

Multi-Agent Large Language Models for Explainable Clinical Decision Support Through Evidence-Based Information Reranking

Elly Schenk

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL,
USA.

schenkelly@uab.edu

Blazej Burnat

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV,
USA.

burnatblazej@unr.edu

Abstract

Clinical decision support systems increasingly integrate large language models, yet single-model pipelines often struggle with evidence traceability, contradictory guidelines, and governance accountability. This paper examines multi-agent large language model architectures that use evidence-based information reranking to produce explainable clinical recommendations. The argument is system-level: explanation quality is not solely a property of model interpretability but emerges from orchestration, retrieval infrastructure, evidence hierarchy, audit logging, and human oversight. The paper conceptualizes clinical decision support as a socio-technical infrastructure in which multiple agents assume specialized roles such as query formulation, evidence retrieval, critical appraisal, patient-context alignment, safety review, and communication. Reranking is positioned as a governance-relevant mechanism because it makes evidence selection explicit and contestable, allowing clinicians to inspect why one source outranks another. The paper analyzes architectural trade-offs involving latency, cost, resilience, data provenance, privacy, and equity. It discusses failure modes including retrieval drift, hallucinated synthesis, automation bias, and opaque escalation. It also considers deployment and sustainability challenges such as model churn, maintenance of clinical knowledge bases, regulatory adaptation, and environmental cost. The discussion concludes that multi-agent reranking can improve explainability only when embedded in auditable, federated, and policy-aligned systems with clear accountability.

Keywords

multi-agent systems; large language models; clinical decision support; evidence-based medicine; information reranking; explainable artificial intelligence; governance; health informatics.

1. Introduction

The integration of artificial intelligence into clinical decision support has moved from narrow predictive models toward generative and multi-agent systems that promise more flexible interaction with clinical knowledge. Early work on high-performance medicine emphasized the convergence of human expertise, data-intensive computation, and clinical workflows, suggesting that machine intelligence would augment rather than replace professional

judgment [1]. Subsequent research demonstrated both the potential and the hazards of algorithmic decision-making in health systems. Population health algorithms have been shown to encode racial bias when proxy variables substitute for need [2], while deep learning on electronic health records has achieved scalable prediction across diverse clinical tasks [3]. Medical imaging systems have also reached specialist-level performance in constrained domains [4]. These developments established that artificial intelligence can be technically effective, but they also revealed that effectiveness alone does not guarantee safety, fairness, or institutional trustworthiness.

The emergence of foundation models and large language models has expanded the design space for clinical decision support by enabling natural language interaction, summarization, reasoning, and synthesis across heterogeneous sources [5]. However, large language models are not neutral knowledge repositories. They are trained on vast, uneven, and historically situated corpora, and they can reproduce social biases, fabricate plausible but unsupported claims, and obscure the provenance of their outputs [6]. In clinical contexts, these limitations are consequential because recommendations may influence diagnosis, treatment, resource allocation, and patient consent. A single large language model responding to a clinical query may appear coherent while failing to distinguish among guidelines, systematic reviews, institutional protocols, and weak observational evidence. The problem is therefore not merely hallucination in the abstract but the absence of a defensible evidence hierarchy at the point of recommendation.

This paper argues that multi-agent large language models combined with evidence-based information reranking offer a promising architecture for explainable clinical decision support, but only if explainability is treated as a system property rather than a model feature. In such architectures, several agents perform specialized functions: interpreting the clinical query, retrieving candidate evidence, appraising source quality, aligning recommendations with patient context, checking safety constraints, and communicating uncertainty to clinicians. Reranking is central because it determines which evidence is elevated, which is suppressed, and which is presented as the basis for a recommendation. A multi-role collaborative reranking workflow illustrates how large language models can coordinate retrieval, ranking, and synthesis across roles [7]. The contribution of this paper is conceptual and system-oriented. It examines the architectural, infrastructural, governance, and policy implications of using multi-agent large language models for explainable clinical decision support through evidence-based reranking.

2. Clinical Decision Support as a Socio-Technical System

Clinical decision support is often described as a computational function embedded in electronic health records, but it is more accurately understood as a socio-technical system composed of people, protocols, data infrastructures, regulatory constraints, and institutional routines. Large language models that encode clinical knowledge can support examination preparation, differential reasoning, and patient communication, yet their outputs must be validated against local formularies, clinical guidelines, and patient preferences [8]. The broader adoption of large language models in medicine has generated enthusiasm about improved access to information, but it has also raised concerns about overconfidence, liability, and the erosion of clinical reasoning [9]. These concerns are amplified when models are trained on electronic health record data that are sparse, irregular, and shaped by billing and documentation practices rather than by a complete representation of patient health [10].

