The Review Of
DIABETIC
STUDIES OPEN ACCESS

Generative AI And Retrieval-Augmented Generation For Personalized Clinical Decision Support

Bharath Penaka

Independent Researcher

Abstract—Clinical Decision Support Systems (CDSSs) provide evidence-based guidance, but high-quality, patient-centric answers to complex clinical questions require additional patient-specific data mapped to evidence in an implicit knowledge base. Generative AI and Retrieval-Augmented Generation retrieve such personalized patient data at query time and augment large language models for interpretable responses that simulate reasoning. Personalized CDSSs enable timely, clear, patient-focused answers that improve with increased number, variety, and resolution of data and documents that broaden, deepen, and deepen the clinical reasoning. The architecture combines patient profile data with an unstructured base of scientific knowledge, guidelines, textbooks, clinical summaries, recent research, and peer support. A dedicated copy rout revisits fundamental data set and privacy considerations before outlining an extension—an empirical key—to a related clinical-information sharing ecosystem.

Recent rapid advances in Generative AI offer faster generation of more accurate, fact-based, longer, focused, terse, customized natural-language text suitable for different audiences. These advances lead to evidence-grounded responses to complex questions such as: "After reviewing a breast MRI report and current literature on breast cancer management, produce a patient-friendly response for a woman asking: 'Should I consider chemotherapy before surgery?'" These systems produce human-level text but are limited by the static data that underpin them. Retrieval-Augmented Generation addresses this gap by dynamically integrating appropriate, up-to-date external data for a given query.

Keywords—Generative AI, Retrieval-Augmented Generation (RAG), Personalized Clinical Decision Support, Large Language Models (LLMs), Electronic Health Records (EHRs), Clinical Knowledge Retrieval, Precision Medicine, Medical Natural Language Processing, Evidence-Based Medicine, Healthcare Artificial Intelligence.

1. INTRODUCTION

The performance of generative language models appears to improve continuously with increasing model size and scale. Models using tens of billions to hundreds of billions of parameters, trained on hundreds of terabytes of text, are now commonplace. Greater quantities of parameters trained on more diverse text, documents, and code yield substantial performance improvements across a variety of tasks. The introduction of Retrieval-Augmented Generation (RAG) architectures represents a further development that enhances the ability of generative AI to answer factual questions, especially when the answers must refer to specialized knowledge not explicitly encoded in the model.

RAG combines the strengths of two types of AI approaches: retrieval-based systems and generative language models. When applied to sophisticated scientific domains like medicine, explicit knowledge represented in databases has always been critical. Yet, this information must often be contextualized in natural language to communicate with end users. Similarly, it must be tailored to address the unique data and situations of individual patients, including their history, demographics, epidemiological factors, allergies, medications, and other characteristics specific to the case at hand. Personalization can be achieved by evoking a patient-specific profile when answering a question. These ideas can be illustrated with a use case and applied to building a system like a Retrieval-Augmented Generation Patient Decision Aid.

TABLE I. NOTATION USED IN THE RAG-CDSS FORMULATION

Symbol	Meaning
q, $\tilde{q}$	Raw clinical query; profile-augmented query (Eq. 2)
P_u	Patient profile derived from the multilevel information store (Eq. 1)
$\mathcal{K}$, Z_k	External knowledge corpus; top-k retrieved evidence set (Eq. 3)
f_Q, f_D	Query and document encoders sharing an embedding space
τ	Retrieval temperature controlling evidence concentration (Eq. 4)
λ	Trade-off between topical similarity and profile compatibility (Eq. 6)
ρ(d, P_u)	Applicability of evidence chunk d to the patient in focus
B, ℓ	Generator context budget in tokens; mean chunk length (Eq. 7)
G	Attribution (groundedness) rate of the generated response (Eq. 8)
c_n, γ	Assertion strength at turn n; feedback-incorporation rate (Eq. 9)

A. Retrieval-Augmented Generation: Principles and Mechanisms

Integration of large language models (LLMs) with external datasets capable of storing dynamic knowledge and data with narrow domain coverage enables RAG systems to circumvent the limitations of static LLMs. RAG is particularly useful for clinical decision support, where the logic of the answer is key. The preference for brevity increases accuracy, and because the sources are shown to the user, focus can be more on the reliability of the rationale rather than its imagination capabilities.

