



From Archaeology to Artificial Intelligence: Reconstructing, Verifying, and Interpreting the Historical Past in the Digital Age

Muhammad Bilal Asif (Corresponding Author)

M. Phil Scholar/ Lecturer History, Government Graduate College Samanabad Faisalabad

Email: bilalgcsr@gmail.com

Nauman Tahir

Assistant Professor History, Government Graduate College Samanabad Faisalabad

Email: t.nauman2@gmail.com

Kishwar Sultana

Associate Professor Department of History, Allama Iqbal Open University Islamabad

Email: kishwar.sultana@aiou.edu.pk

Muhammad Usman

M. Phil Scholar, Department of Political Science GC University Faisalabad

Email: usmanluqman108@gmail.com

Abstract

Historical method has repeatedly widened its evidentiary base, moving from narrative and textual criticism to material archaeology, bioarchaeology, remote sensing and, most recently, artificial intelligence (AI). This paper offers a critical, conceptual synthesis of that trajectory and asks what standard of verification historians should require before accepting AI-derived claims about the past. Drawing on peer-reviewed studies of LiDAR, radar and muon imaging, handwriting and text restoration models, and large language models, it argues that AI is best understood not as a single tool but as three modes of work: recovery of evidence, generation of conjectures, and synthesis of narrative or images. Each mode has distinct failure modes. Hallucination and bias are treated as structural properties of probabilistic systems trained on uneven records rather than as occasional malfunctions. Toynbee's Challenge and Response is retained as a heuristic, complemented by Kuhn's account of paradigm change and Daston and Galison's history of epistemic virtues, and its limits are acknowledged. The paper proposes a five-level verification ladder that runs from data provenance through model documentation, reported uncertainty and independent corroboration to human adjudication, and it outlines a testable design for measuring anachronism in generative image models. The argument is conceptual, and the ladder has not yet been empirically validated.

Keywords: epistemology of history, historical methodology, LiDAR, artificial intelligence, hallucination, algorithmic bias, source criticism, digital archaeology, verification

Introduction

History is not merely the recording of events and dates. It is a disciplined inquiry into change and causation in which the historian and the evidence continually shape one another (Carr, 1961). From Herodotus and Thucydides onward, historians have had to decide which accounts deserve trust and why (Burrow, 2009). For most of the



Vol. 4 No. 1 (January) (2026)

discipline's past, the evidence available was written or oral. Such sources were shaped by patronage, ideology and the accidents of survival, so that what is silent in the record can matter as much as what is said.

Over the nineteenth and twentieth centuries historians widened that base. Archaeological excavation, radiocarbon dating, isotope chemistry, ancient DNA and, later, remote sensing supplied physical evidence that could be set against textual claims. Artificial intelligence (AI) now appears to be a further widening. Machine-learning systems can read damaged texts, transcribe handwriting at scale and propose dates and places of origin for inscriptions, while generative models can produce fluent summaries and images of the past on demand. Scholars of digital history have noted that such abundance changes both what can be found and what can be trusted (Milligan, 2019; Putnam, 2016).

Public discussion of AI and history tends to swing between enthusiasm about speed and alarm about fabrication. Both reactions are partly justified, yet neither supplies a standard by which a historian can decide when an AI-derived claim is good enough to use. This paper therefore asks: what level of verification should historians require before accepting AI-derived claims about the past? Three sub questions follow. What kinds of work does AI actually perform in historical research? What are the characteristic failure modes of each kind? And how can the long-standing practice of triangulating textual, material and scientific evidence be formalized for AI-assisted work?

The paper is a critical, conceptual synthesis. It draws mainly on peer-reviewed primary studies, with institutional sources used only where they are the authoritative record of an event, and it reports the performance figures exactly as the original authors report them. Because AI systems change quickly, such figures are time-bound and should be read as illustrations rather than permanent properties of the technology. The argument proceeds in six steps: a theoretical framework; the shift to material and scientific evidence; remote sensing and the principle of triangulation; a taxonomy of AI in historical research; a discussion of hallucination, bias and related risks; and a proposed verification framework with directions for future research.