Explainability in this setting cannot be reduced to post hoc heatmaps or token attributions. Clinicians often need contextual explanations that clarify why a recommendation is relevant to a particular patient, what evidence supports it, what alternatives exist, and what uncertainties remain [16]. A multidisciplinary perspective on explainability in healthcare emphasizes that patients, clinicians, developers, regulators, and institutions may have different explanatory needs [17]. Responsible machine learning in healthcare therefore requires attention to fairness, transparency, validation, and post-deployment monitoring throughout the lifecycle of a system [18]. Ethical and practical challenges in implementation include informed consent, clinical validation, workflow integration, and the distribution of accountability when recommendations cause harm [19].

Within this socio-technical frame, evidence-based information reranking becomes more than an information retrieval operation. It is a mechanism for ordering institutional and scientific authority. If a system elevates a recent randomized controlled trial over an outdated expert opinion, it makes a normative claim about evidence quality. If it prioritizes a local institutional protocol over a national guideline, it negotiates between standardization and contextual adaptation. If it suppresses evidence because of retrieval limitations, it risks narrowing the epistemic basis of care. Explainable clinical decision support therefore depends on making these ordering decisions visible, contestable, and auditable. The system must not only produce a recommendation but also expose the evidentiary pathway that led to it.

3. Multi-Agent Large Language Models: Architecture and Roles

A multi-agent large language model architecture distributes clinical reasoning across specialized agents rather than relying on a single monolithic prompt. One agent may translate a clinical question into structured search intents; another may retrieve candidate documents from guidelines, systematic reviews, drug databases, and institutional policies; another may appraise methodological quality; another may align evidence with patient-specific factors such as age, comorbidity, pregnancy status, renal function, and prior treatments; and a final agent may compose a response with uncertainty markers. This division of labor draws on foundation model capabilities while introducing modular checkpoints for oversight [5]. Reasoning-oriented prompting methods such as chain-of-thought can improve multi-step problem solving, but they also create risks of unfaithful rationales if the generated reasoning does not actually reflect the underlying computation [11]. Agent frameworks that interleave reasoning and action can connect language models to external tools, but they require robust permissioning and monitoring when used in clinical environments [12].

Retrieval-augmented generation provides a foundation for grounding model outputs in external knowledge sources, reducing reliance on parametric memory alone [13]. Dense retrieval methods can identify semantically relevant passages beyond lexical overlap [15], and dedicated reranking models can improve the precision of retrieved evidence before synthesis [14]. A multi-role collaborative reranking workflow extends this logic by assigning different large language model roles to query analysis, evidence ranking, and final response formation [7]. The advantage is not that multiple agents are inherently more accurate, but that their interactions create inspectable decision points. Each agent can be evaluated separately, constrained by different policies, and logged for audit. The disadvantage is that multi-agent systems introduce coordination overhead, error propagation, and new forms of opacity when agents disagree or when arbitration rules are hidden.

Architecturally, the most consequential question is how authority is distributed among agents. A retrieval agent may optimize recall, while an appraisal agent may optimize methodological

rigor, and a safety agent may veto recommendations that conflict with contraindications. If the final communication agent can override earlier agents without explanation, the system loses the benefits of modularity. If earlier agents can block the final agent too easily, the system may become brittle and clinically unhelpful. A balanced design treats agents as participants in a governed workflow, with explicit escalation pathways, confidence thresholds, and human review triggers. The system should make disagreement productive rather than conceal it. For example, if a guideline recommends anticoagulation but a patient-specific agent identifies elevated bleeding risk, the final output should present the tension rather than collapsing it into a single directive.

4. Evidence-Based Information Reranking as an Explanation Mechanism

Reranking is often treated as a technical step for improving retrieval precision, but in clinical decision support it is also an explanatory and governance mechanism. The rank order of evidence communicates importance. When a systematic review appears above a case report, the system signals that aggregated evidence is more authoritative. When a local antimicrobial stewardship policy appears above a general textbook chapter, the system signals that institutional constraints matter. When a recent safety alert outranks older guidance, the system signals temporal relevance. A multi-agent reranking workflow can make these judgments explicit by assigning separate roles to query understanding, candidate scoring, and final ordering [7]. Reranking models such as passage re-rankers can improve relevance estimation [14], and dense retrieval methods can expand the candidate pool beyond exact keyword matches [15]. However, technical relevance is not identical to clinical validity, and the system must distinguish among topical similarity, methodological quality, recency, applicability, and patient fit.

