RESEARCH PAPER

Synergizing AI: A Cross-Sectoral Framework for Learning Systems in Health and Academia

Dr. Kavita Kasliwa[1*], Sejal Mishra[2] , Praveen Singh[3]

[1*]Associate Professor, Arihant College of Management, Indore, India
Email: kavitakasliwal@gmail.com
[2]Academic Associate, Information System, Indian Institute of Management Indore, India.
Email- Sejalmishra423@gmail.com
[3]Academic Associate, Finance and Accounting, Indian Institute of Management Indore, India,
Email: praveenks30799@gmail.com

Received: 15 May 2026 Revised: 11 July 2026 Accepted: 7 August 2026 Available online : 12 August 2026

Abstract
The rapid emergence of Large Language Models (LLMs) is reshaping how educational and clinical institutions approach teaching, learning, decision support, and knowledge dissemination. This paper systematically reviews literature published between 2020 and 2025 on LLMs in higher education, then extends the analysis through a cross-sectoral framework to healthcare, where similar dynamics of personalisation, automation, and ethical risk are unfolding. In higher education, three dominant themes emerge: enhanced personalised learning, automation of academic support functions, and tensions around academic integrity and algorithmic bias. In healthcare, three parallel themes appear, LLMs as tools for medical education and simulated clinical training, their role in clinical decision support and knowledge retrieval, and their use in patient education and health literacy. Findings indicate that while LLMs show strong promise as supplementary tools across both domains, their integration remains uneven, context-dependent, and reliant on robust human oversight. The paper proposes an integrated, layered framework, presented as a flow diagram, tracing LLM capability from foundational model architecture through sector-specific application to human governance and measurable outcomes in classrooms and clinics. It argues that a policy-driven, cross-sectoral approach is essential to realise LLM benefits without compromising educational values or clinical safety, and discusses implications for administrators, faculty, clinicians, and policymakers.

Keywords: Large Language Models, Higher Education, Healthcare, Artificial Intelligence, Personalised Learning, Clinical Decision Support, Algorithmic Bias, Pedagogical Innovation

How to cite this article: Kasliwa K, Mishra S, Singh P. Synergizing AI: A Cross-Sectoral Framework for Learning Systems in Health and Academia. Int J Drug Deliv Technol. 2026;16(76s): 773-785. DOI: 10.25258/ijddt.16.76s.73

Introduction

The landscape of higher education is undergoing a transformation that, in both pace and scope, is difficult to overstate. Institutions that once relied on structured curricula, fixed textbook content, and traditional examination frameworks are increasingly confronting a technological reality that challenges each of these conventions. At the centre of this disruption sits a class of artificial intelligence systems collectively referred to as Large Language Models (LLMs) computational tools trained on vast corpora of text and capable of generating contextually coherent, syntactically sophisticated responses across a wide range of domains. A structurally similar disruption, arriving on an almost identical timeline, is unfolding in the health sector, where LLMs are being trialled as tutors for medical trainees, as aids to clinical reasoning, and as translators of complex medical information for patients.

The public release of systems such as GPT-3 in 2020 and the subsequent introduction of ChatGPT in late 2022 brought this technology into the daily lives of students, educators, clinicians, and researchers in ways that few had anticipated. Within weeks of ChatGPT's launch, universities across the globe were confronting the dual reality of its potential -as a tutoring assistant, writing aid, and research companion -and its risks, particularly around questions of originality, verification, and intellectual development. Almost simultaneously, medical researchers began reporting that the same general-purpose model could perform credibly on professional licensing examinations, a finding that carried immediate implications for how future clinicians are trained and assessed (Kung et al., 2023). The conversation that followed, in both sectors, was not merely technical but deeply philosophical, touching on what education and clinical practice are for, and what each should protect.

In this context, the Information Systems community has a particularly productive vantage point. The study of how technology interacts with human behaviour, organisational processes, and institutional structures sits precisely at the intersection where LLM adoption in education, and now in healthcare learning systems, raises its most pressing questions. It is from this vantage point

that the present paper proceeds, treating higher education and healthcare not as unrelated case studies but as two instances of a common institutional problem: how a knowledge-intensive profession absorbs a general-purpose reasoning technology without eroding the judgement, integrity, and trust that the profession exists to protect.

The purpose of this to synthesise current scholarly understanding of LLMs in higher education through a structured literature review, and to extend that synthesis into a second, closely related domain -the health sector - through a dedicated review and an integrated conceptual framework. The review is guided by four central research questions. First, in what ways are LLMs being applied within higher education environments, and with what effects on learning outcomes? Second, what barriers and risks accompany LLM integration in academic settings? Third, in what ways do these applications, barriers, and risks recur, transform, or intensify when LLMs are applied to medical education, clinical decision support, and patient education within the health sector? Fourth, what directions should future research and institutional policy pursue in order to generate more rigorous, contextually grounded, and cross-sectorally informed knowledge in this domain .

2. Methodology

This paper adopts a systematic literature review methodology, guided by the PRISMA framework (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). A systematic approach was chosen over a narrative review in order to ensure transparency in source selection, reproducibility of the search process, and rigour in the thematic analysis of findings. Because the paper spans two related but distinct literatures -higher education and healthcare -the search and screening process was conducted in two linked stages, described below.

2.1 Search Strategy

The primary literature search was conducted across three major academic databases: Scopus, Web of Science, and Google Scholar. The following search string was employed for the higher-education corpus: ("Large Language Model" OR "LLM" OR "ChatGPT" OR "GPT-4" OR "Generative AI") AND ("higher education" OR "university" OR "academic" OR "pedagogy" OR "learning"). The search was restricted to peer-reviewed journal articles and conference proceedings published in English between January 2020 and December 2024.

A supplementary search was subsequently conducted to support the health-sector extension of the review, using the string: ("Large Language Model" OR "LLM" OR "ChatGPT" OR "Generative AI") AND ("medical education" OR "clinical decision support" OR "healthcare" OR "patient education" OR "health literacy"). This second search was conducted across PubMed, Scopus, and arXiv, reflecting the disciplinary norms of the health-informatics and clinical literature, where preprint servers play a comparatively larger role in early dissemination. The supplementary search was restricted to English-language records published between January 2022 and mid-2025, a narrower window reflecting the later onset of empirical work in this sub-field relative to the general higher-education literature.

2.2 Inclusion and Exclusion Criteria

Articles were included in the higher-education corpus if they addressed LLM applications in higher education, examined student or faculty perceptions of AI-assisted learning, or analysed policy and ethical dimensions of LLM use in academic contexts. Articles were excluded if they focused solely on K–12 education, lacked empirical or conceptual grounding, or were available only in non-English languages. Book reviews, editorials, and commentary pieces were also excluded. For the health-sector corpus, articles were included if they examined LLM applications in medical or health-professions education, clinical decision support, knowledge retrieval, or patient-facing education and engagement, articles confined to purely technical model-architecture questions with no educational or clinical-practice dimension were excluded, as were single-institution case reports lacking methodological detail.

