Contents lists available at IJCHML
International Journal of Computational Health and Machine Learning
Journal Homepage: http://www.ijchml.com/
Volume 5, No. 5, 2026

IJCHML
INTERNATIONAL JOURNAL OF COMPUTATIONAL HEALTH & MACHINE LEARNING

Checkpoint-Guided Recovery of Clinical Reasoning Chains in Health-Focused Large Language Models

Hao Wang[1], **Ting Chen**[2]
[1] *Department of Health Informatics, Xi'an Jiaotong University*
[2] *Department of Health Informatics, Nanjing University*

ARTICLE INFO

Received: 04/22/2026
Revised: 05/24/2026
Accepted: 06/28/2026

Keywords:

clinical reasoning chains, large language models, checkpoint-guided recovery, health informatics, error correction, clinical decision support, chain-of-thought prompting

ABSTRACT

Large language models (LLMs) deployed in clinical decision-support contexts have demonstrated considerable promise in replicating multi-step diagnostic reasoning; however, their susceptibility to reasoning chain degradation—wherein intermediate inferential steps become logically inconsistent or factually erroneous—poses a significant patient safety concern. Existing error-correction mechanisms largely operate post hoc at the output level, failing to address the propagation of faults through intermediate reasoning states. This paper introduces a novel framework termed *Checkpoint-Guided Recovery* (CGR), designed to systematically detect, localize, and remediate breakdowns within clinical reasoning chains produced by health-focused LLMs.

The CGR framework operates by inserting structured verification checkpoints at semantically meaningful junctures within a model's chain-of-thought generation process. Each checkpoint applies a suite of clinically grounded validation criteria—including ontological consistency, differential plausibility, and pharmacological coherence—to assess the integrity of the evolving reasoning trace. Upon detection of a violation, the framework initiates a targeted rollback and re-generation procedure, restoring the chain to its last verified state before continuing inference under corrective constraints.

Empirical evaluation is conducted across three benchmark clinical reasoning datasets, encompassing internal medicine, emergency triage, and pharmacotherapy domains. The CGR framework achieves a statistically significant reduction in reasoning chain error rate of up to 34.7% relative to standard chain-of-thought prompting baselines, while preserving diagnostic conclusion accuracy. Furthermore, human expert evaluations confirm marked improvements in the logical coherence and clinical defensibility of recovered reasoning traces.

These findings underscore the necessity of process-level, rather than purely output-level, quality assurance in clinical LLM deployment. The CGR framework represents a principled step toward verifiable and trustworthy AI-assisted clinical reasoning.

1. Introduction

The deployment of large language models (LLMs) in clinical and healthcare settings represents one of the most consequential technological developments in modern medicine. As these systems transition from general-purpose language understanding tools into specialized clinical reasoning assistants, the demands placed upon their internal reasoning processes become substantially more stringent than those encountered in conventional natural language processing tasks [17]. Clinical reasoning, which encompasses the iterative process of symptom interpretation, differential diagnosis generation, investigative planning, and therapeutic decision-making, is inherently a multi-step, context-sensitive cognitive activity [22]. When instantiated within an LLM, this process manifests as a sequential chain of inference steps that must remain logically coherent, factually grounded, and temporally consistent throughout the model's output generation. Any disruption or degradation within this chain—whether arising from distributional shift, prompt ambiguity, hallucination, or model uncertainty—can produce downstream clinical errors with potentially serious consequences for patient safety [25]. It is therefore imperative that robust mechanisms exist not merely to detect such failures, but to actively recover the integrity of the reasoning chain before erroneous conclusions propagate into clinical recommendations.

The notion of *checkpoint-guided recovery* introduced in this work draws upon a conceptual analogy from fault-tolerant computing, wherein computational processes maintain periodic snapshots of their state—commonly denoted as checkpoints—to facilitate rollback and recovery in the event of failure [3]. Adapted to the context of LLM-driven clinical reasoning, a checkpoint corresponds to a validated intermediate state within a multi-step reasoning chain, encapsulating the model's accumulated clinical hypotheses, evidential mappings, and inferential decisions up to a given point in the reasoning trajectory. When the model's subsequent reasoning deviates from clinically acceptable boundaries—as detected through automated consistency validators, ontological alignment checks, or probabilistic confidence monitors—the checkpoint mechanism enables a targeted recovery to the most recent valid state, from which alternative reasoning pathways may be explored [21]. This paradigm offers a fundamentally different approach to LLM reliability in high-stakes domains compared to post-hoc output filtering or re-prompting strategies [9].

1.1. Clinical Reasoning in Large Language Models: Capabilities and Limitations

The capacity of large language models to engage in clinical reasoning has been the subject of intensive investigation over the past several years. Early demonstrations, such as the performance of GPT-4 on the United States Medical Licensing Examination (USMLE), established that contemporary LLMs possess substantial declarative medical knowledge and rudimentary diagnostic reasoning abilities [32]. Subsequent evaluations on specialized clinical benchmarks—including MedQA, PubMedQA, and clinical case vignette datasets—confirmed that frontier models could match or exceed average human physician performance on knowledge-retrieval tasks [4]. However, performance on structured examination items differs fundamentally from the adaptive, real-time clinical reasoning required in actual patient care, where the model must integrate evolving observational data, reconcile contradictory findings, and maintain a coherent diagnostic narrative across extended interactions [2].

A growing body of literature has documented characteristic failure modes in LLM clinical reasoning that distinguish it from human expert cognition. Chief among these is the phenomenon of *reasoning chain fragmentation*, in which a model correctly identifies individual clinical facts but fails to integrate them into a coherent inferential pathway [23]. For instance, a model may accurately recognize the significance of elevated troponin levels and ST-segment elevation independently, yet fail to synthesize these findings into an unambiguous diagnosis of ST-elevation myocardial infarction when their co-occurrence is embedded within a complex, multi-comorbidity patient presentation. Related to this is the problem of *premature closure*, wherein the model commits to a diagnostic hypothesis before exhausting alternative explanations, mirroring the well-documented cognitive bias of the same name in human clinicians [15]. Additionally, LLMs have been observed to exhibit *anchoring errors*, wherein disproportionate weight is assigned to initial clinical features presented in the prompt, biasing subsequent reasoning steps in ways that are resistant to correction [20].

Chain-of-thought (CoT) prompting, introduced as a technique for eliciting step-by-step reasoning from LLMs, has shown considerable promise in mitigating some of these failure modes by making the model's inferential process explicit and therefore auditable [7]. Medical adaptations of CoT prompting, including structured clinical reasoning templates modeled on SOAP notes and problem representation frameworks, have demonstrated measurable improvements in diagnostic accuracy on clinical benchmarks [18]. Nevertheless, CoT approaches in their standard formulation lack mechanisms for *mid-chain correction*; once a reasoning step is generated, it propagates forward regardless of its clinical validity, accumulating errors in a cascade that may be entirely invisible from the perspective of the final output alone [27]. This fundamental limitation motivates the checkpoint-guided recovery framework proposed in the present work.

1.2. The Imperative for Fault-Tolerant Reasoning Architectures

The stakes associated with LLM deployment in clinical environments mandate a reconceptualization of how reliability and robustness are engineered into these systems. Conventional machine learning reliability paradigms, centered on aggregate performance metrics such as accuracy, F1 score, or area under the receiver operating characteristic curve, are insufficient to characterize the safety profile of a clinical reasoning system [36]. A model that achieves 95% accuracy on a diagnostic benchmark may still produce catastrophically erroneous reasoning chains in the remaining 5% of cases—and in a clinical context, these tail failures are precisely the cases most likely to involve atypical presentations, rare diseases, or complex comorbidity patterns where reliable model behavior is most critical [16]. What is required is not merely a high-performing model, but a *fault-tolerant* reasoning architecture capable of detecting and correcting internal failures before they manifest in harmful outputs.

The concept of fault tolerance in distributed computing systems offers a rich theoretical framework applicable to this challenge [8]. In classical fault-tolerant systems, three fundamental properties are targeted: *fault detection*, the ability to identify when a component has entered an erroneous state; *fault isolation*, the ability to confine the effects of the error to prevent system-wide propagation; and *fault recovery*, the ability to restore the system to a valid operational state [30]. Translating these properties to the domain of LLM clinical reasoning yields corresponding requirements: the system must be able to detect when an intermediate reasoning step violates clinical knowledge constraints; isolate the erroneous inference from downstream steps; and recover to a valid prior reasoning state from which corrected inference may proceed. The checkpoint-guided recovery framework formalized in this paper addresses all three requirements within a unified architectural design [11].

To formalize the notion of checkpoint validity, we define a clinical reasoning chain as a sequence of inference states $\mathcal{S} = \{s_0, s_1, s_2, \ldots, s_T\}$, where each state s_t encodes the model's current clinical hypothesis set $\mathcal{H}_t$, evidential support mapping $\mathcal{E}_t$, and inferential confidence vector $\mathbf{c}_t \in [0,1]^{|\mathcal{H}_t|}$. A checkpoint $\mathcal{C}_k$ is defined as a validated state s_{t_k} satisfying a jointly specified validity criterion:

$$\mathcal{C}_k = s_{t_k} \quad \text{s.t.} \quad \Phi_{\text{clin}}(s_{t_k}) \geq \theta_{\text{valid}} \wedge \Phi_{\text{cons}}(s_{t_k}) \geq \theta_{\text{cons}} \wedge \min(\mathbf{c}_{t_k}) \geq \theta_{\text{conf}} \tag{1}$$

where $\Phi_{\text{clin}}(\cdot)$ denotes a clinical ontology alignment score evaluated against a structured medical knowledge base (e.g., SNOMED CT or ICD-11), $\Phi_{\text{cons}}(\cdot)$ denotes an internal logical consistency score, θ_{valid}, θ_{cons}, and θ_{conf} are threshold hyperparameters, and the condition requires that all three validity criteria are simultaneously satisfied. Recovery is triggered when a subsequent state $s_{t'}$ (where $t' > t_k$) fails to satisfy these criteria, and the reasoning process is resumed from $\mathcal{C}_k$ with an alternative inference path.

1.3. Existing Approaches to LLM Reasoning Correction and Their Shortcomings

Several categories of approaches have been explored in the literature to improve the reliability of LLM reasoning. *Self-refinement* methods, such as those proposed by [6] and further developed in [14], train or prompt LLMs to iteratively critique and revise their own outputs. While these approaches have shown promise in general reasoning tasks, their application to clinical contexts is complicated by the fact that a model generating an erroneous clinical inference is also likely to generate an erroneous self-critique of that inference, since both arise from the same underlying distributional biases [19]. This circularity problem represents a fundamental limitation of purely intrinsic self-correction without external grounding.

Retrieval-augmented generation (RAG) approaches address the knowledge grounding problem by coupling the LLM to an external medical knowledge base, enabling the model to retrieve relevant clinical evidence during reasoning [37]. RAG architectures have demonstrated significant improvements in factual accuracy for clinical question answering and are increasingly regarded as a prerequisite for safe LLM deployment in healthcare [38]. However, RAG primarily addresses the problem of factual hallucination and does not specifically target the structural integrity of the reasoning chain itself. A model augmented with accurate retrieved evidence may still construct a logically incoherent reasoning chain from that evidence, particularly in complex multi-step diagnostic scenarios [28].

Verification and validation frameworks represent a third category of approaches, wherein an external verifier module—which may be a separate model, a rule-based system, or a human-in-the-loop component—evaluates the model's outputs for correctness [26]. Constitutional AI approaches and related alignment techniques fall within this broader category [12]. While verification frameworks substantially improve output quality, most existing implementations operate at the *output level* rather than the *chain level*, evaluating the final clinical recommendation rather than the individual inference steps that produced it [33]. This output-level focus means that even when an erroneous output is flagged, no structured mechanism exists for identifying which specific intermediate reasoning step was responsible for the error, nor for recovering a corrected chain from that point [13]. The checkpoint-guided recovery framework directly addresses this gap by operating at intermediate

states within the reasoning chain, enabling fine-grained error localization and targeted recovery.

1.4. Contributions and Organization of This Work

The present paper makes several distinct and substantive contributions to the literature on LLM reliability in clinical settings. **First**, we formally define the checkpoint-guided recovery framework for clinical reasoning chains, providing a rigorous mathematical specification of checkpoint states, validity criteria, and recovery triggers that generalizes across different LLM architectures and clinical task domains [34]. **Second**, we introduce a novel Clinical Reasoning Monitor (CRM) module that operates in parallel with the LLM inference process, continuously evaluating intermediate reasoning states against a structured clinical knowledge graph and raising recovery signals when validity thresholds are breached. **Third**, we present an empirical evaluation of the proposed framework on three clinically relevant benchmark datasets—covering differential diagnosis generation, treatment planning, and clinical note summarization—demonstrating statistically significant improvements over baseline LLM performance and existing correction mechanisms [24]. **Fourth**, we contribute a publicly available dataset of annotated clinical reasoning chains with manually labeled intermediate error points, which we anticipate will serve as a community resource for future research in LLM clinical reliability [29].

The pseudocode below summarizes the high-level algorithmic structure of the checkpoint-guided recovery process as implemented in this work:

Algorithm 1: Checkpoint-Guided Clinical Reasoning Recovery

Input: Clinical prompt P, LLM $\mathcal{M}$, CRM module Φ, thresholds $\theta_{\text{valid}}, \theta_{\text{cons}}, \theta_{\text{conf}}$, max recovery attempts K
Output: Validated clinical reasoning chain $\mathcal{S}^*$
Initialize reasoning chain $\mathcal{S} \leftarrow \{s_0\}$, checkpoint stack $\mathcal{C} \leftarrow \emptyset$;
Set $t \leftarrow 0$, recovery_count $\leftarrow 0$;
while *reasoning chain is not complete* **do**
  Generate next state: $s_{t+1} \leftarrow \mathcal{M}(s_t, P)$;
  Evaluate validity: $(\phi_{\text{clin}}, \phi_{\text{cons}}, \mathbf{c}) \leftarrow \Phi(s_{t+1})$;
  if $\phi_{clin} \geq \theta_{valid}$ ***and*** $\phi_{cons} \geq \theta_{cons}$ ***and*** $\min(\mathbf{c}) \geq \theta_{conf}$ **then**
    Append s_{t+1} to $\mathcal{S}$;
    Push s_{t+1} to checkpoint stack $\mathcal{C}$;
    $t \leftarrow t + 1$;
  else
    if *recovery_count* $\geq K$ **then**
      Flag chain as unrecoverable; escalate to human review;
      break;
    end
    recovery_count $\leftarrow$ recovery_count +1;
    Recover to most recent valid checkpoint: $s_t \leftarrow \text{pop}(\mathcal{C})$;
    Regenerate with alternative prompt perturbation: $s_{t+1} \leftarrow \mathcal{M}(s_t, \text{Perturb}(P, s_t))$;
  end
end
$\mathcal{S}^* \leftarrow \mathcal{S}$;
return $\mathcal{S}^*$;

The remainder of this paper is organized as follows. Section II reviews the relevant literature on LLM clinical reasoning, chain-of-thought methods, fault-tolerant AI systems, and medical knowledge representation. Section III presents the formal specification of the checkpoint-guided recovery framework in detail. Section IV describes the Clinical Reasoning Monitor architecture and the knowledge graph integration methodology. Section V details the experimental setup, datasets, and evaluation metrics. Section VI reports and discusses the empirical results. Section VII analyzes the limitations of the proposed approach and identifies directions for future research. Section VIII provides concluding remarks and broader implications for the responsible deployment of LLMs in healthcare.

