Cureus
Part of SPRINGER NATURE

Open Access Review Article

The Role of Natural Language Processing in Graduate Medical Education: A Scoping Review

Ravi K. Janumpally [1]

1. Clinical Informatics, Baylor Scott & White Medical Center, Round Rock, USA

Corresponding author: Ravi K. Janumpally, ravijaua@yahoo.com

Review began 01/26/2025
Review ended 03/16/2025
Published 03/24/2025

DOI: 10.7759/cureus.81078

Abstract

The rapid evolution of artificial intelligence, particularly in the form of natural language processing (NLP) and large language models (LLMs), presents new opportunities to enhance graduate medical education (GME). NLP technologies have the potential to improve residency training programs by automating performance feedback, personalizing learning pathways, and identifying competency gaps. However, the integration of these technologies also raises challenges related to privacy, ethical considerations, and algorithmic bias. This review provides a comprehensive evaluation of the application and impact of NLP in GME.

A scoping review of the literature was conducted following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Relevant studies from 2018 to 2024 were identified using databases such as PubMed, Scopus, Web of Science, and Google Scholar. Inclusion criteria focused on peer-reviewed studies evaluating NLP applications in residency training programs across various specialties. Data were extracted from 20 studies, and key themes were synthesized to assess the educational, technological, and ethical implications of NLP in GME.

The review identified several key areas where NLP is transforming GME. These include automated performance evaluation systems, sentiment analysis of narrative feedback, personalized learning recommendations, and competency assessment algorithms. NLP technologies demonstrated significant potential in reducing administrative workload, improving assessment accuracy, and enhancing the personalization of residency training. However, studies also highlighted concerns regarding algorithmic biases and the need for transparent, ethical frameworks to ensure fair implementation.

The integration of NLP in GME offers significant opportunities to streamline educational processes and enhance trainee development. Automated feedback systems can reduce subjective biases and provide more actionable insights for residents. Additionally, NLP applications can identify early indicators of residents at risk of underperformance and support timely interventions. However, the adoption of these technologies requires careful consideration of ethical and legal implications, particularly around data privacy and fairness.

NLP has the potential to revolutionize GME by improving the quality and efficiency of residency training programs. While the technology offers promising benefits, further research is needed to address ethical challenges and ensure responsible implementation. Interdisciplinary collaboration between educators, informaticians, and ethicists will be critical to fully realize the potential of NLP in medical education.

Categories: Medical Education, Healthcare Technology
Keywords: graduate medical education, large language models, natural language processing, natural language processing (nlp), residency training

Introduction And Background

Natural language processing (NLP) is a subset of machine learning (ML) focused on enabling computers to understand, interpret, and generate human language, often relying on statistical and deep learning techniques. Generative artificial intelligence (AI), a broader category that includes models like the generative pre-trained transformer (GPT), leverages ML, including NLP, to create new text, images, or other content, distinguishing it from traditional ML approaches that primarily focus on pattern recognition and prediction rather than content generation.

The rapid advancement of AI and large language models (LLMs) has precipitated a transformative wave across numerous professional domains, with graduate medical education (GME) emerging as a particularly promising frontier for technological innovation [1]. NLP technologies have demonstrated remarkable potential to revolutionize multiple aspects of residency training programs, encompassing educational strategies, clinical training methodologies, and resident performance assessment [2,3].

GME faces significant challenges in standardizing evaluation, personalizing learning experiences, and efficiently managing the vast amounts of textual data generated during training [4]. Traditional assessment methods often rely on subjective narratives, handwritten evaluations, and time-consuming feedback mechanisms that are prone to human bias, inconsistency, and sometimes gender bias [5,6]. NLP offers unprecedented opportunities to address these systemic limitations by providing sophisticated, data-driven approaches to analyzing and interpreting complex educational and clinical interactions [7].

The application of NLP in residency training programs spans several critical domains. These include automated performance feedback analysis, competency assessment, and the extraction of meaningful insights from resident-generated documentation [8,9]. ML algorithms can now deconstruct narrative evaluations, identify nuanced performance patterns, and generate actionable recommendations that support both individual resident development and programmatic improvement [10].

Recent studies have highlighted the potential of NLP to enhance several key educational processes. For instance, sentiment analysis of evaluation narratives can provide a more nuanced understanding of resident performance beyond traditional grading systems [11,12]. Automated text classification techniques enable more efficient categorization of clinical competencies, reducing administrative burden and improving assessment accuracy [13]. Moreover, natural language understanding technologies can help identify early indicators of resident struggle or exceptional performance, allowing for more proactive educational interventions [14].

Internationally, the adoption of NLP in GME reveals diverse implementation strategies. While United States residency programs have been at the forefront of technological integration, countries like Canada and China have also begun exploring sophisticated NLP applications in medical training [15-17]. These global perspectives underscore the universal potential of AI-driven educational technologies to transform medical professional development.

However, the implementation of NLP in GME is not without challenges. Significant considerations include maintaining patient privacy, ensuring algorithmic fairness, addressing potential biases in training data, and navigating complex ethical landscapes [18]. The responsible development and deployment of these technologies require interdisciplinary collaboration among medical educators, computational experts, and ethicists [19,20].

As LLMs continue to evolve, their potential to revolutionize GME becomes increasingly apparent. The ability to process, understand, and generate human-like text offers unprecedented opportunities for personalized learning, comprehensive assessment, and continuous improvement of medical training programs [21-24]. Research must focus on developing robust, ethically sound NLP frameworks that can be seamlessly integrated into existing educational infrastructures [25,26]. Thus, this review aims to provide a comprehensive, nuanced understanding of NLP's transformative potential in GME, balancing technological innovation with pedagogical effectiveness and ethical considerations.

