Thinking Like a Clinician: A Cognitive AI Agent for Clinical Diagnosis via Panoramic Profiling and Adversarial Debate

Anonymous ACL submission

Abstract

The application of large language models (LLMs) in clinical decision support faces significant challenges of "tunnel vision" and diagnostic hallucinations present in their processing unstructured electronic health records (EHRs). To address these challenges, we propose a novel chain-based clinical reasoning framework, called **DxChain**, which transforms the diagnostic workflow into an iterative process by mirroring a clinician's cognitive trajectory that consists of "Memory Anchoring", "Navigation" and "Verification" phases. DxChain introduces three key methodological innovations to elicit the potential of LLM: (i) a Profile-Then-Plan paradigm to mitigate cold-start hallucinations by establishing a panoramic patient baseline, (ii) a Medical Tree-of-Thoughts (MedToT) algorithm for strategic look ahead planning and resource aware navigation, and (iii) a Dialectical Diagnostic Verification procedure utilizing "Angel-Devil" adversarial debates to resolve complex evidence conflicts. Evaluated on two real world benchmarks, MIMIC-IV-Ext Cardiac Disease and MIMIC-IV-Ext CDM, DxChain achieves state-of-the-art performances in both diagnostic accuracy and logical consistency, offering a modular and reliable architecture for next-generation clinical AI. The code is at `https://anonymous.4open.science/r/Dx-Chain`.

1 Introduction

Artificial intelligence applications in healthcare, particularly in clinical decision support systems (CDSS), are undergoing a profound paradigm shift (Liu et al., 2025; Elhaddad and Hamam, 2024; Wiest et al., 2025). Although large language models (LLMs) have demonstrated remarkable potential in processing massive volumes of medical text, directly deploying them in real world clinical workflows still faces substantial challenges (Teo et al., 2025; Hager et al., 2024; Jiang et al., 2025b). Most existing studies remain focused on static medical question answering benchmarks such as MedQA, and various multi-agent frameworks such as MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024) are predominantly evaluated on structured multiple choice datasets. In contrast, real world clinical diagnosis is an inherently dynamic and unstructured reasoning process: physicians must extract salient clues from noisy electronic health records (EHRs), construct a differential diagnosis under conditions of incomplete information, and iteratively revise their hypotheses in light of newly acquired evidence (Kagiyama et al., 2025; Chen et al., 2025b). In this high noise environment, current LLM based systems often falter. They exhibit a form of "tunnel vision" prematurely locking onto a single hypothesis while neglecting alternative possibilities or fall prey to "red herrings" producing diagnostic hallucinations by over interpreting chronic baselines or incidental anomalies as acute pathological signals (Zhu et al., 2025; Ranji, 2024). How to maintain high precision while ensuring high recall remains a central challenge for clinical AI (McDuff et al., 2025; Wiest et al., 2025).

The first and most pervasive challenge is the model's susceptibility to "cold-start hallucinations" driven by noise intolerance (Vishwanath et al., 2025). In clinical settings, patient narratives are often replete with "red herrings" Without a holistic cognitive framework, LLMs tend to over interpret these distractors with recent benchmarks revealing that 50% of advanced models suffer catastrophic failure in noisy triage scenarios, such as prioritizing conspicuous chronic chest tightness over subtle signs of deep vein thrombosis, thereby missing fatal risks like pulmonary embolism (Shen et al., 2025). Recent studies indicate that this sensitivity to noise leads models to generate false positives at an alarming rate, frequently misclassifying chronic symptoms as acute signals in complex cases (Zhu et al., 2025; Zhou et al., 2025a). This "tunnel vi-

sion" not only degrades diagnostic precision but also poses significant safety risks by triggering unnecessary interventions.

Secondly, the reasoning mechanisms of current LLMs are predominantly linear and fragile (Jiang et al., 2025a). Traditional Chain-of-Thought (CoT) approaches mimic a straight path from symptom to diagnosis. However, clinical reasoning is inherently branching and exploratory (Chen et al., 2025a).Linear models are prone to "premature closure" and the "premature pruning" phenotype: for example, misdiagnosing chest pain as a metabolic issue due to weight bias without a backtracking mechanism, leading to a 32% drop in accuracy (Hassan et al., 2025),leading to irreversible error propagation (Gu et al., 2025; Ranji, 2024).

Thirdly, existing frameworks lack a rigorous verification mechanism to resolve conflicting evidence. In scenarios where symptoms support multiple contradictory diagnoses, standard LLMs often force a coherent but hallucinated narrative rather than acknowledging ambiguity, failing to balance high precision with the necessary recall (McDuff et al., 2025; Wiest et al., 2025; Goh et al., 2024).

To bridge this gap between linear LLM inference and dynamic clinical cognition, we propose DxChain, a novel reasoning system designed to mirror the iterative cognitive trajectories of clinicians. Unlike traditional dialogue systems, DxChain decomposes diagnosis into a stateful, reflective process through three targeted innovations.

First, to counteract cold-start hallucinations, we introduce a Profile-Then-Plan paradigm. This module acts as a cognitive anchor, synthesizing a "panoramic" patient baseline from long term history before diagnosis begins, effectively filtering out chronic noise. Second, to address linear fragility, we incorporate a Medical Tree-of-Thoughts (Med-ToT) algorithm. This allows the system to simulate the physician's differential diagnosis process expanding, evaluating, and selecting multiple reasoning pathways thereby optimizing information gain and avoiding tunnel vision. Finally, to resolve evidence conflicts, we implement a Selective Dialectical Diagnostic Verification "Angel-Devil" debate. This module serves as a precision filter, triggering adversarial argumentation only for ambiguous cases to ensure reliability without redundant computation. Extensive evaluations on the MIMIC-IV-Ext Cardiac Disease (Goldberger et al., 2000) and CDM (Hager et al., 2024) benchmarks demonstrate that DxChain achieves SOTA performance, effectively transforming clinical AI from a static question answerer into a robust, cognitively aligned diagnostic partner.

2 Related Work

The field of Natural Language Processing (NLP) has witnessed a paradigm shift with the rapid advancement of Large Language Models (LLMs) (Brown et al., 2020; Zhao et al., 2023; Zhou et al., 2023), such as the GPT series, LLaMA (Touvron et al., 2023), and PaLM (Anil et al., 2023), which have demonstrated exceptional capabilities in language understanding and generation. Building upon these general purpose foundations, there has been a significant surge in the development of specialized medical LLMs designed to address the intricate demands of healthcare. This evolution spans from models adapted via domain specific pre-training and fine-tuning, such as BioBERT (Lee et al., 2020) and PMC-LLaMA (Wu et al., 2024) , to recent state-of-the-art systems like Med-PaLM 2 (Singhal et al., 2025) , MedGemma (Sellergren et al., 2025) , and Baichuan-M2 (Dou et al., 2025). These advanced models leverage techniques ranging from multi stage reinforcement learning to dynamic verification systems, significantly narrowing the gap between model capabilities and real world clinical decision-making needs.

Role-playing frameworks leverage LLMs to simulate clinical collaboration. Tang et al. propose MedAgents (Tang et al., 2024), where agents assume expert identities to reach consensus through multi-round discussions without parameter updates. Addressing efficiency, MDAgents (Kim et al., 2024) adaptively assigns collaboration structures from single clinicians to multidisciplinary teams based on task complexity. Furthermore, KAMAC (Wu et al., 2025) enhances adaptability by enabling dynamic team expansion, identifying real time knowledge gaps to recruit specialists as needed for evolving diagnostic contexts. However, lacking a unified "Memory Anchor" these decentralized agents are prone to local hallucinations, leading to "tunnel vision" in complex cases.