2.3 Screening and Analysis

For the higher-education corpus, an initial search returned 112 records. After removing duplicates and screening titles and abstracts for relevance, 34 full-text articles were assessed for eligibility. Of these, 31 were retained for inclusion in the final review. For the supplementary health-sector search, an initial search returned 87 records, after de-duplication and title/abstract screening, 34 full-text articles were assessed, of which 22 were retained. The combined corpus underpinning this paper therefore comprises 63 studies. Thematic analysis was conducted using an inductive coding approach, wherein recurring concepts and arguments were grouped into broader themes through iterative reading and refinement. Two rounds of coding were used- an initial open-coding pass generated descriptive labels for each study's principal contribution, followed by an axial-coding pass in which labels were consolidated into the thematic categories reported in Sections 4 and 6.

2.4 Cross-Sectoral Synthesis

Because a central contribution of this paper is the comparison of higher education and health-sector findings, a final synthesis step was added beyond conventional single-domain reviews. Themes generated independently from the two corpora were mapped against one another along three dimensions that recurred in both literatures- (a) the pedagogical or clinical function performed by the LLM (tutor, assistant, or automator), (b) the nature of the principal risk raised (accuracy, bias, or integrity/safety), and (c) the locus of human oversight required to manage that risk (learner, instructor/clinician, or institution). This mapping exercise directly informed the integrated framework presented in Section 5.

2.5 Limitations of the Review Methodology

Several limitations attach to the methodology adopted here and should condition the interpretation of the findings that follow. First, restricting both searches to English-language, peer-reviewed sources almost certainly

excludes relevant practitioner experience documented in institutional reports, non-English journals, and grey literature, particularly from regions under-represented in the corpus, a limitation discussed further in Section 7. Second, the supplementary health-sector search, though methodologically parallel to the primary search, drew on a smaller and more recent literature, since empirical work on LLMs in clinical and medical-education contexts began to appear in meaningful volume only from 2023 onward, conclusions drawn from this sub-corpus should therefore be treated as indicative of an early and rapidly evolving evidence base rather than a mature one. Third, the cross-sectoral synthesis reported in Section 2.4 and Section 6.5 was conducted by mapping themes generated independently within each corpus, an approach that privileges structural comparability over exhaustive coverage of either literature. Readers seeking a fully comprehensive single-domain review of either higher education or clinical LLM applications should consult the source reviews cited throughout Sections 4 and 6 directly.

3. Conceptual Background: Understanding Large Language Models

Large Language Models are a class of deep learning architectures trained using the transformer framework, first introduced by Vaswani et al. (2017). The transformer's central innovation, the self-attention mechanism, allows a model to weigh the relevance of every other token in an input sequence when generating a representation of any given token, dispensing with the sequential, recurrent processing that constrained earlier neural architectures. This parallelisable structure is what has permitted the scaling of these models to hundreds of billions of parameters, trained on datasets encompassing hundreds of billions of tokens drawn from books, websites, academic journals, and other textual sources.

Contemporary LLMs are typically produced in two broad stages. The first, pre-training, exposes the model to a very large, largely unlabelled text corpus and optimises it to predict the next token in a sequence, this stage is responsible for the model's broad linguistic competence and much of its latent world knowledge. The second stage, fine-tuning, adapts the pre-trained model to specific tasks or behavioural norms, frequently through reinforcement learning from human feedback (RLHF), in which human raters score candidate outputs and a reward model is trained to steer the system toward outputs that are judged more helpful, honest, and harmless. In domain-specific applications -including both education and healthcare -a further adaptation step is often layered onto this process, in which the model is exposed to curated academic or clinical corpora, or coupled with retrieval-augmented generation (RAG) systems that ground its outputs in verified reference material at inference time (Vrdoljak et al., 2025).

The resulting systems are capable of performing a range of natural language processing tasks, including text generation, summarisation, translation, question answering, and code completion, with performance that in many benchmarks approaches or exceeds human-level proficiency. What distinguishes contemporary LLMs from earlier generations of AI language systems is not merely their scale but the apparent generalisation of capability across domains. A model trained predominantly on general internet text demonstrates surprising competence in legal reasoning, clinical documentation, mathematical problem-solving, and literary analysis -a phenomenon that has attracted both enthusiasm and scepticism within the research community (Bommasani et al., 2021). For education and health-professions researchers alike, the most relevant characteristic of LLMs may be their conversational fluency. Unlike earlier chatbot systems that relied on rule-based response generation or retrieval-based matching, LLMs produce contextually adaptive, open-ended dialogue that can simulate the kind of exploratory interaction typically associated with human tutoring or, in the clinical context, with a simulated patient encounter. This has led several scholars to situate LLMs within the broader tradition of intelligent tutoring systems (ITS), while simultaneously arguing that they represent a qualitative departure from earlier ITS designs (Kasneci et al., 2023). The same conversational property underpins their use as virtual standardised patients and on-demand clinical knowledge assistants, a parallel developed further in Section 6.

It is important, for the purposes of this paper, to be precise about the parameters that differentiate one LLM deployment from another, since these parameters recur throughout the thematic findings that follow. They include, model scale (the number of parameters and the size of the pre-training corpus, which broadly govern the breadth of latent knowledge available) alignment method (the extent and quality of RLHF or comparable fine-tuning, which governs the reliability and safety of outputs) domain grounding (whether the model operates purely from parametric memory or is coupled to a retrieval system anchored in verified academic or clinical sources) and interface design (whether the model is deployed as an open-ended conversational agent, a constrained tutoring or advising system, or an embedded feature within an existing platform such as a learning management system or electronic health record). Differences along each of these parameters help explain why studies of ostensibly similar LLM applications sometimes report divergent findings regarding accuracy, bias, and user trust.

4. Thematic Findings: Higher Education

4.1 LLMs as Pedagogical Tools: Personalisation and Accessibility

The most extensively documented application of LLMs in higher education concerns their capacity to support personalised learning. Traditional pedagogical models operate under structural constraints -fixed class sizes, standardised curricula, and limited opportunity for individualised feedback -that inevitably compromise the extent to which instruction can be tailored to the specific needs, prior knowledge, or learning pace of individual students. LLMs, by virtue of their ability to generate responsive, contextually nuanced explanations on demand, represent a potential intervention in precisely this domain. Several studies have documented the ways in which students deploy LLMs as on-demand tutors. Malinka et al. (2023) observed that students using ChatGPT for conceptual clarification in computer science

courses reported higher levels of self-efficacy and reduced anxiety around asking questions, particularly among those who had previously hesitated to seek help from instructors during lectures. The asynchronous, non-judgemental quality of LLM interaction appears to lower barriers that social dynamics within the classroom may otherwise erect. This dynamic recurs, in a heightened form, in clinical training contexts, where junior trainees frequently report reluctance to expose gaps in knowledge in front of supervising clinicians.

