Privacy-Preserving Medical Knowledge Retrieval with Self-Supervised Hash Learning and Adversarial Defense for Healthcare AI Agents

Zhoukai Xue
School of Computing, Clemson University, Clemson, SC, USA.
zxue@clemson.edu

Troy Fields
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
troy.work@ku.edu

Ferry Hansson
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
ferry.hansson@buffalo.edu

Dean Wabb
Department of Computer Science, University of North Texas, Denton, TX, USA.
deanwebb90@unt.edu

Abstract

The rapid integration of artificial intelligence agents into clinical decision support systems has intensified the demand for medical knowledge retrieval pipelines that simultaneously ensure high-fidelity semantic access, rigorous privacy protection, and resilience against adversarial manipulation. This paper presents a system-level architectural investigation into privacy-preserving medical knowledge retrieval that unifies self-supervised hash learning with layered adversarial defense mechanisms. By mapping medical knowledge fragments into compact binary codes through self-supervised contrastive objectives, the framework enables efficient approximate nearest-neighbor search over distributed repositories without exposing sensitive clinical content. The proposed architecture couples a differentially private hash encoder with secure index structures and an adversarial sanitization module that monitors query integrity and defends against both input-space perturbations and knowledge-poisoning attacks. We examine structural trade-offs between hash code length, retrieval precision, computational latency, and achievable privacy guarantees, drawing on cross-domain insights from cryptographically secure computation, federated learning, and adversarial robustness literature. Deployment considerations are analyzed within the context of hospital information ecosystems and cross-border regulatory regimes, addressing governance challenges such as algorithmic auditing, fairness across heterogeneous patient populations, and the sustainability of large-scale hash-based retrieval infrastructure. The study further explores forward-looking policy implications for certifying autonomous healthcare AI agents that rely on privacy-preserving retrieval as a core cognitive operation. The analysis demonstrates that self-supervised hashing constitutes a promising foundation for trust-enhancing knowledge access, yet requires careful co-design with adversarial defense, governance frameworks, and lifecycle management to withstand emerging threat surfaces in medical decision-making environments.

Keywords

Privacy-preserving retrieval; self-supervised hashing; adversarial defense; healthcare AI agents; medical knowledge graphs; secure multi-party computation.

1. Introduction

The expanding role of artificial intelligence agents in healthcare is reshaping clinical workflows by automating evidence synthesis, differential diagnosis generation, and personalized treatment recommendation. Such agents depend on high-quality, up-to-date medical knowledge retrieved from diverse repositories, including electronic health records, clinical guidelines, biomedical literature, and knowledge graphs. The retrieval process, however, introduces acute privacy and security concerns because queries and retrieved content may reveal patient-identifying information, institutional care patterns, or exploitable vulnerabilities in decision logic. Simultaneously, the agent itself becomes a target for adversarial interference designed to distort retrieval outcomes, potentially leading to harmful clinical inferences. These intertwined challenges demand retrieval architectures that are privacy-preserving by design and adversarially robust under realistic clinical threat models.

Conventional approaches to privacy-preserving data access in healthcare have relied on federated learning, secure multi-party computation, and differential privacy to train models without exposing raw data [1, 2]. While these techniques protect against certain classes of leakage, they often introduce substantial computational overhead and do not directly address the efficiency requirements of real-time knowledge retrieval at the scale of modern medical knowledge bases. Hashing-based retrieval, which represents documents, concepts, or relational triples as compact binary codes, offers an attractive alternative by enabling fast similarity search with provably sublinear complexity. When hash functions are learned through self-supervised objectives, they can capture high-level semantic relationships without requiring costly manual annotations, thus facilitating deployment across heterogeneous and evolving medical data collections. Recent deep hashing methods that exploit asymmetric semantic excavation have demonstrated that self-supervised hash learning can preserve fine-grained semantic similarity while producing highly compact codes [3].