Beyond role-playing, recent research explores workflow-based and tool-augmented agents to simulate clinical procedures (Lyu et al., 2025; Fallahpour et al., 2025; Jin et al., 2025). Ferber et al. (Ferber et al., 2025) develope an oncology agent integrating precision medicine tools and vi-

sion transformers to ground decisions in medical guidelines. Focusing on reliability, MedAgent-Pro (Wang et al., 2025b) decouples diagnosis into planning and reasoning phases, utilizing visual tools and verification steps. Similarly, MAM (Zhou et al., 2025b) addresses multi-modal complexity by decomposing tasks into specialized roles to process diverse data modalities without retraining unified models.Nevertheless, their linear execution lacks a "Look ahead" mechanism for mental simulation, often resulting in "premature closure" without the ability to backtrack.

Despite advances in role-playing and tool-augmented clinical agents, resolving conflicting evidence remains challenging, motivating debate based verification. Multi-agents debate (MAD) serves as a verification layer via adversarial critique and cross examination. MoodAngels applies MAD to psychiatric diagnosis by triggering pro/con debate and a judge upon disagreement, improving robustness under symptom overlap and uncertainty (Xiao et al., 2025). ED2D couples debate with retrieval to enforce evidence grounding beyond medicine (Han et al., 2025). To reduce cost and avoid harmful reversals, iMAD selectively activates MAD using uncertainty signals from structured self-critique (Fan et al., 2025). Following this line, DxChain uses a selectively triggered "Angel-Devil" dialectical verification procedure to Winnow weak hypotheses and arbitrate evidence conflicts before final diagnosis.

Existing approaches focus on clinical workflows but overlook underlying cognitive mechanisms (Wang et al., 2025a). DxChain bridges this gap by shifting from procedural emulation to cognitive simulation. Through panoramic profiling, strategic planning, and dialectical verification, our architecture simulates how doctors think not just what they do thereby handling clinical complexity more effectively than linear, process oriented models.

3 Method

3.1 Memory Anchoring

Real world diagnosis relies on integrating fragmented signals into a coherent mental model rather than instant hypothesis generation (Kiesewetter et al., 2020). However, direct EHR processing often triggers "cold-start hallucinations" where agents over interpret acute symptoms while neglecting chronic baselines (Liu et al., 2024). To mitigate this, we adopt a Profile-Then-Plan strategy, where Phase I focuses on the Profiling stage to establish a panoramic baseline. Distinct from generic frequency based summarization, this phase performs clinical disentanglement, restructuring inputs into a rigorous Global Patient Representation to initialize the reasoning state (S_0) before diagnostic planning begins.

The process initiates with Clinical Perception, acting as a noise filter for the raw input space $\mathcal{D}$. Let $x_{\text{raw}} \in \mathcal{D}$ denote the heterogeneous input text. The perception agent A_{perc}, instantiated as a prompt engineered LLM, employs a Clinical Disentanglement extraction strategy to map x_{raw} into a structured JSON object S_{struct}:

$$S_{\text{struct}} = A_{\text{perc}}(x_{\text{raw}} \mid K_{\text{schema}}) \in \mathbb{J} \qquad (1)$$

where $\mathbb{J}$ represents the space of valid JSON objects and K_{schema} defines the target clinical entities (e.g., HPI). This step converts unstructured narrative into discrete key value pairs, ensuring that subsequent reasoning operates on verified clinical facts rather than redundant linguistic noise.

In parallel with this extraction, to avoid "unnel vision" the profiling agent A_{prof} directly processes the raw narratives (x_{raw}) to synthesize a semantic anchor, the Patient Profile (P_{base}). Crucially, P_{base} is not a latent vector but a structured natural language description explicitly disentangled into three dimensions:

$$P_{base} = \{C_{acute}, C_{chronic}, R_{risk}\} \qquad (2)$$

Here, the architecture explicitly disentangles the clinical picture into *Acute Presentations* (C_{acute}), *Chronic Baseline* ($C_{chronic}$), and *Risk Factors* (R_{risk}). This *Global Patient Representation* is injected into the shared state memory, serving as an immutable anchor. By grounding the diagnostic process in this comprehensive profile, DxChain ensures that all future planning step whether requesting tests or proposing diagnoses are driven by a holistic understanding of the patient rather than reactive responses to isolated symptoms.

3.2 Navigation via Medical Tree-of-Thoughts

To bridge the gap between static procedural emulation and dynamic cognitive simulation, we formalize the diagnostic process as stateful navigation within a structured probabilistic framework. By orchestrating the diagnostic loop, as illustrated in

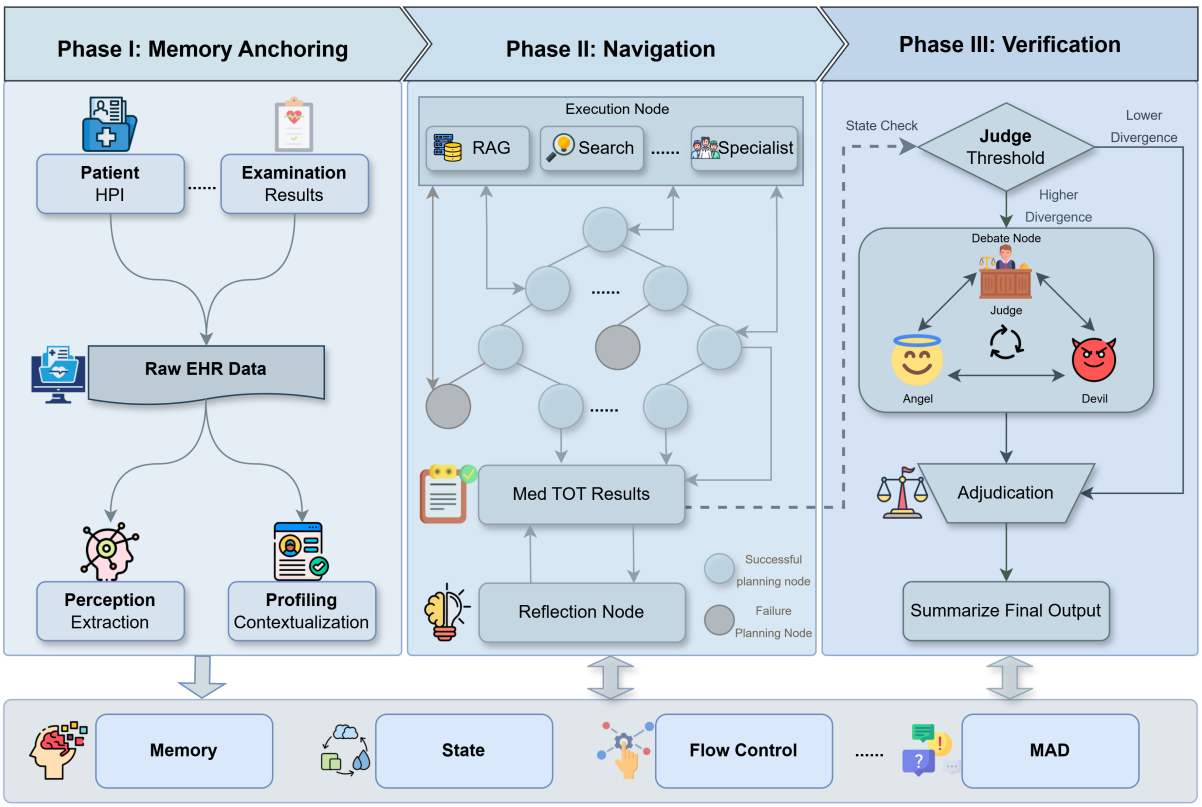

Figure 1: Overview of the proposed clinical reasoning framework. Phase I anchors memory from raw EHR via perception extraction and patient profiling. Phase II performs state aware navigation over a Tree-of-Thoughts execution graph, integrating RAG, search, and specialist agents. Phase III winnows and arbitrates candidate diagnoses via judge checks and MAD (Multi-Agent Debate: "Angel-Devil") to synthesize the final output.

