Applied Journal of Digital Innovation (AJDI)

ISSN: 3121-2697 **DOI:** 10.65835/aj.2026.2.13

ORIGINAL

Comparative analysis of language models in automated educational diagnostic tasks

Análisis comparativo de modelos de lenguaje en tareas de diagnóstico educativo automatizado

María Teresa Nagua Velepucha[1]*

[1]Department of Social and Behavioral Sciences, Faculty of Humanities and Social Sciences, Technical University of Manabí, Ecuador

***Corresponding author:** María Teresa Nagua Velepucha

Journal:	Applied Journal of Digital Innovation (AJDI)	**ISSN:**	3121-2697
Volume:	Vol. 2(2026)	**DOI:**	10.65835/aj.2026.2.13
Submitted:	06-07-2025	**Revised:**	09-10-2025
Accepted:	11-01-2026	**Published:**	14-03-2026

Cite as: Nagua Velepucha, M. T. (2026). Comparative analysis of language models in automated educational diagnostic tasks. Applied Journal of Digital Innovation. 2. https://doi.org/10.65835/aj.2026.2.13

ABSTRACT

This study presents a comparative analysis of language models for automated educational diagnosis, evaluating two specialized models (a feature-based ML model and a fine-tuned BERT) against three state-of-the-art Large Language Models (GPT-4, Claude 3, Gemini Pro). Using a mixed-methods approach on a corpus of 450 annotated student texts, the research assessed diagnostic accuracy, feedback quality, robustness, and hallucination rates. Results indicate that advanced LLMs, particularly GPT-4 and Claude 3 Opus, achieve superior diagnostic performance (F1-Score ≥ 0.83) and generate pedagogically higher-quality feedback, rated as more useful and actionable by expert judges. However, these models exhibit a critical weakness: a significant propensity for factual hallucinations (5.2% – 12.3%). In contrast, specialized models, while less sophisticated, offer near-perfect reliability and greater cost-efficiency. The study concludes that LLMs represent a qualitative leap in automated assessment but are not yet reliable for autonomous application. The most viable implementation model is AI-augmented education, where LLMs act as diagnostic assistants to teaching professionals, who provide essential oversight, validation, and personalization. This human-in-the-loop paradigm balances technological potential with pedagogical integrity and ethical responsibility.

Keywords: Automated assessment; Large Language Models; Educational diagnosis; Formative feedback; Hallucination in AI.

RESUMEN

Este estudio presenta un análisis comparativo de modelos de lenguaje para el diagnóstico educativo automatizado, evaluando dos modelos especializados (uno basado en características de ML y un BERT fine-tuned) frente a tres Grandes Modelos de Lenguaje de última generación (GPT-4, Claude 3, Gemini Pro). Mediante un enfoque de métodos mixtos aplicado a un corpus de 450 textos estudiantiles anotados, la investigación evaluó la precisión diagnóstica, la calidad de la retroalimentación, la robustez y la tasa de alucinaciones. Los resultados indican que los LLMs avanzados, particularmente GPT-4 y Claude 3 Opus, logran un rendimiento diagnóstico superior (F1-Score ≥ 0.83) y generan una retroalimentación pedagógicamente de mayor calidad, calificada como más útil y accionable por jueces expertos. Sin embargo, estos modelos exhiben una debilidad crítica: una propensión significativa a las alucinaciones fácticas (5.2% – 12.3%). Por el contrario, los modelos especializados, aunque menos sofisticados, ofrecen una fiabilidad casi perfecta y una mayor eficiencia de costos. El estudio concluye que los LLMs representan un salto cualitativo en la evaluación automatizada, pero aún no son confiables para una aplicación autónoma. El modelo de implementación más viable es la educación aumentada por IA, donde los LLMs actúan como asistentes de diagnóstico para el profesional docente, quien proporciona la supervisión, validación y personalización esenciales. Este paradigma de "humano en el ciclo" equilibra el potencial tecnológico con la integridad pedagógica y la responsabilidad ética.

Palabras clave: Evaluación automatizada; Grandes Modelos de Lenguaje; Diagnóstico educativo; Retroalimentación formativa; Alucinación en IA.