Theoretical Framework

Challenge and Response as a Heuristic

In *A Study of History*, Toynbee (1934–1961) proposed that civilizations arise and grow when a society meets a challenge with a creative response, typically led by a creative minority. This paper borrows the idea as a heuristic. AI presents historical scholarship with a challenge, namely an abundance of fluent, plausible and un anchored text and imagery, and the creative response is methodological rather than technological. The borrowing has limits that should be stated plainly. Toynbee developed the model for civilizations, not for academic disciplines, so its application here is an analogy and not an explanation. His wider system also drew sustained criticism from contemporaries on empirical and methodological grounds (Geyl et al., 1949; for a biography, see McNeill, 1989). The framework is therefore used to organise the argument, not to carry its evidential weight.

Complementary Lenses

Two further perspectives strengthen the framework. Kuhn (1962) described how scientific communities change the standards by which they judge evidence when anomalies accumulate. History has no single paradigm in Kuhn's strict sense, so the



Vol. 4 No. 1 (January) (2026)

analogy is again partial, but it usefully directs attention to shifting standards of proof. Daston and Galison (2007) offer a closer fit. They show that “objectivity” has a history, and that new instruments and image-making practices have repeatedly forced scholars to rethink which representations may count as evidence. Generative AI poses the same question for images and text of the past.

Constructed Narratives and Empirical Constraint

The claim that history is constructed needs careful handling. White (1973) argued that historical accounts are shaped by narrative form, in other words by the way events are em plotted. That argument concerns interpretation and does not license invention. Evans (1997) defends the view that sources, and the rules of evidence for using them, constrain what a historian may legitimately claim. Read together, these positions clarify the risk examined later in this paper. The danger of generative AI is not that it reveals narrative to be constructed, but that it produces constructions with no evidentiary anchor at all. A well known aphorism attributed to Voltaire, that history is a set of tricks played on the dead, is often invoked in this debate. Its source is uncertain, and it is not relied on here.

Debates in Digital History

Finally, the paper situates itself within debates about digital method. Advocates of distant reading and large-scale analysis argue that computational methods can reveal patterns invisible to close reading (Moretti, 2013; Underwood, 2019) and that historians can use tools as a “macroscope” to move between scales (Graham et al., 2016). Critics and cautious observers point to the shadows cast by digitisation and keyword search (Putnam, 2016), to the challenge of abundance (Milligan, 2019), and to the ways in which academic history writing has been slow to confront digital conditions (Hitchcock, 2013). Kitchin (2014) examines how data-intensive approaches reshape epistemology more broadly. The position taken here lies between enthusiasm and rejection: computational tools are valuable, and their outputs are claims that must be verified.

The First Shift: Material and Scientific Evidence

Before systematic archaeology, historical chronologies rested largely on chronicles, religious writings and epic literature, which combined record with myth and political purpose. Archaeology made it possible to test such texts against surviving physical remains. Comparative studies of early civilizations show both shared patterns and wide variation (Trigger, 2003), a reminder that material evidence qualifies broad generalisations as often as it confirms them.

Bioarchaeology, Isotopes, Radiocarbon and Ancient DNA

Bioarchaeology treats the human body as a primary source. Analysis of skeletons can illuminate diet, workload, disease and injury among people who rarely appear in written records (Larsen, 2015). Stable isotope analysis adds information about movement. Strontium isotope ratios in tooth enamel reflect the geology of the region where a person lived while the tooth formed, and so they indicate place of childhood residence, not nationality or ethnic identity (Bentley, 2006). Radiocarbon dating, developed in the late 1940s, gave archaeology a chronology independent of texts, and its later refinement has been described as transformative for the discipline (Bayliss, 2009). Ancient DNA has extended the toolkit further. Bos et al. (2011), for example, reconstructed a draft genome of the bacterium *Yersinia pestis* from victims of the Black



Vol. 4 No. 1 (January) (2026)

Death buried in London in 1348–1350, linking a documented historical catastrophe to a specific pathogen.