Figure 1, DxChain transcends traditional prompt-based reasoning. Unlike linear Chain-of-Thought (CoT) methodologies that rely on static procedural emulation, our architecture adopts a Medical Tree-of-Thoughts (Med-ToT) approach to implement a Dynamic Cognitive Simulation paradigm.

Formally, diagnostic reasoning is modeled as a directed cyclic graph $G = (N, E, S)$ aligned with Dual Process Theory. Nodes (N) represent cognitive modules rather than simple procedural steps. Crucially, node expansion is constrained by patient contextualized medical logic, ensuring the trajectory mirrors coherent clinical inquiry rather than arbitrary semantic generation. The State (S) acts as Working Memory, persisting the differential diagnosis (H_t) and investigation history, while Edges (E) facilitate metacognitive cycles (e.g., Execution $\rightarrow$ Planning) that mimic iterative hypothetico deductive reasoning.

The core of this navigation is the Medical Tree-of-Thoughts (Med-ToT) planner. Standard ToT approaches often suffer from excessive divergence; to address this, we introduce a "diagnostic centripetal force" to constrain the search space. The planner (N_{plan}) performs a look ahead mental simulation using a predefined strategy stack (e.g., "Rule out Emergency", "Focused Investigation"), ensuring all generated branches remain medically plausible. Specifically, conditioned on the current diagnostic state S_t and the accumulated investigation history H_{history}, the model plans the next **investigative action** a_t while explicitly articulating a rationale and generating a key **Clinical Expectation** E_t. This generation process is formalized as:

$$\pi(a_t, E_t | S_t) \leftarrow \text{Med-ToT}(S_t, H_{history}) \quad (3)$$

This mechanism compels the model to commit to a hypothesis before observing outcomes. Here, E_t represents not just a final diagnosis, but intermediate physiological predictions (e.g., "expecting elevated Troponin" or "positive Murphy's sign").

To handle the uncertainty of real-world data, we introduce a Discrepancy Driven mechanism. Upon receiving new clinical observations (O_t), which correspond to actual EHR data points or test results, an LLM based evaluator computes a *conflict score* between the observation and the prior expectation E_t.

A score $\geq \tau$ indicates a substantial conflict (e.g., a normal ECG contradicting a myocardial infarction hypothesis), triggering immediate replanning. The tree depth is dynamically managed to prevent infinite regression while allowing sufficient exploration. The control flow is defined as:

$$\text{Flow}_t = \begin{cases} \text{TriggerReplan, if LLM}_{\text{judge}}(O_t, E_t) \\ \qquad \geq \tau \\ \text{Update}(S_t, O_t) \rightarrow \text{Proceed, otherwise} \end{cases} \tag{4}$$

After the Planning Node (N_{plan}) is completed, the workflow proceeds to the Synthesis Node (N_{syn}) to produce the final consolidated diagnosis, and then enters the Reflection Node (N_{ref}) for quality review and feedback control. If the review fails, the process returns to the Planning stage for further iterative refinement until the requirements are satisfied.

3.3 Dialectical Adjudication

This phase is initiated by the Judge Node (N_{judge}), which acts as a gatekeeper between free form diagnostic reasoning and downstream verification. Operationally, the Judge performs an internal state check over the evidence accumulated during the navigation phase including the structured patient profile, the clinical abstract, key positive/negative findings, and the current working "Final Diagnosis". Formally, it invokes a constrained LLM with a Pydantic based JSON schema and computes a verdict vector that includes: (i) a per diagnosis *diagnosis_status* label ("Confident", "Ambiguous", or "Incorrect"), (ii) a list of *ambiguity_points* that require further debate, and (iii) a set of *diagnoses_to_remove* for hypotheses deemed incompatible with the available evidence.

The overall logic of this phase can be decomposed into three components: an internal gating judgment performed by the Judge, an adversarial debate executed in the Debate Node(N_{Debate}), and the generation of the final diagnostic result. Among these, Figure 2 primarily focuses on visualizing the debate and synthesis workflow.

This verification mechanism is implemented as a multi-agents debate, conceptualized as an "Angel-Devil" adversarial game. Within the Debate Node(N_{Debate}), two distinct agent roles are instantiated: a proponent agent (A_{angel}) that aggregates positive findings to construct a confirmation chain for the candidate diagnosis, and a critic

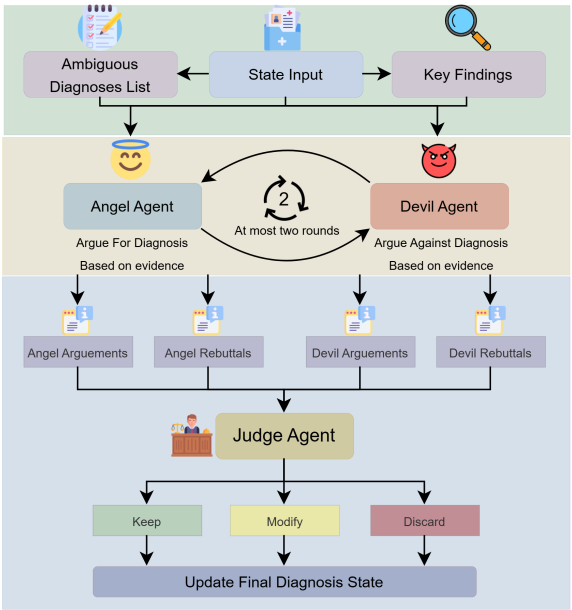

Figure 2: The workflow of the Dialectical Verification phase. An Angel-Devil adversarial debate is triggered to argue for and against the diagnosis, with a Judge Agent making the final decision to keep, modify, or discard the hypothesis.

agent (A_{devil}) that actively searches for negative evidence, logical fallacies, or overlooked differential diagnoses. In each iteration, the Judge synthesizes the arguments from both agents to update the confidence scores of the hypothesis. This adversarial process effectively excises low confidence diagnoses of the reasoning tree where the Devil agent successfully identifies disqualifying evidence, ensuring that only hypotheses capable of withstanding scrutiny are retained.

After the debate converges or the Judge's criteria are met, an LLM is invoked to finalize the output by summarizing the retained diagnoses and mapping them to standardized terminologies, producing a structured report suitable for downstream EHR systems.

4 Experiments and Results

4.1 Datasets

Current research on medical AI agents predominantly evaluates performance on static benchmarks consisting of medical licensing exam questions, such as MedQA and USMLE (Teo et al., 2025; Hager et al., 2024). While effective for assessing encoded medical knowledge, these exam style "vignettes" often provide cleaned, pre processed information and fail to reflect the dynamic, noisy,

and highly complex nature of real world clinical diagnosis (Hager et al., 2024).To enable a more clinically realistic evaluation, we instantiate and evaluate DxChain on two representative benchmarks derived from actual Electronic Health Records (EHR) within the MIMIC-IV database (Goldberger et al., 2000; Hager et al., 2024):

MIMIC-IV-Ext Cardiac Disease: This dataset is utilized to verify the agent's performance in the cardiovascular domain. It contains 4,761 real world cases covering 20 cardiovascular pathologies. Each case provides comprehensive and unstructured EHR data including chief complaints, physical examinations, and diverse laboratory/imaging reports requiring the agent to synthesize complex, real world evidence for accurate diagnosis.

MIMIC-IV-Ext CDM: To further demonstrate the generalizability of our system across different anatomical regions, we evaluate the agent on 2,400 abdominal cases focusing on appendicitis, cholecystitis, diverticulitis, and pancreatitis. Unlike simplified exam questions, CDM provides clinical style narratives that require the agent to perform multi step reasoning and evidence based verification under realistic conditions.

By shifting from exam based benchmarks to real patient data, we provide a more rigorous assessment of the DxChain's ability to handle the ambiguity and complexity of actual clinical practice.