The accessibility dimension is particularly salient for students with disabilities or those studying in non-native languages. Tlili et al. (2023) found that LLMs demonstrated considerable utility as reading comprehension supports for students with dyslexia, and as grammar and vocabulary aids for international students operating in English-medium instruction environments. These findings suggest that LLMs may function as equity tools -democratising access to high-quality explanatory support that would otherwise require expensive human intervention. The parameter that appears to govern the strength of this effect most directly is the granularity of the model's output register: models that can flexibly shift reading level, sentence complexity, and terminological density in response to a learner's stated needs produce measurably larger accessibility gains than models used in a single, fixed register. At the same time, scholars have raised important questions about the depth of learning that LLM-mediated instruction promotes. Dalalah (2023) draw a distinction between performance and understanding -noting that students who rely heavily on LLM-generated explanations may produce correct outputs without developing the underlying conceptual frameworks that education is intended to build. This concern echoes a long-standing debate in educational psychology regarding the difference between surface and deep learning approaches, and it resurfaces, largely unchanged in structure, in the medical-education literature reviewed in Section 6.1, where the concern is reframed as a distinction between passing an examination and possessing durable clinical reasoning.

4.2 Automation of Academic Support Functions

Beyond direct pedagogical interaction, LLMs are being deployed to automate a range of administrative and support functions within higher education institutions. These include automated essay feedback, plagiarism detection supplementation, library research assistance, student advising through conversational agents, and the generation of course materials such as syllabi, question banks, and instructional summaries. Each of these functions can be understood as substituting LLM-generated output for a task that would otherwise consume scarce instructor or administrator time, and each therefore raises a parallel question about the acceptable trade-off between efficiency gains and the loss of a human evaluative or advisory presence.The automation of feedback provision has attracted particular scholarly attention. Traditional essay grading is among the most time-intensive activities in higher education, and the quality of feedback students receive is often constrained by the practical limits of instructor workload. Researchers including Kuhail et al. (2023) have explored the extent to which LLM-generated feedback on student writing compares with that provided by human evaluators, finding that on dimensions such as coherence assessment and grammatical correction, LLM feedback was rated by students as broadly comparable in utility, though lacking the contextual sensitivity that experienced instructors bring to evaluative commentary. This gap in contextual sensitivity is a recurring parameter across the literature reviewed in this paper. LLMs perform well on tasks with clearly specified, rule-governed evaluation criteria, and less reliably on tasks requiring judgement calibrated to an individual student, or an individual patient's, particular circumstance .The deployment of LLM-powered chatbots for student advising represents another active area of experimentation. Institutions in the United States, United Kingdom, and Australia have piloted systems capable of addressing queries related to course registration, academic policies, and mental health referrals. While these systems reduce the administrative burden on human advisors, their limitations in handling ambiguous, emotionally sensitive, or contextually complex queries have been consistently noted as constraints on their applicability (Okonkwo & Ade-Ibijola, 2021). The same limitation, in a higher-stakes form, governs the applicability of LLM-based systems to patient-facing advisory roles in healthcare, discussed in Section 6.3.

4.3 Academic Integrity and Ethical Challenges

No theme in the LLM–higher education literature has attracted more sustained attention than that of academic integrity. The capacity of LLMs to generate plausible, stylistically sophisticated academic writing on demand has fundamentally altered the conditions under which student assessment operates. Essays, reports, research summaries, and even code submissions that would previously have required hours of student effort can now be produced in seconds, raising questions about how institutions can meaningfully assess learning when the instrumental demands of assessment can be so easily outsourced. The institutional response to this challenge has been varied and, in many cases, inconsistent. Some universities have moved to ban LLM use in assessed work entirely, while others have adopted disclosure policies requiring students to acknowledge any AI assistance received. A third position, advocated by researchers such as Lodge et al. (2023), argues that the more productive response is to redesign assessment tasks in ways that are more resistant to LLM substitution-prioritising oral examinations, process-based portfolios, and contextually situated problem-solving exercises that require the application of knowledge in conditions LLMs cannot easily replicate. This assessment-redesign argument anticipates, almost precisely, the argument later made in the medical-education literature that examinations easily passed by an LLM (Section 6.1) call the underlying assessment construct into question rather than merely the honesty of the candidate.

The question of bias embedded within LLMs presents a further ethical dimension, one that recurs with amplified stakes in the health sector. These systems are trained on datasets that reflect existing inequalities in whose

knowledge, whose language, and whose perspectives have historically been documented and valued. Research by Bender et al. (2021) demonstrated that LLMs trained on English-dominant internet corpora systematically disadvantage users whose primary linguistic and cultural contexts are not well-represented in that data. For institutions committed to diversity and inclusion, the uncritical adoption of LLMs risks encoding and amplifying precisely the disparities that higher education has a responsibility to address. As Section 6.4 discusses, the same underlying mechanism -biased training data producing biased outputs -carries materially higher risk when the output in question informs a clinical judgement rather than a graded assignment.

4.4 Faculty Perceptions and Institutional Readiness

A small but growing body of research has examined how faculty members perceive LLMs and what conditions facilitate or hinder their productive integration into pedagogical practice. The findings are mixed. Lim et al. (2023) conducted a survey of 117 academics across multiple disciplines and found that while awareness of LLMs was nearly universal, confidence in using them purposefully in teaching was considerably lower, particularly among faculty in humanities and social science disciplines who reported feeling insufficiently prepared to Evaluate LLM outputs within their own areas of expertise.Institutional readiness emerges as a critical mediating variable. Universities with established digital education infrastructure, dedicated academic technology support units, and clear institutional policies on AI use appear better positioned to support faculty in navigating the pedagogical opportunities and risks that LLMs present. Where such infrastructure is absent, individual faculty members are left to develop their own, often ad hoc, responses -a pattern that produces inconsistency and inequity in the student experience. This institutional-readiness parameter reappears as a decisive variable in the health-sector literature, where McCoy et al. (2024) similarly find that the presence or absence of a structured institutional curriculum, rather than individual clinician enthusiasm, is what determines whether LLM competence is developed systematically across a training programme.

5. An Integrated Cross-Sectoral Framework for Responsible LLM Adoption

The thematic findings summarised in Section 4, and the parallel findings developed for the health sector in Section 6, share a common underlying structure that is not always made explicit in single-domain reviews. To render that structure visible, and to provide a practical planning tool for administrators and clinician-educators alike, this paper proposes an integrated, four-layer framework, presented in Figure 1. The framework is deliberately generic at its foundation and outer layers, and deliberately bifurcated at its application layer, in order to represent both what higher education and healthcare share and where their pathways diverge.

