Personalized Healthcare Recommendation Systems Based on Patient Narratives and Large Language Model-Based Self-Representation Learning

Arjun Kerhra
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
arjunm@binghamton.edu

Dustin Page
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
dpage@uab.edu

Abstract

The increasing availability of unstructured patient-generated health narratives, ranging from symptom diaries to psychosocial reflections, presents a transformative opportunity for precision medicine. This paper investigates the design, infrastructure, and governance of personalized healthcare recommendation systems that integrate patient narratives with large language model (LLM)-based self-representation learning. Rather than focusing narrowly on model accuracy, we offer a system-level analysis of the architectural trade-offs, deployment challenges, and regulatory implications inherent in such platforms. We conceptualize self-representation as a dynamic, linguistically encoded construct that LLMs can extract and refine through iterative narrative processing, enabling recommendations that align not only with clinical histories but also with an individual's evolving self-concept, values, and experiential context. The paper examines the complete pipeline from narrative ingestion and semantic segmentation to the construction of a personal knowledge graph fused with a continuously updated latent self-representation. Key system dimensions are dissected: the tension between centralized and federated architectures for data governance, the robustness of self-representation against distributional shifts and adversarial noise, fairness considerations when modeling subjective identity constructs, and the sustainability of deploying compute-intensive LLMs in health systems. We further address policy implications surrounding data ownership, the right to explanation in algorithmically derived self-models, and the need for auditability. By integrating perspectives from human-computer interaction, clinical informatics, and responsible AI, this work provides a comprehensive framework for building narrative-driven recommendation systems that are clinically safe, ethically grounded, and structurally resilient. The analysis highlights how the fusion of patient narratives and self-representation learning can shift recommendation paradigms from population-based protocols toward truly person-centered care, while also outlining the sociotechnical guardrails necessary to realize that vision.

Keywords

personalized healthcare, recommendation systems, patient narratives, large language models, self-representation learning, fairness, system architecture, governance.

1. Introduction

The digital transformation of healthcare is generating an unprecedented volume of unstructured patient data, much of it in the form of free-text narratives. These narratives encompass symptom descriptions, daily reflections, psychosocial stressors, treatment experiences, and personal goals, often captured through patient portals, mobile health apps, and conversational agents. Unlike structured electronic health records, such texts encode nuanced, temporally evolving, and deeply subjective dimensions of illness and wellness that are critical for personalization yet remain underexploited by current decision support systems. In parallel, large language models have matured to a point where they can parse, summarize, and reason over lengthy textual inputs with semantic depth that rivals human capabilities in specific domains [1, 2]. The convergence of these trends invites a rethinking of healthcare recommendation systems, from rule-based or purely data-driven engines to platforms that derive patient models from the very language individuals use to articulate their lived experiences.

Recent foundation models, such as GPT-4 and LLaMA, have demonstrated the capacity to generate coherent, context-aware responses across specialized medical examinations and patient inquiries [3, 4]. Meanwhile, domain-adapted variants like BioBERT and ClinicalBERT have refined clinical named entity recognition and relation extraction, proving the feasibility of mining actionable signals from clinical notes [5]. Despite this progress, the integration of patient-authored narratives into operational recommendation pipelines remains nascent. Extant systems typically treat narratives as supplementary textual features for classification tasks, rather than as primary substrates for constructing holistic self-representations that evolve with the patient. This paper contends that the next frontier lies in LLM-based self-representation learning—the process of distilling a dynamic, high-dimensional latent representation of a patient's self-concept, preferences, and values directly from their longitudinal narrative stream, and using that representation to steer personalized recommendations.

