AUTOMATED EVALUATION OF HISTORICAL TEXTS: AN AI-BASED MODEL WITH STRUCTURED HISTORIOGRAPHIC CRITERIA

AVALIAÇÃO AUTOMATIZADA DE TEXTOS HISTÓRICOS: UM MODELO BASEADO EM INTELIGÊNCIA ARTIFICIAL E CRITÉRIOS HISTORIOGRÁFICOS ESTRUTURADOS

Antonio Carrasco-Rodríguez
PhD in Modern History from the University of Alicante (San Vicente del Raspeig/Espanha).
E-mail: antonio.carrasco@ua.es

Humberto Álvarez Sepúlveda
Doutor em História pela Universidade de Alicante (Alicante/Espanha).
Professor na Universidade de Alicante (Alicante/Espanha).
E-mail: halvarez@ucsc.cl

Recebido em: 12 de outubro de 2025
Aprovado em: 1 de dezembro de 2025
Sistema de Avaliação: Double Blind Review
RPR | a. 23 | n. 1 | p. 114-141 | jan./jun. 2026
DOI: https://doi.org/10.25112/rpr.v1.4254

Revista Prâksis

p-ISSN: 1807-1112
e-ISSN: 2448-1939

ABSTRACT

This article proposes an artificial intelligence (AI)-based model, specifically a GPT assistant, designed to automate the evaluation of historical texts using structured historiographic criteria. Historical quality assessment frequently encounters challenges, such as cognitive biases and subjective inconsistencies. The proposed model incorporates seven evaluation criteria (disciplinary, epistemological, ethical, technical, pedagogical, reliability, and practical utility) aimed at ensuring a more objective, transparent, and consistent assessment. The article argues that a suitably trained GPT assistant can significantly mitigate common issues found in traditional manual evaluations, streamlining processes and improving overall consistency. The paper discusses potential advantages, acknowledges inherent limitations, and outlines avenues for future research, emphasizing the need for rigorous training data and critical historiographic supervision.

Keywords: artificial intelligence; historiographic evaluation; ChatGPT; historical methodology; digital history.

RESUMO

Este artigo propõe um modelo baseado em inteligência artificial (IA), especificamente um assistente GPT, para avaliar textos históricos utilizando critérios historiográficos estruturados. A avaliação de textos históricos costuma enfrentar dificuldades devido à subjetividade, à falta de critérios uniformes e ao consumo excessivo de tempo. A proposta metodológica apresentada define sete critérios fundamentais (disciplinares, epistemológicos, éticos, técnicos, pedagógicos, fiabilidade e utilidade prática), concebidos para orientar uma avaliação mais objectiva, transparente e coerente. Argumenta-se que um assistente GPT devidamente treinado pode ajudar a superar desafios como preconceitos cognitivos e inconsistências na avaliação manual tradicional, bem como agilizar significativamente o processo de avaliação. Da mesma forma, refletimos criticamente sobre possíveis limitações do modelo, incluindo a necessidade de validar empiricamente a sua eficácia em contextos reais.

Palavras-chave: avaliação historiográfica; inteligência artificial; ChatGPT; metodologia histórica; história digital.

1 INTRODUCTION

Historiographic evaluation plays a crucial role in historical research and teaching, as it validates the quality and coherence of studies (BLOCH, 2023). However, it faces challenges such as the subjectivity of the evaluator, which can lead to divergent interpretations (WHITE, 1992). The lack of unified criteria complicates the comparison of studies across different academic contexts. Lee (2005) highlights the diversity of historiographic traditions, while Carr (2017) points out that historical interpretation is always situated within a particular conceptual framework, leading to inconsistencies and limiting the rigor of analysis. Theoretical, ideological, and methodological biases can also distort interpretations of the past (TOSH, 2015). In university settings, the absence of clear parameters can result in inequalities in assessing student performance, and in popular dissemination, there is a risk of spreading inaccurate or oversimplified views.

In response to these challenges, artificial intelligence has emerged as an alternative to optimize evaluation processes. Models such as GPT can process large volumes of text and provide preliminary analyses of argumentative coherence and source validity (BUDGE, 2021). However, their use must be accompanied by critical oversight to avoid reproducing biases (SAMIRA, 2024)..

However, the emergence of these technologies in the historical discipline should not be understood solely as an instrumental advancement. The ability of large language models to generate complex narratives and perform preliminary analyses poses an epistemological challenge that requires historiography to critically reflect on its own foundations. In this regard, this article not only presents an evaluation model but also engages with the current debate on the methodological and ethical implications of using AI in the construction of historical knowledge, encouraging the necessary critical oversight to align its potential with the discipline's rigor.

This article presents a GPT-based historiographic evaluation model, structured around seven categories (disciplinary, epistemological, ethical, technical, pedagogical, reliability, and practical utility criteria). Its aim is not to replace expert judgment but to complement it with a faster, more consistent, and objective initial analysis.

Furthermore, it proposes a critical discussion of the ethical and methodological implications of using AI in historical studies, encouraging future research that refines the model and explores new ways of integrating technology into the discipline.

2 THEORETICAL FRAMEWORK: TRADITIONAL HISTORIOGRAPHICAL EVALUATION AND CURRENT CHALLENGES

The evaluation of historiography is a fundamental element in the historical discipline, as it helps determine the quality, validity, and relevance of research in different contexts. However, this process has been marked by ongoing debates about methodological diversity, subjectivity in interpreting facts, and the absence of unified criteria. According to Burke (2008), history has always been interpreted in diverse ways, leading to tensions among different academic traditions. In this theoretical framework, we first examine the conventional criteria used in traditional historiographic evaluation and then address the current challenges that demand a more structured and critical approach.

2.1 TRADITIONAL HISTORIOGRAPHICAL EVALUATION

Since its origins, historiography has employed various forms of evaluation for its intellectual products. Traditionally, this process has been dominated by the figure of the expert, whose academic judgment and experience defined the validity of a study. Consequently, Western historiography has tended to reproduce evaluative models based on the European canon, limiting the inclusion of non-Western perspectives (MASOOD, 2025; NAGRE, 2025). Momigliano (1990) emphasizes that these processes have focused on the verification of sources and narrative coherence, underscoring the importance of rigorous documentary analysis. Carr (2017), for his part, argues that objectivity in history is relative, as interpretations are always influenced by the historian’s context.

The lack of universal criteria has led to methodological fragmentation and difficulties in comparing studies, perpetuating the absence of clear standards. As a result, traditional historiographic evaluation has relied on implicit norms and flexible criteria that vary depending on the academic community.

2.1.1 Common Criteria and Subjectivity

Traditional historiographic evaluation emphasizes documentary rigor, factual accuracy, internal coherence, and interpretative originality. However, these criteria are applied differently depending on the evaluator’s experience and perspective, introducing subjectivity into the process (WHITE, 1992).

This is evident in the dominance of certain historiographic approaches over others. For example, Marxist perspectives typically prioritize the analysis of socioeconomic structures (HOBSBAWM, 2000; THOMPSON, 2002), while cultural history places more emphasis on discourses and symbolic representations (BURKE, 2008; CHARTIER, 2021). This methodological diversity enriches the discipline but complicates the comparison and consistency of evaluations.