The first layer, the Data and Model Foundation layer, captures the pre-training, transformer-based architecture, and RLHF fine-tuning processes described in Section 3, together with the domain-adaptation step -exposure to curated academic corpora on one branch, curated clinical and biomedical corpora on the other -that precedes deployment in either sector. This layer is where the parameters of model scale, alignment quality, and domain grounding identified in Section 3 are set, and it is therefore the layer at which upstream interventions (better curation of training data, more rigorous RLHF, tighter retrieval grounding) can most efficiently reduce downstream risk in both sectors simultaneously.

The second layer, the Application layer, is where the two tracks diverge. On the higher-education branch, the model is deployed for personalised tutoring and conceptual support, academic-support automation, and assessment and accessibility-related functions, corresponding to the themes developed in Sections 4.1 and 4.2. On the health-sector branch, the structurally equivalent functions are medical education and simulated-patient interaction, clinical decision support and knowledge retrieval, and patient education and health-literacy support, corresponding to the themes developed in Sections 6.1 through 6.3. The framework deliberately places these two branches in visual parallel to emphasise that they are functionally analogous responses to a shared underlying capability, deployed in institutions with different risk tolerances and different professional norms.The third layer, Human-in-the-Loop Governance, is shared across both branches and is where the ethical themes developed in Sections 4.3, 4.4, and 6.4 converge structurally. It comprises faculty or clinician oversight (the exercise of professional verification and judgement over model output), institutional policy and training (the acceptable-use rules, disclosure requirements, and digital- or AI-literacy training discussed throughout this paper), and a dedicated ethical review function addressing bias mitigation, data privacy and confidentiality, academic or clinical integrity, and equity of access. The framework positions this layer as a genuine bottleneck rather than an afterthought: outputs from the application layer are represented as passing through, rather than around, governance before they are permitted to influence a documented outcome.

The fourth layer, Outcomes, records the ultimate object of concern in each sector -learning outcomes such as competence, self-efficacy, and equitable access on the education branch, and patient safety and health outcomes such as diagnostic accuracy, health literacy, and patient trust on the health branch. A feedback loop, shown on the left margin of Figure 1, returns information from the outcome layer to the foundation layer, representing the continuous-monitoring obligation -addressing phenomena such as dataset shift, the gradual drift in model performance as the underlying clinical or academic reality changes -that both sectors literatures identify as a precondition for the safe, sustained use of LLMs (McCoy et al., 2024 Vrdoljak et al., 2025).

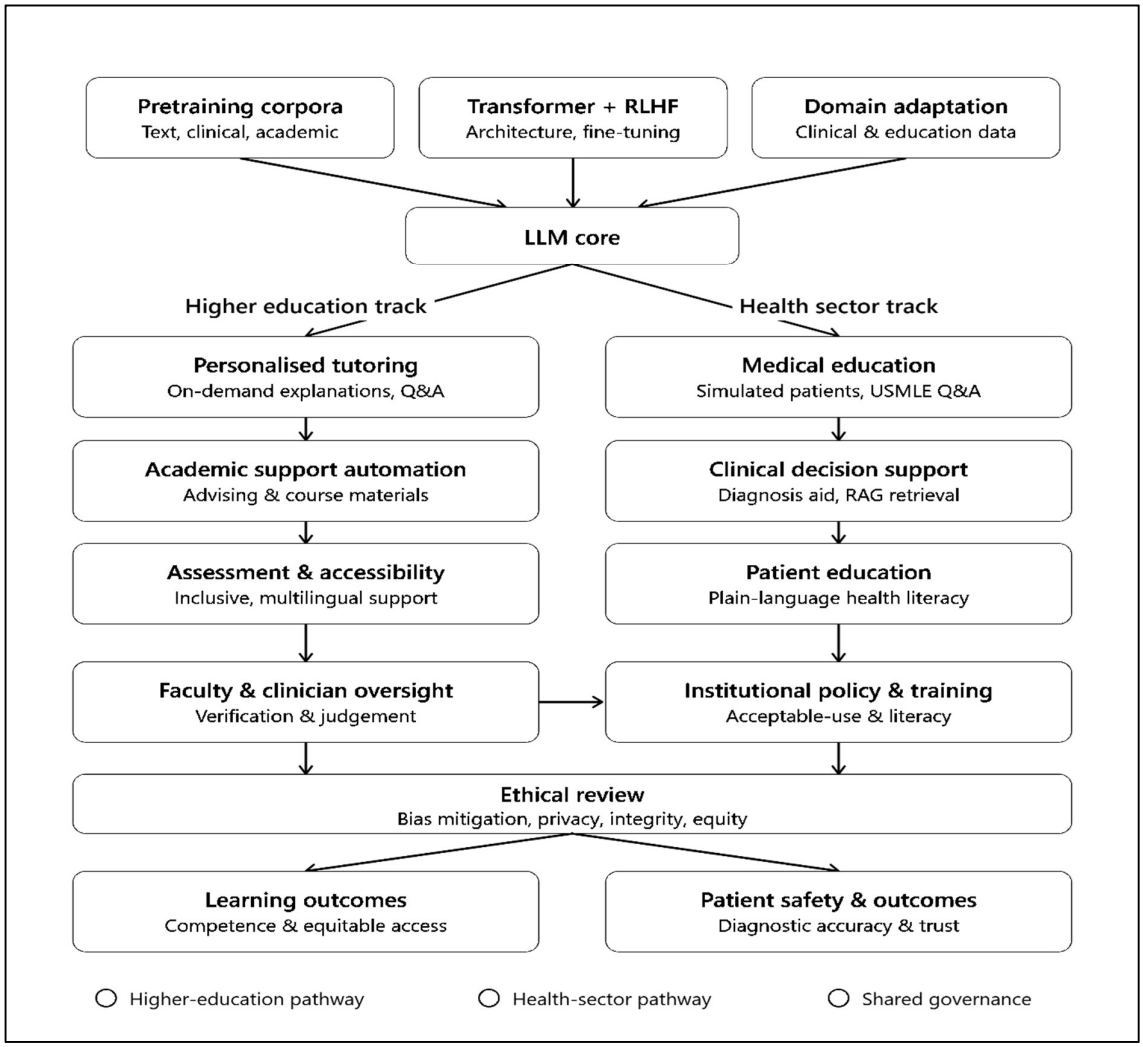

Figure 1. An integrated, four-layer framework for responsible LLM adoption across higher education and health-sector learning systems. The blue pathway traces the higher-education application branch, the teal pathway traces the health-sector application branch, the gold layer represents the shared human-in-the-loop governance function through which both branches must pass before contributing to documented outcomes.