Adversarial threat surfaces in medical AI systems extend well beyond simple input perturbations; they include crafted queries that exploit semantic blind spots of the hash encoder, poisoning attacks that inject spurious relationships into the knowledge base, and backdoor triggers that cause systematic mis-retrieval under specific conditions [4, 5]. When retrieval is embedded within an agent that uses retrieval-augmented generation, as commonly practiced with large language models, the downstream impact of adversarial retrieval becomes particularly severe, as incorrect evidence can propagate unchecked into clinical conversations [6]. Therefore, the architectural integration of adversarial defense within the retrieval pipeline is not an optional add-on but a foundational requirement for healthcare AI agents operating in safety-critical settings.

This paper contributes a system-level framework that unifies privacy-preserving hash-based retrieval with comprehensive adversarial defense tailored to medical knowledge domains. We examine the interplay between self-supervised representation learning, cryptographic privacy mechanisms, and robustness engineering through an architectural lens that prioritizes deployment feasibility, governance alignment, and long-term sustainability. The analysis is structured to illuminate the structural trade-offs that arise when these distinct technical

components are integrated, offering insights that span from hash code design to regulatory policy.

2. Background and Related Work

The intersection of privacy, retrieval, and robustness in medical AI brings together several research streams that have largely evolved in parallel. Privacy-preserving machine learning in healthcare has been dominated by federated architectures that keep data at the edge while aggregating model updates through secure protocols [1]. Differential privacy has been applied to clinical text and structured health data to provide formal guarantees that individual contributions remain indistinguishable [2, 11]. These mechanisms, however, focus on model training rather than on retrieval over distributed knowledge repositories, leaving a gap in privacy-preserving inference-time access. Secure multi-party computation frameworks for search allow encrypted queries over sensitive indexes, but scaling them to high-dimensional medical embeddings remains challenging [7].

On the retrieval side, deep hashing has emerged as a powerful technique for large-scale similarity search in vision and text domains. Learned hash functions map high-dimensional representations into low-dimensional Hamming space, where distances can be computed efficiently using bitwise operations. Self-supervised hashing removes the dependence on labeled similarity pairs by exploiting data augmentation strategies and contrastive objectives that pull semantically related content closer in the hash space while pushing unrelated content apart [9, 15]. In medical contexts, deep hashing has been applied to histopathological image retrieval and cross-modal alignment of imaging and textual reports, demonstrating that binary codes of 64 to 256 bits can achieve competitive retrieval accuracy while enabling orders-of-magnitude speedups [10]. The introduction of asymmetric semantic excavation and margin-scalable constraints has further improved hash code quality by respecting the fine-grained similarity structure often present in medical taxonomies and ontologies [3].

Adversarial robustness in textual and retrieval domains has received increasing attention, with studies exposing the brittleness of neural encoders to carefully crafted perturbations that are perceptually imperceptible to clinicians yet drastically alter retrieval rankings [4, 5, 16]. In the medical domain, adversarial attacks can exploit the sensitivity of clinical language models to synonym substitution, negation insertion, or numerical value changes, leading to unsafe knowledge retrieval [8]. Defenses such as adversarial training, input purification, and certified robustness provide varying degrees of protection but have only been preliminarily studied in the context of retrieval systems. The emergence of large language model agents in medical decision-making tasks has further elevated the stakes, as retrieval serves as the agent's epistemic grounding mechanism; recent work on security enhancement for such agents emphasizes the need for multi-layered defenses that cover both the generative and retrieval components [17].

Several works have begun to connect privacy with retrieval. Secure k-nearest neighbor computation over encrypted databases and privacy-preserving multi-keyword ranked search over cloud data provide cryptographically grounded primitives for search without plaintext access [7, 20]. Federated retrieval architectures, where indexes are distributed and queries are resolved through secure aggregation, represent a nascent but promising direction [24]. However, the integration of these cryptographic privacy layers with learned hash functions and adversarial defense has not been systematically addressed, especially under the unique semantic and regulatory constraints of healthcare.