Explainability through reranking requires that the system provide reasons for ordering decisions. Clinicians should be able to see why a source was included, why another was excluded, how quality was assessed, and whether the patient context altered the ranking. This form of explanation aligns with calls for contextual explainability in clinical machine learning, where explanations must support action rather than merely satisfy technical curiosity [16]. It also responds to the multidisciplinary demand that explanations be meaningful to different audiences, including clinicians, patients, quality improvement teams, and regulators [17]. If reranking is opaque, the system may still produce a correct answer while failing to provide a defensible rationale. If reranking is transparent but clinically uninformed, it may provide a rationale that is irrelevant or misleading. The challenge is to design reranking criteria that are both computationally tractable and clinically legitimate.

Evidence-based reranking also has implications for uncertainty communication. Clinical evidence is often conflicting, incomplete, or indirect. A multi-agent system can represent uncertainty by presenting competing guidelines, noting limitations in study populations, and flagging when evidence is extrapolated. It can distinguish between strong recommendations based on high-quality evidence and conditional recommendations based on weaker evidence. It can also identify when no adequate evidence exists and recommend consultation rather than speculation. These distinctions are difficult for a single prompt to maintain consistently, but they can be operationalized through agent roles and ranking policies. The reranking layer thus becomes a place where epistemic humility is encoded into system behavior. Explainability then means not only showing sources but also showing the limits of those sources.

5. System Architecture, Data Flows, and Infrastructure

A deployable multi-agent clinical decision support system requires an infrastructure that supports secure data ingestion, low-latency retrieval, versioned knowledge bases, and comprehensive audit logging. Electronic health record data are heterogeneous, often fragmented across laboratories, medications, notes, imaging, and claims, and deep learning models trained on such data must contend with missingness, irregular sampling, and documentation bias [3]. Retrieval-augmented generation can connect language models to external evidence, but the retrieval corpus must be curated, licensed, updated, and access-controlled [13]. Privacy considerations are substantial because medical big data can be reidentified, combined with other datasets, and used in ways that patients may not anticipate [20]. Infrastructure design therefore cannot be separated from privacy law, institutional policy, and data governance.

From a system architecture perspective, the evidence pipeline might include a clinical query interface, a terminology normalization service, a federated search layer across guideline repositories, a vector index for semantic retrieval, a reranking service, a patient-context service, a safety rule engine, and an audit store. Each component introduces latency and failure risk. If retrieval is slow, clinicians may bypass the system. If reranking is computationally expensive, the system may be limited to non-urgent settings. If audit logging is incomplete, accountability is weakened. If knowledge bases are not versioned, recommendations may change without a traceable reason. The infrastructure must therefore be designed for graceful degradation, with clear fallbacks when evidence retrieval fails or when confidence is low. In urgent care, a system that cannot retrieve evidence quickly should default to concise safety guidance or human escalation rather than generate unsupported recommendations.

Regulatory oversight of large language models in healthcare adds another infrastructural layer. Generative systems may require premarket review, postmarket surveillance, cybersecurity controls, and transparency about training data and performance boundaries [21]. A multi-agent architecture complicates regulation because the overall system behavior emerges from interactions among components that may be updated independently. Regulators may need to evaluate the orchestration logic, not only individual models. They may also need access to logs that show which evidence was retrieved, how it was ranked, and what human oversight occurred. The infrastructure should therefore be designed with regulatory observability in mind. Auditability is not an afterthought but a core architectural requirement. Without it, explainability claims remain difficult to verify.

6. Robustness, Safety, and Failure Modes

Multi-agent large language models introduce distinctive failure modes. Retrieval drift occurs when the underlying evidence base changes, when indexing lags behind new guidelines, or when semantic search returns topically similar but clinically inappropriate sources. Hallucinated synthesis occurs when a language model combines retrieved passages into a recommendation that no single source supports. Reasoning-action agents may take unnecessary or unsafe actions if tool permissions are too broad [12]. Chain-of-thought prompting may produce plausible rationales that do not faithfully represent the model's decision process [11]. Large language models trained on electronic health record data may inherit documentation artifacts and fail to generalize across institutions [10]. These failures are not merely technical defects; they are threats to clinical safety and institutional legitimacy.