Review

Research aims, objectives, and framework

Overall Research Question

The research investigated the current landscape, effectiveness, and potential impact of NLP technologies in GME, specifically within residency training programs, across educational, clinical training, and assessment domains.

Specific Objectives

This study aimed to comprehensively explore the role of NLP technologies in residency training by addressing several dimensions. It sought to conduct technological mapping, identifying and categorizing the various NLP technologies utilized in medical education. This included analyzing their technological sophistication and computational approaches to better understand how they supported residency training programs. The study also focused on evaluating the educational impact of NLP tools in improving critical aspects of medical education, including their role in resident performance assessment, the development of personalized learning pathways, enhancements in curriculum design, and the refinement of feedback mechanisms. A comparative analysis was performed to investigate NLP implementation strategies across different medical specialties and geographic regions. This analysis identified variations in technological adoption and educational outcomes, providing a global perspective on the integration of NLP into medical education. Additionally, the study addressed ethical considerations and implementation challenges by examining barriers to adoption and proposing solutions to the technological and pedagogical challenges that arose during integration. It also explored the future trajectory of NLP in GME by predicting emerging trends in its application and recommending evidence-based strategies to ensure successful technological integration in the years to come. Through these efforts, the project aspired to provide a comprehensive framework for understanding and advancing NLP technologies in medical education. The overall research

question and specific objectives of the study are outlined in the PICO (population, intervention, comparison, and outcomes) framework in Table *1*.

PICO component	Detailed description
P (Population)	- Medical residents in graduate training programs
	- Specific specialties: Internal medicine, surgery, pediatrics, psychiatry, emergency medicine
	- Geographic focus: United States, Canada, United Kingdom, Australia, European Union
	- Exclusion: Undergraduate medical students, fellows, practicing physicians
I (Intervention)	- Natural language processing (NLP) technologies
	- Large language models (LLMs)
	- Artificial intelligence-driven educational tools
	- Specific applications:
	* Automated performance evaluation systems
	* Intelligent feedback generation
	* Personalized learning path recommendations
	* Clinical documentation analysis
	* Competency assessment algorithms
C (Comparison)	- Traditional medical education assessment methods
	- Manual evaluation processes
	- Conventional feedback mechanisms
	- Human-only performance assessment
	- Pre-NLP educational technologies
O (Outcomes)	Primary outcomes:
	- Accuracy of resident performance assessment
	- Quality and specificity of educational feedback
	- Personalization of learning experiences
	- Reduction in administrative workload
	- Improvement in resident skill development
	Secondary outcomes:
	- Technological adoption rates
	- Cost-effectiveness of NLP implementations
	- User satisfaction (residents and educators)
	- Ethical considerations and potential biases
	- Long-term educational and clinical performance impacts

TABLE 1: PICO framework.

PICO: population, intervention, comparison, and outcomes.

Methodology

The scoping review was conducted and reported in accordance with the Preferred Reporting Items for

Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) 2018 guidelines [27]. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) diagram is presented in Figure *1*. The study employed rigorous eligibility criteria to ensure the inclusion of relevant and high-quality research on the application of NLP and LLMs in GME. The target population included residents enrolled in GME programs worldwide, with a primary focus on residency training programs in the United States, Canada, and the United Kingdom. Studies involving undergraduate medical education were excluded from the scope. The interventions of interest included studies that applied NLP and LLM technologies to resident education, clinical training, performance assessment, curriculum development, and feedback mechanisms. Eligible study types comprised peer-reviewed original research, including retrospective and prospective studies, randomized controlled trials, observational studies, and quasi-experimental studies. Only studies published in English between January 2018 and October 2024 and appearing in peer-reviewed journals were included.

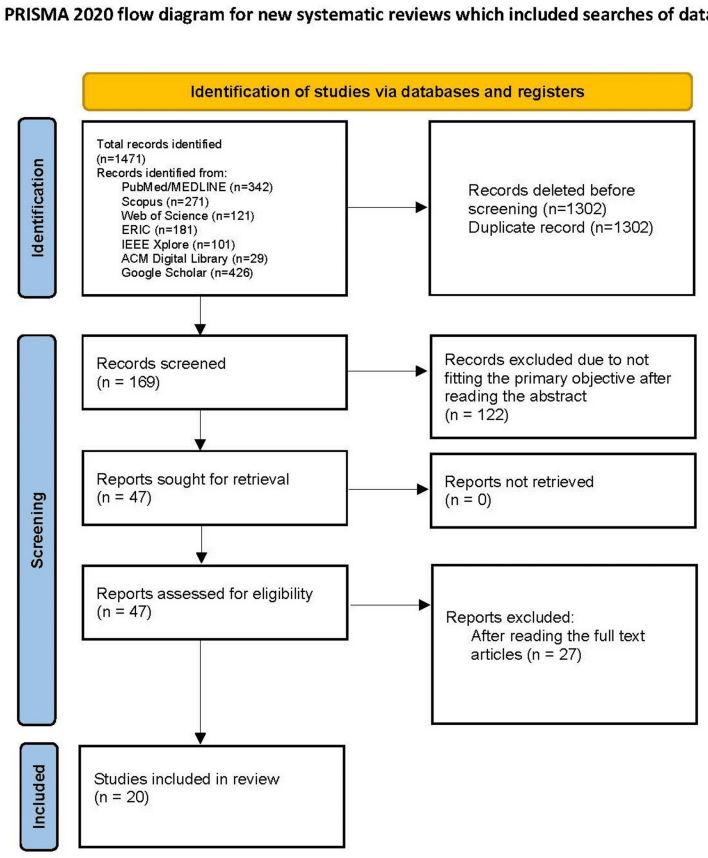

FIGURE 1: PRISMA diagram.

