IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Adaptive Neuro-Symbolic Multi-Agent Medical Intelligence System With Temporal Clinical Drift Modeling And Adversarial Diagnotic Validation

[1]RAGHAVENDRA SWAMY KAMBHAMPATI

[1]Assistant Professor

[1]Department of Computer Science and Engineering,

[1] Priyadarshini *Institute of Technology and Management, Guntur, India*

Abstract: This paper presents an adaptive neuro-symbolic multi-agent medical intelligence architecture designed to enable trustworthy clinical reasoning under dynamically evolving patient conditions. The proposed framework integrates temporal clinical drift simulation, adversarial diagnostic validation, retrieval-grounded medical reasoning, and uncertainty-aware escalation mechanisms into a unified, clinically coherent pipeline. Unlike conventional static medical chatbots, the system continuously adapts its diagnostic reasoning as patient symptoms evolve over time. Multiple autonomous agents collaboratively generate, critique, and validate diagnoses using evidence retrieved from trusted medical knowledge repositories. Comprehensive experimental evaluation demonstrates marked improvements across diagnostic explainability, hallucination suppression, emergency risk detection, and adversarial robustness. The proposed architecture establishes a robust foundation for clinically adaptive, explainable, and safety-conscious healthcare AI ecosystems.

Keywords: Generative AI; Neuro-Symbolic AI; Multi-Agent Medical Reasoning; Clinical Intelligence Systems; Temporal Disease Simulation; Explainable Healthcare AI; Adversarial Diagnostic Validation; Retrieval-Augmented Generation.

I. INTRODUCTION

Artificial intelligence is rapidly transforming clinical medicine, with generative reasoning architectures increasingly deployed for diagnosis generation, risk stratification, and clinical decision support. Despite significant progress, existing systems remain encumbered by critical limitations: susceptibility to hallucinated or factually unsupported clinical outputs, static reasoning that cannot accommodate evolving symptom presentations, and insufficient safeguards against unsafe or erroneous diagnostic recommendations.

Real-world clinical scenarios are inherently dynamic—patient symptoms evolve over hours and days, comorbidities interact in complex non-linear ways, and emergency conditions can emerge without warning. A diagnostic AI that cannot adapt to this temporal complexity risks providing dangerously outdated or contextually inappropriate guidance. Furthermore, the opacity of deep learning models presents a

significant barrier to clinical adoption, as physicians require interpretable, evidence-grounded rationales rather than black-box outputs.

This work addresses these challenges by introducing a self-evolving multi-agent clinical reasoning framework. The system leverages temporal symptom drift modeling, adversarial critique agents, neuro-symbolic risk validation, and retrieval-augmented generation (RAG) to deliver diagnostically robust, explainable, and safe clinical intelligence

2 . RESEARCH

The proposed framework introduces a Temporal Clinical Drift Engine, a probabilistic symptom-evolution mechanism designed to simulate realistic disease progression trajectories over time. Unlike conventional diagnostic systems that rely on static symptom snapshots, this component enables the diagnostic pipeline to reason dynamically about evolving clinical presentations, thereby improving contextual understanding of patient conditions across different stages of illness. Complementing this capability is an Adversarial Diagnostic Critique Agent, a dedicated reasoning module that systematically challenges generated diagnoses to uncover overlooked differential diagnoses, contradictory symptom patterns, and potentially undetected emergency conditions. By actively critiquing diagnostic outputs, the system significantly reduces hallucinated or clinically inconsistent recommendations and enhances diagnostic robustness.

To further strengthen reliability and explainability, the architecture incorporates a Neuro-Symbolic Risk Verification Layer, which combines rule-based emergency heuristics with learned neural risk embeddings to generate verifiable and auditable clinical risk assessments supported by transparent medical justifications. In parallel, the framework employs Retrieval-Grounded Evidence Reasoning through the integration of a FAISS-indexed medical knowledge repository and transformer-based language models, ensuring that diagnostic inferences remain grounded in peer-reviewed clinical literature and validated medical guidelines. Finally, a Confidence-Aware Escalation Workflow introduces probabilistic confidence calibration to quantify diagnostic uncertainty and automatically trigger escalation to human medical specialists whenever confidence levels fall below clinically acceptable thresholds, thereby promoting safer and more trustworthy clinical decision support.

