MedicareAI: A Neuro-Symbolic Agentic Framework for Secure, Multi-Modal Remote Triage and Longitudinal Patient Care

Kulkarni Shreyash[1], Jamma Atharva[2], Inde Samarth[3], Gudhate Samarth[4]
[1,2,3,4]*Department of Computer Science and Engineering, Swami Vivekanand Institute of Technology, Solapur, Gat No. 16/2, Solapur-Barshi Road, A/P. Khed, Taluka- North Solapur, Dist- Solapur, Maharashtra, India – 413255*

***Abstract* - The integration of Artificial Intelligence into healthcare faces significant barriers due to the "Black Box" problem and Large Language Model (LLM) hallucination tendencies. MedicareAI presents a diagnostic support ecosystem that addresses these challenges by combining Neuro-Symbolic AI with multi-modal deep learning. Unlike conventional chatbots, MedicareAI utilizes a LangGraph-based state machine that enforces clinical protocols through a deterministic pathway: Symptom Extraction, Differential Diagnosis, Risk Assessment, and Triage. The system integrates EfficientNet-B0 for X-ray analysis with Grad-CAM explainability mechanisms. An asynchronous "Write-Behind" architecture using Redis and Python worker queues manages longitudinal patient memory without latency. This paper details mathematical models for risk assessment, temporal memory decay, and computer vision pipelines, demonstrating a scalable approach to remote health management.**

***Index Terms* - Artificial Intelligence, Clinical Decision Support, Explainable AI, Large Language Models, Medical Triage, Neuro-Symbolic AI.**

I. INTRODUCTION

Healthcare stands at an inflection point where intelligent system integration has transitioned from speculation to operational necessity. Rising patient volumes, constrained resources, and increasing medical knowledge complexity create conditions where AI-assisted decision support provides substantial value. MedicareAI addresses specific requirements of remote patient triage and longitudinal care management while maintaining rigorous safety standards.

The fundamental premise is that medical AI reliability can be enhanced by constraining probabilistic models within deterministic frameworks. Rather than permitting unconstrained LLM responses, MedicareAI implements a state machine that decomposes diagnostic consultation into discrete phases with defined inputs, outputs, and validation criteria. This ensures predictable, auditable behavior.

Multi-modal capability represents another foundational design consideration. Clinical decision-making frequently requires synthesizing information across heterogeneous data modalities—textual symptoms, radiographic imagery, laboratory results, and temporal health records. MedicareAI integrates specialized processing pipelines for each modality, with Grad-CAM visualizations ensuring clinicians can verify anatomical bases for image-based suggestions.

II. BACKGROUND DATA

Contemporary healthcare faces unprecedented challenges in delivering timely medical triage services. Traditional delivery models, predicated on in-person consultations, demonstrate significant limitations when confronting scenarios demanding rapid scalability. The integration of AI into healthcare workflows has emerged as a promising avenue, yet deployment in clinical contexts reveals substantial obstacles.

The "Black Box" problem represents a formidable barrier to clinical AI adoption. Deep neural networks operate through millions of interconnected parameters that defy intuitive interpretation. When an AI classifies a radiographic image as pathological, clinicians cannot ascertain which features informed

this determination. This opacity undermines trust between healthcare providers and algorithmic tools.

Large Language Models introduce additional complications. These models, trained on vast textual corpora, demonstrate impressive natural language capabilities yet remain susceptible to "hallucination"—generating confident but incorrect assertions. In medical contexts, where erroneous recommendations may cause patient harm, this tendency poses unacceptable risks. Furthermore, LLMs lack inherent mechanisms for enforcing clinical protocols.

III. TECHNOLOGIES USED

The MedicareAI ecosystem leverages a carefully curated technology stack designed to balance computational efficiency, clinical reliability, and operational scalability.

A. Large Language Model Integration

The NLP backbone employs transformer-based architectures fine-tuned on medical corpora. LLM components are constrained to operate within structured output schemas, ensuring extracted information conforms to predefined clinical data models.

