Context-Aware Multimodal Recommendation Framework for Personalized Healthcare Information Retrieval Using Large Language Models

Isaac C. Andrews
Department of Computer Science, University of North Texas, Denton, TX, USA.
icandrews@unt.edu

Ranjemin Lawrence
School of Computing, Clemson University, Clemson, SC, USA.
renjamin.work@clemson.edu

Kiran A. Batra
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
kiranabatra@colostate.edu

Arthur Baker
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
contactarthur@uab.edu

Abstract

The rapid proliferation of large language models has catalyzed a transformation in healthcare information retrieval, yet most systems remain tethered to unimodal textual inputs and static relevance models that disregard temporal, situational, and individual patient contexts. This paper presents a context-aware multimodal recommendation framework that integrates clinical text, medical imaging, structured patient records, and environmental sensor streams into a unified retrieval and personalization pipeline orchestrated by large language models. Drawing upon system-level design principles, the framework employs modular encoders, cross-modal attention fusion, and adaptive context encoders to generate rich representations of user information needs. A dynamic personalization layer continuously refines recommendations by modeling longitudinal health profiles, activity patterns, and domain-specific preferences without retaining raw sensitive data. The architecture is evaluated through the lenses of scalability, infrastructure governance, fairness, robustness, and sustainability. We discuss structural trade-offs between centralized large-model inference and edge-distributed fine-tuning, the integration of federated learning with differential privacy, and the challenges of maintaining retrieval quality under evolving clinical knowledge. Policy implications surrounding algorithmic accountability, data sovereignty, and equitable access are examined, emphasizing that technical design choices directly shape socio-technical outcomes. The framework is positioned as a composite infrastructure that reconciles cutting-edge multimodal understanding with the stringent demands of healthcare ecosystems, offering a reference architecture for deployable, responsible, and personalized medical information systems.

Keywords

multimodal retrieval, large language models, healthcare personalization, context-aware systems, federated learning, algorithmic fairness, infrastructure governance.

1. Introduction

Contemporary healthcare systems generate an overwhelming volume of heterogeneous data, including clinical notes, diagnostic images, laboratory reports, genomic sequences, and continuous streams from wearable devices. Clinicians and patients alike face the formidable challenge of distilling relevant evidence from this multidimensional information landscape to support timely and accurate decision-making. Large language models have recently demonstrated remarkable capabilities in understanding and generating natural language, and their adaptation to the biomedical domain has opened new frontiers for medical question answering, literature summarization, and clinical decision support [1, 2, 3]. However, the majority of these applications operate within a unimodal paradigm, processing text alone and neglecting the rich contextual signals embedded in non-textual modalities and in the dynamic circumstances of the information seeker.

Retrieval-augmented generation architectures have emerged as a promising approach to ground language model outputs in external knowledge bases, thereby mitigating hallucination and improving factual reliability [4]. When extended to healthcare, such architectures can retrieve scientific articles, clinical guidelines, or similar patient cases to inform model responses. Yet they rarely incorporate non-textual data during the retrieval phase, nor do they adapt the retrieval strategy to the user's evolving health condition, location, device constraints, or privacy preferences. The result is a one-size-fits-all relevance model that fails to capitalize on the full spectrum of available context, potentially leading to suboptimal recommendations or information overload.

This paper proposes a context-aware multimodal recommendation framework designed to address these limitations by treating healthcare information retrieval as a dynamic, multi-faceted personalization problem. The framework centers on a large language model that serves as both a reasoning engine and a multimodal orchestrator, coordinating specialized encoders for text, images, structured records, and time-series data. A context encoding subsystem captures situational, temporal, and user-specific factors, modulating the retrieval and ranking processes. Personalization is achieved through a longitudinal learning component that updates user profiles in a privacy-preserving manner, leveraging federated techniques to keep sensitive health data on local devices. By synthesizing insights from systems engineering, machine learning, and socio-technical scholarship, the paper offers a comprehensive analysis of the architectural decisions, trade-offs, and governance implications that arise when deploying such a framework in real-world healthcare settings.

The remainder of the paper is organized as follows. Section 2 surveys related work in multimodal retrieval, large language models in healthcare, and personalization. Section 3 presents the system architecture and component interactions. Section 4 details multimodal context modeling and fusion. Section 5 discusses personalization mechanisms and adaptation strategies. Section 6 addresses infrastructure, deployment, and sustainability considerations. Section 7 analyzes fairness, accountability, and policy dimensions. Section 8 reflects on evaluation methodologies and robustness challenges. Finally, Section 9 concludes with a synthesis of findings and directions for future work.

