Generative Artificial Intelligence as an aid for the assessment of Early Modern History student work in higher education

A Inteligência Artificial Generativa como recurso na avaliação de trabalhos de estudantes de História Moderna na educação superior

Antonio Carrasco-Rodríguez [*1] **and Humberto Álvarez Sepúlveda** [†2]

[1] *Universidad de Alicante, Facultad de Filosofía y Letras, Alicante, España.*
[2] *Universidad Católica de la Santísima Concepción, Facultad de Educación, Concepción, Chile.*

Abstract

This article examines the use of generative artificial intelligence (AI) as a supportive instrument in the assessment of undergraduate academic work in Early Modern American History. Drawing on a hybrid evaluation model that integrates human judgment with AI-assisted assessment, the study implemented three successive stages of correction: an initial assessment based on a basic rubric, a second evaluation using an advanced analytical rubric, and a final phase involving the critical recalibration of prior results. The corpus consisted of 21 research-based assignments produced by fourth-year History students and evaluated according to technical, historiographical, and critical-thinking criteria. The findings show that, once properly calibrated, the AI system was able to discriminate effectively between different levels of academic quality and to identify recurring patterns of historical reasoning in student work. A comparison between human-generated and AI-generated grades revealed a high degree of convergence, alongside significant divergences. While the AI demonstrated greater sensitivity to methodological rigor and critical engagement, human evaluators tended to prioritize formal presentation and technical aspects of writing. Rather than replacing instructor judgment, the proposed model reframes assessment as a more rigorous, equitable, and reflexive pedagogical process. The study suggests that, when embedded within demanding, transparent, and reviewable pedagogical frameworks, generative artificial intelligence can operate as an epistemic agent within the Humanities, contributing meaningfully to the evaluation of complex historical learning outcomes.

Keywords: Early Modern History. Hybrid assessment model. Generative Artificial Intelligence. Higher education. Analytical rubrics.

Textolinne
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2026.59410

Session:
Articles

Corresponding author:
Antonio Carrasco-Rodríguez

Section Editor:
Danniervelin Pereira
Layout editor:
Saula Cecília

Received on:
May 29, 2025
Accepted on:
January 6, 2026
Published on:
April 1, 2026

Resumo

Este artigo examina o uso da inteligência artificial (IA) generativa como instrumento de apoio na avaliação de trabalhos acadêmicos de graduação em História da América Moderna. Com base em um modelo híbrido de avaliação que integra o julgamento humano com a aferição assistida por IA, o estudo implementou três etapas sucessivas de correção: uma avaliação inicial baseada em uma rubrica básica, uma segunda avaliação com o uso de uma rubrica analítica avançada e uma fase final de recalibração crítica dos resultados anteriores. O *corpus* consistiu em 21 trabalhos de caráter investigativo elaborados por estudantes do quarto ano do curso de História, avaliados segundo critérios técnicos, historiográficos e de pensamento crítico. Os resultados demonstram que, uma vez devidamente calibrado, o sistema de IA foi capaz de discriminar eficazmente entre diferentes níveis de qualidade acadêmica e de identificar padrões recorrentes de raciocínio histórico nos trabalhos analisados. A comparação entre as notas atribuídas por avaliadores humanos e aquelas geradas pela IA revelou um elevado grau de convergência, juntamente com divergências significativas. Enquanto a IA demonstrou maior sensibilidade ao rigor metodológico e ao engajamento crítico, os avaliadores humanos tenderam a priorizar a apresentação formal e os aspectos técnicos da escrita. Em vez de substituir o julgamento docente, o modelo proposto reconceptualiza a avaliação como um processo pedagógico mais rigoroso, equitativo e reflexivo. O estudo sugere que, quando integrada a marcos pedagógicos exigentes, transparentes e passíveis de revisão, a inteligência artificial generativa pode atuar como um agente epistêmico no âmbito das

*Email: antonio.carrasco@ua.es
†Email: halvarez@ucsc.cl

Humanidades, contribuindo de forma significativa para a avaliação de resultados complexos de aprendizagem histórica.

Palavras-chave: História Moderna. Modelo híbrido de avaliação da aprendizagem. Inteligência Artificial Generativa. Educação superior. Rubricas analíticas.

1 Introduction

The integration of digital technologies into university education has profoundly transformed pedagogical dynamics, opening up a fertile field of possibilities while simultaneously generating new ethical and educational dilemmas (Scott; Smith, 2024; Tondeur *et al.*, 2025; Zhou; Smith; Al-Samarraie, 2024). Among the most significant challenges arising from this process is the emergence of generative artificial intelligence (AI) tools, particularly within the Humanities, where competencies such as argumentation, critical interpretation, and the construction of historical knowledge are central to students' academic development. Since the launch of ChatGPT in November 2022, and especially during its consolidation in educational contexts throughout the 2023–2024 academic year, debates surrounding the role of AI in higher education have intensified. While some approaches emphasize potential risks – such as the depersonalization of learning or the rise in academic plagiarism – more constructive perspectives underline its potential as a didactic, creative, and evaluative resource, provided that its implementation is framed within ethical, critical, and context-sensitive frameworks (Evangelista, 2025; Gammoh, 2025; Luo; Zou, 2025).

The advent of generative language models such as ChatGPT has profoundly reshaped discussions around authorship, creativity, learning, and modes of assessment in higher education. While students have rapidly incorporated these tools into their everyday academic practices, university faculty have been compelled to critically reassess their teaching strategies and evaluation criteria, as well as to reconsider the dynamics of the pedagogical relationship itself. A growing body of research has begun to examine the opportunities that artificial intelligence offers for university education, while simultaneously warning of the risks associated with its uncritical, automated, or ethically unanchored use (Gammoh, 2025; Holmes; Bialik; Fadel, 2019; Knox, 2020; Luckin; Holmes, 2016; Ofem *et al.*, 2025; Selwyn, 2021).

