REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

O IMPACTO DE GRANDES MODELOS DE LINGUAGEM COMO TUTORES CLÍNICOS: ESTUDO PILOTO EXPLORATÓRIO SOBRE A VIABILIDADE E ACEITABILIDADE DA AVALIAÇÃO AUTOMATIZADA DE ANAMNESES E FEEDBACK A ESTUDANTES DE MEDICINA

LARGE LANGUAGE MODELS AS CLINICAL TUTORS: AN EXPLORATORY PILOT STUDY OF FEASIBILITY AND ACCEPTABILITY IN AUTOMATED CLINICAL NOTE EVALUATION AND STUDENT FEEDBACK

LOS MODELOS DE LENGUAJE A GRAN ESCALA COMO TUTORES CLÍNICOS: UN ESTUDIO PILOTO EXPLORATORIO SOBRE LA VIABILIDAD Y ACEPTABILIDAD DE LA EVALUACIÓN AUTOMATIZADA DE HISTORIAS CLÍNICAS Y RETROALIMENTACIÓN A ESTUDIANTES DE MEDICINA

Renato Faria da Gama
Doutor em Cognição e Linguagem pela Universidade Estadual do Norte Fluminense Darcy Ribeiro, UENF, Brasil
E-mail: renatofgama@gmail.com

Beatriz Lobo-da-gama
Discente do Curso de Medicina pela Faculdade Metropolitana São Carlos, FAMESC, Brasil
E-mail: beatrizgama04@gmail.com

Isadora Ladislau Meireles
Discente do Curso de Medicina pela Universidade Federal do Rio de Janeiro, Macaé, RJ, Brasil

E-mail: isadoramed30@ufrj.br

Juliane Cunha de Oliveira
Discente do Curso de Medicina pela Universidade Federal do Rio de Janeiro, Macaé, RJ, Brasil
E-mail: juliecunha29@gmail.com

Maria Eduarda Martins
Discente do Curso de Medicina pela Universidade Federal do Rio de Janeiro, Macaé, RJ, Brasil
E-mail: maria.ed2703@gmail.com

Maria Eduarda Rodrigues Machado
Discente do Curso de Medicina pela Universidade Federal do Rio de Janeiro, Macaé, RJ, Brasil
E-mail: madurmachado@gmail.com

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

Mayara Milagre de Araújo
Discente do Curso de Medicina pela Universidade Federal do Rio de Janeiro, Macaé, RJ, Brasil
E-mail: mayaramilagre@ufrj.br

Claudia Gusmão
Mestre em Ciências Ambientais pelo Instituto de Biodiversidade e Sustentabilidade, NUPEM, Brasil.
E-mail: claudia.gusm@hotmail.com

Resumo

Introdução e Objetivo: A correção manual de anamneses na educação médica enfrenta desafios como tempo limitado de preceptoria, feedback insuficiente e subjetividade na avaliação. Este estudo piloto exploratório avaliou a viabilidade, aceitabilidade e impacto preliminar da utilização de um Grande Modelo de Linguagem (LLM) como tutor clínico para avaliação automatizada de anamneses e feedback a estudantes de medicina. **Metodologia:** Estudo observacional, longitudinal e retrospectivo, com abordagem quantitativa e qualitativa. Participaram cinco estudantes do 4º período, acompanhadas em sete ciclos avaliativos, totalizando 35 anamneses. A intervenção utilizou a plataforma NotebookLM, estruturada pela técnica Chain-of-Thought. O LLM avaliou o rigor propedêutico, aplicou checklist baseado em evidências e forneceu feedback com reforço positivo. Todos os relatórios foram revisados por um preceptor sênior antes do envio às estudantes. Ao final, realizou-se um grupo focal. **Resultados:** Os escores evoluíram de 6,0–8,0 (média±DP: 7,0±0,82) para 9,6–10,0 (9,88±0,13). As participantes relataram maior segurança e boa aceitação do feedback estruturado. O preceptor modificou 14 dos 35 relatórios (40%), principalmente para corrigir rigidez algorítmica e acrescentar contexto clínico. As limitações incluíram amostra reduzida, ausência de validação independente das notas e falta de grupo comparador. **Conclusão:** O estudo demonstra viabilidade técnica e aceitabilidade preliminar da tutoria clínica baseada em LLM, desde que associada à supervisão docente (Human-in-the-Loop). Estudos maiores e controlados são necessários para confirmar sua eficácia e orientar sua implementação na educação médica.

Palavras-chave: Educação Médica; Inteligência Artificial; Anamnese; Processamento de Linguagem Natural; Avaliação Educacional.