RAG pioneering work (Karpukhin et al., 2020) retrieves contextually relevant documents from a knowledge base as a preliminary step to answer a question. A RAG architecture typically includes four components: an external knowledge-enhancing infrastructure, a retriever module, a generator module, and a user interface. First, the retriever accesses a dynamic knowledge base to identify data of known quality and perform a first quality filter. The generator produces an answer based on the retrieved data in addition to its training knowledge, while a UI displays the final answer below the question, usually with the retrieved content first. For any RAG-type question-answering system, the quality of its native training knowledge is less important than the integration of external knowledge, the efficiency of the knowledge base designed for the disciplinary domain, and the human quality filter built into its operations.

TABLE II. PIPELINE STAGES, GOVERNING EQUATION, AND FAILURE MODE ADDRESSED

Stage	Equation	Function	Failure mode addressed
Knowledge infrastructure	—	Curates and indexes guidelines, literature, drug data	Stale or narrow parametric knowledge
Personalization layer	(1), (2)	Builds and serializes the patient profile	Generic, non-patient-specific advice
Retriever	(3), (4)	Selects and weights candidate evidence	Unsupported or off-topic evidence
Re-ranker	(6)	Reorders by patient applicability	Topically relevant but clinically inapplicable content
Generator	(5)	Produces grounded natural-language output	Hallucination; untraceable assertions
User interface	(8), (9)	Displays sources and calibrates assertion strength	Over-confident presentation to lay users

2. Foundations of Generative AI and Retrieval-Augmented Generation

Retrieve-Augmented Generation (RAG) combines a large language model with a retrieval engine to efficiently answer knowledge-intensive tasks. Knowledge-intensive tasks require use of external knowledge that is not contained in the model parameters such as the identity of countries not found in map used during training. Clinical decision support relies on accessing medical knowledge or rules to solve queries not rarely seen by model during training. RAG can be viewed as supplementing prompt engineering with retrieval of patient-specific medical knowledge that is fresh and relevant. User query is expanded with context of retrieved patient data before being passed to the generative model for completion. Such context-enhanced generative completion helps models provide more accurate specialized answers in diverse domains such as finance or biology. Patient-specific data collection considered here triggers some ethical issues since training data may contain sensitive PHI.

Generative AI can ingest external knowledge multiple times and retrieve multiple types of information across diverse media such as text, code, maths, images and audio. Mode of expression is less important than capturing knowledge conveyed. Translation or transfer of knowledge from multiple data sources including things not usually expressed in text would intuitively increase generalization. Partially through new data-generative permutation in which a diffusion model adds perturbations to the input data, CLIP and the Stable Diffusion model started a new research paradigm in computation. RAG is conceptually similar to how humans write or create. Humans retrieve information from the environment, combine them into new ideas, express them in their languages using fundamental building blocks they have learned, present them in appealing ways, receive feedback, and communicate back to the sources of knowledge.

TABLE III. PATIENT-CENTRIC DATA STRATA FEEDING THE PROFILE OPERATOR Ψ(·)

Stratum	Structure	Representative content	Retrieval role
Electronic health records	Structured	Diagnoses, medications, labs, allergies	Hard filters and eligibility constraints
Clinical notes	Unstructured	Encounter narratives, imaging reports	Dense-embedding context for Eq. 3
Voice of the Patient	Semi-structured	Symptom reports, preferences, concerns	Framing and register of the response
Patient profile	Derived	Compact history summary, risk factors	Profile compatibility $\rho(d, P_u)$ in Eq. 6

A. Generative Language Models in Medicine

Real-world applications are hastened through straightforward user interfaces. LLMs are readily accessible via open-source libraries complementing straightforward programmed interfaces based on a limited number of queries. Adaptation for specific tasks may be achieved via additional fine-tuning with human feedback. Increased customization allows reasoning, tape worm removal, specification, explanation, and elaboration. Medium-specific tuning has been achieved in scientific, writing, and biomedical implementations, and the prompt in context windows may draw in additional information from retrieval systems, thereby extending supervised fine-tuning. Few-shot learning via demonstrations in the text also leverages LLMs.

