Personalized Healthcare Recommendation via LLM-Based Patient Memory Modeling and Medical Query Understanding

Dend Oliavoer
School of Computing, Clemson University, Clemson, SC, USA.
hellodean@clemson.edu

Neeriaj Mathiur
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
neeraj498@buffalo.edu

Abstract

The integration of large language models into healthcare recommender systems represents a paradigm shift in the delivery of personalized medical information and decision support. Traditional clinical recommendation engines often rely on structured electronic health records and static rule-based logic, failing to capture the fluid, longitudinal narrative of a patient's lived health experience. This paper presents a system-level investigation into a novel architecture that constructs a dynamic patient memory model using the generative and contextual capabilities of large language models, coupled with advanced medical query understanding. Rather than proposing a single algorithm, we examine the structural trade-offs, integration pathways, and socio-technical implications of embedding such memory-augmented models within real-world healthcare infrastructures. The discussion extends to the architecture's handling of temporal dynamics, multi-modal health signals, and the disambiguation of clinically ambiguous language in patient queries. A central theme is the governance of such systems, addressing fairness across heterogeneous populations, interpretability for clinical stakeholders, and the sustainability of large-scale deployment under computational and energy constraints. Through conceptual analysis and cross-domain comparison with consumer personalization systems, the paper identifies unique challenges in the medical domain, including the irreversibility of harmful recommendations and the imperative of value alignment between patient preferences and evidence-based medicine. We argue that the memory model must not be a passive log but an active, ethically-aware construct that supports continuity of care while respecting privacy boundaries and evolving patient identities. The analysis further explores federated learning paradigms to distribute model training across institutions without sharing sensitive data, and examines policy implications for regulatory approval of continuously learning recommendation agents. The paper concludes with a forward-looking perspective on how patient-centric memory architectures could reshape chronic disease management, remote monitoring, and shared clinical decision-making, ultimately contributing to more resilient and equitable health systems.

Keywords

large language models, patient memory modeling, medical query understanding, personalized healthcare recommendation, health AI governance, socio-technical systems.

1. Introduction

The contemporary healthcare landscape is characterized by an unprecedented proliferation of both biomedical knowledge and patient-generated data. Electronic health records, wearable sensor streams, genomic profiles, and natural language clinical notes converge into a complex and fragmented representation of an individual's health trajectory. Within this data-rich environment, the challenge of delivering timely, contextually appropriate, and personalized health recommendations remains largely unsolved. Early clinical decision support systems operated on deterministic rules and curated knowledge bases, providing value in well-defined diagnostic pathways but failing when confronted with the nuanced, multi-morbid realities of modern patient populations. The emergence of large language models, trained on vast corpora that include biomedical literature and clinical texts, offers a transformative opportunity to bridge this gap by enabling a more conversational, context-aware, and memory-driven approach to healthcare recommendation. However, the leap from a generic conversational agent to a trustworthy, personalized health advisor demands a fundamental reconceptualization of how patient history is modeled, how queries are interpreted, and how the system is situated within the broader socio-technical fabric of care delivery.

The primary motivation for this work stems from the recognition that healthcare decisions are inherently longitudinal and narrative in nature. A patient's current symptom cannot be meaningfully interpreted without understanding the sequence of past diagnoses, treatments, lifestyle changes, and even psychological states that form a continuous health story. Traditional recommender systems in e-commerce or media streaming excel at modeling user preferences through collaborative filtering and short-term behavioral signals, yet these techniques translate poorly into medicine. A movie watched in the past might inform taste, but a past adverse drug reaction is a life-critical constraint that must not be overridden by population-level trends. Thus, the vertical transfer of personalization architectures from consumer domains into healthcare demands not merely adaptation but a profound re-engineering of memory representation. Large language models, with their ability to perform in-context learning and retrieve information from expansive unstructured text, provide a novel substrate for constructing a patient memory that is simultaneously structured enough to prevent catastrophic omission and flexible enough to accommodate the ambiguities of human health narratives.

