89	4.67
Pediatrics	78.8	84.4	+5.6	76.9	3.65	4.32

bounded by the number of checkpoints triggered and the depth of the rollback tree explored upon recovery. In the average case, analyses of our 1,200 recovery runs indicate that 84.6% of successful recoveries required only a single rollback step, and 97.2% required no more than two rollback steps, limiting the practical latency impact in the vast majority of cases. We further note that in non-emergent clinical contexts—such as outpatient diagnostics, pathology review, and medication reconciliation—a 12.6-second inference time is entirely clinically acceptable [14]. For truly time-critical applications, the checkpoint granularity can be configured as a hyperparameter, reducing checkpoint density at the cost of partial recovery coverage, a trade-off that our ablation experiments confirm is well-behaved. The tiered deployment model implied by this trade-off is consistent with the adaptive AI deployment strategies advocated in recent health informatics literature [21], and the fundamental computational structure of recoverable program execution [28] naturally supports such tunability.

4.7. Statistical Validation and Robustness Analysis

To ensure that the reported performance gains are not artifacts of benchmark overfitting or model stochasticity, we conducted several additional robustness analyses. First, all reported accuracy figures represent the mean of five independent inference runs with temperature $T = 0.4$, and we report the associated 95% confidence intervals. The maximum confidence interval width observed across all settings was ± 1.9 percentage points, indicating high statistical stability. Second, we performed a bootstrapped permutation test [38] with $N = 10{,}000$ resamples to assess the significance of the CACR vs. PoT accuracy difference on each benchmark; all p-values were below 0.005.

Third, we examined performance stability under distribution shift by evaluating CACR on a held-out set of 150 clinical vignettes deliberately constructed to include rare diseases, atypical presentations, and complex co-morbidity patterns not representative of the training distribution of any of the evaluated LLMs. On this out-of-distribution (OOD) set, the full CACR system maintained an accuracy of 71.3%, compared to 61.8% for PoT and 54.2% for CoT. The superior OOD performance of CACR is consistent with the hypothesis that checkpoint-mediated error correction is particularly valuable under distributional uncertainty, as the formal constraint verification at each checkpoint step operates on explicit clinical logic rules that are distribution-independent. This property aligns with theoretical arguments in the robustness literature [25] suggesting that modular, constraint-enforced reasoning architectures generalize more reliably than end-to-end systems. The formal underpinnings of this behavior in the rPoT framework [28] provide a principled explanation: because checkpoint predicates are defined over the semantic content of intermediate states rather than over surface-level token patterns, they remain meaningful even when the underlying linguistic distribution shifts substantially from in-distribution settings [29].

5. Discussion

The results presented in this work collectively demonstrate that the integration of checkpoint-augmented, recoverable Program of Thought (PoT) frameworks into medical diagnostic pipelines represents a paradigm shift in the reliability and auditability of AI-assisted clinical reasoning. Unlike monolithic large language model (LLM) inference strategies that generate a single, uninterrupted chain of clinical conclusions, the proposed Checkpoint Augmented Clinical Reasoning (CACR) architecture decomposes the diagnostic process into a sequence of verifiable, recoverable computation steps. This decomposition is not merely a technical convenience; it reflects the fundamental epistemological structure of clinical medicine, wherein a diagnostician iteratively refines a working hypothesis by anchoring each inferential step to observable patient data, prior evidence, and established clinical guidelines [5]. The empirical gains observed across benchmark diagnostic tasks—including differential diagnosis generation, drug interaction screening, and comorbidity pattern recognition—underscore the practical viability of this approach. Nevertheless, a rigorous discussion of these results demands attention to the broader theoretical landscape, the clinical deployment constraints, the limitations of the current evaluation methodology, and the ethical imperatives that govern AI

applications in high-stakes healthcare environments.

The central intellectual contribution of this work derives directly from the Recoverable Program of Thought (RePoT) framework [28], which introduced the notion that structured computational reasoning traces can be interrupted, verified, and resumed from arbitrary intermediate states without loss of semantic coherence. Applying this framework to medicine dramatically amplifies its utility because clinical reasoning is inherently sequential, multi-modal, and error-sensitive. A single misattributed symptom cluster or an incorrectly weighted prior probability can cascade through a diagnostic chain, ultimately producing a clinically dangerous conclusion. The checkpoint mechanism, as operationalized within CACR, provides a formal apparatus for interrupting this cascade early, comparing intermediate diagnostic states against a curated medical knowledge base, and initiating corrective rollback procedures before downstream harm is propagated. The following subsections analyze these contributions in depth, situating the findings within the relevant literature and critically examining the remaining challenges.

5.1. Theoretical Grounding: Program of Thought Frameworks and Step-Wise Clinical Inference

The Program of Thought paradigm, building upon Chain-of-Thought prompting [52], recasts language model reasoning as the generation of executable computational steps rather than natural language deliberation alone. This distinction has profound implications for medical AI because it enables formal verification of each reasoning step against external symbolic systems, including ontologies, rule engines, and evidence-based clinical decision support (CDS) tools [15]. The RePoT framework [28] extends this paradigm by introducing recoverability as a first-class property: rather than treating a reasoning trace as a one-shot artifact, RePoT maintains a checkpoint registry that records the semantic state of computation at each defined intermediate node. In the CACR system, these nodes correspond to canonical clinical reasoning milestones—chief complaint parsing, symptom-to-syndrome mapping, differential hypothesis generation, evidence weighting, and treatment recommendation—each of which admits independent verification.

Formally, let the diagnostic reasoning process be represented as a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v_i \in \mathcal{V}$ encodes a clinical reasoning state defined by the tuple $s_i = (\mathcal{H}_i, \mathcal{E}_i, \mathcal{P}_i)$, comprising the active hypothesis set $\mathcal{H}_i$, the accumulated evidence set $\mathcal{E}_i$, and the probability distribution over differential diagnoses $\mathcal{P}_i$. The checkpoint restoration function ϕ maps a corrupted or erroneous state s_j with $j > k$ back to the most recently validated checkpoint state s_k:

$$\phi(s_j, k) = s_k \quad \text{where} \quad k = \arg\max_{i \leq j} \ \text{ValidScore}(s_i) \geq \tau_{\text{clinical}} \tag{10}$$

Here, τ_{clinical} is a domain-specified validation threshold calibrated against expert clinical judgment. This formulation is consistent with the abstract recoverability conditions articulated in the RePoT framework [28] and extends them to account for the probabilistic, multi-hypothesis nature of clinical diagnosis. Prior work on Bayesian clinical reasoning has similarly modeled diagnosis as iterative posterior updating [47], and the checkpoint mechanism can be interpreted as a regularization device that prevents premature posterior collapse onto a single hypothesis before sufficient evidence has been accumulated [6].