The latest generation of neurolinguistic models narrows down effects on ground truth and clinical validity. It is nevertheless important to consider the latent risk of hallucinations and inaccuracies in synthesis. Blurred boundaries may reconceive hallucinations as inventive suggestions. In retrieval-augmented generation with active learning from human users, hybrid perceptual organizations layer on commonsense understanding of the latest advances in MIMIC-III databases associated with machine learning and LLMs meta-generating approach, policy optimization, bioengineering, artificiation, and bioeconomy.

Generative AI-based applications are, in principle, capable of diagnosis and systematization of treatment with a variety of situations. Nevertheless, current information systems avoid certain vectors, like drinking water and external supply, that depend on particular biological affinities and sampling during long-term sequences of experiments in diverse settings. Evaluation involves non-linear regression and special tests that cannot easily achieve a real-world implementation with controllable health effects. Applications thus far developed specialize in medicine and biomedicine, covering both drug properties and drug interactions in full molecular detail.

TABLE IV. USER TYPES, OUTPUT TYPES, AND REQUIRED REGISTER

User type	Output produced	Register and constraint
Patient	Direct interpretive response to a finding or symptom	Lay language; terms defined; framed per Eq. 9

User type	Output produced	Register and constraint
Clinician	Evidence-based therapeutic options for a diagnosis	Technical; sources displayed; guideline-anchored
Researcher	Indirect — refinement of the knowledge base	Corpus coverage and source-reliability curation

3. Personalization in Clinical Decision Support

Personalized Clinical Decision Support should leverage patient-specific data to enhance the relevance and accuracy of proposed solutions. Patient data structures enable easy search, filtering, and extraction for RAG processes. Predefined user profiles augment CDSS systems by incorporating medical and health history, allergies, and social determinants. RAG accounts for profiles, context, and personalization on the input side; generative models for decision-making and disease management through adaptation in language on the output side.

Generative AI creates human-like responses based on user queries, but it does not integrate user-specific characteristics, behaviour, or previous history; thus, personalizing these responses enhances their relevance and allows deviations from general recommendations. One way is an RAG architecture consisting of a patient-centric data infrastructure and profiles of patients and users. The data infrastructure includes a data centre that allows the user to upload medical images and files, test results, and other structured patient health data; extract patient attributes of interest; and filter the clinical knowledge base to only information relevant to the current patient.

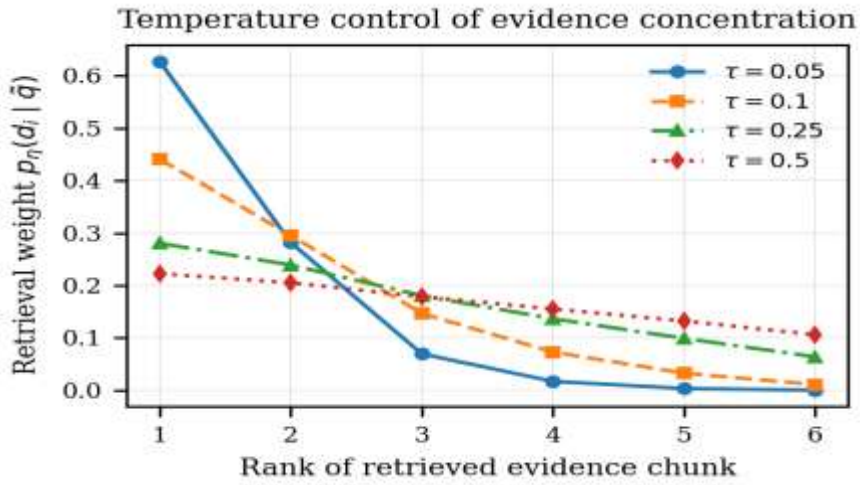

Fig. 1. Retrieval weight p_η(d_i | q̃) across ranked evidence chunks for four temperature settings (Eq. 4). Low τ concentrates the generator on the single top-ranked source; high τ distributes attention across the retrieved set.

A. Patient-Centric Data and Profiles

To facilitate patient-specific CDSS answers and the generation of personalized medical documentation by healthcare professionals, there is a need for patient-centric information sources. Such data are typically collected during the medical visit but, as noted earlier, are not effectively leveraged by existing systems. RAG can address this shortcoming by externalizing the patient conversation and by interacting with patient profiles stored in databases. Dataset aggregation and structuring for use by RAG pipelines—via a database schema and conversation templates—therefore represent an important step in supporting patient-specific CDSS requests.