2. Related Work

The convergence of deep learning and healthcare informatics has spurred extensive research into representation learning, retrieval, and personalization. Foundational language representation models such as BERT and its biomedical variants demonstrated that domain-adapted pretraining yields substantial gains on clinical natural language processing tasks [5]. Subsequent scaling of transformer architectures led to general-purpose language models capable of performing medical question answering with accuracies approaching those of licensed clinicians when appropriately prompted and fine-tuned [2, 3]. Yet these systems largely rely on static textual corpora and do not incorporate visual or sensor-based inputs that are essential for holistic patient understanding.

Multimodal learning has advanced significantly through architectures that jointly model vision and language. Models like ViLBERT, LXMERT, and CLIP pioneered cross-modal pretraining objectives that align image regions with textual descriptions, enabling downstream retrieval tasks where queries may be expressed in one modality and answers retrieved from another [6, 7]. Extensions to the medical domain have yielded promising results in radiology report generation and chest X-ray interpretation, yet integration with large-scale retrieval pipelines remains nascent. For retrieval, dense passage retrieval and late-interaction models such as ColBERT have demonstrated that encoding queries and documents into dense vectors drastically improves recall compared to sparse methods, especially when combined with approximate nearest neighbor search [8, 11]. Context-aware multimodal retrieval frameworks have also been explored for cross-domain personalization, where local categorical query understanding is fused with user context to disambiguate information needs [9].

Personalization in information systems has historically relied on collaborative filtering and content-based profiling. With the advent of deep learning, sequence models and attention mechanisms have enabled the capture of long-term user interests and session-based dynamics. In healthcare, personalization must navigate stricter privacy constraints, because user profiles encode sensitive attributes. Federated learning offers a paradigm in which model updates are computed locally and aggregated without raw data leaving the device, and when combined with differential privacy, it provides formal guarantees against membership inference [13, 14]. Numerous studies have also highlighted the risks of bias propagation in clinical language models, where disparities in training data can lead to unequal recommendation quality across demographic groups [15]. These threads of research collectively motivate the integrated framework presented here, which unifies multimodal retrieval, context awareness, and privacy-preserving personalization under a single architectural vision.

3. System Architecture

The proposed framework is designed as a modular, extensible architecture composed of five primary subsystems: multimodal encoders, a context encoder, a retrieval and ranking engine, a personalization module, and a large language model orchestrator. Each subsystem is responsible for a well-defined transformation of data, and their interactions are governed by clearly specified interfaces that facilitate independent scalability and maintenance.

The multimodal encoding layer ingests heterogeneous inputs, including free-text clinical notes, medical images from modalities such as magnetic resonance imaging and computed tomography, structured laboratory values, and continuous physiological signals from wearables. Domain-specific encoders are pretrained or fine-tuned on relevant corpora: a biomedical language model processes text, a vision transformer handles imaging data, and specialized time-series encoders capture trends in electrocardiogram or glucose monitoring streams. The outputs of these encoders are fixed-dimensional embeddings that reside in distinct latent spaces. To unify them, the system employs a cross-modal fusion module based on transformer attention that projects all modality embeddings into a shared representation space, preserving modality-specific nuances while enabling cross-modal associations.

The context encoder extracts situational, temporal, and user-specific features that modulate the retrieval process. Time of day, clinical setting, user role, and device type are transformed into a context vector that conditions the fusion module and the retrieval ranking. For instance, a query initiated by a cardiologist in an intensive care unit should prioritize recent critical care literature over general patient education materials, even if the textual query is identical to one submitted by a patient at home. The context encoder is implemented as a lightweight network that can be updated frequently without retraining the entire pipeline.

The retrieval and ranking engine leverages a two-stage architecture. A fast approximate nearest neighbor search retrieves a broad set of candidate documents from a continuously updated knowledge base of scientific literature, clinical guidelines, and anonymized case repositories. The candidates are then re-ranked by a cross-encoder that receives the fused multimodal query representation concatenated with the context vector and each candidate document. This re-ranking step employs the language model orchestrator as a scoring function, evaluating the relevance, recency, and trustworthiness of each source. The orchestrator, instantiated as a large language model with instruction-following capabilities, also generates natural language explanations and summaries of the retrieved evidence, tailoring the output to the user's expertise level.

