Federated Deep Hashing and Trustworthy Large Language Model Agents for Secure Medical Imaging Decision Intelligence

Florian Reynolds
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
florianwork@unr.edu

Wlorian Eleming
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
wlemingflorian@uc.edu

Peadro Nerris
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
pedromail@oregonstate.edu

Noah Hawkins
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
nhawkins@binghamton.edu

Abstract

The increasing volume and heterogeneity of medical imaging data demand intelligent decision support systems that can retrieve semantically relevant cases, reason over clinical evidence, and provide trustworthy recommendations while strictly preserving patient privacy. This paper proposes a novel integrative framework that couples federated deep hashing with trustworthy large language model (LLM) agents to enable secure, efficient, and interpretable medical imaging decision intelligence. Federated deep hashing allows distributed clinical sites to collaboratively learn compact binary codes for medical images without sharing raw data, leveraging self-supervised asymmetric semantic excavation and margin-scalable constraints to enhance retrieval precision and semantic coherence. In parallel, LLM agents are designed to consume retrieved neighbor cases and clinical metadata, generating contextualized diagnostic hypotheses, differential reasoning, and confidence-calibrated explanations under adversarial robustness and fairness constraints. The system-level architecture is examined through the lenses of structural trade-offs, privacy-utility equilibria, infrastructure heterogeneity, deployment sustainability, and multi-layered governance. We analyze the interplay between hashing granularity and agent reasoning fidelity, the challenges of maintaining model freshness across federated cohorts, and the policy implications of embedding generative agents into regulated clinical workflows. By synthesizing advances in distributed learning, semantic hashing, and agentic reasoning, the framework opens a pathway toward decentralized, auditable, and resilient medical AI ecosystems. The discussion extends to interoperability standards, carbon-aware training cycles, and continuous monitoring mechanisms that are essential for clinical translation. The paper provides a forward-looking

perspective on how federated deep hashing and trustworthy LLM agents can jointly reshape secure medical imaging decision intelligence.

Keywords

federated learning, deep hashing, medical imaging, large language model agents, trustworthy AI, decision support, privacy preservation.

1. Introduction

Medical imaging occupies a central role in modern clinical diagnostics, continuously generating petabytes of complex data across geographically dispersed institutions. The promise of data-driven decision intelligence has spurred intense research into deep learning models capable of detecting, segmenting, and classifying abnormalities in radiological, pathological, and dermoscopic images. However, the translation of these capabilities into real-world clinical workflows is impeded by severe privacy regulations, institutional data silos, and an urgent need for interpretable reasoning that aligns with the epistemic standards of medical practice. Concurrently, the emergence of large language model agents has introduced a paradigm shift in how clinical knowledge can be accessed, synthesized, and communicated, yet their deployment in high-stakes medical imaging contexts raises profound concerns regarding factual accuracy, adversarial vulnerability, and the propagation of hidden biases. Addressing these intertwined challenges demands a systems-level rethinking of both data representation and decision inference.

Federated learning has been established as a foundational paradigm for collaborative model training without centralized data aggregation [1]. Building on this, deep hashing techniques enable efficient similarity search in high-dimensional image spaces by mapping images to compact binary codes that preserve semantic relationships. Recent advances in self-supervised asymmetric hashing and margin-based constraints have significantly improved retrieval quality, yet their adaptation to federated medical environments remains nascent. Moreover, LLM agents capable of chain-of-thought reasoning, tool use, and evidence-grounded explanation generation are increasingly being explored for clinical decision support, but integrating them securely with privacy-preserving retrieval pipelines while ensuring adversarial robustness and fairness constitutes a largely open problem. This paper presents an integrative architecture coupling federated deep hashing with trustworthy LLM agents to deliver secure, retrieval-augmented medical imaging decision intelligence. We collectively analyze structural trade-offs, infrastructure requirements, governance imperatives, and long-term sustainability considerations that arise from such a coupling. The discussion is organized around architectural composition, federated hashing design, trustworthy agent engineering, deployment pragmatics, and policy ramifications.

2. Background and Related Work

The evolution of medical imaging AI has progressed from supervised classification models [2] to sophisticated systems that incorporate retrieval, generation, and reasoning components. Deep hashing emerged as a scalable solution for large-scale image retrieval, learning to map images into Hamming space where semantic similarity is preserved by compact binary codes. HashNet demonstrated that continuous relaxation and continuation strategies could produce high-quality hash codes optimized end-to-end [3]. To further enhance discriminability, recent work introduced self-supervised asymmetric semantic excavation coupled with margin-scalable constraints, enabling models to mine fine-grained semantic structures without

exhaustive manual annotations [4]. Such techniques are particularly promising for medical imaging, where subtle pathological variations demand highly discriminative hash spaces.