2.1.2 Problems of Replicability and Bias

One of the main problems of traditional historiographic evaluation is the lack of replicability. History is not an exact science, and its interpretations depend on the sociocultural context of production (TOSH, 2015), making it difficult to reach stable consensus.

Evaluation also tends to incorporate personal, cultural, and ideological biases, perpetuating Eurocentric views that marginalize other narratives (DABAT, 2024; BELIEIRO, 2024). This restricts epistemological diversity and limits the acceptance of approaches that challenge established paradigms. These problems call for a critical reassessment and the adoption of more inclusive and transparent criteria.

2.1.3 Impact on Research, Teaching, and Public History

The shortcomings of traditional evaluation significantly impact academic research, university teaching, and the public dissemination of history. In research, the lack of clear criteria undermines fairness in project funding, journal article selection, and dissertation evaluation. Ginzburg (2010) notes that excessive subjectivity can lead to arbitrariness, while Carr (2017) insists on the need for methodological rigor to ensure the validity of history as a source-based construction. These shortcomings create uncertainty for young scholars and limit the emergence of innovative approaches.

In teaching, the absence of common standards can produce inequality in student grading, making it difficult to establish objective parameters. Wineburg (2001) and Carretero and Gartner (2024) emphasize that evaluation should focus on historical thinking rather than subjective instructor interpretations. Without clear criteria, discrepancies in grading can affect student motivation, learning, and perceptions of fairness.

In dissemination, the lack of rigorous mechanisms can promote the spread of inaccurate or oversimplified narratives, undermining public trust in historical knowledge. Ricoeur (2008) warns that collective memory and history mutually influence each other, which can result in manipulated interpretations of the past, while Nora (1989) observes that "sites of memory" can reinforce biased views if not subjected to rigorous historiographic evaluation.

These limitations highlight the need to review and improve historiographic evaluation methods. Promoting more objective, replicable, and transparent models would preserve the discipline's epistemological richness while reducing subjectivity and uncertainty in current practices.

2.2 CURRENT CHALLENGES IN HISTORIOGRAPHY REGARDING TEXTUAL EVALUATION

Contemporary historiography faces new challenges in the evaluation of texts, driven by methodological evolution, digital technologies, and current social demands. These changes require a critical reassessment of traditional models and the adoption of more systematic, structured, and transparent approaches that ensure methodological rigor, interpretative diversity, and ethical responsibility.

One of the main obstacles is the lack of universally accepted criteria that allow for objective and replicable evaluations of historical texts. Burke (2008) underscores that while methodological diversity enriches the discipline, it complicates comparisons across different contexts. This issue is especially relevant in international academic settings, where theoretical and methodological differences generate disagreements about the validity and relevance of scholarly work. The inherent diversity of historiography reinforces the urgency of establishing shared evaluative principles that respect this plurality without homogenizing it. Ginzburg (2010) insists that plurality must be accompanied by evaluation mechanisms that ensure methodological coherence, rigorous use of sources, and robust historical analysis. Defining minimum criteria is essential to guarantee transparency and effectiveness.

Equally important is the ethical challenge of ensuring fair and balanced historical representation. Narratives have often privileged certain voices while marginalizing the experiences of subaltern groups such as women, indigenous communities, and ethnic minorities. This bias affects not only the understanding of the past but also contemporary identities and political discourses. To address this, it is essential to include diverse sources, critically analyze hegemonic discourses, and challenge narratives that perpetuate inequalities (SCOTT, 1999; CHAKRABARTY, 2020). An ethical evaluation must ensure equitable representation, avoiding exclusions or ideological biases that distort interpretation.

Finally, the current context—marked by the proliferation of information in digital environments—has heightened the demand for objectivity and transparency in historiographic production. As Carr (2017) warns, objectivity in history does not mean the absence of interpretation, but rather the rigorous application of methods based on verifiable evidence. It is therefore crucial for historians to make explicit their methodologies and the criteria they use to evaluate both their own work and that of others. Moreover, public access to historiography has become a central issue in the digital age. As Briseño (2021) notes, democratizing historical knowledge involves not only the dissemination of rigorous research but also accessible explanations of evaluation and validation processes. To address this challenge, integrating technologies such as artificial intelligence into historiographic analysis could provide tools that ensure replicable evaluation criteria and clear quality metrics in academic production.

2.3 ARTIFICIAL INTELLIGENCE AS AN EPISTEMOLOGICAL CHALLENGE FOR CONTEMPORARY HISTORIOGRAPHY

The integration of Artificial Intelligence (AI) into contemporary historiography goes beyond its use as a mere optimization tool. It constitutes an epistemological challenge that forces us to rethink fundamental concepts of the discipline, such as authorship, interpretation, and the nature of the historical source (FRONTONI et al., 2024). Generative AI, by moving beyond simple data processing to create coherent narratives, ceases to be only an "assistant" and potentially becomes an "agent" in the construction of historical accounts. Today, machines already participate actively in the elaboration of stories, generating much of the everyday narratives at a global level (HUGHES-WARRINGTON; MARTIN; O'BRIEN, 2024). This positions algorithms within the Digital Humanities not only as tools that uncover patterns in large textual corpora, but also as actors that intervene in the synthesis and writing of history, opening a new frontier in digital historiography. Some scholars have even asked whether machines could be as effective as humans in crafting historical narratives, foregrounding the debate on authorship and agency in AI-assisted history writing. In short, algorithms are moving from being passive tools to becoming active agents in the elaboration of historical discourse.

This new scenario requires a renewed critical framework. As several authors warn, the apparent objectivity of AI models may conceal biases inherited from their training data, thereby perpetuating hegemonic or simplified views of the past (HUGHES-WARRINGTON; MARTIN; O'BRIEN, 2024). When AI systems are trained on partial or incomplete historical records, they risk amplifying inequities and consolidating problematic narratives for the future. Digital historian Jo Guldi (2023) has emphasized that applying computational methods without proper historiographical training is "dangerous", since many analysts lack awareness of what can go wrong when archives are biased or incomplete. Her proposal is to develop a "hybrid knowledge", that is, combining traditional historical methods with algorithmic analysis. Practically, this means fostering a stronger dialogue between historiographical tradition—with its emphasis on source criticism and contextual rigor—and new digital technologies. The question is no longer simply whether AI can help us write history, but how the computational logic of these tools is shaping our very understanding and representation of the past. As recent research stresses, historical knowledge production remains an interpretive process that cannot be reduced to algorithms or mechanical rules (HUGHES-WARRINGTON; MARTIN; O'BRIEN, 2024). Using AI in historiography thus requires reflecting critically on how historical knowledge itself is configured when automated processes of analysis and narrative generation are involved.

This debate is crucial for the training of twenty-first-century historians, who must learn to use these tools effectively while also questioning their implications. Human supervision is not merely a matter of quality assurance but a methodological imperative to ensure that technology enriches rather than impoverishes the complexity of historical interpretation. As Henriot (2025) argues, AI models do not replace the historian's expertise, but can expand their capacity to process and interpret large corpora only if rigorous quality controls, verification procedures, and academic standards are maintained. In other words, the critical intervention of the historian remains indispensable: only by incorporating disciplinary expertise—contextual, ethical, and interpretive—can AI applications contribute meaningfully to historical research, honoring the richness and diversity of the past instead of reducing it (FRONTONI et al., 2024). This combination of computational tools and strict historiographical scrutiny allows AI to enhance the exploration of the past without compromising the interpretive complexity that defines good historiography.