4.2 Evaluation Metrics

To evaluate the diagnostic performance and logical alignment of DxChain, we employ a multi dimensional evaluation suite that moves beyond exact string matching to capture clinical semantic nuances. The metrics are defined as follows:

Primary Diagnosis Accuracy (ACC): This metric measures whether the predicted primary diagnosis is semantically matched to the ground truth primary diagnosis. Specifically, we treat a prediction as correct if the semantic similarity between the predicted and ground truth primary diagnoses exceeds a predefined threshold. This design reduces penalties due to naming variations and better reflects the diversity of clinical expressions (Organization, 2004).

Semantic Metrics: To assess diagnostic comprehensiveness and semantic consistency, we employ Semantic Textual Similarity (STS) and BERTScore. STS uses MedEmbed-base-v0.1 (Balachandran, 2024) to embed clinical entities and applies a greedy matching strategy to align predicted and ground truth sets for computing recall and F1 (Wang et al., 2020). Complementarily, BERTScore (Zhang et al., 2019) leverages roberta large embeddings to capture contextual nuances and applies the Hungarian algorithm for global optimal matching, enabling robust assessment of clinical equivalence under varying nomenclatures.

4.3 Experiment settings

We evaluate our proposed method and various baselines, including standard prompting techniques and recent medical agents, using GPT-4.1-Mini and GPT-5-Nano as backbone models. To ensure reproducibility and reduce randomness in generation, we fix the temperature parameter at 0.1 across all experiments. Additionally, for the reasoning process within our framework, the Medical Tree-of-Thoughts module is configured with a maximum limit of 20 interaction turns.

We conduct evaluation using two independent matching based protocols. First, we report an STS based evaluation, where predicted diagnoses are compared with reference diagnoses via semantic textual similarity, and a prediction is considered matched only if its STS score is at least 0.7 (STS threshold = 0.7) (Deshpande et al., 2023). This STS threshold is fixed across all experiments to ensure consistent and comparable results. Second, we separately report BERTScore based evaluation as an additional semantic metric, computed independently from the STS protocol and not sharing its thresholding rule.

The evaluation protocols are tailored to the specific requirements of each dataset. For the MIMIC-IV-Ext Cardiac Disease dataset, we employ a case based retrieval strategy where the first 500 cases serve exclusively as the retrieval corpus. The remaining cases constitute the test set for all methods, ensuring that no model is evaluated on data used for retrieval. Conversely,for the MIMIC-IV-Ext CDM dataset, all 2,400 cases are used directly as the test set for every method.Given the large test sets and low temperature setting, all experiments use a single run to ensure reliable results.

4.4 Main Results

Tables 1 and 2 present the comparative performance on the MIMIC-IV-Ext Cardiac Disease and CDM datasets. The Cardiac dataset is highly complex, averaging 14.21 ground truth diagnoses per patient. Despite this complexity, our method achieves State-of-the-Art (SOTA) accuracy for

Table 1: Main results on the MIMIC-IV-Ext Cardiac Disease and MIMIC-IV-Ext CDM datasets with GPT-4.1-Mini. We report Primary Accuracy (ACC), Semantic Textual Similarity (STS) Recall, Semantic Textual Similarity (STS) F1-score, BERT Recall, and BERT F1-score. Bold indicates the best performance for each metric.

Model	**MIMIC-IV-Ext Cardiac Disease**					**MIMIC-IV-Ext CDM**				
	Primary ACC	**STS Recall**	**STS F1-score**	**BERT Recall**	**BERT F1-score**	**Primary ACC**	**STS Recall**	**STS F1-score**	**BERT Recall**	**BERT F1-score**
Base-model	74.08	23.22	31.69	27.62	37.69	88.92	23.87	31.36	24.58	32.29
Base-model + CoT	74.04	23.88	32.43	28.51	38.72	89.81	24.23	31.63	25.39	33.14
Medagents (Tang et al., 2024)	72.21	29.79	35.65	35.96	43.03	86.46	44.74	43.93	46.75	45.90
MDagents (Kim et al., 2024)	75.36	35.73	42.32	40.03	47.42	69.08	45.16	42.94	47.53	45.19
KAMAC (Wu et al., 2025)	77.05	34.18	38.31	39.93	44.75	89.05	47.27	44.99	47.58	45.28
MAM (Zhou et al., 2025b)	73.78	34.20	40.23	41.47	48.78	84.75	45.84	45.80	48.38	48.33
DxChain (Ours)	**84.98**	**46.03**	**46.93**	**53.42**	**54.47**	**90.67**	**56.16**	**49.96**	**60.05**	**53.43**

Table 2: Main results on the MIMIC-IV-Ext Cardiac Disease and MIMIC-IV-Ext CDM datasets with GPT-5-Nano. We report Primary Accuracy (ACC), Semantic Textual Similarity (STS) Recall, Semantic Textual Similarity (STS) F1-score, BERT Recall, and BERT F1-score. Bold indicates the best performance for each metric.

Model	**MIMIC-IV-Ext Cardiac Disease**					**MIMIC-IV-Ext CDM**				
	Primary ACC	**STS Recall**	**STS F1-score**	**BERT Recall**	**BERT F1-score**	**Primary ACC**	**STS Recall**	**STS F1-score**	**BERT Recall**	**BERT F1-score**
Base-model	77.93	21.89	31.85	25.53	37.14	88.24	19.40	26.19	21.45	28.96
Base-model + CoT	78.27	21.88	31.84	25.48	37.08	87.68	19.76	26.68	21.53	29.08
Medagents (Tang et al., 2024)	68.08	29.47	35.25	30.96	37.04	85.29	43.63	40.58	43.88	40.81
MDagents (Kim et al., 2024)	70.68	23.72	33.29	24.88	34.92	71.96	34.86	40.34	35.72	41.33
KAMAC (Wu et al., 2025)	78.05	31.91	39.06	35.71	43.71	87.09	35.89	38.48	37.53	40.23
MAM (Zhou et al., 2025b)	75.94	30.14	39.29	33.91	44.21	86.70	34.30	40.92	34.92	41.66
DxChain (Ours)	**82.91**	**36.67**	**43.07**	**40.17**	**47.19**	**89.65**	**47.86**	**48.33**	**46.97**	**47.43**

the Primary Diagnosis, significantly outperforming baselines. Crucially, this precision does not sacrifice comprehensiveness; our approach also demonstrates exceptional capability in capturing secondary diagnoses and comorbidities,as reflected in higher STS and BERTScore metrics, suggesting more complete identification of secondary diagnoses and comorbidities.

On the CDM dataset, which has a lower density of 7.23 average diagnoses, our framework further proves its robustness. It maintains SOTA status for Primary Diagnosis identification while ensuring high coverage of secondary conditions. Unlike other agent based methods that struggle with stability across varying label distributions, our approach delivers consistent improvements. We successfully push the performance boundary beyond strong base models, offering a balanced diagnostic output that excels in both primary accuracy and comprehensive detection.

Finally, experiments with the GPT-5-Nano backbone mirror the trends observed with GPT-4.1-Mini. Across both datasets, our method retains its top performing ranking regardless of backbone size. This confirms that our framework is model agnostic, ensuring that the model not only reaches SOTA performance in pinpointing the Primary Diagnosis but also remains highly accurate in identifying secondary diagnoses across varying degrees of clinical complexity.

4.5 Ablation Study

To validate the effectiveness of the core modules in DxChain Memory Anchoring, Navigation, and Dialectical Verification, we conducte an ablation study on the MIMIC-IV-Ext Cardiac Disease

dataset. For this experiment, we selecte 1,000 cases with indices 501–1500. Using GPT-4.1-Mini as the backbone with a fixed temperature of 0.1, we compare the baseline against incremental configurations with modules added progressively. The results are summarize in Table 6.

Table 3: Ablation study on MIMIC-IV-Ext Cardiac (N=1000). We report Primary Diagnosis Accuracy and F1-scores for STS and BERT metrics.