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

A comprehensive search strategy was employed using several electronic databases, including PubMed/MEDLINE, Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and Google Scholar. The search strategy combined Boolean operators and Medical Subject Heading (MeSH) terms to identify studies relevant to NLP and LLM technologies in residency training. The search string in PubMed included the terms “Natural Language Processing” OR “Large Language Models” AND combined with “Graduate Medical Education“ OR "Residency” OR “Assessment” OR “Performance”. The study selection process followed a structured workflow. An initial database search identified potential studies, followed by duplicate removal. Screening occurred in two stages: title and abstract screening and a full-text review. One reviewer, the author RJ, conducted all stages of the review and extraction for this study. Unfortunately, multiple reviewers were not available for both screening stages. Full-text articles of potentially relevant studies were retrieved based on predefined criteria. Reasons for exclusion were documented.

Data extraction used a standardized and piloted form to capture key information from selected studies, including study identification details, study design, population characteristics, NLP methodology, specific technologies or models used, educational domains, key findings, limitations, and ethical considerations. The analysis involved qualitative synthesis, which employed thematic analysis to identify recurring patterns, context-specific insights, and trends in technological innovation. Together, these analyses provided a

comprehensive understanding of the role and impact of NLP and LLM technologies in GME.

Results

Table *2* summarizes the key information from the 20 included studies [16,28-46].

Sr. No.	Author(s)/year	Type of study	Study material	Purpose of study	Outcomes of study	Findings of study	Sample size (N)
1	Nori et al. (2023) [16]	Model evaluation study	Capabilities of GPT-4 on medical challenge problems	Evaluation of GPT-4's performance on USMLE Step 1–3 and other medical benchmarks	GPT-4 exceeded USMLE Step 3 passing score by over 20 points; improved calibration and reasoning vs. GPT-3.5	Demonstrates strong potential for GPT-4 as a tool in graduate medical education, especially for Step 3 preparation and AI-assisted learning	For Step 3, n = 267
2	Yilmaz et al. (2022) [28]	Retrospective	Workplace-based assessment (WBA) narrative comments	To explore NLP and machine learning (ML) applications for identifying trainees at risk using a large WBA narrative comment data set associated with numerical ratings.	NLP was performed on a data set of narrative comments derived from WBAs, and supervised ML models using linear regression were trained with the quantitative ratings to output a prediction of whether a resident fell into the category of at-risk or not at risk.	The ML models can accurately identify underperforming residents via narrative comments provided for WBAs, and the words generated in WBAs can be a worthy data set to augment human decisions for educators tasked with processing large volumes of narrative assessments.	N = 7199
3	Spadafore et al. (2024) [29]	Retrospective	2,500 entrustable professional activity (EPA) assessments from emergency medicine residency programs	To develop an NLP model for applying the Quality of Assessment for Learning (QuAL) score to narrative supervisor comments.	The NLP model trained on QuAL-rated EPA comments to predict overall and subcomponent scores, with performance assessed across improvement suggestions and performance link dimensions.	The model predicted the exact or near-exact QuAL score (±1) in 87% of cases, performing well in evaluating feedback quality, improvement suggestions, and the link between comments and resident performance.	N = 2,500
4	Solano et al. (2021) [30]	Retrospective	Surgical resident feedback transcripts	To validate the performance of a natural language processing (NLP) model in characterizing the quality of feedback provided to surgical trainees.	The NLP model classified the quality (high vs. low) of 2,416 narrative feedback transcripts with an accuracy of 0.83, sensitivity of 0.37, specificity of 0.97, and an area under the receiver operating characteristic curve of 0.86.	The NLP model classified the quality of operative performance feedback with high accuracy and specificity, offering residency programs the opportunity to efficiently measure feedback quality.	N = 2416
5	Sarraf et al. (2021) [31]	Retrospective	Letters of recommendation (LoRs) for general surgery residency candidates	To examine gender bias in LoRs written for surgical residency candidates across three decades at one institution.	LoRs were analyzed using artificial intelligence (AI) and natural language processing (NLP) to conduct sentiment analysis and detect gender bias.	AI and computer-based algorithms detected linguistic differences and gender bias in LoRs, even following stratification by clerkship grades and when analyzed by decade.	N = 611
6	Gudgel et al. (2021) [32]	Retrospective	Ophthalmology residency applications	To identify application attributes correlated with successful ophthalmology residency performance.	Residents were subjectively ranked into tertiles and top and bottom deciles based on residency performance, and various application attributes were analyzed for associations.	Many metrics traditionally felt to be predictive of residency success (USMLE scores, AOA status, and research) did not predict resident success. Core clerkship grades and medical school ranking were associated with high subjective ranking.	N = 76
						The NLP model using a support vector machine	