In response to this evolving landscape, various initiatives have emerged to explore creative and responsible applications of AI, particularly in areas related to academic writing, simulation, and formative feedback (Bauer *et al.*, 2025; Luckin; Holmes, 2016; Topping *et al.*, 2025). However, most of these experiences have been concentrated within the fields of science, technology, engineering, and mathematics (STEM), leaving the potential applications of AI within the Humanities comparatively underexplored. Although debates on the impact of digital technologies have given rise to a robust theoretical corpus within the field of Digital Humanities (Chen; Witt; Lin, 2025; Svensson, 2012; Toktas, 2025), the specific role of artificial intelligence as a resource for critical pedagogy remains at an early stage of conceptual development. This is particularly evident with regard to the cultivation of skills such as analytical reading, reflective writing, and the reconstruction of complex historical contexts. In the field of History education, these challenges are especially acute: it is essential to prevent AI from reducing historical processes to simplistic linear narratives or from substituting critical problematization and analysis with automated and uncritical content (Alvarez, 2023; Carrasco, 2023, 2024; Kansteiner, 2022; Tirado *et al.*, 2023; Washburn; McCutchen, 2024).

Against this backdrop, research on university assessment has increasingly emphasized the importance of developing clear, formative instruments aligned with learning objectives, such as analytical rubrics. These tools not only facilitate fair and transparent evaluation (Andrade, 2019; Nicol; Macfarlane, 2006; Panadero; Jonsson, 2013), but also function as frameworks for self-regulation, peer assessment, and continuous improvement in learning—even when integrated with technological resources such as artificial intelligence (Kaldaras; Yoshida; Haudek, 2022; Martin; Kranz; Graulich, 2024).

There is an urgent need to design assessment strategies that are technically effective while fostering the development of research skills and historical thinking competencies within ethical, critical, and

transparent frameworks of practice. From this perspective, the present article documents an innovative pedagogical experience developed in the course America: History from Colonization to the Present, taught in the fourth year of the History Degree Program at the University of Alicante (Spain). For this course, a hybrid evaluation system was designed, combining instructor-led mediation with the use of a customized AI assistant. As part of the activity, students working in pairs were tasked with creating prompts capable of simulating conversations with significant historical figures included in the syllabus, thereby promoting the development of key competencies: documentary research, critical problematization, rigorous historical narration, and ethical self-reflection on the use of AI in learning processes.

The article pursues a dual objective. First, it seeks to document and analyze a real-world experience of university assessment in an Early Modern History course that combined the use of artificial intelligence with an advanced rubric to evaluate complex practical assignments. Second, it aims to theorize and systematize this hybrid evaluation model as a methodology applicable to other teaching contexts, particularly within the Humanities. To this end, the article first presents the pedagogical design of the implemented activity, focusing on the creation and evaluation of simulated historical figures through conversational prompts. It then details the evolution of the rubric employed, the AI-assisted evaluation process, the instructor's intervention, and the subsequent recalibration of results. The results section analyzes the grades obtained through different rubrics and levels of rigor, the patterns identified, and the adjustments made based on comparisons between human and AI evaluations. Finally, the article discusses the achievements, limitations, and potential of the model from a critical pedagogical perspective, offering recommendations for its future implementation in similar contexts.

Far from viewing artificial intelligence as a substitute for instructor judgment, this article advocates its use as a complementary tool capable of enriching the assessment process, provided that it is guided by clear criteria, structured through robust pedagogical instruments, and subjected to ongoing reflective review (Evangelista, 2025; Luo; Zou, 2025; Selwyn, 2021).

2 Methodology

2.1 Teaching context and student profile

The experience described in this article was carried out during the 2024-2025 academic year as part of the course America: History from Colonization to the Present. The primary objective of this course is to provide students with conceptual, methodological, and critical tools for analyzing the historical processes that have shaped the American continent from the earliest human settlements to the present day.

The participating group consisted of 42 students with prior experience in document analysis, critical reading, and academic essay writing, as well as an advanced level of digital competence in the educational use of generative artificial intelligence. This competence had been developed through previous workshops and related learning activities.

2.2 Activity design

The activity "Conversing with the Past", which constitutes the empirical basis of this study, was designed as a paired assignment and accounted for a significant proportion (20%) of the final course grade. Characterized by a strong experimental and innovative component, the activity aimed to foster autonomous research, methodological creativity, and critical reflection on the use of generative artificial intelligence in historical learning contexts. Students were tasked with designing and developing a ChatGPT prompt capable of simulating conversations with a historical figure from colonial America. Accordingly, the assignment integrated the development of historiographical skills (research, critical analysis, and contextualization), technical skills (prompt drafting, conversational testing, and tone and style adjustment), and reflective skills (evaluation of the learning experience and critical appraisal of AI use within the activity).

Each project was required to include the following sections: an introduction (justifying the selection of the historical figure and their relevance), a biographical profile, a historical context section, a prompt explanation (detailing the creation process, structure, tests conducted, iterations, and adjustments),

a conclusion (offering a critical reflection on the activity and on the use of AI for learning History), a bibliography (formatted according to APA 7th edition guidelines), and annexes (including the complete prompt and an example of a conversation conducted with the simulated historical figure).

As support resources, students were provided with several customized GPT assistants. These included an assistant specialized in the History of pre-Columbian and colonial Hispanic America, trained on the course's didactic materials (lecture notes, supplementary content, and visual presentations) and aligned with the course syllabus (learning objectives, competencies to be developed, and evaluation criteria). In addition, students had access to a GPT assistant designed to improve basic prompts, another aimed at fostering historical thinking, and a third dedicated to the evaluation of historical narratives.

Students were free to select their historical figure, provided that the choice fell within the chronological and geographical scope of colonial Hispanic America. This flexibility allowed them to develop conversation simulators featuring canonical figures (such as Fray Bartolomé de las Casas, Malintzin, or Pedro de Valdivia), lesser-known historical actors (such as Catalina de Erauso or Francisca Pizarro Yupanqui), and even fictional characters constructed within historically plausible frameworks (such as a Jesuit missionary or an Afro-descendant freedwoman).

2.3 Hybrid assessment model

The evaluation of the assignments was conducted through a hybrid model that combined the course's customized GPT assistant with the expert judgment of the teaching staff. This system was designed to ensure a rigorous, precise, and formative assessment process, capable of identifying differences in assignment quality beyond superficial features and of generating feedback grounded in clearly defined criteria. The evaluation process and the specific tasks carried out by the two instructors involved are detailed below.

