AI & Antiquity: *Journal of Teaching and Technology in Ancient Studies*
Volume 2, Issue 1, 2026, pp. 4-10.
ISSN: 3081-4553

When References Mislead: Verification, AI Attribution, and Academic Bullying in Scholarly Evaluation

Carlos Heredia Chimeno

Editor-in-Chief
Universitat Autònoma de Barcelona
carlos.heredia@uab.cat
0000-0003-2866-5883

1. A Silent Problem in plain sight

If the first volume of *AI & Antiquity* was born from the need to rethink how we teach and learn in the age of artificial intelligence, this second volume arises from a more uncomfortable realisation: AI does not only challenge assessment or authorship; it challenges the very infrastructure of scholarly trust, and the centrality of source verification.

Among the many issues raised by the widespread use of generative AI in academic contexts, one has proven particularly insidious: the creation of bibliographic references that do not exist. These are not necessarily fabricated with malicious intent, nor always introduced by students attempting to deceive. More often, they are generated silently, plausibly, and convincingly by systems trained to reproduce the surface features of scholarly discourse: titles that sound real, journals that almost exist, DOIs that follow correct patterns but lead nowhere, and authors whose names resonate with the field but have never written the cited work.

What makes this phenomenon especially troubling is not its novelty, but its subtlety. Unlike plagiarism, which leaves detectable traces, or factual errors that may be identified through disciplinary knowledge, false bibliography often passes unnoticed precisely because it looks correct (see Walters and Wilder, 2023; Liu, D'Elia and Palermo, this issue). Worse still, it frequently emerges in contexts perceived as low-risk: when asking AI to "format references", "complete missing citations", or "standardise a bibliography according to a style guide". These are tasks that many of us, as researchers, reviewers, or editors, have already delegated at least once, often assuming that the operation is merely cosmetic. This, ultimately, is the crux of the problem: bibliographic hallucination is not merely a student issue. It is a structural risk that affects the entire academic ecosystem, precisely because it exploits the conventions on which scholarship relies.

Actually, generative AI models are not databases. They do not "know" whether a reference exists; they predict text. When prompted to generate or adjust bibliographic entries, they do exactly what they are designed to do: produce something that looks right. The danger lies in our tendency to confuse formal correctness with epistemic reliability, as if the mere presence of a well-formed citation were synonymous with traceability and truth.

In the Humanities, and particularly in Ancient Studies, bibliography is not an accessory. It is the backbone of scholarly argumentation. References are not decorative footnotes; they are claims of verification—signals that knowledge can be traced, contested, and reassessed. When that chain is broken, even unintentionally, credibility collapses, and the reader is deprived of the very mechanism that makes scholarship cumulative rather than rhetorical. What is new in the AI context is the scale and speed at which such errors can propagate. A single fabricated reference, once cited, reformatted, exported to reference managers, or reused in teaching materials, can circulate widely before anyone notices the absence of an original source. This dynamic is amplified by the growing reliance on automated workflows, citation generators, and AI-assisted writing environments that prioritise efficiency over verification. The result is a paradox that we can no longer ignore: tools designed to assist scholarship can, if used uncritically, undermine the very standards they are meant to support.

2. VERIFICATION, DUE PROCESS, AND AI-RELATED ACADEMIC BULLYING

The response to this situation cannot be a return to technological rejection, nor can it rely solely on warnings directed at students. Instead, it requires a collective recalibration of scholarly habits—one that recognises that, in an AI-mediated environment, verification must be treated not as an optional final step but as the core scholarly practice linking responsibility, traceability, and human accountability. In this respect, few approaches are more instructive than the perspective advanced by Solga and Sarwar (2026, this issue) whose emphasis on procedural transparency and evidentiary anchoring foregrounds verification as a constitutive scholarly act rather than a post *hoc* corrective.