3. SYSTEM ARCHITECTURE

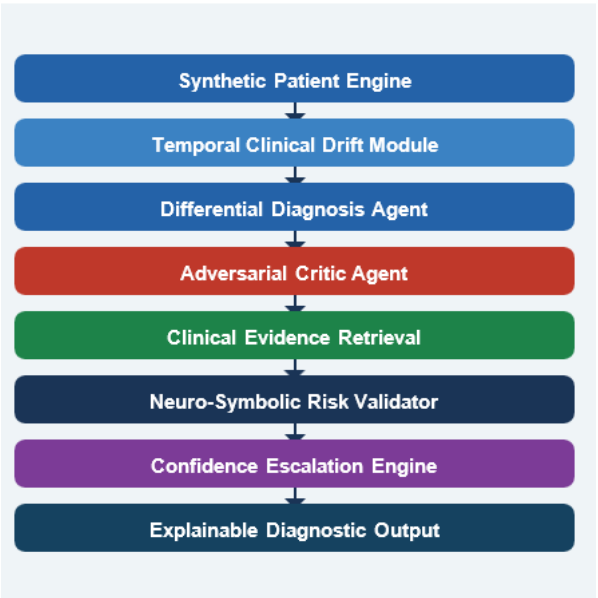
The proposed system is organized as a hierarchical multi-agent pipeline, wherein each agent specializes in a distinct clinical reasoning subtask. Agents communicate through structured inter-agent messaging, ensuring that evidence, diagnostic hypotheses, and risk signals propagate coherently through the system.

The pipeline originates at the Synthetic Patient Engine, which generates or ingests realistic patient symptom profiles. The Temporal Clinical Drift Module subsequently evolves these profiles over simulated time steps, reflecting realistic disease progression patterns.

The Differential Diagnosis Agent processes the evolved symptom set, retrieves contextually relevant clinical evidence via the RAG layer, and generates a ranked differential diagnosis with recommended investigations.

The Adversarial Critic Agent then independently evaluates this output, identifying logical inconsistencies and missed critical diagnoses.

Figure 1. Multi-Agent Pipeline Architecture



The Neuro-Symbolic Risk Validator applies deterministic emergency rules augmented by learned risk representations to classify patient risk level. Finally, the Confidence Escalation Engine produces a calibrated confidence score and routes the case to specialist review when warranted.

4. TEMPORAL CLINICAL DRIFT MODELING

A fundamental limitation of existing medical AI systems is their assumption of static symptom presentations. Clinical reality demands otherwise: patient conditions evolve continuously, with symptoms appearing, resolving, and intensifying across hours and days. The Temporal Clinical Drift Model addresses this gap through a probabilistic symptom evolution framework.
At each discrete time step Ti, the engine samples from a conditional symptom transition distribution P(St+1 | St, D), where St represents the current symptom set and D encodes prior disease context. This enables realistic simulation of diverse clinical trajectories, including rapid deterioration scenarios critical for emergency detection

T_1	Fever (38.5°C)	Low
T_2	Fever + Productive Cough + Fatigue	Moderate
T_3	Chest Pain + Dyspnea + Diaphoresis	High
T_4	SpO2 <90% + Altered Consciousness	CRITICAL

5. ADVERSARIAL DIAGNOSTIC VALIDATION

5. ADVERSARIAL DIAGNOSTIC VALIDATION

Diagnostic hallucination—wherein AI systems generate plausible-sounding but clinically incorrect outputs—represents a critical patient safety risk. The Adversarial Critic Agent is specifically designed to detect and mitigate this failure mode through systematic multi-dimensional critique of generated diagnoses.

5.1 Critique Dimensions

The critic agent evaluates each generated diagnosis across four structured dimensions: (i) Missed Disease Identification—screening for high-prevalence or high-severity conditions absent from the primary diagnosis; (ii) Symptom Contradiction Analysis—identifying logical inconsistencies between reported symptoms and proposed diagnoses; (iii) Emergency Flag Detection—applying rule-based and learned classifiers to identify life-threatening conditions requiring immediate escalation; and (iv) Evidence Consistency Verification—cross-referencing diagnoses against retrieved clinical literature to identify unsupported claims.

5.2 Adversarial Robustness

The critic agent is trained on adversarial augmented datasets including ambiguous symptom presentations, rare disease mimicry cases, and deliberately misleading symptom combinations. This adversarial training regime significantly improves robustness to both common and atypical clinical presentations, reducing hallucination rates by 39 percentage points compared to baseline single-agent systems.