This paper is structured as a system-level inquiry into a proposed architecture that couples an LLM-based patient memory module with a specialized medical query understanding pipeline. We deliberately avoid algorithmic formalism, focusing instead on structural trade-offs, architectural patterns, governance frameworks, and the operational realities of deploying such a system at scale. Section 2 situates our discussion within related work on clinical NLP, memory networks, and retrieval-augmented generation. Section 3 elaborates the design and maintenance of the patient memory model, emphasizing its temporal, multimodal, and ethical dimensions. Section 4 addresses the challenge of medical query understanding, including intent disambiguation and the management of clinically ambiguous language. Section 5 examines the integration of these modules into a cohesive recommendation engine, while Section 6 delves into governance, fairness, and accountability. Section 7 explores deployment architectures, sustainability, and infrastructure considerations. Section 8 presents a cross-domain evaluation perspective, and Section 9 concludes with implications for future health systems.

2. Background and Related Work

The intellectual lineage of our proposed system draws from several converging streams of research. Foundational advances in neural sequence modeling, epitomized by the transformer architecture, have reshaped natural language processing [1]. The subsequent development of pre-trained language models such as BERT and its biomedical derivatives like BioBERT and PubMedBERT demonstrated that domain-adaptive pre-training yields substantial gains on clinical information extraction tasks [2, 3]. These models, however, are static in the sense that they do not maintain a persistent memory of individual users across sessions. The more recent scaling of autoregressive language models, exemplified by GPT-3, revealed emergent capabilities in few-shot reasoning and task generalization, pointing toward a future where a single model could serve multiple clinical functions through natural language prompting [4]. Yet, the fundamental limitation of finite context windows and the absence of long-term personalization remained.

Parallel to the evolution of language models, the field of memory-augmented neural networks offered architectural solutions for persistent state. Early memory networks and differentiable neural computers introduced external memory banks that could be read from and written to during reasoning [5]. While these models were not initially designed for patient-specific longitudinal data, the conceptual framework of separating computation from storage directly inspired architectures for personalized dialogue systems. Retrieval-augmented generation further advanced this paradigm by combining parametric knowledge stored in model weights with non-parametric knowledge retrieved from external corpora at inference time [6]. Applied to healthcare, this suggests a pathway where a patient's historical data serves as a private knowledge base, continuously queried by the language model during recommendation generation, thereby grounding responses in verifiable personal evidence.

On the recommender systems front, a vast literature exists on collaborative filtering, content-based methods, and deep learning approaches for product and content recommendation [7]. Healthcare recommendation, however, introduces unique constraints: the need for strict safety guarantees, the handling of sparse and irregular longitudinal data, and the integration of clinical guidelines that may conflict with individual preferences. Recent works have explored health recommendation using graph neural networks to model patient-disease-treatment relationships and using reinforcement learning for dynamic treatment regimes [8, 9]. Yet these models typically require structured inputs and do not natively accommodate free-text patient narratives or conversational queries. The confluence of advances in clinical NLP, memory systems, and safe recommendation therefore creates a fertile ground for a novel integrative architecture that directly leverages unstructured patient histories.

Crucially, the literature on AI fairness and explainability in medicine provides essential guardrails. Studies have systematically documented how clinical predictive models can exhibit racial, socioeconomic, and gender biases when trained on biased historical data [10]. The move toward LLM-based memory models amplifies these concerns, as language models may absorb and regenerate subtle biases embedded in clinical language. Concurrently, the development of post-hoc explanation methods such as SHAP and LIME, alongside attention-based interpretability, has offered partial visibility into model reasoning [11, 12]. These techniques, however, must be re-evaluated in the context of generative, memory-laden systems whose decisions depend on a vast, opaque aggregation of personal history. The system-level challenge is therefore not only to build an accurate recommender but to embed mechanisms for auditability, contestability, and ongoing fairness monitoring directly within the architectural workflow.

3. System Architecture: LLM-Based Patient Memory Modeling

The core innovation of the proposed framework lies in the design of a dynamic patient memory model that serves as the persistent, evolving representation of an individual's health narrative. Unlike a static feature vector extracted from electronic health records, this memory model is conceptualized as a structured yet narrative-aware knowledge store, continuously updated through interactions with the patient, healthcare providers, and ambient health data streams. Architecturally, the memory system is decoupled from the query understanding and recommendation generation modules, allowing each to evolve independently and be scaled according to distinct requirements. The memory model is anchored in a large language model that acts not as the primary knowledge base itself, but as a semantic encoder and retrieval coordinator, translating raw heterogeneous data into a unified internal representation.