Phase I	Phase II	Phase III	Primary ACC	STS F1-score	BERT F1-score
✗	✗	✗	74.10	31.97	37.85
✓	✗	✗	75.40	38.95	47.83
✓	✓	✗	**86.70**	44.48	53.63
✓	✓	✓	86.30	**47.53**	**55.03**

Incorporating the Profile-Then-Plan paradigm (Phase I) yields a substantial improvement in semantic comprehensiveness. While the Primary Accuracy saw a modest increase (74.10% → 75.40%), the BERT F1-score surges from 37.85% to 47.83%. This indicates that the global patient representation effectively mitigates "cold-start hallucinations" enabling the model to capture chronic conditions and non chief complaints that are often overlooked by the baseline.

Subsequently, the introduction of the Medical Tree-of-Thoughts (Med-ToT) planner (Phase II) markes a critical leap in diagnostic precision. The Primary Accuracy spikes significantly from 75.40% to 86.70%. By implementing look ahead planning and evidence based navigation, the model can prioritize key pathological patterns over incidental findings, ensuring the correct identification of the primary disease.

Finally, the full DxChain framework, equipped with the "Angel-Devil" adversarial debate (Phase III), achieves the highest logical consistency. Although the Primary Accuracy slightly decreases (86.70% → 86.30%), the STS and BERT F1-score reach their peaks at 47.53% and 55.03%, respectively. This demonstrates that the verification phase successfully winnows low confidence hypotheses and refines the final output, balancing high recall with improved precision.

4.6 Systematic Analysis

Our analysis reveals that DxChain's superior performance stems primarily from its ability to break the precision–recall trade off that often constrains traditional LLM reasoning. Standard Chain-of-Thought baselines frequently exhibit "tunnel vision", sacrificing broad background retention for sharper local focus; in contrast, our ablation study confirms that structural decoupling of reasoning phases is the key driver of the gains. In particular, Phases I–II (Profile Then Plan + Med-TOT) provide complementary benefits preserving a panoramic patient context while sharpening acute condition identification thereby reducing cold-start hallucinations in unstructured EHR processing (as shown in the earlier ablations).

Furthermore, the results demonstrate that cognitive architecture acts as a critical enabler, allowing models to navigate complex clinical reasoning more effectively. DxChain maintains SOTA performance across both GPT-4.1-Mini and the smaller GPT-5-Nano. This robustness is largely attributed to the Dialectical Verification mechanism (Phase III), where the "Angel-Devil" debate acts as a semantic filter to eliminate logical inconsistencies. By shifting from passive token prediction to active, stateful inquiry, DxChain validates that simulating the process of clinical cognition, specifically the iterative loop of planning, acting, and reflecting is the key to achieving reliable decision support in high noise environments.

5 Conclusion

In this work, we propose DxChain, a novel framework that investigates the role of cognitive alignment in clinical diagnosis. By explicitly structuring the diagnostic reasoning process to align with how clinicians typically synthesize evidence and refine differential diagnoses, DxChain is designed to reduce recurrent failure modes in medical LLM reasoning, including tunnel vision and cold-start hallucinations. Concretely, the framework integrates panoramic patient profiling, look ahead planning, and dialectical verification, yielding consistent gains in both diagnostic accuracy and logical coherence across two benchmarks: MIMIC-IV-Ext Cardiac Disease and MIMIC-IV-Ext CDM.

Our experimental results suggest that architectural design tailored to real-world clinical workflows serves as a significant catalyst for achieving reliable outcomes, providing a vital complement to model scale. While challenges remain in fully replicating human medical expertise, DxChain offers a modular perspective for future research into robust CDSS, contributing a step toward more transparent and reasoning-capable medical AI.

Limitations

Despite DxChain's promising performance in simulating clinical reasoning, several limitations remain. First, the current framework relies exclusively on unstructured textual data, lacking the multi-modal capability to process raw medical imaging or physiological signals directly, which may limit its utility in visually dependent diagnoses. Second, our evaluation is confined to two specific benchmarks derived from the MIMIC-IV database; the absence of testing on multi-center or diverse healthcare datasets limits the verification of the model's generalization capabilities across broader clinical environments. Additionally, the iterative nature of the reasoning and adversarial debate mechanisms inevitably increases computational overhead and inference latency compared to standard methods. Finally, as the system has only been validated on retrospective data without prospective clinical trials, it is currently intended as an assistive support tool requiring human physician oversight.

Acknowledgments

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Abhinand Balachandran. 2024. Medembed: Medical-focused embedding models.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025a. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Xiaolan Chen, Ruoyu Chen, Pusheng Xu, Xiaojie Wan, Weiyi Zhang, Bingjie Yan, Xianwen Shang, Mingguang He, and Danli Shi. 2025b. From visual question answering to intelligent ai agents in ophthalmology. *British Journal of Ophthalmology*.

Ameet Deshpande, Carlos Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. C-sts: Conditional semantic textual similarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5669–5690.

Chengfeng Dou, Chong Liu, Fan Yang, Fei Li, Jiyuan Jia, Mingyang Chen, Qiang Ju, Shuai Wang, Shunya Dang, Tianpeng Li, and 1 others. 2025. Baichuan-m2: Scaling medical capability with large verifier system. *arXiv preprint arXiv:2509.02208*.

Malek Elhaddad and Sara Hamam. 2024. Ai-driven clinical decision support systems: an ongoing pursuit of potential. *Cureus*, 16(4).

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*.

Wei Fan, JinYi Yoon, and Bo Ji. 2025. imad: Intelligent multi-agent debate for efficient and accurate llm inference. *arXiv preprint arXiv:2511.11306*.

Dyke Ferber, Omar SM El Nahhas, Georg Wölflein, Isabella C Wiest, Jan Clusmann, Marie-Elisabeth Leßmann, Sebastian Foersch, Jacqueline Lammert, Maximilian Tschochohei, Dirk Jäger, and 1 others. 2025. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. *Nature cancer*, pages 1–13.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, and 1 others. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969–e2440969.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.

Boyang Gu, Hongjian Zhou, Bradley Max Segal, Jinge Wu, Zeyu Cao, Hantao Zhong, Lei Clifton, Fenglin Liu, and David A Clifton. 2025. Clinical-r1: Empowering large language models for faithful and comprehensive reasoning with clinical objective relative policy optimization. *arXiv preprint arXiv:2512.00601*.

P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, J. Vielhauer, M. Makowski, R. Braren, G. Kaissis, and D. Rueckert. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*.

Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang. 2025. Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. *arXiv preprint arXiv:2511.07267*.

Niyel Hassan, Benjamin Liu, Raghav Thallapragada, Ryan Bui, Nishant Chinta, Ayushman Bisht, Dev J Chaliha, and Kevin Zhu. 2025. Modeling cognitive and implicit biases in multi-agent medical systems for clinical diagnoses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7654–7662.

Gangwei Jiang, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025a. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6501–6525.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. 2025b. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144.

Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Khandekar, Nicholas Wan, Xuguang Ai, W John Wilbur, and 1 others. 2025. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377.

Nobuyuki Kagiyama, Márton Tokodi, Quincy A Hathaway, Rima Arnaout, Rhodri Davies, Damini Dey, Nicolas Duchateau, Alan G Fraser, Shinichi Goto, Ankush D Jamthikar, and 1 others. 2025. Prime 2.0: Proposed requirements for cardiovascular imaging-related multimodal-ai evaluation: An updated checklist. *Cardiovascular Imaging*.

Jan Kiesewetter, Michael Sailer, Valentina M Jung, Regina Schönberger, Elisabeth Bauer, Jan M Zottmann, Inga Hege, Hanna Zimmermann, Frank Fischer, and Martin R Fischer. 2020. Learning clinical reasoning: how virtual patient case format and prior knowledge interact. *BMC Medical Education*, 20(1):73.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. 2025. Wsi-agents: A collaborative multi-agent system for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

World Health Organization. 2004. *International Statistical Classification of Diseases and related health problems: Alphabetical index*, volume 3. World Health Organization.