The personalization module maintains a longitudinal user profile that captures stable health characteristics, interaction history, and feedback signals. Critically, this profile is stored and updated via federated learning protocols that ensure raw data remains on local devices or within institutional firewalls. Only encrypted model gradients are transmitted to a central aggregation server. The personalization module influences both the fusion weights of multimodal embeddings and the ranking model's scoring function, effectively personalizing the entire retrieval stack without compromising privacy. The architecture's modularity ensures that components can be replaced or upgraded as model capabilities evolve, supporting long-term sustainability.

4. Multimodal Context Modeling

Effective healthcare retrieval demands a representation that captures not only the semantic content of a query but also the rich contextual envelope in which the query arises. The framework's multimodal context modeling subsystem performs this function by constructing a unified context-aware representation that feeds into every downstream decision. The process begins with the ingestion of raw modalities. For text, a specialized biomedical transformer tokenizes clinical language and handles domain-specific abbreviations and negations. For medical images, a vision transformer pretrained on large-scale radiology datasets extracts regional features that correlate with pathological findings. Structured data, such as laboratory results and medication lists, are encoded through entity-aware embeddings that preserve semantic relationships among clinical concepts.

A critical design choice is the fusion strategy. Early fusion approaches that concatenate raw modality inputs before encoding can lead to prohibitively high-dimensional representations and obscure modality-specific patterns. Late fusion, where decisions are combined after independent processing, risks losing cross-modal interactions. The framework adopts a mid-fusion approach based on cross-attention layers that allow each modality to condition the representations of others. For example, image features of a lung nodule can attend to textual mentions of smoking history, and structured lab values can modulate the interpretation of textual symptom descriptions. This fusion process is conditioned by the context vector, which acts as an additional input to the attention mechanisms, scaling the contribution of each modality based on the clinical scenario. In a telemedicine triage setting, where images may be unavailable, the system dynamically down-weights the visual pathway and relies more heavily on textual and structured data.

The context vector itself is the product of a dedicated context encoder that transforms categorical and numerical situational variables into a dense embedding. Environmental variables such as ambient noise level, which can indicate a chaotic emergency department, or the user's recent sleep patterns from a wearable device, can be fused to further personalize the information need. The framework is designed to accept these signals as optional inputs, gracefully degrading to baseline behavior when they are absent. The resulting context-aware multimodal representation is then used both as the query embedding for dense retrieval and as the conditioning signal for the language model orchestrator's generation, ensuring that the entire pipeline remains aligned with the user's immediate circumstances and long-term health trajectory.

5. Personalization and Adaptation Mechanisms

Personalization in healthcare information retrieval is not a luxury but a necessity, as identical clinical questions can have vastly different correct answers depending on patient history, comorbidities, contraindications, and treatment goals. The framework implements a multi-tier personalization strategy that operates at the level of retrieval, ranking, and explanation generation. At the core of this strategy lies a continuously updated user profile that encodes both static attributes and dynamic behavioral patterns. The profile includes a vector representation of the individual's medical history, encoded through a graph neural network that captures relationships among diagnoses, procedures, and medications. It also integrates interaction-level signals such as click-through rates, dwell time on retrieved articles, and explicit relevance feedback provided by clinicians or patients. These signals are aggregated over time using a recurrent architecture with exponential decay, allowing the system to adapt to shifting information needs while gradually forgetting outdated patterns.

To protect sensitive health information, the personalization module is implemented under a federated learning paradigm. Each user's profile is maintained on a local device or within a secure hospital enclave. When the global retrieval and ranking models require updates, local stochastic gradient descent steps are performed on mini-batches of private interaction data, and only the encrypted gradients or model deltas are transmitted to a central coordinator. The coordinator applies secure aggregation to combine updates from multiple users without inspecting individual contributions. Differential privacy is enforced by clipping gradient norms and adding calibrated Gaussian noise during aggregation, providing formal privacy guarantees against adversaries with auxiliary information [14]. This architecture ensures that the benefits of personalization do not come at the expense of patient confidentiality, a non-negotiable requirement in healthcare domains governed by regulations such as HIPAA and GDPR.