The first structural decision concerns the granularity and modality of stored information. The memory model must accommodate episodic clinical events such as diagnoses, procedures, and medication changes, alongside continuous signals from wearable devices, patient-reported outcomes, and even unstructured journal entries. A naive approach of concatenating all data into a chronological text stream and feeding it to a language model would violate context length constraints and obscure critical relationships across time. Instead, we propose a hierarchical memory architecture comprising a short-term buffer for recent events, a long-term semantic index of historically significant episodes, and a concept graph that captures medically meaningful relationships such as causality, comorbidity, and treatment contraindications. The LLM serves to encode raw entries into summary representations that populate these structures, a process that can be seen as a form of structured abstraction wherein clinically salient information is distilled while preserving uncertainty when present [13]. The literature on memory systems for user modeling has similarly emphasized the need for abstraction layers to support efficient retrieval and reasoning for personalization tasks, informing the design of such hierarchical stores [14].

A further critical dimension is the temporality of memory. Health states evolve with complex periodicity and irreversibility; a resolved childhood condition may hold little relevance to an adult query, whereas a past cardiac event remains a lifelong risk factor. The memory model must therefore support temporally-weighted retrieval, where recent and chronic conditions receive higher salience, yet distant but potentially relevant events are not entirely suppressed. This can be achieved through a decay function that interacts with clinical significance weights, such that a remote cancer diagnosis persists with high relevance indefinitely. The language model's role includes assessing the contextual relevance of retrieved memories to the current query, a process that resembles cross-attention over a personalized memory bank. In this way, the system avoids both the rigidity of fixed time windows and the noise of indiscriminate historical recall.

Privacy and data sovereignty are non-negotiable constraints that fundamentally shape the memory architecture. All patient-specific data must remain within the patient's sphere of control, ideally stored in a personal health cloud or under the custodianship of a trusted healthcare provider. The memory model should be deployable in a federated topology, where the language model's encoding and retrieval components operate on local data, and only de-identified, aggregated gradients or model updates are shared across institutions. This federated learning paradigm, initially proposed for mobile keyboard prediction, has been extended to healthcare settings to enable collaborative model improvement without centralizing sensitive records [15]. The challenge is compounded by the need to maintain a

coherent memory representation that can be portably accessed across different care settings; interoperability standards such as HL7 FHIR become essential, with the memory model acting as a semantic overlay that enriches structured data with narrative depth.

Finally, the memory model must be contestable and editable, not a black-box inscription. Patients and their authorized clinicians should be able to correct factual errors, update outdated information, and adjust the salience of certain memories in light of new understanding. An immutable, purely algorithmic memory would erode trust and could propagate harmful inaccuracies. The design therefore incorporates a human-in-the-loop editing interface, guarded by access control mechanisms, where modifications are logged and versioned. This transforms the memory from a static repository into a co-constructed health biography, aligning with the principles of participatory medicine.

4. Medical Query Understanding and Intent Disambiguation

A healthcare recommendation system is only as effective as its capacity to accurately interpret the patient's query, which may range from a precisely phrased clinical question to an emotionally laden, vague description of discomfort. Medical query understanding extends far beyond keyword matching or intent classification; it demands the resolution of clinical ambiguity, the contextualization of symptoms within the patient's unique physiology and history, and an assessment of the urgency and risk profile of the expressed concern. In the proposed architecture, the query understanding module operates in tight coupling with the patient memory model, but retains its own specialized processing pipeline to ensure that the nuances of language are fully respected.

The first layer of query understanding involves syntactic and semantic parsing using a domain-adapted language model. While general-purpose LLMs exhibit impressive zero-shot medical reasoning, their performance can be brittle in the face of rare disease terminology, regional dialects, or lay descriptions that map imperfectly to biomedical concepts. We advocate for a hybrid approach wherein a fine-tuned biomedical encoder first maps the query to a structured intermediate representation of clinical entities, negation cues, and temporal references, following the lineage of clinical named entity recognition and relation extraction research [16]. Simultaneously, the raw query is passed through a larger generative model that captures the patient's affective tone, urgency cues, and conversational context. The fusion of these signals enables the system to distinguish, for example, between a patient casually inquiring about dietary adjustments and a patient expressing alarm about a potential drug interaction.