7	Ötleş et al. (2021) [33]	Retrospective	Surgical trainee feedback comments	To evaluate whether NLP can be used to automatically classify the quality of surgical trainee evaluations.	NLP models were trained to automatically classify the quality of feedback across 4 categories (effective, mediocre, ineffective, or other).	algorithm yielded a maximum mean accuracy of 0.64 in classifying feedback quality. When the classification task was modified to distinguish only high-quality vs. low-quality feedback, the maximum mean accuracy was 0.83.	N = 600
8	Abbott et al. (2021) [34]	Retrospective	End-of-rotation assessments and clinical competency committee (CCC) assessments for surgical residents	To examine whether NLP can be used to estimate CCC ratings.	Models of end-of-rotation assessment ratings and text were created to predict dichotomized CCC assessment ratings for 16 ACGME milestones, with and without predictors derived from NLP of end-of-rotation assessment text.	NLP can identify language correlated with specific ACGME milestone ratings, and faculty could use information automatically extracted from text to focus attention on residents who might benefit from additional support and guide the development of educational interventions.	N = 691
9	Andrews et al. (2021) [35]	Retrospective	Internal medicine resident evaluations	To examine differences in word use, competency themes, and length within written evaluations of internal medicine residents, considering the impact of both faculty and resident gender.	Researchers developed a sentiment model to assess the valence of evaluation responses and used NLP to evaluate whether female versus male residents received more positive or negative feedback and if that feedback focused on different ACGME core competencies.	When examined at scale, quantitative gender differences were not as prevalent as has been seen in qualitative work, suggesting that further investigation of linguistic phenomena (such as context) is warranted to reconcile this finding with prior work.	N = 3864
10	Stahl et al. (2021) [36]	Retrospective	Entrustable professional activity (EPA) micro-assessments	To analyze actual EPA assessment narrative comments using NLP to enhance our understanding of resident entrustment in actual practice.	Latent Dirichlet allocation (LDA), a machine learning algorithm, was used to identify latent topics in the documents associated with a single EPA, which were then reviewed for interpretability by human raters.	LDA is capable of identifying topics relevant to progressive surgical entrustment and autonomy in EPA comments, providing insight into key behaviors that drive different levels of resident autonomy and may allow for data-driven revision of EPA entrustment maps.	N = 1015
11	Geller et al. (2022) [37]	Retrospective	Orthopedic surgery residency program social media presence	To analyze the changes in social media usage by orthopedic surgery programs in response to the COVID-19 pandemic.	Social media data (Instagram and Twitter) were collected and analyzed, including the total number of followers, accounts following, tweets, likes, date of account creation, hashtags, and mentions. Natural language processing was utilized for tweet sentiment analysis.	The study demonstrated substantial growth of Instagram and Twitter presence by orthopedic surgery residency programs during the COVID-19 pandemic, suggesting that programs have utilized social media as a new way to communicate with applicants and showcase their programs.	N = 85
12	Ortiz et al. (2023) [38]	Retrospective	Neurosurgery residency applications	To compare the performance of machine learning models trained on applicant narrative letters of recommendation (NLORs) and demographic data to predict match outcomes and investigate whether narrative language is predictive of	Logistic regression models using the least absolute shrinkage and selection operator were trained to predict match outcomes using applicant NLOR text and demographics, and another model was trained on NLOR text	Both the NLOR and demographics models were able to discriminate similarly between match outcomes. Words including "outstanding," "seamlessly," and "AOA" were predictive of match success, and words like "highest," "outstanding,"	N = 1498

				standardized letter of recommendation (SLOR) rankings.	to predict SLOR rankings.	and "best" were predictive of the top 5% SLORs.	
13	Drum et al. (2023) [39]	Retrospective	Internal medicine-pediatrics residency applications	To develop a machine learning model (MLM) that can identify several values important for resident success in internal medicine-pediatrics programs.	Expert reviewers annotated text snippets from the narrative sections of applications into specific values (e.g., academic strength, compassion, communication) associated with resident success. The authors then applied the MLM to prospective applications and compared the output with a manual holistic review.	The MLM had a sensitivity of 0.64 and a specificity of 0.97, and the mean total number of annotations per application was significantly correlated with invited for interview status.	N = 11,333
14	Mahtani et al. (2023) [40]	Retrospective	Residency application experience entries	To develop an NLP-based tool to automate the review of applicants' narrative experience entries and predict interview invitations.	Experience entries were extracted from residency applications, combined at the applicant level, and paired with the interview invitation decision. NLP identified important words (or word pairs), which were used to predict interview invitation using logistic regression.	The NLP model had an AUROC of 0.80 and AUPRC of 0.49, showing moderate predictive strength. Phrases indicating active leadership, research, or work in social justice and health disparities were associated with the interview invitation.	N = 188,500
15	Gray et al. (2023) [41]	Retrospective	Letters of recommendation (LOR) for pediatric surgical fellowship applications	To analyze the prevalence and type of bias in LORs using NLP.	Demographics were extracted from submitted applications. The Valence Aware Dictionary for Sentiment Reasoning (VADER) model was used to calculate polarity scores, and the National Research Council dataset was used for emotion and intensity analysis.	Differences in the intensity of emotions, particularly "anger," were statistically significant between racial and gender groups. While the types of emotions identified were highly similar, the intensity of these emotions revealed differences that may influence an individual's likelihood of successful matching.	N = 701
16	Guillen-Grima et al. (2023) [42]	Retrospective	Spanish Medical Residency Entrance Examination (MIR) questions	To assess the performance of the GPT-3.5 and GPT-4 language models in passing the MIR examination.	The LLMs were presented with 182 MIR examination questions in Spanish and English, and their performance was analyzed, including across different medical specialties, between theoretical and practical questions, and in terms of error proportions and severity.	GPT-4 outperformed GPT-3.5, scoring 86.81% in Spanish. While the models performed well, with error rates below 13.2%, understanding the severity of errors is critical when considering AI's potential role in real-world medical practice and its implications for patient safety.	
17	Booth et al. (2024) [43]	Retrospective	Anesthesiology graduate medical education narrative feedback	To fine-tune several large language models (LLMs) to explore the tradeoff between model complexity and performance while classifying narrative feedback on trainees into the ACGME sub-competencies.	The authors fine-tuned transformer-based LLMs (BERT-base, BERT-medium, BERT-small, BERT-mini, BERT-tiny, and SciBERT) to predict ACGME subcompetencies on a dataset of 10,218 feedback comments and compared their performance to a previous FastText model.	No models were superior to FastText, but BERT-mini performed comparably while being 94% smaller, which may allow for faster deployment and enhanced data privacy.	
			Cardiovascular medicine questions from	To analyze the ability of	The performance of ChatGPT (versions 3.5 and 4), internal medicine residents, and internal	ChatGPT-4 demonstrated an accuracy of 74.5%, comparable to internal medicine residents and attendings but significantly	