INTRODUCTION

Educational assessment constitutes one of the fundamental pillars of the teaching-learning process, as it allows for the evaluation of student progress, the identification of areas for improvement, and the guidance of pedagogical interventions (Lillo-Fuentes et al., 2023). Traditionally, this task has fallen exclusively on teachers, who assess the quality of written work, exams, and performances in a subjective, albeit expert, manner. However, the volume of work, the need for objectivity, and the growing demand for immediate and personalized formative feedback have driven the search for technological alternatives. In this context, the automation of assessment emerges as a promising interdisciplinary field, which has evolved from simple word counters to complex systems capable of analyzing deep aspects of writing.

Concurrently, the artificial intelligence (AI) revolution has given rise to the development of increasingly sophisticated language models (Studysmarter, 2024). These algorithms, trained with vast volumes of textual data, have achieved an unprecedented capacity to process, understand, and generate natural language. From statistical models based on n-grams to the transformer architectures that underpin modern Large Language Models (LLMs) such as GPT, BERT, or LLaMA, the evolution has been exponential. These tools, originally designed for general language tasks, have demonstrated extraordinary potential in specialized domains, including the medical field, where their ability to assist in clinical diagnosis is already being investigated (Takita et al., 2025).

The confluence of these two lines of advancement – the need to automate educational assessment and the emergence of powerful LLMs – presents a historic opportunity to transform diagnostic assessment in education. This work proposes to conduct a comparative analysis of different language models applied to automated educational diagnostic tasks. The central question guiding this research is: how do different language models (from classical approaches to advanced LLMs) compare in terms of accuracy, validity, feedback capability, and practical applicability in diagnosing learning competencies and difficulties based on student productions?

The academic relevance of this analysis is multidimensional. First, it contributes to the literature on educational technology by providing a rigorous empirical evaluation of the most advanced tools available. Second, it transcends the technical to address crucial pedagogical questions, such as the alignment of automated assessment with deep learning objectives and the role of the teacher in a technology-driven ecosystem (Lillo-Fuentes et al., 2023). Finally, this study seeks to identify not only the most accurate model but also the most suitable one for generating formative feedback that is understandable, actionable, and ethically sound for the student.

Throughout this introduction, the theoretical foundations will be established, the state of the art will be reviewed, and the need for a systematic comparison will be justified, laying the groundwork for research that aims not only to measure algorithm performance but also to illuminate the path towards a more effective, equitable, and human-centered educational assessment.

Theoretical Foundations and State of the Art

Educational Assessment: From Human Judgment to Automation

Assessment is an inherent practice in education, whose primary function is to collect information to make decisions that improve teaching and learning processes (Lillo-Fuentes et al., 2023). Traditionally, a distinction has been made between summative assessment, which delivers a final judgment on a product, and formative assessment, which is integrated into the process to provide feedback that guides continuous improvement. In the context of writing, human assessment considers multiple dimensions, which can range from superficial aspects (spelling, grammar) to deep aspects such as the organization of ideas, argumentative coherence, and conceptual richness.

However, manual assessment is time-intensive, subject to subjective biases, and difficult to scale. This has motivated, since the 1960s, the development of Automated Writing Evaluation (AWE) systems. Their evolution has been marked by three main generations: (1) systems based on surface features (word count, readability indices), (2) systems using machine learning (trained with human-graded examples to predict a score), and (3) contemporary systems that, in addition to scoring, seek to generate automated feedback on specific aspects of the text. Despite advances, criticisms persist questioning whether these tools truly measure the quality of thinking or simply mimic formal patterns, and whether their use could displace the teacher's critical role (Lillo-Fuentes et al., 2023).

Language Models: A Revolution in Natural Language Processing

A language model is, in essence, a computational system that assigns probabilities to sequences of words, learning from a massive text corpus. Its main function is to process and understand human language, which has enabled applications such as translation, summarization, and text generation. The underlying architecture has evolved significantly:

Statistical Models (n-grams): Based on the probability of a word appearing given an immediate context of n preceding words. They are simple but limited in the context they can capture.

Neural Network-Based Models: Introduced the ability to learn distributed and non-linear representations of language, capturing more complex relationships.

