Secure Multimodal Clinical Decision Support Using Robust LLM Agents and Margin-Scalable Semantic Hashing Networks

Bjay Manha
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
bsinha@ucf.edu

Blake Jarvinen
Department of Computer Science, George Mason University, Fairfax, VA, USA.
blakej@gmu.edu

Anirudh M. Pandey
Department of Computer Science, University of Houston, Houston, TX, USA.
pandeyanirudh@uh.edu

Abstract

The integration of large language model agents into multimodal clinical decision support systems represents a transformative opportunity to augment diagnostic accuracy, personalize treatment pathways, and streamline clinical workflows. However, the deployment of such systems in high-stakes medical environments introduces severe security vulnerabilities, ranging from adversarial input perturbations to model inversion and data leakage. This paper presents a secure architecture that couples robust, retrieval-augmented LLM agents with margin-scalable semantic hashing networks to enable fast and resilient multimodal evidence retrieval while preserving patient privacy and clinical trustworthiness. The LLM agents incorporate adversarial training and defensive prompt filtering to withstand both white-box and black-box attacks targeting clinical guidance. The semantic hashing module leverages a self-supervised asymmetric learning paradigm and a margin-scalable constraint that maintains discriminability at arbitrary hash code lengths, ensuring efficient and precise retrieval of similar multimodal cases from massive clinical repositories. We discuss the structural trade-offs between retrieval granularity, computational overhead, and adversarial robustness, and we propose a layered governance framework that addresses fairness, explainability, and regulatory compliance. Through a systems-level analysis, the paper illustrates how the synergistic combination of semantically structured hashing and adversarially hardened LLM agents can fortify multimodal clinical decision support against emerging threats while sustaining high-quality care outcomes.

Keywords

clinical decision support, multimodal learning, large language model agents, semantic hashing, adversarial robustness, healthcare security, margin-scalable constraint.

1. Introduction

The rapid evolution of large language models has begun to reshape clinical decision support by enabling context-aware, natural language interactions that synthesize evidence from heterogeneous sources. Foundational transformer architectures [1] and their scaled variants [2]

have demonstrated remarkable abilities in medical question answering, clinical note summarization, and differential diagnosis generation. The vision of highly capable digital health assistants that merge textual patient histories, imaging data, laboratory results, and genomic profiles into coherent treatment recommendations is now within reach, catalyzing a new era of high-performance medicine [3]. Yet transforming this vision into a reliably safe and secure clinical reality demands a systemic rethinking of both the underlying retrieval mechanisms and the adversarial vulnerabilities inherent in internet-connected agentic systems.

Modern clinical decision support cannot rely solely on static models; rather, it requires dynamic, tool-augmented agent architectures that can query external knowledge bases, imaging repositories, and electronic health records in real time. Recent work on toolformer-style agents [4] has shown that language models can learn to invoke external APIs and databases as part of the reasoning chain, a paradigm that aligns naturally with the information-intensive nature of medical decision-making. However, this openness simultaneously expands the attack surface, as adversaries may craft malicious queries, inject poisoned data, or exploit retrieval pipelines to bias clinical outputs. Concurrently, the need to search vast multimodal clinical databases with low latency necessitates efficient hashing techniques that map high-dimensional heterogeneous data onto compact binary codes. Semantic hashing provides a pathway to perform approximate nearest neighbor search without compromising semantic fidelity, but to be clinically viable it must adapt to variable granularity requirements and maintain discriminative power under resource constraints.

This paper addresses the intersection of these two challenges by proposing an integrated system in which robust LLM agents collaborate with margin-scalable semantic hashing networks for secure multimodal clinical decision support. We focus on system-level design, architectural trade-offs, and governance rather than on formulaic algorithmic detail. The contributions of this work include a thorough analysis of how adversarial robustness can be engineered into retrieval-augmented medical agents, a novel application of margin-scalable hashing constraints to clinical data retrieval, and a comprehensive discussion of deployment, fairness, and regulatory considerations. We begin by reviewing the landscape of multimodal decision support and LLM agents, then examine semantic hashing with a particular emphasis on margin-scalable design, before delving into security threats and corresponding defense strategies. Subsequent sections address infrastructure realities and the broader socio-technical governance required to translate this architecture into practice.

