The evolution of artificial intelligence in healthcare education: A temporal and methodological analysis

Jodie Finn[1,†], Srivarshini Sankar[1,†], Taha Farooqi[1], Mohamed Elhassadi[1] and Ali Hasnain[1]

[1]Royal College of Surgeons in Ireland, 123 St Stephen's Green, Dublin 2, Ireland

Abstract

Artificial intelligence (AI) has rapidly emerged as a transformative technology in higher education, with particularly significant implications for healthcare training. The recent introduction of large language models (LLMs), such as ChatGPT and GPT-4, has accelerated experimentation with AI-assisted learning tools, yet the structural evolution of research in this domain remains poorly understood. This study conducts an exploratory analysis of published research examining AI applications in healthcare education to identify temporal trends, shifts in model adoption, methodological patterns, and thematic emphases in reported outcomes. A curated dataset of studies published between 2015 and 2025 was analysed using descriptive statistics, cross-tabulation, and chi-square tests. Results demonstrate a pronounced increase in publication activity after 2022, coinciding with the public release of ChatGPT. GPT-based systems dominated the literature, accounting for the majority of studies, while traditional machine learning approaches declined proportionally. Despite this rapid expansion, methodological rigour has not increased significantly, with descriptive study designs remaining predominant and randomised controlled trials relatively rare. Thematic analysis further indicates a strong emphasis on performance-related outcomes, with comparatively limited attention to ethical considerations, academic integrity, and governance issues. These findings suggest that while AI-driven educational research is expanding rapidly, methodological maturation and critical evaluation frameworks have yet to develop at the same pace.

Keywords

Artificial Intelligence, Healthcare Education, Medical Education, Educational Technology, Research Trends, Generative AI, Large Language Models, Research Trend Analysis

1. Introduction

Education is fundamentally relational, with effective learning dependent on meaningful interaction between educators and students [1]. Digital transformation within higher education has facilitated increasingly dynamic and technology-enhanced instructional approaches, and recent work in higher-education research documents rapid uptake of AI tools for teaching, learning, and assessment [2, 3]. In their effort to bring technology into the classroom, many international organisations and institutions have emphasised the importance of digitally competent educational agents who can design, orchestrate, and critically oversee AI-mediated learning environments [3, 4].

Within higher education, AI has emerged as a transformative technology influencing pedagogical design and instructional delivery, for example through early-warning systems, adaptive learning, predictive analytics, and AI-supported feedback [5, 6]. These developments have simultaneously raised concerns regarding academic integrity, ethical governance, and the pedagogical implications of automated systems, particularly with the advent of generative AI and conversational agents such as ChatGPT [7, 8]. In healthcare, AI is already augmenting clinical decision-making, image interpretation, and workflow efficiency, prompting calls for "high-performance medicine" in which human clinicians and AI systems collaborate rather than compete [9, 10]. This convergence of human and artificial intelligence in clinical practice paves the way for graduates who can understand, evaluate, and use AI tools appropriately in daily decision making[11].

SeWebMeDA-2026: 9th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 10, 2026, Dubrovnik, Croatia

[†]These authors contributed equally and share first authorship.

jodiefinn25@rcsi.ie (J. Finn); srivarshinisanka25@rcsi.ie (S. Sankar); tahafarooqi24@rcsi.ie (T. Farooqi); mohamed.elhassadi@rcsi.ie (M. Elhassadi); alihasnain@rcsi.ie (A. Hasnain)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Among higher-education domains, medical education represents a particularly high-stakes environment due to its direct implications for patient safety. The implementation of AI in medical education offers transformative possibilities, particularly through its ability to integrate traditionally distinct fields of medical science, including anatomy, biochemistry, pathology, and clinical medicine, into more coherent, case-based frameworks [12]. AI has the potential to assist educators in the creation of integrated and case-based learning experiences, relate previously disparate subject areas, and support students in developing more sophisticated understandings of complex medical concepts and their interrelations [13, 14]. AI-driven case simulations and virtual patients can provide adaptive, interactive scenarios that foster experiential learning and clinical decision-making, while AI-based analytics can be used to detect knowledge gaps and personalise learning trajectories, supported by real-time feedback [15, 16].

Currently, there are significant pedagogical and practical obstacles to incorporating AI into medical-school curricula. The majority of medical programmes are rigid and content-heavy, making it difficult to integrate cutting-edge fields like AI without exacerbating overload [17]. With the exponential growth of medical knowledge and medical data, the conventional focus on memorisation is increasingly unsustainable, and there is growing emphasis on developing competencies in data interpretation, evidence-based decision-making, and human-AI collaboration [18, 19]. This paradigm shift requires re-evaluating educational priorities, strengthening data and AI literacy, and adopting multidisciplinary approaches that involve health-care professionals, data scientists, engineers, and clinicians in curriculum design [20, 21].

The prospective uses, advantages, and limitations of AI in healthcare education have been widely discussed in the literature, yet there is still no empirical mapping of how the research landscape has evolved in response to recent advances in AI, particularly large language models [22, 23]. It remains uncertain whether the rise of LLMs represents a fundamental paradigm shift in methodological focus and thematic emphasis, or an incremental continuation of prior AI-assisted educational research [24, 11]. Furthermore, although ethical and academic-integrity issues are frequently invoked, their relative prominence and treatment within empirical medical-education studies have not been systematically investigated [25, 26]. Several recent scoping and bibliometric reviews have begun to map AI in medical education more broadly, summarising applications across the continuum of undergraduate, postgraduate, and continuing education and highlighting emerging challenges and opportunities [27, 28]. However, these studies typically emphasise thematic categorisation of AI use cases and outcome types, with less attention to temporal inflection points around the release of large language models or to the methodological quality of empirical evaluations. They also give limited quantitative treatment to how frequently ethics, academic integrity, and governance considerations are addressed within the empirical literature.