Taken together, textual, material and scientific evidence form a system of mutual checks. Each can correct, qualify or extend the others, and none is authoritative alone.

Remote Sensing and the Principle of Triangulation

From the late twentieth century, remote sensing allowed landscapes to be studied at scale without excavation. Light Detection and Ranging (LiDAR) fires laser pulses from aircraft and records their returns, allowing analysts to model the ground surface beneath vegetation. Radar satellites and ground-penetrating methods can reveal buried features in arid terrain. These techniques address a problem long familiar to historians, namely that the physical record is fragmentary and mostly invisible from the surface.

LiDAR in Southeast Asia and Mesoamerica

Evans et al. (2013) used airborne LiDAR at Angkor to uncover previously unmapped features of the settlement landscape and extend knowledge of its layout. In Guatemala, the Pacunam LiDAR Initiative surveyed 2,144 km² of the Maya Biosphere Reserve and identified 61,480 ancient structures, from which Canuto et al. (2018) extrapolated a lowland population in the millions, unevenly distributed. These findings altered regional-scale interpretations of Maya settlement. The population figures are extrapolations from a sample, and they remain hypotheses that require continued field validation.

Giza: Convergent Lines of Evidence

Giza illustrates triangulation well. Morishima et al. (2017) reported a large void inside Khufu's Pyramid using cosmic-ray muon radiography, confirmed with three independent detector technologies. Its function remains undetermined. On the landscape scale, Ghoneim et al. (2024) combined radar satellite imagery, geophysical data and deep soil coring to identify a buried Nile channel, which they named the Ahramat Branch, and argued that it served as a waterway for people and building materials. Sheisha et al. (2022) reconstructed fluctuations of the Khufu branch of the Nile from pollen records and concluded that it remained at high water during the reigns of Khufu, Khafre and Menkaure. These physical results sit alongside the Merer papyri, a logbook type text from the reign of Khufu that records the transport of limestone to Giza (Tallet, 2017). The papyri are not disputed. They are a contemporary written source whose account of water transport is consistent with the independent geophysical and environmental findings. The authors of the physical studies present their conclusions as inferences, and the convergence is persuasive precisely because the lines of evidence are independent.

This convergence supplies the principle on which the rest of the paper depends: a historical claim is strongest when evidence of different kinds, produced by different methods, points the same way. The question that follows is how AI-generated outputs can be brought into that system.

Artificial Intelligence in Historical Research

In this paper, artificial intelligence refers to computational systems that learn statistical patterns from data. Machine learning covers methods such as convolutional neural networks and transformers, and generative AI denotes models that produce new text or images. The distinctions matter because they correspond to different kinds of



Vol. 4 No. 1 (January) (2026)

historical work and different kinds of error. Table 1 proposes a taxonomy of three modes.

Table 1

Three Modes of AI Work in Historical Research

Mode	Typical tasks and examples	Characteristic failure	Verification route
Recovery of evidence	Detecting ink in CT scans of sealed scrolls (Parker et al., 2019); handwritten text recognition (Nockels et al., 2022)	Missed or invented strokes; weaker results on scripts and hands unlike the training examples	Specialist palaeographic or papyrological reading; comparison with independent imaging or transcription
Conjecture	Restoring, dating and locating inscriptions (Assael et al., 2022); suggesting artefact type or date (Bickler, 2021)	Plausible but wrong suggestions; skew toward well-documented types; over-confidence	Ranked hypotheses with stated uncertainty; corroboration by parallels, context and scientific dating
Synthesis	Language-model summaries and translations; generated images and 3D reconstructions	Fabricated citations and events; anachronism; stereotyped content (Walters & Wilder, 2023)	Check every claim and citation against primary or scholarly sources; label output as illustration, not evidence

Note. The taxonomy is the personal analytic proposal. Examples are drawn from the cited literature.