Intent disambiguation is particularly challenging in open-domain health dialogue because patients often present multiple, nested goals within a single utterance. A query such as "my back hurts again and I've been feeling tired, should I stop the statins?" simultaneously expresses a symptom report, a fatigue concern, a medication question, and an implicit request for risk-benefit evaluation. Decomposing such compound intents requires a hierarchical attention mechanism that weighs each sub-intent against the patient's memory. The system must prioritize safety-critical intents—such as potential adverse drug reactions—while not dismissing the secondary concerns that contribute to the patient's overall well-being. This design draws inspiration from multi-task learning frameworks in dialogue systems, but is uniquely constrained by the clinical imperative to never miss a safety signal.

An underappreciated dimension of query understanding is the management of epistemic uncertainty in language. Patients frequently use qualifiers ("I think," "maybe," "sort of") that

convey diagnostic uncertainty, and these verbal cues carry vital information for clinical reasoning. A recommendation system that treats all patient statements as equally factual may generate overconfident, potentially dangerous advice. The query understanding module must therefore output not only a point estimate of meaning but a calibrated uncertainty distribution over possible interpretations. When uncertainty exceeds a threshold, the system's recommendation should explicitly acknowledge the ambiguity and advise professional consultation rather than venturing a speculative answer. This design principle directly operationalizes the precautionary principle in AI safety [17].

The architecture further incorporates a feedback loop from the patient memory model to refine query interpretation in real time. For instance, a complaint of "chest tightness" will be interpreted with different urgency for a patient with a recent history of myocardial infarction versus a patient with documented anxiety and normal cardiac workup. The memory-conditioned query understanding thus achieves a form of personalized semantic grounding, where the same linguistic surface form yields divergent internal representations based on the individual's longitudinal profile. This represents a departure from one-size-fits-all NLP pipelines and aligns with the broader movement toward context-aware language understanding in critical domains.

5. Personalized Recommendation Engine and Integration

Sitting atop the memory and query understanding modules, the recommendation engine synthesizes a personalized response that balances clinical evidence, patient preferences, and safety constraints. The engine operates as a generative process guided by retrieved memory snippets, structured clinical guidelines, and the resolved query intent. Critically, the recommendation is not a monolithic text generation but a composite output that may include direct advice, educational content, disclaimers, risk stratifications, and actionable next steps such as scheduling an appointment or adjusting a medication dose within pre-authorized parameters.

The central structural tension in the recommendation engine is the alignment between patient preferences and evidence-based medicine. A purely preference-driven model might accommodate a patient's request to avoid certain medications or procedures, but in doing so could steer toward suboptimal care if not properly bounded. Conversely, a strictly guideline-concordant engine might disregard the patient's values and life context, leading to non-adherence and poor health outcomes. Our approach posits a negotiation layer within the recommendation engine, formalizing the interaction as a constrained optimization where clinical guidelines define a safe region of options and patient preferences guide selection within that region. When the patient's request falls outside the safe region, the system does not simply refuse but engages in an explanatory dialogue, elucidating the rationale behind the clinical boundary and exploring acceptable compromises. This is reminiscent of shared decision-making frameworks in clinical practice, translated into an algorithmic mediation process [18].

To support this negotiation, the engine accesses a curated, versioned repository of clinical practice guidelines and drug interaction databases. Unlike a static knowledge base, this repository must be continuously updated and annotated with evidence grades and conflict-of-interest disclosures. The LLM serves as an interface to this structured knowledge, providing natural language justifications that cite specific guideline paragraphs and explain how they apply to the patient's case. The combination of retrieval-augmented generation with formalized clinical knowledge ensures that recommendations are not hallucinated but

grounded in verifiable sources, reducing the risk of model confabulation that has been a significant concern in medical LLM applications [19].

The recommendation engine also manages the temporal dynamics of health advice. A recommendation delivered today must be coherent with past recommendations and anticipatory of future health trajectories. For patients with chronic conditions such as diabetes, the system must track medication adjustments over time and prevent abrupt, unsupported escalations or de-escalations. The memory model provides the historical context, while the recommendation engine enforces temporal coherence through a state transition model that flags anomalous recommendation patterns. For example, if the patient's memory indicates a prior adverse reaction to a drug class, any recommendation involving that class must trigger an override requiring explicit confirmation. This failsafe mechanism mirrors the clinical practice of allergy cross-checking but extends it to a broader set of temporal contradictions.

