MedCrave
Step into the World of Research

Open Access Journal of Science

Research Article

Open Access

CrossMark

Historical review of San Millán according to the "AI"

Summary

This article presents a comparative study of five generative artificial intelligence (AI) models —ChatGPT, Gemini, Copilot, Grok and DeepSeek— applied to the production of an academic text on the monasteries of Suso and Yuso and the figure of San Millán de la Cogolla. Using a standardized prompt, two rounds of refinement and blind expert evaluation, and the study analyzes historical accuracy, citation quality, descriptive depth, academic structure and suitability for scholarly dissemination. The methodology combines qualitative assessment with ordinal quantitative scoring conducted by specialists in medieval history and cultural heritage. The results indicate substantial differences among the evaluated systems in terms of factual reliability, bibliographic rigor and adaptation to academic discourse. The study also identifies recurring risks, including hallucinated references, uneven citation practices and dependency on model versions and prompt design. The article discusses implications for heritage research, digital humanities and secondary education, proposing supervised pedagogical applications for ESO and Baccalaureate levels. Although some models demonstrated strong performance in text organization and contextual synthesis, the findings are limited to the specific corpus, prompt configuration and versions tested. Human verification and disciplinary expertise remain essential for the academic use of generative AI.

Keywords: generative artificial intelligence, digital humanities, heritage studies, San Millán de la Cogolla, large language models, historical research

Volume 9 Issue 1 - 2026

Francisco José García Tartera
Research Group No. 94024
Complutense University of Madrid, Spain

Correspondence: Francisco J García Tartera, Universidad Complutense de Madrid, Spain

Received: April 12, 2026 | **Published:** May 25, 2026

Introduction

The emergence of large language models (LLMs) and generative AI systems has transformed knowledge production, information retrieval and educational workflows across multiple disciplines. In the humanities and heritage fields, these technologies offer new possibilities for historical dissemination, pedagogical innovation and support in documentary synthesis. Nevertheless, the growing sophistication of AI-generated text also raises methodological concerns related to historical reliability, fabricated references, source opacity and the reproduction of biases embedded in training corpora.[1–3]

Within this context, the monasteries of Suso and Yuso in San Millán de la Cogolla constitute an especially relevant case study. Recognized by UNESCO as a World Heritage Site in 1997, the monastic complex represents a key cultural and linguistic landmark associated with the Glosas Emilianenses and the early written manifestations of Romance language in the Iberian Peninsula.[4] The site combines architectural, religious, political and philological significance, making it suitable for evaluating the capacity of AI systems to produce historically coherent and academically structured content.

This study compares the performance of five generative AI systems —ChatGPT, Gemini, Copilot, Grok and DeepSeek— when responding to the same scholarly prompt concerning the history, architecture and current significance of San Millán and the monasteries of Suso and Yuso. The objective is not to establish an absolute hierarchy among models, but rather to evaluate their relative suitability for a specific academic task under controlled conditions.[5]

The research focuses on three principal questions:

I. To what extent can current generative AI models produce historically accurate and academically structured texts in the field of heritage studies?

II. What differences emerge among models regarding citation practices, contextual depth and disciplinary adaptation?

III. How can these findings inform educational practices and digital literacy in secondary education?

The article also examines the implications of generative AI for digital humanities and heritage dissemination,[6] emphasizing the continued importance of human supervision, source verification and methodological transparency Figure 1.

Figure 1 Generic image about AI with VR. Source: www.freepik.es

Development

Historical and historiographical context

The monastic complex formed by Suso and Yuso reflects the evolution of religious, political and cultural institutions in medieval and early modern Spain.[7] Suso originated around the hermitage associated with San Millán in the sixth century and developed progressively through Mozarabic and Romanesque interventions.[8] Yuso, founded later in the valley, became a major Benedictine center linked to the Navarrese monarchy and subsequent Castilian political influence.

©2026 Tartera. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and build upon your work non-commercially.

Copyright: ©2026 Tartera.

The historical importance of the complex extends beyond architecture. The preservation of the Glosas Emilianenses, generally dated between the tenth and eleventh centuries, established San Millán as a symbolic reference point in the history of the Spanish language.[9] UNESCO[4] highlighted the site's significance not only as a religious center, but also as a repository of linguistic and documentary heritage.