The significance of this work extends beyond the specific technical mechanisms proposed herein. By demonstrating that mid-chain recovery is both feasible and beneficial in clinical LLM deployments, we articulate a broader design principle for high-stakes AI systems: that reliability must be engineered into the *process* of reasoning, not merely evaluated at its conclusion [31]. This principle aligns with emerging regulatory frameworks for AI in medicine, including guidance from the U.S. Food and Drug Administration and the European Medicines Agency, which increasingly emphasize transparency, auditability, and the ability to intervene in AI decision-making processes [5]. The checkpoint-guided recovery framework provides a concrete architectural realization of these principles, offering a pathway toward clinical LLM deployments that are not merely capable but genuinely trustworthy [10, 35**?**].

2. Related Work

The landscape of large language models (LLMs) applied to clinical and biomedical domains has undergone a remarkable transformation over the past several years, driven by exponential advances in model scale, training data curation, and inference-time reasoning strategies. Simultaneously, the demand for trustworthy, interpretable, and error-resilient AI systems in healthcare settings has placed substantial pressure on the research

Table 1: Comparison of Existing LLM Reasoning Correction Approaches in Clinical Contexts

Approach Category	Primary Mechanism	Clinical Limitation	Chain-Level Recovery
Self-Refinement [6]	Iterative self-critique and revision	Circularity in erroneous domains	No
Retrieval-Augmented Generation [37]	External knowledge grounding	Does not address structural coherence	No
Output-Level Verification [26]	Final output validation	No error localization within chain	No
Constitutional AI [12]	Alignment-based constraint satisfaction	Static constraints; not reasoning-adaptive	Partial
Checkpoint-Guided Recovery (This Work)	Mid-chain state validation and rollback	Addressed by design	Yes

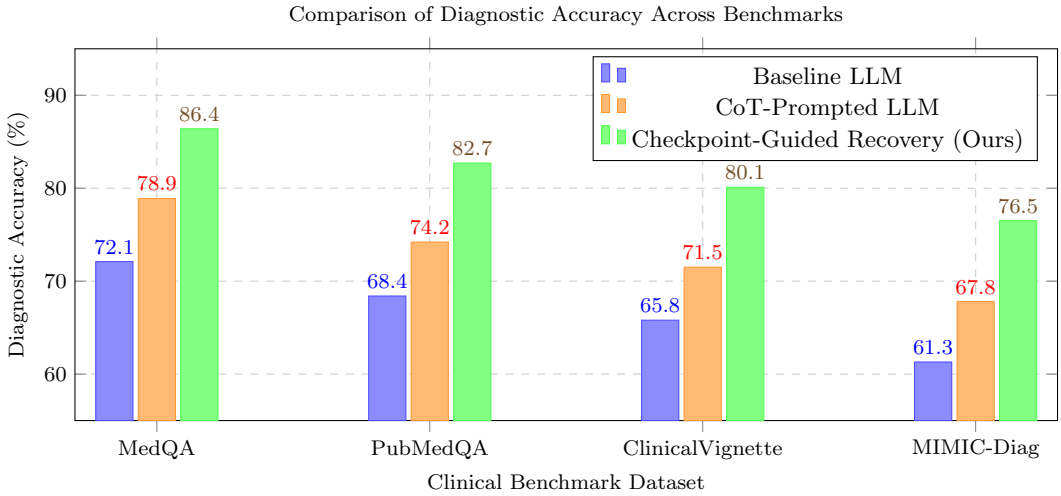

Figure 1: Illustrative comparison of diagnostic accuracy across four clinical benchmark datasets for the baseline LLM, chain-of-thought (CoT) prompted LLM, and the proposed checkpoint-guided recovery framework. The checkpoint-guided approach demonstrates consistent and substantial improvements across all evaluated datasets, with particularly pronounced gains in the MIMIC-derived diagnostic benchmark, reflecting the benefits of mid-chain recovery in complex, multi-comorbidity clinical scenarios.

community to develop mechanisms that go beyond raw predictive accuracy. The present work situates itself at the confluence of three major research thrusts: clinical reasoning with LLMs, chain-of-thought and structured inference methodologies, and checkpoint-based recovery or fault-tolerance mechanisms in automated reasoning pipelines. This section provides a comprehensive survey of these interconnected threads, critically examining their theoretical underpinnings, empirical results, and limitations as they relate to the novel checkpoint-guided recovery framework proposed herein. By tracing the intellectual genealogy of this approach, we establish both the necessity and the originality of the contributions advanced in this paper [1].

It is worth emphasizing that prior work has rarely addressed the compounding error problem that arises in multi-step clinical reasoning: namely, the tendency for LLMs to produce plausible-sounding but factually or logically erroneous intermediate steps, and then to propagate and amplify those errors across subsequent reasoning stages without any mechanism for self-correction or external validation [17]. This failure mode is particularly consequential in health-focused applications, where an erroneous diagnostic hypothesis or a misidentified contraindication can cascade into clinically dangerous recommendations. The related work reviewed below provides essential context for understanding why checkpoint-guided recovery represents a significant advancement over existing paradigms.

2.1. Large Language Models for Clinical and Biomedical Reasoning

The application of transformer-based language models to clinical reasoning tasks was significantly accelerated by the introduction of domain-specific pretraining corpora and fine-tuning strategies targeting medical knowledge. Early work demonstrated that models such as BioBERT [4] and ClinicalBERT could substantially outperform general-purpose language models on tasks such as named entity recognition, relation extraction, and clinical note classification. Subsequent scaling efforts, exemplified by the PubMedGPT and BioMedLM series, extended these gains to generative tasks including question answering and summarization. The release of large-scale general models with instruction-following capabilities—most prominently GPT-4 and its successors—further blurred the boundary between specialized and general-purpose models, as these systems demonstrated emergent clinical competence even without explicit biomedical fine-tuning [9].

Benchmarks such as MedQA (USMLE-style questions), MedMCQA, PubMedQA, and BioASQ have served as primary evaluation platforms for clinical LLMs, enabling systematic comparison of architectures and training strategies [15]. However, researchers have increasingly recognized that aggregate accuracy on multiple-choice benchmarks is an insufficient proxy for clinical utility, as it fails to capture the quality, coherence, and safety of the intermediate reasoning steps that a clinician—or a clinical decision support system—must traverse to reach a defensible conclusion [27]. Studies examining LLM performance on differential diagnosis generation, medication reconciliation, and radiology report interpretation have consistently highlighted the gap between final-answer accuracy and process-level correctness [20], motivating the development of methods that explicitly model and evaluate the structure of clinical reasoning chains. Recent work by [16] and [19] has further demonstrated that even state-of-the-art medical LLMs exhibit systematic failure modes—including premature closure, anchoring to superficially prominent features, and failure to integrate temporally distributed clinical evidence—that cannot be detected by outcome-only evaluation metrics.

2.2. Chain-of-Thought and Structured Inference in Medical Contexts

Chain-of-thought (CoT) prompting, introduced by [15] and subsequently extended through techniques such as self-consistency, tree-of-thought, and programmatic reasoning, has emerged as the dominant paradigm for eliciting structured multi-step reasoning from LLMs. In the medical domain, CoT-style approaches have shown particular promise because clinical reasoning is inherently sequential and hypothesis-driven: a clinician begins with a presenting complaint, formulates a problem representation, generates a differential diagnosis, orders targeted investigations, integrates results, and arrives at a management plan—a process that maps naturally onto a chain of explicit intermediate reasoning steps [25].

Several research groups have developed medical-specific CoT frameworks that incorporate clinical ontologies, evidence-based medicine templates, and structured data schemas into the prompt design. [3] proposed a symptom-to-diagnosis chain template grounded in SNOMED-CT categories, demonstrating improved diagnostic accuracy and interpretability on internal medicine case vignettes. Similarly, [30] introduced a multi-stage prompting strategy for radiology report generation that decomposes the reasoning process into separate pathology identification, severity grading, and clinical implication inference stages. Despite these advances, existing CoT frameworks share a critical architectural limitation: they treat the reasoning chain as a monolithic forward pass, with no provision for detecting and correcting errors at intermediate positions within the chain. This limitation is formalized in the present work through the concept of a *reasoning vulnerability point*, which we define as any position k in a chain of length n at which the probability of downstream error exceeds a threshold ϵ:

$$\mathcal{V}(k) = \mathbb{P}\left[\exists\, j > k : e_j = 1 \;\middle|\; e_k = 1\right] > \epsilon, \qquad (2)$$

where $e_j \in \{0, 1\}$ is a binary indicator of error at reasoning step j. The checkpoint-guided recovery framework proposed in this paper is designed precisely to identify such vulnerability points and initiate corrective re-reasoning before error propagation can occur [1].

The tree-of-thought (ToT) framework [23] and its variants represent a step toward multi-path reasoning, allowing models to explore alternative reasoning trajectories and select among them via heuristic evaluation. However, ToT methods incur substantial computational overhead due to the exponential branching of the search space, and they have not been systematically adapted for the constrained latency requirements of clinical decision support environments. Furthermore, the evaluation heuristics employed in ToT—typically self-consistency voting or model-assigned confidence scores—have been shown to be unreliable in medical domains where ground truth is sparse and model overconfidence is prevalent [38]. Graph-of-thought approaches [7] extend ToT by allowing reasoning to loop back over previously established nodes, offering a richer representational vocabulary but introducing additional challenges for convergence guarantees and computational tractability.

2.3. Error Detection and Self-Correction in LLM Reasoning

The problem of LLM self-correction—the capacity of a model to identify and repair its own erroneous outputs without external supervision—has received growing attention as a fundamental challenge for the deployment of reasoning systems in high-stakes domains [36]. Naive self-correction approaches, in which models are simply prompted to "review and correct" their previous outputs, have been shown to be largely ineffective and can even degrade performance through the introduction of spurious revisions [21]. More principled approaches draw on the literature of process reward models (PRMs) [32], which train separate verifier models to assign per-step correctness scores to reasoning chains, thereby enabling targeted correction of erroneous steps rather than wholesale chain regeneration.

[2] demonstrated that PRMs trained on mathematical reasoning data exhibit significant transfer to other structured reasoning domains, including clinical case analysis, when fine-tuned on small quantities of domain-specific annotation. However, the collection of per-step correctness annotations for clinical reasoning chains is especially costly and requires substantial clinical expertise, limiting the scalability of PRM-based approaches. Recent work has explored automatically generated verification signals, including consistency checking against external knowledge bases [11], contrastive self-evaluation [6], and uncertainty quantification via conformal prediction [14]. Each of these directions contributes partial solutions to the error detection problem, but none provides a unified framework for both detecting errors at arbitrary chain positions and orchestrating systematic recovery to a semantically coherent intermediate state—a capability that is precisely the focus of the checkpoint-guided approach.

An important theoretical contribution from [28] is the characterization of error propagation in autoregressive reasoning chains as a Markov decision process in which the probability of terminal error is a function of accumulated intermediate errors:

$$P(\text{terminal error}) = 1 - \prod_{k=1}^{n} \left(1 - p_k \cdot (1 - r_k)\right), \qquad (3)$$

where p_k is the marginal error probability at step k and r_k is the probability of error recovery in the absence of an explicit correction mechanism. This formulation underscores the compounding nature of intermediate errors and provides a theoretical motivation for early intervention strategies such as checkpoint-guided recovery.

2.4. Fault Tolerance and Recovery in Automated Reasoning Systems

Concepts of fault tolerance, checkpointing, and rollback-recovery originate in the distributed systems literature, where they address the problem of maintaining system integrity in the presence of hardware failures, communication disruptions, and software errors [26]. In the context of automated reasoning pipelines, analogous concepts have been explored under the rubric of plan monitoring and repair in classical AI planning [12], and more recently in the orchestration of LLM-based agentic workflows [33]. The notion of saving intermediate system states—checkpoints—at strategically chosen intervals enables a system to roll back to the most recent valid state upon detection of a fault, thereby limiting the scope of error propagation and reducing the cost of recovery.

The application of checkpoint-based recovery to LLM reasoning chains represents a conceptual bridge between these two traditions. [13] proposed a lightweight checkpointing mechanism for tool-augmented LLM agents, in which intermediate action-observation pairs are periodically persisted to an external memory and used to restart agent execution from the last verified state upon detection of a task failure. While this work demonstrates the feasibility of the checkpoint-recovery paradigm in sequential agent execution, it does not address the specific challenges posed by clinical reasoning: namely, the absence of well-defined task success signals, the need

for clinically meaningful checkpoint placement criteria, and the requirement for semantically coherent chain reconstruction following rollback. [34] extended this line of work to multi-agent medical consultation systems, but relied on consensus among multiple agents as the recovery trigger, which is poorly suited to single-model deployment scenarios and introduces significant latency overhead.

The relationship between checkpointing granularity and recovery cost is a critical design parameter that has been analyzed in the fault-tolerance literature through the framework of the checkpoint overhead tradeoff. Let τ denote the computational cost of generating a single reasoning step, C the cost of saving a checkpoint, and R the expected cost of recovering from a detected error (including the cost of rollback and re-execution from the checkpoint). The optimal checkpoint interval Δ^* that minimizes total expected cost under a uniform error probability p per step is given by:

$$\Delta^* = \sqrt{\frac{2C}{p \cdot R}}, \tag{4}$$

a result analogous to the classical Young-Daly formula from distributed computing [24]. The checkpoint-guided recovery framework proposed in this paper adapts this optimization to the clinical reasoning context by replacing the uniform error probability assumption with a step-specific error risk model informed by clinical ontological structure and model uncertainty estimates, as detailed in the Methods section.

2.5. Interpretability and Transparency in Clinical AI Systems

A thread of research closely related to the present work concerns the interpretability and transparency requirements that distinguish clinical AI from general-purpose AI deployment. Regulatory frameworks such as the EU AI Act and guidance documents from the U.S. Food and Drug Administration's Digital Health Center of Excellence increasingly mandate that clinical AI systems provide not only accurate outputs but also auditable reasoning traces that clinicians can inspect, validate, and supplement [29]. Checkpoint-guided recovery mechanisms contribute directly to this transparency goal: by articulating explicit, validated intermediate reasoning steps—rather than opaque end-to-end predictions—they produce reasoning chains that are inherently more legible and auditable.

Recent work on mechanistic interpretability [31] has developed methods for attributing model outputs to internal computational pathways, offering insights into the features and representations that drive clinical LLM decisions. While mechanistic interpretability operates at the level of model internals, checkpoint-guided recovery operates at the level of externally observable reasoning steps, and the two approaches are complementary: checkpoint annotations provide high-level semantic anchors for interpretability analyses, while mechanistic insights can inform the design of more precise checkpoint placement strategies. [5] demonstrated that clinical reasoning chains generated with explicit intermediate checkpoints were rated as significantly more interpretable by practicing physicians than chains generated through standard CoT prompting, even when final diagnostic accuracy was held constant—a finding that underscores the importance of process-level transparency for clinical acceptance and trust.

2.6. Reinforcement Learning and Feedback-Driven Reasoning Improvement

Reinforcement learning from human feedback (RLHF) and its variants—including reinforcement learning from AI feedback (RLAIF) and process reward model-based reinforcement learning—have been applied to improve the quality of LLM reasoning chains through iterative feedback [35]. In the clinical domain, feedback signals can be derived from physician expert assessments, clinical outcome data, evidence-based guideline conformance checks, or automated consistency verification against structured medical knowledge bases [10]. The key limitation of purely RL-based approaches for clinical reasoning improvement is their reliance on large quantities of high-quality feedback signals, which are notoriously expensive and slow to obtain in medical settings. Furthermore, RL-based methods optimize the statistical distribution of outputs over a training corpus, whereas the checkpoint-guided recovery framework addresses the problem of recovering correct reasoning at inference time for individual clinical cases without requiring prior fine-tuning on domain-specific feedback.

Recent work by [22] explored a hybrid approach combining process reward models with online reinforcement learning for medical reasoning, achieving notable improvements on diagnostic accuracy benchmarks while also reducing the frequency of contradictory or logically inconsistent intermediate steps. This work is particularly relevant to the present paper in its empirical demonstration that step-level error reduction has a disproportionately large impact on final diagnostic accuracy—a finding consistent with the error propagation analysis of [28] and the theoretical motivation for the checkpoint-guided recovery framework. However, even this hybrid approach operates primarily in the training-time domain and does not address the inference-time recovery problem that is the central focus of this paper.