B. LangGraph State Machine Framework

LangGraph provides orchestration governing consultation flow. By treating diagnostic interactions as state machines, LangGraph enables definition of discrete states—symptom collection, differential generation, risk calculation, triage recommendation—with explicit transition conditions.

C. EfficientNet-B0 for Radiographic Analysis

Medical image analysis leverages EfficientNet-B0, a CNN architecture optimized for compound scaling of network depth, width, and resolution. The network processes radiographic inputs through depthwise separable convolutions, producing classification outputs with associated confidence scores.

D. Qdrant Vector Database and Redis

Semantic memory persistence utilizes Qdrant for high-performance vector similarity search. Redis serves dual roles: high-speed cache for session data and message broker for asynchronous worker queues, providing sub-millisecond latency for cache operations.

IV. OBJECTIVES

The MedicareAI project was conceived with clearly articulated objectives guiding architectural decisions:

1) Develop a diagnostic support system mitigating LLM uncertainties through deterministic reasoning components. Symbolic validation layers operate independently of probabilistic LLM, ensuring clinical safety rules apply consistently.

2) Address explainability requirements essential for clinical AI acceptance. MedicareAI incorporates Grad-CAM visualizations and structured output schemas providing evidentiary bases for informed decision-making.

3) Manage longitudinal patient interactions across extended time horizons through dual-layer memory architecture, maintaining coherent patient histories without succumbing to context window limitations.

4) Implement comprehensive security architecture including multi-factor authentication, encrypted communications, and audit logging for Protected Health Information compliance.

5) Achieve operational scalability through asynchronous processing architectures separating latency-sensitive user interactions from computationally intensive background operations.

V. METHODOLOGY

A. Agentic Workflow Architecture

The MedicareAI ecosystem operates through an orchestrated network of specialized computational agents. The diagnostic branch initiates with symptom NLU, where LLMs extract structured clinical information from free-text inputs. Extracted symptoms are encoded into standardized JSON schemas facilitating deterministic processing.

The symbolic safety layer operates as deterministic validation applying clinical contraindication rules through explicit programmatic logic. This layer cross-references medication histories with extracted symptoms, scanning for chronic disease indicators. Conditions like diabetes, hypertension, and cardiovascular disease trigger enhanced risk scoring.

B. Router Logic and Intent Classification

The central router performs intent classification determining appropriate processing pathways.

Diagnostic consultations trigger the full state machine workflow. Informational queries activate the medical search agent employing polymorphic search strategies across Wikipedia and DuckDuckGo APIs. Document analysis requests engage the RAG agent processing uploaded PDFs while cross-referencing with chronic condition history.

C. Asynchronous Worker Architecture

MedicareAI implements a "Write-Behind" pattern separating latency-sensitive API layers from computationally intensive background operations. The API handler performs NLU inference and immediately enqueues non-blocking jobs to Redis message queues. Dedicated worker processes consume tasks executing operations: chat log persistence, PDF report generation, semantic vector embedding, and secure document uploads.

D. Dual-Layer Memory Architecture

Standard LLMs encounter limitations maintaining coherent context across extended interactions. MedicareAI addresses this through dual-layer memory separating episodic records from semantic fact storage. Semantic memory utilizes Qdrant storing patient facts as high-dimensional vectors. When complex queries arise, semantic search retrieves relevant past facts based on vector similarity. Episodic memory maintains chronological records as TimeLineEvents in PostgreSQL with Redis caching.

VI. MATHEMATICAL MODELING

A. Clinical Risk Calculation Model

Patient acuity determination employs weighted linear combination of static and dynamic variables. The risk coefficient R is computed as:

$$R = S(severity) + A(age) + \alpha \cdot C \cdot \beta$$

Where S maps symptom severity to [0.2, 1.0], A adds demographic adjustments (0.2 for pediatrics, 0.1-0.3 for geriatrics), C is chronic disease flag count, and α, β are clamping constants preventing comorbidities from overwhelming acute symptom signals.