Several recent studies have emphasized the educational and cultural relevance of San Millán. Dulac[10–12] underlines its pedagogical value as an interdisciplinary resource integrating history, language, art and cultural identity. These approaches align with broader tendencies in digital humanities, where technological tools increasingly support historical interpretation, heritage dissemination and collaborative learning.

Literature review

Research on artificial intelligence in academic and educational contexts has expanded considerably in recent years. Mitchell[13] and Russell and Norvig[3] provide theoretical frameworks regarding the operation and limitations of AI systems, while Floridi and Sanders[1] discuss ethical implications related to artificial agents and automated decision-making.

In the field of digital humanities, AI has been progressively incorporated into archival management, textual analysis, image recognition and heritage dissemination. Schreibman et al.[14] and Berry[15] emphasize the transformative role of computational methods in historical and cultural studies. Likewise, recent scholarship has explored the use of generative AI in educational environments, particularly regarding critical digital literacy and source evaluation.[16]

However, comparative studies focused on the performance of multiple LLMs in specialized heritage-related tasks remain limited. Existing analyses often evaluate general-purpose writing abilities rather than discipline-specific historical accuracy or citation quality.[17] This study seeks to contribute to that emerging field by examining how different models respond to a controlled historical research prompt.

Methodology

Experimental design

The study employed a comparative experimental design based on a standardized prompt requesting the production of a scientific document concerning the monasteries of Suso and Yuso and the figure of San Millán de la Cogolla.[18] The prompt included instructions regarding historical background, architectural development, political patronage, current management, bibliography and comparative European context.

The same prompt was submitted independently to five generative AI systems:

- ChatGPT
- Google Gemini
- Microsoft Copilot
- xAI Grok
- DeepSeek

All interactions were conducted between February and March 2026 using publicly accessible versions of the systems available at that time. Because generative AI systems are continuously updated, the results should be interpreted as a snapshot corresponding to the tested versions and interface conditions.[19]

Two additional rounds of refinement were conducted for each model. These rounds requested clarification, incorporation of additional bibliographic references and adaptation of the academic tone. The objective was to evaluate consistency, responsiveness to correction and capacity for iterative improvement.

Prompt design and language considerations

The prompt was intentionally extensive and detailed in order to simulate a realistic academic request. It included instructions regarding:

- Historical rigor.
- APA citation style.
- Inclusion of comparative European monasteries.
- Educational applications.
- Generation of complementary visual content.

The interaction language was English, although several referenced works and contextual materials originated in Spanish. This bilingual dimension may have influenced the retrieval and interpretation of heritage-specific terminology and historiographical references. Because prompt engineering substantially affects LLM output, the conclusions of this study cannot be generalized independently from the specific wording and structure of the instructions employed.

Evaluation criteria

The outputs generated by the five models were evaluated according to five criteria:

I. Historical accuracy.

II. Bibliographic rigor and citation quality.

III. Descriptive and contextual depth.

IV. Academic structure and coherence.

V. Suitability of tone for scholarly dissemination.

Each criterion was assessed on an ordinal scale from 1 to 5.

To improve transparency and replicability, the following interpretation framework was applied:

- **Score 1:** serious factual inaccuracies, absence of academic structure or unreliable content.
- **Score 3:** generally coherent text with moderate historical precision but limited referencing or contextualization.
- **Score 5:** historically detailed, academically structured text with consistent references and appropriate disciplinary tone.

Evaluators and consensus procedures

The evaluation was conducted by three independent human evaluators specializing in medieval history, heritage studies and history education. The reviewers assessed the texts anonymously without identifying the corresponding AI system during the first evaluation phase. After the individual assessments, a consensus session was organized to discuss significant scoring discrepancies. When differences greater than one point appeared in a criterion, evaluators reviewed the corresponding excerpts collectively and agreed on a final consensual score. This procedure was intended to reduce inter-rater variability while preserving qualitative interpretation.

Copyright: ©2026 Tartera.