6. Extending the Review to the Health Sector

Having established the dominant themes in higher education and the framework that structures their comparison with an adjacent domain, this section applies the same systematic-review logic to the health sector, drawing on the 22 studies retained from the supplementary search described in Section 2.1. The health-sector literature is organised around three themes that map directly onto the three higher-education themes developed in Section 4, medical education and simulated clinical training (paralleling personalisation and accessibility), clinical decision support and knowledge retrieval (paralleling automation of academic support), and patient education and health literacy (a hybrid theme with no exact higher-education analogue, discussed on its own terms). A fourth theme, ethical, safety, and governance considerations, closes the section and connects directly to the governance layer of the framework in Section 5.

6.1 LLMs in Medical Education and Simulated Clinical Training

The single most widely cited finding in this sub-literature concerns the performance of general-purpose LLMs on professional licensing examinations. Kung et al. (2023) evaluated ChatGPT on the United States Medical Licensing Examination (USMLE), comprising Step 1, Step 2CK, and Step 3, and found that the model performed at or near the passing threshold on all three examinations without any examination-specific fine-tuning, and that its

explanatory outputs demonstrated a notable degree of internal concordance and reasoning insight. This finding, published within months of ChatGPT public release, functioned for the health sector in much the same way that early reports of fluent essay generation functioned for higher education, it converted an abstract technological possibility into an immediate, discipline-specific institutional concern. Subsequent systematic review work has consolidated and extended this early finding. Lucas, Upperman, and Robinson (2024), in a systematic review published in Medical Education, document a rapidly expanding set of applications in which LLMs support medical curriculum development, the generation of teaching materials and question banks, and personalised study planning for individual learners, while cautioning that the same properties that make LLMs useful for generating plausible teaching content also make them liable to generate plausible but incorrect clinical content, a risk with a materially higher cost than the analogous risk in a general higher-education essay. Vrdoljak et al. (2025), in a broader review spanning medical education, clinical decision support, and healthcare administration, report that LLMs show particular promise as virtual or simulated patients-conversational agents that role-play a clinical presentation so that trainees can practise history-taking and clinical reasoning in a low-stakes, infinitely repeatable setting and as personalised tutors capable of adapting explanatory depth to a trainee's level of seniority.

The parameter that most strongly differentiates successful from unsuccessful deployments in this sub-literature is the degree of domain grounding described in Section 3, models coupled to retrieval-augmented generation over curated, current clinical guidelines consistently outperform models relying purely on parametric memory, particularly on questions concerning treatment protocols that have changed since the underlying training corpus was assembled (Vrdoljak et al., 2025). This is the direct clinical analogue of the accessibility-register parameter identified in Section 4.1, in both sectors, the configuration of the model, rather than its raw scale, is what principally determines its educational value. As in the higher-education literature reviewed in Section 4.1, scholars in the medical-education literature draw a sharp distinction between examination performance and durable competence. McCoy et al. (2024) argue that the relative ease with which LLMs pass the USMLE calls into question the structure of an examination process that still allocates significant study time to the rote memorisation of factual material now readily accessible at the point of care, and they advocate a broader shift in medical-education emphasis from knowledge recall toward the skills of evaluation, verification, and judgement that will be required of clinicians working alongside these systems. This argument is, in essence, the health-sector counterpart of the assessment-redesign position articulated by Lodge et al. (2023) for higher education in Section 4.3, and its recurrence across two independently developed literatures lends it considerable weight.

6.2 Clinical Decision Support and Knowledge Retrieval

Beyond the educational setting narrowly defined, LLMs are increasingly positioned as tools for clinical decision support -assisting practising clinicians, as well as trainees, with diagnostic reasoning, treatment-option retrieval, and the rapid synthesis of a fast-growing medical literature. Vrdoljak et al. (2025) report that LLMs exhibit genuine potential in diagnostic assistance, treatment recommendation, and medical knowledge retrieval, but that performance varies considerably across clinical specialties and task types, and that techniques such as retrieval-augmented generation are increasingly used specifically to mitigate the risk of outdated or fabricated clinical content.

This function is the closest health-sector analogue to the academic-support automation theme developed in Section 4.2 for higher education, and it exhibits the same underlying trade-off, LLM-based decision support can substantially reduce the time clinicians spend on literature search and documentation-related tasks, freeing capacity for direct patient interaction, but the same efficiency gain creates institutional pressure to increase clinical throughput rather than to reinvest the recovered time in care quality, a dynamic McCoy et al. (2024) explicitly caution against. The parallel with student-advising chatbots in Section 4.2 is direct: efficiency gains realised through automation are not self-evidently beneficial unless an institution makes a deliberate choice about how the recovered capacity is redeployed. A further parallel concerns the boundary of reliable applicability. Just as LLM-powered advising chatbots in higher education struggle with ambiguous, emotionally sensitive, or contextually complex student queries (Okonkwo & Ade-Ibijola, 2021), clinical decision-support systems built on LLMs perform comparatively well on well-specified, guideline-governed questions and comparatively poorly on complex, multi-morbid presentations that require the integration of idiosyncratic patient history, social context, and values -precisely the domain in which McCoy et al. (2024) argue the clinician's evaluative and empathic role will remain indispensable rather than diminish.

6.3 Patient Education and Health Literacy

A theme with no direct counterpart in the higher-education corpus, but one of clear relevance to an Information Systems audience, concerns the use of LLMs to support patient-facing education and engagement. Aydin, Karabacak, Vlachos, and Margetis (2024), in a scoping review of applications in medicine, find that LLMs are increasingly used to translate complex clinical information into plain-language explanations for patients, to answer patient questions between clinical visits, and to support health-literacy interventions for populations that have historically had limited access to individualised explanatory support -a direct echo of the equity-tool argument made for LLM-assisted accessibility in higher education (Section 4.1, Tlili et al., 2023). The promise of this application is tempered by risks that are specific to the patient-facing context. Where a student who receives a subtly inaccurate LLM explanation of a concept faces, at worst, a delayed or distorted understanding correctable at the next assessment, a patient who receives a subtly inaccurate LLM explanation of a diagnosis or treatment

option may act on that information directly, with immediate consequences for health behaviour and, in some cases, safety. This asymmetry of consequence is the single most important reason the health-sector application of LLMs cannot simply adopt the governance norms developed for higher education wholesale, and it is the primary justification, developed further in Section 6.4, for treating human-in-the-loop verification as a non-negotiable rather than an optional layer of the framework proposed in Section 5.