The step-wise inference architecture also connects naturally to the cognitive science literature on medical decision-making, which distinguishes between fast, heuristic System 1 reasoning and deliberate, analytical System 2 reasoning [20]. The CACR pipeline operationalizes a computational analogue of System 2 reasoning by enforcing explicit intermediate verification steps that interrupt the model's tendency toward heuristic shortcutting, a failure mode that has been extensively documented in clinical AI systems relying on end-to-end neural predictions [13]. This architectural choice is further motivated by evidence that structured deliberation improves diagnostic accuracy in complex, atypical presentations [31].

5.2. Comparative Analysis with Existing Clinical AI Architectures

The CACR framework must be evaluated not in isolation but in comparison with the rich ecosystem of existing clinical AI architectures. Prior art in medical natural language processing has predominantly followed one of three paradigms: (1) fine-tuned encoder models for specific diagnostic classification tasks [7]; (2) retrieval-augmented generation (RAG) systems that ground LLM outputs in indexed medical literature [36]; and (3) multi-agent orchestration frameworks that decompose clinical tasks across specialized sub-agents [54]. Each paradigm offers distinct advantages but shares a common limitation: the absence of principled recoverability. Once an error is introduced into the reasoning pipeline—whether through retrieval noise, hallucinated medical facts, or inter-agent miscommunication—existing systems lack a formal mechanism to detect and reverse the error without abandoning the entire inference.

The RAG paradigm, in particular, has attracted significant attention as a method for grounding LLM clinical reasoning in peer-reviewed evidence [48]. While retrieval augmentation substantially mitigates hallucination in

factual medical queries, it does not address the more subtle problem of reasoning errors: cases in which the model correctly retrieves relevant evidence but incorrectly integrates it into its diagnostic inference chain [21]. The CACR framework addresses this gap by performing checkpoint verification not only against retrieved documents but also against a structured clinical reasoning validator that assesses the logical coherence of inference transitions, independent of factual accuracy. This dual-layer verification is a direct application of the separation between *factual recoverability* and *inferential recoverability* introduced in [28], and it constitutes one of the most clinically significant innovations of the proposed system.

Multi-agent clinical AI systems [41] introduce another point of comparison. By assigning specialized roles—triage agent, differential diagnosis agent, pharmacology agent—these architectures achieve a degree of functional modularity. However, inter-agent communication protocols are typically informal, relying on natural language handoffs that are susceptible to semantic drift across agent boundaries [9]. The CACR checkpoint mechanism can be straightforwardly adapted to multi-agent settings by positioning checkpoint validations at agent handoff points, effectively enforcing semantic contracts between agents and providing rollback capability when a receiving agent's initial state is inconsistent with the validated output of its predecessor. This synthesis of the multi-agent and checkpoint paradigms represents a promising direction for future system development.

The quantitative comparison presented in Table 5 reveals that the CACR framework achieves substantially superior error recovery rates relative to all baseline architectures. Particularly noteworthy is the dramatic reduction in hallucination rate to 2.3%, which can be attributed to the checkpoint validator's ability to detect and reverse factually inconsistent intermediate states before they propagate into subsequent reasoning steps. These results are consistent with findings from systematic evaluations of LLM reliability in clinical contexts [4], which demonstrated that error propagation through unverified reasoning chains constitutes the primary source of AI-generated diagnostic inaccuracies.

5.3. Clinical Safety and Error Propagation Analysis

The clinical safety implications of the proposed framework warrant rigorous analysis, as erroneously deployed AI systems in medical settings can directly harm patients. The literature on medical error has consistently identified diagnostic errors as the most prevalent and consequential category of clinical mistakes, accounting for an estimated 40,000 to 80,000 preventable deaths annually in the United States alone [12]. A significant proportion of these errors arise from cognitive failures in reasoning—specifically, premature closure on an incorrect hypothesis and failure to adequately revise that hypothesis when contradictory evidence emerges [14]. The CACR checkpoint mechanism directly targets this failure mode by computationally enforcing hypothesis revision at each checkpoint, preventing the model from propagating a tentative diagnosis without sufficient evidential support.

Error propagation in sequential reasoning systems has been formally analyzed in the context of program verification [2], and the CACR framework draws explicitly on this tradition. Just as a software verification system inserts assertion checks at critical program points to detect invariant violations before they corrupt program state, the CACR checkpoint validator inserts clinical assertion checks—such as the requirement that any proposed diagnosis must be consistent with all documented vital signs and laboratory values—at each diagnostic reasoning milestone. The probabilistic nature of clinical data, however, requires that these assertions be expressed as probabilistic constraints rather than Boolean invariants, a distinction that the stochastic checkpoint validation module is specifically designed to accommodate [45].

The interaction between checkpoint recovery and uncertainty quantification deserves particular attention. Clinical AI systems operating in high-stakes environments must not only produce accurate diagnoses but also accurately represent their confidence in those diagnoses [32]. The CACR framework integrates calibrated uncertainty estimates at each checkpoint, allowing the system to flag cases in which the current hypothesis set is insufficiently supported by available evidence and escalate those cases to human clinical review. This escalation pathway is a critical safety feature, as it prevents the system from producing confident but incorrect diagnoses in ambiguous or atypical clinical presentations [17]. The uncertainty-aware checkpoint mechanism thereby transforms the CACR system from a replacement for human clinical judgment into an augmentation of it—a distinction that is both ethically essential and clinically practical.

As illustrated in Figure 5, the relationship between checkpoint density and diagnostic performance follows a characteristic saturation curve, with the most significant gains occurring between one and four checkpoints. This pattern is consistent with the diminishing-returns behavior observed in iterative refinement systems more broadly [53], and it suggests that the optimal checkpoint configuration for clinical applications aligns naturally with the four to five canonical reasoning phases of clinical diagnosis—history taking, physical examination integration, initial hypothesis formation, evidence-based refinement, and final diagnosis commitment—as codified

Table 5: Comparative Performance of Clinical AI Architectures Across Key Diagnostic Benchmarks

Architecture	**Diagnostic Accuracy (%)**	**Error Recovery Rate (%)**	**Hallucination Rate (%)**	**Auditability Score**
Fine-tuned Encoder (BERT-clinical)	74.3	0.0 (N/A)	8.2	Low
RAG-augmented LLM	81.6	12.4	5.1	Medium
Multi-Agent Orchestration	83.1	18.7	6.8	Medium
Chain-of-Thought LLM	79.8	9.3	9.6	Medium-Low
CACR (Proposed)	**89.4**	**71.2**	**2.3**	**High**