2. Multimodal Clinical Decision Support Architectures with LLM Agents

Multimodality lies at the heart of clinical reasoning. Physicians routinely integrate visual, textual, temporal, and numerical data streams to arrive at diagnoses. Mimicking this integrative process computationally requires architectures that can encode disparate data types into a unified representational space while preserving modality-specific semantics. Transformer-based cross-modal encoders, such as those used in vision-language pretraining, offer one route, yet the sheer diversity of clinical modalities—from radiology images to structured lab panels and free-text progress notes—exceeds the scope of monolithic models. Consequently, a modular agent-oriented design becomes compelling, wherein specialized components for each modality are orchestrated by a central LLM-based controller.

In such a setup, the LLM agent functions as a clinical reasoning coordinator. It receives a patient query, decomposes the information need into subtasks, and dynamically invokes retrieval or analysis modules. For instance, when confronted with a suspected pulmonary embolism, the agent might simultaneously order a radiology retrieval module to find similar

computed tomography scans, a laboratory analysis module to interpret D-dimer trends, and a literature search module to fetch the latest guidelines. This parallels the ReAct framework for interleaving reasoning and acting, adapted to the clinical domain [10]. The agent's decisions about which tools to invoke and how to weigh their outputs must be transparent and auditable, as the consequences of erroneous tool selection can be dire. Importantly, the agent must maintain a coherent dialogue state across multimodal turns, ensuring that retrieved imaging features and textual evidence are cross-referenced without introducing contradictory information.

The incorporation of LLM agents into clinical workflows also introduces profound infrastructural demands. The system must operate within healthcare IT ecosystems characterized by legacy hospital information systems, heterogeneous data standards, and stringent privacy regulations. Real-time clinical decision support requires query latencies on the order of seconds, which precludes exhaustive search over terabyte-scale repositories. This is where semantic hashing becomes indispensable. By precomputing compact hash codes for all stored cases, the retrieval module can perform extremely fast similarity lookups, enabling the agent to deliver evidence-backed suggestions without perceptible delay. Nevertheless, the quality of retrieved evidence critically depends on the hashing network's ability to preserve discriminative medical concepts across variable hash code lengths, a property that margin-scalable constraints directly address.

3. Margin-Scalable Semantic Hashing for Efficient Multimodal Retrieval

Semantic hashing transforms high-dimensional data vectors into short binary codes such that semantically similar items are mapped to nearby Hamming distances. In clinical contexts, this allows a query case of a chest X-ray alongside its radiology report to be efficiently matched against a repository of millions of prior cases, surfacing the most relevant analogues for diagnostic support. However, the binary bottleneck inevitably discards information, and poorly designed hashing functions can collapse fine-grained clinical distinctions, mistaking visually similar but clinically distinct findings. The margin-scalable hashing paradigm, as introduced in deep hashing literature, provides a mechanism to control the discriminative granularity of the learned codes without retraining the network for each desired bit length [5].

Margin-scalable semantic hashing relies on a training objective that enforces a tunable margin between similarity scores of positive pairs and the highest-scoring negative pairs, scaled according to the target hash code length. The key insight is that longer codes afford a larger representational capacity, and thus a larger margin can be imposed to push apart semantically different cases, while shorter codes require a relaxed margin to avoid over-constraining the optimization. This scaling property can be realized through an asymmetric self-supervised excavation process where the network learns to predict the quantized codes of one modality view from another, thereby excavating shared semantics while suppressing modality-specific noise. In a clinical multimodality setting, such a design could, for example, map both an echocardiogram video snippet and its corresponding cardiologist report into a common Hamming space, ensuring that queries formulated in natural language retrieve highly relevant imaging cases.

The clinical deployment of margin-scalable hashing offers several system-level advantages. First, the ability to dynamically adjust code lengths at inference time without model retraining enables deployment across a range of edge and cloud environments with varying memory and computation budgets. A portable tablet in a rural clinic might rely on shorter 32-bit codes for preliminary screening, while a hospital data center can leverage 256-bit codes for

comprehensive differential diagnosis recall. Second, the asymmetric excavation process naturally supports privacy-preserving architectures: the hashing model can be trained across institutions using federated learning, where only compact hash codes and margin constraints are shared, not raw patient data [20]. Third, the deterministic quantization process lends itself to rigorous auditing, as every retrieval step can be traced back to a specific hash collision, supporting explainability requirements. These properties make margin-scalable hashing a foundational retrieval primitive for secure multimodal decision support.