Integration into clinical workflows is a non-trivial socio-technical challenge. The recommendation engine must not exist as a standalone application but as a service that interoperates with electronic health record systems, patient portals, and telehealth platforms. The architectural choice of a microservices-based deployment, with well-defined APIs for memory retrieval, query processing, and recommendation generation, enables incremental adoption. A primary care physician may initially use the system as a second-opinion tool, reviewing AI-generated recommendations before endorsing them to the patient, transitioning gradually to direct patient-facing use for low-stakes queries as trust and validation accumulate. This graded autonomy model reduces risk and allows continuous learning from clinician feedback, which is captured and used to fine-tune the engine through reinforcement learning from human preferences.

6. Governance, Fairness, and Ethical Implications

The deployment of a personalized healthcare recommendation system with persistent patient memory raises profound questions of governance, fairness, and ethics that must be addressed at the architectural level rather than retrofitted post hoc. The system's recommendations can influence life-altering decisions, and any systematic bias against particular demographic groups or disease profiles could exacerbate health disparities. Governance structures must therefore ensure transparency, accountability, and equitable outcomes across diverse patient populations.

Fairness in this context has multiple dimensions. Distributive fairness concerns whether the quality of recommendations is uniform across groups stratified by race, gender, socioeconomic status, and geography. Because the memory model encodes the patient's past interactions with the healthcare system, it inevitably inherits the biases present in that history. Patients from marginalized communities may have less comprehensive records due to fragmented care, or their symptoms may have been documented with dismissive language that the model could propagate. Mitigating these biases requires a combination of de-biasing techniques during memory encoding, fairness-constrained fine-tuning of the language model, and adversarial testing against validated clinical vignettes that vary only by demographic attributes. The literature on fairness in clinical AI has demonstrated that simplistic demographic blindness is often counterproductive, as it removes the context needed to redress disparities; a more nuanced approach involves fairness through awareness, using demographic information only to detect and correct for inequities [20].

Procedural fairness demands that patients and clinicians can understand and contest the system's decisions. To this end, the architecture incorporates an explanation layer that generates human-readable accounts of the key factors driving a recommendation, grounded in specific memory entries and guideline citations. These explanations must be tailored to the health literacy level of the recipient, avoiding both patronizing simplification and incomprehensible jargon. The challenge of generating faithful explanations from large language models is an active research area, with current evidence suggesting that post-hoc attention maps may not reliably reflect true reasoning processes [21]. Our design therefore emphasizes a modular approach where the explanation is generated by a separate, constrained module that directly references the retrieval traces and decision boundaries, rather than relying on introspective model confabulation.

A further governance imperative is the management of patient identity and consent over time. A memory model that accumulates decades of health data raises questions about the right to be forgotten, particularly for sensitive episodes such as mental health crises, substance use history, or genetic information. The architecture must support granular consent controls, allowing patients to restrict access to specific memory segments or to purge them entirely. Technically, this requires cryptographic mechanisms for data compartmentalization, such that the language model can retrieve only the subset of memory for which active consent exists during a given session. The interplay between persistent memory and the right to identity self-determination is a distinctive challenge that sets healthcare personalization apart from consumer domains, where erasure of past behavioral data is often permissible and uncontentious.

Accountability structures must be clearly defined for adverse outcomes. If a patient experiences harm following a system recommendation, the questions of liability involve the model developers, the deploying healthcare institution, and the treating clinician who may have over-relied on the AI. We argue for a collaborative governance framework inspired by the safety practices in aviation, where confidential incident reporting, root cause analysis, and systemic remediation are prioritized over individual blame. The system's logging infrastructure must capture a full audit trail of every recommendation, including the memory state, query interpretation, retrieved evidence, and generated output, enabling retrospective analysis without compromising patient privacy. Regulatory pathways for adaptive AI systems, which learn and evolve continuously, are still nascent; the U.S. Food and Drug Administration's evolving framework for software as a medical device offers a foundation, but must be extended to address the unique challenges of LLM-based, memory-augmented systems that straddle the boundary between clinical decision support and autonomous recommendation [22].

7. Deployment, Scalability, and Sustainability

Translating the proposed architecture from a research concept into a live healthcare environment requires careful attention to deployment topology, computational resource management, and long-term sustainability. The computational demands of large language models are substantial, and serving millions of patients with low-latency, memory-intensive inference poses a non-trivial engineering challenge. The architecture must be designed to scale down for resource-constrained settings as well, ensuring that the benefits of personalized recommendation are not limited to well-funded academic medical centers.