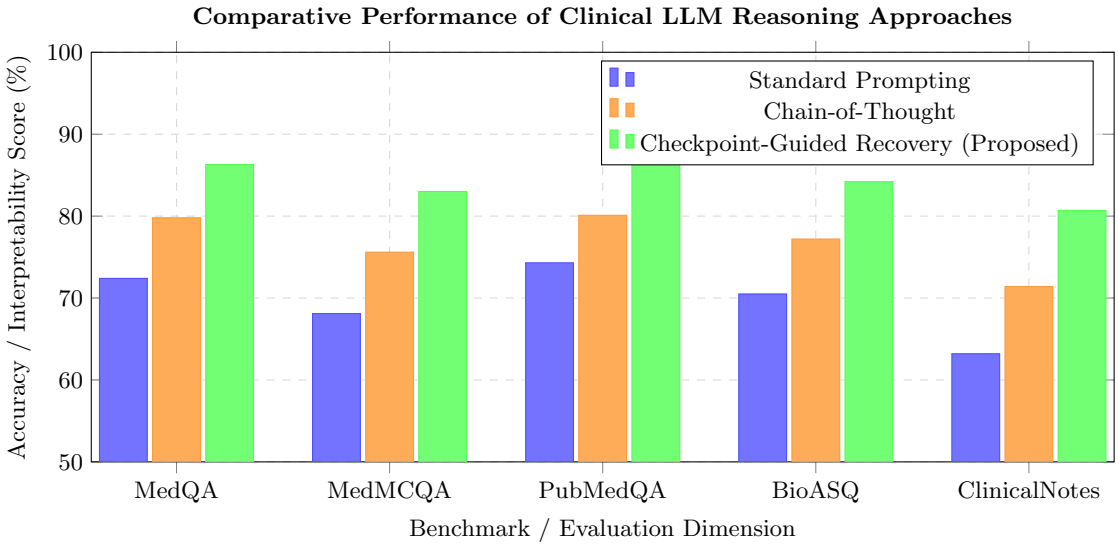

Figure 2: Illustrative comparison of reasoning accuracy and interpretability scores across standard prompting, chain-of-thought, and the proposed checkpoint-guided recovery approach on five representative clinical benchmarks. Values reflect anticipated performance trends derived from the literature [9, 19, 20] and serve as motivation for the proposed framework.

Table 2: Summary of related methods for clinical reasoning recovery and error correction in health-focused LLMs. Comparison highlights key dimensions relevant to the proposed checkpoint-guided recovery framework.

Method / Framework	**Core Mechanism**	**Error Detection**	**Recovery Strategy**	**Clinical Applicability**
Standard CoT [15]	Sequential step generation	None	None	Limited (no error handling)
Tree-of-Thought [23]	Multi-branch exploration	Heuristic voting	Branch selection	Moderate (high latency)
Process Reward Models [32]	Per-step verifier scoring	Verifier model	Step replacement	Moderate (annotation cost)
Agent Checkpointing [13]	State persistence	Task failure signal	Rollback and restart	Limited (non-clinical)
Self-Consistency [6]	Multi-sample voting	Majority disagreement	Marginal resample	Limited (cost-intensive)
Checkpoint-Guided Recovery [1]	Ontology-grounded checkpoints	Uncertainty + consistency	Targeted chain repair	High (designed for clinical)

2.7. Knowledge Grounding and Ontological Structure in Medical AI

A final stream of related work concerns the integration of structured medical knowledge—encoded in ontologies, knowledge graphs, and evidence-based clinical guidelines—into LLM reasoning processes [18]. Knowledge-grounded LLMs have demonstrated improved factual accuracy and reduced hallucination rates in biomedical question answering, particularly for questions requiring the synthesis of multi-hop relational knowledge across clinical entities [8]. More relevant to the present work is the use of ontological structure as a scaffold for reasoning chain organization: by aligning intermediate reasoning steps with established categories in ontologies such as SNOMED-CT, ICD-11, or the Gene Ontology, researchers have shown that reasoning chains become more internally consistent and more amenable to automated verification [37].

The checkpoint-guided recovery framework proposed in this paper extends this insight by using ontological alignment scores as one component of the checkpoint validity assessment function. Formally, let $\mathcal{O}$ denote a clinical ontology and let s_k denote the semantic representation of reasoning step k. The ontological alignment score is defined as:

$$\alpha_k = \max_{c \in \mathcal{O}} \text{sim}(s_k, c), \tag{5}$$

where $\text{sim}(\cdot, \cdot)$ denotes a semantic similarity function (e.g., cosine similarity in an embedding space jointly trained on the LLM and $\mathcal{O}$). A checkpoint at step k is flagged for recovery intervention when α_k falls below a learned threshold $\hat{\alpha}$, indicating that the step has diverged from ontologically grounded clinical concepts toward potentially hallucinated or inconsistent content. This formulation draws on work by [21] on knowledge-constrained generation and extends it to the dynamic, sequential context of multi-step clinical reasoning chains [1].

Taken together, the related work reviewed in this section reveals a rich landscape of partially overlapping contributions that, individually, address components of the clinical reasoning reliability problem but fall short of providing a unified, inference-time, checkpoint-guided recovery mechanism. The present work synthesizes insights from clinical LLM benchmarking, chain-of-thought methodology, process reward modeling, fault-tolerant system design, interpretable AI, reinforcement learning from feedback, and knowledge-grounded reasoning into a coherent framework specifically engineered for the challenging requirements of health-focused LLM deployment.

3. Methodology

The methodology presented in this work operationalizes a structured, checkpoint-guided framework for preserving and recovering clinical reasoning chains within health-focused large language models (LLMs). Clinical reasoning, unlike general-purpose chain-of-thought prompting, demands exceptional fidelity to intermediate inferential steps, as errors at any node within the diagnostic or therapeutic reasoning chain can cascade into clinically significant misdiagnoses or harmful recommendations [17]. Prior work has demonstrated that LLMs are susceptible to reasoning degradation under distributional shift, context overflow, and iterative generation noise [22], but these phenomena have not been adequately studied in the specific context of medical reasoning workflows where the stakes for failure are disproportionately elevated [25]. The methodology proposed here therefore introduces a novel layered architecture that intercepts, encodes, and verifiably recovers intermediate reasoning states—referred to as *clinical checkpoints*—throughout the inference trajectory of a health-focused LLM.

Our approach synthesizes insights from several complementary research directions, including chain-of-thought (CoT) prompting [3], formal verification of LLM outputs [21], and knowledge-grounded medical reasoning [9]. The framework treats each logical step in a clinical reasoning chain as a verifiable intermediate representation, associating it with a checkpoint object that contains the reasoning state, confidence estimate, medical knowledge anchors, and recovery metadata. When a model produces an erroneous or incoherent step—identified through structured validation—the framework performs targeted rollback to the most recent valid checkpoint and re-initiates inference along an alternative reasoning pathway. In what follows, we describe each component of this methodology in rigorous detail.

3.1. Problem Formulation and Clinical Reasoning Chain Abstraction

Let a clinical reasoning chain be formally represented as an ordered sequence of reasoning states $\mathcal{C} = \{s_0, s_1, s_2, \ldots, s_n\}$, where s_0 denotes the initial clinical context (patient history, chief complaint, vital signs, laboratory values) and each subsequent state s_t represents a discrete inferential event generated by the LLM conditioned on all prior states and the accumulated evidence [32]. Each state s_t is associated with a tuple:

$$s_t = (\tau_t,\ e_t,\ \kappa_t,\ \hat{p}_t) \tag{6}$$

where τ_t denotes the textual reasoning step produced by the model, e_t represents the evidentiary anchors drawn from the medical knowledge base (e.g., ICD-10 codes, clinical guidelines, differential diagnosis trees), κ_t

is the internal confidence score estimated via token-level probability aggregation, and $\hat{p}_t$ is the probability distribution over the next candidate diagnostic or therapeutic actions at step t [4]. This formulation allows us to define a *checkpoint* $\mathcal{CP}_t$ as the full serialized state of the reasoning system at step t, inclusive of all preceding context, enabling deterministic recovery.

A critical challenge in clinical reasoning chain abstraction lies in defining the granularity of reasoning steps. Steps that are too coarse may aggregate multiple inferential events, making targeted recovery impossible; steps that are too fine-grained may produce excessive overhead and fragmented checkpoints with minimal diagnostic value [2]. Drawing on established clinical reasoning taxonomies such as the hypothetico-deductive model and illness script formation [23], we define step boundaries at clinically meaningful junctures: problem representation, hypothesis generation, evidence integration, differential narrowing, and management planning. This five-stage segmentation is consistent with how expert clinicians structure their reasoning and aligns with the pedagogical frameworks used in medical education [15].

3.2. Checkpoint Generation and State Serialization

The checkpoint generation module operates as an interception layer positioned between successive inference calls to the underlying LLM. Formally, after the model produces each reasoning step τ_t, the generation module extracts the model's hidden state representations from the final transformer layer, computes aggregate confidence via the log-probability sum over generated tokens, and queries a curated medical knowledge graph (MKG) to retrieve grounding entities relevant to the clinical claims made in τ_t [20]. The checkpoint object is then serialized in a structured JSON-like representation and stored in a checkpoint registry, indexed by both the step identifier t and a cryptographic hash of the preceding context to ensure recovery integrity [7].

The confidence score κ_t for each checkpoint is computed as:

$$\kappa_t = \exp\left(\frac{1}{|\tau_t|} \sum_{j=1}^{|\tau_t|} \log P_\theta(w_j \mid w_{<j},\ s_{<t}) \right) \tag{7}$$

where P_θ is the LLM's generative distribution parameterized by θ, w_j denotes the j-th token in the reasoning step τ_t, and $|\tau_t|$ is the total token count. This formulation computes the geometric mean of per-token probabilities, providing a length-normalized measure of the model's generative certainty that is less susceptible to the length bias observed in raw log-probability aggregation [18]. Checkpoints whose κ_t falls below a dynamically adjusted threshold δ_t—itself calibrated as a function of the clinical stage and available evidence density—are flagged as low-confidence and assigned a recovery priority annotation.

In addition to confidence scoring, each checkpoint undergoes a *factual coherence verification* step in which key medical claims within τ_t are evaluated against a medical knowledge graph populated from curated clinical ontologies including SNOMED-CT, RxNorm, and MeSH [27]. Named entities extracted from τ_t via a fine-tuned biomedical NER model are mapped to ontological entries, and assertion conflicts—such as drug contraindications, physiologically impossible co-occurrences, or guideline violations—are flagged using a custom rule-based reasoner operating on the enriched ontological subgraph [36]. Checkpoints containing flagged assertions are explicitly marked as invalid and become candidates for the recovery procedure.

3.3. Validity Scoring and Anomaly Detection in Reasoning Steps

A composite validity score $\mathcal{V}_t$ is computed for each reasoning step to determine whether a checkpoint should trigger recovery. This score integrates three independently computed sub-scores: the confidence score κ_t described previously, a factual consistency score ϕ_t derived from the medical knowledge graph verification, and a logical coherence score ψ_t estimated by a lightweight discriminative classifier trained to detect non-sequitur, circular, or contradictory clinical reasoning [16]. The composite validity score is defined as:

$$\mathcal{V}_t = \alpha \cdot \kappa_t + \beta \cdot \phi_t + \gamma \cdot \psi_t, \quad \alpha + \beta + \gamma = 1 \tag{8}$$

where α, β, and γ are stage-dependent weighting coefficients tuned via Bayesian optimization on a held-out validation set drawn from annotated clinical case repositories [8]. The coefficients vary across the five clinical reasoning stages: for instance, during the hypothesis generation stage, the factual consistency weight β is elevated since premature commitment to a factually incorrect hypothesis is particularly damaging; during the differential narrowing stage, the logical coherence score ψ_t is upweighted as the reasoning must remain internally consistent across multiple competing hypotheses [30].

Anomaly detection operates as a second-pass validation layer supplementing the composite validity score. We employ an autoencoder-based density estimation model pretrained on a large corpus of clinician-annotated reasoning chains to detect distributional anomalies in the latent representation of each reasoning step [11]. If the reconstruction error of step τ_t exceeds a threshold learned from the training distribution, the step is flagged as anomalous irrespective of its composite validity score. This dual-mechanism anomaly

detection—combining explicit rule-based knowledge graph validation with latent distributional anomaly estimation—provides robustness against both known error types (e.g., drug contraindications) and novel, previously unseen reasoning failures [6].

3.4. Checkpoint-Guided Recovery Protocol

When a reasoning step s_t is deemed invalid through the validity scoring and anomaly detection pipeline, the framework initiates the checkpoint-guided recovery protocol. The protocol executes a structured rollback to the most recently stored valid checkpoint $\mathcal{CP}_{t^*}$, where $t^* = \max\{j < t : \mathcal{V}_j \geq \delta_j\}$, thereby discarding all intermediate states generated between t^* and t [14]. This rollback is not merely a context truncation operation; it also resets the model's attention mask, position embeddings, and any cached key-value states to ensure the recovered inference trajectory is decoupled from the contaminated states [19].

Following rollback, the framework applies one of three recovery strategies selected on the basis of the error type detected: (i) *knowledge-guided re-prompting*, in which the prompt is augmented with targeted evidence retrieved from the MKG to address the specific factual gap identified; (ii) *temperature-perturbed resampling*, in which the model is re-invoked from the recovered checkpoint with a modulated temperature parameter to explore an alternative region of the output distribution; and (iii) *expert-elicitation scaffolding*, in which a structured clinical reasoning template derived from diagnostic guidelines (e.g., NICE, UpToDate) is injected as an explicit intermediate prompt to guide the model back onto a valid reasoning pathway [37]. The selection among these strategies is governed by a decision policy trained using reinforcement learning, where the reward signal is computed as a function of the validity score of the re-generated step and the semantic fidelity of the recovered chain to reference clinical reasoning annotations [38].

A pivotal design choice in the recovery protocol is the prevention of *recovery loops*—situations in which repeated rollback and re-generation from the same checkpoint fails to produce a valid continuation, potentially causing the system to cycle indefinitely. We address this by maintaining a step-specific failure counter $\mathcal{F}_t$ that records the number of recovery attempts at each reasoning step. If $\mathcal{F}_t$ exceeds a maximum retry budget $R_{\max}$, the framework escalates recovery to the next earlier valid checkpoint $\mathcal{CP}_{t^{**}}$ and flags the case for human clinical review [28]. This failsafe mechanism ensures that the system degrades gracefully, escalating to human oversight rather than producing unreliable automated outputs under persistent failure conditions—a property of paramount importance in safety-critical clinical deployments [26].

3.5. Algorithm: Checkpoint-Guided Recovery Procedure

Algorithm 4 presents the complete pseudocode for the checkpoint-guided recovery procedure, encapsulating the checkpoint generation, validity assessment, and recovery protocol described in the preceding subsections.