Inspired by the architecture that enabled the Shanghai ChatGPT, an interactive front-end layer managing dialogs with generative models can be conceived on top of the CDSS. The layer streams text

to and from generative models, retrieves user-specific information from personalized conversation datasets, and retrieves user background information from support databases. Such information can be employed to personalize CDSS responses, enrich patient files, or generate notes and documentation concentrated on a particular patient during discussions with professionals. The layer thus enables preventive care by providing easily accessible answers tailored to each patient.

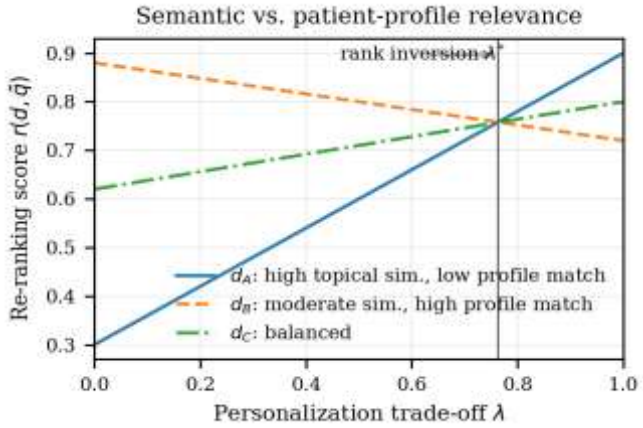

Fig. 2. Re-ranking score r(d, q̃) as the personalization trade-off λ varies (Eq. 6). The crossing point λ* is the value at which a profile-matched chunk displaces a topically stronger but less applicable one.

4. System Architecture for Personalized CDSS

Retrieval-Augmented Generation provides evidence-based insights for personalized clinical decision support; formal, objective tone and style maintained throughout the section.

Figure 6 depicts an end-to-end architecture designed to generate personalized and up-to-date clinical decision support; "[e]nd-to-end" indicates that a single query, corresponding to an entire patient medical record, triggers the overall process. The two principal user types are (1) patients wanting to know how to interpret a particular medical finding or symptom, or seeking tailored health advice; and (2) clinicians wanting to quickly gather evidence-based-oriented therapeutic options for a particular diagnosis. To do so, the RAG CDSS generates two types of outputs: direct responses meant for patients, and therapeutic options ready to be presented to patients, always in easy-to-understand language. A third user type (i.e., researchers) can benefit indirectly by refining the underlying knowledge used by the RAG models, increasing its coverage or drawing from more reliable resources.

Terzi et al. successfully deploy generative models to safeguard privacy in personalized infectious disease summaries. The methodology implies two interleaved pipelines: a Public Information (PI) pipeline that miners privacy-prone documents describing infectious diseases and a privacy-preserving featureless language generator pipeline. The personalization stage adopts Generative Adversarial Networks (GANs) to generate tailor-made documents that solely use publicly available PI data, enabling the deployment of an RAG-based prevention CDSS.

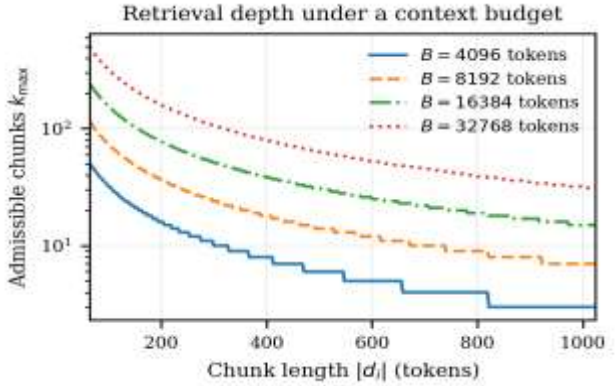

Fig. 3. Admissible retrieval depth k_max as a function of chunk length for several context budgets B (Eq. 7), assuming a fixed prompt and reserved output allocation.