First, we must acknowledge that bibliographic vigilance is now a core scholarly skill, on par with source criticism or methodological transparency. Checking references is no longer a cosmetic final operation; it is the epistemic labour that distinguishes human scholarship from automated plausibility. Second, we must resist the temptation to treat AI as a neutral technical assistant. Delegating tasks such as citation completion, bibliography generation, or reference formatting without verification is not a time-saving shortcut; it is a transfer of responsibility to a system that cannot bear it. The historian, archaeologist, or philologist remains accountable for every reference that appears under their name, just as they remain accountable for the evidence, reasoning, and interpretative claims constructed from those references. Third, this issue must be addressed explicitly in teaching. Students should not only be warned that "AI invents sources", but guided through practical exercises that expose how and why this happens. Asking students to verify AI-generated bibliographies, to track down nonexistent references, or to compare AI outputs with real databases can transform a risk into a powerful learning opportunity, reinforcing critical thinking and digital literacy simultaneously.

For academic journals, the implications are equally serious. Peer review alone is no longer sufficient if reviewers assume that bibliographies are mechanically correct. Editorial workflows must adapt, incorporating explicit checks and fostering a shared awareness that formal plausibility is no guarantee of authenticity. This does not mean increasing bureaucratic burden or policing authors through suspicion. It means recognising that we are operating in a transformed epistemic environment, where traditional signals of reliability can no longer be taken for granted and must be actively verified. Indeed, innovation in teaching and research must go hand in hand with methodological rigour and ethical responsibility. Embracing AI critically does not mean trusting it blindly; it means understanding its limitations and designing practices that compensate for them.

It is essential to make one point absolutely clear: the presence of incorrect (or even nonexistent) bibliographic references should not automatically be framed as academic misconduct, particularly in this transitional phase of widespread AI adoption. In many cases, such errors do not stem from an intention to deceive, fabricate data, or manipulate scholarly discourse, but from the uncritical or partial use of tools whose overall utility far exceeds these specific risks. To reduce AI to a mere channel of misconduct would be both inaccurate and counterproductive, especially when these systems are already embedded in everyday academic practice, from language revision and structural clarity to accessibility and pedagogical design. Moreover, editorial policies that reflexively equate bibliographic hallucinations with fraud risk producing unjust outcomes, especially for early-career researchers, non-native speakers, and scholars working under intense publication pressure. Punitive approaches may, paradoxically, discourage transparency, pushing authors to conceal AI use rather than engage with it openly and responsibly.

A growing number of authors are already encountering the consequences of this transitional moment firsthand. In several reported cases, irregularities in citations, source selection, or textual reconstruction—sometimes linked to AI-assisted workflows, sometimes to misguided methodological shortcuts—have been interpreted not as errors requiring correction but as prima facie evidence of deliberate scientific misconduct. What is striking is not the identification of problems themselves, which is both legitimate and necessary, but the evaluative leap that follows: suspicion is rapidly converted into moral judgement. Irregular or inappropriate citations are treated as proof of intent; methodological weaknesses are reframed as fabrication; and correctable failures of verification are recoded as indicators of dishonesty.

This progression collapses the distinctions on which responsible scholarly assessment depends. Errors, negligence, and fraud—categories that require different evidentiary thresholds and call for different responses—are flattened into a single presumption of bad faith. The result is a shift from procedural evaluation to moralisation, in which the possibility of clarification, revision, or methodological correction is displaced by the language of sanction. In some instances, authors are not only rejected but threatened with institutional reporting without any prior stage of dialogue, documentation of evidence, or opportunity for response, thereby short-circuiting the processes that ensure proportional and evidence-based judgement.

Documented episodes within recent scholarly communication further illustrate this dynamic. Public attributions of AI authorship, initially advanced on the basis of alleged bibliographic anomalies, have in some cases been shown—following systematic verification of sources, citations, and editorial records—to be factually incorrect and unsupported by evidence. Nevertheless, the original claims circulated rapidly through private communication networks and informal scholarly channels, acquiring credibility through repetition rather than documentation and producing reputational effects without any procedural mechanism for clarification or right of reply. Within competitive and precarious academic environments, such dynamics intersect with factional alignments and reputational economies, where AI suspicion may be mobilised strategically as a tool of exclusion. When unsubstantiated allegations are disseminated in ways that are difficult to contest, detached from formal review, and capable of producing demonstrable professional harm, the boundary between critical evaluation and reputational aggression becomes blurred. In their most problematic form, these practices constitute a mode of academic bullying: not through overt hostility, but through the circulation of unverifiable claims that stigmatise authors while evading the evidentiary standards that govern scholarly critique.