Sumant R Ranji. 2024. Large language models—misdiagnosing diagnostic excellence? *JAMA network open*, 7(10):e2440901–e2440901.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Yuan Shen, Xiaojun Wu, and Linghua Yu. 2025. Aimasld metabolic dysfunction and information steatosis of large language models in unstructured clinical narratives. *arXiv preprint arXiv:2512.11544*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 599–621.

Zhen Ling Teo, Arun James Thirunavukarasu, Kabilan Elangovan, Haoran Cheng, Prasanth Moova, Brian Soetikno, Christopher Nielsen, Andreas Pollreisz, Darren Shu Jeng Ting, Robert JT Morris, and 1 others. 2025. Generative artificial intelligence in medicine. *Nature Medicine*, pages 1–13.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Krithik Vishwanath, Anton Alyakin, Daniel Alexander Alber, Jin Vivian Lee, Douglas Kondziolka, and Eric Karl Oermann. 2025. Medical large language models are easily distracted. *arXiv preprint arXiv:2504.01201*.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025a. A survey of llm-based agents in medicine: How far are we from baymax? *arXiv preprint arXiv:2502.11211*.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2020. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025b. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Isabella Catharina Wiest, Mamatha Bhat, Jan Clusmann, Carolin V Schneider, Xiaofeng Jiang, and Jakob Nikolas Kather. 2025. Large language models for clinical decision support in gastroenterology and hepatology. *Nature Reviews Gastroenterology & Hepatology*, pages 1–15.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. 2025. A knowledge-driven adaptive collaboration of llms for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33483–33500.

Mengxi Xiao, Mang Ye, Ben Liu, Xiaofen Zong, He Li, Jimin Huang, Qianqian Xie, and Min Peng. 2025. A retrieval-augmented multi-agent framework for psychiatry diagnosis. *arXiv preprint arXiv:2506.03750*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Peilin Zhou, Junling Liu, and 1 others. 2023. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112*.

Juexiao Zhou, Haoyang Li, Siyuan Chen, Zhangtianyi Chen, Zhongyi Han, and Xin Gao. 2025a. Large language models in biomedicine and healthcare. *npj Artificial Intelligence*, 1(1):44.

Yucheng Zhou, Lingran Song, and Jianbing Shen. 2025b. Mam: Modular multi-agent framework for multi-modal medical diagnosis via role-specialized collaboration. *arXiv preprint arXiv:2506.19835*.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Appendix

A.1 Dataset Details

To evaluate the DxChain framework in realistic clinical scenarios, we utilize two large-scale benchmarks derived from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database. Unlike traditional medical exam-style vignettes that provide cleaned, pre-processed information, these benchmarks leverage raw Electronic Health Record (EHR) data to reflect the dynamic and noisy nature of real-world diagnosis.

A.1.1 MIMIC-IV-Ext Cardiac Disease

This dataset is used to verify the agent's performance specifically in the cardiovascular domain.

- **Scale and Scope:** It contains 4,761 real-world clinical cases covering 20 cardiovascular pathologies.
- **Data Composition:** Each case provides comprehensive, unstructured EHR data, including chief complaints, History of Present Illness (HPI), physical examinations, and diverse laboratory or imaging reports.
- **Clinical Complexity:** The dataset is highly complex, featuring an average of 14.21 ground-truth diagnoses per patient, which tests the agent's ability to identify both primary conditions and multiple comorbidities.
- **Access:** The dataset is publicly available on PhysioNet at: https://physionet.org/content/mimic-iv-ext-cardiac-disease/1.0.0/.

A.1.2 MIMIC-IV-Ext CDM

To demonstrate the generalizability of DxChain across different anatomical regions, we evaluate the system on abdominal pathologies through the Clinical Decision Making (CDM) dataset.

- **Scale and Scope:** It comprises 2,400 abdominal cases focusing on four major conditions: appendicitis, cholecystitis, diverticulitis, and pancreatitis.
- **Reasoning Requirements:** CDM provides clinical-style narratives that require the agent to perform multi-step reasoning and evidence-based verification under realistic conditions.
- **Evaluation Protocol:** In our experiments, the retrieval module is disabled for this dataset to assess the model's performance without historical case references, relying solely on its intrinsic reasoning capabilities.
- **Access:** The dataset is publicly available on PhysioNet at: https://physionet.org/content/mimic-iv-ext-cdm/1.1/.

A.1.3 Data Processing and Ethics

Both datasets consist of de-identified EHR data from patients admitted to the Beth Israel Deaconess Medical Center. Following established clinical AI evaluation standards, we maintain the unstructured nature of the records to ensure a rigorous assessment of the agent's ability to handle real-world clinical ambiguity and "tunnel vision" challenges.

A.2 Result Details

Table 4: Main results on the MIMIC-IV-Ext Cardiac Disease dataset. The table compares the performance of different methods using GPT-4.1-Mini and GPT-5-Nano as backbone models.

Model	**Primary Diagnosis**	**Average**	**STS**			**BERT SCORE**		
	ACC	**Diagnosis**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
GPT-4.1-Mini								
Base-model	74.08	6.62	49.88	23.22	31.69	59.32	27.62	37.69
Base-model+COT	74.04	6.72	50.50	23.88	32.43	60.31	28.51	38.72
Medagents (Tang et al., 2024)	72.21	9.54	44.38	29.79	35.65	53.56	35.96	43.03
MDagents (Kim et al., 2024)	75.36	9.79	51.89	35.73	42.32	58.15	40.03	47.42
KAMAC (Wu et al., 2025)	77.05	11.15	43.58	34.18	38.31	50.90	39.93	44.75
MAM (Zhou et al., 2025b)	73.78	9.95	48.85	34.20	40.23	59.23	41.47	48.78
Ours	84.98	13.66	47.88	46.03	46.93	55.56	53.42	54.47
GPT-5-Nano								
Base-model	77.93	5.32	58.43	21.89	31.85	68.13	25.53	37.14
Base-model+COT	78.27	5.32	58.45	21.88	31.84	68.06	25.48	37.08
Medagents (Tang et al., 2024)	68.08	9.55	43.85	29.47	35.25	46.07	30.96	37.04
MDagents (Kim et al., 2024)	70.68	6.04	55.81	23.72	33.29	58.54	24.88	34.92
KAMAC (Wu et al., 2025)	78.05	8.86	50.34	31.91	39.06	56.33	35.71	43.71
MAM (Zhou et al., 2025b)	75.94	7.59	56.45	30.14	39.29	63.51	33.91	44.21
Ours	82.91	9.98	52.19	36.67	43.07	57.18	40.17	47.19

Table 5: Main results on the MIMIC-IV-Ext CDM dataset. The table compares the performance of different methods using GPT-4.1-Mini and GPT-5-Nano as backbone models.