A promising deployment strategy leverages a cascade of model sizes. The full-scale LLM with expansive memory retrieval capability may be reserved for complex, high-acuity queries,

while a distilled, edge-deployable model handles routine interactions such as appointment scheduling or medication reminders. This tiered approach mirrors the clinical workforce model, where general practitioners manage common cases and specialists are consulted for complex problems. The distillation process must ensure that safety-critical behaviors are preserved in the smaller model, a process that can be guided by knowledge distillation with a focus on boundary-case performance. Moreover, the memory model's hierarchical structure enables partial caching at edge nodes, with the most frequently accessed memory segments stored locally on the patient's device and less common entries fetched from a secure cloud service on demand.

Sustainability concerns extend beyond computational efficiency to the environmental footprint of large-scale AI in healthcare. The carbon impact of training and serving LLMs has drawn increasing scrutiny, and healthcare organizations committed to the principle of "do no harm" must weigh the environmental externalities of their technology choices [23]. Our architecture mitigates this through several mechanisms: the use of retrieval-augmented generation reduces the need for ever-larger parametric models by externalizing knowledge to memory stores; federated training reduces the centralization of computation; and the tiered deployment model channels the majority of interactions through lightweight models. System designers should publish transparent estimates of energy consumption per patient interaction, enabling healthcare systems to make informed procurement decisions that align with their sustainability commitments.

Interoperability with existing health IT infrastructure is a practical prerequisite for deployment. The system must integrate with heterogeneous electronic health record systems, each with its own data model and API conventions. Adopting the Fast Healthcare Interoperability Resources standard as a canonical data representation layer allows the memory model to ingest and output FHIR-compliant resources, while the language model handles the mapping between FHIR structures and narrative text. This decoupling ensures that the system can be plugged into diverse healthcare ecosystems without bespoke integration for each site. As the healthcare industry gradually moves toward a more open, API-driven ecosystem, architectures that anticipate and embrace interoperability will be better positioned for widespread adoption.

The economic sustainability of personalized recommendation systems hinges on demonstrating measurable improvements in patient outcomes and reductions in healthcare costs. While the promise of preventing adverse drug events, reducing unnecessary emergency department visits, and improving chronic disease management is compelling, rigorous health economic evaluations are needed. The architecture's design for incremental deployment and A/B testing within ethical boundaries enables phased clinical trials that compare outcomes between patients receiving AI-augmented recommendations and those receiving standard care. Long-term viability will also depend on business models that do not exploit patient data for advertising or price discrimination; a model of institutional subscription or value-based reimbursement, where the system developer is compensated based on achieved health improvements, aligns incentives more closely with patient welfare.

8. Evaluation and Cross-Domain Perspectives

Evaluating a system of this complexity requires a multi-faceted framework that extends beyond traditional accuracy metrics. In consumer recommendation, success is often measured by click-through rates or user engagement, metrics that are inappropriate and potentially dangerous in health contexts. A personalized health recommendation system must be assessed

on safety, clinical accuracy, alignment with guidelines, patient satisfaction, health outcomes, and equity across populations. We advocate for a layered evaluation methodology, beginning with offline benchmarking against curated clinical case repositories, progressing through simulated patient interactions with clinician review, and culminating in prospective clinical trials.

The offline evaluation phase employs datasets such as MedQA, clinical natural language inference benchmarks, and de-identified patient trajectory data to measure the system's ability to retrieve relevant history, disambiguate queries, and generate guideline-concordant responses. Crucially, these benchmarks must include counterfactual test cases where the patient's history contains conflicting or misleading information, to stress-test the system's robustness. The retrieval component can be evaluated using ranking metrics such as recall at k for clinically salient events, while the generated recommendations are assessed by automated metrics complemented by expert clinician panel judgments.

Cross-domain comparison with personalization systems in other high-stakes domains, such as legal advice or financial planning, reveals common architectural patterns and divergent requirements. In legal recommendation, as in medicine, the system must navigate statutory frameworks and avoid the unauthorized practice of law, paralleling the clinical need to avoid practicing medicine without a license [24]. Both domains require meticulous citation of sources and clear delineation of the system's limitations. However, healthcare introduces additional layers of physiological urgency and the irreversibility of certain outcomes, which demands more conservative recommendation thresholds. The memory modeling paradigm developed for healthcare, with its emphasis on temporality, consent, and ethical constraints, may in turn inform the design of AI systems in other sensitive domains where persistent user models are deployed.