Safety design must therefore include multiple layers of constraint. Retrieval should be limited to approved sources. Reranking should incorporate quality appraisal, recency checks, and

contraindication screening. Synthesis should be required to cite retrieved passages and should be prohibited from introducing unsupported claims. A safety agent should be able to veto or escalate recommendations that conflict with hard rules. Human oversight should be mandatory for high-stakes decisions, ambiguous evidence, or vulnerable populations. Responsible machine learning in healthcare emphasizes prospective validation, monitoring, and governance across the lifecycle [18]. Regulatory oversight of generative artificial intelligence in healthcare further suggests that systems should be evaluated not only for accuracy but also for transparency, traceability, and accountability [21]. Robustness is thus a socio-technical achievement, not a property that can be tested once at deployment.

Another failure mode is automation bias, where clinicians defer to system outputs even when they are uncertain or incorrect. Explainability can reduce automation bias only if explanations are genuinely informative and if workflows encourage critical engagement. If clinicians are presented with a polished recommendation and a long list of citations, they may assume the evidence has been correctly appraised. If they are presented with too many competing sources, they may ignore the system entirely. The design challenge is to calibrate explanation depth to clinical task, urgency, and user expertise. A multi-agent system can support this calibration by producing layered explanations: a concise recommendation, a summary of key evidence, a detailed evidence table, and an audit trail. The same underlying reranking logic can generate different views for different users while preserving a common evidentiary basis.

7. Fairness, Governance, and Policy

Fairness in clinical decision support is not limited to balanced accuracy across demographic groups. It also concerns whose evidence is included, whose data are used, whose preferences are represented, and whose risks are prioritized. Algorithmic bias can arise from proxy variables, historical inequities, and deployment contexts that differ from training environments [2]. Large language models can amplify bias through training corpora and through the stochastic reproduction of dominant narratives [6]. Governance must therefore address data provenance, model evaluation, stakeholder inclusion, and redress mechanisms. A governance model for artificial intelligence in healthcare should specify roles for clinicians, administrators, developers, ethicists, patients, and regulators [22]. It should also define how incidents are reported, how models are updated, and how affected patients are notified.

Policy implications include transparency mandates, clinical validation requirements, liability frameworks, and procurement standards. If a hospital purchases a multi-agent decision support system, it needs to understand how evidence is ranked, how updates are managed, and whether the system can be audited. Regulators need to determine whether the system is a medical device, a clinical decision support tool, or an administrative aid, because these categories carry different obligations. Professional bodies need to define appropriate use, documentation standards, and training expectations. Patients need accessible explanations of how their data and preferences influence recommendations. These policy questions are not external to system design; they shape the architecture. For example, a requirement for auditability may lead to immutable logs and explainable reranking, while a requirement for privacy may lead to federated retrieval and on-premises inference.

Fairness also requires attention to resource constraints. Health systems with fewer technical resources may be unable to maintain complex multi-agent infrastructures, creating a two-tier system in which well-resourced institutions benefit from advanced decision support while under-resourced clinics rely on less safe alternatives. Sustainable governance must therefore consider equity in deployment, not only equity in model performance. Shared evidence

repositories, open standards, and regional infrastructure cooperatives may reduce duplication and improve access. However, shared infrastructure creates new dependencies and governance challenges. Who controls the evidence corpus? Who decides which guidelines are authoritative? How are local adaptations represented? These questions are political as well as technical, and they must be addressed explicitly in policy and system design.

8. Deployment, Sustainability, and Lifecycle Management

Deployment of multi-agent clinical decision support is a long-term commitment rather than a one-time installation. Clinical guidelines change, drug labels are updated, new evidence emerges, and institutional protocols evolve. If the system's knowledge base is not maintained, it will gradually become unsafe. If it is updated too frequently without validation, it may destabilize clinical workflows. Lifecycle management must therefore include version control, change impact assessment, staged rollout, user training, performance monitoring, and incident response. The convergence of human and artificial intelligence in medicine suggests that successful systems will augment clinical work rather than replace it, which means deployment must be co-designed with clinicians and patients [1]. Large language models in medicine have the potential to improve information access, but their integration requires careful attention to workflow, trust, and accountability [9].

