Adversarially Robust Cross-Modal Medical Image Retrieval via Self-Supervised Deep Hashing and Large Language Model Agents

Hristopher Gell
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
gellhristopher@uab.edu

Rennis Karlsson
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
rennis2005@missouri.edu

Enzo M. Koskinen
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
koskinen538@binghamton.edu

Abstract

Cross-modal medical image retrieval supports clinical decision-making by enabling semantically grounded queries across heterogeneous data sources, yet its deployment in real-world settings faces substantial challenges from adversarial perturbations, modality gaps, and governance demands. This paper advances a system-level framework that integrates self-supervised deep hashing with large language model (LLM) agents to achieve adversarially robust, semantically precise retrieval. The architecture couples a self-supervised hashing backbone that learns modality-agnostic binary codes from unpaired radiological reports and imaging studies with an LLM-based semantic mediator that reformulates queries, validates retrieved candidates, and injects domain constraints. A threat model encompassing modality-specific gradient-based perturbations, linguistic prompt injection, and distributional drift is formally characterized. Defense mechanisms are woven throughout the retrieval pipeline, including adversarial hashing via margin-scalable constraints, randomized smoothing for certified robustness, and prompt sanitization layers within the LLM agent. The discussion emphasizes structural trade-offs among retrieval latency, bit-width efficiency, and robustness guarantees. Governance implications are analyzed with regard to fairness across patient subpopulations, audit trails for retrieval decisions, energy sustainability of dual-stage architectures, and compliance with evolving regulatory frameworks for AI-enabled medical devices. By treating robustness and governance not as afterthoughts but as design constraints embedded within the self-supervised training loop and the agent orchestration, the framework aims to bridge the gap between laboratory validation and trustworthy clinical deployment. This systems-oriented synthesis highlights open challenges including continual adaptation under domain shift, scalable federated hash learning, and the need for standardized benchmark protocols that jointly evaluate retrieval accuracy, adversarial resilience, and fairness metrics in cross-modal medical search.

Keywords

adversarial robustness, cross-modal retrieval, medical imaging, self-supervised learning, deep hashing, large language model agents, healthcare AI governance.

1. Introduction

Cross-modal medical image retrieval constitutes a cornerstone capability for data-driven clinical workflows, linking visual findings in radiological scans with textual reports, genomic profiles, and structured electronic health records. The clinical utility of such systems hinges on their ability to retrieve semantically congruent instances across heterogeneous modalities under strict latency constraints, thereby supporting differential diagnosis, treatment planning, and case-based reasoning. Despite decades of progress in content-based retrieval, the translation into routine clinical practice has been hampered by vulnerabilities that are often overlooked in academic benchmarks. Most critically, the reliance on deep neural encoders exposes retrieval pipelines to adversarial perturbations, where imperceptibly crafted image noise or subtly altered textual queries can catastrophically degrade retrieval quality. Safety-critical medical environments demand a shift from performance-centric optimization toward robustness-first architectural design, wherein adversarial resilience, fairness, and governance are integral system properties rather than auxiliary evaluations.

The convergence of self-supervised deep hashing and large language model agents offers a promising pathway to reconcile efficiency, semantic expressiveness, and adversarial robustness. Hashing-based retrieval compresses high-dimensional multimodal representations into compact binary codes, enabling sub-linear search over massive clinical repositories. Self-supervised paradigms leverage the natural co-occurrence of images and reports without requiring costly manual annotation, making large-scale pre-training on institutional data feasible under privacy constraints. Concurrently, LLM agents introduce a semantic mediation layer capable of reformulating ambiguous clinical queries, performing chain-of-thought validation over retrieved candidates, and enforcing domain-specific safety constraints. At the same time, these agents expand the threat surface through prompt injection and adversarial linguistic perturbations, motivating a holistic treatment of robustness that spans both sensory and symbolic modalities. This paper articulates a system-level framework that integrates these components under a unified threat model and governance architecture, examining the structural trade-offs, deployment considerations, and policy implications that arise when adversarial robustness is elevated from a post-hoc patch to a first-class design principle in cross-modal medical retrieval.

2. Background and Related Work

Content-based medical image retrieval has evolved from handcrafted feature pipelines to end-to-end deep learning systems that jointly model visual and textual modalities. Early reviews established the clinical motivations, noting that similar-appearing lesions in imaging repositories often correlate with similar diagnoses, treatment responses, and outcomes, making retrieval a natural paradigm for evidence-based reasoning [1]. However, the vulnerability of deep visual encoders to adversarial examples was starkly demonstrated in medical contexts, where barely perceptible modifications to chest radiographs could invert malignancy predictions and misdirect retrieval results, underscoring the acute safety implications for machine learning in clinical environments [2]. This dual concern—semantic richness and adversarial fragility—sets the stage for contemporary system design.