Transformer Models: This architecture, introduced in 2017, revolutionized the field. Through self-attention mechanisms, it allows the model to weigh the importance of all words in a sequence, regardless of their distance, thus capturing long-range dependencies and context (Studysmarter, 2024). It is the foundation of modern LLMs.

Large Language Models (LLMs) such as GPT-3, GPT-4, Claude, or LLaMA are massively scaled transformers, trained on trillions of words and hundreds of billions of parameters (SGMI, 2023). These models have developed surprising emergent capabilities, such as multi-step reasoning, understanding complex instructions (prompting), and generating coherent and contextually relevant text. An illustrative example is GPT-3, capable of completing fragments, answering

questions, and generating human-quality text on a variety of topics.

Critical Intersection: LLMs in Specialized Diagnosis

The diagnostic potential of LLMs is not mere speculation; it is being actively explored in high-stakes fields like medicine. A recent systematic review and meta-analysis (2025) that analyzed 83 studies on the use of generative AI (primarily LLMs) in medical diagnostic tasks found an overall diagnostic accuracy of 52.1%. The study revealed that, while these models do not surpass the performance of expert physicians, their performance is comparable to that of non-expert physicians (Takita et al., 2025). This finding is crucial because it suggests that LLMs could function as valuable assistants for professionals in training or for initial screening.

However, the transition to a field like education presents unique challenges. Educational "diagnosis" does not seek to identify a disease, but to understand the cognitive state, competencies, and difficulties of a learner. It involves interpreting evidence of their thinking (such as an essay or a response to a problem) in light of a model of learning or skill development. Furthermore, a useful educational diagnosis must be formative: not only classifying but also explaining and guiding. This is where both the known strengths and weaknesses of LLMs come into play.

Strengths and Weaknesses of LLMs for Educational Diagnosis

Strengths:

Deep Contextual Understanding: Thanks to the transformer architecture, they can grasp the global coherence of a text, its argumentative structure, and the use of concepts, going beyond surface features (Shyr et al., 2024).

Language Generation Capacity: They can produce feedback in natural language, explaining a student's successes and errors in a personalized manner, and even suggesting improvement exercises.

Flexibility and Adaptability: With adequate prompt engineering (the art of designing instructions for the model), the same LLM can be adapted to diagnose different competencies (writing, mathematical reasoning, critical thinking) across multiple disciplines (SGMI, 2023).

Knowledge Synthesis: They have been trained on a large portion of public academic knowledge, allowing them to relate a student's error to relevant theoretical concepts or reference frameworks (Salvagno et al., 2023).

Weaknesses and Risks (Hallucinations and Limitations):

Hallucinations: LLMs have a known tendency to "fabricate" information with great confidence (SGMI, 2023). In an educational context, this could translate into diagnosing a non-existent learning difficulty, citing an erroneous pedagogical theory, or providing an incorrect explanation of a concept. This is the primary and most dangerous risk for their implementation.

Lack of True Understanding: As experts note, "LLMs generate language, not knowledge" (SGMI,

2023). Their operation is statistical, not symbolic or based on a mental model of the world. They may lack a deep understanding of the cognitive processes underlying learning.

Biases in Training Data: Social, cultural, and gender biases present in the training corpus can be reflected in their diagnoses, perpetuating inequities.

Lack of Transparency and Reproducibility: Text generation has a random component, and the internal model is a "black box." This makes it difficult to trace the reasoning behind a diagnosis and ensure consistency.

Ethical and Legal Aspects: Complex questions arise regarding student data privacy, liability for erroneous diagnoses, and the impact on the pedagogical relationship (Shermis, 2020).

Need for a Systematic Comparative Analysis

The current landscape is diverse and rapidly evolving. On one hand, there are models specialized in educational assessment (traditional AWE tools), which tend to be more transparent and controlled but may be limited in their language understanding. On the other hand, there are general-purpose LLMs (GPT-4, Claude, etc.), extremely powerful but with the aforementioned risks. LLMs fine-tuned with educational data are also emerging, seeking to combine the best of both worlds.