In this paper, we intend to fill these gaps through an empirical analysis of temporal publication trends in AI-related medical-education research, with particular attention to changes before and after the public release of ChatGPT in 2022, the types of AI models employed, and the methodological rigour of the underlying study designs. By combining temporal, structural, and thematic analyses, including explicit coding of ethical and academic-integrity related content, this study provides a structured overview of how medical education research is evolving in the age of generative AI. Specifically, the paper examines temporal publication trends, patterns of AI model adoption, study design quality across the literature, and thematic emphases within reported findings. Overall, we believe that identifying these patterns provides useful insights for educators, researchers, and policymakers seeking to incorporate AI into medical curricula and prepare future medical professionals for human–AI collaborative healthcare environments. Given the modest sample size, the analyses are intended as an exploratory mapping of trends rather than definitive causal inference, but they nonetheless offer an empirically grounded baseline against which future, larger-scale syntheses can be compared.

The remainder of this paper is organised as follows. Section 2 reviews related studies on artificial intelligence in healthcare education and situates the present work within the existing literature. Section 3 describes the methodology, including data collection, preprocessing, and analytical procedures. Section 4 presents the results of the temporal, structural, and thematic analyses. Section 5 discusses the implications of these findings for research and practice in healthcare education. Finally, Section 6

summarises the main contributions of the study and outlines directions for future work.

2. Related Work

Artificial intelligence has become an increasingly prominent topic in medical education research. Bibliometric analyses show a rapid expansion in publications related to AI in medical education between 2015 and 2025, reflecting growing academic interest and technological development [29]. These analyses also highlight emerging research clusters centred on machine learning, virtual simulation, and large language models in medical training [30]. Similar landscape studies report increasing global research output and expanding interdisciplinary collaboration within this field [31].

Scoping and systematic reviews demonstrate the breadth of AI applications across medical education, including clinical simulation platforms, automated assessment systems, intelligent tutoring tools, and decision-making training environments [28]. Despite this diversity, the evidence base remains heterogeneous and is largely dominated by exploratory studies rather than controlled experimental evaluations [27]. Consequently, these reviews consistently emphasise the need for stronger empirical validation and clearer frameworks to assess the educational impact of AI-based interventions. Recent scoping reviews have mapped AI applications across the continuum of medical education, identifying use cases in admissions, teaching, assessment, and clinical reasoning, while also proposing reporting frameworks to guide future research. Although these syntheses provide a useful overview of application domains and outcome types, they offer limited quantitative analysis of study design quality or changes in methodological rigour over time. Reviews focused specifically on AI efficacy likewise highlight promising early findings but conclude that robust evidence on educational outcomes remains limited and inconsistent, reinforcing the need for larger, well-designed, and validated studies.

More recent work has focused specifically on the emergence of large language models within medical education research. Bibliometric analyses indicate that LLM-related studies have increased rapidly in recent years, particularly following the release of publicly accessible generative AI systems [30]. These models are increasingly explored for educational tasks such as automated feedback, content generation, and support for clinical reasoning exercises. Narrative and scoping reviews of ChatGPT and related tools in healthcare education similarly describe diverse opportunities and challenges, but they largely concentrate on pedagogical use cases and perceived risks rather than systematically quantifying changes in research design or outcome emphasis over time

AI-assisted feedback systems have also emerged as an important research theme within medical education. Bibliometric analysis of this area shows growing interest in AI-driven formative feedback tools designed to support student learning and performance monitoring [32]. Such systems aim to provide scalable feedback mechanisms while complementing traditional instructor-led teaching approaches.

Although these studies provide valuable insights into the growth and thematic structure of AI research in medical education, they primarily focus on publication trends and conceptual mapping of the field. Less attention has been given to examining how methodological approaches, AI model architectures, and reported educational outcomes have evolved together over time.

The present study builds on these prior landscape analyses by jointly examining temporal publication trends, AI model adoption (including LLMs), study design quality, and thematic emphases within reported findings. By analysing how methodological approaches and AI technologies have co-evolved between 2015 and 2025, this work provides a more fine-grained empirical perspective on how medical education research is responding to recent advances in generative AI. A key limitation of earlier reviews is that they typically characterise technological applications and educational themes without systematically linking them to the rigour of study design or to the treatment of ethics and academic integrity. In contrast, the central question in the current analysis is whether the transition to LLM-based systems has been accompanied by changes in evaluation maturity and thematic focus, and where gaps remain.

3. Methodology

A comprehensive literature search was conducted on PubMed (n = 1323), Semantic Scholar (n = 475), and Scopus (n = 393), yielding a total of 2,191 records. After removal of 663 duplicates, 1,528 distinct studies remained for screening. Titles and abstracts were systematically evaluated for relevance, resulting in 147 full-text publications being assessed for eligibility. Conference abstracts and non-empirical publications without full-text access were excluded. In addition, studies were excluded if they addressed only theoretical aspects of AI, explored AI applications outside formal healthcare education, surveyed opinions about AI without evaluating a specific AI-based educational intervention, or did not use artificial intelligence. Applying these criteria, 66 empirical studies were included in the final dataset for analysis. Given this modest sample size, subsequent analyses are framed as exploratory mappings of trends rather than definitive causal inferences.

3.1. Data Collection

Exploratory analysis was conducted on a curated dataset of published studies examining the use of artificial intelligence in formal healthcare education. The dataset was derived from a previously completed systematic review and included bibliographic and study-level variables extracted during screening and data extraction. The literature search was completed in June 2025; consequently, publications from the second half of 2025 are not fully represented, and counts for 2025 should be interpreted as partial-year data rather than full-year totals. Variables available for analysis included publication year, AI model type, study design, educational setting, AI application category, and summaries of key findings. The present analysis focused on temporal publication trends, patterns of model adoption, distribution of study design quality, structural alignment between application category and study design, and thematic emphases in reported outcomes, including ethics- and academic-integrity-related themes. An overview of the analytical workflow used in this study is illustrated in Figure 1.