Recovery of Evidence

The Herculaneum papyri were carbonized by the eruption of Vesuvius in 79 CE and cannot be unrolled without destroying them. Parker et al. (2019) built a machine-learning pipeline around a three dimensional convolutional neural network that detects the faint signature of carbon ink in micro-CT data, a signal long assumed to be undetectable. The Vesuvius Challenge later set a public target of four passages of 140 characters each, with at least 85% of characters recoverable, and awarded its 2023 Grand Prize to a team that met it, revealing text from inside a rolled scroll for the first time (Vesuvius Challenge, 2024). The achievement is real, but its limits should be stated carefully. Such a model infers ink from subtle volumetric patterns rather than photographing it, so readings require papyrological scrutiny, and the passages recovered so far are partial.

Handwritten text recognition (HTR) offers a second case. A systematic review of the Transkribus platform describes HTR as a mature machine-learning tool that speeds the transcription of primary sources and makes full-text searching and analysis of historic texts possible at scale (Nockels et al., 2022). Because such models learn from transcribed examples, their reliability for unusual hands or scripts depends on the availability of comparable training data, which is a reason for specialist checking rather than a reason to avoid the technology.



Conjecture

Ithaca, a deep neural network trained on ancient Greek inscriptions, restores missing text and estimates the location and date of an inscription. In the authors' evaluation, Ithaca alone achieved 62% accuracy in restoring damaged text, attributed inscriptions to their original location with 71% accuracy, and dated them to within 30 years of the ground-truth range. When historians used Ithaca, their own restoration accuracy rose from 25% to 72% (Assael et al., 2022). Two features of this study deserve emphasis. First, a restoration is a ranked, probabilistic hypothesis rather than recovered evidence, which places it alongside traditional epigraphic conjecture and not above it. Second, the strongest result concerns collaboration, not replacement, a point echoed by a broad survey of machine learning for ancient languages (Sommerschild et al., 2023).

Similar logic applies to artefact analysis. Reviews of machine learning in archaeology describe its use for classifying artefacts and detecting sites in imagery, and they stress its dependence on training data (Bickler, 2021). A classifier can compare an object with thousands of reference examples in seconds and propose a probable type, date or cultural affiliation, thereby saving researchers considerable time. The gain is speed and consistency at scale. What the classifier cannot do is establish that its answer is right, because it can only choose among the categories it has seen. An object from a type poorly represented in the reference collection may be confidently assigned to a familiar one. Classifications should therefore be treated as probabilities that specialists then confirm.

Synthesis

Generative systems can summarize sources, produce translations, and render images or three-dimensional models of vanished buildings and landscapes. Such reconstructions can be valuable for hypothesis testing about sight lines, spatial layout or acoustics, but only to the extent that the inputs and assumptions behind them are documented. Their outputs are plausible constructions, not observations, and this is the mode in which the risks discussed below are most acute.

Risks: Hallucination, Bias, and the Degradation of the Record

Hallucination and Fabricated Sources

Large language models generate text by predicting likely continuations of a sequence, and they do so without reference to meaning or to an external record of the past (Bender et al., 2021). The resulting errors are commonly called hallucinations, defined in a major survey as generated content that is nonsensical or unfaithful to its source (Ji et al., 2023). Some philosophers object that the term wrongly implies a perceiving mind and argue that such output is better described as indifferent to truth in the sense of Frankfurt's bullshit (Hicks et al., 2024). The distinction has practical weight for historians. If the system's objective is plausibility, then fluency is not evidence of accuracy, and errors will appear in exactly the places where the record is thin.

Evidence on fabricated references is concrete. Walters and Wilder (2023) prompted two versions of ChatGPT to write short literature reviews on 42 topics and checked the 636 citations produced. Of the citations from GPT-3.5, 55% were fabricated, as were 18% from GPT-4, and among the real citations 43% and 24% respectively contained substantive errors. Newer systems may perform differently, and studies specific to historical topics are still needed. The finding nevertheless establishes the risk



Vol. 4 No. 1 (January) (2026)

of phantom citations, and it explains why a history paper that reports AI-assisted findings must trace every reference to a real source.