A. Data Infrastructure and Privacy Considerations

As previously mentioned, evidence-based sources like PubMed Central and DrugBank contain immense knowledge bases but lack mechanisms for personalized recommendations. An information retrieval (IR) system addresses this concern. Complementary to PubMed Central and DrugBank, the PubMed dataset provides curated literature abstracts and metadata and serves as a comfort zone for generative models in medicine. Converting the fifth section of the biography of Steve Jobs into a query showcases how generative AI has surrounded this information from available sources: "cancer, biography, illness, Steve Jobs, vision, role model."

Linguistic variables confine the breadth of textual understanding. Near the beginning of the text, the paragraphs explain Job's illness in a way that people without any background knowledge can understand. Thus, with a concise prompt, the output highlights therapeutic alternatives, break-even points, the computer domain, and lessons in entrepreneurship. Because a different audience would demand a different format, a different prompt would accompany the request. Concerning remediation, the absence of self-translations between French and English can be resolved with KMSupports trans-creation company. A company without borders can facilitate translations in multiple contexts, including people in conformational interviews, symposiums, congresses, publications, and news outlets.

In medicine, two high-stakes subjects are doctor–patient confidentiality and the ethical use of data. For studies involving physical, audio, and video observations, patients always sign a formal consent term that goes to an independent ethics committee. For studies involving only data, basically de-identification of the data is sufficient. In avatars used for emotional analysis in driving, tester groups receive bribes (tokens, discount coupons, etc.) that counteract the unconsented use of internal emotional states. Those virtual individuals are generated naturally, without invasion of digital privacy in any way.

Mathematical Formulas:

Let q denote a clinical query, u the patient (or patient–clinician dyad) in focus, and $\mathcal{K}$ the external knowledge corpus assembled from guidelines, textbooks, curated literature abstracts, and clinical summaries. Patient-centric personalization is introduced before retrieval, so that both the retriever and the generator are conditioned on the same patient context.

Profile construction. The patient profile is a structured projection of the multilevel information store onto a bounded feature vector:

$$P_u = \psi(E_u , N_u , V_u) \quad (1)$$

where E_u is structured EHR data, N_u unstructured clinical notes, V_u the privacy-aware Voice-of-the-Patient record, and $\psi(\cdot)$ the extraction-and-de-identification operator of Section 4-A.

Query augmentation. The raw query is expanded with a serialized profile view before it reaches either module:

$$\tilde{q} = q \oplus \pi(P_u) \quad (2)$$

with $\oplus$ denoting concatenation in the prompt context and $\pi(\cdot)$ a task-conditioned profile serializer.

Dense retrieval. Evidence chunks are scored by inner product in a shared embedding space and the top-k set is returned by maximum inner-product search:

$$s(\tilde{q}, d) = \langle f_Q(\tilde{q}) , f_D(d) \rangle , \quad Z_k = \text{argtop-k}_{d\in\mathcal{K}}\, s(\tilde{q}, d) \quad (3)$$

Retrieval distribution. Scores are normalized into a distribution over the retrieved set, with temperature τ controlling how sharply mass concentrates on the highest-ranked evidence (Fig. 1):

$$p_\eta(d \mid \tilde{q}) = \exp(s(\tilde{q},d) / \tau) / \Sigma_{d'\in Z_k} \exp(s(\tilde{q},d') / \tau) \quad (4)$$

Grounded generation. The response is produced by marginalizing the token-level generator over retrieved evidence, which is what makes each assertion traceable to a displayed source:

$$p(y \mid \tilde{q}) = \Sigma_{d\in Z_k} p_\eta(d \mid \tilde{q}) \cdot \Pi_t p_\theta(y_t \mid \tilde{q} , d , y_{<t}) \quad (5)$$

Personalized re-ranking. Topical similarity alone under-weights patient-specific applicability (comorbidity, allergy, contraindication, social determinants). A convex combination with a profile-compatibility term $\rho(d, P_u) \in [0,1]$ restores it (Fig. 2):

$$r(d, \tilde{q}) = \lambda \cdot \bar{s}(\tilde{q},d) + (1 - \lambda) \cdot \rho(d, P_u) , \quad \lambda \in [0,1] \quad (6)$$

Context budget. Retrieval depth is bounded by the generator context window B, giving the feasibility constraint plotted in Fig. 3:

$$\Sigma_{i=1}^{k} |d_i| \le B - |\tilde{q}| - y_{max} \;\Rightarrow\; k_{max} = \lfloor (B - |\tilde{q}| - y_{max}) / \ell \rfloor \quad (7)$$