Deep hashing emerged as a practical response to the scalability demands of large-scale retrieval, learning to map high-dimensional features into compact Hamming space where

similarity is preserved and search is accelerated via bitwise operations [3]. The application of self-supervised learning further reduced dependency on curated annotations by designing pretext tasks that exploit inherent multimodal alignment, enabling models to capture clinically meaningful structure from unlabeled or weakly paired data at scale [4]. In parallel, large language models have demonstrated remarkable capacities for encoding clinical knowledge and performing zero-shot reasoning over medical narratives [5]. These capabilities suggest a natural role for LLM agents as query understanding components, yet their deployment in healthcare also raises ethical challenges concerning accountability, bias amplification, and the opacity of chain-of-thought rationales [6].

Within the specific subfield of cross-modal hashing, architectures that learn aligned binary spaces for images and text have shown substantial improvements in retrieval accuracy on benchmark datasets [7]. Recent work has also begun to address the adversarial resilience of LLM-based agents in medical decision-making tasks, providing security enhancement methods that harden these agents against intentional input perturbations without sacrificing clinical utility [8]. On the hashing side, self-supervised techniques incorporating asymmetric semantic excavation and margin-scalable constraints have pushed the state of the art in code quality, demonstrating that robust hash functions can be learned without exhaustive pairwise supervision [9]. A broader line of adversarial robustness research has produced foundational techniques such as projected gradient descent training and certified defenses, which are applicable across vision and language domains but have rarely been jointly applied to integrated retrieval-agent architectures [10]. Moreover, equity-oriented analyses have revealed that clinical decision support algorithms can perpetuate racial and socioeconomic disparities when trained on historically skewed data, highlighting the necessity of embedding fairness constraints directly into the retrieval and ranking pipeline [11]. Privacy-preserving collaborative learning approaches, notably federated learning frameworks tailored for medical imaging, have begun to address the tension between data centralization for robust model training and the regulatory imperative to keep sensitive patient data within institutional perimeters [12]. Together, these strands of research provide the enabling technologies and cautionary insights that inform the integrated architecture proposed in this work.

3. System Architecture and Adversarial Threat Model

The proposed framework is structured as a three-layer pipeline: a self-supervised cross-modal hashing encoder, a binary code index spanning distributed repository partitions, and an LLM agent serving as a semantic mediator at the query interface. The hashing layer ingests unpaired chest radiographs, computed tomography slices, and corresponding radiology reports, projecting each modality into a shared Hamming space of configurable bit width, typically ranging from 64 to 256 bits. The index layer maintains multiple locality-sensitive hash tables to support approximate nearest neighbor search with sub-linear complexity, while the LLM agent intercepts user-supplied queries expressed in natural clinical language, expands them with synonym sets, resolves acronyms through context-aware prompt engineering, and generates a shortlist of candidate binary codes for verification against inclusion-exclusion criteria derived from clinical guidelines. This layered separation enforces a clean abstraction boundary between representational learning, efficient retrieval mechanics, and cognitive-semantic reasoning, facilitating modular audits and incremental upgrades without destabilizing the entire system.

Adversarial threats are categorized along three axes: visual perturbation attacks, linguistic prompt injection, and distributional drift. Visual adversaries craft imperceptible gradient-

based noise targeted at the image encoder, aiming to push the resulting hash code away from its semantically correct neighborhood while remaining undetectable to human readers. Prompt injection attacks exploit the LLM mediator by inserting concealed directives that override the intended retrieval semantics, potentially exfiltrating patient data or bypassing safety filters. Distributional drift encompasses systematic shifts in imaging protocols, demographic composition, or report phrasing that degrade model performance over time without any explicit malicious intervention. The threat model assumes a gray-box adversary with partial access to the encoder architecture but no knowledge of the hashing random seed or the LLM's dynamic system prompt. This adversary can query the retrieval API and observe ranked results but cannot directly read the index or modify stored codes. Defensive measures are distributed across the architecture: adversarial training of the hashing backbone using perturbed image-report pairs, margin-scalable constraints that bound the sensitivity of binary codes to input variations, input sanitization and output verification layers within the LLM agent, and cryptographic hash-based integrity checks on the index to detect tampering or poisoning. By modeling threats at each architectural boundary, the design moves beyond holistic black-box testing toward component-level resilience guarantees.