Bias and Uneven Knowledge

Bias is the second structural risk, and it is distinct from hallucination because it can occur even when outputs are technically accurate. Training and reference data are uneven. Joshi et al. (2020) showed that the vast majority of the world's languages are poorly served by natural language processing, and Putnam (2016) demonstrated that what is digitized and text-searchable shapes which histories are found. Jo and Gebru (2020) argued that machine-learning datasets practice should learn from archival traditions of consent, inclusivity and transparency. More broadly, critical scholarship holds that algorithmic systems can reproduce social hierarchies while appearing neutral (Benjamin, 2019; Crawford, 2021; Noble, 2018).

A benchmark of language-model historical knowledge illustrates the unevenness directly. Hauser et al. (2024) tested seven models on expert-curated data from the Seshat Global History Databank. Balanced accuracy on four-choice questions ranged from 33.6% to 46%, above the 25% chance level but short of expert comprehension, and the more advanced models performed worst on Oceania and Sub-Saharan Africa. For historians working on under-documented regions, languages or periods, this pattern is a warning that AI assistance is least reliable where the need for care is greatest. By extension, reference collections of artifact that reflect histories of excavation and collecting are likely to carry comparable skews, which is a further reason to report classifications as probabilistic and dataset dependent.

Anachronism in Generated Reconstructions

Image and video generators raise a related concern. A model asked to depict a Late Bronze Age scene may blend armour, architecture or dress from other periods, or default to the conventions of popular visual culture. This hypothesis is plausible, but the author is not aware of a systematic quantification of the problem in the sources reviewed, so it is stated here as a hypothesis and not as an established finding. The section on future research proposes a design to test it.

Recursive Training and the Loss of Rare Histories

A further risk concerns the future of the record itself. Shumailov et al. (2024) showed that models trained recursively on data generated by earlier models degrade, progressively losing the tails of the original distribution. As synthetic text spreads online, future models may be trained on material increasingly dominated by earlier model output. The rare and minority elements of the historical record, which already attract less data, are the elements most exposed to this loss.

Re framing the Epistemological Risk

These risks are sometimes presented as a revival of radical scepticism about history. A more precise framing follows from the framework above. The danger is not that narrative construction has been exposed, but that generative systems can produce text and imagery at industrial scale with no evidentiary anchor, which strains the practices of source criticism that distinguish interpretation from invention (Evans, 1997; Milligan, 2019; Putnam, 2016). In Toynbee's terms, the challenge is a flood of unanchored plausibility, and the creative response is methodological discipline.



Toward a Verification Framework

From Digital Source Criticism to a Verification Ladder

Traditional Quellenkritik asks who produced a source, when, for whom and with what bias. Applied to AI, the same questions extend to the data a model learned from, the model itself, and the output. Table 2 proposes a five-level verification ladder that turns this extension into a practical protocol. It builds on documentation practices such as datasheets for datasets (Gebru et al., 2021), on the principle of applicability in ethical frameworks for AI (Floridi & Cowls, 2019), and on the triangulation logic illustrated above.

Table 2

A Five-Level Verification Ladder for AI-Derived Historical Claims

Level	Verification question	Minimum practice
1. Data provenance	What data did the system learn from or analyse, and who or what is missing?	Document source, digitisation, and language and regional coverage (Gebru et al., 2021; Jo & Gebru, 2020)
2. Model documentation	Which model and version produced the output, and what are its known limits?	Record model, version, settings and prompt; cite published evaluation evidence
3. Reported uncertainty	How confident is the output, and what alternatives were possible?	Report accuracy or confidence and ranked alternatives, as in Assael et al. (2022)
4. Independent corroboration	Does evidence not produced by AI point the same way?	Cross-check against material, scientific, spatial or documentary evidence
5. Human adjudication and transparency	Who judged the claim, on what grounds, and can others retrace the steps?	Named expert review; disclosure of AI use; archived inputs and outputs

Note. The ladder is the author's proposal and has not been empirically validated.