3 THEORETICAL FOUNDATIONS OF ARTIFICIAL INTELLIGENCE APPLIED TO HISTORIOGRAPHY

The introduction of artificial intelligence into the humanities is transforming traditional practices, such as the creation and evaluation of texts. In the field of history, these technologies open up new methodological possibilities for the automated generation of texts and the critical evaluation of historical content (CARRASCO-RODRÍGUEZ, 2023).

3.1 ARTIFICIAL INTELLIGENCE AND GPT GENERATIVE MODELS

Artificial intelligence, a branch of computer science focused on developing algorithms that simulate human cognitive processes, has made notable advances with generative natural language models. Based on Transformer architectures, these models use attention mechanisms to analyze complex word relationships, enhancing precision in language interpretation.

Among them, OpenAI's GPT models (Generative Pre-trained Transformer) stand out for their ability to generate coherent, original texts after training on vast amounts of internet-based text. The latest versions, such as GPT-4, handle hundreds of billions of parameters, greatly expanding their capacity to grasp linguistic nuances and complex narrative structures.

A crucial aspect for historiography is the ability to adapt these models through fine-tuning—a supervised process that allows general models to be specialized in a specific domain, incorporating methodological and ethical criteria defined by experts. This ensures they can meet the discipline's standards for methodological coherence and rigorous documentation.

However, it is important to stress that these models do not replace the interpretative work of historians. While they provide a powerful technological foundation for processing large volumes of information and detecting textual patterns, they always require critical human oversight to ensure validity, relevance, and consistency. The combination of GPT's technological capabilities with rigorous expert supervision is essential for achieving methodologically sound and ethically responsible results, especially in sensitive fields like historiography.

3.2 CURRENT STATE OF GPT APPLICATION IN ACADEMIC EVALUATION

The application of GPT generative models in academia, though still in its early stages, has shown notable progress in essay evaluation, formative feedback, and the generation of pedagogical content. In automated essay scoring, for instance, these models have begun to surpass previous methods in consistency and accuracy, particularly when detailed rubrics developed by expert instructors are employed (BUI; BARROT, 2024; HUSSEIN; HASSAN; NASSEF, 2019; LATIF; ZHAI, 2024; LEE; LATIF; WU; LIU; ZHAI, 2024; MIZUMOTO; EGUCHI, 2023). These applications provide immediate and objective feedback, freeing teachers to focus on more complex and creative aspects of learning.

Additionally, GPT-powered educational chatbots offer real-time, personalized feedback, identifying argumentative errors and suggesting specific improvements to enhance critical skills among students in the humanities and social sciences (BERTRAM; WEISS; ZACHRICH; ZIAI, 2021; REDONDO-DUARTE; MARTÍNEZ-REQUEJO; JIMÉNEZ-GARCÍA; RUIZ-LÁZARO, 2023; YIN; GOH; HU, 2024).

However, these tools have limitations. A key ethical concern is the risk of automation reducing student autonomy or impoverishing the evaluative process (JINCUÑA HUALLPA, 2023). Moreover, the quality of evaluations can be compromised by biases or limitations in training data, affecting their reliability in specific contexts (VARSHA, 2023). These concerns underscore the need for critical oversight to ensure technological tools complement—but do not replace—the interpretative and methodological roles of human instructors.

Despite these challenges, empirical evidence suggests that GPT models applied to academic evaluation hold considerable potential. With proper human supervision and methodological adjustments, they can significantly optimize processes that would otherwise be manual, time-consuming, and heavily influenced by subjectivity.

3.3 SPECIFIC CHALLENGES IN THE HISTORIOGRAPHICAL APPLICATION OF ARTIFICIAL INTELLIGENCE

The application of GPT models in historiography faces several specific challenges. The first is the need to ensure rigorous factual accuracy, given that these models can generate inaccurate data if they have not been trained on reliable historical sources.

Another significant challenge is interpretative bias: these models learn from vast amounts of text, primarily from Western contexts, which can reinforce Eurocentric or biased narratives. To counter this, fine-tuning processes are required to expose the models to diverse and balanced perspectives.

In addition, a deep understanding of complex historical contexts remains a problem for these systems, as historiography requires not only factual knowledge but also critical interpretation and nuanced analysis.

Finally, there are important ethical issues concerning authorship and academic integrity. It is crucial to define guidelines that regulate the role of artificial intelligence in historiographic production and ensure adherence to the ethical and methodological principles of the discipline. Active expert oversight is essential to ensure that these technologies complement, rather than replace, human judgment in historical research and teaching.

4 METHODOLOGICAL DESIGN OF THE CUSTOMIZED GPT ASSISTANT: CRITERIA AND PROCEDURES

This study proposes a methodological design based on explicit and rigorous criteria that guide the evaluative activity of the GPT assistant. The central goal is to ensure a systematic, transparent, and consistent evaluation of historical texts, complementing and enriching the expert judgment of the historian, but without replacing it.

4.1 DEFINITION AND JUSTIFICATION OF STRUCTURED HISTORIOGRAPHICAL CRITERIA

The proposed criteria for the customized GPT *Historical Text Evaluation Assistant* (https://chatgpt.com/g/g-67386e7e6dfc8191adc1f22b852a3be6-historical-text-evaluation-assistant) are organized into seven categories that allow for the assessment of the fundamental characteristics of a historiographic text: disciplinary, epistemological, ethical, technical, pedagogical, reliability, and practical utility criteria.

4.1.1 Disciplinary Criteria

The evaluation of historical texts requires a methodological framework that guarantees the rigor and solidity of historiographic analysis. Disciplinary criteria make it possible to assess the quality of a text by considering the accuracy of facts, the substantiation of causal relationships, appropriate contextualization, and the inclusion of plural and critical perspectives (BURKE, 2008; TOSH, 2015).

Historical concepts are divided into two levels: first-order and second-order concepts. First-order concepts include factual data (dates, names, places), whose precision is fundamental to the credibility of the text. Second-order concepts provide an essential analytical framework for interpreting historical processes. These concepts foster the development of argumentative and historical thinking skills in students and researchers (LEE; DICKINSON; ASHBY, 2004; SEIXAS; MORTON, 2012; ÁLVAREZ, 2023).

Among the second-order concepts, historical causality is essential for understanding cause-and-effect relationships and avoiding simplistic interpretations; continuity and change offer a dynamic view of processes; and context is indispensable for interpreting facts within their sociopolitical and cultural frameworks. Multiperspectivity promotes balanced interpretations and avoids biases (RÜSEN, 2005; RÜSEN, 2010).

Historical significance helps evaluate the relevance of events within broader temporal frameworks. The criterion of progress and decline facilitates a critical analysis of social transformations. Global interconnectedness underscores the relationship between local processes and global trends, highlighting transnational history (MAZOWER, 2012). The use of historical evidence is a key principle: any interpretation must be based on verifiable and consistent sources (BLOCH, 2001).

Historical agency focuses on the capacity of individuals and groups to influence history (SCOTT, 1999). The analysis of structures and systems highlights the impact of institutions and power dynamics (BRAUDEL, 1990). Conflict is understood as a driver of many transformations, and historical memory examines how memories of the past shape collective identities (NORA, 1989).