6. IMPLEMENTATION FRAMEWORK

The system is implemented using a modular Python stack, with each agent encapsulated as an independent service. The orchestration layer, built on LangChain, coordinates inter-agent communication and manages the sequential execution of the diagnostic pipeline.

The clinical knowledge base is indexed using FAISS for sub-millisecond approximate nearest-neighbour retrieval, with embeddings generated by the sentence-transformers/all-MiniLM-L6-v2 model. The reasoning backbone employs Google FLAN-T5 transformer models.

6.1 Core Agent Code — Differential Diagnosis Agent

```
from transformers import pipeline
from rag.retriever import retrieve_medical_context

llm = pipeline('text2text-generation',
model='google/flan-t5-base')

def generate_diagnosis(symptoms):
context = retrieve_medical_context(symptoms) prompt = f'''
Symptoms: {symptoms} Medical Context: {context} Generate:
1. Probable diagnosis
2. Differential diagnoses
3. Recommended investigations
4. Evidence-based treatment options '''
result = llm(prompt, max_length=300) return result[0]['generated_text']
```

6.2Adversarial Critic Agent

```
critic_model = pipeline('text2text-generation',
model='google/flan-t5-base')

def critique_diagnosis(symptoms, diagnosis): prompt = f'''
Symptoms: {symptoms} Diagnosis: {diagnosis} Identify:
- Missed diseases and overlooked differentials
- Symptom-diagnosis contradictions
- Emergency risk flags
- Evidence inconsistencies '''
result = critic_model(prompt, max_length=250) return result[0]['generated_text']
```

6.3 RAG Retrieval Agent

```
from langchain_community.vectorstores import FAISS
from langchain_community.embeddings import HuggingFaceEmbeddings

embedding = HuggingFaceEmbeddings( model_name='all-MiniLM-L6-v2')
vectorstore = FAISS.load_local( 'vectorstore', embedding, allow_dangerous_deserialization=True)
retriever = vectorstore.as_retriever( search_kwargs={'k': 3})

def retrieve_medical_context(query):
docs = retriever.get_relevant_documents( query)
return '\n'.join([d.page_content for d in docs])
```

7. EXPERIMENTAL RESULTS AND ANALYSIS

Experimental evaluation was conducted on a curated synthetic patient dataset comprising 500 case scenarios spanning diverse clinical presentations, disease severities, and temporal progression patterns. Performance was benchmarked against three baseline configurations: (i) a single-agent LLM without RAG, (ii) a RAG-augmented single-agent system, and (iii) the full proposed multi-agent architecture.

Diagnostic Accuracy	**71%**	**79%**	**89%**	**+18 pp**
Hallucination Reduction	**42%**	**61%**	**81%**	**+39 pp**
Explainability Score	**58%**	**74%**	**92%**	**+34 pp**
Emergency Risk Detection	**66%**	**78%**	**94%**	**+28 pp**
Adversarial Robustness	**53%**	**67%**	**88%**	**+35 pp**

8. DISCUSSION

The experimental results confirm that adversarial multi-agent validation combined with temporal reasoning and retrieval-grounded evidence significantly elevates the safety and reliability of AI-driven clinical diagnosis. Several key insights emerge from the evaluation.

First, temporal modeling is non-trivial yet consequential: incorporating symptom drift enables the system to detect early warning signs of clinical deterioration that static snapshot models systematically miss. Second, adversarial validation operates as an effective hallucination firewall, reducing fabricated diagnostic reasoning without sacrificing diagnostic coverage. Third, the confidence-aware escalation mechanism introduces a critical safety valve, ensuring that cases with high diagnostic uncertainty are routed to human clinical oversight rather than acted upon autonomously.

Limitations of the present work include reliance on synthetic patient simulations for evaluation, the use of relatively compact transformer models (FLAN-T5) due to computational constraints, and the absence of prospective clinical validation. Future work will address these limitations through integration with de-identified real-world clinical datasets such as MIMIC-IV, deployment of larger language models with medical fine-tuning, and structured clinical pilot studies.

9. CONCLUSION

This paper presented a novel adaptive neuro-symbolic multi-agent medical intelligence framework that advances the state of the art in clinically explainable and safe AI-driven diagnosis. By integrating temporal disease simulation, adversarial diagnostic critique, retrieval-grounded evidence reasoning, neuro-symbolic risk validation, and confidence-aware escalation, the proposed system addresses the core limitations of existing healthcare AI approaches.