Groundedness. For a response segmented into atomic assertions $S = \{s_1,\ldots,s_m\}$, the attribution rate quantifies how much of the output is supported by retrieved evidence rather than parametric memory:

$$G(y, Z_k) = (1/m)\ \Sigma_{j=1}^{m}\ 1[\ \exists\ d \in Z_k : d \vDash s_j\ ] \quad (8)$$

Graduated uncertainty. Section 4 proposes that early turns be framed as hypotheses and later turns, informed by patient feedback, as recommendations. A monotone assertion-strength schedule formalizes this (Fig. 4):

$$c_n = 1 - (1 - c_0) \cdot e^{-\gamma n}\ ,\quad n = 0,1,2,\ldots \quad (9)$$

where c_0 is the initial assertion strength, γ the feedback-incorporation rate, and a threshold c^* marks the transition from hypothesis framing to recommendation framing.

Privacy constraint. All personalization is required to operate on a de-identified or synthetically generated view, so that no protected health identifier enters the prompt:

$$I(\ \pi(P_u)\ ;\ PHI_u\ ) \leq \varepsilon \quad (10)$$

with ε the admissible residual-identifiability budget agreed by the governing ethics framework.

5. Conclusion

Generative AI and Retrieval-Augmented Generation enable evidence-based patient-centric Clinical Decision Support through personalized natural-language responses. As a representative implementation for illustrative purposes, the system harnesses the capabilities of a generative language model—ChatGPT—and a suite of disease-agnostic clinical classification models within a unified architecture. To provide pertinent information, a multilevel information store supplies a diverse array of patient-centered data, including structured Electronic Health Records data, unstructured Clinical Notes, privacy-aware Voice of the Patient data, and patient profiles that summarize the patient's medical history while giving due consideration to privacy.

Incorporating Retrieval-Augmented Generation into Clinical Decision Support offers a clear pathway towards personalization by aligning the response generation process with the patient in focus. The system generates patient-centric natural-language responses based on a combination of patient-contextualized information retrieved from different repositories and patient-specific information distilled into a profile through an intermediate personalization framework. This architecture allows for a plethora of patient-centric response types, thus addressing a critical area in Clinical Decision Support Systems, and privacy-preserving mechanisms inherent in the construction of the Voice of the Patient repository further enhance patient privacy.

REFERENCES

[1] Amugongo, L. M., Mascheroni, P., Brooks, S., Doering, S., & Seidel, J. (2025). Retrieval augmented generation for large language models in healthcare: A systematic review. PLOS Digital Health, 4(6), e0000877.

[2] Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., & Akalin, A. (2025). Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. npj Digital Medicine, 8, 263.

[3] Miao, Y., Zhao, Y., Luo, Y., Wang, H., & Wu, Y. (2025). Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: Scoping review. Journal of Medical Internet Research, 27, e80557.

[4] Wiest, I. C., Bhat, M., Clusmann, J., Schneider, C. V., Jiang, X., & Kather, J. N. (2025). Large language models for clinical decision support in gastroenterology and hepatology. Nature Reviews Gastroenterology & Hepatology, 22(11), 773–787.

[5] Yang, R., Ning, Y., Keppo, E., Liu, M., Hong, C., Bitterman, D. S., Ong, J. C. L., Ting, D. S. W., & Liu, N. (2025). Retrieval-augmented generation for generative artificial intelligence in health care. npj Health Systems, 2, 2.

[6] Mangalampalli, B. M. (2024). AI-Enhanced Data Governance: Automating Compliance In Healthcare Analytics Platforms. The Review Of DIABETIC STUDIES OPEN ACCESS.

[7] Delourme, S., Redjdal, A., Bouaud, J., & Seroussi, B. (2024). Leveraging guideline-based clinical decision support systems with large language models: A case study with breast cancer. Methods of Information in Medicine, 63(3–4), 85–96.

[8] Gilbert, S., Kather, J. N., & Hogan, A. (2024). Augmented non-hallucinating large language models as medical information curators. npj Digital Medicine, 7, 100.

[9] Haltaufderheide, J., & Ranisch, R. (2024). The ethics of ChatGPT in medicine and healthcare: A systematic review on large language models (LLMs). npj Digital Medicine, 7, 183.