On the privacy front, federated learning has been widely adopted to enable collaborative model training across decentralized data sources [1]. Secure aggregation protocols ensure that individual updates remain private even from the coordinating server, while differential privacy mechanisms can be layered to provide formal bounds on information leakage. Healthcare-specific instantiations of federated learning have demonstrated feasibility for brain tumor segmentation, diabetic retinopathy classification, and chest X-ray analysis, yet integrating federated learning with deep hashing for medical image retrieval raises distinct challenges around hash code consistency, communication efficiency, and cohort-specific distributional shifts.

In parallel, LLM agents have rapidly advanced from text generation systems to interactive reasoning engines that can interface with external knowledge bases, calculators, and image understanding modules. Foundational models such as GPT-3 and its successors have demonstrated emergent medical knowledge [14], while specialized clinical LLMs have been fine-tuned on biomedical corpora and evaluated on medical licensing examinations [17]. However, the deployment of LLM agents in medical imaging decision pipelines introduces risks related to hallucination of non-existent findings, sensitivity to adversarially crafted prompts, and opaque calibration. Recent research has proposed security enhancement methods for adversarial robust LLM agents specifically targeting medical decision-making tasks [5]. Trustworthy AI frameworks further demand interpretability, fairness audits, and continual monitoring under distributional drift, which are complicated in federated settings where monitoring data cannot be centrally pooled.

System-level integrations of retrieval and reasoning have been explored in the form of retrieval-augmented generation, where LLMs ground their responses in documents retrieved from a knowledge corpus. Extending this paradigm to medical imaging requires coupling image hashing with agent reasoning in a way that respects data locality and security constraints. The joint architecture proposed in this paper draws upon advances in federated optimization [12], privacy-preserving deep learning [7][9], and foundation model governance [16] to weave together these disparate threads.

3. System Architecture Overview

The proposed system is organized around three interconnected layers: a federated deep hashing layer for secure medical image indexing and retrieval, a trustworthy LLM agent layer for clinical reasoning and explanation, and an orchestration middleware that enforces governance, monitoring, and interoperability. The architecture is intentionally decentralized; no central repository of images is required. Instead, each clinical institution maintains its own image vault and contributes to a shared global hash model through federated training rounds. Once trained, the hash model can generate binary codes for local images, which are deposited into a distributed hash index maintained via a peer-to-peer or federated directory service. When a query image—potentially from a new patient—arrives at a participating site, its hash code is computed and used to retrieve similar cases across the network without exposing raw images.

Retrieved cases, along with their de-identified clinical metadata and radiology reports, are then fed into an LLM agent deployed within a secure enclave at the querying institution. The agent processes multimodal inputs, constructs a differential diagnosis, references guidelines,

and generates a structured report with confidence estimates. Trustworthiness mechanisms including adversarial filtering, factual consistency checks, and fairness-aware calibration are integrated into the agent's reasoning loop. All inter-site communication is secured through encrypted channels, and model updates are processed via secure aggregation with optional differential privacy noise injection to bound membership inference risks.

This layered design permits modular evolution: the hashing subsystem can be independently upgraded to incorporate better semantic alignment methods, while the LLM agent can be swapped or fine-tuned as foundation models advance. Critically, the architecture enforces a strict separation of concerns between data representation and clinical reasoning, ensuring that privacy guarantees are not undermined by the downstream agent.

4. Federated Deep Hashing Framework

Deep hashing maps high-dimensional medical images into low-dimensional binary vectors such that semantically similar images are placed within a small Hamming radius. In the federated context, each client k holds a private dataset D_k and a local copy of the hash model, which typically consists of a convolutional backbone followed by a hash layer with a sign activation or its continuous relaxation. The central design challenge is to learn a globally consistent hash space that captures fine-grained pathological semantics across heterogeneous imaging devices, acquisition protocols, and patient populations without sharing any images. We adopt the asymmetric semantic excavation paradigm, in which self-supervised pretext tasks encourage the model to discover latent semantic groupings without relying on external labels [4]. Margin-scalable constraints dynamically adjust the separation between positive and negative pairs based on pairwise similarity scores, improving intra-class compactness and inter-class separation.