Framing the challenge as a system-level problem shifts the emphasis from isolated model metrics toward a broader set of concerns. The ingestion of deeply personal narratives demands infrastructure that upholds privacy, consent, and data sovereignty while enabling the computational scaling required for transformer-based inference. The semantic richness of narratives introduces new forms of bias, because self-representation is mediated by language, culture, and health literacy, making fairness and representation core engineering constraints. Moreover, deploying such systems in clinical workflows necessitates robust governance frameworks that regulate how self-models are stored, updated, and explained. This paper offers a comprehensive interdisciplinary analysis of these dimensions, drawing on systems engineering, human-computer interaction, and policy studies to articulate a blueprint for narrative-driven personalized recommendation systems. We structure the discussion around architectural design, infrastructure and scalability, robustness and fairness, and sustainability and policy, after first unpacking the conceptual underpinnings of patient narratives and self-representation learning.

2. Patient Narratives and Self-Representation Learning

Patient narratives represent a distinct category of health data that is inherently interpretive, temporally contextualized, and performative. Drawing on traditions of narrative medicine, scholars have long argued that the way patients story their illness shapes identity, coping, and health behaviors [6]. In computational terms, these narratives form a high-dimensional text stream from which patterns of symptom expression, emotional valence, and personal agency

can be extracted using natural language processing. Early work in digital phenotyping demonstrated that language features derived from social media and journaling correlate with mental health states, offering a glimpse into the diagnostic potential of everyday expressive language [7]. However, moving from correlation-based phenotyping to causal recommendation requires a deeper representational substrate: a model of the self that captures stable traits, dynamic states, and value hierarchies.

Self-representation learning, as conceptualized in this work, is the process of encoding a patient's multi-faceted identity—their health beliefs, goals, cultural context, and emotional landscape—into a continuous vector space that can be fed into downstream recommendation modules. Large language models are uniquely suited for this task because their pretraining on diverse corpora endows them with a broad semantic palette for interpreting subjective experience. By fine-tuning on longitudinal patient narratives while preserving a privacy-preserving architecture, an LLM can learn to disentangle stable aspects of self-representation from transient mood or acute symptom spikes. The resulting latent space can be structured to respect meaningful clinical dimensions, such as readiness to change in behavioral interventions or perceived social support in chronic disease management [8, 9].

Crucially, self-representation is not a fixed essence but a fluid construct that is continuously renegotiated through narrative acts. A patient might reframe their cancer journey from one of victimhood to one of resilience, or revise their health goals after a life event. Traditional collaborative filtering or static profile-based recommender systems are poorly equipped to track such ontological shifts, because they treat user models as slowly drifting parameters. In contrast, an LLM-based self-representation module can attend to rhetorical markers, tense shifts, and causal attributions that signal identity transformation, updating the internal model in near real-time. This dynamic capability aligns with insights from HCI research indicating that people actively construct their self-identity through reflective writing and that digital tools can either support or distort this process [16]. Designing systems that honor the agency of the patient in shaping their own self-representation thus becomes both a technical and an ethical imperative.

The challenge of grounding self-representation learning in clinical validity requires bridging subjective narratives with objective health records. One promising approach is to anchor LLM-derived representations to established psychological constructs, such as the transtheoretical model of behavior change or the common-sense model of illness self-regulation [10, 11]. By projecting narrative embeddings onto a semantic space defined by these theoretical frameworks, the system can generate representations that are interpretable to clinicians and patients alike. This hybrid approach reduces the black-box nature of pure neural representations and facilitates the integration of self-representation scores into clinical decision support rules. However, it also raises questions about the cultural universality of psychological models and the risk of imposing normative frameworks on diverse narrative expressions, a tension that must be managed through inclusive design and continuous community feedback.

3. Architectural Design of Narrative-Driven Recommendation Systems

Designing a recommendation system that ingests patient narratives and leverages LLM-based self-representations demands a layered architecture that separates concerns of data ingestion, representation learning, recommendation reasoning, and user interaction. A canonical architecture can be conceptualized as four interconnected layers. The narrative ingestion layer handles secure collection, pseudonymization, and structured segmentation of free-text inputs;

the self-representation layer employs a fine-tuned transformer backbone to generate and update a personal knowledge graph coupled with a continuous embedding; the recommendation engine fuses this representation with evidence-based clinical guidelines, drug interaction databases, and population health models to produce personalized suggestions; and the interaction layer delivers recommendations through explainable interfaces that incorporate narrative justifications.