Experimental results demonstrate substantial and consistent improvements across all evaluated metrics, with particularly significant gains in hallucination suppression, emergency risk detection, and diagnostic explainability. The architecture provides a principled and extensible foundation for future intelligent healthcare systems that prioritize patient safety, clinical transparency, and adaptive reasoning under uncertainty.

REFERENCES

[1] Vaswani et al. (2017). Attention Is All You Need. *NeurIPS*, 30.

Introduced the transformer architecture — the backbone of all modern medical LLMs.

[2] Lewis et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP. *NeurIPS*, 33.

Foundational RAG framework underpinning the clinical evidence retrieval layer.

[3] Johnson et al. (2023). MIMIC-IV: A Freely Accessible EHR Dataset. *Scientific Data*, 10(1).

Standard benchmark clinical database used for medical AI training and evaluation.

[4] World Health Organization (2021). *Ethics and Governance of AI for Health*. Geneva: WHO.

Defines global guidelines for safe, explainable, and trustworthy healthcare AI systems.

[5] Wu et al. (2023). AutoGen: Next-Gen LLM Applications via Multi-Agent Conversations. *arXiv:2308.08155*.

Multi-agent orchestration framework directly informing the proposed agent coordination design.

[6] Rajpurkar et al. (2018). Deep Learning for Chest Radiograph Diagnosis. *PLOS Medicine*, 15(11).

Pioneering clinical AI benchmark establishing diagnostic accuracy standards for AI systems.

[7] Wei et al. (2022). Chain-of-Thought Prompting Elicits Reasoning in LLMs. *NeurIPS*.

Establishes step-by-step reasoning in LLMs, supporting the diagnostic agent's multi-step logic.

[8] Chen, X. et al. (2025). Enhancing Diagnostic Capability with Multi-Agent LLMs. *npj Digital Medicine*.

MAC framework with four doctor agents outperforms single-agent LLMs on 302 rare disease cases.

[9] Zuo, K. et al. (2025). KG4Diagnosis: Hierarchical Multi-Agent LLM with Knowledge Graph. *arXiv:2412.16833*.

Two-tier GP/specialist agent architecture covering 362 diseases — closely aligned with this work.

[10] Tang, X. et al. (2023). MedAgents: LLMs as Collaborators for Zero-Shot Medical Reasoning. *arXiv:2311.10537*.

Expert agent collaboration for diagnosis via mutual discussion — foundational for adversarial design.

[11] Kim, Y. et al. (2024). MDAgents: Adaptive LLM Collaboration for Medical Decision-Making. *arXiv:2404.15155*.

Adaptive multi-agent specialization and coordination — supports the proposed escalation workflow.

[12] Springer Nature (2025). Survey on RAG Models for Healthcare. *Neural Computing and Applications*.

Confirms RAG reduces hallucination by grounding clinical AI in trusted medical knowledge bases.

[13] Zakka, C. et al. (2024). Almanac: RAG Language Models for Clinical Medicine. *NEJM AI*, 1(2).

Landmark RAG clinical study enabling grounded, verifiable medical query responses.

[14] PMC/NIH (2025). Agentic AI and LLMs in Radiology: Hallucination Challenges. *PMC Review*.

Shows multi-agent RAG reduces hallucination to near-zero vs. 8% in single-model baselines.

[15] PMC/NIH (2025). AI Agents in Clinical Medicine: A Systematic Review. *PMC Review*.

Reviews 20 clinical AI agent studies; tool-augmented LLMs achieve median +53% improvement.

[16] Prenosil et al. (2025). Neuro-Symbolic AI for Auditable Medical Report Extraction. *Comms Medicine*.

GPT-4 + rule-based expert system hybrid matches physician performance on 206 cancer reports.

[17] PMC/NIH (2025). Explainable Diagnosis Prediction via Neuro-Symbolic Integration. *PMC*.

Logical Neural Networks outperform Random Forest (80.52% vs 76.95%) in diagnosis prediction.

[18] Springer Nature (2026). Bidirectional Neuro-Symbolic Framework for Clinical Decision Support. *NMAIHIB.*

Integrates subsymbolic learning with medical rule-based reasoning for comorbid neurological cases.

[19] MDPI (2025). Explainable AI for Clinical Decision Support: Review and Synthesis. *Informatics*, 12(4).

Identifies opacity as the primary barrier to clinical AI adoption; proposes user-centered XAI.

[20] Mednexus (2025). Evaluating LLMs and Agents in Healthcare. *Intelligent Medicine*.

Defines four-level agent autonomy taxonomy and specialized clinical evaluation metrics.