[10] Kaiser, K. N., Hughes, A. J., Yang, A. D., Turk, A. A., Mohanty, S., Gonzalez, A. A., Patzer, R. E., Bilimoria, K. Y., & Ellis, R. J. (2024). Accuracy and consistency of publicly available large language models as clinical decision support tools for the management of colon cancer. Journal of Surgical Oncology, 130(5), 1104–1110.
[11] Kwong, J. C. C., Wang, S. C. Y., Nickel, G. C., Cacciamani, G. E., & Kvedar, J. C. (2024). The long but necessary road to responsible use of large language models in healthcare research. npj Digital Medicine, 7, 177.
[12] Levin, C., Kagan, T., Rosen, S., & Saban, M. (2024). An evaluation of the capabilities of language models and nurses in providing neonatal clinical decision support. International Journal of Nursing Studies, 155, 104771.
[13] Omar, M., Nadkarni, G. N., Klang, E., & Glicksberg, B. S. (2024). Large language models in medicine: A review of current clinical trials across healthcare applications. PLOS Digital Health, 3(11), e0000662.
[14] Polineni, T. N. S., Kumar, A. S., Maguluri, K. K., Koli, V., Valiki, D., & Ravikanth, S. (2024, November). A Scalable and Robust Framework for Advanced Semi Supervised Learning Supporting Universal Applications. In Proceedings of the 3rd International Conference on Optimization Techniques in the Field of Engineering (ICOFE-2024).
[15] Oniani, D., Wu, X., Visweswaran, S., & Kapoor, S. (2024). Enhancing large language models for clinical decision support by incorporating clinical practice guidelines. Proceedings of the IEEE International Conference on Healthcare Informatics, 2024, 694–702.
[16] Perlis, R. H., Goldberg, J. F., Ostacher, M. J., & Schneck, C. D. (2024). Clinical decision support for bipolar depression using large language models. Neuropsychopharmacology, 49, 1412–1416.
[17] Ranji, S. R. (2024). Large language models—Misdiagnosing diagnostic excellence? JAMA Network Open, 7(10), e2440901.
[18] Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. npj Digital Medicine, 7, 339.
[19] Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyarov, N., Visweswaran, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. npj Digital Medicine, 7, 258.
[20] Kolla, S. K. (2024). Clinical Knowledge Intelligence through Deep Learning and Natural Language Understanding in Healthcare Platforms. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 7(3), 14088.
[21] Wang, D., & Zhang, S. (2024). Large language models in medical and healthcare fields: Applications, advances, and challenges. Artificial Intelligence Review, 57, 299.
[22] Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H., Cool, J. A., Kanjee, Z., Parsons, A. S., Ahuja, N., Horvitz, E., Yang, D., Chen, J. H., & Milstein, A. (2024). Large language model influence on diagnostic reasoning: A randomized clinical trial. JAMA Network Open, 7(10), e2440969.
[23] Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Medicine, 30, 2613–2622.
[24] Hirosawa, T., Harada, Y., Yokose, M., Sakamoto, T., Kawamura, R., & Shimizu, T. (2024). Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nature Communications, 15, 2050.
[25] Liu, J., Wang, C., & Liu, S. (2023). Utility of ChatGPT in clinical practice. Journal of Medical Internet Research, 25, e48568.
[26] Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., & Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health, 2(2), e0000198.
[27] Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. New England Journal of Medicine, 388(13), 1233–1239.

[28] Mbakwe, A. B., Lourentzou, I., Celi, L. A., Mechanic, O. J., & Dagan, A. (2023). ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digital Health, 2(2), e0000205.
[29] Reddy, V. A. R. (2024). Generative Intelligence for Healthcare Claims Processing and Personalized Benefits Management. International Journal of Advanced Engineering Science and Information Technology (IJAESIT), 7(3), 14099.
[30] Rao, A., Pang, M., Kim, J., Kamineni, M., Lie, W., Prasad, A. K., Landman, A., Dreyer, K., & Succi, M. D. (2023). Assessing the utility of ChatGPT throughout the entire clinical workflow: Development and usability study. Journal of Medical Internet Research, 25, e48659.
[31] 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., Karthikesalingam, A., & Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180.
[32] 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, 1930–1940.