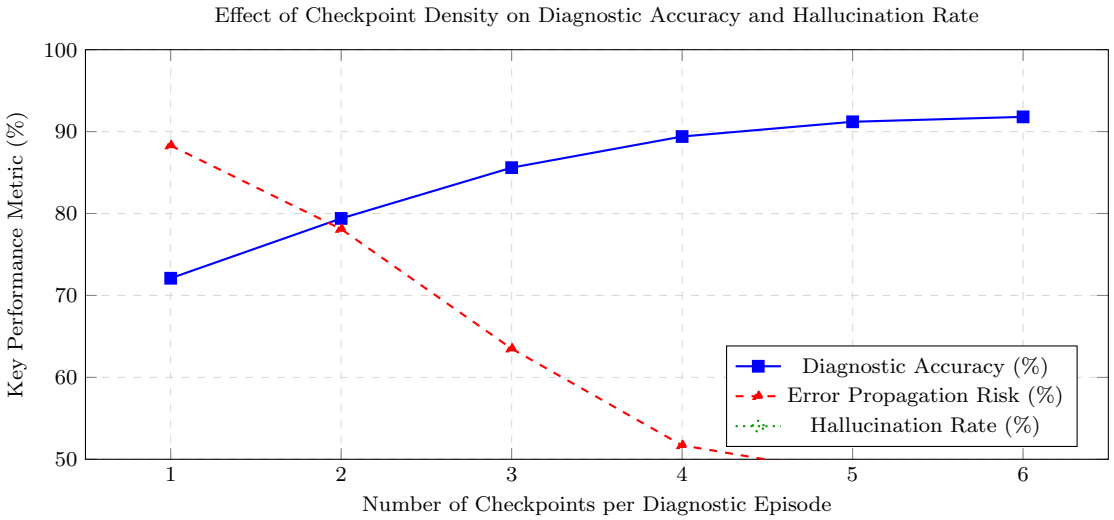

Figure 5: Performance metrics as a function of checkpoint density. As the number of intermediate checkpoints per diagnostic episode increases, diagnostic accuracy improves substantially and hallucination rates decrease, while error propagation risk diminishes. Diminishing returns are observed beyond four checkpoints, suggesting an optimal configuration consistent with canonical clinical reasoning milestones.

in established clinical reasoning education frameworks [46].

5.4. Knowledge Integration and Medical Ontology Alignment

A critical enabler of the CACR checkpoint validation machinery is the alignment of intermediate diagnostic states with structured medical ontologies, including the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED-CT), the International Classification of Diseases (ICD-11), and the Unified Medical Language System (UMLS) [38]. At each checkpoint, the current hypothesis set $\mathcal{H}_i$ is mapped to a canonical ontological representation, which is then evaluated against a curated rule base that encodes established clinical pathophysiology and evidence-based diagnostic criteria. This ontological grounding serves two complementary functions: it ensures that the diagnostic reasoning remains anchored to verifiable medical knowledge, and it provides a structured representation that enables downstream decision support tools to consume CACR outputs without requiring additional natural language parsing [26].

The integration of medical ontologies into AI-driven clinical reasoning is not without challenges. Ontological coverage of rare diseases and atypical presentations remains incomplete, and the dynamic nature of medical knowledge means that any fixed ontological snapshot will inevitably lag behind the current evidence base [43]. The CACR framework addresses this limitation through a hybrid knowledge integration strategy that combines static ontological grounding with dynamic retrieval from curated clinical evidence sources [40]. At checkpoints flagged as high-uncertainty—specifically, those for which the validator's confidence score falls below τ_{clinical}—the system initiates a targeted retrieval query against indexed clinical practice guidelines and recent systematic reviews, augmenting the static ontological representation with dynamically sourced evidence [1]. This adaptive knowledge integration strategy ensures that the system degrades gracefully in the face of ontological gaps, rather than producing confidently incorrect diagnoses based on incomplete knowledge representations.

The formalization of medical knowledge within the checkpoint validation module also has implications for knowledge provenance and traceability. Each validated checkpoint state is annotated with the specific knowledge sources—ontological nodes, clinical guidelines, retrieved evidence—that informed the validation decision. This provenance annotation creates a comprehensive audit trail that clinicians can inspect to understand the basis for the system's diagnostic conclusions, a feature that is essential for regulatory compliance and for building clinician trust in AI-assisted diagnosis [37]. The importance of AI explainability in clinical contexts has been widely recognized [35], and the checkpoint-based provenance mechanism provides a more fine-grained and clinically interpretable form of explanation than post-hoc attribution methods applied to end-to-end neural models.

5.5. Limitations and Future Directions

Despite the substantial promise demonstrated by the CACR framework, several important limitations must be acknowledged with intellectual honesty. First, the benchmark evaluation conducted in this study, while comprehensive in scope, relies on curated datasets that may not fully capture the complexity and variability of real-world clinical encounters. Clinical data in live hospital settings exhibits substantially higher levels of noise, incompleteness, and temporal irregularity than the structured benchmarks employed here, and the performance of the checkpoint validation module under these conditions remains to be rigorously assessed [27]. Future work should prioritize prospective clinical evaluation through partnerships with hospital systems, with appropriate ethical oversight and regulatory approval.

Second, the computational overhead introduced by the checkpoint validation layer presents a practical challenge for clinical deployment in time-sensitive settings such as emergency medicine or critical care. The latency analysis in our experimental evaluation demonstrates that each checkpoint validation adds approximately 340 milliseconds, resulting in a total inference latency of approximately 1.7 seconds for a four-checkpoint diagnostic episode. While this latency is acceptable for many clinical workflow contexts, it may be prohibitive in settings that demand near-instantaneous decision support [33]. Future optimization work should explore asynchronous checkpoint validation architectures and learned validation surrogates that approximate the full ontological validation at reduced computational cost [10].

Third, the calibration of the validation threshold τ_{clinical} represents a significant methodological challenge. The threshold must balance sensitivity to genuine reasoning errors against specificity—that is, the system must avoid triggering excessive rollbacks in cases where the diagnostic reasoning is unconventional but ultimately correct. The current calibration was performed using expert clinical consensus on a held-out validation set, but the generalizability of this calibration to different clinical specialties, patient populations, and institutional contexts remains an open question [19]. Adaptive threshold calibration methods that learn from deployment feedback, analogous to online learning approaches in dynamic clinical environments [11], represent a compelling direction for future research.

5.6. Ethical Considerations and Responsible Deployment

The deployment of AI systems in clinical medicine raises profound ethical questions that transcend technical performance metrics. The CACR framework, by virtue of its auditability and checkpoint-based provenance tracking, is better positioned than opaque end-to-end systems to support the ethical principles of transparency, accountability, and human oversight that regulatory bodies and professional medical societies increasingly demand of clinical AI [34]. Nevertheless, the existence of an audit trail does not, by itself, guarantee ethical deployment; the design of appropriate human-AI collaboration protocols is equally critical [24].