The personalization layer also addresses the cold-start problem by initializing user profiles with demographic and condition-based priors derived from population-level statistics, with careful auditing to avoid perpetuating harmful stereotypes. As the user interacts with the system, the prior is progressively overridden by individual evidence. Adaptation mechanisms are designed to respect the temporal dynamics of health: a sudden change in a patient's condition, such as a new diagnosis, triggers a rapid decay of older profile components to prevent outdated information from anchoring recommendations. The interplay between personalization and context awareness ensures that the system remains responsive both to long-term health trajectories and to acute situational demands, embodying a holistic understanding of information retrieval in clinical environments.

6. Infrastructure and Deployment Considerations

Deploying a context-aware multimodal recommendation framework in healthcare requires careful orchestration of computational resources, data governance policies, and system robustness. The framework's heavy reliance on large language model inference introduces significant latency and throughput challenges that must be reconciled with clinical workflows, where delays can directly impact patient outcomes. To address this, the architecture supports a tiered deployment model. Large, general-purpose models reside in cloud-based data centers equipped with high-performance accelerators, while smaller, distilled variants are deployed at the edge within hospital networks or on mobile devices. Inference routing decisions are made dynamically based on query complexity, required modality processing, and privacy constraints. Simple queries that involve only text and a stable context can be handled locally, whereas complex multimodal analyses that require retinal scans and genomic data are escalated to the cloud.

Data residence and sovereignty impose additional constraints. Many healthcare institutions mandate that patient data remain within jurisdictional boundaries. The framework accommodates this through a federation of model instances that exchange encrypted updates without raw data transfer. To reduce communication overhead, compression techniques such as gradient sparsification and quantization are employed, and participating nodes can schedule updates during off-peak hours. The knowledge base over which retrieval is performed is similarly federated; each institution maintains its own document index of locally relevant guidelines and anonymized case repositories, while a global index aggregates public biomedical literature. Cross-institutional retrieval is mediated by a trusted broker that performs privacy-preserving set intersection to identify relevant documents without revealing query specifics.

Sustainability considerations are integral to the design. The carbon footprint of training and serving large language models has drawn increasing scrutiny, and healthcare systems must balance clinical benefit against environmental impact. The framework's modularity allows for model compression, selective activation of modality encoders, and caching of frequent queries to reduce computational load. Federated learning further distributes the training burden, and models are periodically pruned and quantized to extend their operational life on aging hardware. Robustness is ensured through redundant retrieval pipelines that fall back to rule-based systems or precomputed guidelines when model inference fails or confidence scores are low. Continuous monitoring of distribution shifts in clinical data triggers model retraining or retrieval index refresh cycles, preventing performance degradation as medical knowledge evolves. These infrastructure decisions collectively enable a scalable, resilient, and ethically grounded deployment that can adapt to the heterogeneous landscapes of global healthcare.

7. Governance, Fairness, and Policy Implications

The introduction of context-aware multimodal recommendation systems into healthcare raises profound governance challenges that extend far beyond technical performance metrics. Algorithmic bias in clinical language models can reproduce and amplify disparities in healthcare access and quality. If training data over-represents certain demographic groups or clinical conditions, the resulting retrieval and recommendation outputs may systematically disadvantage underrepresented populations. The framework incorporates fairness-aware ranking mechanisms that adjust relevance scores to ensure equitable exposure of information across protected attributes. These adjustments are not post-hoc patches but are integrated into the training objective of the reranking module, using adversarial debiasing techniques and calibrated equalized odds constraints. Transparency is promoted through explanation modules that surface the provenance and demographic characteristics of retrieved documents, enabling clinicians to critically assess potential biases.

Accountability in such systems is distributed across model developers, healthcare providers, and regulatory bodies. When a flawed recommendation leads to patient harm, the chain of responsibility must be clearly defined. The framework includes an audit trail that records the context vector, retrieved document identifiers, personalization profile influence, and model version used for every recommendation. This logging infrastructure enables retrospective forensic analysis and supports compliance with the EU Artificial Intelligence Act and evolving FDA guidelines for software as a medical device. Data governance extends to informed consent procedures, wherein patients must be clearly informed about how their data contributes to personalization and model improvement, with meaningful opt-out mechanisms that do not degrade baseline care quality.

The policy implications also encompass the digital divide. Multimodal systems that require high-bandwidth internet and modern sensor-equipped devices may inadvertently widen gaps between well-resourced and underserved communities. The framework's tiered deployment strategy partially mitigates this by supporting lightweight, unimodal fallback modes on low-end devices, but deeper structural inequities in digital infrastructure remain a societal challenge that technology alone cannot resolve. International collaboration is essential to harmonize privacy regulations, data sharing agreements, and model evaluation standards, enabling the benefits of such systems to cross borders without compromising sovereignty. The governance layer is thus conceived as a living component of the framework, updated as societal norms and legal mandates evolve, ensuring that the technical system remains aligned with the values of the communities it serves.