Another valuable cross-domain lens is the comparison with ambient intelligence systems in elder care, which continuously monitor activity patterns and generate alerts. These systems similarly balance beneficence with privacy, and their failure modes—such as false alarms eroding trust or missed events leading to injury—mirror the challenges of health recommendation. The lesson is that any memory-augmented system must be designed for graceful degradation; when confidence is low or data is missing, the system should default to cautious inaction rather than proactive speculation. This design ethos, sometimes termed the "do-nothing" baseline, is surprisingly robust in high-risk settings and should be a conscious architectural choice rather than an afterthought.

Looking forward, the frontier of evaluation will involve continuous post-market surveillance once the system is deployed in live settings. Unlike a static drug or device, an LLM-based memory system will evolve as the underlying model is updated, as the patient's memory grows, and as clinical guidelines change. This necessitates an ongoing monitoring infrastructure that tracks recommendation patterns, detects drift in safety or fairness metrics, and triggers re-evaluation when significant changes occur. Regulatory science is only beginning to grapple with the implications of such learning health AI systems, and close collaboration between system architects, clinicians, and regulators will be essential to craft sensible, adaptive oversight mechanisms [25].

9. Conclusion

The convergence of large language models, longitudinal patient data, and novel memory architectures opens a transformative chapter in personalized healthcare recommendation. This

paper has articulated a system-level vision for an LLM-based patient memory model coupled with deep medical query understanding, emphasizing that the transition from generic AI assistants to trusted health advisors requires more than better token prediction. It demands a thoughtful interweaving of narrative memory, clinical knowledge, ethical governance, and sustainable engineering. We have argued that the memory model must be hierarchical, temporally-aware, contestable, and privacy-preserving; that query understanding must handle ambiguity with calibrated uncertainty and personalized semantic grounding; and that the recommendation engine must mediate between patient preferences and evidence-based medicine through transparent negotiation mechanisms.

The structural trade-offs examined throughout this paper—between model scale and sustainability, between personalization and fairness, between autonomy and safety—are not obstacles to be overcome but enduring design tensions that will shape the evolution of health AI for years to come. The paper's governance analysis underscores that accountability, consent management, and equitable access are not external policy decorations but integral to the architecture's technical core. As healthcare systems worldwide grapple with aging populations, rising costs, and workforce shortages, well-designed AI recommendation systems can augment clinical capacity and empower patients in self-management. However, the realization of this potential hinges on a rigorous, interdisciplinary approach that treats these systems not as mere tools but as socio-technical actors embedded in the moral and relational fabric of care. The patient memory model, in its deepest sense, becomes a digital extension of the human capacity for health narrative, a co-constructed artifact that supports continuity of identity and informed decision-making across a lifetime of health journeys.

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, 5998–6008.

2. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240.

3. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23.

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

5. Sukhbaatar, S., Weston, J., Fergus, R., et al. (2015). End-to-end memory networks. Advances in Neural Information Processing Systems, 28, 2440–2448.

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

7. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. ACM Computing Surveys, 52(1), 1–38.

8. Yin, F., Wang, L., Liu, J., Zhou, X., & Li, Q. (2021). GAMENet: Graph augmented memory network for medication recommendation. IEEE Transactions on Knowledge and Data Engineering, 35(1), 699–712.

9. Yu, C., Ren, G., & Liu, J. (2020). Deep reinforcement learning for treatment recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 1184–1191.

10. 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.

11. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.

12. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

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

14. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.

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

16. Jagannatha, A. N., & Yu, H. (2016). Bidirectional RNN for medical event detection in electronic health records. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 473–482).

17. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

18. Elwyn, G., Frosch, D., Thomson, R., Joseph-Williams, N., Lloyd, A., Kinnersley, P., ... & Barry, M. (2012). Shared decision making: A model for clinical practice. Journal of General Internal Medicine, 27(10), 1361–1367.

19. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.

20. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214–226).

21. Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3543–3556).

22. U.S. Food and Drug Administration. (2021). Artificial intelligence and machine learning in software as a medical device. FDA.

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

24. Susskind, R., & Susskind, D. (2015). The future of the professions: How technology will transform the work of human experts. Oxford University Press.

25. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38.