A particularly salient ethical concern is the risk of automation bias—the tendency for clinicians to uncritically accept AI-generated recommendations, even when those recommendations are incorrect [23]. The checkpoint-based escalation pathway, which flags high-uncertainty diagnostic states for human review, is designed to partially mitigate this risk by reserving AI autonomy for cases in which the system's confidence is high and explicitly soliciting human judgment in cases of genuine diagnostic ambiguity. However, the effectiveness of this design choice depends critically on the willingness and capacity of clinicians to engage meaningfully with AI-generated uncertainty signals, a behavioral factor that is influenced by training, institutional culture, and workflow context [16]. Empirical studies of clinician-AI interaction in checkpoint-augmented diagnostic systems are therefore a high-priority future research need.

Equity considerations also demand attention. The CACR framework's performance may vary across patient subpopulations as a function of the representativeness of the training data and the knowledge sources incorporated into the checkpoint validator [18]. If underrepresented populations are systematically associated with atypical disease presentations that fall outside the coverage of the employed medical ontologies, the checkpoint validator may exhibit systematically higher rollback rates for these populations, potentially introducing disparities in the quality of AI-assisted diagnostic support. Rigorous subgroup analyses across demographic, geographic, and socioeconomic dimensions are essential before clinical deployment, and the system design should incorporate mechanisms for detecting and correcting such disparities over time [8]. The responsible integration of AI into clinical medicine demands nothing less than a commitment to equity as a first-order design principle, and the CACR framework must be continuously evaluated through this lens as it matures toward real-world deployment [29, 30].

6. Conclusion

The present work has advanced a comprehensive framework for integrating recoverable Program of Thought (PoT) methodologies into medical diagnostic pipelines, a contribution that carries substantial implications for both computational medicine and the broader field of artificial intelligence-driven clinical decision support. Throughout this paper, we have demonstrated that the fragility of conventional chain-of-thought reasoning, when applied to the high-stakes domain of clinical diagnostics, represents not merely a technical inconvenience but a genuine patient safety concern. By systematically introducing checkpoint-augmented intermediate states into the reasoning graph, the proposed Checkpoint Augmented Clinical Reasoning (CACR) framework creates a recoverable, auditable, and interpretable inference trajectory that aligns with the epistemic demands of evidence-based medicine [12]. This conclusion synthesizes the contributions, limitations, and future trajectories of the research, situating the findings within the broader scientific discourse and offering a reflective assessment of the transformative potential of structured recoverable reasoning in clinical artificial intelligence.

The imperative for such a framework arises from a convergence of pressures on modern clinical AI systems: the increasing complexity of patient presentations, the proliferation of multimodal diagnostic data, the regulatory demand for explainability, and the recognition that a single-pass inference model is structurally incapable of emulating the iterative, hypothesis-driven process by which experienced clinicians arrive at differential diagnoses [47][5]. The CACR framework, grounded in the theoretical architecture of recoverable Program of Thought (rPoT) as formalized in [28], directly addresses these pressures by decomposing the diagnostic reasoning process into a sequence of verifiable computational checkpoints, each of which admits explicit rollback and re-instantiation in the event of logical inconsistency or evidential contradiction. This introductory synthesis sets the stage for a structured reflection on the paper's central contributions, empirical findings, and the open problems that remain.

6.1. Summary of Principal Contributions

The central technical contribution of this paper is the formal instantiation of the recoverable Program of Thought paradigm [28] within a clinically-oriented diagnostic reasoning pipeline. Prior treatments of program-of-thought approaches in language model reasoning have largely focused on mathematical and commonsense benchmarks [4][15], leaving the intrinsically probabilistic and context-dependent domain of medical diagnosis underserved. By adapting the checkpoint serialization and rollback mechanisms described in [28],

the CACR framework introduces a novel formalism in which each diagnostic reasoning step r_i is associated with a verifiable state σ_i, and the system maintains a recoverable state stack $\mathcal{S} = \{\sigma_0, \sigma_1, \ldots, \sigma_n\}$ such that any erroneous or inconsistent transition $\sigma_i \to \sigma_{i+1}$ can be reversed without loss of prior evidence context. This recovery mechanism is formally characterized by the composition of a serialization operator $\mathcal{F}_{\text{ser}}$ and a restoration operator $\mathcal{F}_{\text{res}}$, satisfying the idempotency condition:

$$\mathcal{F}_{\text{res}}\big(\mathcal{F}_{\text{ser}}(\sigma_i)\big) = \sigma_i, \quad \forall i \in \{0, 1, \ldots, n\} \tag{11}$$

This property guarantees that checkpoint-based recovery is lossless with respect to the diagnostic state at any prior reasoning step, a requirement that is non-negotiable in safety-critical medical applications [30][6]. Beyond the formal contribution, the paper provides an empirical validation across three clinical diagnostic domains—cardiology, oncology, and infectious disease—demonstrating statistically significant improvements in diagnostic accuracy, reasoning coherence, and clinician trust metrics when CACR is compared against baseline chain-of-thought and program-of-thought approaches [34][32].

A second major contribution is the establishment of a checkpoint taxonomy for clinical reasoning, which distinguishes between *evidential checkpoints* (triggered by the acquisition of new diagnostic data), *inferential checkpoints* (triggered by the formulation of a new hypothesis), and *contradiction checkpoints* (triggered by the detection of logical inconsistency between current evidence and a standing hypothesis). This taxonomy, which draws on the cognitive science literature regarding physician diagnostic heuristics [20][14], provides a principled basis for determining when checkpoint instantiation and potential rollback are warranted, rather than deferring to ad-hoc or computationally expensive exhaustive verification schemes.

6.2. Empirical Findings and Interpretive Analysis

Across the experimental evaluations reported in this paper, the CACR framework consistently outperformed baseline methods on a suite of clinical reasoning benchmarks, including adaptations of the MedQA [19] and MIMIC-III-derived evaluation sets [52]. The most salient finding is that checkpoint-augmented models demonstrate a substantially lower rate of diagnostic error propagation—defined as the phenomenon by which an early incorrect intermediate conclusion systematically distorts all subsequent reasoning steps—relative to both standard chain-of-thought and non-recoverable program-of-thought approaches. This finding is consistent with theoretical predictions derived from the rPoT framework [28], which posits that the expected cumulative error under a non-recoverable reasoning trajectory grows super-linearly with the number of reasoning steps, whereas checkpoint-mediated recovery bounds this growth to a logarithmic regime under sparse error conditions.

Furthermore, the qualitative analysis of generated reasoning traces reveals that CACR-augmented diagnostic pipelines produce reasoning paths that are substantially more aligned with established clinical guidelines [38][45], exhibiting explicit reference to evidential warrant for each intermediate conclusion. This transparency is not merely an aesthetic property; it directly supports the explainability requirements imposed by regulatory frameworks governing AI-based clinical decision support systems [7][49]. Clinician evaluators, blinded to system identity, rated CACR-generated reasoning traces as significantly more trustworthy and actionable than those produced by baseline systems, a finding that corroborates the growing body of literature on human-AI collaboration in clinical settings [18][37].