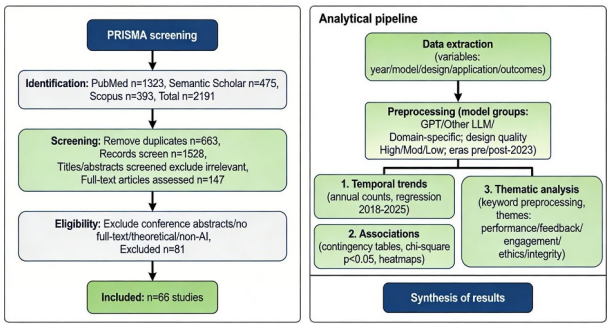

Figure 1: Overview of the analytical workflow used in this study, including data collection, preprocessing, analysis, and visualisation stages.

3.2. Data Preprocessing

Publication year was extracted from bibliographic metadata and converted to numerical format to enable temporal analysis. Studies were grouped by calendar year between 2018 and 2025. For structural comparisons, publications were additionally categorised into pre- and post-2023 eras to explore shifts following the widespread adoption of large language models.

Due to heterogeneity in reported AI model nomenclature, model names were manually standardised to ensure comparability. Variants such as “Chat GPT,” “ChatGPT-4,” “GPT 4,” proprietary platforms, and multi-modal systems were harmonised into three mutually exclusive model groups. The *GPT derivatives* group comprised OpenAI GPT-based systems including GPT-3, GPT-3.5, GPT-4, GPT-4 Turbo, GPT-4o, ChatGPT variants, and custom GPT implementations. The *Other LLM* group comprised non-OpenAI large language models including Gemini, Claude, Bard, LLaMA, Qwen, and Mistral. The *Domain-specific*

AI group comprised task-specific machine learning systems such as clinical decision-support tools, virtual patient simulators, automated assessment systems, natural language processing chatbots, and machine learning classifiers including XGBoost-based models. Where multiple models were evaluated within a single study, classification was based on the dominant model family described.

For broader paradigm-level analysis, models were additionally grouped into LLM-based systems, defined as GPT derivatives and other LLMs, and Non-LLM systems, defined as domain-specific AI systems.

Study designs were categorised into predefined levels of study design quality. High-quality designs were defined as quantitative randomised controlled trials. Moderate-quality designs comprised quantitative non-randomised and mixed-method studies. Lower-quality designs comprised quantitative descriptive and qualitative studies. For inferential testing involving sparse cell counts, study design quality was further dichotomised into High versus Not High to improve statistical stability.

Application categories were defined as AI-generated content in healthcare education, AI in learning and feedback, and AI in OSCE or clinical skills. These categories were used to examine structural alignment between educational application and study design. All analyses were conducted in Python using the `pandas`, `scipy`, and `seaborn` libraries.

3.3. Data Analysis

3.3.1. Temporal and Structural Trend Analysis

Annual publication counts were calculated to assess growth patterns over time. Year-over-year percentage change was computed to evaluate acceleration trends. Linear regression analysis was used to examine the relationship between publication year and annual publication count. Linear regression is commonly used to model relationships between variables and identify temporal trends in quantitative datasets [33]. Given the limited number of yearly observations and the modest overall sample size, inferential results were interpreted cautiously and are intended to support exploratory rather than definitive conclusions.

Structural shifts in model adoption were examined using stacked visualisations representing both raw counts and proportional distributions of LLM-based versus Non-LLM systems across years. Comparative descriptive analyses were conducted between pre-2023 and post-2023 periods to explore potential changes associated with the wider adoption of large language models.

3.3.2. Association and Study Design Quality Analysis

Descriptive statistics were used to summarise the distribution of AI model groups and study designs. Contingency tables were constructed to examine associations between model group and study design quality, publication era and study type, and application category and study design.

Associations between categorical variables were evaluated using Pearson's chi-square test of independence. This statistical test assesses whether two categorical variables are significantly associated by comparing observed frequencies with expected frequencies under the assumption of independence [34]. In this study, the chi-square test was used to examine relationships between AI model groups, study design quality, publication era, and application categories. Where expected cell frequencies were low, study design quality was dichotomised (High versus Not High) to ensure sufficient cell counts for statistical testing, as sparse contingency tables can reduce the reliability and power of association tests. Statistical significance was defined as $p < 0.05$, and effect sizes and patterns in the cross-tabulations were interpreted with caution in light of the small number of studies in some subgroups. Cross-tabulations were visualised using stacked bar charts and heatmaps to identify structural clustering patterns.

3.3.3. Thematic Analysis of Reported Findings

To assess thematic emphasis in reported outcomes, summaries of key findings were analysed using a structured keyword-based thematic approach. Text data were preprocessed through lowercasing

and removal of punctuation and stopwords. Keywords were grouped into thematic domains reflecting common outcome emphases, including performance outcomes, feedback mechanisms, engagement, ethical considerations, and academic integrity. The keyword lists for each domain were developed iteratively, informed by prior literature on AI in medical and healthcare education and by an initial reading of the included studies. Thematic coding was based on the presence of one or more domain-specific keywords within the extracted summaries; frequency distributions were then used to identify dominant themes across the dataset. Because this approach relies on keyword matches rather than full qualitative coding of complete articles, it may underestimate more implicit or nuanced discussions of ethics and academic integrity, and related findings are therefore interpreted as conservative estimates of thematic attention.