One fundamental architectural trade-off resides at the boundary between narrative processing and model inference. Patient narratives are inherently longitudinal, with varying update frequencies and lengths, which poses challenges for batch-oriented LLM inference pipelines. An event-driven microservice architecture can decouple narrative ingestion from representation recalculation, allowing real-time streaming of patient entries while scheduling resource-intensive LLM refreshes during low-demand windows [12]. The representation layer must maintain versioned snapshots of the self-representation to support retrospective auditability and to enable the recommendation engine to trace how a particular suggestion was influenced by the patient's narrative history. This design choice introduces storage overhead but is indispensable for explainability and regulatory compliance.

The recommendation engine itself must balance multiple competing objectives: maximizing clinical benefit, respecting patient preferences encoded in the self-representation, and avoiding oversteering based on linguistically amplified but clinically minor signals. A common strategy is to decompose the recommendation problem into a multi-stage pipeline in which a candidate generation phase exploits narrative similarity and self-representation alignment to retrieve a broad set of interventions, followed by a ranking phase that applies clinical constraints, contraindication filters, and fairness adjustments. The modularity of this design allows domain experts to inject medical logic at multiple checkpoints, thereby preventing the LLM-derived representation from acting as an unchecked oracle. It also enables the incorporation of federated learning schemes, where the self-representation model is trained across institutional boundaries without centralizing raw narrative data, preserving patient privacy while enriching the latent space with diverse linguistic and cultural patterns [13].

A related architectural decision concerns the placement of the LLM backbone. In a fully on-device paradigm, a compressed or distilled version of the model runs on the patient's smartphone or home hub, processing narratives locally and transmitting only anonymized representation vectors to the recommendation server. This architecture provides the strongest privacy guarantees and reduces network dependency but significantly constrains model capacity and update frequency. A hybrid option deploys a larger LLM in a trusted execution environment at the healthcare provider's edge cloud, balancing computational power with strict data residency requirements. The choice between these configurations is not merely technical; it is entangled with jurisdictional data protection laws, institutional risk tolerance, and patient digital literacy, making it a quintessential sociotechnical design problem.

4. Infrastructure, Scalability, and Deployment Considerations

Scaling narrative-driven recommendation systems to health-system-wide deployment introduces infrastructure challenges that transcend conventional machine learning operations. The computational footprint of transformer-based inference is considerable, especially when processing lengthy, conversational narrative streams for thousands of patients with low-latency requirements. Techniques such as distillation, quantization, and speculative decoding have made strides in reducing inference costs, yet the healthcare context demands careful

trade-offs against accuracy, as even small degradation in representation quality could propagate into unsafe recommendations [14]. Infrastructure planners must therefore design elastic compute clusters capable of absorbing diurnal and seasonal variations in narrative input rates, while maintaining strict service level agreements for critical alerts such as medication reminders or mental health crisis escalations.

Data governance infrastructure constitutes another pillar of deployability. Patient narratives are among the most sensitive forms of personal data, often containing identifiable details about family members, locations, and traumatic events. Architectures that embrace data minimization principles must implement robust de-identification pipelines that operate on structured and unstructured text. Modern LLMs can both assist in de-identification—by generating privacy-preserving paraphrases—and inadvertently reintroduce identifiable information during representation learning if not carefully controlled through differential privacy mechanisms [15]. A mature deployment therefore layers formal privacy frameworks, such as k-anonymity on metadata and epsilon-differential privacy on gradient updates during fine-tuning, atop comprehensive consent management systems that allow patients to set granular permissions for narrative use.

Integration with existing health information systems presents additional deployment friction. Most hospital ecosystems rely on HL7 FHIR interfaces for structured data exchange, a paradigm that does not natively accommodate high-dimensional narrative embeddings. Middleware adapters must translate between the self-representation vector space and the categorical ontology of electronic health records, for instance by projecting the representation onto standardized clinical vocabulary codes for problem lists and care plans. This ontological bridging is not lossless, and the design of the mapping function—whether linear projection with concept bottleneck layers or post-hoc labeling by a clinician-in-the-loop system—directly affects the fidelity and auditability of recommendations. Furthermore, network resilience becomes a safety-critical property, because a breakdown in the streaming pipeline during a patient's narrative entry could delay the detection of acute deterioration. Hospital-grade deployments must therefore implement buffering, retry logic, and fallback to simpler rule-based recommenders when LLM services are degraded, ensuring that the system fails safely.