The ladder yields three possible statuses for an AI-derived claim. A claim that passes all five levels may be reported as evidence. A claim that fails independent corroboration at Level 4 remains a hypothesis and should be labeled as one. An output whose provenance cannot be established, such as a citation that cannot be found in any catalogue, should be discarded or used only as illustration.

Two Worked Illustrations

Consider first an Ithaca restoration of a damaged inscription. The training corpus and model are documented and published (Levels 1 and 2), the system reports ranked alternatives and has a known accuracy (Level 3), the historian can test each alternative against parallels and context (Level 4), and the final judgement rests with a named scholar (Level 5). This is a case in which AI output can enter scholarship as a well-bounded hypothesis. Consider by contrast a reference generated by a language model. Its training data are opaque and no confidence is reported, so Levels 1 to 3 give little support, and Level 4 is the decisive test: the work must be found in a catalogue or



Vol. 4 No. 1 (January) (2026)

database before it is cited. The fabrication rates reported by Walters and Wilder (2023) show why that check cannot be skipped.

Implications for Practice

Several practical recommendations follow. Historians should disclose any use of AI and archive the prompts, model versions and outputs on which a claim depends. They should treat language-model summaries as leads to be followed up in the sources, not as sources in their own right. They should report uncertainty in the form scholars already use for conjecture. And they should collaborate with domain specialists and computer scientists, as the Ithaca results suggest, rather than delegating judgement to the tool. Training in source criticism for new researchers should now include criticism of datasets and models.

Limitations and Directions for Future Research

This study has several limitations. It is conceptual and does not report original empirical data. It relies mainly on English language, peer-reviewed sources and therefore inherits some of the unevenness it criticizes. The performance figures quoted belong to particular systems at particular dates, and they will change. The verification ladder is a proposal, not a validated instrument.

Two lines of research would address these limitations. The first is an anachronism audit of generative image models. A study would prompt several models with a fixed set of period-specific scenarios, such as Late Bronze Age Aegean battles, rituals and settlements, and record the prompts and model versions. Two or more specialists would then independently code each output for anachronism in armour, weaponry, architecture and dress, with inter coder agreement reported. Error rates by category and by model would turn the hypothesis stated earlier into evidence. The second line is a validation of the verification ladder. It could apply the ladder retrospectively to published AI assisted historical studies, or ask panels of historians to use it on sample claims and compare the decisions reached with and without it. Researchers working with regional archives should also test existing handwriting and language models on under-represented languages and scripts, since the sources reviewed here suggest that this is where performance gaps are likely to be largest.

Conclusion

The history of historical method is a history of widening evidence and of the practices developed to check it. Material archaeology, bio archaeology and remote sensing each tested textual claims, and where textual and physical evidence converge, as at Giza, confidence rises. AI continues this widening, and it does so in three distinct ways: recovering evidence, proposing conjectures and synthesizing narrative and images. Each carries its own benefits and its own errors. The speed with which models can read scrolls, transcribe archives and classify artefact is a genuine advance, and the Ithaca study shows that human-machine collaboration can improve scholars' own accuracy.

Yet fabrication, bias and the erosion of rare histories are structural properties of probabilistic systems trained on uneven records, and they cannot be assumed away. The appropriate response is neither rejection nor trust but verification: documented data, documented models, reported uncertainty, independent corroboration and accountable human judgement. In Toynbee's terms the challenge is abundance without anchor, and the creative response is a discipline of digital source criticism. Historians who adopt it



Vol. 4 No. 1 (January) (2026)

can use AI to extend the reach of the evidence while preserving the standard that makes history a discipline of inquiry and not of invention.