Phase 1. Creation of a basic rubric and initial AI-based evaluation. The instructors developed an initial rubric consisting of four criteria (historical rigor, prompt quality, critical and reflective capacity, and formal and communicative quality) and a total of eleven evaluable indicators. Instructor #1 subsequently uploaded the 21 assignments to the course's customized GPT assistant and requested their evaluation according to the newly created rubric.

Phase 2. Development of an advanced rubric and second AI-based evaluation. After reviewing the results of the initial automated evaluation, the teaching team sought to enhance both the human and AI-assisted assessment processes by developing a more advanced rubric. This revised rubric comprised five general criteria (historical rigor and academic relevance, technical design of the prompt, dialogue quality, critical and reflective capacity, and overall presentation), sixteen indicators, and a seven-level evaluation scale ranging from "very poor" to "excellent", with corresponding numerical scores from 1 to 100. Once finalized, Instructor #1 asked the course GPT assistant to re-evaluate the 21 assignments using this updated framework.

Phase 3. Independent human evaluation. Without access to the automated evaluation results, Instructor #2 independently assessed the assignments using the advanced rubric. This evaluation included minor discretionary adjustments not explicitly specified in the rubric, such as deductions for spelling errors or issues related to writing quality.

Phase 4. Review of evaluations, recalibration, and final validation. Following a comparison of the human and automated evaluations, and after observing a tendency for scores to cluster toward the higher end of the scale, the instructors decided to recalibrate the evaluation standards embedded in the AI rubric in order to better differentiate between excellent, good, and acceptable assignments. Once the recalibration criteria were defined, the course GPT assistant conducted a new round of evaluations accordingly.

Phase 5. Assignment of final grades. The instructors compared the recalibrated AI-generated evaluations with the human evaluations and determined that the final grade for each assignment would be calculated as the average of both scores. This decision was intended to reduce the potential influence of subjective bias in human assessment.

From the instructors' perspective, the implementation of this hybrid assessment system not only

enhanced evaluative precision but also created new spaces for reflection on the act of assessment itself. The collaboration between AI-based analysis and instructor judgment did not diminish the critical role of faculty; rather, it reinforced it by introducing contrast, systematization, and objective data to support informed decision-making.

2.4 Development and application of the rubrics

2.4.1 The basic rubric

As previously indicated, the starting point of the assessment process was the design of a basic rubric intended to facilitate a rapid, automatable, and sufficiently structured evaluation capable of distinguishing the overall quality of the submitted assignments. Its primary purpose was diagnostic, making it particularly useful during an initial phase of review and preliminary classification.

This basic rubric comprised four criteria and a total of eleven specific indicators, each assigned a percentage weight contributing to the final grade. The criteria employed are detailed below:

1. Historical rigor (30%)
   (a) Factual accuracy (15%)
   (b) Analysis and problematization (10%)
   (c) Relevance and contextualization of the historical figure (5%)
2. Prompt quality (30%)
   (a) Technical design (10%)
   (b) Character coherence in AI interactions (10%)
   (c) Meaningful interaction (10%)
3. Critical and reflective capacity (20%)
   (a) Explanation of the process (10%)
   (b) Evaluation of AI usage (5%)
   (c) Self-assessment of learning (5%)
4. Formal and communicative quality (20%)
   (a) Writing and structure (10%)
   (b) References and formatting (10%)

The total score was calculated on the basis of the weighted average of all indicators. This rubric was employed by the AI assistant during the initial automated evaluation. Its application yielded useful results for distinguishing between higher- and lower-performing assignments; however, it also revealed several limitations:

- Limited discriminative capacity in the upper range of scores (90–100), where a large number of formally correct assignments clustered despite significant differences in analytical depth and critical engagement.
- Absence of defined achievement levels, as the rubric did not incorporate a qualitative scale to differentiate levels of performance within each criterion.
- Dependence on implicit evaluator judgment, given the lack of explicit guidelines specifying, for example, what should meaningfully distinguish a score of 85 from one of 95.

These shortcomings ultimately motivated the development of a second, more sophisticated rubric, characterized by greater granularity, higher evaluative demands, and enhanced analytical precision.

2.4.2 The advanced rubric

Following the application of the basic rubric in the AI-assisted assessment, it became evident that, although it provided a useful initial classification of the assignments, its discriminative power was limited, particularly within the upper range of grades. A considerable number of formally correct assignments – well written, complete, and well structured – received scores close to or above 90, despite substantial differences in analytical depth, use of sources, and technical handling of the prompt. Moreover, the instructors identified an academic need to refine the rubric before proceeding with the human evaluation of the assignments. Consequently, they undertook the design of a more detailed rubric capable of identifying and rewarding genuinely outstanding work across multiple dimensions, distinguishing more precisely between satisfactory, good, and excellent assignments, and fostering

deeper reflection both in the correction process and in the instructors' own understanding of the evaluation outcomes.

The result of this reflective process on the most appropriate assessment criteria was the development of an advanced rubric. A new criterion – analysis of dialogue quality – was incorporated. In parallel, the existing indicators were revised, their descriptions refined, and five additional indicators were introduced. To further enhance the evaluation process, a seven-level performance scale was established, ranging from "very poor" to "excellent".

This new rubric was employed both in the second AI-assisted evaluation and in the independent human evaluation, thereby enabling quantitative and qualitative comparisons between the two assessment processes.