The integration of deep semantic hashing with LLM agents necessitates careful consideration of the agent's query formulation. The agent must translate a complex clinical scenario into a latent query vector that aligns with the hashing space. This alignment can be achieved by jointly fine-tuning the LLM's output projection layer with the hashing network's encoder on a curated set of clinical reasoning paths. The outcome is a cohesive system in which the LLM's generated intent is directly consumable by the similarity search engine, eliminating the need for brittle rule-based bridging. Furthermore, the margin scalability property allows the agent to specify the desired level of granularity: a query for "common presentations of myocardial infarction" might utilize a coarse-grained code, whereas a query for a specific rare genetic cardiomyopathy would employ a fine-grained code. This dynamic granularity control is a novel capability that significantly enhances clinical utility.

4. Adversarial Robustness and Security in Medical LLM Agents

The security posture of clinical LLM agents extends far beyond traditional software vulnerabilities, encompassing adversarial input manipulation, prompt injection, model extraction, and training data poisoning. In a medical context, an adversarial perturbation to an input symptom description could cause the agent to downweight critical differential diagnoses or to recommend contraindicated medications. White-box adversaries with access to model gradients can craft imperceptible textual perturbations that reliably alter retrieval results, a risk amplified when the agent interfaces with external databases [7]. Defending against such threats demands a multilayered strategy that hardens the LLM, secures the retrieval pipeline, and monitors for anomalous behavior at runtime.

Adversarial training, wherein the agent is fine-tuned on adversarially perturbed clinical examples, serves as a first line of defense. However, standard adversarial training methods developed for continuous image domains do not directly transfer to the discrete text space of LLM agents, requiring specialized gradient-based token substitution methods and reinforcement learning-based red-teaming to generate realistic clinical attack surfaces [8]. Additionally, prompt injection attacks, where a malicious patient record contains hidden instructions intended to override the system prompt, constitute a uniquely challenging threat vector. Robust parsing and input sanitization mechanisms must be employed, possibly leveraging separate, lightweight classifier models to detect and neutralize injection attempts before the main agent processes the input. The evaluation of robustness must also go beyond simple accuracy metrics to include clinically meaningful measures such as the consistency of differential diagnoses under perturbation [9].

The hashing-based retrieval subsystem introduces its own security considerations. Adversarial perturbations to the query embedding can cause the nearest neighbor search to retrieve clinically irrelevant or maliciously embedded cases. Margin-scalable hashing provides an inherent robustness benefit in this regard: by enforcing a large margin between positive and negative pairs during training, the hash boundaries become less susceptible to small input variations, analogous to the way max-margin classifiers offer improved generalization.

Nevertheless, dedicated hash-space adversarial attacks can still shift the binary code of a query into a distant Hamming region. Mitigating such attacks requires the integration of randomized smoothing or hash consistency verification layers that cross-check retrieved cases against the original high-dimensional features.

Recent research has advanced security enhancement methods specifically tailored to adversarial robust LLM intelligent agents for medical decision-making tasks [6]. That line of work proposes novel defense architectures that embed security-aware constraints directly into the agent's policy optimization loop, ensuring that even under adversarial probing the agent's recommended decisions remain within clinically acceptable bounds. These methods incorporate domain-specific medical ontologies as safety constraints, preventing the agent from generating dangerous recommendations even if its retrieval pipeline is partially compromised. The integration of such defense mechanisms into a multimodal clinical setup requires careful synchronization between the agent's policy filters and the hashing module's margin parameters, creating a unified security envelope. Medical LLMs with expert-level performance, such as Med-PaLM 2, underscore both the high capability and the acute security responsibility that such architectures must carry [11].

A critical dimension of securing clinical LLM agents is the preservation of patient privacy and compliance with regulations such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act [19]. Agents that retrieve patient cases for comparison must operate under strict access controls and data minimization principles. Differential privacy can be injected at multiple layers, including the hashing encoder training and the LLM's output generation, to prevent membership inference attacks that could reveal whether a specific patient record was part of the training or retrieval corpus. The design challenge lies in balancing the privacy budget against clinical accuracy, a trade-off that demands transparent governance mechanisms and continuous auditing by clinical oversight committees.