4. Self-Supervised Deep Hashing for Cross-Modal Embedding

The hashing backbone is trained under a multi-task self-supervised objective that combines contrastive alignment across modalities with intra-modal invariance constraints. Pairs of radiology images and their corresponding reports are treated as positive examples, while randomly sampled mismatches form negatives. The encoder, structured as a dual-branch network with a vision transformer for imaging and a clinical text transformer for reports, projects both modalities into a common latent space prior to binarization. Crucially, the training loop does not require explicit disease labels, instead exploiting the natural co-occurrence structure of Picture Archiving and Communication Systems (PACS) data, where images and reports are linked by accession number. This self-supervised design allows the system to leverage the vast repositories of historical studies accumulated within academic medical centers without imposing retrospective annotation burdens, which are often infeasible at scale.

Binary code generation employs a sign function during inference but is approximated by a continuous relaxation during training to permit gradient flow. The hashing loss explicitly enforces a margin between the similarity scores of relevant pairs and the highest-scoring irrelevant pair within a mini-batch, with an adaptive scaling factor that increases penalization as the model converges, tightening the decision boundary. This margin-scalable scheme, informed by principles that have proven effective in asymmetric self-supervised hashing, reduces the incidence of semantically ambiguous codes near the Hamming boundaries, thereby improving robustness to minor input perturbations. Moreover, the bit width serves as a tunable structural parameter that governs the fundamental trade-off between retrieval granularity and the adversarial attack surface: narrower codes are inherently more resistant to bit-flip attacks because each dimension aggregates information from larger input regions, but they sacrifice discriminative capacity for fine-grained clinical distinctions. Experimental regimes examining bit widths between 48 and 256 bits reveal a U-shaped risk profile, where the optimal operating point depends on the anticipated perturbation budget and the tolerance for false positives in high-stakes diagnostic contexts.

To counteract modality-specific adversarial perturbations, the hashing backbone incorporates an adversarial training phase in which image and text inputs are perturbed under projected

gradient constraints while maximizing the Hamming distance between the original and perturbed codes. At the same time, semantic consistency is enforced by requiring that the perturbed instance remain aligned with the paired modality from the other branch. This dual-constraint formulation ensures that robustness improvements do not come at the expense of cross-modal alignment fidelity. The resulting hash functions exhibit Lipschitz-continuous behaviour in Hamming space, meaning that small input changes cannot cause disproportionate code flips, a property that can be empirically verified through local sensitivity analysis over clinical validation sets. The self-supervised pre-training also facilitates rapid domain adaptation when institutional imaging protocols change, as the model need only be fine-tuned on unlabeled new-domain pairs, a pragmatic advantage for sustained deployment.

5. Large Language Model Agents for Semantic Mediation

The LLM agent serves as an intelligent intermediary, interpreting free-text queries that reflect the natural diversity of clinical language and translating them into structured retrieval directives. A cardiologist might query “find similar cases to this echo with septal wall motion abnormality and mild TR,” while a resident in a different institution might phrase an equivalent information need using distinct terminology. The agent’s role encompasses canonicalization of clinical entities to standard ontologies, expansion using hierarchical concept relationships, and dynamic weighting of retrieval criteria based on the perceived diagnostic intent. This mediation layer not only improves recall for under-specified queries but also provides an opportunity to embed clinical safety rules—for example, verifying that retrieved cases are drawn from populations demographically relevant to the patient at hand, or flagging when a query inadvertently requests comparisons across incompatible anatomies.

However, the integration of LLM agents introduces novel adversarial vulnerabilities that differ in nature from pixel-space perturbations. Adversaries may embed malicious instructions within query text, attempting to override system prompts and compel the agent to ignore safety filters, disclose institutional private information encoded in its few-shot examples, or retrieve results that are medically misleading. Prompt injection resilience is achieved through a multi-layered defense: an input sanitizer normalizes encoding, strips non-printable characters, and applies a learned filter that detects semantic divergence from expected clinical query distributions; the agent’s core prompt includes immutable guardrails that cannot be overridden by concatenated user input; and a verification stage re-evaluates whether the agent’s final retrieval plan is consistent with the original query semantics, using a secondary lightweight intent classifier. These measures draw from recent advances in secure agent design tailored to medical contexts, where the stakes of adversarial manipulation include patient harm [8]. The architectural separation of the LLM agent from the hashing index also restricts the blast radius of a compromised agent, as the agent cannot directly alter stored binary codes or index parameters.