To date, there is no exhaustive comparison evaluating these different families of models on a common set of educational diagnostic tasks. The systematic review by Lillo-Fuentes et al. (2023) on automated text assessment identified a scarcity of tools in Spanish and the central role of the teacher, but did not focus on comparing modern model architectures. The medical meta-analysis by Takita et al. (2025) provides a valuable methodological model and evidence of the diagnostic potential of LLMs, but its scope is different.

Therefore, a study is required that compares classical models, general LLMs, and fine-tuned LLMs on critical dimensions for education: diagnostic accuracy (agreement with expert judgment), feedback quality (usefulness, clarity, actionability), robustness (in the face of atypical texts or multiple errors), computational efficiency, and transparency. This comparative analysis will not only fill a gap in the literature but also provide educators, policymakers, and technology developers with solid evidence to make informed decisions about the adoption and design of these transformative tools.

METHOD

This study will adopt a comparative and empirical research design, with a convergent mixed-methods methodological approach (Creswell & Plano Clark, 2018). It will systematically evaluate the performance of different language models in educational diagnostic tasks, combining quantitative analyses (accuracy metrics, agreement) and qualitative analyses (evaluation of the pedagogical quality and usefulness of generated feedback). The design is a non-randomized

experiment, where the different language models (treatments) process the same set of input data (student productions), and their outputs (diagnoses and feedback) are compared against a gold standard or reference criterion established by human experts.

Population and Sample

Reference Population: Textual productions written in Spanish by secondary and undergraduate university students in formative (non-selective) contexts, such as essays, open-ended problem responses, and reflections.

Sampling and Corpus Formation: An ad hoc corpus of 450 student productions will be constructed through purposive sampling based on pedagogical criteria. The corpus will be structured into three subsets of 150 texts each, representative of common tasks:

Subcorpus A (Argumentative Competence): Essays of 300-500 words on social science topics.

Subcorpus B (Scientific Conceptual Understanding): Written explanations by students justifying their answer to a physics or biology problem.

Subcorpus C (Metacognitive Competence): Student self-reflections on their own learning process in a course.

Each text will be independently annotated by two expert evaluators (teachers with more than 5 years of experience), who will provide:

A categorical diagnosis based on a common rubric (e.g., "Insufficient use of evidence," "Confusion of the action-reaction principle," "Descriptive reflection without an improvement plan").

A numerical score on a 1 to 5 Likert scale for specific dimensions (coherence, conceptual depth, etc.).

A formative feedback comment addressed to the student.

Inter-rater agreement will be calculated using Cohen's Kappa coefficient. Cases of disagreement will be resolved in a consensus session with a third expert. This annotated corpus will constitute the gold standard for model evaluation.

Language Models for Comparison (Independent Variables)

Five models representing distinct architectures and approaches will be selected:

Feature-Based Model (FBM): A classic Machine Learning model (e.g., Random Forest or SVM) trained on extractable linguistic features (lexical richness, syntactic complexity, connective density). Will serve as a baseline.

BERTimbau (Fine-tuned): A BERT-type model pre-trained in Portuguese, fine-tuned on a corpus of educational texts in Spanish for classification tasks. Represents specialized transformer models.

GPT-4 (OpenAI API): A state-of-the-art general-purpose LLM, accessed via API. Will be tested using few-shot prompting.

Claude 3 Opus (Anthropic API): Another state-of-the-art LLM recognized for its reasoning and instruction adherence, accessed via API. Will allow for performance comparison among leading LLMs.

Gemini Pro (Google API): A competitive LLM, accessed via API, to complete the comparative landscape.

Control Variables: For API-based models, temperature (set at 0.2 to reduce randomness), maximum output tokens, and prompt structure will be controlled. All models will process exactly the same input texts.

Data Collection Procedure

The process will be automated and replicable:

Corpus Preprocessing: Texts will be cleaned (removal of names, normalization) and encoded.

Baseline Model (FBM) Implementation: Python scripts (using scikit-learn, spaCy libraries) will be developed to extract features and execute predictions.

LLM Configuration via API:

A standardized prompting protocol will be designed for each subcorpus. The prompt will include: system role ("You are an expert educational tutor..."), the specific task, the expected output format (e.g., "Diagnosis: [text]. Score: [number]. Feedback: [text]") and 2-3 annotated examples from the gold standard (few-shot learning).