5. Robustness, Fairness, and Governance

The reliance on linguistic self-representation introduces robustness challenges that are distinct from those in imaging or structured lab data. Language is inherently ambiguous, context-dependent, and susceptible to adversarial manipulation, whether intentional or arising from cognitive decline, affective extremes, or social desirability biases. A patient in a depressive episode may narrate their health status in globally negative terms that exaggerate functional impairment, leading the self-representation model to mistakenly downgrade their readiness for rehabilitative exercise recommendations. Conversely, a patient seeking a specific medication might inadvertently craft narratives that align their self-representation with the clinical criteria for that drug, a form of steering that could undermine the system's clinical integrity. Robustness strategies must therefore encompass input validation through semantic consistency checks, temporal smoothing of representation updates to dampen acute narrative volatility, and ensemble methods that cross-reference self-representation signals with objective data streams such as wearable sensor readings.

Fairness in narrative-based recommendation is a multifaceted problem that intersects with language standardization, cultural representation, and health equity. LLMs pretrained

predominantly on English-language internet corpora encode sociolinguistic biases that can systematically distort self-representations for patients who speak non-dominant dialects, use minority narrative genres, or express health distress through culturally specific metaphors [17, 18]. If a system interprets low health literacy or non-standard syntax as a sign of lower self-efficacy, it may recommend less ambitious treatment plans, perpetuating health disparities. Mitigation requires deliberate efforts in dataset curation, dialect-inclusive fine-tuning, and the development of fairness metrics that capture cultural fidelity rather than simply demographic parity. Participatory design with patient communities can surface narrative styles that the base model misinterprets, informing targeted data augmentation and representation recalibration.

Governance of LLM-based self-representation systems must address the dual nature of the self-representation as both a computational artifact and a deeply personal construct. Patients should have the right to access, correct, and even delete their self-representation, mirroring GDPR-inspired data subject rights but extended to latent model states [19]. This creates a technical challenge because a high-dimensional embedding cannot be intuitively inspected or edited without specialized interfaces. Governance frameworks must therefore mandate that any recommendation derived from a self-representation be accompanied by a human-readable narrative rationale, possibly generated through chain-of-thought distillation, which reconstructs the chain of reasoning from the patient's own words to the suggestion. Institutional review boards and regulatory bodies like the FDA, which are beginning to regulate AI/ML-based software as a medical device, will need to evolve protocols for pre-market review that encompass continuous learning systems whose behavior changes as a patient's narrative self-representation evolves over time [20, 21].

6. Sustainability and Policy Implications

The sustainability of narrative-driven healthcare recommendation systems extends beyond environmental considerations to encompass long-term economic viability, technical maintenance, and societal trust. The energy demands of large-scale LLM inference have been extensively documented, and a system that continually reprocesses full narrative histories for millions of patients would carry a significant carbon footprint unless augmented with caching, sparse activation, and renewable energy-powered inference clusters [22]. System designers must incorporate carbon-aware scheduling algorithms that align compute-intensive representation updates with periods of low grid carbon intensity, alongside efforts to develop smaller, domain-specialized models that achieve comparable self-representation fidelity with orders of magnitude fewer parameters. Economic sustainability also depends on reimbursement models; if narrative-driven personalization reduces hospital readmissions or improves chronic disease management, value-based payment arrangements can offset the infrastructure costs, but these outcomes must be demonstrated through rigorous health economics research.