18	Milutinovic et al. (2024) [44]	Retrospective	the Medical Knowledge Self-Assessment Program (MKSAP)	ChatGPT to answer board-style cardiovascular medicine questions.	medicine and cardiology attendings in answering 98 multiple-choice questions from the Cardiovascular Medicine Chapter of MKSAP was evaluated.	lower than cardiology attendings (85.7%). Subcategory analysis revealed no statistical difference between ChatGPT and physicians, except in valvular heart disease and heart failure.	
19	Vasan et al. (2024) [45]	Retrospective	Otolaryngology residency application letters of recommendation (LORs)	To utilize natural language processing and machine learning (ML) models using LOR text to predict interview invitations for otolaryngology residency applicants.	LORs were retrospectively retrieved, preprocessed, and vectorized using three techniques (CountVectorizer, TF-IDF, and Word2Vec), and then trained and tested on five ML models (logistic regression, naive Bayes, decision tree, random forest, and support vector machine).	The two best-performing ML models in predicting interview invitations were the TF-IDF vectorized decision tree and CountVectorizer vectorized decision tree models, providing better-than-chance predictions.	N = 1642
20	Heath et al. (2024) [46]	Retrospective	Work-based assessments (WBAs) of standardized resident-patient encounters	To explore gender differences in WBA ratings as well as narrative comments when scripted performance was known.	Multivariable regression was used to assess gender differences in mean entrustment ratings, and natural language processing categories (masculine, feminine, agentic, and communal words) were used to determine associations of word use in the narrative comments with resident gender, race, and skill level, as well as faculty demographics.	Significant differences in entrustment ratings were found between women and men standardized residents, and feminine terms were more common in comments about what women residents did poorly, even after adjusting for the faculty's entrustment ratings.	N = 1527

TABLE 2: Summary of included studies pertaining to the role of NLP and LLM in graduate medical education.

NLP: natural language processing; LLM: large language model; USMLE: United States Medical Licensing Examination; AOA: alpha-omega-alpha; ACGME: Accreditation Council for Graduate Medical Education; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision-recall curve; BERT: Bidirectional Encoder Representations from Transformers; TF-IDF: term frequency-inverse document frequency.

The studies summarized in the table make valuable contributions to the growing body of research on the application of NLP and ML in medical education and graduate medical training. These computational approaches have demonstrated the ability to analyze diverse data sources, identify biases, automate assessment processes, and provide insights into the factors that influence trainee development and performance. As the field continues to evolve, the integration of these technologies into educational practices holds the promise of enhancing the assessment and support of medical trainees.

The papers reviewed in the provided table highlight the growing application of NLP and ML techniques in medical education and graduate medical training. These studies demonstrate the potential of these computational approaches to analyze various types of data, including residency applications, letters of recommendation, assessment narratives, and social media presence.

Several key findings emerge from the synthesis of results in the following domains.

Predicting Residency Performance and Matching

Studies by Ortiz et al. (2023), Drum et al. (2023), and Vasan et al. (2024) utilized NLP and ML models to analyze application materials, such as narrative letters of recommendation and experience entries, to predict interview invitation and match outcomes for residency applicants [38,39,45]. These models showed moderate to high predictive accuracy, suggesting the value of language-based features in assessing candidate potential.

Identifying Bias in Assessments

Researchers have employed NLP techniques to investigate potential gender and racial biases in residency evaluations and letters of recommendation. Studies by Sarraf et al. (2021), Gray et al. (2023), and Heath et al. (2024) found linguistic differences and variations in the intensity of emotions expressed in these materials, which may influence the perceived performance and success of trainees from underrepresented groups [31,41,46].

Automating Assessment Processes

NLP models have demonstrated the ability to classify the quality of feedback provided to trainees, as shown in the studies by Solano et al. (2021), Ötleş et al. (2021), and Spadafore et al. (2024). These findings suggest that automated systems could assist in the efficient review and improvement of assessment practices, potentially complementing human evaluations [29,30,33].

Understanding Entrustment and Competency Development

Researchers have used NLP to analyze narrative comments from workplace-based assessments and entrustable professional activity (EPA) evaluations, as reported by Stahl et al. (2021), Yilmaz et al. (2022), and Milutinovic et al. (2024) [28,36,44]. These studies provide insights into the factors that influence trainee autonomy and the development of competencies, informing educational interventions.

Evaluating AI Capabilities

The recent study by Guillen-Grima et al. (2023) explored the performance of LLMs, such as GPT-3.5 and GPT-4, in passing a Spanish medical residency entrance examination [42]. While the models performed well, the authors emphasize the importance of understanding the severity of errors when considering the potential role of AI in medical practice.

Discussion

The studies summarized in the table demonstrate the growing application of NLP and ML techniques in the realm of medical education and graduate medical training. These computational approaches have the potential to enhance various aspects of the assessment and selection processes for medical trainees.

One of the key areas where NLP and ML have shown promise is in predicting residency performance and match outcomes. Studies by Ortiz et al. (2023), Drum et al. (2023), and Vasan et al. (2024) utilized these techniques to analyze application materials, such as narrative letters of recommendation and experience entries, and achieve moderate to high predictive accuracy for interview invitations and match success [38,39,45]. This aligns with the findings of Mahtani et al. (2023), who employed NLP to analyze the relationship between applicant's narrative experience entries and subsequent interview invitation, highlighting the potential of language-based features in assessing candidate potential [40].