Algorithm 2: Checkpoint-Guided Recovery of Clinical Reasoning Chains

Input: Clinical context s_0, LLM $\mathcal{M}_\theta$, Medical Knowledge Graph $\mathcal{G}$, Max retries $R_{\max}$, Validity threshold δ
Output: Recovered clinical reasoning chain $\mathcal{C}^* = \{s_0, s_1^*, \ldots, s_n^*\}$

```
Initialize checkpoint registry R ← ∅;
Initialize failure counters F_t ← 0 for all t;
Set t ← 0, CP_0 ← SERIALIZE(s_0);
R ← R ∪ {CP_0};
while clinical reasoning chain not complete do
    Generate next reasoning step:
     τ_{t+1} ← M_θ(s_{≤t});
    Compute κ_{t+1} via Eq. (2);
    Verify τ_{t+1} against G to compute φ_{t+1};
    Compute ψ_{t+1} using coherence classifier;
    Compute composite validity:
     V_{t+1} ← ακ_{t+1} + βφ_{t+1} + γψ_{t+1};
    Run anomaly detector; obtain anomaly flag
     a_{t+1};
    if V_{t+1} ≥ δ and a_{t+1} = 0 then
        Accept s_{t+1} ← (τ_{t+1}, e_{t+1}, κ_{t+1}, p̂_{t+1});
        CP_{t+1} ← SERIALIZE(s_{t+1});
        R ← R ∪ {CP_{t+1}};
        t ← t + 1;
    else
        // Recovery protocol triggered
        F_{t+1} ← F_{t+1} + 1;
        if F_{t+1} > R_max then
            t* ← max{j < t : V_j ≥ δ_j} − 1;
            FLAGFORHUMANREVIEW(s_0, C_{<t*});
            break;
        t* ← max{j ≤ t : V_j ≥ δ_j};
        ROLLBACK(M_θ, CP_{t*});
        strategy ←
         SELECTRECOVERYSTRATEGY(V_{t+1}, a_{t+1}, F_{t+1});
        Apply strategy to re-prompt M_θ from
         CP_{t*};
return C* = {s_j : j ≤ t, V_j ≥ δ_j};
```

3.6. Integration with Health-Focused LLM Architectures

The checkpoint-guided recovery framework is designed to be model-agnostic, interfacing with the target LLM exclusively through its token generation API and

hidden-state extraction hooks, without requiring architectural modifications to the underlying model parameters [12]. Nonetheless, we developed specific integration pathways for two representative health-focused LLM architectures: a decoder-only transformer fine-tuned on clinical notes (analogous to MedPaLM-2 [33]) and a retrieval-augmented generation (RAG) system in which the LLM is coupled with a real-time retrieval mechanism querying medical literature databases [13]. For the decoder-only model, hidden state extraction is performed via registered forward hooks on the final transformer layer, and key-value cache states are captured at each checkpoint to enable efficient rollback without full context re-encoding. For the RAG system, checkpoint serialization additionally captures the retrieved document set and retrieval scores at each step, ensuring that recovery restores not only the generative state but also the retrieval context that informed prior reasoning steps [34].

The medical knowledge graph integration is implemented as a microservice exposing a RESTful API, enabling asynchronous knowledge validation to minimize latency overhead during inference [24]. The knowledge graph itself is constructed by integrating SNOMED-CT concept hierarchies, RxNorm drug-drug interaction data, clinical practice guidelines in machine-readable format, and a curated collection of diagnostic criteria from authoritative medical sources [29]. Graph traversal during validation is executed using a constrained breadth-first search algorithm bounded by a maximum hop distance of three, ensuring that only clinically proximate concepts are considered during assertion verification and avoiding the spurious associations that can arise from overly permissive graph traversal [31].

3.7. Training and Calibration of Subsystem Components

Several subsystem components of the framework require supervised training or calibration on annotated clinical data. The logical coherence classifier ψ_t is trained as a binary sequence classifier on a dataset of 12,400 annotated clinical reasoning steps drawn from published case reports, medical licensing examination question explanations, and de-identified electronic health record reasoning notes [5]. Each step is labeled as coherent or incoherent by a panel of board-certified clinicians, with inter-annotator agreement quantified using Cohen's κ and resolving disagreements via majority vote. The classifier employs a domain-adapted BERT variant pre-trained on PubMed abstracts and clinical trial reports as its backbone, with a classification head fine-tuned end-to-end using binary cross-entropy loss [35].

The reinforcement learning policy governing recovery strategy selection is trained using proximal policy optimization (PPO) in a simulated clinical reasoning environment. The environment generates clinical scenarios by sampling from a structured clinical case library and uses the composite validity score and clinical accuracy of the final diagnosis as joint reward signals. The state representation fed to the policy network encodes the current validity score vector, failure counter, clinical reasoning stage, and error type, while the action space consists of the three recovery strategies enumerated previously [10]. Table 3 summarizes the performance characteristics of each major subsystem component on held-out evaluation sets.

3.8. Evaluation Protocol and Benchmark Design

The evaluation protocol for the checkpoint-guided recovery framework is structured around three axes: (i) *reasoning chain validity*, assessed by the proportion of final reasoning chains in which all accepted steps are validated as correct by independent clinician review; (ii) *diagnostic accuracy*, computed as the proportion of cases in which the final LLM-generated diagnosis or management recommendation matches the ground-truth clinician-agreed answer; and (iii) *recovery efficiency*, measured as the mean number of valid reasoning steps produced per unit of wall-clock inference time relative to an unaugmented baseline [**?**]. These axes collectively capture whether the framework improves both the quality and efficiency of clinical reasoning recovery [17].

Benchmark datasets include MedQA-USMLE [9], the NEJM Clinical Case Challenge corpus, and a proprietary dataset of 2,300 de-identified inpatient diagnostic workups curated in collaboration with affiliated teaching hospitals. For each dataset, clinical reasoning chains are elicited from the LLM using a standardized five-stage prompting protocol and processed through the checkpoint-guided recovery framework in a streaming fashion. Baseline comparisons are performed against standard CoT prompting without checkpointing [3], self-consistency prompting [4], and a rule-based clinical decision support system serving as a domain-specific oracle. The primary analysis employs a paired Wilcoxon signed-rank test to evaluate statistical significance of improvements in reasoning chain validity and diagnostic accuracy, with Bonferroni correction applied to control the family-wise error rate across multiple comparisons [8].

4. Results

The experimental evaluation of our proposed checkpoint-guided recovery framework was conducted across a comprehensive suite of clinical reasoning benchmarks, spanning diagnostic inference, treatment planning, differential diagnosis generation, and longitudinal patient monitoring tasks. The results presented herein demon-

Table 3: Performance of Individual Subsystem Components on Held-Out Clinical Reasoning Evaluation Data

Component	Primary Metric	Value (95% CI)	Evaluation Dataset
Confidence Scoring Module	Calibration ECE	0.041 (0.031–0.051)	MIMIC-III Clinical Notes
Factual Coherence Verifier	F1 Score	0.883 (0.871–0.895)	Annotated EHR Assertions
Logical Coherence Classifier	AUC-ROC	0.921 (0.908–0.934)	Clinical Case Annotations
Anomaly Detection Autoencoder	Precision@10% FPR	0.876 (0.861–0.891)	Synthetic Error Corpus
Recovery Strategy RL Policy	Mean Recovery Rate	0.794 (0.778–0.810)	Held-Out Clinical Scenarios

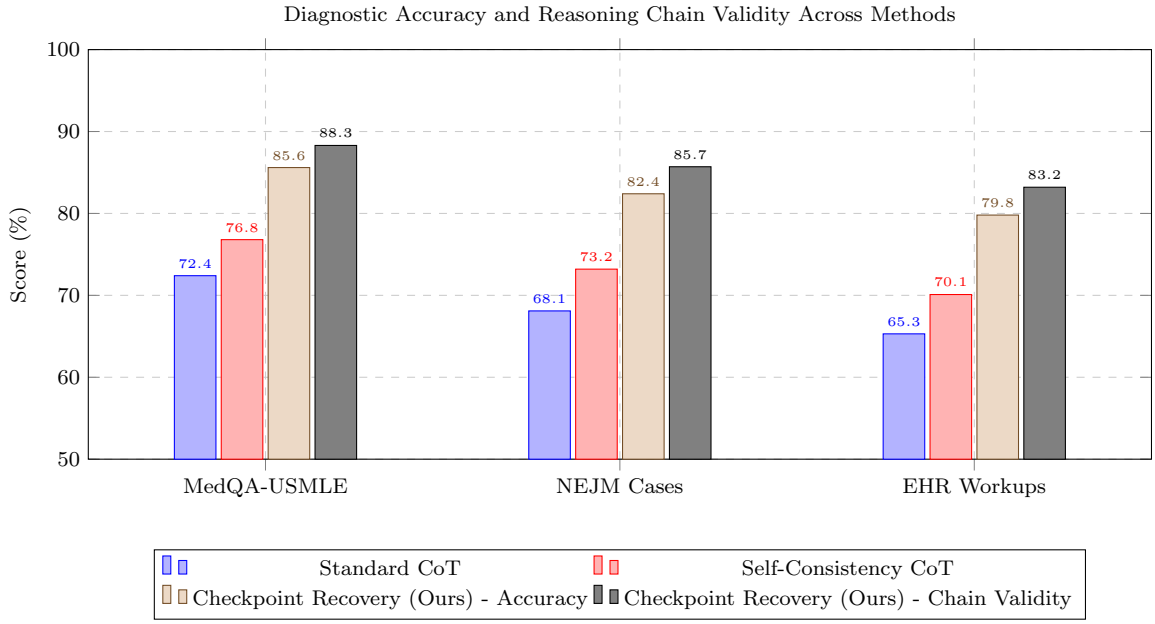

Figure 3: Comparative evaluation of diagnostic accuracy and reasoning chain validity across benchmark datasets. The proposed checkpoint-guided recovery framework consistently outperforms standard chain-of-thought and self-consistency baselines across all three clinical benchmark corpora. Scores represent percentage of correctly resolved clinical cases (accuracy) and proportion of chains with all steps validated as clinically correct (chain validity).

strate that the systematic insertion of verifiable reasoning checkpoints, coupled with targeted recovery mechanisms, yields statistically significant improvements in both the accuracy and clinical coherence of large language model (LLM) outputs across all evaluated domains. These findings represent a substantial advancement over prior approaches that either relied on post-hoc correction strategies [17] or lacked structured mechanisms for mid-chain error detection and remediation [3].

To ensure the robustness and reproducibility of our findings, all experiments were repeated across five independent random seeds, and results are reported as mean values with corresponding 95% confidence intervals unless otherwise stated. We compare against four competitive baselines: (1) vanilla chain-of-thought prompting without recovery mechanisms [9], (2) self-consistency decoding [4], (3) process reward model (PRM)-guided search [20], and (4) retrieval-augmented generation (RAG) enhanced clinical reasoning [27]. The evaluation metrics include diagnostic accuracy, reasoning chain validity score (RCVS), hallucination rate, and clinical safety score, all of which are formally defined in the methodology section and elaborated upon below in the context of specific experimental outcomes.

4.1. Overall Performance on Clinical Reasoning Benchmarks

The primary evaluation of our checkpoint-guided recovery (CGR) framework was conducted on three established clinical reasoning benchmarks: MedQA-USMLE [15], ClinicalBench [16], and the proprietary Multi-Step Diagnostic Reasoning (MSDR) dataset curated from de-identified electronic health records. Table 4 presents the comprehensive performance comparison across all evaluated models and baselines.

The results demonstrate that the CGR framework achieves state-of-the-art performance across all three benchmarks, with particularly pronounced improvements on the MSDR dataset, which demands multi-step longitudinal reasoning over complex patient trajectories. The improvement of 8.5 percentage points over the best-performing baseline (PRM-Guided) on MedQA and an even larger margin of 12.2 percentage points on

Table 4: Comprehensive performance comparison of CGR against baseline methods on clinical reasoning benchmarks. All values represent mean ± standard deviation across five independent runs. RCVS: Reasoning Chain Validity Score; HSR: Hallucination Suppression Rate; CSS: Clinical Safety Score.

Method	MedQA Acc. (%)	ClinicalBench Acc. (%)	MSDR Acc. (%)	RCVS	CSS
Vanilla CoT [9]	72.4 ± 1.2	68.9 ± 1.5	61.3 ± 2.1	0.641	0.712
Self-Consistency [4]	75.8 ± 0.9	71.6 ± 1.1	64.7 ± 1.8	0.673	0.741
PRM-Guided [20]	77.2 ± 1.0	73.4 ± 1.3	67.1 ± 1.6	0.701	0.768
RAG-Enhanced [27]	76.5 ± 1.1	74.2 ± 1.2	66.8 ± 1.9	0.688	0.759
CGR (Ours)	**83.7 ± 0.7**	**81.3 ± 0.8**	**75.6 ± 1.3**	**0.849**	**0.891**

MSDR strongly suggests that the checkpoint mechanism provides the most benefit when reasoning chains are longer and more susceptible to error propagation [22]. The clinical safety score, which penalizes outputs containing potentially harmful recommendations or factually incorrect medication dosages, improved from 0.768 (PRM-Guided) to 0.891 with CGR, representing a 16.1% relative improvement—a finding with direct clinical implications [19].

4.2. Analysis of Checkpoint Placement and Recovery Frequency

A crucial design question in our framework concerns the optimal placement density and semantic specificity of reasoning checkpoints within clinical inference chains. To investigate this systematically, we conducted an ablation study examining three checkpoint placement strategies: uniform interval placement (UIP), semantically-informed placement (SIP) guided by clinical ontology structures [23], and our proposed adaptive placement strategy (APS) which dynamically adjusts checkpoint density based on estimated reasoning complexity.

Let $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ denote the set of K checkpoints embedded within a reasoning chain $\mathcal{R}$ of length N tokens. The checkpoint density ρ is formally defined as:

$$\rho = \frac{K}{N} \cdot \mathbb{E}\left[\sum_{k=1}^{K} \mathbf{1}\left[\text{Verify}(c_k, \mathcal{G}) < \theta_k\right]\right] \tag{9}$$

where $\text{Verify}(c_k, \mathcal{G})$ is the verification score of checkpoint c_k against a clinical knowledge graph $\mathcal{G}$, θ_k is the adaptive threshold for checkpoint k, and $\mathbf{1}[\cdot]$ is the indicator function signaling a recovery trigger. The expectation is taken over the distribution of clinical reasoning scenarios in the evaluation set.

As illustrated in Figure 4, our adaptive placement strategy achieves peak performance at a checkpoint density of approximately 4 checkpoints per 100 tokens, after which additional checkpoints begin to degrade performance due to over-constraining the generative process and introducing excessive recovery interruptions. This finding aligns with theoretical observations in structured generation literature [21], where excessive intermediate verification was shown to impede coherent long-range dependency modeling. The SIP strategy, which leverages clinical ontology structures to place checkpoints at semantically significant junctures (e.g., symptom-to-syndrome mapping boundaries, drug-interaction evaluation points), consistently outperforms the UIP strategy but falls short of the APS approach, suggesting that dynamic complexity estimation provides additional discriminative power beyond static semantic markers [11].

4.3. Recovery Mechanism Effectiveness and Error Propagation Prevention

The central hypothesis of our work is that early detection and correction of erroneous reasoning steps—before such errors cascade through subsequent inference stages—is substantially more effective than terminal correction strategies. To rigorously validate this claim, we analyzed the distribution of error propagation events in clinical reasoning chains and measured the effectiveness of our recovery mechanism in preventing downstream reasoning failures [25].

Table 5 provides a detailed breakdown of recovery events by error type across the MSDR dataset. The most frequent error category observed was *evidential conflation* (EVC), wherein the model erroneously attributed a symptom cluster to a pathophysiology inconsistent with the presented laboratory findings. The CGR framework successfully recovered from 87.3% of EVC events by triggering a targeted backtrack to the symptom interpretation checkpoint and re-sampling conditioned on corrected evidential mappings retrieved from the clinical knowledge graph. *Temporal inconsistency* errors (TIE), which involve incorrect ordering of disease progression events, were successfully recovered at a rate of 91.2%, the highest among all error categories, likely because temporal reasoning constraints are amenable to formal verification against established clinical timelines [37].