Abstract

Introduction and Objective: Manual evaluation of medical histories in medical education faces challenges such as limited preceptorship time, insufficient feedback, and subjectivity in assessment. This exploratory pilot study evaluated the feasibility, acceptability, and preliminary impact of using a Large Language Model (LLM) as a clinical tutor for automated medical history evaluation and feedback to medical students. **Methodology:** This observational, longitudinal, retrospective study used quantitative and qualitative approaches. Five fourth-semester medical students participated in seven evaluation cycles, totaling 35 medical histories. The intervention used the NotebookLM platform structured with the Chain-of-Thought technique. The LLM assessed clinical rigor, applied an evidence-based checklist, and provided feedback using positive reinforcement. All reports were reviewed by a senior preceptor before being sent to the students. A focus group was conducted at the end of the study. **Results:** Scores improved from 6.0–8.0 (mean±SD: 7.0±0.82) to 9.6–10.0 (9.88±0.13). Participants reported increased confidence and good acceptance of the structured feedback. The preceptor modified 14 of the 35 reports (40%), mainly to correct algorithmic rigidity and add clinical context. Limitations included the small sample, lack of independent validation of the scores, and absence of a comparison group. **Conclusion:** This study demonstrates the technical feasibility and preliminary acceptability of LLM-based clinical tutoring when combined with faculty

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

supervision (Human-in-the-Loop). Larger controlled studies are needed to confirm its effectiveness and guide its implementation in medical education.

Keywords: Medical Education; Artificial Intelligence; Medical History Taking; Natural Language Processing; Educational Assessment.

Resumen

Introducción y Objetivo: La evaluación manual de historias clínicas en la educación médica enfrenta desafíos como el tiempo limitado de tutoría, la retroalimentación insuficiente y la subjetividad en la evaluación. Este estudio piloto exploratorio evaluó la viabilidad, aceptabilidad e impacto preliminar del uso de un Modelo de Lenguaje a Gran Escala (LLM) como tutor clínico para la evaluación automatizada de historias clínicas y la retroalimentación a estudiantes de medicina. **Metodología:** Se realizó un estudio observacional, longitudinal y retrospectivo, con enfoques cuantitativo y cualitativo. Participaron cinco estudiantes de cuarto semestre, seguidas durante siete ciclos de evaluación, totalizando 35 historias clínicas. La intervención utilizó la plataforma NotebookLM, estructurada mediante la técnica Chain-of-Thought. El LLM evaluó el rigor clínico, aplicó una lista de verificación basada en evidencia y proporcionó retroalimentación mediante refuerzo positivo. Todos los informes fueron revisados por un preceptor senior antes de enviarse a las estudiantes. Al finalizar, se realizó un grupo focal. **Resultados:** Las puntuaciones aumentaron de 6,0–8,0 (media±DE: 7,0±0,82) a 9,6–10,0 (9,88±0,13). Las participantes informaron mayor seguridad y buena aceptación de la retroalimentación estructurada. El preceptor modificó 14 de los 35 informes (40%), principalmente para corregir la rigidez algorítmica y añadir contexto clínico. Las limitaciones incluyeron el tamaño reducido de la muestra, la ausencia de validación independiente de las puntuaciones y la falta de un grupo comparador. **Conclusión:** Este estudio demuestra la viabilidad técnica y la aceptabilidad preliminar de la tutoría clínica basada en LLM, siempre que esté acompañada de supervisión docente (Human-in-the-Loop). Se necesitan estudios controlados y con muestras mayores para confirmar su eficacia y orientar su implementación en la educación médica.

Palabras clave: Educación Médica; Inteligencia Artificial; Historia Clínica; Procesamiento del Lenguaje Natural; Evaluación Educativa.

1. Introduction

The capacity to approach, interpret, and document a patient's clinical history constitutes one of the most complex and fundamental pillars of medical education and clinical-diagnostic reasoning. Historically, the teaching of semiology has been grounded in classical texts that emphasize the exhaustiveness of the clinical history and anatomoclinical correlation, consolidating the principle that medicine is both a science and an art directed toward the human being in their totality. However, the traditional process of evaluation and correction of clinical narratives—such as in the Objective Structured Clinical Examination (OSCE) and

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

chart review—faces substantial practical challenges. It demands significant preceptor availability, is subject to human fatigue, and frequently employs evaluation criteria that may be inconsistent or unfair, resulting in student dissatisfaction and loss of motivation [1] [2].

In this context, the emergence of Natural Language Processing (NLP) and Large Language Models (LLMs) offers a new paradigm for medical education. Recent literature demonstrates that artificial intelligence models can evaluate clinical notes with notable precision, assisting in education and diagnostic reasoning [3] [4]. LLM-based models have demonstrated the capacity to match or exceed human evaluators in specific clinical documentation tasks, significantly reducing correction time while maintaining high levels of reliability [5]. The application of LLMs in undergraduate medical education has expanded, encompassing feedback generation to the simulation of complex clinical scenarios [6] [7].

The rationale for this study is grounded in the application of this emerging technology not merely as a spelling or structural corrector, but as an “Empathic Clinical Tutor.” Using the Chain-of-Thought (CoT) prompt engineering technique, the system replaces punitive judgment with positive reinforcement, transforming “errors” into “Learning Opportunities.” This approach actively mitigates common cognitive biases in early-stage students, such as premature diagnostic closure and neglect of social determinants of health. Additionally, the system offers real-time mentorship by inserting “Clinical Pearls”—knowledge snippets related to pathophysiology and epidemiology—and an etymological literacy engine that accelerates technical erudition [8]. The intervention seeks to balance high algorithmic precision with the necessity of faculty supervision (Human-in-the-Loop) to avoid ethical pitfalls such as mechanization of medical writing and loss of human nuance [9] [10].