4. Results

This section presents the main findings of the analysis, including the distribution of AI model groups, study design quality across the literature, temporal publication trends between 2015 and 2025, and thematic patterns in reported study outcomes.

4.1. AI Model Distribution

GPT derivatives accounted for the majority of included studies (n = 43), followed by domain-specific AI systems (n = 13) and other large language models (n = 10) (Figure 2). This distribution indicates a strong concentration of empirical research within the OpenAI GPT ecosystem relative to other model groups.

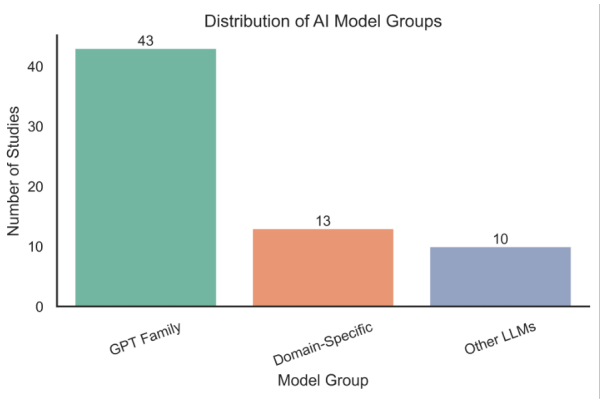

Figure 2: Distribution of included studies by AI model group. GPT-family models accounted for the majority of studies (n=43), followed by domain-specific models (n=13) and other large language models (n=10).

4.2. Model Group and Study Design Quality

Across all model groups, lower-quality study designs predominated (defined in Section 3.3.2), as illustrated in Figure 3. Within the GPT derivatives group, 6 studies (14.0%) were classified as high-quality designs, 12 (27.9%) as moderate-quality, and 25 (58.1%) as lower-quality designs. Domain-specific AI systems included 2 high-quality studies (15.4%), 1 moderate-quality study (7.7%), and 10 lower-quality studies (76.9%). Among other LLMs, 1 study (10.0%) was high-quality, 4 (40.0%) moderate-quality, and 5 (50.0%) lower-quality. Relative proportions across groups are shown in Figure 4.

Using the statistical procedures described in Section 3.3, a Pearson chi-square test showed no significant association between AI model group and study design quality ($\chi^2 = 3.441$, $p = 0.487$). Because several expected cell counts were small, study quality was also analysed using a binary classification (High vs Not High). This analysis likewise showed no significant association ($\chi^2 = 0.150$, $p = 0.928$). These results indicate that the likelihood of implementing a high-quality experimental design does not differ systematically across AI model groups.

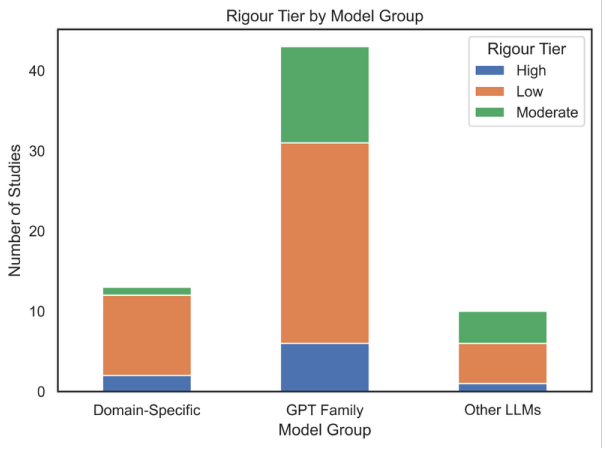

Figure 3: Study design rigour stratified by AI model group, shown as absolute counts across three tiers (high, moderate, low). Across all model groups, low-rigour designs predominated, particularly within GPT-family studies. High-rigour studies (e.g. controlled or randomised designs) were limited in number across all categories.

4.3. Model Group and Application Category

AI models were applied across three educational application categories: AI-generated educational content, learning and feedback tools, and OSCE or clinical skills training contexts.

GPT derivatives were primarily used for AI-generated educational content (23 of 43 studies, 53.5%) and learning or feedback applications (17 of 43 studies, 39.5%). Only three studies (7.0%) used GPT derivatives in OSCE or clinical skills contexts.

Domain-specific AI systems showed a more balanced distribution. Seven studies (53.8%) focused on learning and feedback applications, while three studies each (23.1%) examined AI-generated content and OSCE-based training.

Other LLMs demonstrated proportionally greater representation in OSCE or clinical skills applications (4 of 10 studies, 40.0%) compared with GPT derivatives.

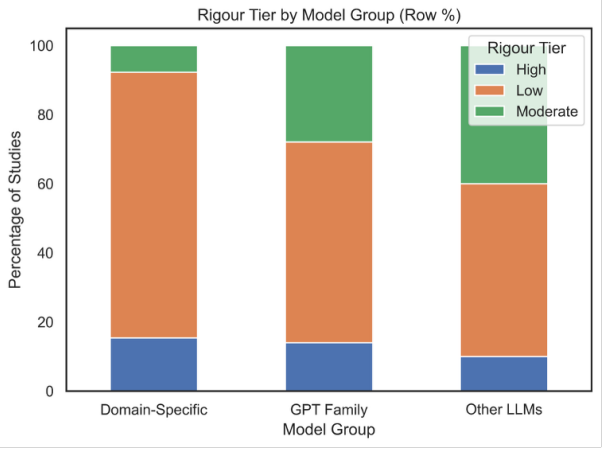

Figure 4: Proportional distribution of study design rigour within each AI model group. Values are normalised to 100% per group to enable comparison independent of sample size.

4.4. Growth of Research Activity (2015–2025)

Publication activity increased substantially over the study period (Figure 5). Between 2015 and 2021, annual publication counts remained low, ranging from one to four studies. Beginning in 2022, publication volume increased rapidly. A pronounced increase was observed in 2023 (19 studies), representing a 171% rise compared with the previous year. The highest number of publications occurred in 2024 (22 studies).