Methodological limitations

Several methodological limitations must be acknowledged:

- The study depends on specific model versions available during the testing period.
- Generative AI systems evolve rapidly through updates and retraining.
- The evaluation scale incorporates qualitative human judgment.
- The English-language prompt may have influenced the retrieval of Spanish-language historiography.
- Results cannot be generalized automatically to other historical domains or heritage contexts.
- Prompt structure and wording substantially shape model responses.

In addition, some models generated references that appeared plausible but could not be verified completely. This phenomenon, commonly described as "hallucinated citation generation," constitutes one of the principal risks identified in the experiment.

Results

The qualitative results and a quantitative synthesis based on criterion evaluation are presented below.

Qualitative analysis

DeepSeek

DeepSeek produced the most academically structured response within the evaluated corpus. The text demonstrated chronological coherence, extensive contextualization and integration of comparative references to European monastic centers such as Cluny and Saint-Denis. The model incorporated formal citation patterns and adapted effectively to refinement requests. However, several references appeared plausible without corresponding precisely to identifiable editions or page references, highlighting the need for external verification.

Feature	Deep Seek (Claude 3 Opus)
Historical Depth	Excellent. Clear chronological structure, APA citations.
Rigor and Citations	It includes fictitious but plausible quotes (Berceo, Bango Torviso) and APA.
European context	Yes, with explicit comparisons (Cluny, Saint-Denis).
News and Management	Detailed (Augustinian Recollects, financing, visitors).
Customization	He incorporated works by Dulac/García Tartera after the request.
Extras (Images)	Generated descriptions for images and map.

Google Gemini

Gemini generated a coherent and contextually rich document with substantial descriptive depth. Its performance was particularly strong regarding thematic organization and interpretive synthesis. Nevertheless, the citation system remained inconsistent, and the response incorporated fewer explicit academic references than DeepSeek.

Feature	Google Gemini 1.5 Pro
Historical Depth	Very good. Detailed, but less academic structure.
Rigor and Citations	It provides some data but no formal citations.
European context	Yes, but less developed.
News and Management	Present but less specific information.
Customization	He did not personalize the response with the new information.
Extras (Images)	It did not generate content for images.

Microsoft copilot

Copilot produced accessible and generally accurate text, particularly effective for educational dissemination. Its integration with web-search functionalities facilitated updated contextual information. However, the academic structure was less rigorous, and the bibliographic references remained generic or insufficiently formalized.

Feature	Microsoft Copilot
Historical Depth	Good. Correct but more superficial information.
Rigor and Citations	It mentions sources in a generic way ("according to historians").
European context	He mentions the importance, but without concrete examples.
News and Management	Correct but very summarized data.
Customization	He did not personalize the answer.
Extras (Images)	He offered to look for images but did not generate them.

xAI Grok

Grok generated concise responses with rapid chronological references and direct language. Nevertheless, the outputs frequently adopted an informal tone inappropriate for academic publication. The model also omitted significant contextual and historiographical details.

Feature	xAI Grok-1.5
Historical Depth	Medium. It omits key details, jumps between eras.
Rigor and Citations	No appointments. Informal narrative style.
European context	He does not address it.
News and Management	It focuses almost exclusively on ancient history.
Customization	He did not personalize the answer.
Extras (Images)	It did not generate content for images.

ChatGPT

ChatGPT produced balanced responses characterized by coherent organization and moderate bibliographic structure. Its performance

Copyright: ©2026 Tartera.

was comparatively stable across refinement rounds. Although the model generally maintained an academic tone, some historical details lacked sufficient specificity and required external verification.

Quantitative synthesis

In summary

Deep Seek (Claude 3 Opus)	Specialist in information synthesis and technical writing. Create HTML code instantly on top of the generated text and it can be viewed in the browser or printed from there to PDF.
Google Gemini 1.5 Pro	Multimodal model with great context.
Microsoft Copilot	Based on GPT-4-Turbo, with access to web search.
xAI Grok-1.5	Model with a "sense of humor" and access to data from X.