References

- Assael, Y., Sommerschild, T., Shillingford, B., Bordbar, M., Pavlopoulos, J., Chatzipanagiotou, M., Androutsopoulos, I., Prag, J., & de Freitas, N. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900), 280–283. <https://doi.org/10.1038/s41586-022-04448-z>
- Bayliss, A. (2009). Rolling out revolution: Using radiocarbon dating in archaeology. *Radiocarbon*, 51(1), 123–147. <https://doi.org/10.1017/S0033822200033762>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. <https://doi.org/10.1145/3442188.3445922>
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity.
- Bentley, R. A. (2006). Strontium isotopes from the earth to the archaeological skeleton: A review. *Journal of Archaeological Method and Theory*, 13(3), 135–187. <https://doi.org/10.1007/s10816-006-9009-x>
- Bickler, S. H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9(2), 186–191. <https://doi.org/10.1017/aap.2021.6>
- Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., & Krause, J. (2011). A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature*, 478(7370), 506–510. <https://doi.org/10.1038/nature10549>
- Burrow, J. (2009). *A history of histories: Epics, chronicles, and inquiries from Herodotus and Thucydides to the twentieth century*. Vintage.
- Canuto, M. A., Estrada-Belli, F., Garrison, T. G., Houston, S. D., Acuña, M. J., Kováč, M., Marken, D., Nondédéo, P., Auld-Thomas, L., Castanet, C., Chatelain, D., Chiriboga, C. R., Drápela, T., Lieskovský, T., Tokovinine, A., Velasquez, A., Fernández-Díaz, J. C., & Shrestha, R. (2018). Ancient lowland Maya complexity as revealed by airborne laser scanning of northern Guatemala. *Science*, 361(6409), Article eaau0137. <https://doi.org/10.1126/science.aau0137>
- Carr, E. H. (1961). *What is history?* Macmillan.
- Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.
- Daston, L., & Galison, P. (2007). *Objectivity*. Zone Books.
- Evans, D. H., Fletcher, R. J., Pottier, C., Chevance, J.-B., Soutif, D., Tan, B. S., Im, S., Ea, D., Tin, T., Kim, S., Cromarty, C., De Greef, S., Hanus, K., Bâty, P., Kuszinger, R., Shimoda, I., & Boornazian, G. (2013). Uncovering archaeological landscapes at Angkor using lidar. *Proceedings of the National Academy of Sciences*, 110(31), 12595–12600. <https://doi.org/10.1073/pnas.1306539110>
- Evans, R. J. (1997). *In defence of history*. Granta Books.
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*,



Vol. 4 No. 1 (January) (2026)