3. System Architecture: Privacy-Preserving Medical Knowledge Retrieval

The proposed architecture is structured around four core subsystems: a knowledge encoding and hashing module, a privacy enforcement layer, a distributed retrieval index, and an AI agent interface augmented with adversarial defense. Medical knowledge sources, including textual clinical guidelines, structured ontology entries, imaging annotations, and relational triples from knowledge graphs, are first transformed into dense vector representations using domain-adapted transformer encoders or graph neural networks that have been pre-trained on de-identified corpora. These representations are subsequently passed through a self-supervised hash encoder that projects them into a binary code of configurable length, typically ranging from 48 to 256 bits depending on the privacy-utility target. The hash encoder is trained using contrastive learning objectives that do not require explicit relevance labels, instead relying on data augmentations such as paraphrasing, entity masking, and graph perturbations to define positive pairs. This self-supervision strategy is particularly well-suited to medical knowledge, where manually curated relevance labels are scarce and expensive to produce.

The privacy enforcement layer operates immediately after hash generation and is responsible for ensuring that the binary codes do not leak information that could be inverted to reconstruct sensitive clinical content. One instantiation applies a randomized response mechanism at the bit level, calibrated to achieve a prescribed local differential privacy guarantee. Alternative instantiations rely on client-side encryption using secure indexing schemes that allow similarity search in encrypted space through homomorphic comparison or secure two-party computation. The trade-off here is fundamental: longer hash codes preserve more semantic detail and improve retrieval accuracy but simultaneously increase the information leakage potential and the computational cost of privacy-enhancing operations. Short codes, in contrast, are inherently more privacy-friendly but may lose the discriminative capacity required for nuanced medical queries, such as differentiating between closely related clinical phenotypes.

The distributed retrieval index is built upon the encrypted or perturbed binary codes and is implemented across institutional data silos to maintain data sovereignty. Queries originating from the healthcare AI agent are hashed using the same pre-trained encoder and privacy layer, then routed to the index nodes via a secure query protocol. Candidate retrieval employs Hamming distance computation, possibly accelerated through multi-index hashing or inverted bit-slice tables, and the top-k candidates are aggregated and re-ranked using a secure comparison protocol if needed. The index partitions can be organized along temporal, geographic, or specialty-based dimensions to reflect natural clinical boundaries, thereby aligning with federated data stewardship principles.

The AI agent interface enriches the retrieval pipeline with an adversarial defense module that inspects the query embedding and its binary projection before index dispatch. This module employs a lightweight detector network trained to distinguish benign query variations from adversarially perturbed inputs that aim to shift retrieval rankings toward harmful or irrelevant documents. Additionally, a consensus mechanism over multiple hash tables derived through independent randomized training runs can be used to validate retrieval consistency and suppress results that appear only under suspect patterns. The defense layer thus serves as a gatekeeper that prevents adversarial manipulation from propagating into the knowledge-grounded reasoning cycles of the clinical agent.

The system architecture embodies multiple structural trade-offs. Increasing the number of hash bits enhances retrieval precision but requires more extensive privacy noise to maintain

guarantees, thereby partially eroding the gain. The adversarial defense module adds latency to each query, which must be balanced against the real-time requirements of clinical workflows. Decentralizing the index across institutional silos improves privacy and data sovereignty but complicates the adversarial defense since a malicious insider could poison a local partition. These tensions are not resolvable through a single optimal configuration; rather, the architecture must support parameterized deployment profiles that can be tuned to the risk tolerance, computational budget, and regulatory context of each healthcare ecosystem.

4. Self-Supervised Hash Learning for Semantic Representation

The hash encoder is the semantic core of the retrieval system, and its design critically determines both the fidelity of knowledge access and the vulnerability surface. Self-supervised learning has emerged as the preferred training paradigm because it enables the encoder to capture rich medical semantics without relying on fragile and labor-intensive supervision labels. The training process constructs positive pairs from augmented versions of the same medical knowledge fragment—for instance, by applying synonym replacement, numerical perturbation within clinically safe bounds, or graph rewiring that preserves logical consistency—while treating other fragments within the batch as negative examples. A contrastive loss, adapted to the Hamming space via a continuous relaxation of binary constraints, encourages the encoder to assign similar binary codes to semantically aligned instances and dissimilar codes to unrelated ones.