The detection of bias in assessments is another area where NLP and ML have been applied. The studies by Sarraf et al. (2021), Gray et al. (2023), and Heath et al. (2024) found linguistic differences and variations in the intensity of emotions expressed in letters of recommendation and performance evaluations, which may influence the perceived success of trainees from underrepresented groups [31,41,46]. These findings align with the broader literature on the prevalence of bias in medical education, as reported by Dayal et al. (2017), Bludevich et al. (2021), and Mueller et al. (2020) [47-49]. The ability of NLP and ML to systematically identify these biases could inform efforts to improve the fairness and equity of assessment practices.

The potential for NLP and ML to automate and enhance assessment processes is also evident in the reviewed studies. Solano et al. (2021), Ötleş et al. (2021), and Spadafore et al. (2024) demonstrated the use of these techniques to classify the quality of feedback provided to trainees and analyze narrative comments from workplace-based assessments and EPA evaluations [29,30,33]. These findings are consistent with the work of Kalet et al. (2010), who explored the use of NLP to provide real-time feedback on clinical performance, and the research of Otleş et al. (2021), who reported an NLP-based system to analyze the quality of feedback in surgical training [33,50].

Furthermore, the studies by Yilmaz et al. (2022), Stahl et al. (2021), and Milutinovic et al. (2024) provide insights into the factors that influence trainee autonomy and the development of competencies, informing educational interventions [28,36,44]. This aligns with the broader efforts to understand and support the progression of medical trainees through the use of NLP and ML, as evidenced by the work of Grządzielewska et al. (2021) and Neves et al. (2021) [10,51].

The recent study by Guillen-Grima et al. (2023) on the performance of LLMs (GPT-3.5 and GPT-4) in passing a Spanish medical residency entrance examination highlights the importance of understanding the severity of errors when considering the potential role of AI in medical practice [42]. This concern is echoed in the broader literature by Cabitza et al. (2017), Knudsen et al. (2024), and Vora et al. (2023), where researchers have cautioned about the need for careful evaluation of the safety and reliability of AI systems in healthcare

settings [52-54].

Limitations and future recommendations

While the studies reviewed in this systematic review demonstrate the potential of NLP and ML in medical education and graduate medical training, there are several limitations and areas for future research.

Generalizability

Most of the studies were conducted at single institutions or within specific medical specialties, which may limit the generalizability of the findings. Expanding the research to diverse educational settings and larger, multi-institutional datasets would help validate the robustness and broader applicability of the computational approaches.

Ethical Considerations

The detection of bias in assessments using NLP and ML raises important ethical concerns regarding the fairness, transparency, and accountability of these systems. Future research should explore the development of ethical frameworks and guidelines to ensure the responsible and equitable deployment of these technologies in medical education.

User Engagement and Collaboration

Effective integration of NLP and ML tools into educational practices requires close collaboration between researchers, educators, and trainees. Future studies should actively involve end-users in the design, development, and evaluation of these systems to ensure they meet the needs and address the concerns of the medical education community.

Longitudinal Impact

The reviewed studies primarily focused on short-term outcomes, such as interview invitations and match success. Longitudinal research is needed to evaluate the long-term impact of NLP and ML-based interventions on trainee performance, well-being, and career trajectories.

Validation and Benchmarking

Standardized datasets and evaluation protocols would help facilitate the comparison and benchmarking of different NLP and ML models, enabling the identification of best practices and the most effective approaches for specific educational tasks.

Interpretability and Explainability

As the complexity of NLP and ML models increases, it is essential to develop methods that enhance the interpretability and explainability of their decision-making processes. This would foster trust, transparency, and the ability to understand the underlying factors driving the models' outputs.

Conclusions

The studies reviewed in this systematic analysis demonstrate the growing application of NLP and ML techniques in the field of medical education and graduate medical training. These computational approaches have shown promise in predicting residency performance and match outcomes, identifying biases in assessments, automating feedback processes, and providing insights into the factors that influence trainee development and competency progression.

The integration of NLP and ML into educational practices holds the potential to enhance the fairness, efficiency, and personalization of the assessment and support systems for medical trainees. However, the successful implementation of these technologies requires addressing key limitations, such as generalizability, ethical considerations, user engagement, longitudinal impact, validation, and interpretability. As the field continues to evolve, future research should focus on addressing these challenges and expanding the evidence base to ensure the responsible and effective integration of NLP and ML in medical education. Collaborative efforts between researchers, educators, and trainees will be crucial in shaping the future of computational tools that support the development and success of the next generation of medical professionals.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Ravi K. Janumpally

Acquisition, analysis, or interpretation of data: Ravi K. Janumpally

Drafting of the manuscript: Ravi K. Janumpally

Critical review of the manuscript for important intellectual content: Ravi K. Janumpally