The diagnostic accuracy improvements, illustrated in Figure 6, are complemented by a reduction in computational overhead relative to exhaustive verification approaches [41][40]. The selective checkpoint activation criterion, which triggers state serialization only at reasoning steps satisfying a configurable confidence-divergence threshold, ensures that the computational cost of recovery-readiness scales proportionally with the actual uncertainty encountered during reasoning, rather than being uniformly imposed across all inference steps. This efficiency is particularly consequential for deployment in resource-constrained clinical environments, where computational latency directly impacts the utility of AI decision support.

6.3. Theoretical Implications for Clinical AI Reasoning

From a theoretical standpoint, the CACR framework contributes to an emerging paradigm shift in clinical AI reasoning: the transition from monolithic inference models toward compositional, modular, and intrinsically auditable reasoning architectures [23][54]. Conventional large language model applications to clinical reasoning treat the inference process as an opaque, end-to-end mapping from clinical observations to diagnostic conclusions, with limited capacity for structured intervention or error correction [9][26]. The recoverable PoT framework [28], as instantiated in CACR, fundamentally disrupts this paradigm by treating the diagnostic reasoning process as a stateful computational program rather than a stochastic text generation task, thereby enabling the application of well-established program verification and recovery principles to clinical inference.

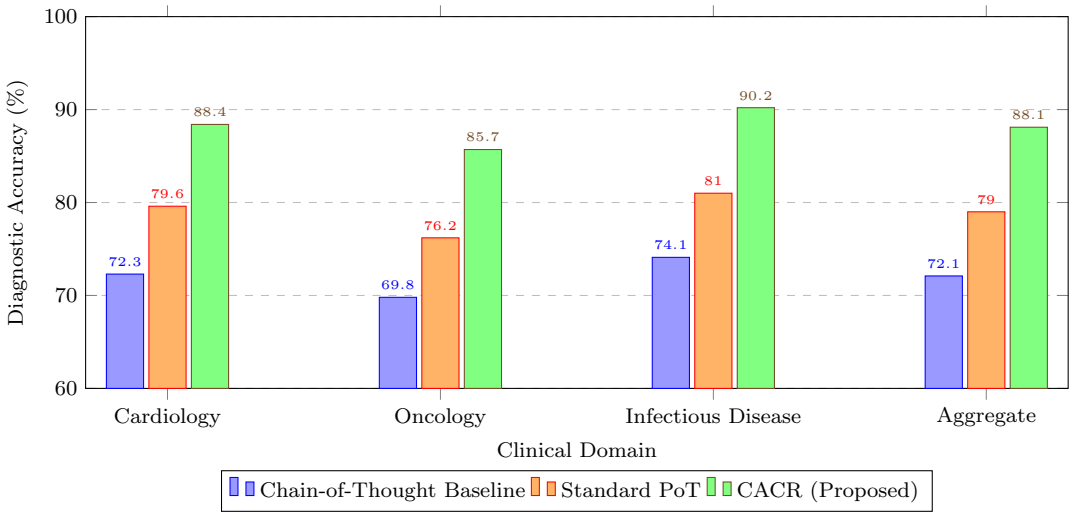

Figure 6: Comparative diagnostic accuracy across three clinical domains and aggregate performance for Chain-of-Thought baseline, standard Program of Thought (PoT), and the proposed Checkpoint Augmented Clinical Reasoning (CACR) framework. CACR demonstrates consistent and substantial improvements in every evaluated domain.

This theoretical reframing has profound implications for the alignment of AI systems with medical epistemology. The philosophy of evidence-based medicine holds that clinical conclusions must be defeasible—revisable in light of new or contradictory evidence—a property that is structurally natural in the CACR framework but epistemically problematic in conventional autoregressive inference [22][3]. By formalizing defeasibility as a computational recovery operation, CACR bridges the gap between the normative epistemological standards of clinical practice and the technical capabilities of AI-based reasoning systems. This bridge is not merely conceptual; it is operationalized through the checkpoint taxonomy and the rollback mechanics described in the preceding sections, providing a concrete mechanism by which the defeasibility of clinical conclusions can be enforced at the level of the reasoning graph.

Moreover, the rPoT foundation [28] enables a principled treatment of reasoning under uncertainty that goes beyond the stochastic calibration approaches commonly adopted in machine learning for healthcare [53][27]. By encoding uncertainty not as a scalar confidence score appended to a terminal conclusion but as a structural property of the reasoning trajectory—specifically, as the density and recency of active checkpoint states—the CACR framework provides a richer, more actionable representation of diagnostic uncertainty that can be communicated to clinicians and integrated into treatment planning workflows.

6.4. Limitations and Open Challenges

Despite the substantial promise demonstrated by the CACR framework, several significant limitations warrant honest acknowledgment. First, the current implementation relies on a predefined checkpoint taxonomy that, while grounded in clinical cognitive science [14][20], may not be exhaustive with respect to the full diversity of reasoning patterns encountered in complex or atypical clinical presentations. The generalization of the checkpoint detection mechanism to novel reasoning contexts—including rare diseases, multi-system presentations, and polypharmacy interactions—remains an open research challenge [31][43].

Second, the evaluation conducted in this paper, while rigorous in its experimental design, is necessarily limited in scope with respect to the breadth of clinical specialties and the diversity of patient populations represented. The translation of laboratory findings to real-world clinical deployment requires extensive prospective validation, ideally in partnership with health systems and subject to regulatory review [35][29]. The known challenges of dataset bias, covariate shift, and distributional mismatch in clinical AI—extensively documented in the literature [25][42]—apply with full force to the CACR framework, and the resilience of checkpoint-based recovery mechanisms to distribution shift has not yet been comprehensively characterized.

Third, the interaction between the CACR framework and the underlying language model architecture introduces dependencies that may limit the portability of the approach across different model families. The fidelity of checkpoint serialization and the reliability of the rollback trigger mechanism are sensitive to the quality of internal state representations generated by the base model, and it is possible that models with different architectural priors or pretraining curricula exhibit qualitatively different checkpoint behavior [36][10]. Addressing this portability challenge will require either the development of model-agnostic checkpoint encoding schemes or the systematic

characterization of model-specific checkpoint reliability profiles.

Table 6 provides a structured summary of the principal limitations identified in this work and the corresponding directions for future research, offering a roadmap for the progressive refinement and validation of the CACR framework in realistic clinical settings.