The advanced rubric, together with its criteria, indicators, descriptions, and weightings, is presented below:

1. Historical rigor and academic relevance (35%)
   (a) Factual accuracy (15%)
   (b) Analysis and problematization (10%)
   (c) Relevance and justification of the historical figure (5%)
   (d) Use of primary and secondary sources (5%)
2. Technical design of the prompt (25%)
   (a) Prompt structure (10%)
   (b) Control of voice and style (5%)
   (c) Technical robustness and testing (5%)
   (d) Originality of approach (5%)
3. Dialogue quality (15%)
   (a) Depth of conversation (6%)
   (b) Evolution and progression of the dialogue (5%)
   (c) Presentation of historical conflicts or dilemmas (4%)
4. Critical and reflective capacity (15%)
   (a) Explanation of the work process (6%)
   (b) Ethical and critical reflection on AI use (5%)
   (c) Self-assessment of learning (4%)
5. Presentation of the assignment (10%)
   (a) Writing, clarity, and structure (6%)
   (b) Citations, bibliography, and formatting (APA 7th edition) (4%)

The advanced rubric employed a seven-level performance scale, each level associated with a numerical score range out of 100. This scale was inspired by the traditional Spanish grading system used in the History Degree Program and enabled a more precise and equitable evaluation, clearly differentiating between excellent, good, and merely acceptable assignments. The defined levels are detailed below:

- Excellent (95–100 points)
- Outstanding (85–94)
- Very Good (70–84)
- High Pass (60–69)
- Pass (50–59)
- Insufficient (30–49)
- Very Poor (1–29)

Following the AI-assisted evaluation using the advanced rubric, a recurring tendency was identified whereby the course's GPT assistant tended to overrate certain assignments, particularly those characterized by highly polished formal presentation but limited critical depth or insufficient historiographical problematization. This pattern underscored the need to introduce a recalibration of rigor levels, aimed at aligning the evaluation process more strictly with the formative objectives of the course.

The recalibration process involved redefining the interpretive criteria associated with each achievement level and adjusting evaluation thresholds across several key dimensions:

- For the first criterion, historiographical depth was prioritized over formal aspects, with particular emphasis placed on engagement with debates, interpretive approaches, historiographical schools, or internal conflicts within the field.
- In the assessment of the prompt, technical execution was balanced with the historical and critical quality of the simulated interaction.
- In AI-mediated dialogues, greater weight was assigned to the capacity to generate tensions, contradictions, and historically grounded dilemmas, rather than merely expository exchanges.
- Within the critical and reflective criterion, the absence of ethical, pedagogical, or epistemic problematization concerning the use of artificial intelligence in History learning was explicitly penalized.
- With regard to the fifth criterion, it was emphasized that while writing quality and organizational clarity are important, they should not obscure deficiencies in analytical depth or independent historical thinking.

Finally, the GPT assistant was retrained using more precise examples and clearer instructions on how to apply the revised evaluation criteria. This recalibration resulted in a fairer and more critical assessment process, more closely aligned with the competency profile expected of fourth-year undergraduate History students.

3 Results

3.1 Overall performance and diversity of outcomes

The dataset analyzed comprises 21 assignments completed by fourth-year History students at the University of Alicante. Each assignment was evaluated a total of four times: three times by the course's customized GPT assistant – first using the basic rubric, subsequently using the advanced rubric, and finally applying the recalibrated version of the advanced rubric – and once by the course instructor. The final grade for each assignment was calculated as the arithmetic mean of the recalibrated AI-generated evaluation and the human evaluation, in accordance with the principles of fairness, contrast, and transparency established in the methodological design (see Section 2.3).

The grades obtained by each assignment across the different evaluation processes are presented below (Table 1).

Table 1. Grades across the different evaluation processes.

Assignment	Human Evaluation	Basic AI Evaluation	Advanced AI Evaluation	Recalibrated AI Evaluation	Final Grade
01.pdf	92.50	89.75	92.06	85.85	89.18
02.pdf	94.50	92.75	94.10	86.69	90.60
03.pdf	75.00	80.75	79.30	78.50	76.75
04.pdf	82.50	77.00	85.56	82.60	82.55
05.pdf	86.50	91.00	91.44	88.47	87.49
06.pdf	87.50	84.75	88.64	86.86	87.18
07.pdf	95.00	92.00	94.75	91.82	93.41
08.pdf	92.50	88.00	88.04	85.54	89.02
09.pdf	83.50	83.50	89.04	84.70	84.10
10.pdf	96.50	95.35	96.33	93.89	95.20
11.pdf	94.50	92.00	95.86	93.74	94.12
12.pdf	96.00	90.50	97.06	92.88	94.44
13.pdf	96.50	92.50	96.85	92.80	94.65
14.pdf	95.00	93.75	96.41	93.86	94.43
15.pdf	91.50	91.00	96.99	88.18	89.84
16.pdf	89.00	93.50	95.89	92.14	90.57
17.pdf	98.00	92.00	97.75	95.70	96.85
18.pdf	93.50	92.75	94.77	91.15	92.33
19.pdf	96.50	97.50	97.91	96.73	96.62
20.pdf	89.00	97.50	98.21	94.38	91.69
21.pdf	86.00	91.25	96.95	90.53	88.27

Source: Own elaboration.

The overall mean of the final grades was 90.44 out of 100, with a standard deviation of 4.98.

The lowest score recorded was 76.75 points, while the highest reached 96.85, resulting in a difference of more than 20 points between the two extremes. This relatively wide range indicates a moderate degree of dispersion and demonstrates that the evaluation model was capable of identifying meaningful qualitative distinctions among assignments that were formally correct but differed in depth, creativity, and critical rigor.

To facilitate qualitative analysis, the final grades were classified according to the seven performance levels defined in the advanced rubric. The distribution was as follows:

- Excellent: 3 assignments
- Outstanding: 15 assignments
- Very Good: 3 assignments
- No lower grades recorded

This distribution reflects an overall high level of performance within the group, which is consistent with its academic profile: fourth-year undergraduate students benefiting from close instructional support, prior training in academic writing, and sustained experience with the use of AI tools in pedagogical contexts. Nevertheless, the presence of three assignments rated as "Very Good" alongside three clearly "Excellent" ones indicates that the evaluation model did not merely confirm formal correctness, but rather succeeded in identifying substantive differences in the quality of student work.

In relation to the objectives defined in Section 1 – particularly the aim of distinguishing between different levels of quality beyond surface-level formal accuracy – the results suggest that the hybrid system, when combined with a demanding rubric and a rigorous recalibration process, was capable of offering a nuanced, differentiated, and formative reading of student performance. Rather than homogenizing grades, this model enhanced evaluative precision, remaining attentive to historiographical content, critical reflection, and the technical construction of academic work.

3.2 Impact of the rubrics and the recalibration process

One of the central components of the hybrid assessment model was the progressive evolution of the automated evaluation instruments, which moved from an initial basic rubric designed for rapid diagnostic purposes, to a more detailed advanced rubric, and finally to a recalibration phase that adjusted levels of rigor. This sequential design made it possible to observe clearly how the grades generated by the GPT assistant evolved over time and to assess the concrete impact that changes in the assessment structure had on evaluation outcomes.