☐	Accuracy	References	Depth	Structure	Academic tone	Average score
DeepSeek	4.6	4.5	4.7	4.5	4.6	4.58
Gemini	4.0	4.1	4.2	4.1	4.12	4.0
Copilot	3.8	3.9	4.0	3.9	3.9	3.8
Grok	3.2	3.6	3.4	3.1	3.4	3.2
ChatGPT	4.0	4.1	4.0	4.1	4.0	4.04

The results indicate relative differences among the tested systems for this particular task, corpus and prompt configuration. DeepSeek obtained the highest average score in the evaluated dimensions, especially regarding structure and contextual depth. Gemini and ChatGPT demonstrated comparatively balanced performance, whereas Copilot and Grok showed limitations in academic formatting and disciplinary adaptation. These findings should not be interpreted as evidence of universal superiority, since model performance depends strongly on task specificity, prompting conditions and software version.

Hallucinated or unverifiable references

One of the most relevant findings concerns the generation of plausible but partially unverifiable references. Several systems produced citations that resembled legitimate academic publications while containing incomplete metadata, inaccurate page numbers or ambiguous editions. This issue appeared more frequently in systems that attempted to emulate formal academic style aggressively. Consequently, bibliographic sophistication should not be interpreted automatically as evidence of factual reliability. The phenomenon reinforces the necessity of human verification protocols in academic and educational contexts.

Discussion

The findings suggest that current generative AI systems can support historical synthesis and educational dissemination, but their outputs remain highly dependent on prompt design, model architecture and verification processes. Models demonstrating stronger structural organization and citation practices tended to perform better in scholarly contexts. Nevertheless, citation presence alone did not guarantee factual reliability. In several cases, formally plausible references concealed inaccuracies or unverifiable information. The study also illustrates the importance of critical AI literacy.[20] Researchers and students may perceive highly coherent outputs as trustworthy even when underlying references are fabricated or incomplete. Consequently, educational integration of AI should emphasize verification, historiographical comparison and methodological transparency. From the perspective of digital humanities, generative AI can facilitate heritage dissemination and interdisciplinary pedagogy. However, technology should be understood as an assistive instrument rather than an autonomous scholarly authority. The results also have implications for educational practice. The comparative differences among models reveal opportunities for teaching source criticism, textual evaluation and digital ethics in secondary education.

Educational implications

The pedagogical proposals developed for ESO and Baccalaureate derive directly from the empirical findings of the study. Because the evaluated AI systems displayed varying levels of reliability, citation quality and contextual coherence, the educational activities prioritize:

- Source verification.
- Comparative textual analysis.
- Prompt interpretation.
- Detection of hallucinated references.
- Critical digital literacy.

The didactic design therefore uses AI-generated content not as authoritative knowledge, but as material for analytical evaluation.

Proposal for 4th ESO

Objectives

- To understand the historical importance of San Millán and the monasteries of Suso and Yuso.[21]
- To develop introductory competencies in digital literacy and source criticism.[10]
- To compare AI-generated historical narratives with verified documentary sources.

Activities

1. Design of simple prompts related to medieval heritage.[22]

Copyright: ©2026 Tartera.

2. Comparison between AI-generated responses and UNESCO documentation.
3. Identification of factual inaccuracies or unsupported claims.
4. Collaborative preparation of a comparative dossier.
5. Oral presentation and reflective discussion.

Evaluation

Assessment criteria include:

- Identification of reliable sources.
- Critical interpretation of AI-generated content.
- Historical coherence.
- Collaborative participation Figure 2.