Given this context, the present exploratory pilot study is guided by the following research questions: “What is the feasibility of implementing an LLM-based clinical tutor for automated feedback on medical student clinical documentation? What is the preliminary acceptability of this intervention among

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

students? What are the perceived risks and necessary adjustments for future implementation?" Rather than claiming to establish causal efficacy, this pilot study aims to document preliminary evidence of feasibility, acceptability, and perceived impact, while identifying limitations and recommendations for future research.

The primary objectives of this research are to (1) assess the technical feasibility of implementing an LLM-based clinical tutor; (2) document preliminary acceptability and perceived impact among students; (3) identify perceived risks, limitations, and necessary adjustments; and (4) provide recommendations for future rigorous studies with adequate controls and independent validation.

2. Methods

2.1 Study Design and Sample

This was an observational, longitudinal, and retrospective study with both quantitative and qualitative components. The study population consisted of a convenience sample of five female medical students in their fourth year of medical school from a single institution. Students were selected based on their enrollment in the clinical semiology course during the study period. No formal inclusion or exclusion criteria were pre-specified. The study evaluated the progression of performance and subjective perceptions of students across seven evaluation cycles. Each student produced seven clinical notes over the intervention period, totaling 35 clinical documents for analysis. The longitudinal design allowed tracking of each student's developmental trajectory, identifying patterns of improvement and persistence of difficulties over time.

2.2 Intervention: Technological Implementation and Prompt Engineering

Clinical note correction was processed by a Large Language Model (LLM) orchestrated to function as an "Empathic Clinical Tutor" and senior preceptor. The system's prompt engineering adopted the Chain-of-Thought (CoT) prompt engineering approach, which structures the input to encourage step-by-step

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

reasoning. It is important to note that CoT is a prompt engineering technique and does not provide direct access to the model's internal reasoning processes. The prompt was designed to request the LLM to execute four sequential analytical steps.

First, Privacy Filtering (Named Entity Recognition - NER): Automatic scanning at the edge level to ensure removal of any real sensitive data or patient identification information. This step is critical for protecting privacy and complying with data protection regulations, preventing any exposure of personal or health information.

Second, Structural Decomposition: Algorithmic segmentation of the text according to classical semiological guidelines: Identification, Chief Complaint (CC), History of Present Illness (HPI), Past Medical History (PMH), Physiological History (PhysH), Family History (FH), Social History (SH), Review of Systems, Vital Signs, General Appearance, and Diagnostic Impression. This decomposition allows granular analysis of each component of the clinical history, identifying gaps and inconsistencies in specific sections, facilitating detection of structural omissions or errors.

Third, Evidence-Based Analysis (Checklist) and Scoring: The student's production was compared against a gold-standard template, with a raw score assigned based on a rigorous checklist with a maximum of 37.0 points. This checklist encompassed criteria for completeness, accuracy, logical structure, and appropriateness to the level of training. The raw score was mathematically converted to the classical 0–10.0 scale, allowing comparability with traditional evaluation systems.

Fourth, Literacy and Etymology Engine: Detection of colloquial expressions to translate and reframe them semantically into corresponding medical jargon, coupled with Greco-Latin etymological recovery of terms. This functionality not only corrects language but educates the student about the origin and precise meaning of medical terms, promoting deeper understanding of terminology.

To abolish the punitive character of evaluation, the prompt was instructed to adopt positive reinforcement, replacing terminology such as "errors" with "Learning

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

Opportunities." In each generated report, the LLM actively included "Clinical Pearls"—knowledge snippets on pathophysiology, epidemiology, and clinical decision-making—tailored to the specific deficiencies identified in the student's note. The system was also configured to flag premature diagnostic closure and omissions of social determinants of health, prompting the student to reconsider multisystemic diagnostic impressions.

2.3 Evaluation Instruments

The evaluation checklist was developed through review of standard semiological teaching guidelines and consensus among three experienced clinicians (mean experience: 18 years). The 37-point checklist assessed: completeness of each clinical history section (identification, HPI, PMH, FH, SH, review of systems), accuracy of documented findings, logical structure and coherence, appropriate use of medical terminology, and incorporation of social determinants of health. The checklist was designed to align with learning objectives for fourth-year medical students in clinical semiology.

The checklist underwent content validity assessment using the Content Validity Index (CVI) methodology. Each item was rated by the three expert clinicians on a 4-point scale (1=not relevant, 2=somewhat relevant, 3=quite relevant, 4=highly relevant). The overall CVI was 0.89, indicating excellent content validity. Items with individual CVI <0.78 were revised or removed.

Rationale for Checklist Weighting: The checklist weights reflect the level of complexity and diagnostic importance inherent to each semiological domain. Sections such as Identification and Chief Complaint consist primarily of standardized questions with minimal cognitive demand and limited diagnostic specificity. Past Medical History and Family History, while important for contextual understanding, require moderate complexity in synthesis but contribute indirectly to diagnostic formulation. In contrast, the History of Present Illness (HPI) and Diagnostic Impression represent the core of diagnostic complexity and clinical reasoning. The revised checklist (37 points) reflects this hierarchy of complexity and diagnostic importance.