8. Evaluation and Sustainability

Evaluating a recommendation framework of this complexity requires a multi-faceted methodology that goes beyond traditional relevance metrics. Offline evaluation on curated biomedical benchmarks, such as those derived from PubMed question-answering datasets or radiology report retrieval tasks, provides initial evidence of retrieval accuracy and modality alignment. However, such benchmarks often fail to capture the contextual and personalization dynamics that are central to the framework's value proposition. Online A/B testing in simulated clinical environments, using realistic patient trajectories generated from de-identified electronic health records, offers a more ecologically valid assessment. Metrics include retrieval precision at varying context shifts, personalization lift measured as improvement in downstream task completion relative to a non-personalized baseline, and latency under different network conditions.

User satisfaction and trust are assessed through structured interviews and surveys with clinicians and patient advocates. Trust is particularly fragile in healthcare; if a system occasionally produces irrelevant or unsafe recommendations, user confidence can rapidly erode. The framework therefore includes a confidence calibration module that provides uncertainty scores alongside each recommendation, enabling users to apply appropriate skepticism. Sustainability evaluation considers both environmental and operational dimensions. Energy consumption per query is logged across cloud and edge deployments, and models are periodically compressed to meet pre-defined carbon budgets. The federated learning process is audited for communication efficiency, and the system's ability to incorporate updated medical knowledge without full retraining is measured through incremental learning benchmarks. Long-term viability is supported by an open governance model that allows multiple stakeholders to contribute new encoders, retrieval indexes, and fairness constraints, transforming the framework into a community-owned infrastructure that evolves with medical science and societal expectations.

9. Conclusion

This paper has presented a context-aware multimodal recommendation framework that reimagines healthcare information retrieval as a personalized, privacy-preserving, and context-driven service orchestrated by large language models. By integrating text, imaging, structured data, and environmental signals within a unified architectural fabric, the framework moves beyond the unimodal paradigms that currently dominate clinical decision support. The discussion traversed structural trade-offs, including mid-fusion multimodal representation learning, federated personalization with differential privacy, tiered deployment models that balance latency and computational cost, and fairness-aware ranking mechanisms embedded from the ground up. Governance and policy analyses revealed that technical architectures are inseparable from the socio-technical systems they inhabit, necessitating ongoing engagement with regulatory frameworks, equity considerations, and global infrastructural disparities.

The framework outlined here is not a finished product but a reference architecture intended to stimulate interdisciplinary collaboration and deliberate design in the rapidly advancing field of medical AI. Future work should investigate tighter integration with real-time clinical workflows, longitudinal field studies that validate personalization claims in diverse patient populations, and the development of formal verification methods for fairness and safety properties. The path toward responsible, effective healthcare AI is long and demanding, but architectures that treat context, multimodality, and personalization as first-class design principles hold significant promise for transforming information retrieval into a truly patient-centered endeavor.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

3. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A. K., 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, 172-180.

4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

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

6. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in Neural Information Processing Systems, 32.

7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748-8763.

8. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 39-48.

9. Yu, X. (2026, January). AI-Driven Personalization across Domains for Local Categorical Query Understanding and Context-Aware Retrieval. In Proceedings of the 2nd International Conference on Artificial Intelligence, Digital Media Technology and Social Computing (pp. 77-82).

10. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 3982-3992.

11. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 6769-6781.

12. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

13. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273-1282.

14. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3-4), 211-407.

15. 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.

16. Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: UNiversal image-text representation learning. European Conference on Computer Vision, 104-120.

17. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. Advances in Neural Information Processing Systems, 34.

18. Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2022). SimVLM: Simple visual language model pretraining with weak supervision. International Conference on Learning Representations.

19. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society. Minds and Machines, 28(4), 689-707.

20. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.

21. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1-21.

22. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daume III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

23. Cutillo, L. A., Molva, R., & Strufe, T. (2009). Privacy preserving social networking through decentralization. Proceedings of the 6th International Conference on Wireless On-Demand Network Systems and Services, 145-152.

24. Barocas, S., & Selbst, A. D. (2016). Big data’s disparate impact. California Law Review, 104(3), 671-732.

25. Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4(2), 1-17.