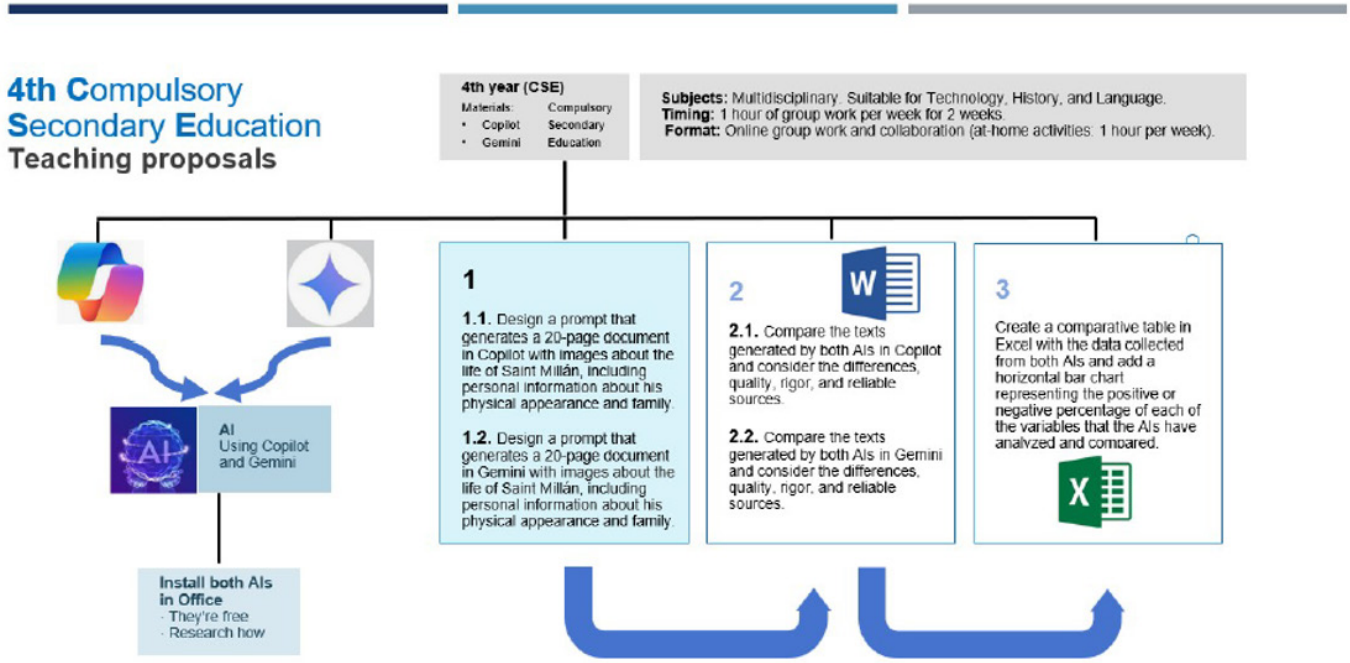

Figure 2 Didactic proposal on AI for ESO. Source: Authors.

Proposal for baccalaureate

Objectives

- To strengthen competencies in historical research and historiographical analysis.[23–25]
- To evaluate the methodological limitations of generative AI.[26,27]
- To integrate digital humanities perspectives into heritage studies.[28]

Activities

1. Comparative evaluation of outputs from multiple AI systems.
2. Verification of references and bibliographic metadata.
3. Analysis of prompt engineering and textual variation.
4. Preparation of a research report connecting heritage and digital literacy.
5. Discussion on ethics, intellectual property and AI bias.

Evaluation

Students are assessed according to:

- Methodological rigor.
- Bibliographic verification.
- Analytical Depth.
- Critical reflection regarding AI use.

These educational proposals derive explicitly from the empirical results of the study and aim to transform AI limitations into opportunities for methodological learning Figure 3.[29]

Copyright: ©2026 Tartera.