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

2.4 Faculty Supervision and Human-in-the-Loop Implementation

All 35 LLM-generated feedback reports were reviewed by a senior faculty preceptor (a clinician with 20 years of experience in clinical semiology and medical education) before transmission to students. The preceptor's review process followed explicit criteria: assessment of whether the LLM feedback was clinically accurate and evidence-based; verification that feedback was developmentally appropriate for fourth-year students; identification of instances where the LLM feedback was overly rigid or missed important clinical nuances; evaluation of whether the "Learning Opportunities" framing was pedagogically sound; and assessment of whether the feedback inadvertently incentivized "teaching to the algorithm" rather than genuine clinical reasoning.

The preceptor intervened and modified the LLM-generated feedback in 14 of 35 instances (40% of total reports), primarily when: (1) Algorithmic Rigidity Correction (n=8 instances): The algorithm penalized clinically appropriate alternative approaches or emphasized checklist compliance over clinical judgment. (2) Clinical Context Addition (n=4 instances): The algorithm provided accurate feedback but missed important clinical context or educational opportunity. (3) Tone and Framing Adjustment (n=2 instances): The algorithm's "Learning Opportunities" framing, while non-punitive, could have been more encouraging or supportive.

2.5 Ethics, Privacy, and Data Governance

This study was conducted as a quality improvement initiative within the institution’s medical education program. All participating students provided informed written consent. As co-authors of the study, they were fully aware of the ethical aspects involved in the analysis of their anonymized data and agreed to its use in the preparation of this manuscript.

Nature of Data: The clinical notes analyzed in this study were based on didactic case scenarios and simulated patient presentations, not real patient cases. All data were de-identified and anonymized prior to analysis. Student identifiers were replaced with alphabetic codes (Student A, B, C, D, E).

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

Data Protection and LGPD Compliance: Data processing complied with the Brazilian General Data Protection Law (Lei Geral de Proteção de Dados - LGPD) and institutional data protection policies. Clinical notes and LLM-generated feedback were stored on secure institutional servers with access restricted to the research team. Data retention period was set at three years, after which all records will be securely destroyed. Feedback to students was delivered through the institutional learning management system (Moodle), ensuring compliance with data protection standards. All communications were encrypted and access-controlled.

Confidentiality and Anonymization: All data presented in this manuscript have been thoroughly anonymized. Student identifiers, specific case details, and any potentially identifying information have been removed or generalized. The focus group discussion was conducted in a confidential setting, and participants were reminded of confidentiality obligations.

2.6 Data Analysis

Quantitative Analysis: Descriptive statistics were calculated for all 35 scores, including mean, standard deviation, median, quartiles, and range for each evaluation cycle. The complete dataset is presented in Table 1. Progressive change in scores across cycles was assessed descriptively; no inferential statistical testing was performed given the small sample size and pilot nature of the study.

Qualitative Analysis: Focus group data were analyzed using thematic analysis methodology. Audio recordings were transcribed verbatim and independently coded by two researchers to identify recurring themes and patterns. Codes were organized into broader thematic categories reflecting student perceptions of the intervention's benefits, limitations, and recommendations. Interrater agreement was assessed using Cohen's kappa (κ=0.82, indicating substantial agreement). Representative quotes were selected to illustrate each theme. The complete thematic analysis is presented in Table 2.

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

3. Results

3.1 Quantitative Findings: Score Progression

All 35 clinical notes were successfully evaluated by the LLM-based tutor across seven evaluation cycles. The complete dataset is presented in Table 1, which includes individual scores for each student across all cycles, along with comprehensive descriptive statistics (mean, standard deviation, median, quartiles, and interquartile range) for each cycle.

Table 1: Complete Dataset of Clinical Note Scores Across Seven Evaluation Cycles with Comprehensive Descriptive

Cycle	Student A	Student B	Student C	Student D	Student E	Mean	SD	Median	Q1	Q3	IQR	Min	Max
1	6.5	7.0	6.0	7.5	8.0	7.0	0.82	7.0	6.5	7.5	1.0	6.0	8.0
2	7.2	7.5	6.8	8.0	8.3	7.56	0.68	7.5	7.2	8.0	0.8	6.8	8.3
3	7.8	8.0	7.5	8.5	8.8	8.12	0.56	8.0	7.8	8.5	0.7	7.5	8.8
4	8.5	8.8	8.2	9.0	9.1	8.72	0.42	8.8	8.5	9.0	0.5	8.2	9.1
5	9.0	9.2	8.9	9.3	9.4	9.16	0.18	9.2	9.0	9.3	0.3	8.9	9.4
6	9.5	9.6	9.4	9.7	9.8	9.60	0.16	9.6	9.5	9.7	0.2	9.4	9.8
7	9.8	9.9	9.7	10.0	10.0	9.88	0.13	9.9	9.8	10.0	0.2	9.7	10.0

Initial scores ranged from 6.0 to 8.0 (mean ± SD: 7.0 ± 0.82; median: 7.0), while final cycle scores ranged from 9.6 to 10.0 (mean ± SD: 9.88 ± 0.13; median: 9.9). A progressive improvement in scores was observed across the seven evaluation cycles, with the standard deviation decreasing over time (from 0.82 to 0.13), suggesting convergence toward higher performance levels. The progressive decrease in standard deviation indicates convergence of student performance toward higher levels. The median scores parallel the mean scores across all cycles, suggesting symmetric distributions without outliers. The interquartile range (IQR) narrowing from 1.0 to 0.2 demonstrates increasing consistency in student performance.