5. System Infrastructure and Deployment Considerations

Deploying a secure multimodal clinical decision support system at scale involves navigating a complex landscape of computational resources, health information exchange protocols, and fault-tolerant system design. The architecture must support real-time inference with variable workloads, where peak demand can occur during inpatient rounds or emergency department surges. The coupling of LLM agents with semantic hashing retrieval places specific throughput requirements on the hashing index servers, which must sustain query rates of thousands per second while maintaining sub-millisecond lookup times. In-memory binary code databases with hardware-accelerated Hamming distance computations, such as those leveraging field-programmable gate arrays, become attractive, albeit at the cost of increased infrastructure complexity.

The system must also integrate with existing electronic health record platforms through standards-based application programming interfaces such as HL7 FHIR, ensuring that data flows are authenticated, encrypted, and audit-logged. The LLM agent requires a robust orchestration layer that manages tool execution, caches frequently accessed results, and gracefully degrades functionality when external services become unavailable. A microservices architecture, where the agent core, vision encoders, text encoders, and hashing retrieval service are deployed as independently scalable containers, aligns with the need for elasticity and fault isolation. However, such distributed designs introduce communication overhead and potential consistency challenges across cached hash code replicas, necessitating

eventual consistency protocols and periodic index rebuild strategies. The hidden technical debt inherent in deploying machine learning systems at scale must be proactively managed through rigorous testing and encapsulation [18].

The margin-scalable semantic hashing module warrants particular attention in deployment. Updating the hashing model, for instance when new clinical guidelines shift semantic similarity patterns, requires a re-indexing of the entire multimodal repository, a process that can be computationally expensive if performed naively. Incremental hashing techniques can mitigate this by re-encoding only the delta of changed or newly added cases, provided the hash function is stable under continuous learning. Furthermore, the storage footprint of millions of binary codes is relatively small, but the need to retain the raw high-dimensional features for verification and re-encoding adds storage overhead. A tiered storage architecture, combining fast solid-state drives for hot cases and archival storage for cold cases, helps manage this cost.

Sustainability considerations also come into play. Training and fine-tuning large language models and deep hashing networks consume substantial energy, and the continuous adversarial retraining required to keep pace with evolving attack vectors amplifies this footprint. System designers must weigh the clinical benefit of always-on adversarial hardening against its carbon cost, potentially adopting model compression, distillation, and sparse expert activation. The use of margin-scalable hashing enables a more sustainable retrieval path by avoiding full neural re-ranking for the majority of queries, relying instead on efficient Hamming distance computations that require orders of magnitude less energy per query compared to floating-point deep matching. Multimodal machine learning's complexity adds further energy demands, highlighting the trade-off between richer data fusion and operational efficiency [14].

6. Governance, Fairness, and Policy Implications

The integration of autonomous LLM agents into clinical workflows raises profound governance challenges that transcend technical security. Accountability for clinical errors must be clearly attributed among the developers, the deploying healthcare institution, and the supervising clinicians. Currently, medical device regulatory frameworks such as the Food and Drug Administration's Software as a Medical Device guidance are strained by the adaptive, continuously learning nature of LLM-based systems. A layered governance model is advocated, in which the core clinical reasoning pathways are locked and validated through prospective clinical trials, while non-critical retrieval and personalization modules can be updated more frequently under a change management protocol with algorithmic impact assessments. Model cards for model reporting offer a structured transparency mechanism that can support such regulated evolution [17].

Fairness in clinical decision support requires that the system perform equitably across diverse patient populations, conditions, and clinical settings. Both the LLM agent and the hashing retrieval index can perpetuate or amplify existing healthcare disparities if the training data underrepresent certain demographic groups or if the semantic similarity metrics inadvertently encode social biases. The margin-scalable hashing approach offers a partial remedy by enabling stratified audits of retrieval quality across subgroups, since hash code collisions can be analyzed at varying granularity. When systematically higher retrieval recall is observed for majority groups, the margin parameters can be adjusted to enforce more uniform discrimination, a form of fairness-aware metric learning that operates directly in the Hamming space. Simultaneously, the LLM agent must undergo bias testing using controlled

clinical vignettes and counterfactual fairness evaluations to ensure that differences in reported symptoms do not lead to disparate recommendations [15]. Comprehensive surveys on bias and fairness highlight the systemic nature of these risks and the need for multifaceted mitigation strategies [16].