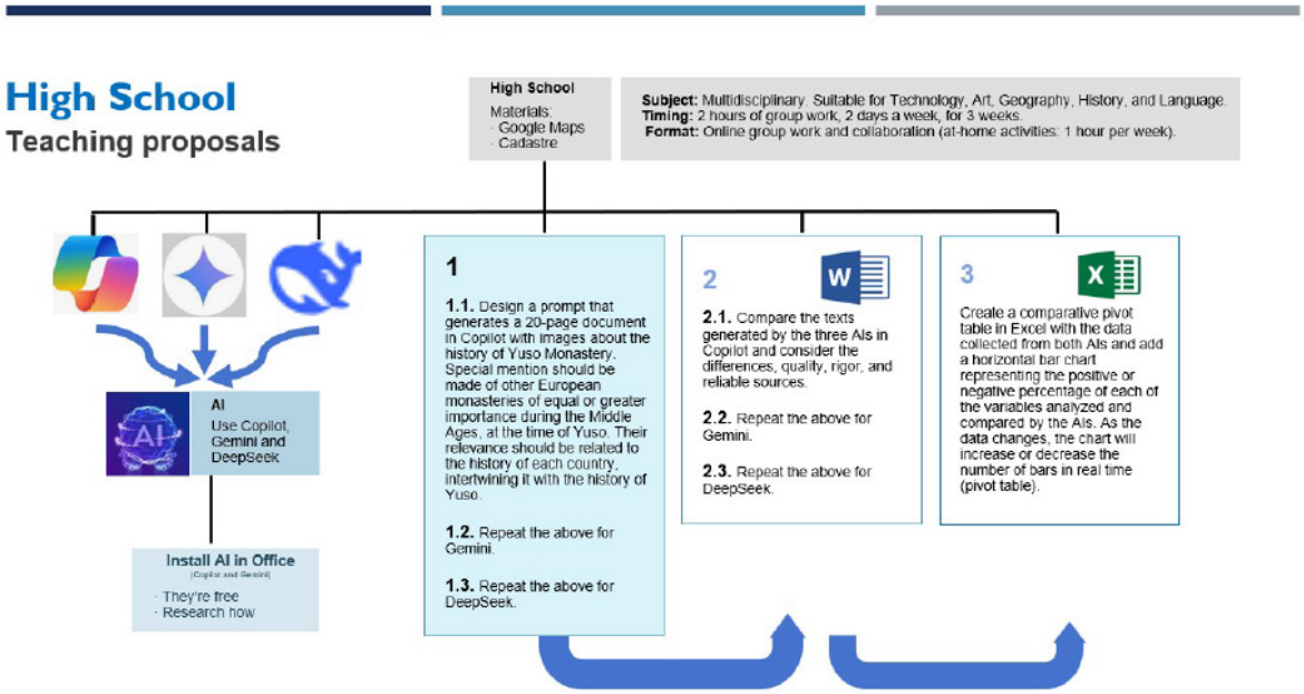

Figure 3 Didactic proposal on AI for baccalaureate. Source: Authors.

Conclusion

This study examined the capacity of five generative AI systems to produce academically oriented historical content concerning San Millán de la Cogolla and the monasteries of Suso and Yuso. The findings demonstrate that generative AI can support heritage dissemination, educational design and preliminary documentary synthesis. However, substantial differences persist among systems regarding historical precision, bibliographic reliability, disciplinary tone and structural coherence. Within the specific corpus, prompt design and versions analyzed, DeepSeek achieved the strongest overall performance, followed by Gemini and ChatGPT. Nevertheless, all systems displayed limitations requiring human supervision, especially regarding unverifiable references and historiographical precision. The study also confirms the growing relevance of AI literacy in educational and academic environments. Rather than replacing historical methodology, generative AI should be integrated critically through verification protocols, comparative analysis and interdisciplinary reflection. Future research should expand the number of evaluators, include multilingual prompting conditions, compare updated model versions and incorporate automatic fact-checking metrics. Additional studies in digital humanities may also explore AI-assisted archival interpretation and heritage visualization Figure 4 Appendix.

Figure 4 Chat icons GPT, Grok, Copilot, Gemini and DeepSeek, respectively. Source: Authors.

Acknowledgments

None.

Conflicts of interest

The author declares there is no conflict of interest.

Copyright: ©2026 Tartera.