B. Temporal Memory Decay Model

Historical context management requires principled retention approaches. The relevance score of event e at time t is:

$$R(e, t) = \lambda \cdot exp(-\gamma \cdot (t - t_e))$$

Where t_e is event timestamp, γ is decay factor (γ→0 for critical events, γ≈0.1 for acute events). Events below threshold θ=0.3 are excluded from LLM working memory, preventing context window pollution.

C. EfficientNet Compound Scaling

EfficientNet-B0 optimizes compound scaling jointly adjusting depth, width, and resolution:

$$d = \alpha^{\varphi},\ w = \beta^{\varphi},\ r = \gamma^{\varphi}$$

The constraint $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ ensures predictable computational scaling, where φ controls overall scaling magnitude.

D. Grad-CAM Explainability Model

Grad-CAM provides interpretability by computing gradients of target class scores with respect to feature map activations:

$$\alpha^k = (1/Z) \Sigma_i \Sigma_j (\partial y^3 / \partial A_{ij}^k)$$

The localization map $L^3 = \mathrm{ReLU}(\Sigma^k \alpha^{k3} \cdot A^k)$ generates heatmaps highlighting regions contributing to predictions, enabling clinicians to verify AI identified clinically relevant structures.

E. NLU Symptom Classification

Symptom classification probability is calculated as:

$$p(c|T) = softmax(W \cdot h(T) + b)^3$$

Where h(T) is the semantic representation from patient complaints, enabling distinction between linguistically similar but clinically distinct presentations.

VII. SECURITY & INFRASTRUCTURE

The protection of Protected Health Information necessitates comprehensive security provisions. The architecture adopts a Zero-Trust model requiring explicit verification for every access request.

Multi-Factor Authentication generates one-time passwords transmitted via Twilio SMS or SMTP email, with Redis providing sub-second OTP verification. Stateless session management employs JSON Web Tokens with expiration enforcement. A PostgreSQL-based token blacklist enables immediate session revocation.

All transmissions use TLS encryption. Data at rest in PostgreSQL and Redis uses industry-standard algorithms. Document uploads are stored in S3-

compatible storage with server-side encryption, accessible only through authenticated worker processes.

VIII. CONCLUSION

MedicareAI demonstrates that medical AI reliability can be substantially enhanced by constraining probabilistic models within deterministic frameworks. By implementing LangGraph state machine governance, the system achieves behavioral predictability essential for healthcare deployment.

The integration of Grad-CAM visualizations addresses "Black Box" concerns, enabling clinicians to verify anatomical bases for diagnostic suggestions. The asynchronous Write-Behind architecture reconciles real-time responsiveness with comprehensive data persistence. The dual-layer memory architecture enables coherent patient relationships across extended time horizons.

IX. FUTURE SCOPE

Future extensions include additional modalities: electrocardiogram signals, dermatological imagery, and genomic data. Federated learning approaches enable collaborative model improvement without centralizing patient data. Integration with EHR systems through HL7 FHIR will provide comprehensive patient histories. Clinical trials comparing AI-assisted triage with standard care will provide evidence for broader adoption.

REFERENCES

[1] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," *Proc. ICML*, PMLR, 2019, pp. 6105-6114.

[2] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks," *Proc. IEEE ICCV*, 2017, pp. 618-626.

[3] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers," *Proc. NAACL-HLT*, 2019, pp. 4171-4186.

[4] A. Vaswani et al., "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," *Proc. EMNLP*, 2019, pp. 3982-3992.

[6] LangChain, "LangGraph: Building stateful, multi-actor applications with LLMs," *Documentation*, 2024. [Online]. Available: https://langchain-ai.github.io/langgraph/

[7] Qdrant Team, "Qdrant: Vector database for AI applications," *Documentation*, 2024. [Online]. Available: https://qdrant.tech/

[8] Redis Labs, "Redis: In-memory data structure store," *Documentation*, 2024. [Online]. Available: https://redis.io/

[9] M. Jones et al., "JSON Web Token (JWT)," *RFC 7519*, IETF, 2015.

[10] D. R. Karger, "Neuro-symbolic AI: The 3rd wave," *Artificial Intelligence*, vol. 314, 103810, 2023.