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Buch VH, Ahmed I, Maruthappu M: Artificial intelligence in medicine: current trends and future possibilities. Br J Gen Pract. 2018, 68:143-4. 10.3399/bjgp18X695213
2. Chen JS, Baxter SL: Applications of natural language processing in ophthalmology: present and future . Front Med (Lausanne). 2022, 9:906554. 10.3389/fmed.2022.906554
3. Ravi A, Neinstein A, Murray SG: Large language models and medical education: preparing for a rapid transformation in how trainees will learn to be doctors. ATS Sch. 2023, 4:282-92. 10.34197/ats-scholar.2023-0036PS
4. Danilovich N, Kitto S, Price DW, Campbell C, Hodgson A, Hendry P: Implementing competency-based medical education in family medicine: a narrative review of current trends in assessment. Fam Med. 2021, 53:9-22. 10.22454/FamMed.2021.453158
5. Hong DZ, Goh JL, Ong ZY, et al.: Postgraduate ethics training programs: a systematic scoping review . BMC Med Educ. 2021, 21:338. 10.1186/s12909-021-02644-5
6. Gold JM, Yemane L, Keppler H, Balasubramanian V, Rassbach CE: Words matter: examining gender differences in the language used to evaluate pediatrics residents. Acad Pediatr. 2022, 22:698-704. 10.1016/j.acap.2022.02.004
7. Ouyang F, Jiao P: Artificial intelligence in education: the three paradigms . Comput Educ Artif Intell. 2021, 2:100020. 10.1016/j.caeai.2021.100020
8. Gordon M, Daniel M, Ajiboye A, et al.: A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024, 46:446-70. 10.1080/0142159X.2024.2314198
9. Jian MJKO: Personalized learning through AI. Adv Eng Innov. 2023, 5:16-9. 10.54254/2977-3903/5/2023039
10. Neves SE, Chen MJ, Ku CM, et al.: Using machine learning to evaluate attending feedback on resident performance. Anesth Analg. 2021, 132:545-55. 10.1213/ANE.0000000000005265
11. Bansal A, Kumar N: Aspect-based sentiment analysis using attribute extraction of hospital reviews . New Gener Comput. 2022, 40:941-60. 10.1007/s00354-021-00141-3
12. Srisankar M, Lochanambal KP: A survey on sentiment analysis techniques in the medical domain . Medicon Agri Environ Sci. 2024, 6:4-9. 10.55162/MCAES.06.157
13. Krive J, Isola M, Chang L, Patel T, Anderson M, Sreedhar R: Grounded in reality: artificial intelligence in medical education. JAMIA Open. 2023, 6:ooad037. 10.1093/jamiaopen/ooad037
14. Ariaeinejad A, Samavi R: A performance predictive model for emergency medicine residents . eHealth. 2017, 1-52.
15. Sun L, Yin C, Xu Q, Zhao W: Artificial intelligence for healthcare and medical education: a systematic review. Am J Transl Res. 2023, 15:4820-8.
16. Nori H, King N, McKinney SM, Carignan D, Horvitz E: Capabilities of GPT-4 on medical challenge problems. [PREPRINT]. arXiv. 2023, 10.48550/arXiv.2303.13375
17. Chary M, Parikh S, Manini AF, Boyer EW, Radeos M: A review of natural language processing in medical education. West J Emerg Med. 2019, 20:78-86. 10.5811/westjem.2018.11.39725
18. Saeidnia HR, Hashemi Fotami SG, Lund B, Ghiasi N: Ethical considerations in artificial intelligence interventions for mental health and well-being: ensuring responsible implementation and impact. Soc Sci. 2024, 13:381. 10.3390/socsci13070381
19. Alowais SA, Alghamdi SS, Alsuhebany N, et al.: Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023, 23:689. 10.1186/s12909-023-04698-z
20. Al Kuwaiti A, Nazer K, Al-Reedy A, et al.: A review of the role of artificial intelligence in healthcare . J Pers Med. 2023, 13:951. 10.3390/jpm13060951
21. Abd-Alrazaq A, AlSaad R, Alhuwail D, et al.: Large language models in medical education: opportunities, challenges, and future directions. JMIR Med Educ. 2023, 9:e48291. 10.2196/48291
22. Naveed H, Khan AU, Qiu S, et al.: A comprehensive overview of large language models. [PREPRINT] . arXiv. 2023, 10.48550/arXiv.2307.06435
23. Taşkın Bedizel NR: Evolving landscape of artificial intelligence (AI) and assessment in education: a bibliometric analysis. Int J Assess Tool Educ. 2023, 10:208-23. 10.21449/ijate.1369290
24. Nazi ZA, Peng W: Large language models in healthcare and medical domain: a review . Informatics. 2024,