Extending beyond query interpretation, the LLM agent contributes to retrieval reliability through chain-of-thought validation. When the initial ranked list is returned, the agent performs a structured review by comparing retrieved study descriptions against the original clinical question, identifying potential false positives arising from confounded anatomy or spurious text-image correlations. This validation step produces a confidence-calibrated re-ranking and, critically, an audit log that documents the clinical rationale for each inclusion or exclusion decision. The audit log serves not only for retrospective accountability but also as a training signal for continual improvement of the hashing encoder, creating a virtuous cycle in

which semantic feedback from the linguistic layer refines the visual-semantic embedding over time. This human-in-the-loop—or more precisely, agent-in-the-loop—design philosophy aligns with emerging regulatory expectations for explainability in medical device software.

6. Adversarial Robustness and Defense Integration

Achieving end-to-end robustness requires coordinated hardening of both the hashing encoder and the LLM mediator under a unified adversarial objective. The hashing backbone is fortified through a combination of adversarial training and certified smoothing. During adversarial training, image report pairs are augmented with worst-case perturbations generated under a constrained optimization that maximizes Hamming distortion while preserving structural medical plausibility, assessed via organ segmentation overlap and anatomical fidelity metrics. For certified robustness, randomized smoothing is applied at hash code inference: multiple noisy copies of the input are hashed, and a majority vote determines each bit, yielding a probabilistic guarantee that the Hamming distance between adversary-perturbed codes remains bounded with high confidence. This smoothing incurs a computational overhead proportional to the number of samples, yet because hashing inference is lightweight, the latency penalty is tolerable within clinical retrieval workflows that already accommodate multi-second turnaround for cognitive reasoning.

On the linguistic side, robustness is enhanced through adversarial augmentation of the agent's training prompts, including paraphrasing attacks and semantically equivalent adversarial rephrasing generated by a white-box LLM adversary during red-teaming exercises. The input sanitizer and intent classifier are co-trained to distinguish legitimate clinical queries from adversarial injections, using a curated dataset that blends authentic radiology requisition phrases with synthetically generated attacks spanning insertion, deletion, and substitution patterns. Certified robustness for the agent is less mature than for vision, but recent techniques based on interpretable region verification around embedding representations provide a path toward formal guarantees against bounded text perturbations, and these can be integrated as a runtime monitor that aborts retrieval when an input falls outside the certified envelope. The modular architecture ensures that advances in certified linguistic robustness can be adopted independently of the hashing substrate.

Beyond reactive defenses, the system incorporates proactive monitoring for distributional drift through statistical process control applied to the distribution of binary code population statistics. A gradual shift in the mean Hamming distance between query codes and index codes across a stream of clinical queries may signal a change in imaging equipment, a new scanner vendor, or a demographic shift, any of which can compromise retrieval accuracy before a full adversarial attack is mounted. When drift is detected, the system triggers automated re-calibration of the hashing encoders via self-supervised fine-tuning on a sliding window of recent unlabeled studies, a mechanism that operationalizes the continual learning capacity inherent to the self-supervised paradigm. This closed-loop adaptation ensures that robustness is maintained as a dynamic property rather than a one-time certification, aligning with the reality of evolving clinical environments.

7. Governance, Fairness, and Deployment Considerations

Deploying an adversarially robust cross-modal retrieval system in a clinical enterprise demands navigation of a complex governance landscape encompassing regulatory approval, algorithmic fairness, data stewardship, and environmental sustainability. Regulatory bodies increasingly view AI-enabled medical retrieval tools as software as a medical device,

requiring evidence of safety and effectiveness under representative adversarial stress conditions. The architecture's audit logging and agent-driven verification directly support compliance with documentation requirements, as every retrieval decision can be traced to a human-readable clinical rationale or a machine-generated confidence metric. Nonetheless, standardization of adversarial resilience benchmarks for retrieval tasks remains an open challenge, as existing radiology AI validation protocols predominantly address classification and segmentation accuracy without systematically evaluating robustness to input perturbations or prompt injection.

Fairness across patient populations is a critical cross-cutting concern. Self-supervised training on institutional PACS data can inadvertently encode historical disparities in imaging utilization, report language, and disease prevalence, causing retrieval systems to exhibit differential performance across racial, gender, and socioeconomic subgroups. Mitigation strategies include stratified sampling during pre-training to balance representation, adversarial debiasing of the shared embedding space to remove spurious correlations between demographic attributes and hash code geometry, and LLM-agent instructions that explicitly prompt consideration of equity when shortlisting cases for clinical comparison. The agent's audit trail serves as a continuous monitoring mechanism, enabling periodic fairness audits that measure recall parity and false positive rate equality across subgroups. Embedding fairness as a requirement at the architectural level—rather than as a post-processing adjustment—reflects a growing consensus that equity must be inseparable from safety in medical AI governance.