6.4 Ethical, Safety, and Governance Considerations

The ethical themes identified in the higher-education literature -bias, integrity, and institutional readiness (Sections 4.3 and 4.4) -recur in the health-sector literature in an intensified form. On the question of bias, the mechanism identified by Bender et al. (2021) for language generally applies with particular force in clinical contexts, where training corpora that under-represent certain populations, languages, or presentations of disease can translate directly into diagnostic or explanatory outputs of uneven quality across patient groups, a concern central to the review by Vrdoljak et al. (2025) and to the narrative review by McCoy et al. (2024). On the question of integrity and reliability, the health-sector literature substitutes the assessment-integrity concern of Section 4.3 with a safety-integrity concern, the risk that a clinician or trainee accepts a fluent but incorrect LLM output without independent verification. McCoy et al. (2024) frame this as a durable requirement for professional scepticism, arguing that clinicians must be trained explicitly to treat LLM output as an input to their own judgement rather than a substitute for it, and must retain accountability for any decision made with LLM assistance a governance principle that corresponds closely to the human-oversight node of Layer 3 in the framework proposed in Section 5.

On the question of institutional readiness, McCoy et al. (2024) develop an argument with a structure almost identical to that of Lim et al. (2023) in the higher-education literature (Section 4.4), they find that medical-education institutions currently offer highly inconsistent, and in many cases negligible, formal instruction on LLM-specific competencies, and they call for the deliberate development of a shared competency framework - addressing what should be taught, who should teach it, and how it should be taught-rather than leaving individual clinician-educators to develop ad hoc responses. Lucas et al. (2024) add a further, health-sector-specific governance dimension absent from the higher-education literature, the confidentiality obligations attached to patient data mean that LLM use in clinical education and practice must additionally satisfy data-protection regimes with no direct counterpart in a university classroom, reinforcing the case for the dedicated ethical-review node depicted in Layer 3 of Figure 1. Taken together, these findings support the central claim of this paper framework, that higher education and healthcare are not merely analogous but structurally coupled in their exposure to LLM-related risk, and that governance mechanisms designed for one sector are frequently transferable, with appropriate strengthening, to the other.

6.5 Comparative Synthesis

Table 1 summarises the cross-sectoral synthesis conducted according to the method described in Section 2.4, mapping each higher-education theme onto its health-sector counterpart along the three dimensions of pedagogical/clinical function, principal risk, and locus of oversight. The table is intended as a practical reference for institutional planners seeking to anticipate, in one sector, risks and mitigations that have already been documented in the other.

Table 1. Cross-sectoral synthesis mapping higher-education themes to health-sector counterparts.

Dimension	Higher Education	Health Sector
Personalisation function	On-demand tutoring, conceptual clarification, accessibility support (Malinka et al., 2023, Tlili et al., 2023)	Simulated patients, personalised medical tutoring, adaptive explanation by seniority (Kung et al., 2023, Vrdoljak et al., 2025)
Automation function	Essay feedback, advising chatbots, course-material generation (Kuhail et al., 2023, Okonkwo & Ade-Ibijola, 2021)	Clinical decision support, medical knowledge retrieval, documentation assistance (Vrdoljak et al., 2025, McCoy et al., 2024)
Principal accuracy risk	Surface-level correctness without deep conceptual understanding (Dalalah & Dalalah, 2023)	Plausible but clinically incorrect output; outdated treatment information without retrieval grounding (Vrdoljak et al., 2025)
Principal integrity/safety risk	Assessment integrity; outsourcing of academic effort (Lodge et al., 2023)	Uncritical clinical acceptance of LLM output, patient-safety consequences (McCoy et al., 2024)
Bias concern	Under-representation of non-English languages and cultural contexts (Bender et al., 2021)	Under-representation of patient populations in training data; differential diagnostic quality (Vrdoljak et al., 2025)
Institutional readiness gap	Inconsistent faculty confidence and infrastructure across disciplines (Lim et al., 2023)	Inconsistent, largely informal LLM-specific curricula across institutions (McCoy et al., 2024, Lucas et al., 2024)
Recommended durable response	Redesign of assessment toward process and applied judgement (Lodge et al., 2023)	Cultivation of clinician verification habits; domain-grounded, retrieval-augmented deployment (McCoy et al., 2024)

7. Discussion

The literature reviewed in this paper does not yield a simple verdict on LLMs in either higher education or healthcare. What emerges instead is a picture of genuine but unevenly distributed promise, set against a backdrop of substantive and, in the health sector particularly, safety-critical unresolved challenges. Several implications warrant attention. First, the case for LLMs as accessibility and equity tools is compelling in both sectors but must be pursued deliberately. The benefits documented in the higher-education literature -reduced anxiety, improved access for non-native speakers, on-demand support and their health-sector counterpart, improved patient health literacy and access to explanatory support, are most likely to materialise in institutional contexts that have thought carefully about which students or patients are being served and which may be disadvantaged by LLM integration. A technology that improves outcomes for already-advantaged students or patients while leaving others behind does not constitute progress in either domain.

Second, the integrity and safety challenges in each sector are unlikely to be resolved through detection or restriction alone. In higher education, the arms race between LLM-generated content and detection tools is one that the former will, over time, win, the more durable response lies in assessment redesign, a pedagogically motivated shift toward evaluating not merely the products of learning but the processes through which understanding develops (Lodge et al., 2023). In healthcare, the equivalent durable response is not the prohibition of clinical LLM use but the cultivation, described by McCoy et al. (2024), of professional scepticism and verification habits among clinicians, paired with domain-grounded, retrieval-augmented model configurations that reduce the base rate of fabricated content in the first place.

Third, the institutional dimension cannot be neglected in either sector. Individual faculty enthusiasm or scepticism, and individual clinician enthusiasm or scepticism, will shape local practice, but it is institutional policy-governing acceptable use, training provision, and infrastructure investment that will determine whether LLM integration is systematic and equitable or fragmented and inequitable. The evidence from both literatures suggests that most institutions, whether universities or health systems, are still in early, exploratory phases of this policy development, a finding that gives the integrated framework proposed in Section 5 practical rather than merely descriptive value, it offers a common vocabulary and a common set of checkpoints that institutions in either sector can use to audit their own state of readiness.

Fourth, the global and equity dimension of this challenge deserves greater attention in both literatures. The majority of studies reviewed in this paper, in both the higher-education and health-sector corpora, draw on institutional contexts in the Global North, principally the United States, United Kingdom, and continental Europe. Research from higher-education and health-system settings in South Asia, Sub-Saharan Africa, and Latin America remains comparatively sparse, despite the fact that these contexts often face distinctive resource constraints, linguistic diversity, and pedagogical or clinical traditions that may interact with LLM capabilities in materially different ways.