The means and standard deviations corresponding to each evaluation phase were as follows:

- AI with basic rubric: mean = 90.43, standard deviation = 5.14
- AI with advanced rubric (unrecalibrated): mean = 93.19, standard deviation = 4.99
- AI with recalibrated advanced rubric: mean = 89.86, standard deviation = 4.72

This pattern reveals three distinct dynamics. First, the transition from the basic rubric to the advanced rubric resulted in an increase in the average grade for 20 of the 21 assignments, with a mean improvement of +3.09 points and six assignments showing gains greater than five points. This general upward trend can be interpreted as the effect of a richer and more detailed rubric, which allowed technical, formal, and structural aspects of student work to be recognized more explicitly than was possible with the basic rubric.

The third evaluation phase – following recalibration – introduced a decisive adjustment. In this stage, all 21 assignments experienced a reduction in their scores relative to the unrecalibrated advanced rubric, with an average decrease of –3.66 points; four assignments declined by more than five points. This reduction is pedagogically significant: rather than reflecting a systematic penalty, it represents a critical refinement of the evaluation criteria, prioritizing historiographical depth, analytical problematization, and reflective capacity over purely formal qualities.

A comparison between the basic rubric and the recalibrated advanced rubric yielded a more balanced outcome. Eight assignments showed an improvement in their grades, while thirteen experienced a decrease; only one assignment increased by more than five points, and only one declined beyond that threshold. The overall average variation was –0.58 points, indicating that the recalibrated system did not globally make the evaluation more stringent, but instead redistributed grades in a more

equitable and meaningful manner. The recalibration phase was therefore not merely an adjustment of score ranges, but a redefinition of how achievement levels were interpreted across key dimensions of the advanced rubric, as outlined in Section 2.4. These adjustments substantially enhanced the system's discriminative capacity: it was no longer sufficient to submit a well-written and well-structured assignment; instead, complex historical reasoning, historiographical decision-making, creativity, and self-regulatory capacity became central evaluative requirements. The GPT assistant was retrained to internalize these criteria, resulting in a more demanding, nuanced, and educationally aligned evaluative tool.

The analyzed data confirm that the introduction of an advanced rubric, in the absence of critical recalibration, may produce inflationary effects on grades, even when the evaluative criteria are more comprehensive. Only a subsequent adjustment process – based on a stricter and more explicit definition of rigor thresholds – can refine the evaluation without compromising fairness.

Finally, the impact of the recalibration was not homogeneous: it did not operate as a blanket penalty but rather redistributed recognition in accordance with clearer academic performance standards. Consequently, the decision to calculate final grades as the average of the human evaluation and the recalibrated AI-generated evaluation is fully justified. This approach integrates the instructor's expert judgment with a critically adjusted automated assessment and ensures that grading reflects a plurality of criteria and perspectives. Overall, the results demonstrate that AI-assisted evaluation can only be considered valid when embedded within a critical, reflective, and demanding pedagogical framework. Its legitimacy depends not on the model itself, but on its continuous review and alignment with formative educational objectives.

3.3 Alignment and Tension between Instructor Judgment and AI Evaluation

One of the most relevant aspects of the hybrid assessment model was the comparison between the grades assigned by the instructors and those generated by the GPT assistant using the recalibrated rubric. The results reveal a notably high level of overall alignment, albeit accompanied by critical nuances that enrich the analysis.

Of the 21 assignments evaluated, four showed a difference of less than or equal to one point between the two evaluations; eight assignments presented a difference between 1 and 2.99 points; and the remaining nine exhibited a discrepancy greater than three points. The mean absolute difference was 2.91 points, with a maximum divergence of 7.81 points and a minimum of 0.10 points. This distribution indicates strong overall convergence, while simultaneously pointing to differential sensitivities to specific evaluative criteria on the part of AI and human assessors.

A closer examination of the cases with the greatest divergences reveals consistent patterns. Assignments 02.pdf and 03.pdf received high scores from the instructors in the initial review, largely due to their careful formal presentation, fluent academic writing, and technically sound prompt design. By contrast, the AI assigned lower scores after identifying limitations in historiographical problematization and critical reflexivity.

The opposite pattern was observed in assignments 20.pdf and 21.pdf, which, despite exhibiting minor formal errors and a less conventional structural organization, stood out for their analytical depth, critical capacity, and disciplinary originality. In these cases, the AI clearly recognized these strengths and awarded higher scores than those initially assigned by the instructors.

This contrast highlights a subtle yet significant bias in the initial human evaluation process: an unconscious preference for formal and technical polish, which may lead to the undervaluation of assignments that are less refined in presentation but richer in substantive content. In the cases noted above, a subsequent rereading by the instructors led to the acknowledgment of the AI's more precise evaluative judgment, thereby validating the pedagogical usefulness of this comparative approach.

From a pedagogical standpoint, the hybrid model demonstrates that both sources of evaluation contribute complementary perspectives. When trained using refined and demanding criteria, the AI proved to be more stringent in assessing disciplinary content, methodological rigor, and critical reflection, whereas instructor judgment added sensitivity to contextual, ethical, and formative dimensions that are not always fully captured by rubric-based assessment. For this reason, calculating the final

grade as the average of the two evaluations is a fully justified methodological decision. This approach mitigates potential biases inherent in each source, broadens the interpretive spectrum, and contributes to a fairer and more pluralistic assessment process. Rather than replacing the instructor, the system complements human judgment with a distinct, fine-tuned, and pedagogically aligned evaluative perspective.

3.4 Qualitative Lessons on Historical Thinking and Critical Evaluation

The qualitative analysis of the assignments reveals clear patterns that offer valuable insights into the development of historical thinking in learning contexts mediated by artificial intelligence. The highest-rated assignments combined deep historiographical reflection, clearly articulated problematizations, and a creative use of prompts, resulting in simulations that incorporated dilemmas, tensions, and intellectually demanding exchanges. These assignments did not merely reproduce factual information; rather, they demonstrated students' capacity to critically engage with the past, interpret its complexities, and construct well-argued historical narratives—reflecting advanced levels of historical thinking as conceptualized by Seixas and Morton (2013).