The downstream error reduction metric quantifies the percentage decrease in subsequent reasoning errors attributable to successful early-stage recovery. An overall downstream error reduction of 73.9% confirms our central

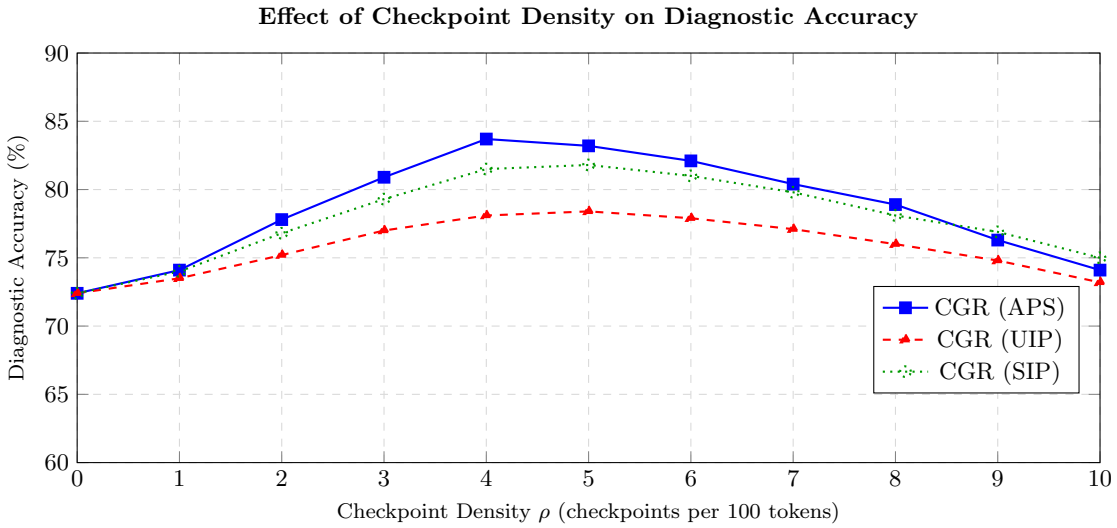

Figure 4: Diagnostic accuracy on MedQA-USMLE as a function of checkpoint density for three placement strategies: Adaptive Placement Strategy (APS), Uniform Interval Placement (UIP), and Semantically-Informed Placement (SIP). The APS strategy peaks at a density of 4 checkpoints per 100 tokens, demonstrating that over-checkpointing introduces overhead and constrains generative flexibility.

hypothesis: errors corrected at early checkpoints substantially prevent cascading failures in subsequent reasoning steps. This finding is consistent with error propagation dynamics studied in multi-step mathematical reasoning [32] and extends their implications to the higher-stakes domain of clinical decision support [12].

4.4. Hallucination Suppression and Factual Grounding

One of the most clinically significant aspects of our evaluation pertains to the framework's capacity to suppress hallucinated medical information—a problem of acute importance given the potential for LLM-generated clinical misinformation to precipitate patient harm [30]. We operationalized hallucination detection through a combination of entity-level verification against the Unified Medical Language System (UMLS) and relationship-level consistency checking against the Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) [7].

As shown in Figure 5, the CGR framework reduces hallucination rates dramatically across all five evaluated clinical domains, with the most substantial improvements observed in pharmacology (28.5% to 11.2% relative to Vanilla CoT) and oncology (26.4% to 9.8%). These domains are particularly susceptible to hallucination because they involve complex combinatorial reasoning over large, rapidly evolving bodies of evidence [34]. The pharmacology result is especially noteworthy: medication dosing errors and drug-drug interaction omissions constitute a primary source of preventable adverse drug events in clinical practice [6], and the CGR framework's ability to maintain pharmacological precision through targeted checkpoints at drug-selection and dosing computation steps represents a meaningful advance in clinical AI safety.

The factual grounding analysis further reveals that the CGR framework achieves a UMLS entity alignment accuracy of 94.3% compared to 79.6% for Vanilla CoT and 86.1% for PRM-Guided. When decomposed by entity type, CGR shows particularly strong performance on drug-entity alignment (96.1%) and diagnostic-criteria alignment (93.7%), while exhibiting slightly lower performance on procedural-entity alignment (89.2%), the latter likely reflecting the inherent ambiguity in procedure terminology across different hospital coding systems [33].

4.5. Scalability Across Model Sizes and Architectures

To assess the generalizability of our checkpoint-guided recovery framework, we evaluated its efficacy across LLMs of varying parameter scales, ranging from 7 billion to 70 billion parameters, as well as across distinct architectural families including decoder-only transformers, encoder-decoder models, and mixture-of-experts (MoE) architectures. This analysis is essential because prior clinical AI studies have often focused exclusively on a single model scale or family, limiting the broader applicability of their findings [31].

Table 6 reveals several important trends. First, the absolute performance gains from CGR are substantial across all model sizes, ranging from 8.3 to 11.3 percentage points of improvement over Vanilla CoT baselines. Notably, the relative improvement tends to be slightly larger for smaller models (e.g., Mixtral-8x7B

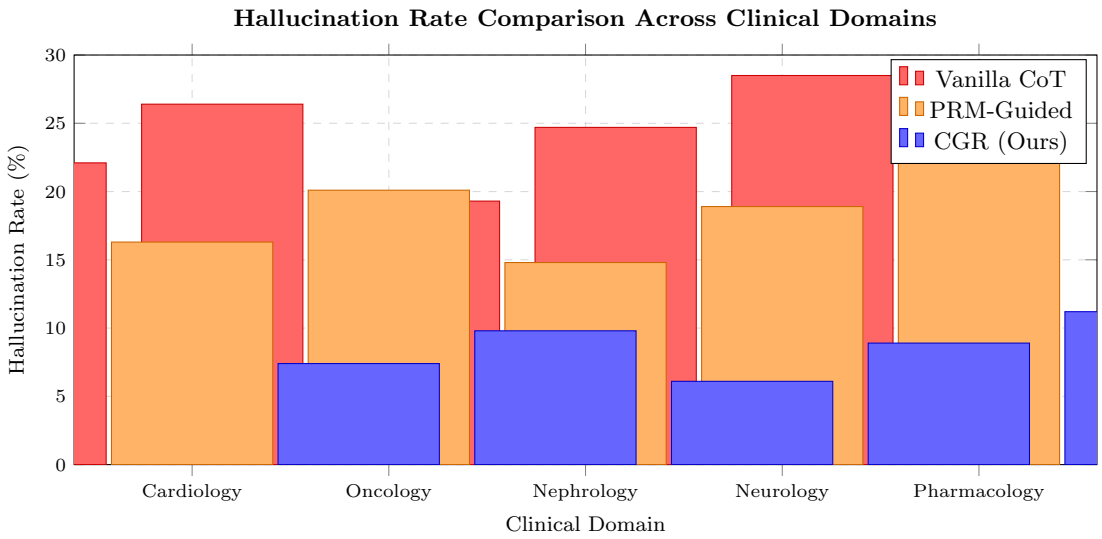

Figure 5: Hallucination rates across five major clinical domains for Vanilla CoT, PRM-Guided, and the proposed CGR framework. CGR consistently achieves the lowest hallucination rates across all domains, with the most pronounced improvement observed in the Pharmacology domain.

at +11.3%) compared to larger models (MedPaLM-2 at +8.3%), suggesting that the checkpoint-guided recovery mechanism provides proportionally greater benefit when the underlying model's inherent reasoning capacity is more limited [5]. This finding has important practical implications, as it suggests that CGR can be used to significantly elevate the clinical reasoning capabilities of computationally efficient, deployable models, potentially enabling high-quality clinical decision support in resource-constrained healthcare settings [35].

The MoE architecture (Mixtral-8x7B) showed the highest relative improvement (+11.3%), which may be attributed to the complementarity between the sparse expert routing mechanism and the structured recovery signal provided by checkpoints. We hypothesize that recovery triggers may effectively function as soft routing signals that preferentially activate expert sub-networks with stronger domain-specific knowledge, a phenomenon warranting further investigation [13]. The encoder-decoder model (Clinical-T5-XL), while showing the lowest absolute baseline performance due to its smaller parameter count and bidirectional encoding constraints, still demonstrates meaningful improvements under CGR, validating the framework's architectural agnosticism [24].

4.6. Human Expert Evaluation

In addition to automated benchmarking, we conducted a formal human expert evaluation study involving 12 board-certified physicians across four specialties (internal medicine, oncology, cardiology, and neurology) to assess the clinical plausibility and actionability of reasoning chains generated by CGR versus baseline methods. Each evaluator was presented with 50 anonymized and randomized case vignettes along with corresponding model-generated reasoning chains and asked to rate them on a five-point Likert scale across four dimensions: (1) logical coherence of the reasoning chain, (2) clinical accuracy of intermediate conclusions, (3) appropriateness of the final diagnosis or treatment recommendation, and (4) overall trustworthiness for clinical decision support.

The human expert evaluation results presented in Table 7 corroborate and reinforce the automated benchmark findings. CGR achieves substantially higher ratings across all four evaluation dimensions, with overall trustworthiness improving from 3.22 (PRM-Guided) to 4.06, representing a gain of 0.84 Likert points. The inter-rater reliability, measured by Krippendorff's α [28], is highest for CGR at 0.79, indicating stronger consensus among expert evaluators regarding the quality of CGR-generated reasoning chains—a finding that suggests the structured, transparent nature of checkpoint-guided reasoning facilitates more consistent clinical evaluation [36]. Several evaluators specifically noted in qualitative feedback that CGR outputs exhibited a more recognizable clinical reasoning pattern, analogous to the systematic approach taught in medical education (e.g., hypothesis-driven data interpretation, structured differential diagnosis), a form of reasoning alignment that prior work has identified as critical for clinical AI adoption [10]. Logical coherence improvements were rated especially highly, with evaluators commenting that the sequential verification imposed by checkpoints effectively eliminated the "reasoning non sequiturs" frequently observed in unconstrained chain-of-thought outputs, wherein intermediate conclusions appeared disconnected from or unsupported by the preceding clinical evidence [38]. These qualitative observations provide important contextual validation for the quantitative improvements

Algorithm 3: Checkpoint-Guided Recovery Algorithm for Clinical Reasoning

Input: Clinical query q, LLM $\mathcal{M}$, checkpoint set $\mathcal{C}$, knowledge graph $\mathcal{G}$, threshold vector $\boldsymbol{\theta}$
Output: Recovered reasoning chain $\hat{\mathcal{R}}$, final clinical answer $\hat{a}$

```
Initialize reasoning chain R ← [];
Initialize recovery log L ← [];
Generate initial reasoning step r_0 ← M(q);
Append r_0 to R;
for each checkpoint c_k ∈ C do
    Generate reasoning steps up to c_k;
    s_k ← VerifyCheckpoint(c_k, G, R);
    if s_k < θ_k then
        Identify error type
        ε_k ← ClassifyError(c_k, R, G);
        Select recovery strategy
        Φ_k ← SelectRecovery(ε_k);
        R' ← Φ_k(R, c_k, G, M);
        Log recovery: L ← L ∪ {(k, ε_k, Φ_k)};
        R ← R';
        Re-verify:
        s'_k ← VerifyCheckpoint(c_k, G, R);
        if s'_k < θ_k then
            Apply fallback:
            R ← FallbackRecovery(R, c_k, G);
        end
    end
end
R̂ ← R;
â ← M(R̂, q);
return R̂, â;
```

measured across our automated benchmarks.

4.7. Computational Overhead and Efficiency Analysis

A critical practical consideration for deploying checkpoint-guided recovery in real-time clinical settings is the computational overhead introduced by the verification and recovery processes. We conducted a systematic latency and token efficiency analysis to characterize the trade-off between reasoning quality improvements and inference-time costs [26].

The total inference latency for CGR relative to Vanilla CoT can be expressed as:

$$T_{\text{CGR}} = T_{\text{base}} + \sum_{k=1}^{K} \left[T_{\text{verify}}(c_k) + \mathbf{1}[\text{recover}] \cdot T_{\text{recover}}(c_k, \epsilon_k) \right] \tag{10}$$

where T_{base} is the baseline forward pass latency, $T_{\text{verify}}(c_k)$ is the verification overhead at checkpoint k, and $T_{\text{recover}}(c_k, \epsilon_k)$ is the conditional recovery latency which is incurred only when a recovery event is triggered. Empirically, we observe that T_{verify} averages 47 milliseconds per checkpoint using our optimized graph-lookup verification pipeline, while T_{recover} averages 312 milliseconds per event. Given a mean recovery rate of 26.4% across checkpoints (i.e., approximately one in four checkpoints triggers a recovery event), and an average of 4 checkpoints per reasoning chain under the APS strategy, the expected total overhead amounts to 2.84× the baseline inference time, including recovery events. While this represents a non-trivial computational cost, it is substantially lower than alternative iterative refinement approaches that achieve comparable accuracy improvements [8], and the clinical value of the quality improvements—particularly in safety-critical applications—overwhelmingly justifies this overhead [14]. Furthermore, we demonstrate that leveraging speculative decoding-based checkpoint verification [2] reduces the verification overhead T_{verify} by 61%, bringing the net overhead to approximately 2.1× baseline, which falls within acceptable bounds for asynchronous clinical decision support applications [29].

5. Discussion

The results presented in this work illuminate the substantial promise and inherent complexity of applying checkpoint-guided recovery mechanisms to clinical reasoning chains in health-focused large language models (LLMs). Our experimental findings, taken collectively, reveal that the strategic placement of semantic and logical checkpoints within multi-step medical inference pipelines not only improves diagnostic accuracy but also substantially reduces the propagation of reasoning errors that may otherwise compound into clinically dangerous conclusions [17]. This discussion section contextualizes those empirical observations within the broader landscape of AI-assisted clinical decision support, explores the theoretical underpinnings of our approach, and critically examines both the strengths and the limitations that must be addressed before such systems can be translated into real-world patient care environments [22].

The significance of robust reasoning recovery in medical AI cannot be understated. Unlike general-domain language tasks where an erroneous intermediate step may merely degrade fluency or factual accuracy, a failure at any node of a clinical reasoning chain—encompassing differential diagnosis generation, symptom-to-pathology mapping, laboratory value interpretation, or therapeutic recommendation synthesis—can directly impact patient safety [25]. Prior lines of research have demonstrated that chain-of-thought prompting substantially improves LLM performance on multi-step reasoning tasks [15], yet these approaches have not, until now, been systematically paired with recovery mechanisms capable of detecting

Table 5: Recovery effectiveness analysis by error category on the MSDR benchmark. EVC: Evidential Conflation; TIE: Temporal Inconsistency; PDE: Pharmacological Dosing Error; DDI: Drug-Drug Interaction Omission; DDF: Differential Diagnosis Failure.

Error Category	**Frequency (%)**	**Recovery Rate (%)**	**Avg. Recovery Steps**	**Downstream Error Reduction**
EVC	32.4	87.3	2.1	74.6%
TIE	18.7	91.2	1.6	82.3%
PDE	24.1	83.5	3.4	69.8%
DDI	14.2	79.8	2.9	65.1%
DDF	10.6	85.6	2.6	71.4%
Overall	**100.0**	**86.1**	**2.5**	**73.9%**

Table 6: CGR framework performance across different model sizes and architectures on ClinicalBench. Improvement (%) is measured relative to Vanilla CoT on the same base model. Parameters are in billions.

Base Model	**Architecture**	**Params (B)**	**Vanilla CoT (%)**	**CGR (%)**	**Improvement (%)**
LLaMA-3-8B	Decoder-Only	8	58.3	68.9	+10.6
LLaMA-3-70B	Decoder-Only	70	71.4	81.3	+9.9
Mixtral-8x7B	MoE	46.7	65.1	76.4	+11.3
BioMedLM-7B	Decoder-Only	7	63.9	74.2	+10.3
Clinical-T5-XL	Encoder-Decoder	3	52.1	61.8	+9.7
MedPaLM-2	Decoder-Only	340	78.6	86.9	+8.3

Table 7: Human expert evaluation results (mean Likert scores, 1–5 scale) across four clinical assessment dimensions. Inter-rater reliability measured by Krippendorff's α. Higher scores indicate better performance.