The ethical issue at stake is therefore not only methodological but procedural. Scholarly disagreement requires argument, documentation, and the possibility of response; reputational claims about authorship or integrity require an equivalent evidentiary burden. To advance allegations of AI use without verifiable criteria, and to disseminate them through informal channels that preclude correction, risks approximating the dynamics of defamation in functional terms, even when framed as methodological concern. Detached from transparent processes, provisional suspicions can harden into durable reputational narratives that are resistant to later clarification. These cases underscore the necessity of shared protocols that distinguish clearly between documented irregularity, correctable error, and demonstrable misconduct, and that subject both texts and accusations to equivalent evidentiary standards as a condition of scholarly integrity.

Instead, academic publishers and editorial boards must recognise bibliographic fabrication by AI as a systemic by-product of a rapidly evolving technology, not as evidence of individual ethical failure *per se*. The appropriate response is not sanction, but shared responsibility: clearer guidelines on acceptable AI use, explicit verification protocols, and constructive communication between editors, reviewers, and authors when such issues arise. This position does not imply lowering standards. On the contrary, it reinforces them. Scholarly rigour is best protected not through suspicion and punishment, but through education, awareness, and procedural adaptation—much as the academic community once learned to integrate digital databases, reference managers, and online archives without criminalising their early misuses. Indeed, recent reporting has made clear that this is no longer an anecdotal classroom issue. Cases are emerging in which real academic outlets cite references that appear to have been fabricated or laundered through AI-mediated workflows, blurring the line between error, automation, and institutional negligence. Bibliographic hallucinations are thus migrating from student submissions into the wider circulation of scholarly writing—precisely because they imitate our conventions so well.

Reactions across scholarly networks are instructive. Some researchers return to traditional, verifiable search practices; others rely on specialised discovery platforms or AI tools designed for bibliographic validation; still others read the phenomenon as a symptom of structural pressures such as citation inflation, paywalls, and editorial workflows that do not systematically verify references. What matters editorially is not which response is "correct", but what they collectively signal: scholarship must now actively defend itself against the production of plausible but unverified knowledge. This requires not only attention to student practice but the refinement of institutional norms, editorial procedures, and everyday research habits.

The task is not to reject AI, but to embed verification as a non-negotiable scholarly standard, ensuring that the ease of generating academically styled text does not weaken the chain of traceability on which humanistic knowledge depends. It is against this backdrop of uncertainty, eroded trust, and renewed demands for evidence that the present issue must be read. Volume 2, Issue 1 approaches artificial intelligence neither as a neutral innovation nor as a threat to be contained, but as a field of practice in which mediation, method, responsibility, and inclusion must be actively negotiated. The contributions gathered here offer not technical fixes but concrete pedagogical, methodological, and historiographical strategies for working with AI without relinquishing scholarly rigour.

3. HIGHLIGHTS OF THIS ISSUE

Taking into account all of the above, *Volume 2, Issue 1* of *AI & Antiquity* has been conceived not as a loose collection of contributions, but as a deliberately structured itinerary, one that moves from the classroom to research practice, from methodological experimentation to cultural responsibility, and ultimately to the ethical question that remains beneath every technological debate: what is the purpose of knowledge, and whose voices does it serve?

The issue opens with a first thematic block centred on AI as a mediator of knowledge, with a clear emphasis on the classroom as the primary laboratory of transformation. In "From Classical Sources to Artificial Intelligence: *Notebook LM* as a Cognitive Mediator in University Teaching of Ancient History", Francisco Javier Catalán González offers an experimental, source-based case study built around a controlled *corpus* of Greco-Roman narratives on the foundation of Rome (Livy, Dionysius of Halicarnassus, Cassius Dio, and Velleius Paterculus, using in this case the Gredos editions). Working within a closed-document environment, the article tests how *Notebook LM* can process and synthesise primary materials and then transform them into university-oriented learning artefacts—summaries, concept maps, flashcards, and multimodal outputs (audio and video)—designed to support cognitive structuring, retention, and active study. Crucially, the study does not present AI as an interpretative substitute, but as a scaffold for engaging with dense textual corpora; at the same time, it identifies two pedagogically significant risks—discursive homogenisation and limited philological sensitivity—thereby reinforcing the issue's wider argument that AI can improve accessibility and comprehension only under explicit human supervision and source-aware teaching design.