Moreover, a critical stance toward artificial intelligence itself emerged as a notable feature of the strongest assignments, as students approached AI not only as a tool but also as an object of reflection. This reflexive dimension, evident in the concluding sections of several assignments, revealed a metacognitive appropriation of the learning process, whereby students did not simply employ AI instrumentally but also problematized its potential, limitations, and epistemological risks. This orientation aligns with recent approaches to critical digital literacy in the Humanities (Carrasco, 2024; Chen; Witt; Lin, 2025; Kansteiner, 2022; Knox, 2020; Selwyn, 2021).

By contrast, lower-scoring assignments tended to present flatter narratives, limited problematization, and a more superficial application of AI, often characterized by generic questioning and a lack of interpretive agency. Although many of these assignments were formally correct, their historical approaches were weaker, revealing a limited understanding of the processes involved in the construction of historical knowledge and an excessive reliance on surface-level information. This pattern confirms that the evaluation model – particularly following rubric recalibration – was capable of distinguishing between performances based on rote reproduction and those demonstrating critical and autonomous historical thinking.

A defining feature of the strongest assignments was the deliberate incorporation of conflicts, ethical dilemmas, and uncertainties inherent to historical knowledge. Students who succeeded in constructing complex dialogues introduced themes such as internal contradictions of historical figures, the political tensions of their time, or the ambiguities of colonial processes. These elements suggest the development of an authentic critical historical consciousness, as described by Rüsen (2005, 2010). This ability to articulate multiple perspectives and to recognize the contingency and interpretive nature of historical events corresponds to higher levels of progression in historical thinking, as proposed by Ashby and Lee (2000), Seixas and Morton (2013) and Alvarez (2021).

In addition, the critical integration of AI into the learning process emerged as a valuable transversal competence. In several reports, students explicitly acknowledged that interacting with an AI assistant compelled them to formulate questions more precisely, refine their research objectives, and anticipate potential technological biases. This finding reinforces the hypothesis that a didactically guided use of AI can not only support learning but also catalyze metacognitive processes and the development of academic self-regulation skills (Holmes; Bialik; Fadel, 2019; Luckin; Holmes, 2016; Luo; Zou, 2025; Tondeur *et al.*, 2025).

With regard to critical evaluation, the experience demonstrated that the combination of demanding analytical rubrics and a reflective recalibration process made it possible to highlight substantive differences in quality that might have remained unnoticed in more traditional assessment frameworks. When properly trained and calibrated, the AI functioned as a critical mirror, not only assessing formal correctness but also detecting nuances of analytical depth, historiographical rigor, and problematization capacity. In this way, it enriched instructor judgment and supported the provision of more formative and targeted feedback.

Finally, it is important to emphasize the need to understand historical thinking as a situated, ethical, and inherently complex practice, particularly within technology-mediated educational environments (Alvarez, 2023; Carrasco, 2024; Tirado *et al.*, 2023). The highest-performing assignments were not those that simply reproduced conventional narratives or achieved stylistically fluent interactions, but rather those that ventured into uncertainty, challenged established interpretations, and critically reflected on the very act of learning and producing historical knowledge in the twenty-first century.

4 Discussion

The experience analyzed in this study demonstrates that the critical integration of artificial intelligence into university-level History assessment can be highly beneficial from a pedagogical standpoint. The overarching objective – namely, to explore the potential of AI as a support tool for the assessment of historical learning through a hybrid model combining instructor judgment with automated evaluation – was fully achieved. The findings support the conclusion that this model fosters more equitable, transparent, and formative assessment practices, in line with the recommendations of Panadero and Jonsson (2013) regarding the development of assessment instruments that promote autonomous learning and self-regulation.

From a quantitative perspective, the implementation of recalibrated AI assessment, underpinned by robust and demanding rubrics, improved the discrimination of performance levels by prioritizing historiographical, critical, and epistemological dimensions over purely formal aspects. This methodological adjustment – reflected in the observed reduction in average scores following recalibration – can be understood as an instance of what Holmes, Bialik, and Fadel (2019, p. 6, 11–12) describe as a "critical appropriation of technology", whereby AI is employed not as a tool for simplification, but as a catalyst for deeper and more rigorous evaluative practices.

With regard to evaluative convergence, the mean difference of 2.91 points between human and automated assessments confirms the viability of carefully designed hybrid systems. As argued by Luckin and Holmes (2016), AI should not seek to replace instructor judgment, but rather to complement and enrich it by broadening evaluative perspectives and enhancing fairness in the assessment process. In this respect, AI – when trained using stringent historiographical and ethical criteria – demonstrated a particular sensitivity to problematization, critical reflection, and argumentative density, dimensions that human judgment may tend to undervalue when formal presentation predominates (Carrasco, 2024; Selwyn, 2021; Toktas, 2025).

From a qualitative standpoint, the experience revealed diverse forms of historical thinking among students. The highest-performing assignments not only demonstrated mastery of historical content, but also an advanced capacity to formulate complex questions, challenge established narratives, recognize ethical dilemmas, and critically interpret the past. These features reflect higher levels of historical thinking progression as described by Seixas and Morton (2013). Moreover, critical engagement with AI – far from being limited to instrumental use – fostered a metacognitive appropriation of the learning process, reinforcing the importance of reflecting on the means through which historical knowledge is produced (Carrasco, 2023; Carretero; Gartner, 2024; Leme, 2023; Tirado *et al.*, 2023).

In terms of contributions to both educational practice and the disciplinary field, the study supports the notion that AI can function as an epistemic agent in the classroom (Holmes; Bialik; Fadel, 2019; Washburn; McCutchen, 2024), provided that it is embedded within a rigorous, ethical, and discipline-sensitive pedagogical framework. Rather than conceiving AI as a mechanism for automation, the model analyzed here positions it as a facilitator of deep, self-regulated, and critical learning, in line with emerging perspectives on the pedagogical use of digital technologies in the Humanities.