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

It is important to emphasize that this improvement may reflect multiple mechanisms, including increased familiarity with the evaluation checklist, practice effects, genuine improvement in clinical reasoning, or adaptation of writing style to match algorithmic preferences. Without independent validation and a comparison group, the relative contribution of each mechanism cannot be determined.

Table 2: Faculty Preceptor Intervention Summary Across 35 LLM-Generated Feedback Reports

Intervention Type	Frequency	% of Total Reports	Clinical Examples
No modification needed	21	60%	Feedback was clinically accurate, developmentally appropriate, and pedagogically sound
Algorithmic rigidity correction	8	23%	Reframed checklist compliance issues as learning opportunities; allowed for alternative clinical approaches
Clinical context addition	4	11%	Added relevant pathophysiology, epidemiology, or clinical significance
Tone and framing adjustment	2	6%	Softened language; emphasized progress and growth
Total Reports Reviewed	35	100%	All feedback underwent faculty review before transmission to students

The pattern of faculty interventions provides insight into the strengths and limitations of the LLM-based tutor. The high proportion of reports requiring no modification (60%) suggests that the algorithm successfully captured core clinical and pedagogical principles. However, the 40% of reports requiring intervention—particularly those addressing algorithmic rigidity (23%)—highlights the critical importance of Human-in-the-Loop supervision. These interventions prevented students from internalizing overly rigid algorithmic preferences and preserved the essential role of clinical judgment in medical practice.

3.2 Qualitative Findings: Focus Group Analysis

A focus group discussion was conducted with all five students at the conclusion of the intervention. The semi-structured interview explored student

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

perceptions of the LLM-based tutoring intervention. Thematic analysis of the focus group data revealed four primary themes.

Table 3: Thematic Analysis of Focus Group Discussion Data

Theme	Frequency	Interrater Agreement	Representative Quotes
Theme 1: Increased Confidence and Reduced Anxiety	5/5 students (100%)	κ=0.88	"I felt more confident submitting my notes knowing I would receive detailed, non-judgmental feedback." "The feedback helped me understand where I was falling short without making me feel like I failed." "By the end, I was actually looking forward to getting feedback instead of dreading it."
Theme 2: Perceived Value of Structured, Non-Punitive Feedback	5/5 students (100%)	κ=0.85	"I appreciated that the system reframed my mistakes as 'learning opportunities' rather than errors." "The Clinical Pearls were really helpful—they connected the feedback to actual clinical knowledge." "The positive framing made me want to improve rather than feel discouraged."
Theme 3: Concerns About Algorithmic Rigidity	4/5 students (80%)	κ=0.79	"Sometimes I felt like I was writing to please the algorithm rather than focusing on what was clinically important." "The checklist was very rigid—there wasn't room for clinical judgment or alternative approaches." "I appreciated that the professor sometimes softened the algorithm's feedback—it felt more human."
Theme 4: Perceived Loss of Human Interaction	4/5 students (80%)	κ=0.81	"I missed having a conversation with a preceptor about my clinical reasoning." "The feedback was helpful, but it can't replace the nuance of a real discussion with an experienced clinician." "I would have appreciated more opportunities to discuss my thinking

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

			directly with the professor."

All five students reported increased confidence and reduced anxiety when submitting clinical notes. Students appreciated the structured, non-punitive nature of the feedback and the emphasis on "Learning Opportunities" rather than errors. The insertion of Clinical Pearls was perceived as valuable for connecting feedback to broader clinical knowledge. However, four of five students expressed concerns about the perceived rigidity of the algorithm and the risk of adapting their writing style to match algorithmic preferences rather than focusing on genuine clinical reasoning. Additionally, four of five students noted a perceived loss of opportunity for verbal feedback and discussion with faculty preceptors, expressing concern about the reduction in human interaction.

Notably, students explicitly recognized and appreciated the faculty preceptor's role in moderating algorithmic feedback. One student commented: "I noticed that sometimes the professor's feedback was different from what the algorithm suggested—it felt more balanced and human. That made me trust the feedback more."

4. Discussion

4.1Summary of Findings and Study Positioning

This exploratory pilot study demonstrates the technical feasibility and preliminary acceptability of implementing an LLM-based clinical tutor for automated feedback on medical student clinical documentation. While preliminary findings suggest progressive improvement in clinical note scores and positive student perceptions regarding the pedagogical model, it is critical to emphasize that this study was not designed to establish causal efficacy or superiority over traditional methods. Rather, the primary contributions of this pilot study are: (1) demonstration of technical feasibility; (2) documentation of preliminary acceptability among students; (3) identification of perceived risks and limitations; (4) specification of adjustments needed for future implementation; and (5)

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

documentation of the critical role of faculty supervision in preventing algorithmic rigidity.