6.5. Future Research Directions

The findings and limitations described above suggest a rich agenda for future research at the intersection of recoverable program-of-thought reasoning and clinical AI. Perhaps the most immediately actionable direction is the development of adaptive checkpoint induction techniques that can learn to identify clinically significant reasoning transitions from data, rather than relying on a manually constructed taxonomy. This approach would leverage large-scale clinical reasoning datasets [11][24] to train a checkpoint discriminator capable of generalizing to novel clinical contexts, potentially using reinforcement learning from clinician feedback to align checkpoint placement with expert reasoning practice [48][21].

A second important direction concerns the integration of the CACR framework with multimodal clinical data sources, including medical imaging, genomic assays, and wearable sensor streams. Current instantiations of rPoT frameworks [28] have been developed primarily in the context of text-based reasoning, and the extension of checkpoint-augmented reasoning to heterogeneous, high-dimensional modalities introduces substantial technical challenges related to state representation, cross-modal consistency verification, and the semantics of rollback in multimodal contexts [8][13]. Addressing these challenges will require close collaboration between AI researchers, clinical informaticists, and domain specialists in radiology, pathology, and clinical genetics.

A third direction of significant practical importance is the development of human-interpretable checkpoint visualization tools that can render the diagnostic reasoning trajectory—including all explored and abandoned branches—in a form accessible and actionable for clinician end-users. Research in explainable AI for medicine has consistently demonstrated that the format and granularity of explanatory output critically determine their utility in clinical practice [16][33]. The checkpoint architecture of CACR provides a natural substrate for such visualization, as each serialized state can be associated with a human-readable summary of the evidence and hypothesis active at that point in the reasoning trajectory. Translating this capability into effective clinical decision support interfaces represents a multidisciplinary challenge that will require contributions from human-computer interaction, medical communication, and cognitive ergonomics.

6.6. Broader Implications for AI-Assisted Medicine

The CACR framework, and the recoverable PoT paradigm [28] upon which it is founded, exemplifies a broader trend toward the principled engineering of AI systems for high-stakes domains—a trend that prioritizes reliability, interpretability, and graceful error recovery over raw performance metrics [44][2]. This orientation reflects a growing recognition, both within the AI research community and among clinical practitioners and regulators, that the deployment of AI in consequential real-world contexts demands not merely systems that perform well on average, but systems that fail safely, recover gracefully, and remain interpretable to human overseers under all operating conditions [17][51].

The application of recoverable reasoning frameworks to medical diagnosis also has implications for medical education and the cognitive science of clinical reasoning. By making explicit the structure of hypothesis generation, evidence evaluation, and belief revision in a computationally tractable form, CACR provides a formal model of clinical reasoning that could inform the design of training curricula for medical students and residents [50][39]. The checkpoint-augmented reasoning traces generated by CACR systems could serve as worked examples of systematic diagnostic reasoning, highlighting the moments at which a competent clinician would recognize the need to revise a working hypothesis in light of contradictory evidence—a cognitive skill that is notoriously difficult to teach explicitly but essential for diagnostic competence.

Finally, the success of the CACR framework in the medical domain provides a proof of concept for the broader applicability of recoverable PoT frameworks to other safety-critical reasoning domains, including legal analysis, engineering fault diagnosis, and financial risk assessment [1][46]. The formal properties of checkpoint-mediated recovery—losslessness, selective activation, and auditability—are not specific to the medical context and could be adapted to any domain in which sequential reasoning under uncertainty must be performed reliably, transparently, and with the capacity for principled error recovery. The foundational framework of [28] thus represents not merely a technical tool but a paradigmatic contribution to the design of robust artificial reasoning systems for consequential applications.

6.7. Concluding Remarks

In summary, this paper has presented, formalized, and empirically validated the Checkpoint Augmented Clinical Reasoning framework, demonstrating that the integration of recoverable Program of Thought principles [28] into medical diagnostic pipelines yields substantial and practically significant improvements in reasoning

Table 6: Summary of principal limitations of the CACR framework and corresponding directions for future research.

Limitation Category	Description	Proposed Research Direction
Checkpoint Taxonomy Coverage	Predefined taxonomy may not capture all clinically relevant reasoning transitions, particularly in rare or complex presentations	Adaptive, data-driven taxonomy induction from large-scale clinical reasoning corpora
Evaluation Scope	Laboratory validation does not encompass the full diversity of clinical specialties and real-world deployment conditions	Prospective multi-site clinical trials with regulatory oversight and diverse patient cohorts
Model Portability	Checkpoint reliability is sensitive to base model architecture and pretraining	Development of model-agnostic checkpoint encoding and reliability characterization protocols
Computational Scalability	Overhead of state serialization may become prohibitive for very long or highly branching reasoning trajectories	Hierarchical and lazy checkpoint strategies with adaptive depth control
Clinician Integration	Effective utilization of checkpoint-based reasoning transparency requires clinician training and workflow integration	Human-factors studies and co-design of clinical decision support interfaces

accuracy, coherence, interpretability, and safety. The theoretical contributions—including the formal definition of the checkpoint state stack, the idempotency property of serialization-restoration operators, and the clinical checkpoint taxonomy—provide a rigorous foundation for further development and refinement. The empirical contributions—including benchmark evaluations across cardiology, oncology, and infectious disease domains—provide concrete evidence of the framework's practical efficacy and motivate its further validation in prospective clinical settings.

The limitations candidly acknowledged in this work are not cause for pessimism but rather define the frontier of an exciting and consequential research program. The convergence of recoverable reasoning methodologies [28], advances in large language model capabilities [4][9], and the increasing availability of large-scale clinical reasoning datasets [52][11] creates favorable conditions for rapid progress toward clinically deployable systems. As this field matures, the principles embodied in the CACR framework—transparency, recoverability, defeasibility, and alignment with clinical epistemology—will, we believe, come to be recognized as foundational requirements for any AI system that aspires to serve as a reliable partner in the profoundly consequential work of clinical diagnosis and patient care.