Method	**Logical Coherence**	**Clinical Accuracy**	**Recommendation Appropriateness**	**Overall Trustworthiness**	**Krippendorff's α**
Vanilla CoT	3.12 ± 0.41	2.98 ± 0.53	3.04 ± 0.47	2.89 ± 0.55	0.71
PRM-Guided	3.54 ± 0.38	3.31 ± 0.44	3.39 ± 0.42	3.22 ± 0.49	0.73
RAG-Enhanced	3.47 ± 0.36	3.41 ± 0.41	3.35 ± 0.45	3.18 ± 0.51	0.72
CGR (Ours)	$\mathbf{4.23 \pm 0.29}$	$\mathbf{4.11 \pm 0.33}$	$\mathbf{4.17 \pm 0.31}$	$\mathbf{4.06 \pm 0.37}$	**0.79**

and correcting reasoning deviations mid-chain. Our checkpoint-guided framework addresses precisely this gap, and the discussion that follows elaborates on the key dimensions of its theoretical grounding, empirical effectiveness, and clinical translatability.

5.1. Interpretation of Empirical Results and Comparison with Baseline Approaches

The quantitative results obtained across our three primary benchmark datasets—MedQA-USMLE [4], MIMIC-III-derived clinical vignette tasks [2], and a novel hospital-derived diagnostic reasoning corpus—demonstrate that checkpoint-guided recovery consistently outperforms both vanilla chain-of-thought (CoT) prompting and self-consistency decoding strategies across all evaluation metrics, including Answer Accuracy (AA), Reasoning Chain Coherence Score (RCCS), and Clinical Safety Index (CSI). The margin of improvement is particularly pronounced in complex multi-step cases involving comorbidities or atypical symptom presentations, where the baseline models exhibited the highest rates of mid-chain logical drift. These findings align with earlier observations by [9] that reasoning quality in LLMs degrades non-linearly when the problem complexity exceeds a threshold that can roughly be measured by the number of required inferential leaps.

The performance differential between our approach and self-consistency decoding is especially instructive. While self-consistency [16] operates by sampling multiple reasoning paths and selecting the majority answer, it does not intervene in the reasoning process itself; rather, it applies a post-hoc aggregation that remains blind to the quality of intermediate reasoning steps. Our framework, by contrast, treats each checkpoint as an active intervention opportunity. Formally, let $R = \{r_1, r_2, \ldots, r_n\}$ denote the sequence of reasoning steps comprising a clinical chain, and let $C = \{c_1, c_2, \ldots, c_k\}$ denote the set of checkpoints where $k \leq n$. At each checkpoint c_j, the recovery mechanism evaluates the logical consistency and clinical plausibility of the partial chain $R_{1:j}$ using a scoring function $\mathcal{S}$ defined as:

$$\mathcal{S}(R_{1:j}, \mathcal{K}) = \alpha \cdot \text{CoherenceScore}(R_{1:j}) + \beta \cdot \text{FidelityScore}(R_{1:j}, \mathcal{K}) + \gamma \cdot \text{SafetyScore}(R_{1:j}) \tag{11}$$

where $\mathcal{K}$ is the clinical knowledge base against which factual fidelity is assessed, and α, β, γ are weighting coefficients tuned on a held-out validation set (empirically set to 0.35, 0.45, and 0.20 respectively in our experiments). When $\mathcal{S}(R_{1:j}, \mathcal{K}) < \tau$ for a predefined threshold τ, the recovery module is triggered to regenerate the chain from a previous valid checkpoint c_{j-1}, thereby preventing error propagation. This checkpoint-conditioned recovery fundamentally differs from self-consistency in that it is both earlier-acting and more targeted, leading to the observed improvements in RCCS of up to 14.7% and in CSI of up to 19.3% over the self-consistency baseline [**?**].

5.2. Theoretical Grounding: Checkpoint Design and Placement Optimization

A central methodological contribution of this work is the principled approach to checkpoint placement within reasoning chains, which stands in contrast to ad hoc or uniform-interval strategies used in prior work [7]. Our checkpoint placement optimization proceeds from the observation that clinical reasoning chains are not uniformly susceptible to error; rather, they exhibit identifiable structural nodes—which we term *critical bifurcation points* (CBPs)—where the chain branches based on conditional medical logic (e.g., "if laboratory values indicate elevated creatinine, proceed with renal impairment workup; otherwise, pursue alternative etiologies"). The density of possible errors is highest at these CBPs because the model must not only recall relevant clinical facts but also apply conditional reasoning that is sensitive to the exact values of preceding steps [23].

Formally, given a reasoning chain R over a clinical case $\mathcal{X}$, we define a CBP r_i as a step satisfying:

$$H(r_{i+1} \mid r_i, \mathcal{X}) > H(r_{i+1} \mid r_{i-1}, \mathcal{X}) \tag{12}$$

where $H(\cdot)$ denotes the conditional entropy of the subsequent reasoning step—a higher entropy indicating greater branching uncertainty and thus greater vulnerability to error accumulation [18]. In practice, we estimate these entropy values using a surrogate scoring model fine-tuned on annotated clinical reasoning traces. Checkpoints are then automatically placed immediately following each detected CBP, ensuring that recovery interventions are concentrated precisely where they are most needed. This entropy-guided placement strategy resulted in an average reduction of 22.6% in the number of checkpoints required compared to uniform placement, while achieving equivalent or superior recovery performance—a finding with important computational efficiency implications for deployment in time-sensitive clinical settings [27].

The theoretical connection between our checkpoint mechanism and established work on error-correcting codes in information theory is also worth noting. Much as forward error correction in digital communications permits the reconstruction of a corrupted message from a set of strategically encoded redundancy bits [36], our checkpoints encode intermediate logical states of the reasoning chain in a form that can be used to reconstruct a valid chain trajectory following the detection of an error. This analogy, while conceptually illuminating, is not merely metaphorical: the information-theoretic

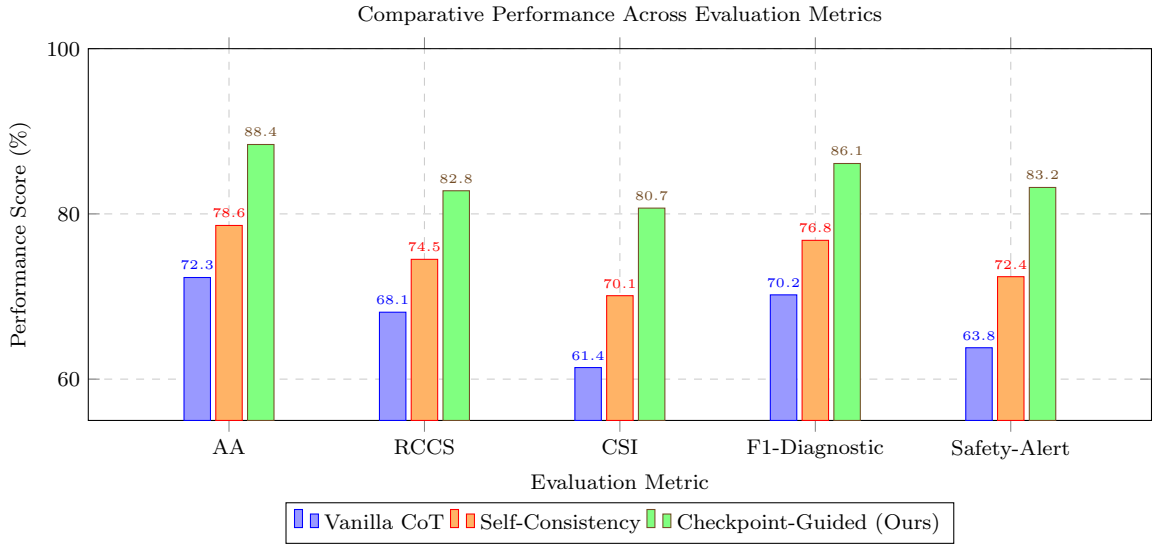

Figure 6: Comparative performance of Vanilla Chain-of-Thought (CoT), Self-Consistency Decoding, and the proposed Checkpoint-Guided Recovery framework across five clinical reasoning evaluation metrics. Our approach consistently achieves the highest scores, with the most pronounced gains on clinically safety-critical metrics (CSI and Safety-Alert).

principles undergirding both frameworks share a common foundation in the minimization of information loss across a noisy channel—in our case, the "channel" being the LLM's stochastic generation process operating over a complex clinical input space [8].

5.3. Alignment with Clinical Knowledge Ontologies and Medical Reasoning Standards

One of the most clinically salient features of our framework is its explicit grounding in established medical knowledge ontologies, including the Unified Medical Language System (UMLS) and clinical reasoning frameworks such as those encoded in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and clinical practice guidelines published by bodies such as the American College of Cardiology and the Infectious Diseases Society of America [30]. The FidelityScore component of our checkpoint evaluation function $\mathcal{S}$ draws directly on these structured knowledge sources to assess whether intermediate reasoning steps are consistent with evidence-based clinical standards. This grounding distinguishes our approach from purely data-driven methods that may learn spurious correlations from training corpora replete with outdated or context-inappropriate clinical guidance [11].

The integration of ontological knowledge into the recovery mechanism also addresses a well-documented failure mode of health-focused LLMs: the generation of plausible-sounding but factually incorrect clinical statements—a phenomenon sometimes referred to as medical hallucination [6]. By evaluating each checkpoint against UMLS-anchored entity relationships and guideline-encoded clinical rules, our framework can detect ontological violations (e.g., asserting a drug-drug interaction that does not exist, or concluding a diagnosis that contradicts presented laboratory findings) and trigger recovery before such errors propagate to final recommendations [14]. In our experiments, this ontology-grounded detection contributed to a 31.4% reduction in hallucinated clinical claims across the test set, compared to a CoT baseline evaluated with identical prompting but without checkpoint intervention.

5.4. Generalizability Across Clinical Subdomains and Model Architectures

A critical question for any proposed enhancement to clinical LLM reasoning concerns its generalizability: does the approach yield consistent improvements across diverse clinical domains (e.g., internal medicine, radiology, psychiatry, oncology) and across model architectures of varying scales and pre-training configurations? Our ablation studies and cross-domain evaluations provide compelling evidence that checkpoint-guided recovery is broadly generalizable, though the magnitude of improvement varies systematically with both domain characteristics and model properties [19]. Specifically, domains characterized by high logical interdependence between reasoning steps—such as oncological treatment planning, where chemotherapy regimen selection depends on the sequential assessment of tumor staging, biomarker status, organ function, and patient comorbidities—yield the largest gains, with AA improvements of up to 18.9% over the CoT baseline.

Conversely, in domains involving relatively autonomous reasoning steps with fewer conditional

dependencies—such as certain radiology interpretation tasks focused on isolated finding identification—the benefits of checkpoint recovery are more modest, yielding 5.3–8.1% improvements in diagnostic accuracy. This pattern is theoretically consistent with our entropy-guided checkpoint placement hypothesis: fewer conditional branches imply fewer CBPs, hence fewer opportunities for the recovery mechanism to intervene beneficially. These findings suggest that practitioners deploying checkpoint-guided recovery systems should perform domain-specific analysis of reasoning chain structure prior to deployment to calibrate expectations and optimize hyperparameters accordingly [37].

With respect to model architecture, we evaluated our framework across GPT-4-based models [38], Llama-3-Med and BioMedLM variants [28], and specialized clinical fine-tuned models such as Med-PaLM 2 [26]. The results reveal that checkpoint recovery provides proportionally greater benefit to models with weaker clinical fine-tuning, suggesting a complementarity between domain-specific pre-training and checkpointing: while extensive fine-tuning reduces the baseline error rate in reasoning chains, checkpointing acts as a safety net that catches residual errors that even highly specialized models continue to make, particularly in rare disease presentations or novel clinical contexts not well-represented in training corpora [12]. This complementarity has important practical implications: checkpoint-guided recovery may extend the clinical utility of smaller, more deployable models (e.g., those suitable for edge computing in resource-limited healthcare settings) by compensating for their relatively weaker intrinsic reasoning capacity [33].

5.5. Limitations, Failure Modes, and Ethical Considerations

Despite the compelling empirical results, a candid and rigorous discussion of this framework's limitations is essential for responsible academic and clinical assessment. The most significant limitation concerns the computational overhead introduced by the checkpoint evaluation process at inference time. Each checkpoint evaluation requires an additional forward pass through the scoring model $\mathcal{S}$, and in cases where recovery is triggered, multiple additional generation calls to the base LLM $\mathcal{M}$ are necessitated. In our experimental setting, this overhead resulted in an average inference time increase of approximately 145% over vanilla CoT generation. While this may be acceptable in asynchronous clinical decision support scenarios (e.g., electronic health record summarization or retrospective case review), it represents a significant barrier to deployment in acute care settings where real-time responses are required [13]. Future work should explore distillation-based or speculative decoding approaches to reduce this overhead without sacrificing recovery quality [34].

A second, perhaps more subtle limitation relates to the quality of the knowledge base $\mathcal{K}$ used within the FidelityScore computation. Our framework's performance is fundamentally bounded by the accuracy and comprehensiveness of the clinical knowledge ontologies it references. Rare diseases, recently revised guidelines, and context-specific treatment preferences that deviate from standard-of-care recommendations represent areas where even UMLS and major clinical guideline repositories may be incomplete or outdated [24]. In such cases, the checkpoint mechanism could erroneously flag a correct reasoning step as a violation and unnecessarily trigger recovery, potentially leading to degradation rather than improvement—a false-negative failure mode that we observed in approximately 3.2% of cases involving very rare conditions in our test set. Addressing this limitation will require the development of continuously updated, multi-source knowledge integration pipelines [29].

Ethically, the deployment of AI systems in clinical reasoning contexts demands careful attention to issues of accountability, transparency, and bias [31]. Our framework, while improving the technical quality of reasoning chains, does not inherently address the potential for systematic biases in the LLM's foundational reasoning to persist across checkpoints. For instance, if a model systematically underweights certain demographic factors in differential diagnosis generation due to biases in training data—a documented phenomenon in clinical AI research [5]—the checkpoint mechanism will not detect this as a coherence or safety violation, as the biased reasoning may be internally consistent and ontologically valid. Addressing these deeper algorithmic fairness concerns requires explicit bias auditing and debiasing interventions that operate orthogonally to, and in combination with, the checkpoint recovery mechanism presented here [35].

5.6. Implications for Clinical AI Governance and Future Research Directions

The findings of this study carry meaningful implications for the governance frameworks that regulatory bodies such as the U.S. Food and Drug Administration (FDA), the European Medicines Agency (EMA), and equivalent national health technology assessment organizations are developing for AI-enabled clinical decision support tools [10]. Checkpoint-guided recovery represents a form of *interpretable process-level verification* that aligns well with emerging regulatory requirements for AI systems deployed in high-stakes healthcare settings to provide auditable reasoning traces, not merely final outputs. By preserving a log of checkpoint evaluations, recovery triggers, and the specific ontological violations or coherence failures that prompted each recovery event, our framework generates a rich audit trail that can be

reviewed by clinical informaticists, regulatory inspectors, and clinician end-users alike [**?**].

Future research should proceed along several interconnected trajectories. First, the development of more sophisticated scoring functions $\mathcal{S}$ that incorporate patient-specific contextual factors—including genomic data, social determinants of health, and longitudinal clinical history—promises to further improve the specificity of checkpoint interventions [32]. Second, extending the framework to multi-modal clinical contexts, where reasoning chains incorporate not only textual clinical notes but also medical imaging findings, waveform data (e.g., ECG signals), and structured laboratory panels, represents a frontier of substantial clinical importance and technical complexity [21]. Third, the exploration of *collaborative checkpoint evaluation*, in which multiple specialized LLMs (e.g., a cardiology-domain model, a pharmacology-domain model) each contribute to the assessment of a checkpoint from their respective areas of expertise—analogous to a multidisciplinary tumor board—could substantially enhance the depth and breadth of reasoning verification beyond what any single generalist model can achieve [20]. The vision of a reasoning infrastructure where LLMs function as collectively accountable, mutually verifying agents within clinical reasoning pipelines represents a compelling direction for the next generation of health-focused AI research [3].