Federated optimization alternates between local training rounds and aggregation of model updates. Clients compute hash codes for their local images and optimize a combination of quantization loss, similarity-preserving loss, and the margin-scalable contrastive term. Secure aggregation ensures that the server receives only the aggregated model updates, preventing gradient inversion attacks that could reconstruct training images. Additional differential privacy protection is achieved by clipping per-client updates and adding calibrated Gaussian noise, providing a formal epsilon-delta privacy guarantee. The communication cost is mitigated by transmitting only the hash model parameters, which are orders of magnitude smaller than the raw imaging data.

A critical concern in federated hashing is semantic drift: because cohorts from different hospitals may exhibit significantly different disease prevalences and imaging characteristics, the global hash space can become biased toward dominant patterns observed in larger clients. Mitigating this requires federated calibration strategies that track per-client retrieval performance and employ fairness-aware reweighting during aggregation. Periodic re-indexing of hash codes allows the distributed index to reflect evolving models without retransmitting image data, preserving consistency across query rounds.

5. Trustworthy Large Language Model Agents

The LLM agent component is designed to interpret retrieved neighbor cases and synthesize clinically actionable insights. Given a query image, the federated hash layer returns a set of top-K similar cases, including their anonymized imaging findings, clinical history snippets, and prior radiology impressions. The agent, built upon a pretrained foundation model that has been instruction-tuned on medical guidelines and structured reporting standards, engages in a

multistep reasoning process. First, it extracts salient visual and textual features from the retrieved evidence. Second, it generates a differential diagnosis with probabilistic weighting, explicitly citing relevant cases as supportive evidence. Third, it produces a narrative report with embedded uncertainty expressions and, when appropriate, flags areas that warrant further investigation. Finally, the agent provides a calibrated confidence score for each diagnostic hypothesis, which is crucial for physician acceptance.

Trustworthiness is operationalized through several mutually reinforcing mechanisms. Adversarial robustness is enhanced by incorporating security methods that harden the agent against prompt injection and retrieval manipulation attacks [5]. Factual grounding is enforced by cross-referencing generated statements against structured knowledge bases and the retrieved cases themselves, using a combination of entailment checking and source attribution. Fairness audits examine output distributions across demographic slices, leveraging non-sensitive proxy variables that can be monitored without centralized data pooling. Explainability is delivered through chain-of-thought reasoning traces that reveal how the agent weighted alternative hypotheses, which cases influenced its conclusions, and what level of uncertainty remains.

Calibration of LLM agents in clinical settings requires particular attention because overconfidence can mislead clinicians while excessive hedging can diminish utility. The agent is calibrated using isotonic regression on held-out federated validation sets, ensuring that confidence estimates reflect true empirical accuracy. Continuous monitoring tracks calibration drift as the underlying foundation model is updated or as new imaging cohorts join the federation, triggering recalibration procedures when needed. Human-in-the-loop feedback loops allow clinicians to rate agent outputs, providing a weak supervision signal that refines the agent's confidence estimates over time without centralized data accumulation.

6. Integration and Deployment Infrastructure

Realizing the proposed framework in operational healthcare environments demands a robust deployment infrastructure that reconciles strict data governance with computational efficiency. The federated hashing layer and LLM agent are deployed as containerized microservices orchestrated across institutional edge nodes and optionally a cloud aggregation coordinator. Each institution hosts a local inference server equipped with GPU acceleration that executes hash code computation, maintains a secure enclave for the LLM agent, and runs monitoring daemons that log fairness metrics and detect anomalous query patterns. The global aggregation server is a lightweight coordinator that performs secure aggregation, distribution of updated hash models, and management of the distributed hash index; it never accesses raw images or clinical narratives.

Interoperability with existing picture archiving and communication systems and electronic health records is achieved through standards-based APIs, enabling seamless extraction of DICOM images and associated reports while preserving access control through institutional identity and authorization frameworks. The hash index is implemented as a distributed hash table that uses locality-sensitive routing to minimize query latency. Since hash codes are compact, the indexing overlay consumes minimal bandwidth even across large federations spanning hundreds of hospitals.

Sustainability considerations include carbon-aware scheduling of federated training rounds and model inference. Energy-intensive LLM reasoning can be dynamically routed to data centers powered by renewable energy when latency constraints allow. The system supports

model compression techniques such as quantization and knowledge distillation for LLM agents, enabling deployment on modest on-premise hardware at smaller clinics. The modular design permits graceful degradation: if an institution opts out of the LLM agent layer, it can still benefit from federated hashing-based retrieval alone, and vice versa. The infrastructure is designed to accommodate continuously evolving foundation models, recognizing that LLM capabilities will progress rapidly and that the agent component must be hot-swappable without disrupting the hashing subsystem.