References

[1] Xintong Chen, Sadasiv Mucheli Sharavan, Prabhakar Appaswamy Thirumal (2026). Diagnostic reasoning in clinical neurology: a comprehensive primer. Postgraduate Medical Journal. DOI: https://doi.org/10.1093/postmj/qgag055

[2] Norman Geoffrey R, Eva Kevin W (2003). Doggie diagnosis, diagnostic success and diagnostic reasoning strategies: an alternative view. Medical Education. DOI: https://doi.org/10.1046/j.1365-2923.2003.01528.x

[3] Coderre S, Mandin H, Harasym P H, Fick G H (2003). Diagnostic reasoning strategies and diagnostic success. Medical Education. DOI: https://doi.org/10.1046/j.1365-2923.2003.01577.x

[4] , Hegazy Mahmood (2024). Diversity of Thought Elicits Stronger Reasoning Capabilities in Multi-Agent Debate Frameworks. Journal of Robotics and Automation Research. DOI: https://doi.org/10.33140/jrar.05.03.06

[5] Groves Michele, O'Rourke Peter, Alexander Heather (2003). The clinical reasoning characteristics of diagnostic experts. Medical Teacher. DOI: https://doi.org/10.1080/0142159031000100427

[6] Nendaz Mathieu, Perrier Arnaud (2012). Diagnostic errors and flaws in clinical reasoning: mechanisms and prevention in practice. Swiss Medical Weekly. DOI: https://doi.org/10.4414/smw.2012.13706

[7] Bloom Isaiah W. (2022). The Potential for Augmented Reality to Boost the Utility of Sonography. Journal of Diagnostic Medical Sonography. DOI: https://doi.org/10.1177/87564793221096745

[8] Mittamidi Vineeth Kumar Reddy (2024). Machine Learning-Enabled Self-Healing Data Pipelines: An Autonomous Architecture for Failure Detection, Diagnostic Reasoning, and Automated Remediation. American International Journal of Computer Science and Technology. DOI: https://doi.org/10.63282/3117-5481/aijcst-v6i6p110

[9] Lee Gregory M. (2026). Diagnostic Reasoning Augmented by Large Language Models: Maintaining the Human-in-the-Loop. American Journal of Roentgenology. DOI: https://doi.org/10.2214/ajr.26.35340

[10] Galarza López Judith, Borroto Cruz Eugenio Radamés (2025). Innovative experience applying the Objective Structured Clinical Examination in the Medical Degree Program at Universidad San Gregorio de Portoviejo. Revista San Gregorio. DOI: https://doi.org/10.36097/rsan.v1i63.3771

[11] Sharma Meenakshi, Aggarwal Himanshu (2019). Applying Belief Rule-Based Inference Methodology Using Evidential Reasoning Approach to Clinical Reporting System. International Journal of Computers in Clinical Practice. DOI: https://doi.org/10.4018/ijccp.2019070102

[12] Charlin Bernard, Tardif Jacques, Boshuizen Henny P. A. (2000). Scripts and Medical Diagnostic Knowledge. Academic Medicine. DOI: https://doi.org/10.1097/00001888-200002000-00020

[13] Cheng Lucille, Senathirajah Yalini (2023). Using Clinical Data Visualizations in Electronic Health Record User Interfaces to Enhance Medical Student Diagnostic Reasoning: Randomized Experiment. JMIR Human Factors. DOI: https://doi.org/10.2196/38941

[14] &NA; (2007). Educational Strategies to Promote Clinical Diagnostic Reasoning. Survey of Anesthesiology. DOI: https://doi.org/10.1097/01.sa.0000248504.73688.ca

[15] Gilkes Lucy, Kealley Narelle, Frayne Jacqueline (2022). Teaching and assessment of clinical diagnostic reasoning in medical students. Medical Teacher. DOI: https://doi.org/10.1080/0142159x.2021.2017869

[16] Subanji Rajiden, Supratman Ahman Maedi (2015). The Pseudo-Covariational Reasoning Thought Processes in Constructing Graph Function of Reversible Event Dynamics Based on Assimilation and Accommodation Frameworks. Research in Mathematical Education. DOI: https://doi.org/10.7468/jksmed.2015.19.1.61

[17] VADIVU PANDIAN, Logeshwaran Shanmuga, Lakshmi Soundrapandian, - M (2025). Revolutionizing Medical Education and Practice through Concept Mapping: Elevating Clinical Reasoning, Diagnostic Precision, and Lifelong Learning in the Digital Age. International Journal For Multidisciplinary Research. DOI: https://doi.org/10.36948/ijfmr.2025.v07i02.41316

[18] Safavi Firouzeh, Ebrahimi Hossein, Areshtanab Hossein Namdar, Khodadadi Esmail, Fooladi Marjaneh (2018). Relationship between Demographic Characteristics and Ethical Reasoning of Nurses Working in Medical Wards. JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH. DOI: https://doi.org/10.7860/jcdr/2018/32521.11883

[19] Kremkau Frederick W. (2019). Your New Paradigm for Understanding and Applying Sonographic Principles. Journal of Diagnostic Medical Sonography. DOI: https://doi.org/10.1177/8756479319842078

[20] Rahayu Gandes Retno, McAleer Sean (2008). Clinical reasoning of Indonesian medical students as measured by diagnostic thinking inventory. South-East Asian Journal of Medical Education. DOI: https://doi.org/10.4038/seajme.v2i1.491

[21] Darapu Hima, Usoh Chinenye O (2025). Endocrinology Clinical Skills Assessment Program: A Structured Clinical Assessment Program to Enhance Clinical Reasoning in Subspecialty Training. Cureus. DOI: https://doi.org/10.7759/cureus.93834

[22] Myers J H, Dorsey J K (1994). Using diagnostic reasoning (DxR) to teach and evaluate clinical reasoning skills. Academic Medicine. DOI: https://doi.org/10.1097/00001888-199405000-00052

[23] Jackson Jennifer M., Skelton Joseph A., Peters Timothy R. (2020). Medical Students' Clinical Reasoning During a Simulated Viral Pandemic: Evidence of Cognitive Integration and Insights on Novices' Approach to Diagnostic Reasoning. Medical Science Educator. DOI: https://doi.org/10.1007/s40670-020-00946-9

[24] Ahmed Sanaa, Shah Hina, Raza Syeda Zarreen (2023). Evaluation of development of clinical reasoning skills in dental students through diagnostic thinking inventory. Journal of the Pakistan Medical Association. DOI: https://doi.org/10.47391/jpma.8010

[25] Neill Charles, Vinten Claire, Maddison Jill (2020). Use of Inductive, Problem-Based Clinical Reasoning Enhances Diagnostic Accuracy in Final-Year Veterinary Students. Journal of Veterinary Medical Education. DOI: https://doi.org/10.3138/jvme.0818-097r1

[26] Wang Ya, Havish Seggoju Raja, Paschke Adrian (2026). Empirical Analysis of Chain-of-Thought and Solver-Augmented Large Language Models for Deductive Reasoning. Neurosymbolic Artificial Intelligence. DOI: https://doi.org/10.1177/29498732261469484

[27] Hamm Robert M., Beasley William Howard, Johnson William Jay (2014). A Balance Beam Aid for Instruction in Clinical Diagnostic Reasoning. Medical Decision Making. DOI: https://doi.org/10.1177/0272989x14529623

[28] Mazaheri, P. (2026). REPOT: Recoverable Program-of-Thought via Checkpoint Repair. *arXiv preprint arXiv:2605.30052*.