The lower count observed for 2025 reflects partial-year data, as the systematic search was completed in June 2025; therefore, 2025 values are included for transparency but should not be interpreted as a decline in research activity.

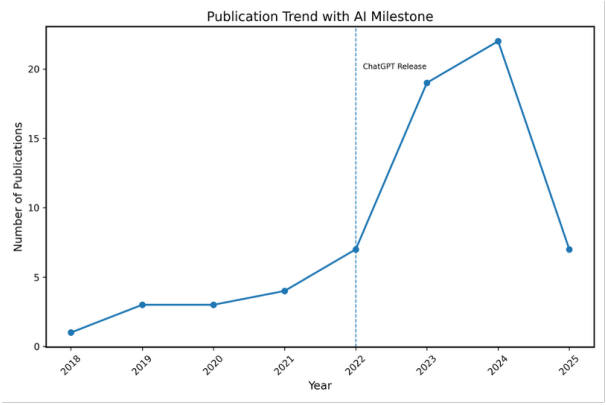

Figure 5: Annual number of included publications from 2018 to 2025, with the release of ChatGPT (late 2022) indicated by a dashed vertical line. A marked increase in publication volume is observed following this milestone, with a sharp rise in 2023 and peak output in 2024. The apparent decline in 2025 reflects partial-year data collection (search completed June 2025) rather than a true reduction in research activity.

4.5. Changes in AI Model Usage

Before 2022, most studies focused on traditional machine-learning systems and other non-generative AI approaches. From 2023 onwards, GPT-based systems became the most frequently studied models, as shown in Figure 6. By 2024, diversification in model usage emerged, including increasing representation of GPT-4 and alternative large language models such as Bard or Gemini.

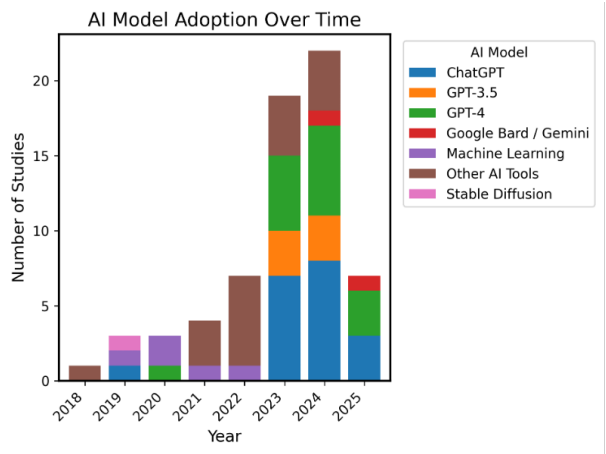

Figure 6: Temporal trends in the adoption of specific AI models across included studies. GPT-based models (ChatGPT, GPT-3.5, GPT-4) show rapid uptake from 2023 onwards, becoming the dominant tools in recent publications.

4.6. Structural Shift: LLM vs Non-LLM Research

Binary classification into LLM-based and Non-LLM groups demonstrates a clear structural shift in research focus (Figure 7). Studies published prior to 2022 were almost exclusively Non-LLM, whereas from 2023 onwards the majority of publications examined LLM-based systems.

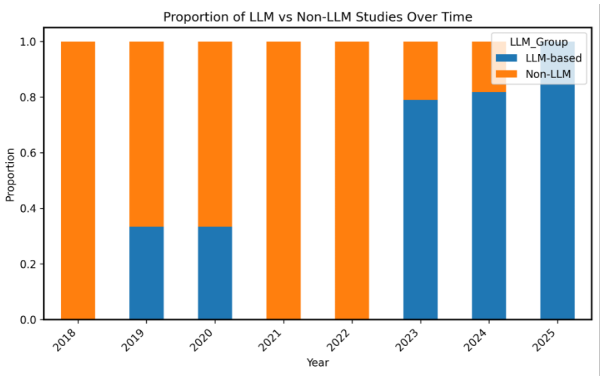

Figure 7: Proportion of studies utilising LLM-based versus non-LLM approaches by publication year. Prior to 2022, studies were predominantly non-LLM. Proportions are presented to account for variation in total study counts across years.

4.7. Thematic Emphasis in Reported Findings

Thematic analysis of reported findings showed a strong emphasis on performance-related outcomes (n = 19), followed by feedback-related benefits (n = 10), as illustrated in Figure 8. Engagement-related outcomes were less frequently reported (n = 3). Ethical considerations appeared rarely in the coded summaries (n = 1), and no explicit references to academic integrity were identified. Given the keyword-based approach used for thematic coding, these counts likely provide conservative estimates of how frequently ethics and academic-integrity issues were addressed.

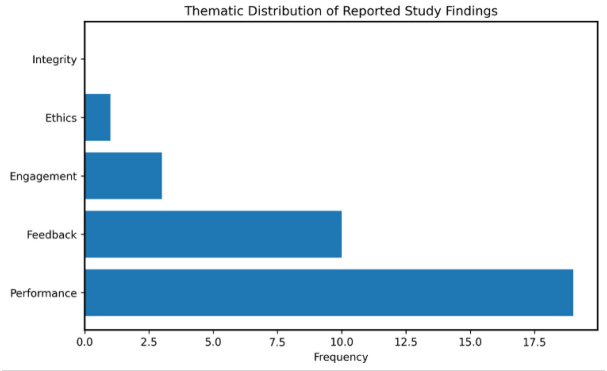

Figure 8: Frequency of reported outcome themes across included studies. Performance-related outcomes were most commonly evaluated, followed by feedback and engagement, while ethical considerations and academic integrity were less frequently addressed.