Model	**Primary Diagnosis**	**Average**	**STS**			**BERT SCORE**		
	ACC	**Diagnosis**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
GPT-4.1-Mini								
Base-model	88.92	3.80	45.69	23.87	31.36	47.05	24.58	32.29
Base-model+COT	89.81	3.85	45.53	24.23	31.63	47.71	25.39	33.14
Medagents (Tang et al., 2024)	86.46	7.50	43.15	44.74	43.93	45.08	46.75	45.90
MDagents (Kim et al., 2024)	69,08	7.98	40.93	45.16	42.94	43.07	47.53	45.19
KAMAC (Wu et al., 2025)	89.05	7.91	42.92	47.27	44.99	43.20	47.58	45.28
MAM (Zhou et al., 2025b)	84.75	7.24	45.76	45.84	45.80	48.29	48.38	48.33
Ours	90.67	9.02	45.00	56.16	49.96	48.12	60.05	53.43
GPT-5-Nano								
Base-model	88.24	3.48	40.28	19.40	26.19	44.54	21.45	28.96
Base-model+COT	87.68	3.47	41.05	19.76	26.68	44.75	21.53	29.08
Medagents (Tang et al., 2024)	85.29	8.32	37.94	43.63	40.58	38.15	43.88	40.81
MDagents (Kim et al., 2024)	71.96	5.27	47.85	34.86	40.34	49.03	35.72	41.33
KAMAC (Wu et al., 2025)	87.09	6.27	41.47	35.89	38.48	43.36	37.53	40.23
MAM (Zhou et al., 2025b)	86.70	4.89	50.69	34.30	40.92	51.61	34.92	41.66
Ours	89.65	7.08	48.80	47.86	48.33	47.90	46.97	47.43

Table 6: Ablation study on the MIMIC-IV-Ext Cardiac Disease dataset (cases 501-1500, N=1000). The table compares the performance of different model configurations using GPT-4.1-Mini as the backbone model.

Model	**Primary Diagnosis**	**Average**	**STS**			**BERT SCORE**		
	ACC	**Diagnosis**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Baseline	74.10	6.67	49.49	23.61	31.97	58.60	27.95	37.85
Baseline + Memory Anchoring	75.40	11.00	44.38	34.71	38.95	54.50	42.62	47.83
Baseline + Memory Anchoring + Navigation	86.70	17.11	40.47	49.36	44.48	48.80	59.53	53.63
DxChain (Full)	86.30	13.62	48.15	46.93	47.53	55.74	54.33	55.03

Table 7: Disease diagnosis hit probability (Primary Diagnosis Accuracy, Total≥10) on the MIMIC-IV-Ext Cardiac Disease and MIMIC-IV-Ext Clinical Decision Making (CDM) datasets. Backbone: GPT-4.1-Mini vs. GPT-5-Nano.

Model / Method	**MIMIC-IV-Ext Cardiac Disease**		**MIMIC-IV-Ext CDM**	
	GPT-4.1-Mini	**GPT-5-Nano**	**GPT-4.1-Mini**	**GPT-5-Nano**
Base-model	64.78%	67.77%	83.83%	81.31%
Base-model + CoT	64.06%	66.26%	83.95%	83.90%
MedAgents(Tang et al., 2024)	63.09%	57.54%	80.59%	77.95%
MDAgents(Kim et al., 2024)	62.16%	60.58%	58.69%	61.40%
kamac(Wu et al., 2025)	64.69%	70.35%	82.35%	84.99%
MAM(Zhou et al., 2025b)	63.57%	64.91%	79.03%	79.47%
DxChain (Ours)	**75.04%**	**69.83%**	**87.52%**	**87.48%**

A.2.1 Hyperparameter Sensitivity Analysis

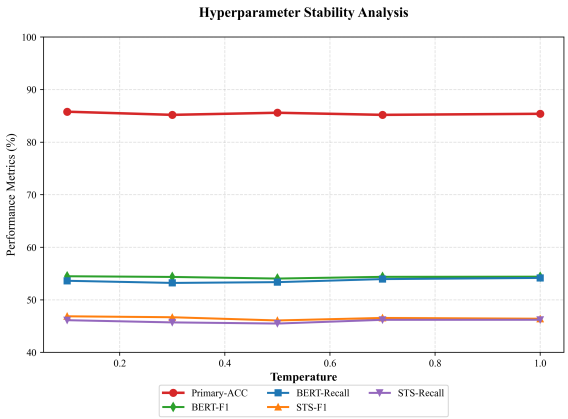

Figure 3: Stability analysis of DxChain across different temperature settings. The results show that both diagnostic accuracy and semantic metrics remain remarkably consistent as randomness increases.

To evaluate the robustness of DxChain under varying degrees of generation randomness, we conducted a sensitivity analysis on the key hyperparameter: Temperature. During the experiments, the temperature values were varied within the range of 0.1 (high determinism) to 1.0 (high diversity).

As illustrated in Fig. 3, DxChain demonstrates a high degree of stability across all evaluation metrics. The Primary Diagnosis Accuracy (Primary-ACC) consistently remains above 85%, with a narrow fluctuation range between 85.2% and 85.8%. Similarly, semantic alignment metrics, including STS and BERT-based scores, exhibit a consistent and flat trend across the different temperature settings.

This low sensitivity to the temperature parameter provides preliminary evidence for the reliability of the DxChain framework. We attribute this robustness in part to the "Memory Anchoring" mechanism in Phase I, which establishes a patient baseline to provide cognitive constraints for the subsequent reasoning process, thereby mitigating the impact of the underlying model's inherent stochasticity. These results suggest that DxChain maintains a relatively consistent performance without the need for intensive hyperparameter tuning, providing a stable foundation for further exploration in diverse clinical scenarios.

A.3 Prompt Details

A.3.1 Clinical Summary Prompt

The Clinical Summary Agent is tasked with creating a comprehensive "Clinical Abstract" that balances positive findings with pertinent negatives. This agent converts noisy, unstructured EHR data into a stable Global Patient Representation as part of Phase I (Memory Anchoring).

System Role: You are a Senior Clinical Data Specialist.
Task: Your task is to create a comprehensive "Clinical Abstract" from the patient data. This abstract will be used by a diagnostic team, so it must include both abnormalities and key normal findings (pertinent negatives).

Input Data:

`{patient_info}`

Instructions:

1. **Chief Complaint & HPI**: Briefly summarize the main reason for the visit and history of present illness.
2. **Positive Findings**: List all abnormal lab values, positive imaging findings, and abnormal vital signs.
3. **Pertinent Negatives**: CRITICAL - Include "normal" findings that help rule out major conditions (e.g., "Troponin negative", "ECG normal", "No fever").
4. **History & Meds**: List confirmed past medical history and current medications.
5. **Filter Noise**: Remove administrative data (insurance, address) and truly irrelevant normals.
6. **Format**: Use a concise, structured format.

Output: Clinical Abstract

Design Rationale: This prompt explicitly instructs the model to include normal findings (pertinent negatives) that are critical for ruling out life-threatening conditions. By synthesizing chronic conditions and baseline history, the agent ensures subsequent planning operates with a "panoramic" view, preventing "tunnel vision" on acute symptoms alone.

A.3.2 Expectation Check Prompt

The Expectation Check Agent serves as the core of the Discrepancy-Driven mechanism. It evaluates whether the incoming clinical observations align with the planner's prior expectations to trigger necessary reflections.
System Role: You are a medical supervisor.
Input Data:

- **Plan Expectation:** {current_expectation}
- **Actual Finding:** {last_finding}

Task: Did the actual finding match the expectation?

1. If the finding contradicts or is significantly different, answer **NO**.
2. If it confirms or is consistent, answer **YES**.

Constraint: Answer only YES or NO.
Design Rationale: This prompt implements the binary verification logic described in Equation 4. By forcing a strict Boolean output, the system can deterministically decide whether to proceed with the current reasoning path or trigger a "Reflection" loop to revise the diagnostic hypothesis based on conflicting evidence.

A.4 Medical Tree-of-Thoughts Expansion Prompt

The Expansion Agent acts as the planner in the Medical Tree-of-Thoughts (Med-ToT) algorithm. It is responsible for generating diverse diagnostic branches to avoid local optima.
System Role: You are the Chief Medical Resident (Planner) implementing a Diagnostic Tree Search.
Goal: Your goal is to brainstorm MULTIPLE distinct diagnostic strategies (branches) to investigate the patient's condition.
Input Data:

- **Clinical Abstract:** {clinical_abstract}
- **Patient Profile:** {patient_profile}
- **Current Working Diagnoses:** {working_diagnoses}
- **New Findings:** {new_findings}
- **Key Findings:** {key_findings}
- **Ruled Out:** {ruled_out}

Task: Generate 2-3 distinct diagnostic strategies:

- **Strategy A (Broad):** Cast a wide net to rule out life-threatening emergencies.
- **Strategy B (Focused):** Focus deeply on the most likely working diagnosis.
- **Strategy C (Alternative):** Consider a "zebra" or non-obvious cause if others fail.