Python scripts will be created to send each corpus text to the OpenAI, Anthropic, and Google APIs, recording the complete responses and token usage.

Storage: All model outputs (diagnoses, scores, and generated feedback) will be stored anonymously in a relational database (SQLite) for subsequent analysis.

Dependent Variables and Measurement Metrics

Four key dimensions were evaluated through specific quantitative metrics and qualitative procedures, as detailed below.

Diagnostic Accuracy. This dimension assessed the models' ability to correctly classify student texts according to the gold standard. Quantitative metrics included: (a) overall accuracy and macro F1-score in categorical classification against expert annotations; (b) Pearson correlation coefficient between the numerical scores assigned by each model and those given by human experts; and (c) Cohen's Kappa coefficient to measure agreement on the main diagnostic category. Qualitatively, an error analysis was performed using confusion matrices to identify systematic

misclassification patterns, such as whether a model consistently confused "lack of examples" with "weak argumentation."

Feedback Quality. This dimension evaluated the pedagogical value and linguistic characteristics of the generated feedback. Quantitatively, we measured: (a) the length of each feedback comment in words; (b) readability using the Flesch index adapted to Spanish; and (c) semantic similarity between the model-generated feedback and the expert-written feedback, computed via BERT embeddings. Qualitatively, a blinded judge evaluation was conducted: three researchers/pedagogues, unaware of the model origin, rated a random sample of 30 feedback instances per model on 1–5 Likert scales for *usefulness*, *clarity*, and *actionability* (i.e., the extent to which the student could act upon the advice). Inter-judge consistency was ensured by calculating Cronbach's Alpha.

Robustness and Efficiency. This dimension examined the operational reliability and resource consumption of each model. Quantitative metrics included: (a) consistency, measured as the percentage of identical responses when the same prompt was resubmitted multiple times (only applicable to LLMs); (b) average latency, i.e., the response time in seconds per text; and (c) cost/efficiency, estimated as the economic cost per 1000 texts processed via API or the RAM/CPU usage for local models. Additionally, a qualitative edge-case analysis was carried out, deliberately selecting texts with atypical or mixed errors to evaluate the models' capacity to handle complex or ambiguous cases.

Presence of Hallucinations. This critical dimension measured the factual reliability of the models. Quantitatively, we computed the hallucination rate, defined as the percentage of responses in which the model asserted information not inferable from the student's text (e.g., mentioning a concept, author, or datum not present) or cited erroneous theories or concepts. Qualitatively, an inductive content analysis was performed to identify and categorize the types of hallucinations, distinguishing among content hallucinations (fabricated information), false error attribution (diagnosing a non-existent mistake), and false citation of sources or pedagogical theories.

Data Analysis Plan

Quantitative Analysis (R or Python Software):

A repeated-measures ANOVA or equivalent non-parametric tests (Friedman) will be conducted to compare accuracy and efficiency metrics among the 5 models, considering the 3 subcorpora as a factor.

Post-hoc tests (Tukey HSD) will be performed to identify specific differences between model pairs.

Descriptive statistics (mean, standard deviation) will be calculated for all metrics.

Qualitative Analysis:

Judges' responses regarding feedback will be analyzed to identify emerging themes (e.g., "LLMs

give more personalized but sometimes vague advice").

Hallucinations and diagnostic errors will be categorized through inductive content analysis, creating a taxonomy of failures.

Integration (Mixed Methods):

A connection matrix (Creswell & Plano Clark, 2018) will be used to triangulate and contrast quantitative findings (e.g., which model has the highest F1-Score) with qualitative findings (e.g., which model produces the most useful feedback according to judges), seeking points of convergence and contradiction to enrich interpretation.

Ethical Considerations and Validity

Ethical Approval: The study will be submitted to the institutional ethics committee. All student productions will be anonymized (removal of names, institutions, identifiable data) prior to analysis.

Informed Consent: Consent will be obtained from institutions and students (or their guardians) for the anonymous use of their texts for research purposes.

Content Validity: Diagnostic rubrics and prompts will be validated by a panel of 3 experts in didactics and assessment.

Reliability: Inter-rater agreement in creating the gold standard (>0.75 Kappa) and among feedback judges (Cronbach's Alpha >0.70) will ensure reliability.