4.8. Randomised Controlled Trials by Publication Year

Across the dataset, the most common study design was quantitative descriptive (n = 35). Quantitative RCTs and quantitative non-RCT designs each accounted for nine studies. Trends in RCT use over time are shown in Figure 9.

Using the statistical procedures described in Section 3.3, comparison of the pre- and post-2023 periods showed no statistically significant association between publication era and study design ($\chi^2 = 4.89$, $p = 0.299$). This suggests that, within this exploratory dataset, the rapid increase in publications after 2022 was not matched by clear evidence of a corresponding increase in experimental study designs.

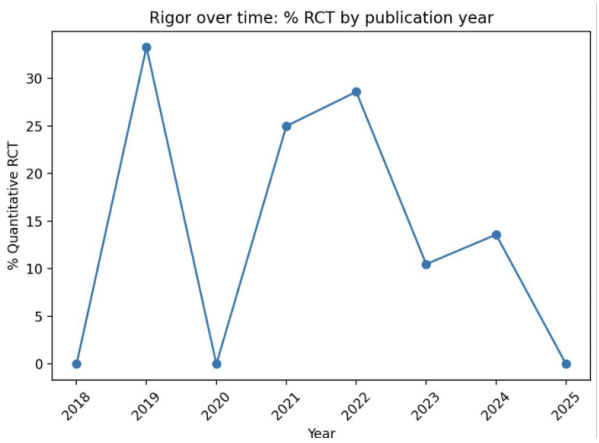

Figure 9: Proportion of studies employing randomised controlled trial (RCT) designs by publication year. Although there are fluctuations over time, no sustained upward trend in high-rigour study designs is observed despite increasing publication volume. This suggests that growth in the field has been driven primarily by exploratory and descriptive studies rather than more rigorous experimental methodologies. The absence of RCTs in certain years highlights inconsistency in methodological advancement.

4.9. Distribution of Study Design Across Application Categories

Study design varied across application domains (Figure 10). Studies examining AI-generated educational content were predominantly descriptive. Research focusing on learning and feedback applications included a relatively higher proportion of experimental or mixed-method approaches. OSCE or clinical skills studies showed the highest proportion of RCT designs.

Using the statistical procedures described in Section 3.3, the association between application category and study design approached but did not reach statistical significance ($\chi^2 = 14.27$, $p = 0.075$). Given the small number of studies within some application categories, this near-threshold result should be interpreted as hypothesis-generating rather than confirmatory, indicating a potential pattern that warrants examination in larger samples.

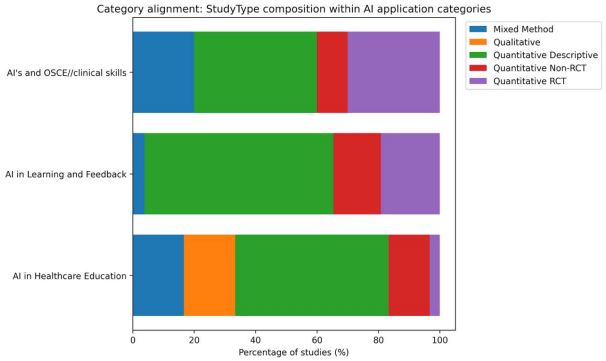

Figure 10: Distribution of study designs within major AI application categories, expressed as percentages.

5. Discussion

This study examined thematic emphases, methodological transitions, and temporal trends in AI-related educational research following the emergence of large language models. Three key patterns were identified: a shift towards LLM-based research, a substantial rise in publication activity after 2022, and a predominance of performance-related outcomes in reported findings.

Publication output increased markedly after 2022, particularly in 2023 following the public release of ChatGPT. The scale and timing of this increase suggest a structural shift rather than gradual growth.

The accessibility and adaptability of LLMs likely reduced barriers to experimentation, enabling rapid adoption across a range of educational contexts. Although interpretation of 2025 trends is limited by partial-year data, the post-2022 surge indicates a clear change in research activity.

Alongside this growth, the analysis identified a broader technological transition from conventional machine-learning models towards LLM-based systems. Earlier AI applications in education commonly focused on predictive analytics, automated assessment, and structured instructional systems. In contrast, LLM-based tools support conversational interaction, adaptive feedback, and dynamic content generation. These capabilities align closely with pedagogical approaches used in medical education, particularly case-based learning and iterative feedback.

Despite these technological developments, the analysis did not demonstrate a corresponding increase in study design quality (defined in Section 3.3.2). Quantitative descriptive designs remained the predominant study type. Statistical analyses described in Section 3.3 showed no significant association between publication era and study design, nor between AI model group and study design quality. This suggests that the rapid expansion of LLM-related research has not yet been accompanied by a shift towards more rigorous experimental evaluation frameworks.

The absence of a statistically significant relationship between AI model group and the likelihood of conducting randomised controlled trials is particularly notable. Although GPT derivatives dominate the literature numerically, they are not disproportionately evaluated using high-quality experimental designs. High-quality studies remain relatively scarce across all model groups. This pattern indicates that the shift in model architecture has not yet translated into a parallel evolution in evaluation standards.

Similarly, descriptive differences were observed across application categories. Learning and feedback contexts showed somewhat greater representation of experimental or mixed-method approaches compared with AI-generated content studies. However, these differences did not reach statistical significance (Section 3.3). This suggests emerging tendencies rather than established differentiation. AI research in healthcare education may therefore still be in an early consolidation phase, where exploratory experimentation precedes the development of stable methodological norms.

These findings highlight an asymmetry between technological advancement and evaluation maturity. The rapid adoption of LLMs reflects increased accessibility and reduced barriers to experimentation. However, rigorous validation frameworks have not expanded at the same pace. In high-stakes domains such as medical education, this gap is particularly consequential. Without robust comparative or longitudinal evaluation, it remains uncertain whether reported performance improvements represent genuine learning gains or short-term engagement effects.