Finally, concepts such as structural change, the relationship between the local and the global, and crisis and resilience enable the examination of broad transformations, interdependencies, and the capacity for social adaptation (KOSSELLECK, 1993). These disciplinary criteria form the basis for a critical and methodologically robust evaluation of historical texts.

4.1.2 Epistemological Criteria

Epistemological criteria are essential for assessing the robustness of a historical text by evaluating its theoretical coherence, its limitations, and its positioning within academic debates.

First, the GPT assistant considers the alignment of texts with recognized historiographic approaches (THOMPSON, 2002; BRAUDEL, 1990; BURKE, 2008; GINZBURG, 2010).

Another key criterion is the identification of interpretative gaps, which makes it possible to acknowledge the limits of the analysis and the existence of documentary gaps or unresolved debates. This practice strengthens academic transparency and promotes future research.

Theoretical and methodological coherence is also a fundamental pillar (CARR, 2017). A text should explain why it adopts a particular approach and how it contributes to interpretation, avoiding the unsubstantiated mixing of incompatible categories.

Originality and historiographic contribution are essential: a text must offer novel interpretations and engage with current debates, not simply reproduce existing narratives (TOSH, 2015). Alongside this, reflexivity—understood as an awareness that all narration is mediated by the author’s context and theoretical framework (WHITE, 1992)—provides rigor and self-criticism.

Another criterion analyzed is the text's ability to address the complexity of historical processes, avoiding reductionisms or determinisms (HOBSBAWM, 2000; SEIXAS; MORTON, 2012). Finally, the integration of multiple disciplines enriches historical analysis, fostering a more comprehensive and well-founded interpretation.

4.1.3 Ethical and Representational Criteria

Ethical and representational criteria are fundamental for evaluating historical texts, as they assess whether these texts reflect diverse perspectives, avoid biases, and respect the cultural sensitivity of the groups studied. In this way, an inclusive and plural historiography is promoted (TROUILLOT, 1995; BURKE, 2008).

A key point is the incorporation of diverse perspectives. Traditional historiography has privileged narratives constructed by dominant groups, marginalizing the voices of women, indigenous communities, and popular sectors. Thompson (2002) highlights the importance of recovering popular experiences, while Scott (1999) emphasizes the gender dimension, and Trouillot (1995) warns how power shapes not only the historical discourse but also the silences within it.

Narrative neutrality is also essential: texts must avoid moralizing judgments or ideological biases that distort interpretation. While historiography inevitably involves interpretation, it must be based on critical sources and rigorous argumentation (RICOEUR, 2008). Bloch (2023) underscores the importance of countering biased narratives through critical stances.

Respect for cultural sensitivity requires the use of precise and respectful language, avoiding the legitimization of hegemonic discourses that marginalize other viewpoints. In this vein, Chakrabarty (2020) advocates for the decolonization of history, recognizing multiple forms of knowledge.

Recognizing historical biases—present both in sources and in narratives—is equally essential. White (1992) argues that historical discourse is conditioned by narrative structures, while Appleby, Hunt, and Jacob (1994) emphasize the need to be aware of one's own subjectivity and the context of knowledge production.

Finally, balance in representation avoids simplistic dichotomies of heroes and villains. Ginzburg (2010) stresses that history is complex, and understanding the motivations and contradictions of historical actors is crucial for offering more complete and rigorous interpretations.

4.1.4 Technical Criteria

Technical criteria are fundamental for evaluating historical texts, as they ensure clarity, coherence, and suitability for academic contexts. These aspects help determine whether the content is accessible and well-structured, ensuring its appropriateness for educational settings.

Narrative organization is a key element, as a text must present its introduction, development, and conclusion in a logical manner, with smooth transitions between sections and without abrupt jumps (RICOEUR, 2008). The structure can be chronological or thematic, depending on the analysis proposed.

Clarity and coherence are equally essential: ideas must be expressed precisely and unambiguously, using language appropriate for the target audience. Internal coherence requires that the parts of the text are logically connected, avoiding contradictions. In this regard, Koselleck (1993) highlights the importance of precise terminology that organizes the temporal experience.

The use of explicit and well-supported references is another pillar of historiography. Every text must be based on rigorous and verifiable sources or, in their absence, on solid inferences (BLOCH, 2023; CARR, 2017; BURKE, 2008). This methodological rigor helps prevent anachronistic or biased interpretations.

Stylistic consistency is also essential. Maintaining a uniform tone, consistent verb tenses, and coherent use of historical terminology reinforces credibility and facilitates reading. Chartier (2021) emphasizes that history is an act of communication, requiring clear and structured language.

In the digital environment, compatibility and accessibility become highly relevant. Texts must be in appropriate formats, optimized for navigation and reading on mobile devices, including headings, tags, and alternative descriptions for images, which promote pedagogical use and inclusion.

Finally, the GPT assistant is capable of considering the integration of recognized historiographic perspectives, evaluating the depth and coherence of relevant historiographic approaches. White (1992) points out that historical narration involves a structured interpretation, while Ginzburg (2010) emphasizes the value of microhistory and Thompson (2002) highlights the need to recover subaltern voices. Veyne (1984) underscores the narrative and interpretative dimension of history, which requires constant methodological debate. The assistant recognizes and weighs these approaches, ensuring a more rigorous and methodologically grounded evaluation. In this way, it analyzes not only the form but also the fidelity to historiographic debates and the argumentative solidity of the text.

4.1.5 Pedagogical Criteria

Pedagogical criteria are fundamental for evaluating the effectiveness of educational texts as learning tools, ensuring they facilitate understanding, promote critical reflection, and adapt to diverse audiences and contexts.

Suitability for the educational level is key and involves adjusting complexity, language, and depth to the cognitive capacities of the audience (COLL, 2017). Well-designed materials must consider the intellectual maturity of the reader to ensure relevance and accessibility (PIAGET, 2005). Clarity and thematic relevance are equally essential, with texts needing to define key concepts and connect with curricular objectives or audience interests. Ausubel (2002) highlights that meaningful learning occurs when new knowledge integrates with what the student already knows, while Bruner (1984) emphasizes that structure should facilitate understanding and the construction of relationships between content.

Adaptability to different audiences is also crucial. Texts should be simplified for basic levels or enriched for advanced ones, serving as "scaffolding" to support student progress (VYGOTSKY, 1978). This dialogic approach encourages active knowledge appropriation (FREIRE, 2005). Promoting critical thinking is another essential element: beyond merely transmitting information, texts should stimulate questions, source comparison, and reflection on historiographic approaches. Wineburg (2001) argues that history teaching should build historical consciousness, encouraging the evaluation of established narratives, while Rüsen (2005) advocates for a shift from a traditional perspective toward a critical historical consciousness that interprets the past in relation to present and future impacts.

Finally, the use of didactic resources and practical activities helps explain complex concepts and promotes active learning, enriching history teaching by moving beyond rote memorization to foster analytical and argumentative skills.

4.1.6 Reliability Criteria

Reliability criteria serve to evaluate the quality of the sources used, the transparency in presenting information, and the accuracy of historical content. Applying these principles is essential to prevent the spread of historiographical errors and to ensure that the texts analyzed are grounded in verifiable evidence.

One of the fundamental aspects of this analysis is source transparency, which involves verifying whether the text explicitly cites its references, indicating their origin and type (primary, secondary, academic, popular, etc.). According to Bloch (2023), a well-grounded historical work should include citations and bibliographic references—or, in the case of texts generated by artificial intelligence, a clear indication of the training framework that supports the information provided. The absence of source transparency compromises the credibility of the analysis, making it difficult to verify the data and assess its reliability.