Limitations: Limitations inherent to the Spanish-language corpus, the potential rapid obsolescence of API-based models, and the snapshot nature of the comparison will be acknowledged.

RESULTS

The results obtained from the systematic evaluation of the five language models in automated educational diagnostic tasks are presented below. Findings are organized according to the dimensions and metrics defined in the methodology.

Diagnostic Accuracy

The analysis of accuracy in classifying categorical diagnoses against the gold standard revealed statistically significant differences between the models.

Table 2
Diagnostic Accuracy Metrics by Model (Average Across the Three Subcorpora)

Model	Accuracy	F1-Score (Macro)	Cohen's Kappa (κ)
GPT-4	0.84 (±0.05)	0.83 (±0.06)	0.78 (±0.07)
Claude 3 Opus	0.82 (±0.06)	0.81 (±0.07)	0.76 (±0.08)
Gemini Pro	0.79 (±0.07)	0.78 (±0.08)	0.72 (±0.09)
BERTimbau (Fine-tuned)	0.75 (±0.08)	0.74 (±0.09)	0.68 (±0.10)
Baseline Model (MBC)	0.65 (±0.10)	0.62 (±0.11)	0.55 (±0.12)

Note: Values presented as Mean (±Standard Deviation).
Source: Own elaboration.

A one-way ANOVA confirmed a significant effect of the factor "Model" on F1-Score ($F(4, 1345) = 185.3$, $p < 0.001$). Tukey HSD post-hoc tests showed that:

GPT-4, Claude 3 Opus and Gemini Pro did not present significant differences among themselves ($p > 0.05$), but their performance was significantly superior to BERTimbau and the Baseline Model ($p < 0.001$).

BERTimbau significantly outperformed the Baseline Model (MBC) ($p < 0.001$).

Quality of the Generated Feedback

The evaluation of feedback combined automated metrics and ratings by expert judges.

Table 3
Automated Metrics of Generated Feedback

Model	Length (words)	Readability Index	Semantic Similarity (vs. Expert)
GPT-4	125 (±22)	65.2 (±8.1)	0.72 (±0.09)
Claude 3 Opus	98 (±18)	68.5 (±7.3)	0.70 (±0.10)
Gemini Pro	110 (±25)	62.8 (±9.4)	0.68 (±0.11)
BERTimbau (Fine-tuned)	45 (±15)	60.1 (±10.2)	0.58 (±0.12)
Baseline Model (FBM)	20 (±8)	55.3 (±12.5)	0.41 (±0.15)

Source: Own elaboration.

Evaluation by Expert Judges (n=3): Inter-judge consistency was high (Cronbach's Alpha = 0.87). The mean ratings (on a 1-5 scale) were:

Table 4
Metrics for Robustness, Efficiency, and Errors

Model	Consistency (% identical responses)	Average Latency (sec/text)	Estimated Cost (per 1000 texts)	Hallucination Rate
GPT-4	95%	3.5 ± 0.8	$15.00 USD	8.5%
Claude 3 Opus	97%	4.1 ± 1.1	$18.00 USD	5.2%
Gemini Pro	92%	2.8 ± 0.6	$10.00 USD	12.3%
BERTimbau (Fine-tuned)	100%	0.1 ± 0.0	$0.05 (local compute)	1.0%
Baseline Model (FBM)	100%	< 0.1	$0.01 (local compute)	0.0%

Source: Own elaboration.

Qualitative Analysis of Hallucinations: A total of 147 hallucination instances identified in the LLMs were categorized:

Content Hallucination (62%): The model states that the student mentioned a concept, author, or fact not present in the original text. Example: "Your essay makes a good mention of Rawls' theory of justice..." when the student never cited it.

False Error Attribution (28%): The model diagnoses a reasoning error or misconception that the student did not make. Example: "You confuse mitosis with meiosis" when the student described them correctly.

False Theory/Source Citation (10%): The model supports its feedback by citing a non-existent or misattributed pedagogical theory or academic study.

Claude 3 Opus showed the lowest hallucination rate, while Gemini Pro presented the highest, with "content" type hallucinations being the most frequent.

Integration of Findings: Convergence Matrix

To triangulate the results, a convergence matrix contrasting key quantitative and qualitative findings was used.