The thematic analysis further showed that reported findings are dominated by performance-related outcomes, followed by feedback-related benefits. Engagement was discussed less frequently, while ethical considerations and academic integrity appeared rarely in empirical findings. Given the prominence of ethical debates surrounding generative AI in public and policy discussions, this disparity is noteworthy. The current literature prioritises measurable academic outcomes over broader questions of governance, reliability, bias, and professional responsibility. Existing frameworks for AI in medicine emphasise that educational deployments should explicitly address issues such as transparency, data protection, fairness, and learner safety, yet these dimensions are only sporadically evaluated in current empirical work.

Several limitations should be acknowledged. The thematic analysis relied on short reported summaries rather than full-text coding, which may underestimate the prevalence of some themes, particularly more nuanced discussions of ethics and academic integrity. In addition, the relatively small number of yearly observations limits the strength of statistical inference regarding temporal trends. The statistical power of the chi-square analyses is constrained by the modest overall sample ($n = 66$) and uneven distribution among model groups. The near-threshold association between application category and study design ($p = 0.075$) should therefore be interpreted as hypothesis-generating rather than confirmatory, as it provides an initial indication of a possible pattern that requires replication in larger datasets. It is also noteworthy that the absolute number of publications analysed has not been normalised with respect to annual output in the wider healthcare education literature, meaning that part of the post-2022 rise could reflect broader increases in scholarly publishing. Applying field-normalised publication indicators in

future work would allow more precise characterisation of whether AI-in-education research is growing faster than the parent field.

Building on these observations, several methodological priorities emerge for future research in AI-assisted healthcare education. First, studies should increasingly use pre-registered designs with clearly defined comparators (for example, standard curricula, non-AI digital tools, or alternative AI systems), so that effect sizes can be interpreted against meaningful benchmarks. Second, outcome measurement should extend beyond immediate performance indicators to include follow-up assessments of knowledge retention, transfer to clinical reasoning, and, where feasible, downstream patient-related proxy outcomes. Third, samples should be recruited across multiple institutions or sites to enhance generalisability and to test whether AI-supported interventions perform consistently across curricular contexts and learner populations. Fourth, structured assessments of bias, reliability, and ethical implications of LLM outputs should be integrated into study protocols rather than treated as secondary considerations. Finally, future synthesis efforts should incorporate discipline-specific analyses and field-normalised publication metrics to clarify which medical subjects are leading AI adoption and whether evaluation maturity is evolving differentially across domains.

6. Conclusion

This study provides a structured empirical overview of how research on artificial intelligence in healthcare education has evolved following the emergence of large language models. The analysis identified a clear increase in publication activity after 2022 and a substantial shift towards LLM-based systems within educational research. However, this rapid technological adoption has not been accompanied by a comparable improvement in study design quality. Most studies remain descriptive, and high-quality experimental designs such as randomised controlled trials remain relatively uncommon across all model groups.

These findings highlight an important imbalance between technological innovation and evaluative maturity within the field. While LLMs have lowered barriers to experimentation and enabled new forms of educational interaction, the empirical evidence supporting their pedagogical effectiveness remains limited. In high-stakes domains such as medical education, stronger empirical validation is required to determine whether reported performance improvements translate into sustained learning gains and meaningful improvements in clinical reasoning or professional competence.

Future research should prioritise rigorous experimental designs, comparative evaluations, and longitudinal studies that assess educational impact beyond short-term performance metrics. Greater attention should also be given to ethical considerations, bias, accountability, and the long-term implications of AI-supported learning environments. Addressing these challenges will be essential to ensure that the integration of AI in healthcare education is guided by robust evidence and responsible practice. Such work should also incorporate systematic assessments of bias, transparency, and academic-integrity safeguards to ensure that evaluation practices keep pace with the rapid evolution of AI technologies in medical education.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-based generative AI tools in order to assist with code troubleshooting, syntax clarification, and minor language editing such as identifying synonyms and improving phrasing. No generative AI tools were used to generate original research data, analyses, figures, or scientific conclusions. All analytical procedures, statistical tests, and interpretations were designed, implemented, and verified by the authors. After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

J.F. and S.S. contributed equally to this work and share first authorship. J.F., S.S., and T.F. contributed to the conceptualisation, methodology development, data analysis, and manuscript preparation. M.E. contributed to data curation, review, and editing of the manuscript. A.H. supervised the project, provided

critical revisions, and contributed to the overall study design. All authors reviewed and approved the final manuscript.

References

[1] A. G. van der Niet, A. Bleakley, Where medical education meets artificial intelligence: Does technology care?, Medical Education 55 (2021) 30–36. doi:10.1111/medu.14131.

[2] P. S. Venkateswaran, F. T. M. Ayasrah, V. K. Nomula, P. Paramasivan, P. Anand, K. Bogeshwaran, Applications of artificial intelligence tools in higher education, in: Advances in Educational Technologies and Artificial Intelligence, IGI Global, 2024. URL: https://www.igi-global.com/chapter/applications-of-artificialintelligence-tools-in-higher-education/335567.

[3] F. Belkhir, Artificial intelligence in higher education, 2024. URL: https://www.researchgate.net/profile/Fatima-Belkhir/publication/380968527_Vol_23_No_3_March_2024/links/6657403222a7f16b4f55f882/Vol-23-No-3-March-2024.pdf.

[4] S. Edwards, A. Cheok, Artificial intelligence in education: A review, Applied Artificial Intelligence (2018). doi:10.1080/08839514.2018.1464286.