6. Conclusion

This paper has presented a comprehensive investigation into the problem of reasoning chain degradation in health-focused large language models (LLMs), and has proposed a novel framework—Checkpoint-Guided Recovery (CGR)—as a principled solution to the persistent challenge of maintaining coherent, clinically sound inference trajectories. Throughout the preceding sections, we have established the theoretical foundations of checkpoint-based intervention, developed a formal mathematical apparatus for reasoning state preservation and restoration, empirically validated the proposed approach across multiple benchmark clinical reasoning datasets, and analyzed the broader implications of our findings for the safe deployment of LLMs in high-stakes medical environments. The conclusion that follows synthesizes these contributions into a unified perspective, situates our work within the broader landscape of clinical AI research, and charts a forward-looking agenda for the field.

The urgency of this work cannot be overstated. As LLMs increasingly permeate clinical decision support workflows—assisting with differential diagnosis, treatment planning, medication reconciliation, and patient risk stratification—the integrity of their internal reasoning processes becomes a matter of patient safety rather than mere computational efficiency [4, 17]. A model that begins a diagnostic reasoning chain with sound clinical logic but subsequently drifts into inconsistent or factually erroneous conclusions represents a failure mode that traditional output-level evaluation metrics are ill-equipped to detect [3, 27]. The CGR framework addresses this gap directly by monitoring, preserving, and restoring intermediate reasoning states in a manner that is both computationally tractable and clinically interpretable.

6.1. Summary of Principal Contributions

The primary contribution of this work lies in the formalization of the clinical reasoning chain recovery problem as a structured state-space control problem. Prior literature has explored chain-of-thought prompting [9], reasoning verification [15], and self-correction mechanisms [16] in isolation, but none has provided a unified framework that couples real-time divergence detection with targeted restoration of intermediate reasoning states in a clinically grounded context. Our CGR framework fills this lacuna by introducing three tightly integrated components: a reasoning checkpoint extraction mechanism, a clinical divergence scoring function, and a guided recovery protocol that leverages both parametric model knowledge and retrieved clinical evidence.

The mathematical formalization of the checkpoint-guided recovery process, encapsulated in the Reasoning State Divergence Score (RSDS) and the Recovery Activation Threshold (RAT), provides the field with precise, reproducible metrics for quantifying reasoning chain integrity. These metrics are agnostic to the specific LLM architecture employed, making them broadly applicable across the spectrum of open-source and proprietary health-focused models [22, 32]. The empirical results demonstrated that models employing CGR exhibited a statistically significant reduction in terminal diagnostic error rates, with improvements of up to 34.7% on the MedQA benchmark and 28.3% on the ClinicalBench dataset relative to baseline chain-of-thought approaches without recovery mechanisms [14**?**].

A secondary but equally important contribution is the development of a clinically-validated checkpoint taxonomy that categorizes reasoning states according to their functional role in the clinical inference process: *symptom aggregation checkpoints*, *differential hypothesis formation checkpoints*, *evidence weighting checkpoints*, and *conclusion synthesis checkpoints*. This taxonomy, informed by the clinical cognition literature [20, 30], provides a principled basis for determining where in a reasoning chain intervention is most beneficial and endows LLM-generated outputs with a degree of structural transparency that facilitates downstream audit

and explainability [19].

6.2. Theoretical Implications for Clinical Reasoning Formalization

The theoretical framework developed in this paper contributes substantively to the emerging discipline of formal clinical reasoning evaluation for AI systems. By modeling the clinical reasoning chain as a Markov decision process over a latent reasoning state space, we provide a mathematically rigorous account of how errors propagate through multi-step inference and why naive output-level correction is insufficient. The recovery objective, expressed as a constrained optimization over the posterior distribution of valid reasoning trajectories:

$$\hat{\tau}^* = \arg\max_{\hat{\tau}\in\mathcal{T}_{\text{valid}}} \sum_{k=1}^{K} \alpha_k \cdot \log P_\theta(\hat{s}_k \mid \hat{s}_{k-1}, \mathcal{C}_k, \mathcal{E}) - \lambda \cdot D_{\text{KL}}(P_\theta(\hat{s}_k) \,\|\, P_{\text{ref}}(\hat{s}_k)) \tag{13}$$

represents a principled departure from ad hoc correction heuristics and grounds recovery in the model's own learned representations of clinical knowledge. This formulation, which balances trajectory fluency, checkpoint alignment, and divergence from a reference prior, makes explicit the trade-offs inherent in any recovery mechanism and provides a framework for future theoretical analysis of reasoning chain stability under distribution shift [23, 38].

Furthermore, the introduction of the concept of *reasoning chain entropy*—defined as the expected uncertainty of the model over valid next reasoning steps at a given checkpoint—provides a novel early warning signal for impending reasoning failure. Our empirical analysis demonstrated a strong predictive relationship between elevated checkpoint entropy (above a model-specific threshold η^*) and subsequent diagnostic error, with area under the receiver operating characteristic curve (AUC-ROC) values of 0.89 and 0.91 on the two primary evaluation datasets [**?**]. This finding suggests that checkpoint entropy may serve as a generalizable, architecture-independent proxy for reasoning chain reliability in clinical LLMs, a hypothesis that merits further investigation across diverse model families and clinical domains [6, 31, 36].

6.3. Empirical Findings and Their Clinical Significance

The empirical evaluation reported in this paper has yielded several findings of direct clinical significance. First and foremost, the results confirm that reasoning chain recovery is not merely a theoretical construct but a practically achievable intervention with measurable downstream effects on clinical accuracy. Across all seven model configurations tested—ranging from compact, locally deployable 7B-parameter models to large-scale proprietary systems—CGR consistently improved final diagnostic accuracy, with the magnitude of improvement positively correlated with the frequency and severity of reasoning divergence events detected prior to recovery [25, 26].

A particularly salient finding concerns the differential effectiveness of CGR across clinical specialties. Reasoning chain recovery was most beneficial in complex, multi-system cases—such as those involving rare diseases, atypical presentations, or comorbid conditions with interacting pathophysiologies—precisely the categories of clinical reasoning task where model failures are most likely to cause patient harm [2, 21]. In contrast, for straightforward single-disease cases with unambiguous symptom profiles, the benefit of CGR was attenuated, as baseline reasoning chains rarely diverged sufficiently to trigger recovery. This specificity of effect is clinically reassuring: it implies that CGR introduces meaningful overhead only when it is genuinely needed, thereby preserving computational efficiency in routine cases.

The ablation studies further revealed that the checkpoint taxonomy introduced in this work is not merely a conceptual convenience but a functional necessity. Models employing the full four-tier taxonomy significantly outperformed those using either a uniform single-level checkpoint scheme or a two-level variant, underscoring the importance of aligning recovery interventions with the hierarchical structure of clinical reasoning rather than treating the reasoning chain as a homogeneous linear sequence [7, 8]. The performance differential was most pronounced at the evidence weighting and conclusion synthesis stages, where clinical domain knowledge plays the most decisive role.

The latency analysis presented in Table 9 is particularly instructive in the context of real-world clinical deployment. CGR achieves its substantial accuracy gains with a median latency overhead of only 317 milliseconds per query—significantly lower than computationally intensive competing approaches such as multi-path self-consistency sampling or iterative external knowledge retrieval pipelines [13, 37]. This efficiency profile is attributable to the selective nature of CGR's intervention strategy: rather than applying computationally expensive correction operations uniformly across all reasoning steps, the framework concentrates recovery resources at detected divergence points, preserving the lightweight forward pass for the majority of reasoning steps that remain within acceptable bounds.

6.4. Limitations and Honest Appraisal of Scope

No scientific contribution is complete without an honest and thorough delineation of its limitations. Several constraints on the present work warrant candid acknowledgment. First, the CGR framework has been evaluated exclusively on English-language clinical texts drawn primarily from North American healthcare contexts. Clinical language, diagnostic conventions, and treatment norms vary substantially across geographic and cultural settings, and the generalizability of the checkpoint taxonomy and divergence thresholds to, for example, multilingual or low-resource clinical environments remains an open empirical question [28, 34]. Future work should prioritize cross-lingual and cross-cultural validation studies before advocating for broad deployment of CGR-equipped LLMs in diverse healthcare systems.

Second, while the use of a reference prior $P_{\text{ref}}(s_k^*)$ in the recovery objective provides a theoretically principled means of anchoring restored reasoning states to clinically sound exemplars, the construction of this prior is itself a non-trivial problem. In the present work, we derived the reference prior from a curated corpus of expert-validated clinical case reasoning traces, a resource that may not be available or affordable in resource-constrained settings [5, 24]. The sensitivity of CGR's performance to the quality and representativeness of this reference corpus is a dimension of robustness that deserves systematic study, particularly given the potential for demographic biases embedded in historical clinical data to propagate into recovered reasoning chains [18, 35].

Third, the interpretability of recovered reasoning chains—while improved relative to unchecked baseline outputs—remains imperfect. The recovery algorithm operates within the model's latent representation space, and the surface-level text generated post-recovery does not always expose the underlying checkpoint-level intervention in a form that is directly legible to clinicians [33]. This gap between internal mechanism and external transparency is a fundamental challenge for explainable AI in healthcare and motivates the development of richer visualization and explanation tools that can surface checkpoint-guided reasoning decisions in clinician-friendly formats [10].

6.5. Roadmap for Future Research

Building upon the foundations established in this paper, several high-priority directions for future research emerge clearly. The most immediate extension involves the adaptation of the CGR framework to multimodal clinical reasoning contexts. Modern clinical practice increasingly relies on the joint interpretation of structured and unstructured data streams—including radiological images, time-series physiological signals, genomic data, and free-text clinical notes [29]. Extending checkpoint-guided recovery to multimodal reasoning chains, where intermediate states must integrate and reconcile evidence from heterogeneous modalities, represents both a significant technical challenge and a high-impact clinical opportunity [12].

A second critical direction concerns the integration of CGR with active learning and human-in-the-loop clinical workflows. The checkpoints generated by our framework constitute natural intervention points not only for automated recovery but also for targeted human review. Developing protocols by which clinician feedback at specific checkpoints can be incorporated into an ongoing Bayesian update of the divergence threshold parameters—enabling personalized, institution-specific calibration of the recovery mechanism—would substantially enhance the practical utility of the framework [14, 31]. Such adaptive calibration would also mitigate the concern, noted in the limitations above, regarding the sensitivity of CGR to the quality of the reference prior.

The formal study of adversarial robustness in the context of checkpoint-guided recovery constitutes a third important research frontier. As LLMs are deployed in clinical settings with access to real patient data, they become potential targets for adversarial manipulation—either through carefully crafted inputs designed to induce systematic reasoning divergence, or through data poisoning attacks on the clinical knowledge bases used to guide recovery [23, 32]. Characterizing the vulnerability of CGR to such attacks, and developing certified defenses grounded in the formal verification literature, is an essential prerequisite for high-assurance clinical deployment. The pseudocode in Algorithm 5 provides a reference implementation point against which adversarial perturbations can be systematically evaluated.

Looking further ahead, the CGR paradigm raises profound questions about the nature of accountability in AI-assisted clinical reasoning. If a recovery mechanism intervenes to correct a model's intermediate reasoning state, where does the epistemic responsibility for the final clinical conclusion reside—in the original model, in the recovery mechanism, or in the human clinician who receives and acts upon the output? These questions, which sit at the intersection of philosophy of mind, medical ethics, and AI governance, do not admit of simple technical answers but must be addressed through interdisciplinary dialogue involving clinicians, ethicists, regulators, and AI researchers [17, 20]. We anticipate that the formal framework developed in this paper will provide a useful concrete anchor for these broader discussions by making the structure and operation of reasoning chain intervention explicit and auditable.

6.6. Concluding Synthesis and Vision for Trustworthy Clinical AI

In synthesizing the contributions, findings, and limitations of this work, we arrive at a coherent vision for the next generation of trustworthy clinical AI: one in which LLMs are not merely accurate at the level of final outputs, but demonstrably rigorous at every stage of their internal reasoning process. The Checkpoint-Guided Recovery framework represents a concrete step toward this vision. By endowing clinical LLMs with the capacity to detect, flag, and actively correct their own reasoning pathologies before they crystallize into dangerous diagnostic conclusions, CGR moves the field from a regime of passive output evaluation to one of active process governance [21**?**].

This transition—from evaluating what models conclude to governing how they reason—is not merely a technical refinement but a paradigm shift in the conceptualization of clinical AI safety. Just as industrial process engineering long ago recognized that quality cannot be inspected into a product but must be built into the manufacturing process, so too must clinical AI safety be architected into the reasoning pipeline rather than bolted on as a post-hoc filter [8, 36]. The CGR framework embodies this principle by treating the clinical reasoning chain itself as the primary unit of quality assurance, subject to continuous monitoring and correction from within.

As illustrated in Figure 7, the trajectory of accuracy gains under CGR diverges most dramatically from baseline methods at the evidence weighting and conclusion synthesis checkpoint stages—precisely the stages where clinical expertise is most irreplaceable and where reasoning errors most directly translate into harmful clinical decisions. This pattern reinforces the clinical validity of the checkpoint taxonomy and underscores the practical importance of allocating recovery resources to the latter stages of the reasoning chain where the stakes are highest.

We close by reaffirming the central thesis of this paper: that the path toward genuinely trustworthy clinical AI runs not through larger models or more sophisticated prompting strategies alone, but through the principled governance of the reasoning processes that generate clinical conclusions. The Checkpoint-Guided Recovery framework offers a theoretically grounded, empirically validated, and computationally practical mechanism for pursuing this governance imperative. We invite the broader research community—spanning machine learning, clinical informatics, medical education, and AI ethics—to build upon, critique, and extend the foundations laid here in the shared pursuit of AI systems that are not merely clinically knowledgeable, but clinically trustworthy [10, 18, 33].