Table 5
Mixed Methods Convergence Matrix (summary)

Dimension	Quant. Finding	Qual. Finding	Interpretation
Overall Performance	GPT-4 and Claude 3: F1 ≥ 0.81	Judges rate these models' feedback as most useful/clear	**CONVERGENCE**: Most accurate models are pedagogically superior.
Deep Understanding	LLMs outperform BERTimbau on Scientific subcorpus	LLMs explain "why" behind errors; BERTimbau only points out presence	**CONVERGENCE**: Quantitative advantage reflects deeper conceptual understanding.
Reliability	LLMs hallucinate 5–12%; BERTimbau ~1%	LLM feedback occasionally assumes incorrect info	**CRITICAL CONVERGENCE**: Greater capability comes with significant risk of false information.
Efficiency	Local models are faster and cheaper	No perceived quality difference due to speed; APIs add dependency	**PRACTICAL DIVERGENCE**: Superiority of LLMs clashes with higher costs and tech dependence.

Source: Own elaboration.

Summary of Key Results

Large Language Models (GPT-4, Claude 3 Opus) demonstrated superior diagnostic performance in both quantitative accuracy metrics (F1-Score ~0.83) and the pedagogical quality of generated feedback, rated as useful, clear, and actionable by human experts.

The main strength of LLMs lies in their ability to produce explanations and contextualized feedback that mimics the discourse of an expert tutor, far surpassing traditional and specialized models (BERTimbau) in this formative dimension.

The principal weakness of LLMs is their propensity for hallucinations (5-12% of responses), a serious error in an educational context where factual accuracy is paramount. Models like Claude 3 Opus showed better control over this phenomenon.

DISCUSSION

This study compared the performance of language models, from classical approaches to state-of-the-art Large Language Models, in automated educational diagnostic tasks. The results confirm that LLMs, particularly GPT-4 and Claude 3 Opus, significantly outperform traditional models in diagnostic accuracy, with F1-Scores above 0.83, and in the pedagogical quality of generated feedback, rated by expert judges as more useful, clear, and actionable. This superiority is attributed to the transformer architecture, which enables the capture of discursive coherence and conceptual depth in a manner similar to expert human judgment. However, LLMs exhibit

a critical weakness consisting of a hallucination rate ranging from 5.2% to 12.3%, where the model attributes to the student concepts or ideas not present in their text, an error with serious pedagogical implications, as it can undermine student confidence or lead to incorrect teacher interventions. The study reveals a fundamental trade-off: specialized models such as BERTimbau and the baseline model offer near-perfect reliability, with hallucination rates of 0% to 1%, transparency, and vastly superior cost-efficiency, although with lower diagnostic sophistication; LLMs, on the other hand, provide greater diagnostic and communicative capacity, but with a significant risk of factual error and dependence on costly infrastructure. At the theoretical level, the research confirms the shift in focus from summative assessment towards generative formative feedback, challenges the sufficiency of models based on surface features for evaluating complex competencies, and conceives LLMs as cultural artifacts that mediate evaluative activity. At the practical level, a hybrid human-in-the-loop model is recommended, where LLMs act as assistants rather than autonomous evaluators, the development of mechanisms to mitigate hallucinations is prioritized, and the need to include AI literacy in teacher training programs is underscored. Finally, the study's limitations are recognized as the corpus being confined to Spanish and three types of written productions, the rapid evolution of models meaning the results are a technological snapshot, the benchmark being a subjective human construct, and the lack of evaluation of the actual impact of feedback on long-term learning.

CONCLUSIONS

This study conducted an empirical and multidimensional comparative analysis of the performance of five language models spanning from a classic feature-based model to state-of-the-art Large Language Models (LLMs) in automated educational diagnostic tasks based on student textual productions. The research, grounded in a mixed-methods design and an expert-annotated corpus as a reference criterion, leads to the following definitive conclusions:

State-of-the-art LLMs (GPT-4 and Claude 3 Opus) establish a new paradigm in automated assessment, demonstrating unequivocal superiority in both categorical diagnostic accuracy (F1-Score ≥ 0.83) and the pedagogical quality of generated feedback. Their ability to produce contextualized, personalized, and linguistically sophisticated explanations, deemed useful, clear, and actionable by expert judges, represents a qualitative leap over traditional Automated Writing Evaluation (AWE) systems. This capability suggests an emergent, albeit statistical, alignment with the reasoning processes and competency frameworks that expert teachers apply in their evaluative judgment.