Nevertheless, the experience also highlights tensions and risks that must be addressed critically. As Wineburg (2001) cautions, authentic historical thinking requires engagement with the complexity, contingency, and ambiguity inherent in the past – dimensions that poorly designed or inadequately calibrated AI systems may inadvertently suppress by privileging simplified or normative narrative structures. Furthermore, there is a risk that automated assessment could normalize particular forms of historiographical reasoning, potentially marginalizing creative, alternative, or less conventional approaches unless such systems are regularly triangulated with instructors' qualitative evaluations.

From an epistemological perspective, the incorporation of AI into assessment raises significant formative challenges. These include ensuring that students do not perceive automated feedback as mechanical or disconnected from their own learning processes (Carretero; Gartner, 2024; Holmes; Bialik; Fadel, 2019), as well as creating spaces for critical reflection on the impact of digital technologies on the production, validation, and interpretation of historical knowledge.

In response to these challenges, the experience documented in this study suggests four guiding principles for advancing a critical approach to AI-based assessment in History education:

- Pedagogical contextualization, embedding the use of AI within didactic frameworks that prioritize the development of critical historiographical competencies.
- Rigorous instrument design, following Panadero and Jonsson (2013) guidelines for constructing clear, specific, and periodically revisable rubrics.
- Active instructor mediation, ensuring the qualitative interpretation of results and preventing the technocratization of evaluative judgment.
- Continuous critical reflection, fostering in students a critical awareness of the role of technologies in historical learning processes.

Finally, as Wineburg (2001) emphasizes, teaching History entails cultivating critical thinkers rather than mere reproducers of established narratives. Within this framework, AI-based assessment should not be reduced to measuring degrees of formal compliance, but should instead aim to foster students' capacity to interpret, problematize, and construct ethical and situated understandings of the past. When properly integrated, AI can act as a catalyst for the development of these higher-order competencies, provided that its use is mediated by emancipatory and reflective educational projects. Accordingly, this experience confirms that hybrid assessment models – centered on historical thinking competencies and critically supported by AI – are not only feasible, but also desirable for educating historians and citizens capable of interpreting and acting within a complex, dynamic, and digitally mediated world (Carrasco, 2024; Carretero; Gartner, 2024; Kansteiner, 2022; Tirado *et al.*, 2023).

5 Conclusions

This study has explored the potential of artificial intelligence as a support tool for the assessment of historical learning in the university context. Through the implementation of a hybrid model combining automated evaluation with instructor judgment, the findings demonstrate that it is possible to construct a fairer, more nuanced, and pedagogically meaningful assessment system. The results confirm that, when critically integrated into the evaluation process, AI not only enhances the differentiation of quality levels but also helps to illuminate deeper dimensions of students' academic work that might otherwise remain less visible.

The proposed model represents a significant innovation within the field of History education and, more broadly, within university assessment practices. By incorporating historiographical, methodological, and critical criteria into automated analysis, AI ceases to function merely as a tool for technical correction and instead becomes an agent that broadens pedagogical perspectives. This experience has shown that assessment can – and should – move beyond formal quality control to operate as a formative process that recognizes the diversity of historical thinking, fosters self-regulation, and stimulates sustained intellectual reflection.

Nevertheless, this model is not without limitations. Artificial intelligence cannot fully capture contextual nuances, nor can it adequately account for individual learning trajectories, personal effort, or singular intellectual developments. Its reliability is contingent upon the rigorous design of analytical rubrics, continuous review, and active instructor mediation. The aim is not to replace instructor judgment, but rather to structure, contrast, and enrich it. The use of AI in assessment therefore entails assuming a methodological and ethical responsibility that cannot itself be automated.

Looking ahead, the model outlined in this study could be adapted to other courses, disciplines, and educational contexts, provided that the conditions for its critical and reflective use are respected. There remains considerable scope for further research into the ways in which the pedagogical use of AI reshapes authorship practices, learning modalities, and disciplinary forms of thinking. Ultimately, evaluating with AI does not mean delegating judgment; rather, it involves rethinking – through

pedagogical rigor – how, why, and according to which criteria learning is assessed in university-level History education.

References

ALVAREZ, Humberto. Evaluación del pensamiento histórico de estudiantes de secundaria a través de la construcción de narrativas históricas sobre los pueblos originarios de Chile. *Años 90*, n. 28, p. 1–28, 2021. Available from: https://doi.org/10.22456/1983-201X.111650.

ALVAREZ, Humberto. La inteligencia artificial como catalizador en la enseñanza de la Historia: Retos y posibilidades pedagógicas. *Revista Docentes 2.0*, v. 16, n. 2, p. 318–325, 2023. Available from: https://doi.org/10.37843/rted.v16i2.426.

ANDRADE, Heidi. A critical review of research on student self-assessment. *Frontiers in Education*, v. 4, p. 87, 2019. Available from: https://doi.org/10.3389/feduc.2019.00087.

ASHBY, Rosalyn; LEE, Peter. Progression in historical understanding among students ages 7-14. *In:* STEARNS, Peter; SEIXAS, Peter; WINEBURG, Sam (eds.). *Knowing, teaching and learning history: National and international perspectives*. New York: New York University Press, 2000. p. 199–222.

BAUER, Elisabeth *et al.* AI-based adaptive feedback in simulations for teacher education: An experimental replication in the field. *Journal of Computer Assisted Learning*, v. 41, n. 1, e13123, 2025. Available from: https://doi.org/10.1111/jcal.13123.

CARRASCO, Antonio. Reinventing the teaching of early Modern History in secondary school: The use of ChatGPT to enhance learning and educational innovation. *Studia Historica: Historia Moderna*, v. 45, n. 1, p. 101–145, 2023. Available from: https://doi.org/10.14201/shhmo2023451101146.

CARRASCO, Antonio. Perceptions of generative artificial intelligence among university early Modern History students. *Tiempos Modernos*, v. 14, n. 49, p. 269–285, 2024.

CARRETERO, Mario; GARTNER, Elisa. Artificial intelligence and historical thinking: A dialogic exploration of ChatGPT. *Studies in Psychology*, v. 45, n. 1, p. 80–102, 2024. Available from: https://doi.org/10.1177/02109395241241379.

CHEN, Chih-Ming; WITT, Barbara; LIN, Chun-Yu. A knowledge graph analysis tool of people and organizations to facilitate digital humanities research. *Data Technologies and Applications*, v. 59, n. 1, p. 82–110, 2025. Available from: https://doi.org/10.1108/dta-01-2024-0009.