Another essential criterion is source reliability. It is not enough for a text to include references; those references must also be recognized and appropriate for the subject at hand. According to Carr (2017), rigorous historiographical analysis should be based on primary sources whenever possible, complemented by high-quality academic studies, while avoiding unverified or methodologically questionable materials. The use of secondary sources must be accompanied by critical analysis in order to avoid reproducing errors or biased interpretations without proper scrutiny.

Acknowledgment of limitations is another key aspect in assessing the reliability of a text. Historiography is built upon sources that are often fragmentary, partial, or open to interpretation. A well-crafted historical text should explicitly acknowledge these limitations, identifying documentary gaps, source biases, and methodological challenges that may affect the analysis. This becomes particularly relevant in topics where the available information is incomplete or contradictory, allowing readers to understand the scope and limitations of the study presented.

Factual validation is another indispensable criterion, referring to the verification of whether the data presented in the text are accurate and consistent with recognized sources. Reliable content should avoid factual errors, anachronisms, or distortions of historical facts. For White (1992), the correct dating of events, precise identification of historical figures, and coherence in the presented information are essential indicators of factual solidity.

Lastly, consistency between sources and analysis is a criterion that helps determine whether the interpretation developed in the text aligns with the evidence used. A rigorous historiographical analysis must establish a clear connection between the arguments presented and the sources that support them, avoiding arbitrary interpretations or claims without documentary support (Tosh, 2015). When a text presents novel hypotheses or interpretations, it must justify them through a solid critical apparatus that validates their credibility.

4.1.7 Practical Utility Criteria

Practical utility criteria make it possible to assess whether a text is applicable in research, teaching, or dissemination, ensuring its relevance and adaptability to specific objectives.

Thematic relevance is key: it means that the content addresses significant issues for the audience, responding to the interests of the academic and student community (CHARTIER, 2021). In this regard, Tosh (2015) emphasizes that history teaching should foster critical reflection on historical processes in relation to current issues.

Ease of adaptation is another essential aspect. Texts should be adjustable to different levels of comprehension, from students to researchers (WINEBURG, 2001), and should employ flexible and accessible language.

Potential for improvement is fundamental: a useful text should open up questions and suggest future lines of research (GINZBURG, 2010; BLOCH, 2023). In this way, the content not only transmits information but also fosters curiosity and the pursuit of new perspectives.

Connection to practical applications is also decisive. Historical texts should stimulate narrative and critical skills through activities such as debates, comparative studies, or reflective exercises (RÜSEN, 2005).

Finally, a useful text should contribute to the development of historical thinking and stimulate research by presenting clear explanations and solid arguments that promote a deeper understanding of historical processes (KOSELLECK, 1993).

4.2 METHODOLOGICAL PROCESS: WORKFLOW OF THE GPT ASSISTANT

The *Historical Text Evaluation Assistant* operates through a structured workflow. The process begins with the initial configuration, during which the assistant asks the user for key information about the text to be evaluated. At this stage, the user provides the document or fragment to be analyzed and defines the purpose of the analysis—whether for research, teaching, or dissemination. In addition, the user specifies whether the text was produced by a human or generated by artificial intelligence, and indicates the target audience along with the intended educational or academic level. Finally, the user selects which of the seven categories of criteria they wish the assistant to evaluate.

Once the analysis is configured, the assistant proceeds to evaluate the text by applying the selected criteria. During this phase, it examines each aspect of the content individually, identifying its strengths, weaknesses, and possible areas for improvement. The assistant conducts a quantitative assessment, assigning scores on a scale from 0 (minimum) to 10 (maximum), and complements these evaluations with qualitative comments.

Upon completing the evaluation, the assistant can generate two types of result reports: a detailed report by categories, which allows for reviewing each criterion separately, or a general report that provides a unified analysis in a single response. In both cases, a weighted score is provided along with a set of specific recommendations to improve the text. In the detailed report, the user has the opportunity to review and discuss each category before moving on to the next, whereas the general report synthesizes the findings in a more compact structure.

A fundamental feature of the assistant is its iterative approach, which enables the progressive improvement of the text based on the results obtained. Once the report is received, the user can modify the text according to the assistant's suggestions and, if desired, resubmit it for evaluation after making the adjustments.

Finally, the assistant offers the option to export the final report in various formats, such as DOCX or PDF. Depending on the user's needs, they can choose a detailed report, which documents the evaluation process step by step, or a global report, which summarizes the key findings and recommendations.

4.3 EVALUATION SYSTEM

The GPT assistant uses a weighted evaluation system that assigns specific weights to each of the criteria analyzed. The weighting model is structured into three levels. First, general weights are established for each of the seven evaluation categories: disciplinary criteria (25%), epistemological (15%), ethical (15%), technical (20%), pedagogical (15%), reliability (5%), and practical utility (5%). These weights reflect the relative importance of each category in a standard evaluation.

However, since historical texts may serve different purposes—research, education, or dissemination—the assistant adjusts the weightings to adapt to each context. In a research-oriented evaluation, for example, more weight is given to disciplinary (30%) and epistemological (20%) criteria, while pedagogical criteria have lesser importance (5%). In contrast, in an educational evaluation, technical (25%) and pedagogical (20%) criteria are prioritized. On the other hand, in a dissemination-oriented evaluation, ethical (20%) and technical (25%) criteria gain greater relevance, ensuring that the text is both understandable and respectful of diverse perspectives. Additionally, the assistant automatically adjusts the assigned percentages based on the evaluation criteria chosen by the user in the configuration phase.

Beyond these general weightings, each criterion has internal weightings that determine the relative importance of its subcriteria. For instance, in disciplinary criteria, 25% of the score corresponds to the accuracy of first-order concepts (facts, dates, and names), while 75% is allocated to second-order concepts. Within the latter, aspects such as causality, context, and continuity and change carry greater weight (15% each). Similarly, epistemological criteria prioritize alignment with historiographic approaches (25%) and theoretical and methodological coherence (20%), while in ethical criteria, the inclusion of diverse perspectives (30%) and narrative neutrality (25%) are the most relevant aspects. In technical criteria, narrative organization and clarity hold the highest weight (30% each), and in reliability criteria, transparency and the reliability of sources account for 60% of the evaluation. In pedagogical criteria, suitability for the educational level (30%) and clarity and thematic relevance (25%) are the aspects with the greatest weight in the evaluation. Finally, in practical utility criteria, thematic relevance (30%) and adaptability (25%) are prioritized. The system also includes adjustments to these internal weightings if the texts are intended for research, teaching, or dissemination purposes.

The weighting system enables the assistant to calculate an accurate score on a scale from 0 (minimum) to 10 (maximum), performing the necessary conversions from the percentages assigned to each category.

Moreover, it offers the possibility of making adjustments according to context, tailoring the evaluation to the different uses of historical texts. In this way, the assistant not only provides a detailed analysis but also ensures that the scores accurately reflect the quality of the text in relation to its specific purpose.

The specification of the percentages and weightings has been the subject of debate and consensus among the authors of the study, who have also collaborated with other history professionals from various universities. Since the question of weightings can be debated and requires broader consensus, the weighting configuration document can be easily and quickly modified or updated within the assistant.