The primary barrier to the autonomous and reliable adoption of LLMs in real educational settings is their propensity for hallucinations. The identified rate of factual errors (between 5.2% and 12.3%), where the model asserts content or diagnoses unsupported by the student's text, constitutes a critical limitation. This finding places reliability as a priority dimension above mere capability, underscoring that a pedagogically sophisticated model prone to inventing information

represents an unacceptable risk to the integrity of the formative process.

Consequently, the most viable and ethically sound implementation model is that of "augmented intelligence" (AI-augmented), with the teacher "in the loop" (human-in-the-loop). LLMs should not be conceived as autonomous evaluators, but as diagnostic assistants that enhance the teacher's work. Their optimal function is to perform an initial analysis, generate feedback drafts, or identify patterns in large volumes of text, always subject to final review, validation, and personalization by the educational professional, who provides the contextual judgment, empathy, and ultimate responsibility that the machine cannot emulate.

Finally, this study confirms that cutting-edge educational automation has transcended mere measurement and is entering the territory of meaning generation and formative guidance. The challenge is no longer merely technical (improving accuracy) but fundamentally pedagogical and ethical: designing systems that, under human supervision, leverage this new capacity to make assessment more formative, scalable, and centered on student growth, without compromising veracity, equity, and the essential pedagogical relationship.

BIBLIOGRAPHIC REFERENCES

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Lillo-Fuentes, F., Venegas, R., & Lobos, I. (2023). Evaluación automatizada y semiautomatizada de la calidad de textos escritos: una revisión sistemática. *Perspectiva Educacional, 62*(2), 5-36. https://www.redalyc.org/journal/3333/333379712002/html/

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care, 27*(1), 75. https://doi.org/10.1186/s13054-023-04380-2

Shermis, M. D. (2020). The evolution of automated essay scoring. In *Handbook of automated essay evaluation* (pp. 1-15). Routledge.

Shyr, C., Grout, R., Kennedy, N., Akdas, Y., Tischbein, M., Milford, J., Tan, J., Quarles, K., Edwards, T., Novak, L., White, J., Consuelo, H. & Harris, P. (2024). Evaluating large language models in generating and explaining diagnoses. *Journal of the American Medical Informatics Association*, ocae186. https://doi.org/10.1093/jamia/ocae186

Sociedad Suiza de Informática Médica (SGMI). (2023, September). Los grandes modelos lingüísticos (LLM) en la práctica clínica diaria [Mentor]. *Medizin Online*. https://medizinonline.com/es/dr-chatgpt-grandes-modelos-lingueisticos-en-la-practica-clinica-diaria/

Studysmarter. (2024, August 30). Modelos del lenguaje: Teóricos y Cognitivo. https://www.studysmarter.es/resumenes/medicina/logopedia/modelos-del-lenguaje/

Takita, H., Kabata, D., Walston, S., Tatekawa, H., Saito, K., Tsujimoto, Y., Miki, Y. & Ueda, D. (2025). A systematic review and meta-analysis of diagnostic accuracy of generative artificial intelligence models in healthcare. *npj Digital Medicine, 8*, 175. https://doi.org/10.1038/s41746-025-01543-z

FINANCING

The author received no funding for the development of this research.

CONFLICT OF INTEREST

The author declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: María Teresa Nagua Velepucha

Data curation: María Teresa Nagua Velepucha

Formal analysis: María Teresa Nagua Velepucha

Investigation: María Teresa Nagua Velepucha

Methodology: María Teresa Nagua Velepucha

Project Management: María Teresa Nagua Velepucha

Resources: María Teresa Nagua Velepucha

Software: María Teresa Nagua Velepucha

Supervision: María Teresa Nagua Velepucha

Validation: María Teresa Nagua Velepucha

Drafting – original draft: María Teresa Nagua Velepucha

Drafting – revision and editing: María Teresa Nagua Velepucha