[5] M. Ciolacu, et al., Artificial intelligence in education, in: IEEE International Conference on Teaching, Assessment, and Learning, 2018. URL: https://ieeexplore.ieee.org/abstract/document/8599203.

[6] I. Gligorea, et al., Artificial intelligence in higher education: A review, Education Sciences 13 (2023). doi:10.3390/educsci13121216.

[7] Z. Aytaç, Using artificial intelligence tools in higher education, in: Artificial Intelligence in Education, Taylor and Francis, 2024. URL: https://www.taylorfrancis.com/chapters/edit/10.1201/9781003425809-11/using-artificial-intelligence-tools-higher-education-zeynep-aytaç.

[8] C. Kooli, Artificial intelligence in education and sustainability, Sustainability 15 (2023). doi:10.3390/su15075614.

[9] E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nature Medicine 25 (2019) 44–56. doi:10.1038/s41591-018-0300-7.

[10] X. Zhang, et al., Artificial intelligence in scientific research, Nature Computational Science (2024). doi:10.1038/s44401-024-00006-z.

[11] L. G. McCoy, S. Nagaraj, F. Morgado, V. Harish, S. Das, L. A. Celi, What do medical students actually need to know about artificial intelligence?, npj Digital Medicine 3 (2020) 86. doi:10.1038/s41746-020-0294-7.

[12] V. B. Kolachalama, P. S. Garg, Machine learning and medical education, npj Digital Medicine 1 (2018) 54. doi:10.1038/s41746-018-0061-1.

[13] A. H. Sapci, H. Sapci, Artificial intelligence education in medicine, JMIR Medical Education 6 (2020) e19285. doi:10.2196/19285.

[14] R. Luckin, M. Cukurova, Designing ai in education, British Journal of Educational Technology (2019). doi:10.1111/bjet.12861.

[15] P. Olla, et al., Artificial intelligence in medical education, 2024. URL: https://www.researchsquare.com/article/rs-3750487/v1.

[16] A. Wood, et al., Artificial intelligence in medical education, Journal of Medical Education and Curricular Development 8 (2021). doi:10.1177/23821205211024078.

[17] K. S. Chan, N. Zary, Applications and challenges of implementing artificial intelligence in medical education, JMIR Medical Education 5 (2019) e13930. doi:10.2196/13930.

[18] N. Schubert, et al., Artificial intelligence in medicine, The Lancet eClinicalMedicine (2024). doi:10.1016/j.eclinm.2024.102547.

[19] J. Lee, et al., Artificial intelligence in medical education, Academic Medicine 99 (2024) 524–531. URL: https://academic.oup.com/academicmedicine/article/99/5/524/8343986.

[20] R. Charow, et al., Artificial intelligence education in medical curricula, JMIR Medical Education 7 (2021) e31043. doi:10.2196/31043.

[21] M. Sánchez-Mendiola, et al., Artificial intelligence education in medical schools, BMC Medical Education 15 (2015) 34. doi:10.1186/s12909-015-0349-7.

[22] J. Lee, A. S. Wu, D. Li, K. M. Kulasegaram, Artificial intelligence in undergraduate medical education: A scoping review, Academic Medicine 96 (2021) S62–S70. doi:10.1097/ACM.0000000000004291.

[23] J. Grunhut, A. T. Wyatt, O. Marques, Educating future physicians in artificial intelligence (ai): An integrative review and proposed changes, Journal of Medical Education and Curricular Development 8 (2021) 23821205211036836. doi:10.1177/23821205211036836.

[24] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, V. Tseng, Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models, PLOS Digital Health 2 (2023) e0000198. doi:10.1371/journal.pdig.0000198.

[25] D. Cirillo, S. Catuara-Solarz, C. Morey, E. Guney, L. Subirats, S. Mellino, A. Gigante, A. Valencia, M. J. Rementeria, A. S. Chadha, N. Mavridis, Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare, npj Digital Medicine 3 (2020) 81. doi:10.1038/s41746-020-0288-5.

[26] L. Weidener, M. Fischer, Teaching ai ethics in medical education: A scoping review of current literature and practices, Perspectives on Medical Education 12 (2023) 399–410. doi:10.5334/pme.954.

[27] K. Shaw, M. A. Henning, C. S. Webster, Artificial intelligence in medical education: A scoping review of the evidence for efficacy and future directions, Medical Science Educator (2025).

[28] M. Gordon, M. Daniel, A. Ajiboye, H. Uraiby, N. Y. Xu, R. Bartlett, J. Hanson, M. Haas, M. Spadafore, C. Grafton-Clarke, R. Y. Gasiea, C. Michie, J. Corral, B. N. Kwan, D. Dolmans, S. Thammasitboon, A scoping review of artificial intelligence in medical education: Beme guide no. 84, Medical Teacher 46 (2024) 446–470. doi:10.1080/0142159X.2024.2306023.

[29] W. Cheng, et al., A bibliometric analysis of artificial intelligence in medical education (2015–2025), Medicine (2025). doi:10.1097/MD.0000000000041236.

[30] K. Lu, et al., Mapping key nodes and global trends in artificial intelligence and large language models in medical education: A bibliometric analysis, Advances in Medical Education and Practice (2025). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12360372/.

[31] Y. Wang, et al., How ai is transforming medical education: A bibliometric analysis, BMC Medical Education (2025). URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12673300/.

[32] S. Yu, J. Liu, Mapping the landscape of ai-assisted formative feedback in medical education: A bibliometric analysis, Medicine (Baltimore) 105 (2026) e47489. doi:10.1097/MD.0000000000047489.

[33] N. Roustaei, et al., Application and interpretation of linear regression analysis, Ophthalmic Epidemiology (2024).

[34] M. L. McHugh, The chi-square test of independence, Biochemia Medica 23 (2013) 143–149. doi:10.11613/bm.2013.018.