4.4 APPLICABILITY OF THE GPT ASSISTANT TO HUMAN-AUTHORED AND AI-GENERATED TEXTS

The GPT assistant has been designed to evaluate both texts written by humans and those generated by artificial intelligence, applying common criteria but adapting them to the particularities of each case. Both types of content must meet historiographic standards: factual accuracy, argumentative coherence, and balanced representation, ensuring their usefulness in teaching, dissemination, or research.

However, AI-generated texts present certain peculiarities. The transparency of sources is limited, as these texts often do not explicitly disclose the origin of the information, requiring the assistant to conduct a more rigorous analysis of the plausibility of the data. Moreover, the narrative coherence of these texts may establish historical relationships that are not always correct, necessitating careful review of causality and the validity of arguments.

Another distinctive aspect of AI-generated texts is their limited interpretative capacity: they lack the critical awareness and intentionality of historians, as they reproduce statistical patterns without a real understanding of the past. Therefore, the assistant checks whether these texts present critical analysis or merely replicate conventional narratives.

In contrast, human-generated texts exhibit subjectivity and a personal style that can introduce conscious or unconscious biases. The assistant analyzes whether they maintain a balance between originality and fidelity to sources, and whether the interpretations are well-founded or merely opinions. It also assesses the reliability of the sources and the validity of interpretations, distinguishing between solid innovative narratives and those lacking historiographic support.

The assistant adjusts its analysis according to the type of text. In AI-generated texts, it emphasizes the verification of anachronisms, the plausibility of historical relationships, and narrative coherence. In human-generated texts, it focuses on the originality of the approach, methodological consistency, and the identification of interpretative biases. In this way, the model ensures an accurate and flexible evaluation, fostering improvements and guaranteeing the historiographic quality of the content.

5 RESULTS AND DISCUSSION

5.1 POTENTIAL OF THE GPT ASSISTANT IN HISTORIOGRAPHICAL EVALUATION

The implementation of a GPT assistant for historiographic evaluation offers a series of significant potentialities, but its use also entails technical, methodological, and ethical challenges that must be considered. This section discusses these aspects, analyzing the expected advantages of the model, its limitations, and the possible acceptance or resistance from the academic community. To begin with the potentialities, an assistant designed with structured criteria offers several key advantages over traditional, manual, and subjective evaluation. First, it ensures evaluative consistency: while traditional evaluation varies depending on the historian's preferences or institutional context, the GPT assistant applies an explicit rubric to produce systematic, coherent, and uniform evaluations. This improves comparability and facilitates standardization in academic settings.

Another significant benefit is the reduction in evaluation time. Manually reviewing extensive texts demands long hours that could be devoted to research or personalized teaching. The GPT assistant processes texts rapidly and generates preliminary evaluations automatically, freeing up time for experts to focus on deeper analysis. It also enhances objectivity and methodological transparency. Manual evaluation can be seen as opaque or arbitrary, especially by early-career researchers. With explicit criteria, the GPT assistant makes evaluations clearer and more justifiable, serving as a solid starting point for further human review.

Perhaps the most important advantage is its ability to perform rapid preliminary evaluations of large volumes of texts, which is particularly useful in large-scale educational contexts or broad historiographic research. Finally, its structured, well-founded report provides immediate and specific feedback on strengths, weaknesses, and areas for improvement, supporting learning and the development of methodological skills in students and early-career researchers. Together, these advantages position the GPT assistant as a valuable complement to the interpretative and critical work of historians, enhancing consistency, objectivity, transparency, and efficiency.

5.2 TECHNICAL AND METHODOLOGICAL LIMITATIONS OF THE GPT ASSISTANT IN HISTORIOGRAPHICAL EVALUATION

The use of the GPT assistant in historiographic evaluation presents technical and methodological limitations that must be critically addressed to ensure the practical effectiveness and academic validity of this tool. One of the main technical limitations lies in the quality and representativeness of the training data used. Since GPT models learn patterns from large corpora of texts, any bias or factual error present in these data can be reproduced in their evaluations. This issue is especially delicate in historiography, which demands sensitivity to historical context and to cultural and epistemological diversity.

Another methodological limitation concerns errors that may arise due to the lack of deep contextual understanding on the part of the AI. Although the assistant can generate coherent text, it lacks the critical interpretive capacity of an expert historian. This is particularly problematic when texts present subtle nuances or implicit references that require specialized knowledge.

Finally, the GPT assistant always requires expert human supervision, which limits the complete automation of the process and underscores the importance of the historian's role in the final validation of the analysis.

5.3 ETHICAL AND EPISTEMOLOGICAL CHALLENGES OF USING ARTIFICIAL INTELLIGENCE IN HISTORIOGRAPHY

The incorporation of the GPT assistant into historiographic evaluation raises ethical and epistemological challenges that must be carefully considered. Ethically, there is the question of responsible AI use and academic authorship. Although designed to support the historian, it is debated to what extent automated evaluations might partially replace human responsibilities. This calls for clear limits on academic AI use and guidelines on the ultimate intellectual responsibility for AI-generated evaluations.

Epistemologically, the challenge is to ensure that automated evaluation does not impoverish the discipline's theoretical and methodological diversity. Critical historiographic evaluation requires interpretative sensitivity that AI cannot always replicate. Therefore, the assistant must complement—but not replace—human interpretative work, preserving the discipline's richness.

These challenges call for critical reflection on the role of the GPT assistant, which should serve as an objective and efficient complement to strengthen historiographic work without diluting its ethical and methodological complexity.

5.4 POSSIBLE ACCEPTANCE AND RESISTANCE WITHIN THE ACADEMIC COMMUNITY

The incorporation of a GPT assistant in historiographic evaluation introduces a crucial social dimension: the acceptance or resistance of the academic community. As with any technological innovation, both support and criticism are expected.

Among the factors fostering positive reception is the growing pressure to optimize academic resources and enhance evaluative efficiency. Universities, research centers, and instructors face increasing demands for scientific productivity and educational outcomes, which could encourage the adoption of tools that streamline routine processes and improve transparency and objectivity.

Another valued advantage is the freeing up of time for more creative and in-depth activities, such as research, pedagogical development, or scientific dissemination. In a resource-limited environment, the partial automation of initial evaluations can be seen as an opportunity to redirect efforts toward more meaningful tasks.

However, alongside these advantages, resistance and criticism are also anticipated. A legitimate concern is that automation might displace human judgment, causing historians to feel that their intellectual autonomy, interpretative sensitivity, and critical capacity are threatened or diminished. This concern intensifies if AI is viewed as a substitute rather than a complement. Additionally, there may be criticisms regarding the reliability and transparency of AI evaluations: the "black box" nature of GPT models raises doubts about the clarity and explainability of their decisions, especially in a field as demanding as historiography.

Further ethical concerns arise from the dependence on technologies developed by private companies (such as OpenAI or Google), which could affect academic autonomy and the ethical and methodological control of evaluation processes. To address these concerns, it is crucial to emphasize that the tool does not replace human work but rather complements it under the active supervision of expert historians.

Therefore, the acceptance of this proposal will require the participation of the historiographic community at all stages of the design, implementation, and validation of the assistant. Well-founded criticisms should be seen as opportunities to strengthen transparency, improve the ethical and methodological quality of the model, and ensure an appropriate balance between automation and human oversight. While the GPT assistant offers advantages for historiographic evaluation, its acceptance will ultimately depend on addressing these concerns collaboratively.