- 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Geyl, P., Toynbee, A. J., & Sorokin, P. A. (1949). *The pattern of the past: Can we determine it?* Beacon Press.
- Ghoneim, E., Ralph, T. J., Onstine, S., El-Behaedi, R., El-Qady, G., Fahil, A. S., Hafez, M., Atya, M., Ebrahim, M., Khozym, A., & Fathy, M. S. (2024). The Egyptian pyramid chain was built along the now abandoned Ahramat Nile Branch. *Communications Earth & Environment*, 5, Article 233. <https://doi.org/10.1038/s43247-024-01379-7>
- Graham, S., Milligan, I., & Weingart, S. (2016). *Exploring big historical data: The historian's microscope*. Imperial College Press.
- Hauser, J., Kondor, D., Reddish, J., Benam, M., Cioni, E., Villa, F., Bennett, J. S., Hoyer, D., François, P., Turchin, P., & del Rio-Chanona, R. M. (2024). Large language models' expert-level global history knowledge benchmark (HiST-LLM). In *Advances in Neural Information Processing Systems*, 37 (Datasets and Benchmarks Track). https://proceedings.neurips.cc/paper_files/paper/2024/file/38cc5cba8e513547b96bc326e25610dc-Paper-Datasets_and_Benchmarks_Track.pdf
- Hicks, M. T., Humphries, J., & Slater, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, 26, Article 38. <https://doi.org/10.1007/s10676-024-09775-5>
- Hitchcock, T. (2013). Confronting the digital: Or how academic history writing lost the plot. *Cultural and Social History*, 10(1), 9–23. <https://doi.org/10.2752/147800413X13515292098070>
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248. <https://doi.org/10.1145/3571730>
- Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 306–316). ACM. <https://doi.org/10.1145/3351095.3372829>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560>
- Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1). <https://doi.org/10.1177/2053951714528481>
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.
- Larsen, C. S. (2015). *Bioarchaeology: Interpreting behavior from the human skeleton* (2nd ed.). Cambridge University Press.
- McNeill, W. H. (1989). *Arnold J. Toynbee: A life*. Oxford University Press.
- Milligan, I. (2019). *History in the age of abundance? How the web is transforming historical research*. McGill-Queen's University Press.
- Moretti, F. (2013). *Distant reading*. Verso.
- Morishima, K., Kuno, M., Nishio, A., Kitagawa, N., Manabe, Y., Moto, M., Takasaki, F., Fujii, H., Satoh, K., Kodama, H., Hayashi, K., Odaka, S., Procureur, S., Attié, D., Bouteille, S., Calvet, D., Filosa, C., Magnier, P., Mandjavidze, I., . . . Tayoubi, M. (2017). Discovery of a big void in Khufu's Pyramid by observation of cosmic-



Vol. 4 No. 1 (January) (2026)

- ray muons. *Nature*, 552(7685), 386–390. <https://doi.org/10.1038/nature24647>
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.
- Nockels, J., Gooding, P., Ames, S., & Terras, M. (2022). Understanding the application of handwritten text recognition technology in heritage contexts: A systematic review of Transkribus in published research. *Archival Science*, 22(3), 367–392. <https://doi.org/10.1007/s10502-022-09397-0>
- Parker, C. S., Parsons, S., Bandy, J., Chapman, C., Coppens, F., & Seales, W. B. (2019). From invisibility to readability: Recovering the ink of Herculaneum. *PLOS ONE*, 14(5), Article e0215775. <https://doi.org/10.1371/journal.pone.0215775>
- Putnam, L. (2016). The transnational and the text-searchable: Digitized sources and the shadows they cast. *The American Historical Review*, 121(2), 377–402. <https://doi.org/10.1093/ahr/121.2.377>
- Sheisha, H., Kaniewski, D., Marriner, N., Djamali, M., Younes, G., Chen, Z., El-Qady, G., Saleem, A., Véron, A., & Morhange, C. (2022). Nile waterscapes facilitated the construction of the Giza pyramids during the 3rd millennium BCE. *Proceedings of the National Academy of Sciences*, 119(37), Article e2202530119. <https://doi.org/10.1073/pnas.2202530119>
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755–759. <https://doi.org/10.1038/s41586-024-07566-y>
- Sommerschild, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I., & de Freitas, N. (2023). Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3), 703–747. https://doi.org/10.1162/coli_a_00481
- Tallet, P. (2017). *Les papyrus de la mer Rouge I: Le « journal de Merer » (Papyrus Jarf A et B)*. Institut français d'archéologie orientale.
- Toynbee, A. J. (1934–1961). *A study of history* (Vols. 1–12). Oxford University Press.
- Trigger, B. G. (2003). *Understanding early civilizations: A comparative study*. Cambridge University Press.
- Underwood, T. (2019). *Distant horizons: Digital evidence and literary change*. University of Chicago Press.
- Vesuvius Challenge. (2024, February 5). *Vesuvius Challenge 2023 Grand Prize awarded: We can read the first scroll!* <https://scrollprize.org/grandprize>
- Walters, W. H., & Wilder, E. I. (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports*, 13, Article 14045. <https://doi.org/10.1038/s41598-023-41032-5>
- White, H. (1973). *Metahistory: The historical imagination in nineteenth-century Europe*. Johns Hopkins University Press.