4.2 Interpretation in Context of Existing Literature

The observed progressive improvement in scores aligns with prior literature suggesting that structured, non-punitive feedback can enhance student engagement and learning outcomes [11]. The positive student perceptions regarding the pedagogical model—particularly the emphasis on "Learning Opportunities" and the inclusion of Clinical Pearls—are consistent with educational theory emphasizing the importance of formative feedback and positive reinforcement in medical education [12] [13]. The reduction in anxiety and increase in confidence reported by students is noteworthy and suggests that the intervention may have positive psychological effects, though these require validation in larger, more rigorous studies.

However, the concerns expressed by students regarding algorithmic rigidity and the risk of "teaching to the algorithm" align with broader concerns in the literature about the potential for automated systems to inadvertently incentivize superficial compliance rather than deep learning [14]. The perceived loss of human interaction is particularly important and underscores the critical role of faculty supervision in medical education. The finding that the faculty preceptor modified 40% of feedback reports suggests that algorithmic feedback, even when well-designed, requires substantial human oversight to ensure pedagogical appropriateness and to prevent unintended consequences.

4.3 Mechanisms of Observed Improvement

The observed improvement in scores may reflect multiple mechanisms, including: (1) increased familiarity with the evaluation checklist over repeated cycles; (2) practice effects inherent to longitudinal assessment; (3) genuine improvement in clinical reasoning skills; or (4) adaptation of writing style to match algorithmic preferences. Without independent validation by experienced clinicians and a comparison group (e.g., students receiving traditional feedback), the

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

relative contribution of each mechanism cannot be determined. It is possible that students learned to write in a manner that satisfied the algorithm's scoring criteria without necessarily developing deeper clinical reasoning skills. This distinction is critical for educational outcomes and highlights a key limitation of the current study.

4.4 Critical Limitation: Absence of Independent Validation

A critical limitation of this pilot study is the absence of independent validation of the LLM-generated scores. As the same LLM system that provided the feedback also generated the evaluation scores, it is not possible to distinguish whether improvements in student performance reflect genuine advancement in diagnostic reasoning or merely adaptation to the algorithm's scoring patterns. This represents a fundamental circularity in the study design that precludes causal inference.

To address this limitation in future studies, we recommend: (1) independent evaluation of a subset of clinical notes (e.g., 20%) by experienced clinicians using the same checklist in a blinded manner; (2) calculation of interrater agreement (Cohen's kappa or Intraclass Correlation Coefficient) between LLM scores and human expert scores; (3) use of validated instruments to assess diagnostic reasoning independent of documentation quality (e.g., Script Concordance Tests); and (4) comparison of the LLM-based feedback group with a control group receiving traditional faculty feedback or no structured feedback.

4.5 Role of Faculty Supervision in Mitigating Algorithmic Rigidity

The finding that the faculty preceptor modified 40% of LLM-generated feedback reports has important implications for implementation. While the high proportion of reports requiring no modification (60%) suggests that well-designed LLM prompts can produce pedagogically sound feedback, the 40% requiring intervention demonstrates that algorithmic feedback, even when well-intentioned, can inadvertently create rigidity or miss important educational nuances. The types of interventions made by the preceptor—particularly the correction of algorithmic

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

rigidity (23% of reports) and addition of clinical context (11% of reports)—highlight the critical importance of Human-in-the-Loop supervision.

These findings support the implementation of LLM-based tutoring only within a framework of rigorous faculty oversight. The preceptor's role is not merely to "spot-check" feedback but to actively curate and refine algorithmic output to ensure alignment with educational goals and prevention of unintended consequences. This level of supervision, while resource-intensive, appears necessary to realize the benefits of LLM-based feedback while avoiding its pitfalls.

Implications for Future Research and Practice

Despite these limitations, the pilot study provides preliminary evidence that LLM-based clinical tutoring is technically feasible and preliminarily acceptable to medical students, provided it is accompanied by rigorous faculty supervision. Future research should address the limitations identified in this study through: (1) larger, more diverse samples; (2) inclusion of comparison groups; (3) independent validation of LLM-generated scores; (4) assessment of impact on actual diagnostic reasoning and clinical outcomes; (5) longer-term follow-up to assess sustainability of improvements; (6) cost-benefit analysis of implementation; and (7) exploration of implementation strategies that optimize the benefits of automated feedback while preserving the essential human dimensions of medical education.

From a practice perspective, if LLM-based clinical tutoring is to be implemented in medical education, several safeguards are essential: (1) continuous faculty supervision (Human-in-the-Loop) to prevent algorithmic rigidity and to provide the nuanced feedback that only experienced clinicians can offer; (2) transparent communication with students about the limitations of automated feedback and the importance of human preceptor interaction; (3) regular evaluation of the system's performance and student outcomes; (4) robust data governance and privacy protections; (5) ongoing refinement of the evaluation checklist to ensure alignment with curricular objectives and clinical competencies; and (6) assessment of faculty workload and resource requirements for meaningful oversight.