Finally, the comparison undertaken in this paper suggests a more general methodological point for future research: because higher education and healthcare are exposed to the same underlying technology through structurally similar institutional mechanisms-personalised instruction, automated support, and integrity or safety risk, all mediated by institutional readiness-single-sector reviews risk reinventing governance solutions that have already been developed, tested, and critiqued in an adjacent domain. Explicit cross-sectoral synthesis, of the kind attempted in Section 6.5, may therefore accelerate the maturation of institutional policy in both fields. A further implication concerns the pace at which institutional policy must be revised relative to the pace of underlying model change. Both corpora reviewed in this paper describe a technology whose capability profile shifted materially within the two-to-three-year window each review covers, a rate of change that outstrips the typical cycle of curriculum revision or clinical-guideline updating in either sector. This mismatch suggests that static, one-time policy statements -a single approved-use list, a single disclosure clause in an academic-integrity code -are unlikely to remain adequate for long, and that the continuous-monitoring feedback loop depicted in Figure 1 should be read as a structural requirement rather than a best-practice aspiration. Institutions that treat LLM governance as a standing agenda item, reviewed on a fixed cycle by the ethical-review function proposed in Section 5, are better positioned to keep policy aligned with capability than institutions that treat governance as a task to be completed once and then set aside.

8. Policy and Institutional Recommendations

Drawing on the framework developed in Section 5 and the comparative findings of Section 7, this paper offers four recommendations addressed to administrators, faculty, and clinician-educators considering LLM integration. First, institutions should treat the four layers of the framework in Figure 1 as a sequential audit checklist rather than a one-time adoption decision: before deploying an LLM-based tool, an institution should be able to specify which foundation model and domain-adaptation approach is in use, which application-layer function the tool performs, which specific human role exercises oversight at the governance layer, and which outcome metric will be used to evaluate the deployment.

Second, institutions in both sectors should invest in domain grounding-retrieval-augmented configurations tied to current, curated academic or clinical sources in preference to relying on unadapted, general-purpose models, since the evidence reviewed in Sections 4 and 6 consistently identifies domain grounding as the parameter most strongly associated with reliable useful output.

Third, assessment and evaluation practices, whether of students or of clinical trainees, should be redesigned around processes and applied judgement rather than solely around outputs that an LLM can readily reproduce,

following the logic articulated by Lodge et al. (2023) and echoed by McCoy et al. (2024).

Fourth, institutions should establish or strengthen a dedicated ethical-review function, corresponding to Layer 3 of the proposed framework, with explicit responsibility for bias auditing, data-privacy compliance, and the ongoing monitoring of model performance over time, since both literatures identify the absence of such a function, rather than the absence of enthusiasm for the technology, as the more consistent predictor of poor institutional outcomes.

9. Conclusion and Future Research Directions

Future research should pursue several directions. Longitudinal studies tracking the effects of LLM use on learning outcomes, and on clinical competence and patient outcomes, over extended periods are conspicuously absent from the current literature in both sectors and represent a priority. Research examining LLM integration in non-Western educational and health-system contexts would significantly broaden the evidentiary base from which policy recommendations are drawn, an opportunity of particular relevance to institutions .There is also a need for controlled experimental designs that isolate the specific mechanisms through which LLM interaction affects cognitive engagement and clinical reasoning, rather than relying predominantly on self-reported perceptions. Finally, the development of pedagogically and clinically grounded frameworks for LLM integration -frameworks that begin from educational and patient-safety goals rather than technological capability -represents perhaps the most urgent need identified in this review. Technology has a consistent history of generating solutions in search of problems, in both higher education and healthcare, the challenge is to ensure that the adoption of LLMs is driven by a clear and evidence-informed understanding of the problems that each domain must solve, and by governance structures, of the kind proposed in Section 5, robust enough to be shared across both.

Bibliographic References:

1. Alkaissi, H., & McFarlane, S. I. (2023). Artificial hallucinations in ChatGPT: Implications in scientific writing. Cureus, 15(2), e35179. https://doi.org/10.7759/cureus.35179
2. Aydin, S., Karabacak, M., Vlachos, V., & Margetis, K. (2024). Large language models in patient education: A scoping review of applications in medicine. Frontiers in Medicine, 11, 1477898. https://doi.org/10.3389/fmed.2024.1477898
3. Behers, B. J., Stephenson-Moe, C. A., Gibons, R. M., Vargas, I. A., Wojtas, C. N., Rosario, M. A., Anneaud, D., Nord, P., Hamad, K. M., & Baker, J. F. (2024). Assessing the quality of patient education materials on cardiac catheterization from artificial intelligence chatbots. Cureus, 16(9), e69996. https://doi.org/10.7759/cureus.69996
4. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM.
5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
6. Chinta, S. V., Wang, Z., Zhang, X., Doan, T. V., Kashif, A., Smith, M. A., & Zhang, W. (2024). AI-driven healthcare: A survey on ensuring fairness and mitigating bias. arXiv preprint arXiv:2407.19655.
7. Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228–239. https://doi.org/10.1080/14703297.2023.2190148
8. Dalalah, D., & Dalalah, O. M. A. (2023). The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT. The International Journal of Management Education, 21(2), 100822.
9. Eriksen, A. V., Möller, S., & Ryg, J. (2023). Use of GPT-4 to diagnose complex clinical cases. NEJM AI, 1(1), AIp2300031.
10. Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 23(120), 1–39.
11. Fink, A., Nattenmüller, J., Rau, S., Rau, A., Tran, H., Bamberg, F., Reisert, M., Kotter, E., Diallo, T., & Russe, M. F. (2025). Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification. European Radiology. https://doi.org/10.1007/s00330-025-11445-z
12. Goto, T., & Tsugawa, Y. (2023). The accuracy and potential racial and ethnic biases of GPT-4 in the diagnosis and triage of health conditions: Evaluation study. JMIR Medical Education, 9, e48003.
13. Han, T., Adams, L. C., Bressem, K. K., Busch, F., Nebelung, S., & Truhn, D. (2024). Comparative analysis of multimodal large language model performance on clinical vignette questions. JAMA, 331(15), 1320–1321. https://doi.org/10.1001/jama.2023.27861
14. Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. EBioMedicine, 90, 104512.
15. Hirosawa, T., Kawamura, R., Harada, Y., Mizuta, K., Tokumasu, K., Kaji, Y., Suzuki, T., & Shimizu, T. (2023). ChatGPT-generated differential diagnosis lists for complex case–