EVANGELISTA, Edmund De Leon. Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology*, v. 17, n. 1, ep559, 2025. Available from: https://doi.org/10.30935/cedtech/15775.

GAMMOH, Leen. ChatGPT risks in academia: Examining university educators' challenges in Jordan. *Education and Information Technologies*, v. 30, n. 3, p. 3645–3667, 2025. Available from: https://doi.org/10.1007/s10639-024-13009-y.

HOLMES, Wayne; BIALIK, Maya; FADEL, Charles. *Artificial intelligence in education: Promises and implications for teaching and learning*. [*S. l.*]: Center for Curriculum Redesign, 2019.

KALDARAS, Leonora; YOSHIDA, Nicholas; HAUDEK, Kevin. Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. *Frontiers in Education*, v. 7, p. 983055, 2022. Available from: https://doi.org/10.3389/feduc.2022.983055.

KANSTEINER, Wulf. Digital doping for historians: Can history, memory, and historical theory be rendered artificially intelligent? *History and Theory*, v. 61, n. 4, p. 119–133, 2022. Available from: https://doi.org/10.1111/hith.12282.

KNOX, Jeremy. Artificial intelligence and education in China. *Learning, Media and Technology*, v. 45, n. 3, p. 298–311, 2020. Available from: https://doi.org/10.1080/17439884.2020.1754236.

LEME, André. Artificial history? Inquiring ChatGPT on historiography. *Rethinking History*, v. 27, n. 4, p. 709–749, 2023.

LUCKIN, Rosemary; HOLMES, Wayne. *Intelligence unleashed: An argument for AI in education*. [*S. l.*]: Pearson, 2016.

LUO, Shuqiong; ZOU, Di. University learners' readiness for ChatGPT-assisted English learning: Scale development and validation. *European Journal of Education*, v. 60, n. 1, e12886, 2025. Available from: https://doi.org/10.1111/ejed.12886.

MARTIN, Paul; KRANZ, David; GRAULICH, Nicole. Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry. *International Journal of Artificial Intelligence in Education*, p. 1–39, 2024. Available from: https://doi.org/10.1007/s40593-024-00440-y.

NICOL, David; MACFARLANE, Debra. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, v. 31, n. 2, p. 199–218, 2006. Available from: https://doi.org/10.1080/03075070600572090.

OFEM, Usani *et al.* Students' perceptions, attitudes and utilisation of ChatGPT for academic dishonesty: Multigroup analyses via PLS-SEM. *Education and Information Technologies*, v. 30, n. 1, p. 159–187, 2025. Available from: https://doi.org/10.1007/s10639-024-12850-5.

PANADERO, Ernesto; JONSSON, Anders. The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review*, n. 9, p. 129–144, 2013. Available from: https://doi.org/10.1016/j.edurev.2013.01.002.

RÜSEN, Jörn. *History: Narration, interpretation, orientation*. New York: Berghahn, 2005.

RÜSEN, Jörn. *Jörn Rüsen e o ensino de história*. Curitiba: Editora UFPR, 2010.

SCOTT, Howard; SMITH, Matthew. Innovation from necessity: Digital technologies, teacher development and reciprocity with organisational innovation. *Open Learning: The Journal of Open, Distance and e-Learning*, v. 39, n. 2, p. 170–187, 2024. Available from: https://doi.org/10.1080/02680513.2024.2307627.

SEIXAS, Peter; MORTON, Tom. *The big six historical thinking concepts*. Toronto: Nelson Education, 2013.

SELWYN, Neil. *Should robots replace teachers? AI and the future of education*. Cambridge: Polity Press, 2021.

SVENSSON, Patrik. Beyond the big tent. *In:* GOLD, Matthew (ed.). *Debates in the digital humanities*. Minneapolis: University of Minnesota Press, 2012. p. 36–49. Available from: https://doi.org/10.5749/minnesota/9780816677948.003.0004.

TIRADO, Sergio *et al.* From human to machine: Investigating the effectiveness of the conversational AI ChatGPT in historical thinking. *Education Sciences*, v. 13, n. 8, p. 803, 2023. Available from: https://doi.org/10.3390/educsci13080803.

TOKTAS, Elif. Future scenarios of digital humanities and post-humanist education. *Journal of Foresight and Public Health*, v. 2, n. 1, p. 21–31, 2025.

TONDEUR, Jo *et al.* Preparing preservice teachers to teach with digital technologies: An update of effective SQD-strategies. *Computers & Education*, p. 105262, 2025. Available from: https://doi.org/10.1016/j.compedu.2025.105262.

TOPPING, Keith *et al.* Enhancing peer assessment with artificial intelligence. *International Journal of Educational Technology in Higher Education*, v. 22, n. 1, p. 3, 2025. Available from: https://doi.org/10.1186/s41239-024-00501-1.

WASHBURN, Jeffrey; MCCUTCHEN, Jennifer. AI meets AI: ChatGPT as a pedagogical tool to teach American Indian history. *Critical Humanities*, v. 2, n. 2, p. 2, 2024. Available from: https://doi.org/10.33470/2836-3140.1037.

WINEBURG, Sam. *Historical thinking and other unnatural acts: Charting the future of teaching the past*. Philadelphia: Temple University Press, 2001.

ZHOU, Xue; SMITH, Christopher; AL-SAMARRAIE, Hosam. Digital technology adaptation and initiatives: A systematic review of teaching and learning during COVID-19. *Journal of Computing in Higher Education*, v. 36, n. 3, p. 813–834, 2024. Available from: https://doi.org/10.1007/s12528-023-09376-z.

Author contributions

Antonio Carrasco-Rodríguez: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing; **Humberto Álvarez Sepúlveda**: Investigation, Supervision, Validation, Visualization, Writing – review and editing.

Ethical considerations

This study adhered to ethical principles based on international guidelines for research in higher education. The assessment activity formed part of the regular curricular programming and did not involve any interventions requiring additional ethical approval. Student participation was voluntary and anonymous, and the confidentiality of all data was ensured throughout the entire analysis process.

Data availability

Research data is available in the repository.