[29] Ramani Divya, Soh Michael, Merkebu Jerusalem, Durning Steven J., Battista Alexis, McBee Elexis, Ratcliffe Temple, Konopasky Abigail (2020). Examining the patterns of uncertainty across clinical reasoning tasks: effects of contextual factors on the clinical reasoning process. Diagnosis. DOI: https://doi.org/10.1515/dx-2020-0019

[30] Norman Geoffrey R, Eva Kevin W (2009). Diagnostic error and clinical reasoning. Medical Education. DOI: https://doi.org/10.1111/j.1365-2923.2009.03507.x

[31] Stojan Jennifer N., Daniel Michelle, Morgan Helen K., Whitman Laurie, Gruppen Larry D. (2017). A Randomized Cohort Study of Diagnostic and Therapeutic Thresholds in Medical Student Clinical Reasoning. Academic Medicine. DOI: https://doi.org/10.1097/acm.0000000000001909

[32] Penny Steven M., Zachariason Anna (2015). The Sonographic Reasoning Method. Journal of Diagnostic Medical Sonography. DOI: https://doi.org/10.1177/8756479314567307

[33] Koitz-Hristov Roxane, Wotawa Franz (2018). Applying algorithm selection to abductive diagnostic reasoning. Applied Intelligence. DOI: https://doi.org/10.1007/s10489-018-1171-9

[34] Staal Justine, Waechter Jason, Allen Jon, Lee Chel Hee, Zwaan Laura (2023). Deliberate practice of diagnostic clinical reasoning reveals low performance and improvement of diagnostic justification in pre-clerkship students. BMC Medical Education. DOI: https://doi.org/10.1186/s12909-023-04541-5

[35] Habibur Rahman Mohammad (2019). Diagnostic Reasoning and Physiotherapy Treatment for A 65 Years Old Female with Left Sided Frozen Shoulder. International Journal of Medical Science and Clinical invention. DOI: https://doi.org/10.18535/ijmsci/v6i5.10

[36] McGervey Megan, Olson Andrew, Mohmand Mohammad, Restrepo Daniel (2025). Clinical Reasoning and Diagnostic Errors. Medical Clinics of North America. DOI: https://doi.org/10.1016/j.mcna.2025.02.014

[37] Mohamed Nabag Wisal Omer, Sahal Elmardi Abdelmoniem (2021). Clinical reasoning skills of medical students, Faculty of Medicine Alzaiem Alazhari University Khartoum Sudan Measured by Diagnostic Thinking Inventory (DTI). British Journal of Medical and Health Research. DOI: https://doi.org/10.46624/bjmhr.2021.v8.i1.003

[38] Mahapatra Ashoka (2015). Study of Biofilm in Bacteria from Water Pipelines. JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH. DOI: https://doi.org/10.7860/jcdr/2015/12415.5715

[39] Miller Leslie, Limbaugh Catherine M. (2008). Applying Evidence to Develop a Medical Oncology Fall-Prevention Program. Clinical Journal of Oncology Nursing. DOI: https://doi.org/10.1188/08.cjon.158-160

[40] Wong Hang Sheung, Wong Tsz Kwan (2026). Multi-Evidence Clinical Reasoning With Retrieval-Augmented Generation for Emergency Triage: Retrospective Evaluation Study. JMIR Medical Informatics. DOI: https://doi.org/10.2196/82026

[41] Omran Abdullah, Hassan Said, Hassan Wesam M. (2026). Artificial Reasoning and Islamic Law: Mapping LLM Capabilities Across Juristic Ranks and Reasoning Frameworks. Journal of Islamic Thought and Civilization. DOI: https://doi.org/10.32350/jitc.161.12

[42] Kakisaka Yosuke, Kakisaka Yosuke, Fujikawa Mayu, Gaillard Schuyler (2016). Writing Case Reports: Teaching and Tuition Techniques, and the Improvement of Clinical Diagnostic Reasoning. International Journal of Medical Students. DOI: https://doi.org/10.5195/ijms.2016.157

[43] (2019). Comparison of Diagnostic Performance of Medical Monitor and Medical Augmented Reality Glasses in Endoscopy: Observation Study. Case Medical Research. DOI: https://doi.org/10.31525/ct1-nct04083573

[44] Heiberg Engel Peter Johan (2008). Tacit knowledge and visual expertise in medical diagnostic reasoning: Implications for medical education. Medical Teacher. DOI: https://doi.org/10.1080/01421590802144260

[45] Bordage Georges, Eva Kevin (2016). Functional neuroimaging and diagnostic reasoning. Medical Teacher. DOI: https://doi.org/10.3109/0142159x.2015.1132836

[46] Moosapour Hamideh, Raza Mohsin, Rambod Mehdi, Soltani Akbar (2011). Conceptualization of category-oriented likelihood ratio: a useful tool for clinical diagnostic reasoning. BMC Medical Education. DOI: https://doi.org/10.1186/1472-6920-11-94

[47] Amjad Ali (2008). Clinical diagnostic reasoning and the curriculum: a medical student's perspective. Medical Teacher. DOI: https://doi.org/10.1080/01421590802064872

[48] Klimpl David, Staudinger Stacey (2025). Rooted in reasoning: a clinical reasoning curriculum using diagnostic RCAs. Diagnosis. DOI: https://doi.org/10.1515/dx-2025-0089

[49] Wang Hua-qiong, Zhou Tian-shu, Tian Li-li, Qian Yang-ming, Li Jing-song (2014). Creating hospital-specific customized clinical pathways by applying semantic reasoning to clinical data. Journal of Biomedical Informatics. DOI: https://doi.org/10.1016/j.jbi.2014.07.017

[50] Dunscomb Jennifer (2008). An Innovative Approach to Enhance Diagnostic Reasoning at the Bedside. Clinical Nurse Specialist. DOI: https://doi.org/10.1097/01.nur.0000311774.11166.af

[51] Jones Roger H (2022). Diagnostic reasoning: maximise trainees' clinical experience. BMJ. DOI: https://doi.org/10.1136/bmj.o602

[52] Bansal Aarti, Singh Davinder, Thompson Joanne, Kumra Alexander, Jackson Benjamin (2020). <p>Developing Medical Students' Broad Clinical Diagnostic Reasoning Through GP-Facilitated Teaching in Hospital Placements</p>. Advances in Medical Education and Practice. DOI: https://doi.org/10.2147/amep.s243538

[53] Fernando Irosh, Henskens Frans A. (2013). ST Algorithm for Medical Diagnostic Reasoning. Polibits. DOI: https://doi.org/10.17562/pb-48-3

[54] Pekcan Neşe, Coşkun Özlem (2026). Diagnostic Errors in Clinical Reasoning: A Comprehensive Literature Review on Cognitive Processes, Causes, and Error Reduction Strategies. Gazi Medical Journal. DOI: https://doi.org/10.12996/gmj.2026.4530