The use of asymmetric semantic excavation addresses a key challenge in medical hashing: the distribution of similarities in medical knowledge is highly skewed, with most documents being semantically distant and only a small fraction being highly relevant to a given query. Traditional symmetric contrastive objectives can produce gradients that are dominated by easy negatives, leading to hash codes that collapse into coarse clusters and fail to capture subtle clinical distinctions. Asymmetric designs that treat the query and document representations through separate pathways, combined with margin-scalable constraints that adjust the penalty based on the relative similarity rank, have been shown to produce hash codes with superior local semantic structure [3]. In the medical domain, this translates into the ability to distinguish between, for example, hypertension management guidelines for patients with and without comorbid diabetes, where the distinction may hinge on a few key sentences.

The hash encoder architecture can be instantiated using a stack of transformer layers followed by a bottleneck projection with sign activation or continuous relaxation during training. Pre-training on large-scale de-identified clinical corpora, including clinical notes and biomedical literature, provides a strong initialization that captures domain-specific terminology, abbreviations, and relational patterns [21]. Fine-tuning through self-supervised hashing objectives on specialized knowledge bases—such as disease-specific guidelines or pharmacological databases—yields compact binary representations that remain aligned with the underlying medical semantics. The bit independence assumption often made in hashing research is, in practice, relaxed; correlated bits can be permitted if the privacy enforcement layer is designed to account for such correlations through appropriate noise calibration.

An important consideration is the tension between hash code specificity and privacy. Self-supervised hashing naturally tends to produce codes that are functions of the semantic content; if the medical knowledge base contains rare concepts or unique phrasing, the corresponding binary codes can become quasi-identifiers. The privacy enforcement layer must therefore be co-optimized with the hashing objective to ensure that the guarantees hold under any plausible adversarial knowledge. This co-design can be operationalized through a training-

time perturbation that simulates the noise that will be added during deployment, effectively regularizing the encoder toward smoother, more privacy-preserving mappings. The result is a hash space where clinically meaningful neighborhoods are preserved while the exact reconstruction of input content becomes provably difficult.

5. Adversarial Defense Mechanisms for Healthcare AI Agents

Adversarial threats in medical knowledge retrieval arise at multiple interfaces. Input-space attacks craft query perturbations that are designed to mislead the hash encoder into producing a binary code that is close to maliciously chosen knowledge entries. For example, an attacker could introduce subtle synonym swaps or numerical modifications that shift a query about safe dosage ranges into the vicinity of documents describing toxic levels. Poisoning attacks target the knowledge base itself, inserting or modifying entries so that legitimate queries retrieve compromised content under certain conditions. Backdoor triggers, often hidden in very specific query patterns, can activate systematic failures that are difficult to detect through random testing. These threat vectors are especially dangerous in healthcare AI agent settings because the agent's subsequent reasoning and natural language generation will treat the retrieved evidence as authoritative, potentially resulting in harmful recommendations that erode clinician trust.

The proposed adversarial defense module is composed of several complementary layers. Query sanitization relies on a detector network trained to flag inputs that lie far from the distribution of benign clinical queries in the encoder's embedding space or that exhibit anomalous hash code sensitivity—measured by the Hamming distance change induced by small controlled perturbations. Such input-level monitoring can catch many gradient-based attacks that rely on subtle but systematic deviations. The second defense layer operates within the retrieval index through robust aggregation: multiple independently trained hash tables derived from different random initializations or data augmentations are queried in parallel, and only documents that consistently rank high across tables are returned. This consensus mechanism raises the cost for an adversary to mount a successful poisoning attack, because the attacker must simultaneously corrupt multiple diverse representations.

Adversarial training of the hash encoder itself provides a further layer of resilience. During self-supervised learning, adversarial perturbations are generated on the fly against the current encoder and included as negatives or as hard positives in the contrastive loss, forcing the encoder to learn representations that are stable under input variation. However, adversarial training in the discrete binary code domain poses unique challenges because gradient propagation through the thresholding operation is ill-defined; methods based on the straight-through estimator with careful noise injection have shown promise in maintaining both robustness and retrieval accuracy. The interplay between adversarial training and privacy is delicate: aggressive adversarial training can sharpen the hash code distributions, potentially weakening differential privacy guarantees unless the noise scale is adjusted accordingly.