Sustainability also has economic and environmental dimensions. Multi-agent systems may require multiple model calls per query, retrieval over large corpora, and continuous monitoring, all of which consume computational resources. Larger foundation models may offer better performance but at higher inference cost and energy use [5]. Health systems must weigh these costs against clinical benefits, especially when resources are scarce. Cost concerns may favor smaller, domain-adapted models for some agents and larger models only for synthesis or arbitration. Energy use may favor efficient retrieval, caching, and batching. Sustainability is not merely an operational concern; it intersects with equity because institutions with limited budgets may be unable to sustain expensive architectures. A responsible deployment strategy should therefore include cost transparency, resource optimization, and long-term funding for maintenance.

Privacy and data protection are also lifecycle concerns. Medical big data can be valuable for research and improvement, but it can also be misused or reidentified [20]. A multi-agent system that logs queries, retrieved evidence, and clinician interactions creates a rich data trail that must be protected. Logs should be minimized, encrypted, access-controlled, and retained only as long as necessary for safety, audit, and quality improvement. If logs are used for model training, patients and clinicians should understand the terms and safeguards. Regulatory oversight of generative artificial intelligence in healthcare is likely to require ongoing monitoring rather than static approval [21]. Deployment should therefore be understood as a continuous process of governance, evaluation, and adaptation.

9. Evaluation and Cross-Domain Lessons

Evaluating multi-agent clinical decision support requires metrics that go beyond answer accuracy. Retrieval systems are often assessed with recall, precision, and ranking metrics [14]. Retrieval-augmented generation is evaluated for faithfulness, relevance, and factuality [13]. Reasoning agents are assessed for task success and tool use [12]. Chain-of-thought methods are assessed for problem-solving accuracy, but their rationales may not be faithful [11]. In clinical settings, these technical metrics must be complemented by human factors evaluation, workflow impact, and patient outcomes. Does the system help clinicians find evidence faster?

Does it improve guideline adherence? Does it reduce diagnostic error? Does it increase or decrease cognitive burden? Does it support shared decision-making? These questions cannot be answered by offline benchmarks alone.

Cross-domain lessons from information retrieval and web search suggest that reranking is most effective when it combines multiple signals rather than optimizing a single relevance score. In clinical decision support, signals might include evidence level, publication recency, guideline endorsement, local applicability, patient-specific contraindications, and equity considerations. A multi-role collaborative reranking workflow demonstrates how large language models can coordinate such signals through role specialization [7]. However, web search is not clinical reasoning. Search engines optimize for user satisfaction and engagement, while clinical systems must optimize for safety, effectiveness, and accountability. Evaluation must therefore include adverse event analysis, subgroup performance, and failure mode reproduction. It must also include contestability: can clinicians challenge a ranking, correct an error, and see the system update? A system that cannot learn from contestation will struggle to maintain trust.

Cross-domain comparisons also highlight the importance of governance. Aviation, nuclear power, and financial services have developed safety cultures that emphasize incident reporting, redundancy, and human factors. Healthcare can learn from these fields without simply importing their models. Clinical decision support involves intimate ethical questions about autonomy, consent, and justice that are not fully captured by engineering safety frameworks. Responsible machine learning in healthcare provides a roadmap that includes transparency, fairness, and monitoring [18]. Ethical machine learning in healthcare emphasizes stakeholder engagement and longitudinal evaluation [19]. These perspectives suggest that evaluation should be participatory and ongoing, involving clinicians, patients, and communities affected by the system.

10. Limitations and Future Research

The multi-agent reranking approach has several limitations. First, evidence-based reranking depends on the quality and accessibility of the underlying evidence corpus. If guidelines are paywalled, outdated, or institution-specific, the system's explanations will be incomplete. Second, multi-agent coordination can introduce latency and cost that may be unacceptable in acute care. Third, language model agents may still produce unfaithful rationales or unsupported synthesis even when retrieval is accurate. Fourth, the system may create a false sense of rigor if the ranking logic is not clinically validated. Fifth, governance structures may lag behind technical capabilities, leaving accountability ambiguous. These limitations suggest that multi-agent reranking should be deployed incrementally, with clear boundaries and human oversight.