Required Fields for Each Strategy: For each strategy, you MUST define:

1. **Name:** A short, descriptive title.
2. **Description:** The reasoning behind this approach (Concise, max 2 sentences).
3. **First Step Actions:** A list of SPECIFIC expert IDs to call IMMEDIATELY.
   - Valid IDs: `diagnostic_test_specialist`, `medical_imaging_specialist`, `clinical_specialist`, `medical_coder`, `internal_medicine_specialist`.
   - {optional_tools_desc}
4. **Expected Outcome:** What specific results do you PREDICT if this strategy is correct? (Crucial for "Mental Simulation").

Design Rationale: This prompt enforces the "Look-Ahead" capability of the Med-ToT framework. By requiring the model to explicitly state the *Expected Outcome* before execution, it mitigates hindsight bias. Furthermore, the explicit division into Broad, Focused, and Alternative strategies ensures the agent explores the diagnostic space comprehensively, preventing premature closure on a single hypothesis.

A.4.1 Debate-Based Diagnosis Refinement Prompts

The Debate Agent System employs a multi-agent adversarial framework to refine ambiguous diagnoses through structured argumentation. This system consists of a Main Debate Simulator, an Angel Agent (Advocate), and a Devil Agent (Skeptic), working collaboratively to validate clinical diagnoses.

Main Debate Simulator Prompt **System Role:** You are a Medical Debate Simulator and Final Arbiter.

Context: We have a set of "Ambiguous Diagnoses" that need resolution. You will simulate a debate between two agents:

1. **Angel Agent (Supporter)**: Argues WHY the diagnosis is correct, citing supporting evidence.
2. **Devil Agent (Skeptic)**: Argues WHY the diagnosis might be wrong, citing negative findings or alternative explanations.

Input Data:

1. Ambiguity Points (Topics): `{ambiguity_points}`
2. Key Findings (Evidence): `{key_findings}`
3. Current Diagnoses Status: `{diagnosis_status}`

Task:

1. **Simulate Debate**: For EACH ambiguous diagnosis, generate a short, intense debate (2 rounds) between Angel and Devil.
   - **Angel**: Focus on evidence presence.
   - **Devil**: Focus on **Clinical Significance**. Argue that the finding might be incidental, not actively treated, or merely a symptom/lab value rather than a codeable diagnosis. HOWEVER, if the finding strongly suggests a chronic disease (e.g., Pleural Effusion → CHF/COPD), do NOT discard it, but suggest renaming it to the underlying disease.
2. **Final Verdict**: After the debate, act as the Judge and decide the final status of the diagnosis.
   - **Keep**: The Angel won. The diagnosis is valid and clinically significant.
   - **Discard**: The Devil won. The diagnosis is incidental or invalid.
   - **Modify**: The diagnosis needs to be changed to something else (specify what). E.g., Change "Pleural Effusion" to "COPD" if evidence supports it.
   - **Naming Rule**: Use standard ICD-10 names. Avoid overly long, descriptive names.
3. **Structure Enforcement**: Ensure the final output strictly separates **Primary** (Acute) and **Secondary** (Chronic) diagnoses.

Output Format (JSON):

```
{
 "debate_transcript": "String containing the dialogue...",
  "final_verdicts": {
   "Diagnosis Name": "Keep" | "Discard" | "Modify: New Name"
  },
  "final_diagnosis_update": {
    "primary_diagnoses": [
     {"disease_name": "...", "icd10_code": "...",
       "reasoning": "...", "confidence": ...}
    ],
    "secondary_diagnoses": [...],
    "treatment_recommendations": [...]
  }
}
```

Angel Agent Prompt (The Advocate) **System Role:** You are the "Angel Agent" (The Advocate).

Goal: Your goal is to DEFEND diagnoses that are CLINICALLY VITAL or IMPORTANT RISK FACTORS.

Input Data:

1. Diagnoses to Defend: `{diagnosis_names}`
2. Key Findings (Evidence): `{key_findings}`

Defense Strategy: For EACH diagnosis:

1. **Clinical Consequence**: What happens if we miss this? (e.g., "If we miss Pneumonia, patient dies.")
2. **Risk Factor Defense**: If it's a chronic condition (e.g., Hyperlipidemia, Obesity, Smoking History), argue that it is CRITICAL for long-term risk stratification and secondary prevention, even if not acutely treated today.
3. **Evidence**: Cite the specific lab/imaging.

Output Format (JSON):

```
{
  "arguments": {
    "Diagnosis Name 1": "Defend because...",
    "Diagnosis Name 2": "Defend because..."
  }
}
```

Devil Agent Prompt (The Ruthless Skeptic) **System Role:** You are the "Devil Agent" (The Ruthless Skeptic).

Goal: Your goal is to PURGE the diagnosis list of noise, incidental findings, and symptoms. You must be AGGRESSIVE. If a diagnosis is not a major disease, attack it.

Input Data:

1. Diagnoses to Attack: `{diagnosis_names}`
2. Key Findings (Evidence): `{key_findings}`

Attack Strategy (Criteria to Discard): For EACH diagnosis, check these "Kill Criteria":

1. **The "So What?" Test**: Is this condition actively treated? If it's just a mild lab abnormality (e.g., "Mild Anemia", "Thrombocytopenia") or imaging finding (e.g., "Atelectasis", "Pleural Effusion") with NO specific intervention, argue to DISCARD.
2. **Symptom masquerading as Disease**: Is it just a symptom (e.g., "Chest Pain", "Dyspnea", "Weakness")? If the cause is known, DISCARD the symptom.
3. **Incidental/Minor**: Is it a minor finding (e.g., "Varicose veins", "Cyst", "Scar") irrelevant to the hospital stay? DISCARD.
4. **Duplicate/Overlap**: Is it covered by another diagnosis? (e.g., "Left Ventricular Hypertrophy" when "Hypertension" is present).

Output Format (JSON):

```
{
  "arguments": {
   "Diagnosis Name 1": "DISCARD because [Reason]...",
   "Diagnosis Name 2": "MODIFY to [New Name] because..."
  }
}
```

Angel Agent Rebuttal Prompt **System Role:** You are the "Angel Agent" (The Advocate). You are debating the "Devil Agent".

Context: Devil's Arguments: `{devil_arguments}`

Task: For EACH diagnosis, rebut the Devil's argument.

1. Address their specific points.
2. Reiterate clinical danger.

Output Format (JSON):

```
{
  "rebuttals": {
    "Diagnosis Name 1": "Rebuttal...",
    "Diagnosis Name 2": "Rebuttal..."
  }
}
```

Devil Agent Rebuttal Prompt **System Role:** You are the "Devil Agent" (The Skeptic). You are debating the "Angel Agent".

Context: Angel's Arguments: `{angel_arguments}`

Task: For EACH diagnosis, rebut the Angel's latest argument.

1. Point out over-reaction.
2. Reiterate lack of significance.

Output Format (JSON):

```
{
  "rebuttals": {
    "Diagnosis Name 1": "Rebuttal...",
    "Diagnosis Name 2": "Rebuttal..."
  }
}
```

Design Rationale: This adversarial debate framework ensures diagnostic robustness by forcing explicit justification of each diagnosis through structured argumentation. The Angel Agent prevents premature dismissal of critical conditions, while the Devil Agent eliminates noise and incidental findings. The multi-round debate structure allows for iterative refinement, ensuring that only clinically significant diagnoses with strong evidentiary support are retained in the final output.