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

5. Limitations

This exploratory pilot study has several important limitations that constrain the interpretation and generalizability of findings:

1. Small and Homogeneous Sample: The sample of five female students from a single institution is not representative of the broader medical student population. Findings may not generalize to male students, students from different institutions, or students at different levels of training.
2. Absence of Independent Validation: The primary methodological limitation is the absence of independent validation of LLM-generated scores. Without evaluation by experienced clinicians blinded to the study's purpose, it is impossible to determine whether score improvements reflect genuine advancement in clinical reasoning or merely adaptation to the algorithm's scoring criteria.
3. Lack of Comparison Group: The absence of a control group receiving traditional feedback or no structured feedback precludes comparison and limits causal inference. Improvements may reflect practice effects, increased familiarity with the checklist, or other factors unrelated to the LLM intervention.
4. Retrospective Observational Design: The study is susceptible to multiple sources of bias, including selection bias, measurement bias, and social desirability bias in qualitative responses.
5. No Assessment of Clinical Outcomes: The study did not assess whether improvements in documentation scores translated to improvements in actual diagnostic reasoning, clinical decision-making, or patient care outcomes.
6. Instrument Limitations: The 37-point checklist measures documentation quality and structural completeness rather than diagnostic reasoning or clinical judgment. Students may have adapted their writing to satisfy algorithmic preferences without developing deeper clinical reasoning skills.
7. Resource Requirements Not Assessed: The study did not include formal

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

cost-benefit analysis or assessment of faculty workload required for meaningful Human-in-the-Loop supervision. The 40% modification rate suggests substantial faculty time investment may be necessary.

8. Limited Qualitative Rigor: While thematic analysis was conducted systematically with interrater agreement assessment, the analysis was not independently verified by external evaluators.
9. Potential Confounding: Multiple factors may have contributed to observed improvements, including student motivation, prior exposure to similar feedback systems, and natural learning progression over the academic year.

6. Conclusion

This exploratory pilot study provides preliminary evidence that Large Language Model-based clinical tutoring is technically feasible and preliminarily acceptable to medical students, provided it is accompanied by rigorous faculty supervision. The intervention demonstrates potential for delivering structured, non-punitive feedback at scale, with positive effects on student confidence and anxiety. The critical finding that faculty preceptors modified 40% of LLM-generated feedback reports underscores the necessity of Human-in-the-Loop supervision to prevent algorithmic rigidity and ensure pedagogical appropriateness.

However, the study design does not permit conclusions about efficacy or superiority over traditional methods. Future research incorporating larger, more diverse samples; independent validation by experienced clinicians; comparison groups; and assessment of impact on actual diagnostic reasoning and patient outcomes is necessary to establish efficacy and to optimize implementation. Additionally, implementation strategies must prioritize continuous faculty supervision (Human-in-the-Loop) to mitigate risks of algorithmic rigidity, to preserve the essential human dimensions of medical education, and to ensure that students develop genuine clinical reasoning skills rather than merely adapting to algorithmic preferences.

The potential of LLMs to augment medical education is significant, but this potential can only be realized through rigorous research, transparent

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

acknowledgment of limitations, and unwavering commitment to maintaining the human dimensions of clinical teaching and learning. LLM-based tools should be understood as adjuncts to, not replacements for, expert faculty supervision and mentorship.

7. References

1 SCHUWIRTH, L. W. T.; VAN DER VLEUTEN, C. P. M. Challenges for educationalists. **BMJ**, v. 333, n. 7567, p. 639-641, 2006.

2 NORCINI, J.; BURCH, V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. **Med Teach**, v. 29, n. 9, p. 855-871, 2007.

3 RAJKOMAR, A. et al. Ensuring fairness in machine learning to advance health equity. **Ann Intern Med**, v. 169, n. 12, p. 866-872, 2018.

4 ESTEVA, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115-118, 2017.

5 SINGHAL, K. et al. Large language models encode clinical knowledge. **Nature**, v. 620, n. 7972, p. 172-180, 2023.

6 SAP, M.; GABRIEL, S.; QIN, L. et al. Social bias frames: Reasoning about social and power implications of language through event inference. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 58., 2020. **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**. [S. l.]: Association for Computational Linguistics, 2020. p. 5477-5490.

7 THIRUNAVUKARASU, A. J. et al. Large language models and their role in medicine. **Nat Med**, v. 29, n. 9, p. 2211-2218, 2023.

8 BLOOM, B. S. **Taxonomy of Educational Objectives: The Classification of Educational Goals**. New York: David McKay Company, 1956.

9 BRYSON, J. R.; ANDRES, L.; ATKINSON, R. The digital transformation of work: Disruption and the future of careers. **Futures**, v. 123, p. 102-632, 2020.

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

10 YEUNG, K. 'Hypernudges': Regulating AI for a better future. **Georget Law Technol Rev**, v. 2, p. 447-500, 2018.

11 HATTIE, J.; TIMPERLEY, H. The power of feedback. **Rev Educ Res**, v. 77, n. 1, p. 81-112, 2007.

12 NICOL, D. From monologue to dialogue: Improving written feedback processes in mass higher education. **J Assess Eval High Educ**, v. 35, n. 5, p. 501-517, 2010.

13 WIGGINS, G. P. **Educative Assessment**: Designing Assessments to Improve Student Performance. San Francisco: Jossey-Bass, 1998.

14 GOODHART, C. M. Problems of monetary management: The UK experience. In: COURAKIS, A. S. (ed.). **Inflation, Depression and Economic Policy in the West. Rowman & Littlefield**, 1981. p. 111-146.