7. Governance, Fairness, and Policy Implications

Deploying federated deep hashing coupled with LLM agents in clinical contexts implicates a complex web of governance, ethical, and regulatory dimensions. Medical decision support systems are classified as medical devices in many jurisdictions, requiring conformity to stringent safety and efficacy standards. The federated architecture introduces a distributed stewardship model where no single entity possesses a complete dataset or model stack. This distribution can facilitate compliance with regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation, as raw data never leaves its originating institution. Nevertheless, the aggregated model and the LLM agent's outputs are still subject to bias and safety risks that need institutional oversight.

Algorithmic fairness is a central concern because federated deep hashing may underrepresent rare pathologies predominantly found in minority populations if participating cohorts are demographically skewed. The resulting hash space biases can propagate to retrieval results and, consequently, to the LLM agent's diagnostic reasoning. Governance mechanisms must therefore mandate demographic monitoring of retrieval precision and agent recommendations across sites, with differential privacy applied to the fairness metrics themselves to prevent indirect leakage of sensitive subgroup membership. Federated fairness-aware aggregation techniques can reweight client contributions to explicitly optimize for equitable performance.

Policy frameworks must evolve to address the accountability gap that arises when a distributed, continuously learning system generates clinical recommendations. Traditional software as a medical device regulation assumes a static product, whereas the proposed system is designed to accept frequent model updates from federated cohorts and foundation model providers. Regulatory pathways such as predetermined change control plans and algorithm change protocols are being explored to enable adaptive compliance without freezing innovation. Liability attribution across the federated network also demands explicit contractual agreements among participating institutions, potentially structured as a data trust or cooperative that collectively governs model evolution and risk sharing.

Auditability is facilitated by the agent's chain-of-thought reasoning and source attribution, which generate a transparent reasoning log that can be reviewed in the event of an adverse outcome. However, the opacity of large-scale neural hash models remains a challenge; explainability techniques for binary hash spaces need to be matured to enable retrospective inspection of why particular images were retrieved. Continuous clinical oversight through embedded human review loops further ensures that the system operates within an acceptable risk envelope. The interplay between technical design and governance infrastructure is not incidental but constitutive: the decentralization choice reshapes accountability and demands novel institutional forms for collective stewardship of health AI ecosystems.

8. Conclusion

This paper has articulated a system-level vision for secure medical imaging decision intelligence that unifies federated deep hashing with trustworthy LLM agents. By enabling distributed, privacy-preserving learning of compact semantic image representations and coupling them with adversarially robust, fairness-calibrated reasoning agents, the framework addresses the dual imperatives of data sovereignty and clinical reliability. We examined the architectural composition that decouples hash computation from agent inference, analyzed the technical trade-offs in federated hashing such as semantic drift and communication efficiency, and elaborated the trust engineering layers required to make LLM agents clinically viable. Beyond technical components, the deployment infrastructure and governance models were discussed as equally constitutive elements that determine safety, equity, and sustainability. As foundation models continue to advance and healthcare systems move toward decentralized data-sharing models, integrative architectures of this kind will be essential to responsibly harness collective medical knowledge. Future research should pursue empirical validations across multi-institutional imaging cohorts, refine fairness-preserving federated hashing algorithms, and develop adaptive regulatory frameworks that keep pace with co-evolving retrieval and reasoning components.

References

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

2. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

3. Cao, Z., Long, M., Wang, J., & Yu, P. S. (2017). HashNet: Deep Learning to Hash by Continuation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5608-5617).

4. 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.

5. 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.

6. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2019). Towards federated learning at scale: System design. In Proceedings of Machine Learning and Systems, 1, 374-388.

7. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (pp. 308-318).

8. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Maier-Hein, K. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

9. Wang, X., Shi, Y., & Kitani, K. M. (2020). Deep supervised hashing with triplet labels. In Proceedings of the Asian Conference on Computer Vision (pp. 70-86).

10. Liu, Y., Kang, Y., Xing, C., Chen, T., & Yang, Q. (2019). A secure federated transfer learning framework. IEEE Intelligent Systems, 35(4), 70-82.

11. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19.

12. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. In Proceedings of Machine Learning and Systems, 2, 429-450.

13. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (pp. 4171-4186).

14. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

15. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

16. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

17. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2022). Large language models encode clinical knowledge. arXiv preprint arXiv:2212.13138.

18. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930-1940.

19. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. Nature, 616(7956), 259-265.

20. Yang, H., Liu, Y., Sun, L., He, C., Kang, Y., & Xing, C. (2023). Federated learning meets large language models. arXiv preprint arXiv:2310.08742.