6 FINAL CONCLUSIONS

This article has presented a methodological proposal for using a GPT assistant in historiographic evaluation through structured and explicit criteria. After analyzing its theoretical foundations, methodological design, expected results, and challenges, several key conclusions emerge.

First, it is clear that traditional historiographic evaluation—based on subjective and implicit criteria—has limitations that affect the quality and efficiency of historical analysis. Subjectivity generates inconsistencies and biases that hinder replicability and obstruct the establishment of evaluation standards.

In response, this proposal offers a structured model supported by clearly defined historiographic criteria and a GPT assistant trained to perform objective, systematic, and replicable preliminary evaluations of historical texts. This approach does not aim to replace the historian's work but to complement it, freeing up time for more complex tasks.

Another important conclusion is that, if used responsibly and critically, artificial intelligence can enhance methodological transparency, objectivity, and efficiency in historical text evaluation. Its ability to process large volumes of information with clear criteria provides unprecedented levels of replicability and transparency, especially valuable in higher education, multinational research, and public dissemination.

Nevertheless, significant methodological, technical, and ethical challenges remain. The quality and representativeness of the assistant's training data are crucial to avoid cultural, historical, or ideological biases. Expert human supervision is also indispensable to correct errors, ensure factual accuracy, and interpret complex nuances that AI cannot fully master.

These challenges underscore the need for interdisciplinary collaboration between historians and AI researchers, as well as continuous ethical and methodological oversight to ensure academic quality in evaluation processes.

In short, this work proposes a model that, by offering a practical and methodologically rigorous response to the epistemological challenges posed by the irruption of artificial intelligence, can serve as a valuable tool for strengthening the historical discipline in its academic, educational, and social dimensions. The key lies in embracing AI not as a threat but as a critical and ethical complement to improve the quality and transparency of historiographic evaluation.

7 FUTURE PERSPECTIVES AND OPEN LINES OF RESEARCH

Building on this methodological model, several future perspectives arise that will enable the validation and refinement of artificial intelligence in historiographic evaluation.

First, empirical validation is essential. Pilot studies in universities and educational centers are needed to compare the GPT assistant's evaluations with those by human experts. This will provide solid evidence of its practical applicability and inform adjustments to the proposed criteria.

Another crucial line of research involves technically improving GPT models through fine-tuning specifically for historiography. This requires selecting rigorous historiographic sources, creating diverse and representative datasets, and eliminating historical or cultural biases, all through interdisciplinary collaboration among historians, educators, and AI specialists.

Advancing interpretative transparency and explaining the GPT assistant's decision-making are also vital. Although these models produce coherent results, their internal logic remains opaque. Investigating mechanisms to clarify automated evaluations will be key to strengthening academic trust and epistemological legitimacy.

Finally, studying the perception and acceptance of this tool among historians and students is important. Surveys and qualitative studies can identify cultural or methodological barriers, fostering interdisciplinary collaboration and ensuring the critical and ethical integration of AI into historiography.

8 PRACTICAL RECOMMENDATIONS FOR IMPLEMENTING ARTIFICIAL INTELLIGENCE IN ACADEMIC HISTORIOGRAPHY

In addition to future perspectives, it is essential to establish practical recommendations to effectively implement the GPT assistant in academic contexts, ensuring ethically and methodologically sound results. Constant human supervision is crucial, as the assistant should be integrated into a hybrid model in which the automated preliminary evaluation is reviewed, validated, and enriched by expert historians. Specific training for historians is also essential to ensure they can use AI ethically, critically, and effectively. The creation of explicit rubrics, designed in collaboration with experts in didactics, ethics, and history, will provide clear, consensual, and ethically responsible criteria that reinforce the quality and validity of automated evaluations. The model should be subject to continuous evaluation and adjustment, informed by feedback from diverse national and international contexts. Additionally, clear guidelines on the ethical and academic use of AI must be established, ensuring responsible authorship and methodological integrity. The historiographic community must lead this process, ensuring that AI complements—rather than replaces—the critical and ethical richness of the discipline. In short, while the proposed methodological model offers an opportunity to improve the quality and transparency of historiographic evaluation, its practical success will depend on how limitations, ethical concerns, and the active integration of expert supervision are managed. Only in this way can AI contribute to enriching the quality and social relevance of historical knowledge.

ACKNOWLEDGMENTS

The authors would like to thank the Agencia Nacional de Investigación y Desarrollo de Chile (ANID), Proyecto Fondecyt de Iniciación 2023 en Investigación, Folio 11230035, “Evaluation of Historical Thinking in Future Primary and Secondary Education Teachers through the Construction of Historical Narratives on the Social Uprising (2019–2022)”.

REFERENCES

ÁLVAREZ, H. El laboratorio histórico como estrategia de indagación para desarrollar el pensamiento histórico en la formación del profesorado de Historia. **Interciencia**, v. 48, n. 5, p. 245-251, 2023.

APPLEBY, J.; HUNT, L.; JACOB, M. **Telling the truth about history**. New York: W.W. Norton, 1994. 336 p.

AUSUBEL, D. **Adquisición y retención del conocimiento: Una perspectiva cognitiva**. Barcelona: Paidós, 2002. 328 p.

BELIEIRO, T. Jamais fomos decoloniais: A crítica à história eurocêntrica na historiografia do ensino de história a partir das apropriações e interlocuções com as teorias pós-coloniais e o pensamento decolonial (2015-2024). **Revista de Teoria da História**, v. 27, n. 2, p. 41-64, 2024.

BERTRAM, C. et al. Artificial intelligence in history education: Linguistic content and complexity analyses of student writings in the CAHisT project (Computational assessment of historical thinking). **Computers and Education: Artificial Intelligence**, n. 100038, 2021. https://doi.org/10.1016/j.caeai.2021.100038

BLOCH, M. **Apologie pour l'histoire ou métier d'historien**. Paris: SHS Éditions, 2023. 140 p.

BRAUDEL, F. **La Historia y las Ciencias Sociales**. Madrid: Alianza Editorial, 1968. 221 p.

BRISEÑO, L. Los retos de la historia académica en la era digital. **Historia y memoria**, n. 22, p. 161-195, 2021.

BRUNER, J. **Actual minds, possible worlds**. Cambridge: Harvard University Press, 1987. 222 p.

BUI, N. M.; BARROT, J. S. ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring. **Education and Information Technologies**, v. 30, p. 2041–2058, 2025. https://doi.org/10.1007/s10639-024-12891-w.

BURKE, P. **What is Cultural History?** Cambridge: Polity, 2008. 179 p.

CARR, E. H. **¿Qué es la historia?** Barcelona: Ariel Historia, 1998. 228 p.

CARRASCO-RODRÍGUEZ, A. Reinventando la enseñanza de la Historia Moderna en Secundaria: La utilización de ChatGPT para potenciar el aprendizaje y la innovación docente. **Studia Histórica, Historia Moderna**, v. 45, n. 1, p. 101-145, 2023. https://doi.org/10.14201/shhmo2023451101146

CARRETERO, M.; GARTNER, E. Artificial Intelligence and historical thinking: A dialogic exploration of ChatGPT. **Studies in Psychology**, v. 45, n. 1, p. 80-102, 2024. 10.1177/02109395241241379

CHAKRABARTY, D. **Provincializing Europe: Postcolonial thought and historical difference**. Princeton: Afterall Books, 2020. 301 p.