From a policy standpoint, the deployment of systems that derive intimate self-models from patient stories challenges existing regulatory categories. Is the self-representation a medical record, a behavioral inference, or a new class of digital personhood data? Different jurisdictions may answer this question divergently, creating a fragmented compliance landscape that could stifle cross-border health platforms. Clear legislative guidance is needed to define the rights and responsibilities associated with algorithmic self-representations, including protocols for mandatory disclosure when a self-representation is used to deny or modify care. Additionally, policies must address the risk of dual use, such as insurers or employers leveraging self-representation data to discriminate or to apply subtle nudges that

undermine patient autonomy, mandating that these systems are deployed exclusively within fiduciary clinical relationships.

Interoperability standards will play a pivotal role in policy-driven adoption. Without common formats for exchanging narrative annotations and self-representation vectors, the ecosystem risks balkanization, where each vendor's self-model becomes a walled garden that locks patients and providers into proprietary platforms. Organizations such as HL7 and IEEE are exploring extensions to FHIR that could accommodate AI-derived representations, but these efforts must be accelerated and coupled with certification requirements that validate the safety, fairness, and transparency of recommendation engines before they receive interoperability credentials. International harmonization, while ambitious, would reduce the regulatory burden and foster an open research community that can independently audit and improve these sociotechnical systems.

7. Conclusion

Personalized healthcare stands at a turning point, where the richness of patient narratives can be unlocked through large language model-based self-representation learning to drive recommendations that are attuned to the whole person, not merely a collection of symptoms and lab values. This paper has argued that realizing this vision requires a deliberate shift toward system-level thinking, one that accounts for the architectural interdependencies among narrative ingestion, self-representation modeling, and recommendation delivery, as well as the broader fabric of infrastructure, fairness, robustness, governance, and policy. By analyzing the design trade-offs between centralized and edge-based deployment, the challenges of distilling culturally situated self-representations without perpetuating bias, and the imperative of building semantically aligned governance structures, we have charted a pathway for constructing narrative-driven platforms that are both technically sound and ethically defensible. The integration of dynamic self-representations into clinical workflows holds the promise of transforming healthcare recommendation from a static, population-averaged exercise into a continuously adaptive dialogue between patients and the systems that serve them. Achieving that transformation demands interdisciplinary collaboration, longitudinal real-world validation, and an unwavering commitment to preserving patient dignity in the algorithms that learn from their stories.

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. In Advances in Neural Information Processing Systems (pp. 5998–6008).

2. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shinn, 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. In Advances in Neural Information Processing Systems (pp. 1877–1901).

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

4. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Cantón, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

5. 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.

6. Charon, R. (2001). Narrative medicine: A model for empathy, reflection, profession, and trust. JAMA, 286(15), 1897–1902.

7. Insel, T. R. (2017). Digital phenotyping: Technology for a new science of behavior. JAMA, 318(13), 1215–1216.

8. McAdams, D. P. (2001). The psychology of life stories. Review of General Psychology, 5(2), 100–122.

9. Prochaska, J. O., & Velicer, W. F. (1997). The transtheoretical model of health behavior change. American Journal of Health Promotion, 12(1), 38–48.

10. Leventhal, H., Phillips, L. A., & Burns, E. (2016). The Common-Sense Model of Self-Regulation (CSM): a dynamic framework for understanding illness self-management. Journal of Behavioral Medicine, 39(6), 935–946.

11. Pennebaker, J. W., & Chung, C. K. (2011). Expressive writing: Connections to physical and mental health. In H. S. Friedman (Ed.), The Oxford Handbook of Health Psychology (pp. 417–437). Oxford University Press.

12. Hohman, F., Wongsuphasawat, K., Kery, M. B., & Patel, K. (2021). Understanding and visualizing data iteration in machine learning. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1–13). ACM.

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

14. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

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

16. Solanki, D., Hsu, H. M., Zhao, O., Zhang, R., Bi, W., & Kannan, R. (2020, July). The way we think about ourselves. In International Conference on Human-Computer Interaction (pp. 276-285). Cham: Springer International Publishing.

17. 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.

18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

19. European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.

20. World Health Organization. (2021). Ethics and governance of artificial intelligence for health. World Health Organization.

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

22. 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). Association for Computational Linguistics.