Infrastructure sustainability introduces another layer of structural trade-offs. The dual-stage design incurs energy costs both from training the self-supervised hashing backbone and from inference-time LLM calls. While hashing inference is lightweight, the LLM mediation layer's energy footprint is non-trivial, especially when chain-of-thought reasoning and generation of multiple candidate re-rankings are factored in. Strategies to mitigate this include caching common clinical query templates, distilling the LLM into a smaller specialized model fine-tuned on institutional query logs, and routing simple retrieval requests directly to the index without agent mediation when the query intent is unambiguously classifiable. These operational decisions have sustainability implications that must be disclosed in environmental impact assessments accompanying large-scale clinical AI deployments.

The federated architecture further complicates governance, as hash code indices may be distributed across hospital systems for privacy-preserving search. Federated deep hashing, where encoders are trained collaboratively without raw data exchange, holds great promise but introduces new threats from model inversion and gradient leakage that must be modelled alongside the primary retrieval adversary. The integration of robust aggregation techniques and differential privacy into the federated hashing protocol ensures that the adversarial robustness guarantees extend to multi-institutional query settings without compromising patient confidentiality. Policy frameworks such as the European Health Data Space and evolving FDA guidance on adaptive AI systems will heavily influence the pace and shape of deployment, underscoring the need for close collaboration between system architects and clinical governance bodies throughout the design lifecycle.

8. Conclusion

This paper has presented a cohesive framework for adversarially robust cross-modal medical image retrieval that embeds self-supervised deep hashing and LLM-based semantic agents within a unified threat-aware architecture. By jointly considering visual perturbation robustness, prompt injection resilience, fairness constraints, and governance infrastructure,

the approach moves beyond isolated accuracy benchmarks toward a systems-level conception of trustworthy clinical retrieval. The structural separation of hashing, indexing, and semantic mediation enables modular adoption of robustness enhancements while preserving the latency and storage efficiencies essential for practical deployment. Self-supervised training paradigms reduce annotation costs and facilitate continuous adaptation under distributional drift, while LLM agents provide interpretable validation loops that simultaneously improve retrieval quality and generate audit-ready clinical rationales. Adversarial training, certified smoothing, and prompt sanitization are interwoven as defense layers, each addressing a specific threat surface within the end-to-end pipeline. The discussion of fairness, sustainability, and federated governance highlights that technical robustness cannot be divorced from the broader socio-technical environment in which medical AI operates. Future work should pursue standardized adversarial benchmark suites for cross-modal retrieval that encompass demographic fairness and linguistic injection attacks alongside pixel-space perturbations, as well as novel protocols for certifying the joint vision-language robustness of integrated retrieval-agent systems. By treating adversarial resilience as a foundational design principle rather than a post-hoc hardening task, the field can advance toward retrieval systems worthy of the trust that clinical users and patients place in them.

References

1. Müller, H., Michoux, N., Bandon, D., & Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. International Journal of Medical Informatics, 73(1), 1–23.

2. 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.

3. Liu, H., Wang, R., Shan, S., & Chen, X. (2016). Deep supervised hashing for fast image retrieval. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2064–2072.

4. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Kornblith, S., Chen, T., & Norouzi, M. (2021). Big self-supervised models advance medical image classification. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3478–3488.

5. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Schärli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

6. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.

7. Jiang, Q. Y., & Li, W. J. (2017). Deep cross-modal hashing. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3232–3240.

8. 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.

9. 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.

10. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. Proceedings of the International Conference on Learning Representations (ICLR).

11. 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.

12. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

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

14. Wang, J., Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2018). A survey on learning to hash. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4), 769–790.

15. Yang, X., Chen, A., PourNejatian, N., Shin, H. C., Smith, K. E., Parisien, C., Compas, C., Martin, C., Flores, M. G., Zhang, Y., Magoc, T., Harle, C. A., Lipori, G., Mitchell, D. A., Hogan, W. R., Shenkman, E. A., Bian, J., & Wu, Y. (2023). Large language models in medicine: the potentials and pitfalls. arXiv preprint arXiv:2309.10980.

16. Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. Proceedings of the International Conference on Machine Learning (ICML), 1310–1320.

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

18. 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 (ACL), 3645–3650.

19. Eslami, S., de Melo, G., & Meinel, C. (2023). MedCLIP: Contrastive learning from unpaired medical images and text. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 3876–3887.

20. Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S. M., Leskovec, J., Kundaje, A., … Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. Proceedings of the International Conference on Machine Learning (ICML), 5637–5664.