11:57. 10.3390/informatics11030057
25. Pregowska A, Perkins M: Artificial intelligence in medical education: typologies and ethical approaches . Ethics Bioeth. 2024, 14:96-113. 10.2478/ebce-2024-0004
26. Masters K: Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158 . Med Teach. 2023, 45:574-84. 10.1080/0142159X.2023.2186203
27. Tricco AC, Lillie E, Zarin W, et al.: PRISMA Extension for Scoping Reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018, 169:467-73. 10.7326/M18-0850
28. Yilmaz Y, Jurado Nunez A, Ariaeinejad A, Lee M, Sherbino J, Chan TM: Harnessing natural language processing to support decisions around workplace-based assessment: machine learning study of competency-based medical education. JMIR Med Educ. 2022, 8:e30537. 10.2196/30537
29. Spadafore M, Yilmaz Y, Rally V, et al.: Using natural language processing to evaluate the quality of supervisor narrative comments in competency-based medical education. Acad Med. 2024, 99:534-40. 10.1097/ACM.0000000000005634
30. Solano QP, Hayward L, Chopra Z, et al.: Natural language processing and assessment of resident feedback quality. J Surg Educ. 2021, 78:e72-7. 10.1016/j.jsurg.2021.05.012
31. Sarraf D, Vasiliu V, Imberman B, Lindeman B: Use of artificial intelligence for gender bias analysis in letters of recommendation for general surgery residency candidates. Am J Surg. 2021, 222:1051-9. 10.1016/j.amjsurg.2021.09.034
32. Gudgel BM, Melson AT, Dvorak J, Ding K, Siatkowski RM: Correlation of ophthalmology residency application characteristics with subsequent performance in residency. J Acad Ophthalmol (2017). 2021, 13:e151-7. 10.1055/s-0041-1733932
33. Ötleş E, Kendrick DE, Solano QP, et al.: Using natural language processing to automatically assess feedback quality: findings from 3 surgical residencies. Acad Med. 2021, 96:1457-60. 10.1097/ACM.0000000000004153
34. Abbott KL, George BC, Sandhu G, et al.: Natural language processing to estimate clinical competency committee ratings. J Surg Educ. 2021, 78:2046-51. 10.1016/j.jsurg.2021.06.013
35. Andrews J, Chartash D, Hay S: Gender bias in resident evaluations: natural language processing and competency evaluation. Med Educ. 2021, 55:1383-7. 10.1111/medu.14593
36. Stahl CC, Jung SA, Rosser AA, et al.: Natural language processing and entrustable professional activity text feedback in surgery: a machine learning model of resident autonomy. Am J Surg. 2021, 221:369-75. 10.1016/j.amjsurg.2020.11.044
37. Geller JS, Massel DH, Rizzo MG, Schwartz E, Milner JE, Donnally CJ 3rd: Social media growth of orthopaedic surgery residency programs in response to the COVID-19 pandemic. World J Orthop. 2022, 13:693-702. 10.5312/wjo.v13.i8.693
38. Ortiz AV, Feldman MJ, Yengo-Kahn AM, Roth SG, Dambrino RJ, Chitale RV, Chambless LB: Words matter: using natural language processing to predict neurosurgical residency match outcomes. J Neurosurg. 2023, 138:559-66. 10.3171/2022.5.JNS22558
39. Drum B, Shi J, Peterson B, Lamb S, Hurdle JF, Gradick C: Using natural language processing and machine learning to identify internal medicine-pediatrics residency values in applications. Acad Med. 2023, 98:1278-82. 10.1097/ACM.0000000000005352
40. Mahtani AU, Reinstein I, Marin M, Burk-Rafel J: A new tool for holistic residency application review: using natural language processing of applicant experiences to predict interview invitation. Acad Med. 2023, 98:1018-21. 10.1097/ACM.0000000000005210
41. Gray GM, Williams SA, Bludevich B, et al.: Examining implicit bias differences in pediatric surgical fellowship letters of recommendation using natural language processing. J Surg Educ. 2023, 80:547-55. 10.1016/j.jsurg.2022.12.002
42. Guillen-Grima F, Guillen-Aguinaga S, Guillen-Aguinaga L, et al.: Evaluating the efficacy of ChatGPT in navigating the Spanish Medical Residency Entrance Examination (MIR): promising horizons for AI in clinical medicine. Clin Pract. 2023, 13:1460-87. 10.3390/clinpract13060130
43. Booth GJ, Hauert T, Mynes M, Hodgson J, Slama E, Goldman A, Moore J: Fine-tuning large language models to enhance programmatic assessment in graduate medical education. J Educ Perioper Med. 2024, 26:E729. 10.46374/VolXXVI_Issue3_Moore
44. Milutinovic S, Petrovic M, Begosh-Mayne D, Lopez-Mattei J, Chazal RA, Wood MJ, Escarcega RO: Evaluating performance of ChatGPT on MKSAP cardiology board review questions. Int J Cardiol. 2024, 417:132576. 10.1016/j.ijcard.2024.132576
45. Vasan V, Cheng CP, Lerner DK, Pascual K, Mercado A, Iloreta AM, Teng MS: Machine learning for predictive analysis of otolaryngology residency letters of recommendation. Laryngoscope. 2024, 134:4016-22. 10.1002/lary.31439
46. Heath JK, Kogan JR, Holmboe ES, Conforti L, Park YS, Dine CJ: Gender differences in work-based assessment scores and narrative comments after direct observation. J Gen Intern Med. 2024, 39:1795-802. 10.1007/s11606-024-08645-6
47. Dayal A, O'Connor DM, Qadri U, Arora VM: Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. JAMA Intern Med. 2017, 177:651-7. 10.1001/jamainternmed.2016.9616
48. Bludevich B, Irby I, Chang H, Danielson PD, Gonzalez R, Snyder CW, Chandler NM: Letters of recommendation for pediatric surgery fellowship: analysis of linguistic differences based on gender of the applicant. J Pediatr Surg. 2021, 56:1299-304. 10.1016/j.jpedsurg.2021.02.049
49. Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM: Gender differences in attending physicians' feedback to residents: a qualitative analysis. J Grad Med Educ. 2017, 9:577-85. 10.4300/JGME-D-17-00126.1
50. Kalet AL, Gillespie CC, Schwartz MD, et al.: New measures to establish the evidence base for medical education: identifying educationally sensitive patient outcomes. Acad Med. 2010, 85:844-51. 10.1097/ACM.0b013e3181d734a5
51. Grządzielewska M: Using machine learning in burnout prediction: a survey . Child Adolesc Soc Work J. 2021, 38:175-80. 10.1007/s10560-020-00733-w

52. Cabitza F, Rasoini R, Gensini GF: Unintended consequences of machine learning in medicine . JAMA. 2017, 318:517-8. 10.1001/jama.2017.7797
53. Knudsen JE, Ghaffar U, Ma R, Hung AJ: Clinical applications of artificial intelligence in robotic surgery . J Robot Surg. 2024, 18:102. 10.1007/s11701-024-01867-0
54. Vora LK, Gholap AD, Jetha K, Thakur RR, Solanki HK, Chavda VP: Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics. 2023, 15:1916. 10.3390/pharmaceutics15071916