This pedagogical entry point is then expanded and reframed through "Let's Chat About Archaeology: Responsible and Thoughtful Use of AI Tools in the Classroom, A Case Study" by Yusi Liu, Daniel D'Elia, and Rocco Palermo. Based on their Bryn Mawr course, the authors design an assignment in which students work both as AI-assisted writers and as editors who annotate, fact-check, and revise chatbot outputs using citations and tracked changes. By linking prompt design, collaborative verification, and the identification of inaccuracies and bias to the wider problem of pseudo-archaeological misinformation, the study provides a transferable model for integrating AI into teaching while foregrounding accountability, transparency, and critical evaluation.

From the classroom, the issue then moves into research practice with "Evaluating Generative AI in Historical Research: A Comparative Study on Identifying Primary Source Evidence in Ancient History" by Raymond S. Solga and Mohammed J. Sarwar. Through a paired case-study design, the authors contrast four human-led investigations with four AI-assisted inquiries conducted across GPT-4, Claude 2, Gemini, and Perplexity. The comparison is structured around explicitly historiographical criteria—temporal framing, provenance transparency, genre differentiation, linguistic fidelity, and evidentiary reliability—making visible both the exploratory speed of generative systems and their persistent weaknesses in verification, manuscript lineage, and genre control.

The next step in the issue's arc functions as a conceptual hinge. "Algorithmic Memory: Towards Reflexive Authenticity in Cultural Heritage" by Menna Salah shifts the focus from research method to the public life of the past. The article argues that artificial intelligence is not merely a tool of preservation but a force that reshapes how cultural memory is curated, legitimised, and shared. Through the notion of *reflexive authenticity*, authenticity is redefined as transparency rather than curatorial authority as participatory rather than hierarchical.

"Recovering the Voices of Silence in Ancient Historiography: A Re-reading of the *Shiji* through the Lens of Inclusive Artificial Intelligence" by Samandar Ruziboev and Noyibjon Khudoyorov brings the volume to its most explicitly historiographical horizon. Focusing on Sima Qian's *Shiji* (ca. 145–86 BCE) within the political context of the early Han empire, the article re-reads silence and diplomacy not as narrative absences but as structuring features of historical writing. AI is employed not as an interpretive authority but as an analytical instrument capable of tracing patterns that elude close reading, opening a methodological space in which questions of power, language, and representation can be reconsidered. The contribution thus moves the debate from procedures to meaning, asking how inclusive and reflexive approaches reshape whose voices are transformed in ancient historiography.

After the peer-reviewed articles, the volume turns to a section that brings these debates into the sphere of Public History and contemporary narratives of Antiquity. The contributions by Iban Martín, Patricia González, and Mario Agudo, presented in the seminar on *Ancient World Today* organised with Marc Mendoza at the UAB (27 November 2025), show how the questions raised throughout the issue move beyond the classroom and research settings into public discourse, where Antiquity is told in ways that shape contemporary understanding. What emerges is not only a methodological concern, but the persistent unevenness of the archive itself—those presences that survive only in fragments and those that never entered the record, including many women whose experiences remain largely inaccessible. Making those absences visible requires not only new tools, but renewed commitments to verification, proportionality, and the rejection of bullying practices in scholarly evaluation.

Bellaterra (Barcelona), February 2026

BIBLIOGRAPHY

Liu, Y., D'Elia, D. and Palermo, R. (2026) 'Let's Chat About Archaeology: Responsible and Thoughtful Use of AI Tools in the Classroom, A Case Study', *AI & Antiquity*, 2(1), this issue.

Solga, R. and Sarwar, M. J. (2026) 'Evaluating Generative AI in Historical Research: A Comparative Study on Identifying Primary Source Evidence in Ancient History', *AI & Antiquity*, 2(1), this issue.

Walters, W. H. and Wilder, E. I. (2023) 'Fabrication and errors in the bibliographic citations generated by ChatGPT', *Scientific Reports*, 13, article 14045. doi: 10.1038/s41598-023-41032-5.