The security of the large language model agent that consumes the retrieval results must be considered as an integral part of the defense architecture. Recent investigations into security enhancement for adversarial robust LLM agents in medical decision-making highlight that retrieval serves as both a strength and a vulnerability point [17]. If the agent has been fine-tuned to rely heavily on retrieval outputs, an attacker who compromises the retrieval pipeline can manipulate the agent's entire decision trajectory. Mitigation strategies include retrieval confidence calibration, where the agent learns to down-weight retrieved evidence that

originates from index partitions with suspicious patterns, and explicit reasoning traces that expose the retrieval evidence for human or automated audit in high-stakes scenarios.

A particularly subtle class of attacks targets fairness through retrieval manipulation. An adversary could engineer queries or poison indexes to systematically disadvantage specific demographic groups—for example, by suppressing evidence necessary for correct diagnosis in maternal health conditions that predominantly affect certain populations. Defending against fairness-oriented attacks requires the retrieval system to incorporate demographic-aware auditing that continuously tests for disparate retrieval quality across population subgroups. This defensive objective links robustness directly to governance concerns and motivates the integration of fairness metrics into the adversarial defense monitoring loop.

6. Governance, Fairness, and Deployment Considerations

Deploying privacy-preserving hash-based retrieval with adversarial defense in real-world healthcare settings introduces governance challenges that span technical, organizational, and regulatory domains. The architecture's reliance on distributed index partitions across institutional silos aligns with the data sovereignty requirements of regulations such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act, but it simultaneously complicates the task of auditing the retrieval pipeline for compliance and safety. Healthcare providers, regulators, and patients require verifiable assurances that the retrieval decisions are explainable, free from unlawful bias, and resilient to manipulation. Providing such assurances without exposing the underlying raw data or the exact hash codes—which could themselves be considered protected health information under certain interpretations—demands new approaches to zero-knowledge auditing and secure logging.

Algorithmic fairness in medical knowledge retrieval is a multidimensional concern. Self-supervised hash learning can inadvertently amplify existing biases in medical knowledge bases, such as the overrepresentation of research findings from high-income countries or the historical underrepresentation of certain demographic groups in clinical trials. If these biases are embedded in the binary code similarity structures, queries about conditions that manifest differently across populations may retrieve evidence that is less relevant or even misleading for underrepresented groups. Addressing this requires bias-aware self-supervised objectives that incorporate fairness constraints during training, as well as post-deployment monitoring that tracks retrieval disparities using available demographic metadata in a privacy-compliant manner [18]. Fairness interventions must themselves be privacy-preserving; for example, stratified auditing can be performed using secure aggregation to compute per-group retrieval metrics without revealing individual-level attributes.

The sustainability of large-scale hash-based retrieval infrastructure is an underappreciated dimension. Medical knowledge evolves rapidly, with new clinical guidelines, drug approvals, and research findings emerging continuously. The hash encoder and index must support incremental updates without full retraining, yet introducing new knowledge fragments or pruning outdated ones can shift the hash code distribution and degrade retrieval quality over time. The energy footprint of maintaining and querying an encrypted distributed hash index is non-trivial, and the computational cost of adversarial defense must be weighed against the environmental and financial sustainability targets of healthcare institutions [22]. Green retrieval strategies, including knowledge distillation of hash encoders and adaptive query routing that minimizes index node activation, represent important research directions.

Governance frameworks for healthcare AI agents must evolve to explicitly address the retrieval component. Current regulatory guidance predominantly focuses on the decision-making model itself, often treating the retrieval step as an internal implementation detail. However, when retrieval meaningfully shapes clinical recommendations, it warrants independent scrutiny. Certification schemes could require standardized adversarial stress testing of retrieval pipelines, privacy loss accounting over query histories, and fairness impact assessments tailored to the knowledge domains served. The establishment of multi-stakeholder governance bodies that include clinicians, patients, informaticians, and security experts would support the creation of context-sensitive standards that respect clinical workflow constraints while maintaining rigorous protections.