Future research should investigate clinically grounded reranking criteria that combine methodological quality, patient applicability, and equity impact. It should develop evaluation methods for explanation faithfulness in clinical workflows, not only for model outputs. It should explore federated and privacy-preserving architectures that allow institutions to share evidence and learning without sharing patient data. It should examine how clinicians actually use layered explanations and whether they improve or undermine judgment. It should study the environmental and economic sustainability of multi-agent systems at scale. It should also investigate policy mechanisms for auditing agentic systems whose behavior changes with model updates. The most important research question is not whether multi-agent large

language models can produce convincing clinical recommendations, but under what institutional, technical, and ethical conditions they can be trusted to support care.

11. Conclusion

Multi-agent large language models offer a promising architecture for explainable clinical decision support because they can separate query formulation, evidence retrieval, critical appraisal, patient-context alignment, safety review, and communication into inspectable roles. Evidence-based information reranking is central to this architecture because it makes the ordering of clinical evidence explicit and contestable. When reranking is transparent, versioned, and governed, it can help clinicians understand not only what a system recommends but why it prioritizes certain sources over others. When reranking is opaque, it can conceal bias, outdated evidence, and unsupported synthesis behind a veneer of technical sophistication. The difference depends less on model size than on system design.

Explainability in this context is a socio-technical property. It emerges from data provenance, retrieval infrastructure, audit logs, human oversight, regulatory policy, and institutional culture. A multi-agent system that cannot be audited, updated, contested, or constrained will not be trustworthy merely because it provides citations. A system that is designed for accountability, however, can support clinicians without displacing professional judgment. It can present uncertainty rather than false confidence, escalate rather than overreach, and learn from use rather than ossify. The future of clinical decision support lies not in a single omniscient model but in governed infrastructures that combine machine capability with human responsibility. Evidence-based reranking can be one part of that infrastructure, provided it is embedded in a broader commitment to safety, fairness, sustainability, and public accountability.

References

1. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25(1), 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
2. 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. <https://doi.org/10.1126/science.aax2342>
3. Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., Liu, P. J., Liu, X., Marcus, J., Sun, M., Sundaram, V., Yee, H., Zhang, K., Zhang, Y., Flores, G., Duggan, G. E., Irvine, J., Le, Q. V., Litsch, K., ... Dean, J. (2018). Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1, 18. <https://doi.org/10.1038/s41746-018-0029-1>
4. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. <https://doi.org/10.1038/nature21056>
5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
6. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM*

Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>

7. Jin, Can, et al. "Rankflow: A multi-role collaborative reranking workflow utilizing large language models." *Companion Proceedings of the ACM on Web Conference 2025*. 2025.
8. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole, E., Pfohl, S., Payne, P., Seneviratne, M., Gamble, P., Kelly, C., Babiker, A., Scharli, N., Chowdhery, A., Mansfield, P., Demner-Fushman, D., ... Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, 620(7972), 172–180. <https://doi.org/10.1038/s41586-023-06291-2>
9. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, 29(8), 1930–1940. <https://doi.org/10.1038/s41591-023-02448-8>
10. Wornow, M., Xu, Y., Thapa, R., Patel, B., Steinberg, E., Fleming, S., Pfeffer, M. A., Fries, J., & Shah, N. H. (2023). The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6, 135. <https://doi.org/10.1038/s41746-023-00879-8>
11. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837. <https://doi.org/10.48550/arXiv.2201.11903>
12. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*. <https://arxiv.org/abs/2210.03629>
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474. <https://doi.org/10.48550/arXiv.2005.11401>
14. Nogueira, R., & Cho, K. (2019). Passage re-ranking with BERT. *arXiv*. <https://arxiv.org/abs/1901.04085>
15. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). Association for Computational Linguistics. Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. In *Proceedings of the 4th Machine Learning for Healthcare Conference* (Vol. 106, pp. 359–380). PMLR.
16. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20, 310. <https://doi.org/10.1186/s12911-020-01332-6>
17. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D. C., Saeed, M., Ossorio, P. N., Thadaneys-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340. <https://doi.org/10.1038/s41591-019-0548-6>

18. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science*, 4, 123–144. <https://doi.org/10.1146/annurev-biodatasci-092820-114757>
19. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25(1), 37–43. <https://doi.org/10.1038/s41591-018-0272-7>
20. Gilbert, S., Harvey, H., Melvin, T., Vollebregt, E., & Wicks, P. (2023). Large language model AI chatbots require approval as medical devices. *Nature Medicine*, 29(10), 2396–2398. <https://doi.org/10.1038/s41591-023-02412-6>
21. Morley, J., Machado, C. C. V., Burr, C., Cows, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, 260, 113172. <https://doi.org/10.1016/j.socscimed.2020.113172>