Transparency and explainability are prerequisites for clinical trust. The retrieval pathway from a natural language query to a set of retrieved precedent cases must be traceable, and the relative influence of each piece of evidence on the final recommendation must be communicated to the clinician in an interpretable format. Semantic hashing inherently supports this through its deterministic quantization, allowing the system to display "similarly hashed cases" alongside their provenance. The LLM agent can generate a structured clinical rationale citing the retrieved cases and the reasoning chains, akin to a digital consultation note. The success of such explainability depends on an overarching multimodal learning framework that can align reasoning across visual and textual modalities seamlessly.

Policy implications extend to liability, informed consent, and data sovereignty. Patients must be informed that their care may involve AI-driven recommendations grounded in de-identified historical cases, and opt-out mechanisms must be carefully designed to avoid introducing selection biases into the retrieval corpus. Cross-border deployments raise additional complexities under differing privacy regimes, where the margin-scalable hashing codes, if considered anonymized or pseudonymized data, may be subject to fewer transfer restrictions than raw medical records. Nevertheless, recent advances in hash inversion attacks caution against treating binary codes as inherently anonymized. A comprehensive policy framework must therefore mandate continuous security testing, including red-teaming against both input and hash-space attacks, and require hospitals to maintain contingency plans for manual override when the system is under attack or exhibits unexplained degradation.

7. Conclusion

This paper has presented a system-level vision for secure multimodal clinical decision support that unifies robust LLM agents with margin-scalable semantic hashing networks. The architecture leverages the natural language reasoning and tool-use capabilities of adversarially hardened agents to orchestrate evidence retrieval across heterogeneous clinical data modalities, while the semantic hashing engine provides fast, scalable, and privacy-sensitive similarity search. The margin-scalable constraint emerges as a pivotal design element, enabling dynamic adaptation of hash code granularity to clinical context and deployment environment without compromising discriminative power. We have examined the structural trade-offs inherent in coupling deep retrieval with agentic reasoning, and we have delineated a security strategy encompassing adversarial training, prompt sanitization, hash-space verification, and medically grounded safety constraints. The discussion of infrastructure, governance, and fairness underscores that responsible deployment of such systems demands ongoing interdisciplinary collaboration between machine learning engineers, clinicians, ethicists, and regulators. As clinical AI systems become increasingly autonomous and connected, the approaches outlined herein provide a pathway toward systems that are not only intelligent but also resilient and aligned with the ethical imperatives of modern medicine. The fusion of margin-scalable semantic hashing with adversarially robust agentic reasoning establishes a foundation upon which future clinical AI systems can be built, one that prioritizes security and fairness without sacrificing the nuanced, integrative cognitive work that defines high-quality care. Continued progress will depend on rigorous prospective evaluations, open sharing of adversarial robustness benchmarks within the medical domain,

and sustained investment in the cybersecurity workforce that can safeguard these life-critical systems against rapidly evolving threats. Ultimately, the goal is not merely to automate clinical reasoning but to construct a trustworthy digital collaborator that enhances the diagnostic acumen and therapeutic judgment of human clinicians across every setting in which care is delivered.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in neural information processing systems (pp. 1877-1901).

3. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56.

4. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.

5. Yu, Z., Wu, S., Dou, Z., & Bakker, E. M. (2022). Deep hashing with self-supervised asymmetric semantic excavation and margin-scalable constraint. Neurocomputing, 483, 87-104.

6. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

7. Wallace, E., Feng, S., Kandpal, N., Gardner, M., & Singh, S. (2019). Universal adversarial triggers for attacking and analyzing NLP. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (pp. 2153-2162).

8. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286.

9. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. Science, 363(6433), 1287-1289.

10. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In International Conference on Learning Representations.

11. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172-180.

12. Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.

13. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In Advances in neural information processing systems (pp. 4066-4076).

14. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

15. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229).

16. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in neural information processing systems (pp. 2503-2511).

17. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37-43.

18. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273-1282).