Cross-border deployment introduces additional complexity because the legal norms governing health data privacy and algorithmic accountability differ substantially across jurisdictions. An architecture that separates the hash encoder training—performed within a trusted computing environment in one region—from the distributed index hosting and query execution—which may be geographically dispersed—must navigate conflicting legal requirements. Emerging techniques in policy-aware cryptographic protocols may allow the system to dynamically enforce jurisdictional rules by restricting certain retrieval operations based on the query's provenance. Such policy enforcement layers must be engineered with the same rigor as privacy and defense mechanisms to ensure that compliance does not become a second-class design consideration.

The interplay between human oversight and autonomous retrieval is a critical governance dimension. In high-acuity clinical settings, the retrieval pipeline should support an adjustable autonomy model where the level of human review scales with the risk associated with the retrieved evidence. Adversarial defense metrics, such as the consensus score across hash tables and the detector's confidence, can serve as signals that trigger human review thresholds. Designing this human-machine interface in a way that does not induce alert fatigue while maintaining meaningful oversight requires careful human factors engineering, a consideration that system architects must incorporate from the earliest design phases [12]. Ultimately, the goal is to embed the retrieval system within a broader socio-technical fabric that respects clinical judgment, patient rights, and the evolving landscape of medical knowledge.

7. Conclusion

Privacy-preserving medical knowledge retrieval for healthcare AI agents demands an integrated approach that treats self-supervised hash learning, cryptographic privacy enforcement, and adversarial defense as co-equal pillars of a trustworthy system. The architectural investigation presented in this paper reveals that the design space is characterized by fundamental trade-offs: the bit length and specificity of hash codes influence both retrieval accuracy and privacy leakage; adversarial robustness measures increase computational cost and must be carefully balanced against real-time clinical requirements; and distributed index designs that respect data sovereignty complicate the auditing and fairness verification processes. Self-supervised hash learning, particularly when enhanced by asymmetric semantic excavation and margin-scalable constraints, provides a strong semantic foundation that can be hardened through differential privacy noise injection and secure index protocols. Adversarial defense, spanning query sanitization, consensus retrieval, and encoder adversarial training, guards against the growing threat surface faced by retrieval-dependent clinical agents. The path to deployment requires not only technical refinement but also the parallel development of governance structures, auditing methodologies, and sustainability

practices that align with the norms and values of healthcare communities. By framing retrieval as a systemic capability that must be designed, monitored, and governed through a combination of machine learning, cryptography, and policy, this work contributes a holistic perspective on building the next generation of privacy-respecting, robust, and equitable healthcare AI infrastructures.

References

1. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3, 119.

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

3. 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.

4. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers), 31–36.

5. Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8018-8025.

6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

7. Wong, W. K., Cheung, D. W., Kao, B., & Mamoulis, N. (2009). Secure kNN computation on encrypted databases. Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data, 139-152.

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

9. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the 37th International Conference on Machine Learning, 1597-1607.

10. Zhang, H., Zhang, J., Lu, G., & Zhang, D. (2021). Asymmetric deep hashing for large-scale histopathological image retrieval. Computers in Biology and Medicine, 137, 104809.

11. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Theory of Cryptography Conference, 265-284.

12. He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nature Medicine, 25(1), 30-36.

13. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. Proceedings of the 38th IEEE Symposium on Security and Privacy, 19-38.

14. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

15. Shen, D., Su, Q., Chapfuwa, P., Wang, W., Wang, G., Henao, R., & Carin, L. (2018). NASH: Toward end-to-end neural architecture for generative semantic hashing. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 795-805.

16. Yang, E., Liu, T., Deng, C., & Tao, D. (2020). Adversarial examples for image retrieval. IEEE Transactions on Image Processing, 29, 7565-7577.

17. 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.

18. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

19. Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260, 113172.

20. Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2014). Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Transactions on Parallel and Distributed Systems, 25(1), 222-233.

21. Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. B. A. (2019). Publicly available clinical BERT embeddings. Proceedings of the 2nd Clinical Natural Language Processing Workshop, 72-78.

22. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645-3650.

23. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. Proceedings of the 1st IEEE European Symposium on Security and Privacy, 372-387.

24. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175-1191.

25. Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. Scientific Reports, 7, 5994.