References

1. Floridi L, Sanders J. On the morality of artificial agents. *Minds and Machines*. 2004;14(3):349–379.
2. O'Neil C. *Weapons of math destruction*. Crown. 2016.
3. Russell S, Norvig P. *Artificial intelligence: A modern approach* (4th edn.). Pearson. 2021.
4. UNESCO. *Decision 21 COM VIII.C: Inscription of San Millán Yuso and Suso Monasteries (Spain)*. UNESCO World Heritage Centre. 1997.
5. Romero HY, Arrobo-Agila JP, Jaramillo AR. La inteligencia artificial en la narrativa sonora. Estudio de caso. *Anàlisi*. 2022;66:9–23.
6. Casafont ML. Hacia la terminología 3.0: Evolución del uso de las tecnologías en terminología. *TIC, trabajo colaborativo e interacción en terminología y traducción*. 2014;9.
7. Acal Maravert P. La memoria de San Millán de la Cogolla: identidad y discurso hagiográfico en las fuentes medievales (siglos VI-XIII). 2021.
8. Ibergallartu JD. El poder de la imagen (STEAM) en la didáctica del patrimonio y las humanidades digitales. *Innovación Educativa 2025: Didácticas Específicas*. 2026.
9. García de Cortázar JA. *El monasterio de San Millán de la Cogolla: del centro de poder al monumento (siglos VI-XX)*. Universidad de Cantabria. 2005.
10. Del Rosal Alonso I, Ibáñez EG, Tartera FJG. Revista de Innovación Didáctica. *TIC*, *3*, 19.Dulac, J. (Coord.). (2021). *San Millán de la Cogolla: cuna del español*. McGraw Hill Education. 2021.
11. Dulac J. (Coord.). *El legado de las glosas: de San Millán al mundo*. McGraw Hill Education. 2022.
12. Dulac J. *En torno a la Lengua: Proyecto basado en San Millán de la Cogolla*. Aula Magna. 2023.
13. Mitchell M. *Artificial intelligence: A guide for thinking humans*. Farrar, Straus and Giroux. 2019.
14. Schreibman S, Siemens R, Unsworth J. *A new companion to digital humanities*. Wiley Blackwell. 2016.
15. Berry DM. *Understanding digital humanities*. Palgrave Macmillan. 2012.
16. Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences.* 2023;103:102274.
17. Tascón M, Montolío E. *El derecho a entender: la comunicación clara, la mejor defensa de la ciudadanía*. Los libros de la Catarata. 2020.
18. Blázquez Jiménez J. El uso de las nuevas tecnologías para la protección y difusión del patrimonio cultural: Ávila y el proyecto Smart Heritage City. 2023.
19. Sánchez MÁC. Las humanidades digitales como expresión y estudio del patrimonio digital. *Ediciones de la Universidad de Castilla La Mancha*. 2021;31.
20. Dans Álvarez de Sotomayor I. Educación digital: formación contra la desinformación. *Innovación Educativa Pluma y Arroba 2022: Competencia Digital. Aulas del Futuro. Sostenibilidad. Metaverso*. 2023.
21. Korro Bañuelos J. La gestión documental de la conservación-restauración del patrimonio arquitectónico: implementación, mantenimiento y repercusión en la puesta en valor del patrimonio y de los proyectos de conservación. 2024.
22. Bailey R. Exploring design process learning through two reflective prompts. *The International journal of engineering education*. 2020;36(2):568–573.
23. Černín D. Historical methodology and critical thinking as synergised concepts. *Disputatio*. 2020;9(13):349–382.
24. Chapman AJ, Burn K, Kitson A. What is School History For? British Student-teachers' Perspectives//¿ Para Qué Sirve La Enseñanza De La Historia? Perspectivas De Docentes y Estudiantes Británicos. *Arbor: Revista de Ciencia, Pensamiento y Cultura.* 2018;194(788).
25. Tartera FJG. Ingeniería en San Millán de la Cogolla desde la época medieval. In *En torno a la lengua: proyecto basado en San Millán de la Cogolla*. Aula Magna. 2023. p. 187–204.
26. García Beltrán E. No es magia, es prompting: el diseño de prompts como competencia emergente en la formación docente. Un estudio desde el modelo CRETA+ R. 2026.
27. Tartera FJG. *Competencias digitales en la docencia universitaria del siglo XXI* (Doctoral dissertation, Universidad Complutense de Madrid). 2017.
28. Mavrommati M, Repoussi M. 'Something was wrong with the movie': formal analysis of historical films and the development of historical literacy ('Había algo fuera de lugar en la película': análisis formal de las películas históricas y desarrollo de la alfabetización histórica). *Journal for the Study of Education and Development*. 2020;43(3):574–605.
29. Berceo G. *Vida de San Millán de la Cogolla*. In: JA Pascual, editor. Crítica. (Original work written ca. 1230). 2000.