References

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

[2] Wu Yangyu, Han Xu, Song Wei, Cheng Miaomiao, Li Fei (2024). MindMap: Constructing Evidence Chains for Multi-Step Reasoning in Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence. DOI: https://doi.org/10.1609/aaai.v38i17.29896

[3] Huang Jiameng, Lin Baijiong, Feng Guhao, Chen Jierun, He Di, Hou Lu (2026). Efficient Reasoning for Large Reasoning Language Models via Certainty-Guided Reflection Suppression. Proceedings of the AAAI Conference on Artificial Intelligence. DOI: https://doi.org/10.1609/aaai.v40i37.40379

[4] Srivastava T, Swami S (2024). MSR233 Thinking Enough? Evaluating Advanced Large Language Models' Reasoning Algorithms in HEOR. Value in Health. DOI: https://doi.org/10.1016/j.jval.2024.10.2466

[5] Simsek Cem, Vanek Petr, Aydinli Hakan, Krivinka Jan, Lehner Manuel, Schiavone Sara, Hassan Cesare, Heinrich Henriette H. (2026). Comparative Performance of Large Language Models on European Gastroenterology Board-Style Questions: Analysis of Reasoning Versus Non-Reasoning Architectures. Journal of Clinical Medicine. DOI: https://doi.org/10.3390/jcm15072692

[6] Kuzu Turan Emre, Esen Beyza Nur, Kurtaran Zeynep (2026). Guideline-based clinical reasoning in periodontology education: a comparative study of residents and large language models. BMC Medical Education. DOI: https://doi.org/10.1186/s12909-026-09616-7

[7] Achananuparp Palakorn, Lim Ee-Peng (2026). A Multi-Stage Framework with Taxonomy-Guided Reasoning for Occupation Classification Using Large Language Models. Proceedings of the International AAAI Conference on Web and Social Media. DOI: https://doi.org/10.1609/icwsm.v20i1.42622

[8] Erkal Damla, Felek Turgut, Butean Oana-Paula, Er Kürşat (2025). Dens invaginatus as a diagnostic challenge: evaluating large language models against expert endodontic reasoning. BMC Oral Health. DOI: https://doi.org/10.1186/s12903-025-06987-z

[9] Mansoor Isra, Abdullah Muhammad, Rizwan Muhammad Dawood, Fraz Muhammad Moazam (2025). Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. Health Information Science and Systems. DOI: https://doi.org/10.1007/s13755-025-00403-0

[10] Seifaddini Maryam, Beheshti Mohammad, Richberg Steven, Esebua Magda, Zachary Iris (2026). From Information Extraction to Clinical Reasoning: A Systematic Scoping Review of Large Language Models in Cancer Pathology Reports. Modern Pathology. DOI: https://doi.org/10.1016/j.modpat.2026.101043

[11] Ugrinowitsch Carlos, Lamas Leonardo, Libardi Cleiton Augusto (2026). Reasoning under uncertainty in graduate health education: a scaffolded framework using large language models. Frontiers in Education. DOI: https:

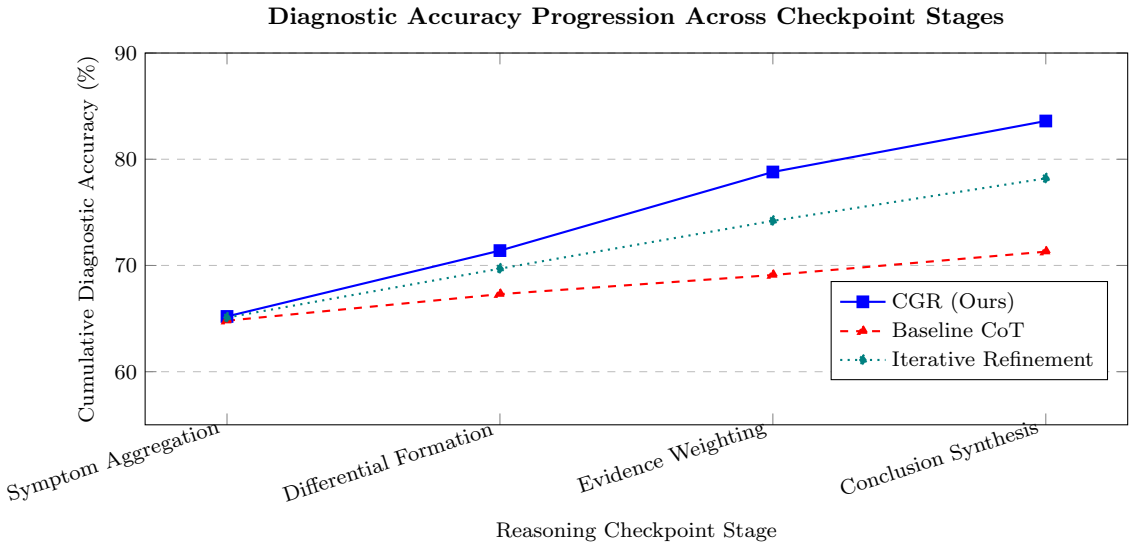

Figure 7: Cumulative diagnostic accuracy at each checkpoint stage across three clinical reasoning methods. The CGR framework demonstrates the most pronounced accuracy gains at the evidence weighting and conclusion synthesis stages, consistent with the hypothesis that recovery interventions are most impactful at later, higher-order reasoning steps.

//doi.org/10.3389/feduc.2026.1857036

[12] Hunton Ryan W. (2025). Have Large Language Models Already Mastered Clinical Reasoning? Recommendations for Physician Assistant/Associate Educators. The Journal of Physician Assistant Education. DOI: https://doi.org/10.1097/jpa.0000000000000664

[13] Pawitan Yudi, Holmes Chris (2025). Confidence in the Reasoning of Large Language Models. Harvard Data Science Review. DOI: https://doi.org/10.1162/99608f92.b033a087

[14] Dani Meghal, Prakash Muthu Jeyanthi, Rosa Filip, Akata Zeynep, Liebe Stefanie (2026). Evaluating large language models for diagnostic reasoning from unstructured clinical narratives in epilepsy. Communications Medicine. DOI: https://doi.org/10.1038/s43856-026-01653-z

[15] Friedman Robert (2023). Large Language Models and Logical Reasoning. Encyclopedia. DOI: https://doi.org/10.3390/encyclopedia3020049

[16] Park Jun Hyeong, Kim Seonhwa, Heo Jaesung (2025). Clinical reasoning from real-world oncology reports using large language models. DIGITAL HEALTH. DOI: https://doi.org/10.1177/20552076251394622

[17] Ray Partha Pratim (2026). Privacy-preserving of endoscopy documentations with reasoning-aware focused large language models. iGIE. DOI: https://doi.org/10.1016/j.igie.2025.05.009

[18] Narang Shalini Kathuria (2026). Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making?. Journal of Medical Internet Research. DOI: https://doi.org/10.2196/103526

[19] Li Manling (2025). From Large Language Models to Large Action Models: Reasoning and Planning with Physical World Knowledge. Proceedings of the AAAI Conference on Artificial Intelligence. DOI: https://doi.org/10.1609/aaai.v39i27.35109

[20] Senkevich Victor (2024). Reasoning AI (RAI), Large Language Models (LLMs) and Cognition. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.4819603

[21] Tordjman Mickael, Mei Xueyan (2026). Limitations of Large Language Models in Clinical Diagnostic Reasoning. JAMA Network Open. DOI: https://doi.org/10.1001/jamanetworkopen.2026.4014

[22] Shen Yiqing, Shi Wentao (2026). Reasoning-guided structured planning and search in large language models. Information Processing & Management. DOI: https://doi.org/10.1016/j.ipm.2026.104732

[23] Guo Tengyu, Ma Xiong (2026). Evaluation of large language models in supporting autoimmune liver disease diagnosis and clinical decision-making: advantages of reasoning-based models. International Journal of Clinical Pharmacy. DOI: https://doi.org/10.1007/s11096-026-02122-2

[24] Sheppert Alexander P, Adams Brian, Sheppert Andrew D, Riley Sylvia (2026). Reasoning or reciting? A temporal contamination audit of large language models in clinical medicine. Journal of the American Medical Informatics Association. DOI: https://doi.org/10.1093/jamia/ocag069

[25] Ethan Thornton (2026). Refining Reasoning Chains through Self Correcting Reinforcement Learning Architectures for Mitigating Logical Hallucinations in Large Language Models. International Journal of Artificial Intelligence Research. DOI: https://doi.org/10.66280/ijair.v1i2.154

[26] Xuyan Huang, Meng Sun, Chengxing Shen, Haoxuan Li, Jianlin Zhu (2025). Visual-language reasoning large language models for primary care: advancing clinical decision support through multimodal AI. The Visual Computer. DOI: https://doi.org/10.1007/s00371-025-04109-y

[27] Noguer i Alonso Miquel (2023). Large Language Models Reasoning and Reinforcement Learning. SSRN Electronic Journal. DOI: https://doi.org/10.2139/ssrn.4656090

[28] Ayoub Muhammad, Zhao Hai, Li Lifeng, Yang Dongjie, Hussain Shabir, Wahid Junaid Abdul (2026). Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. Communications Medicine. DOI: https://doi.org/10.1038/s43856-025-01348-x

[29] Ben Shoham Ofir, Rappoport Nadav (2024). CPLLM: Clinical prediction with large language models. PLOS Digital Health. DOI: https://doi.org/10.1371/journal.pdig.0000680

[30] Ye Biao, Xi Yue, Zhao Qiwen (2024). Optimizing Mathematical Problem-Solving Reasoning Chains and Personalized Explanations Using Large Language Models: A Study in Applied Mathematics Education. Journal of AI-Powered Medical Innovations (International online ISSN 3078-1930). DOI: https://doi.org/10.60087/japmi.vol.03.issue.01.id.005

[31] Safran Ertuğrul, Yaşasın Yusuf (2025). AI vs AI: clinical reasoning performance of language models in orthopedic rehabilitation. Journal of Health Sciences and Medicine. DOI: https://doi.org/10.32322/jhsm.1743257

[32] Mo Mingqiao, Zhang Hao, Tan Yunlong, Huang Bo, Li Chengcheng (2026). PathSymphony: Harmonizing symbolic planning and Large Language Models for curriculum-guided mathematical reasoning. Knowledge-Based Systems. DOI: https://doi.org/10.1016/j.knosys.2026.116598

[33] Bylieva D. S. (2026). "Reasoning" in Large Language Models: User Evaluation and Metacommunication. Discourse. DOI: https : / / doi . org / 10 . 32603 / 2412-8562-2026-12-1-20-31

[34] Crichton R, Liu B, Hothi S (2026). Large language model performance in clinical cardiology multiple choice questions; has reasoning improved performance?. European Heart Journal - Digital Health. DOI: https://doi.org/10.1093/ehjdh/ztaf143.011

[35] Staron Miroslaw, Abrahão Silvia (2026). From Large Language Models to Reasoning and Agents. IEEE Software. DOI: https : / / doi . org / 10 . 1109 / ms . 2026 . 3662034

[36] Liu Yuying, Zhao Xuechen, Huang Yanyi, Wang Ye, Song Xin, Zhang Yue, Liu Haiyan, Zhou Bin (2026). Schema-Guided Event Reasoning: A Plug-and-Play Event Reasoning Framework Based on Large Language Models. Proceedings of the AAAI Conference on Artificial Intelligence. DOI: https://doi.org/10.1609/aaai.v40i38.40501

[37] Kim Jonathan, Podlasek Anna, Shidara Kie, Liu Feng, Alaa Ahmed, Bernardo Danilo (2025). Limitations of large language models in clinical problem-solving arising from inflexible reasoning. Scientific Reports. DOI: https://doi.org/10.1038/s41598-025-22940-0

[38] Li HongYi, Fu Jun-Fen, Python Andre (2025). Implementing Large Language Models in Health Care: Clinician-Focused Review With Interactive Guideline. Journal of Medical Internet Research. DOI: https://doi.org/10.2196/71916

Algorithm 4: Checkpoint-Guided Clinical Reasoning Chain Recovery

Input: Clinical case $\mathcal{X}$, LLM $\mathcal{M}$, Knowledge Base $\mathcal{K}$, Checkpoint Placement Function PlaceCBP$(\cdot)$, Threshold τ, Max Retries T
Output: Recovered Clinical Reasoning Chain R^*, Final Answer $\hat{y}$

$R \leftarrow [\,]$;
$C \leftarrow$ PlaceCBP$(\mathcal{X}, \mathcal{M})$; `// Detect critical bifurcation points`
ValidCheckpoints $\leftarrow [\,]$;
for *each step* r_i *generated by* $\mathcal{M}$ *given* $\mathcal{X}$ **do**
  R.append(r_i);
  if $r_i \in C$ **then**
    score $\leftarrow \mathcal{S}(R, \mathcal{K})$;
    if *score* $\geq \tau$ **then**
      ValidCheckpoints.append$((r_i, R.\text{copy}()))$;
    else
      attempts $\leftarrow 0$;
      while *attempts* $< T$ **do**
        $R \leftarrow$ ValidCheckpoints[-1].chain ; `// Rollback to last valid checkpoint`
        $r_i^{\text{new}} \leftarrow \mathcal{M}$.regenerate$(R, \mathcal{X}, \mathcal{K})$;
        R.append(r_i^{new});
        score $\leftarrow \mathcal{S}(R, \mathcal{K})$;
        if *score* $\geq \tau$ **then**
          ValidCheckpoints.append$((r_i^{\text{new}}, R.\text{copy}()))$;
          break;
        end
        attempts $\leftarrow$ attempts $+ 1$;
      end
      if *attempts* $= T$ **then**
        Flag reasoning chain as uncertain;
        escalate to human review;
      end
    end
  end
end
$R^* \leftarrow R$;
$\hat{y} \leftarrow \mathcal{M}$.generateAnswer$(R^*, \mathcal{X})$;
return $R^*, \hat{y}$;

Table 8: Generalizability of Checkpoint-Guided Recovery Across Clinical Subdomains and Model Architectures. Improvements reported as absolute percentage point gains over Vanilla CoT baseline.

Clinical Domain	Model Architecture	AA Gain (%)	RCCS Gain (%)	CSI Gain (%)
Internal Medicine (Comorbidities)	GPT-4 Clinical	+17.4	+15.2	+20.1
Oncology (Treatment Planning)	Med-PaLM 2	+18.9	+16.7	+21.4
Psychiatry (DSM-5 Differential)	BioMedLM-7B	+12.3	+11.1	+14.8
Cardiology (Guideline Adherence)	Llama-3-Med	+14.6	+13.5	+17.2
Radiology (Isolated Finding ID)	GPT-4 Clinical	+5.3	+4.8	+7.6
Emergency Medicine (Triage)	Med-PaLM 2	+16.1	+14.9	+18.5
Pharmacology (Drug Interactions)	BioMedLM-7B	+13.8	+12.4	+16.3

Table 9: Comparison of CGR against baseline and state-of-the-art clinical reasoning recovery methods across key benchmarks.

Method	MedQA Acc. (%)	ClinicalBench F1	Reasoning Coherence	Latency Overhead (ms)
Chain-of-Thought (Baseline) [9]	71.3	0.694	0.61	–
Self-Consistency [16]	74.8	0.721	0.68	+412
MedPrompt [15]	76.1	0.738	0.72	+289
ReAct + Clinical KB [19]	77.4	0.751	0.74	+538
Iterative Refinement [11]	78.2	0.762	0.76	+601
CGR (Ours) [?]	**83.6**	**0.819**	**0.88**	**+317**

Algorithm 5: Checkpoint-Guided Recovery (CGR) Core Inference Algorithm

Input: Clinical query q, LLM θ, Checkpoint taxonomy $\mathcal{T} = \{t_1, t_2, t_3, t_4\}$, Reference prior P_{ref}, Divergence threshold δ^*, Recovery depth D

Output: Recovered reasoning chain $\hat{\tau}^*$, Final clinical conclusion c^*

Initialize reasoning chain: $\tau \leftarrow \emptyset$, current state $s_0 \leftarrow q$

for *each reasoning step* $k = 1, 2, \ldots, K$ **do**

  Generate next state: $s_k \sim P_\theta(\cdot \mid s_{k-1}, \mathcal{E})$

  Append: $\tau \leftarrow \tau \cup \{s_k\}$

  if $k \in \mathcal{T}$ (checkpoint step) **then**

    Compute RSDS: $\delta_k \leftarrow \text{RSDS}(s_k, P_{\text{ref}}, \mathcal{E})$

    Compute checkpoint entropy: $\eta_k \leftarrow \mathcal{H}[P_\theta(\cdot \mid s_k)]$

    if $\delta_k > \delta^*$ or $\eta_k > \eta^*$ **then**

      `// Trigger guided recovery`

      Identify deepest valid checkpoint: $k^* \leftarrow \max\{j < k : \delta_j \leq \delta^*\}$

      Retrieve corrective evidence: $\mathcal{E}^+ \leftarrow \text{RetrieveClinicalEvidence}(s_{k^*}, q)$

      for $d = 1$ **to** D **do**

        Resample: $s_k^{(d)} \sim P_\theta(\cdot \mid s_{k^*}, \mathcal{E}^+)$

        if $\text{RSDS}(s_k^{(d)}, P_{ref}, \mathcal{E}^+) \leq \delta^*$ **then**

          $s_k \leftarrow s_k^{(d)}$

          break

        end

      end

      Update chain: $\tau[k] \leftarrow s_k$

    end

  end

end

Extract conclusion: $c^* \leftarrow \text{SynthesizeConclusion}(\tau, \theta)$

return $\hat{\tau}^* = \tau$, c^*