CHARTIER, R. **El mundo como representación: Estudios sobre historia cultural**. Barcelona: Gedisa, 1999. 286 p.

COLL, C. **La personalización del aprendizaje escolar**. México: Ediciones SM, 2017. 93 p.

DABAT, C. As áreas econômicas-chave segundo Ji Chaoding: o desafio discreto de um marxista chinês ao eurocentrismo historiográfico. **História da Historiografia: International Journal of Theory and History of Historiography**, v. 17, p. 1-29, 2024. https://doi.org/10.15848/hh.v17.2183

FREIRE, P. **Pedagogía del oprimido**. Madrid: Siglo XXI, 2023. 256 p.

FRONTONI, E. et al. Editorial: Artificial intelligence: the new frontier in digital humanities. **Frontiers in Computer Science**, v.6, 1529826, 2024. https://doi.org/10.3389/fcomp.2024.1529826

GINZBURG, C. **El hilo y las huellas: lo verdadero, lo falso, lo ficticio**. México: Fondo de Cultura Económica, 2010. 492 p.

GULDI, J. **Towards a Practice of Text-Mining to Understand Change Over Historical Time: The Persistence of Memory in British Parliamentary Debates in the Nineteenth Century**. Lecture at UC Berkeley, 8 Mar. 2023. [Transcript available via Social Science Matrix and D-Lab].

HENRIOT, C. **The AI-Augmented Research Process: A Historian's Perspective**. Preprint (arXiv:2508.01779), 2025.

HUGHES-WARRINGTON, M.; MARTIN, A.; O'BRIEN, L. Y. **Artificial Historians**. London: Routledge, 2025. 206 p.

HUSSEIN, M.; HASSAN, H.; NASSEF, M. Automated language essay scoring systems: A literature review. **PeerJ Computer Science**, v. 5, e208, 2019. https://doi.org/10.7717/peerj-cs.208

HOBSBAWM, E. **Historia del siglo XX**. Barcelona: Crítica, 2000. 656 p.

JINCHUÑA HUALLPA, J. Exploring the ethical considerations of using Chat GPT in university education. **Periodicals of Engineering and Natural Sciences**, v. 11, n. 4, p. 105-115, 2023. https://doi.org/10.21533/pen.v11.i4.200

KOSSELLECK, R. **Futuro pasado: Para una semántica de los tiempos históricos**. Barcelona: Paidós, 1993. 370 p.

LATIF, E.; ZHAI, X. Fine-tuning ChatGPT for automatic scoring. **Computers and Education: Artificial Intelligence**, v. 6, 100210, 2024. https://doi.org/10.1016/j.caeai.2024.100210

LEE, P.; DICKINSON, A.; ASHBY, R. Children's ideas about historical explanation. In: DICKINSON, A.; GORDON, P.; LEE, P. (Eds.). **International review of history education. Volume 3: Raising standards in history education**. London: Woburn Press, 2004. p. 97-115.

LEE, P. Putting principles into practice: Understanding history. In: DONOVAN, M. S.; BRANSFORD, J. D. (Eds.). **How students learn: History, mathematics, and science in the classroom**. Washington, D.C.: National Academies Press, 2005. p. 31-77.

LEE, G. G. et al. Applying large language models and chain-of-thought for automatic scoring. **Computers and Education: Artificial Intelligence**, v. 6, 100213, 2024. https://doi.org/10.1016/j.caeai.2024.100213

MASOOD, M. D. The effect of cultural turn in western historical discourse: Historiographic shift in truth, tropology, and subjectivity. **Journal of European Studies (JES)**, v. 41, n. 1, p. 17-17, 2025. 10.56384/jes.v41i1.361

MAZOWER, M. **Governing the world: The history of an idea, 1815 to the Present**. London: Penguin Books, 2012. 496 p.

MIZUMOTO, A.; EGUCHI, M. Exploring the potential of using an AI language model for automated essay scoring. **Research Methods in Applied Linguistics**, v. 2, n. 2, 100050, 2023. https://doi.org/10.1016/j.rmal.2023.100050

MOMIGLIANO, A. **The classical foundations of modern historiography**. Berkeley: University of California Press, 1992. 180 p.

NAGRE, K. (Mis)educating England: Eurocentric narratives in secondary school history textbooks. **Race Ethnicity and Education**, v. 28, n. 1, p. 134-153, 2025. https://doi.org/10.1080/13613324.2023.2192945

NORA, P. **Les lieux de mémoire**. Paris: Gallimard, 1989. 664 p.

PIAGET, J. **The psychology of intelligence**. London: Routledge, 2001. 202 p.

REDONDO-DUARTE, S. et al. The potential of educational chatbots for the support and formative assessment of students. In: IBRAHIM, M.; AYDOĞMUȘ, M.; TÜKEL, Y. (eds.). **New trends and promising directions in modern education: «AI in education»**. Konya: Palet Yayınları, 2023. p. 105-136. https://hdl.handle.net/20.500.14352/101695

RICOEUR, P. **La memoria, la historia, el olvido**. Madrid: Editorial Trotta, 2003. 688 p.

RÜSEN, J. **History: Narration, interpretation, orientation**. New York: Berghahn, 2005. 236 p.

RÜSEN, J. **Jörn Rüsen e o ensino de história**. Curitiba: Editora UFPR, 2019. 152 p.

SAMIRA, A. Employing artificial intelligence applications in historical research: Ancient Rome as a model. **Psychology and Education**, v. 61, n. 8, p. 604-617, 2024.

SCOTT, J. W. **Gender and the politics of History**. New York: Columbia University Press, 2018. 288 p.

SEIXAS, P.; MORTON, T. **The big six historical thinking concepts**. Toronto: Nelson, 2012. 200 p.

THOMPSON, E. P. **The making of the English working class**. Reino Unido: Penguin Books Limited, 1991. 960 p.

TOSH, J. **The pursuit of History**. New York: Taylor & Francis, 2015. 316 p.

TROUILLOT, M.-R. **Silencing the past: Power and the production of History**. Boston: Beacon Press, 1995. 191 p.

VARSHA, P. S. How can we manage biases in artificial intelligence systems – A systematic literature review. **International Journal of Information Management Data Insights**, v. 3, n. 1, 100165, 2023. https://doi.org/10.1016/j.jjimei.2023.100165

VEYNE, P. **Cómo se escribe la historia**. Madrid: Fragua, 1972. 367 p.

VYGOTSKY, L. S. **Mind in society: The development of Higher Psychological Processes**. Cambridge: Harvard University Press, 1978. 159 p.

WHITE, H. **Metahistoria: La imaginación histórica en la Europa del siglo XIX**. México: Fondo de Cultura Económica, 1992. 432 p.

WINEBURG, S. **Historical thinking and other unnatural acts**. Philadelphia: Temple University Press, 2001. 280 p.

YIN, J.; GOH, T. T.; HU, Y. Using a chatbot to provide formative feedback: A longitudinal study of intrinsic motivation, cognitive load, and learning performance. **IEEE Transactions on Learning Technologies**, v. 17, p. 1378-1389, 2024. https://doi.org/10.1109/TLT.2024.3364015