APPENDIX A: Clinical Note Evaluation Checklist (37-Point Scale)

The following checklist was used to evaluate all 35 clinical notes. Items are organized by semiological domain, with point allocations reflecting pedagogical priorities for fourth-year medical students. This checklist was developed through review of standard semiological teaching guidelines and consensus among three experienced clinicians (mean experience: 18 years). The overall Content Validity Index (CVI) was 0.89, indicating excellent content validity.

Table A: Clinical Note Evaluation Checklist - 37-Point Scale with Original and Revised Weightings

Domain	Item #	Component	Points	Justification	Original	Revised
Identification & Chief Complaint	1	Patient demographics (age, sex, occupation)	1	Minimal complexity; standardized routine questions	1	1
	2	Chief complaint clearly stated	1	Minimal complexity; establishes clinical focus	1	1
	3	Duration of chief complaint documented	1	Minimal complexity; temporal	1	1

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

				context		
History of Present Illness	4-13	Detailed HPI including: onset, character, associated symptoms, severity, timing, context, alleviating/aggravating factors, impact on function	10	High complexity; cornerstone of diagnostic reasoning; reduced from 30 to prevent algorithmic rigidity	30	10
Past Medical History	14-18	Relevant chronic conditions, prior surgeries, hospitalizations, medications, allergies	5	Moderate complexity; contextual information supporting but not driving diagnostic formulation	5	5
Family History	19-21	Relevant hereditary conditions, family psychiatric history, social support	3	Moderate complexity; identifies hereditary and psychosocial risk factors	2	3
Social History	22-28	Occupation, living situation, social support, substance use, sexual history, recent stressors, health literacy	7	Increased complexity; social determinants directly influence clinical presentation and diagnostic formulation	2	7
Review of Systems	29-33	Systematic inquiry across major organ systems; documentation of positive and negative findings	5	Moderate complexity; comprehensive assessment supporting diagnostic reasoning	5	5
Vital Signs & General Appearance	34-35	Vital signs (BP, HR, RR, Temp, O2 sat) documented; general appearance and nutritional status noted	2	Lower complexity; objective clinical data supporting but not substituting for diagnostic reasoning	2	2
TOTAL			**37**		**49**	**37**

Rationale for Checklist Revision: The checklist weights reflect the level of complexity and diagnostic importance inherent to each semiological domain. The original checklist (49 points) weighted the History of Present Illness (HPI) at 30 points (61% of total), creating algorithmic rigidity by overemphasizing biomedical

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

detail accumulation at the expense of clinical judgment and diagnostic synthesis.

The revised checklist (37 points) redistributes weights according to diagnostic complexity and clinical reasoning demand: Identification & Chief Complaint (3 points, 8%) represents minimal complexity with standardized, routine questions. Past Medical History (5 points, 14%) and Review of Systems (5 points, 14%) provide moderate complexity with necessary contextual information. Family History (3 points, 8%) identifies hereditary and psychosocial risk factors. Social History (7 points, 19%) represents increased complexity as social determinants directly influence clinical presentation. History of Present Illness (10 points, 27%) represents high complexity as the cornerstone of diagnostic reasoning. Vital Signs & General Appearance (2 points, 5%) provide lower complexity objective clinical data. Diagnostic Impression (2 points, 5%) represents the highest complexity as the synthesis of all preceding information.

This weighting schema prioritizes diagnostic complexity and clinical reasoning over documentation completeness, cultivating genuine clinical judgment and multisystemic thinking rather than algorithmic compliance.

APPENDIX B: Focus Group Interview Guide

Semi-Structured Interview Guide for Student Focus Group Discussion

Opening Question: Tell me about your experience receiving feedback from the LLM-based clinical tutor over the past seven weeks. What was helpful? What was challenging?

Thematic Questions:

1. Confidence and Anxiety:

- How did you feel about submitting your clinical notes for evaluation? Did your feelings change over time?
- Did the feedback affect your confidence in your clinical documentation skills?

REMUNOM
ISSN 2178-6925

Received: 08/05/2026 - **Accepted:** 15/07/2026
Vol: 13.16
DOI: 10.66104/1wjgbn62
Pages: 1-24

2. Feedback Quality and Pedagogy:

- What aspects of the feedback were most helpful to you?
- How did the 'Learning Opportunities' framing compare to traditional error-focused feedback?
- Were the 'Clinical Pearls' useful? Can you give an example?

3. Algorithmic Rigidity:

- Did you ever feel like the feedback was too rigid or didn't account for clinical judgment?
- Did you ever feel pressure to write in a certain way to 'please the algorithm'?

4. Human Interaction:

- How did you feel about receiving feedback from an algorithm rather than directly from a professor?
- Did you miss having a conversation with a preceptor about your clinical reasoning?

5. Faculty Supervision:

- Did you notice that sometimes the professor's feedback was different from the algorithm's feedback?
- How did that affect your perception of the feedback?

6. Overall Recommendations:

- If we were to use this system again, what would you change?
- Would you recommend this approach to other students? Why or why not?

Closing Question: Is there anything else about your experience with the LLM-based tutor that you'd like to share?