derived clinical vignettes: Diagnostic accuracy evaluation. JMIR Medical Informatics, 11, e48808. https://doi.org/10.2196/48808
16. Hirosawa, T., Mizuta, K., Harada, Y., & Shimizu, T. (2023). Comparative evaluation of diagnostic accuracy between Google Bard and physicians. American Journal of Medicine, 136(11), 1119–1123. https://doi.org/10.1016/j.amjmed.2023.08.003
17. Jones, C., Thornton, J., & Wyatt, J. C. (2023). Artificial intelligence and clinical decision support: Clinicians' perspectives on trust, trustworthiness, and liability. Medical Law Review, 31(4), 501–520. https://doi.org/10.1093/medlaw/fwad013
18. Kanjee, Z., Crowe, B., & Rodman, A. (2023). Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA, 330(1), 78–80. https://doi.org/10.1001/jama.2023.8288
19. Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274. https://doi.org/10.1016/j.lindif.2023.102274
20. Kelly, A., Noctor, E., Ryan, L., & van de Ven, P. (2025). The effectiveness of a custom AI chatbot for type 2 diabetes mellitus health literacy: Development and evaluation study. Journal of Medical Internet Research, 27, e70131. https://doi.org/10.2196/70131
21. Kohandel Gargari, O., & Habibi, G. (2025). Enhancing medical AI with retrieval-augmented generation: A mini narrative review. Digital Health, 11. https://doi.org/10.1177/20552076251337177
22. Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2023). Interacting with educational chatbots: A systematic review. Education and Information Technologies, 28(2), 973–1018.
23. 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. https://doi.org/10.1371/journal.pdig.0000198
24. Laverde, N., Grévisse, C., Jaramillo, S., & Manrique, R. (2025). Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training. Computational and Structural Biotechnology Journal, 27, 2481–2491.
25. Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. Computers and Education: Artificial Intelligence, 6, 100213.
26. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.
27. Li, Y., Zeng, C., Zhong, J., Zhang, R., Zhang, M., & Zou, L. (2024). Leveraging large language model as simulated patients for clinical education. arXiv preprint arXiv:2404.13066.
28. Liang, Y., Zou, D., Xie, H., & Wang, F. L. (2023). Exploring the potential of using ChatGPT in physics education. Smart Learning Environments, 10(1), 52.
29. Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The International Journal of Management Education, 21(2), 100790.
30. Lodge, J. M., Howard, S., Bearman, M., & Dawson, P. (2023). Assessment reform in the age of artificial intelligence: A call for research. Assessment & Evaluation in Higher Education, 48(7), 943–953.
31. Lucas, H. C., Upperman, J. S., & Robinson, J. R. (2024). A systematic review of large language models and their implications in medical education. Medical Education, 58(11), 1276–1285. https://doi.org/10.1111/medu.15402
32. Maity, S., & Deroy, A. (2024). Generative AI and its impact on personalized intelligent tutoring systems. arXiv preprint arXiv:2410.10650.
33. Malinka, K., Peresini, M., Firc, A., Hujnak, O., & Jurnecka, F. (2023). On the educational impact of ChatGPT: Is artificial intelligence ready to obtain a university degree? In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education (pp. 47–53). ACM.
34. McCoy, L. G., Ng, F. Y. C., Sauer, C. M., Legaspi, K. E. Y., Jain, B., Gallifant, J., McClurkin, M., Hammond, A., Goode, D., Gichoya, J., & Celi, L. A. (2024). Understanding and training for the impact of large language models and artificial intelligence in healthcare practice: A narrative review. BMC Medical Education, 24, 1096. https://doi.org/10.1186/s12909-024-06048-z
35. McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., et al. (2023). Towards accurate differential diagnosis with large language models. arXiv preprint arXiv:2312.00164.
36. Nazi, Z. A., & Peng, W. (2024). Large language models in healthcare and medical domain: A review. Informatics, 11(3), 57.

37. Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
38. Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. Computers and Education: Artificial Intelligence, 2, 100033.
39. Omar, M., et al. (2025). Large language models are highly vulnerable to adversarial hallucination attacks in clinical decision support: A multi-model assurance analysis. medRxiv. https://doi.org/10.1101/2025.03.18.25324184
40. Omiye, J. A., Lester, J. C., Spichak, S., Rotemberg, V., & Daneshjou, R. (2023). Large language models propagate race-based medicine. NPJ Digital Medicine, 6, 195.
41. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.
42. Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., Tomasev, N., Rashid, Q. M., Azizi, S., Rostamzadeh, N., et al. (2024). A toolbox for surfacing health equity harms and biases in large language models. Nature Medicine, 30(12), 3590–3600.
43. Preiksaitis, C., Ashenburg, N., Bunney, G., Chu, A., Kabeer, R., Riley, F., Ribeira, R., & Rose, C. (2024). The role of large language models in transforming emergency medicine: Scoping review. JMIR Medical Informatics, 12, e53787.
44. Rutledge, G. W. (2024). Diagnostic accuracy of GPT-4 on common clinical scenarios and challenging cases. Learning Health Systems. https://doi.org/10.1002/lrh2.10438
45. Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., et al. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. https://doi.org/10.1038/s41586-023-06291-2
46. Singhal, K., et al. (2025). Toward expert-level medical question answering with large language models. Nature Medicine. https://doi.org/10.1038/s41591-024-03423-7
47. Strong, E., DiGiammarino, A., Weng, Y., et al. (2023). Chatbot vs medical student performance on free-response clinical reasoning examinations. JAMA Internal Medicine, 183(9), 1028–1030.
48. Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. Nature Medicine, 29(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8
49. Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. Smart Learning Environments, 10(1), 15.
50. Ueda, D., Mitsuyama, Y., Takita, H., et al. (2023). ChatGPT's diagnostic performance from patient history and imaging findings on the Diagnosis Please quizzes. Radiology, 308(1), e231040. https://doi.org/10.1148/radiol.231040
51. 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, 30. Curran Associates.
52. Vrdoljak, J., Boban, Z., Vilović, M., Kumrić, M., & Božić, J. (2025). A review of large language models in medical education, clinical decision support, and healthcare administration. Healthcare, 13(6), 603. https://doi.org/10.3390/healthcare13060603
53. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., et al. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229).
54. Weissburg, I., Anand, S., Levy, S., & Jeong, H. (2025). LLMs are biased teachers: Evaluating LLM bias in personalized education. arXiv preprint arXiv:2410.14012.
55. Wollny, S., et al. (2021). Are we there yet? A systematic literature review on chatbots in education. Frontiers in Artificial Intelligence, 4, 654924.
56. Wu, F., Dang, Y., & Li, M. (2025). A systematic review of responses, attitudes, and utilization behaviors on generative AI for teaching and learning in higher education. Behavioral Sciences, 15(4), 467. https://doi.org/10.3390/bs15040467
57. Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. British Journal of Educational Technology, 55. https://doi.org/10.1111/bjet.13370
58. Zack, T., Lehman, E., Suzgun, M., Rodriguez, J. A., Celi, L. A., Gichoya, J., Jurafsky, D., Szolovits, P., Bates, D. W., Abdulnour, R.-E. E., Butte, A. J., & Alsentzer, E. (2024). Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: A model evaluation study. The Lancet Digital Health, 6(1), e12–e22.
59. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education, 16(1), 39. https://doi.org/10.1186/s41239-019-0171-0

60. Zhou, C., et al. (2024). A comprehensive survey on pre-trained foundation models: A history from BERT to ChatGPT. International Journal of Machine Learning and Cybernetics.