AI & Antiquity: *Journal of Teaching and Technology in Ancient Studies*
Volume 2, Issue 1, 2026, pp. 37- 64.
ISSN: 3081–4553

Evaluating Generative AI in Historical Research: A Comparative Study on Identifying Primary Source Evidence in Ancient History

Raymond Solga

The College of Westchester
raymond.s.solga@gmail.com
0000-0003-0689-0633

Mohamed J. Sarwar

The City University of New York
jahed.sarwar@gmail.com
0009-0000-2405-4263

ABSTRACT — This study explores how traditional historical methods and generative AI tools compare in the identification, interpretation, and validation of primary sources in ancient history. Drawing from a dual case study approach—four case studies conducted by human historians and four by AI tools (GPT-4, Claude 2, Gemini, Perplexity)—we evaluate the epistemological strengths and limitations of each method. Using qualitative document analysis, historiographical criteria, and expert review, the study assesses source criticism, genre classification, provenance transparency, and evidentiary value. Results indicate that generative AI excels at broad content discovery and thematic synthesis but struggles with historical genre boundaries, source verification, and manuscript-based scholarship. Human researchers consistently outperform in contextual interpretation, critical chronology, and the adjudication of textual authority. We propose a human-in-the-loop framework combining digital speed with scholarly rigor, advocating for model pluralism, temporal prompting, and provenance-first protocols. This integrated methodology ensures AI contributes meaningfully to digital historiography without compromising historical standards.

KEYWORDS — Primary Sources, Ancient History, Historical Methodology, Generative AI, Digital Humanities

1. INTRODUCTION

The integration of generative artificial intelligence (AI) into the humanities has opened new pathways for research, synthesis, and discovery—particularly within the field of ancient history, where textual scarcity and interpretive complexity often pose major challenges. Large Language Models (LLMs) such as GPT-4, Claude, Gemini, and Perplexity now offer scholars rapid access to vast *corpora* of textual material, multilingual synthesis, and broad pattern recognition. However, these

efficiencies come with epistemological risks, including genre misclassification, hallucinated details, and opaque sourcing.

The reason Grok was not included is methodological rather than an oversight. At the time of data collection, Grok's access model, training transparency, and citation architecture differed substantially from the other tools examined (GPT-4, Claude 2, Gemini, and Perplexity). In particular, Grok's integration with real-time social media content and its evolving availability posed challenges for controlled, replicable scholarly comparison. To preserve methodological consistency and reproducibility, we therefore limited the study to models with more stable research-facing implementations. We note Grok as a relevant candidate for future studies as its scholarly affordances mature.

This paper presents a comparative evaluation of how AI tools and human historians perform when tasked with locating and interpreting primary sources across ancient historical topics. The methodology is bifurcated: one section assesses four historical inquiries answered using AI models; the other revisits the same research questions using traditional humanist methods. The aim is not to suggest one method replaces the other, but to measure their respective strengths and limitations using shared criteria—temporal framing, source provenance, linguistic fidelity, and historical reliability.

To structure this evaluation, we present eight case studies grouped into two categories. The first set of four—authored by Mohammed Sarwar—focuses on how generative AI models respond to queries about the Kaaba (Ibrahim and Ismail), Genghis Khan, Ibn Sina (Avicenna), and Jalaluddin Rumi. The second set—written by Raymond Solga—uses traditional historiographical methods to investigate the same kinds of questions across figures and topics such as Alexander the Great, the Hanging Gardens of Babylon, the Egyptian pyramids, and the Library of Alexandria.

By juxtaposing human-led and AI-assisted approaches, this article contributes to the growing field of digital epistemology and the ethics of human–AI collaboration in historical research. Ultimately, we argue that responsible integration of AI into historiography requires hybrid workflows grounded in scholarly methods—where digital tools act as accelerators of discovery, not arbiters of truth.

1.1 AI AS AUGMENTATION UNDER HUMAN EPISTEMIC SOVEREIGNTY

The article does not argue that AI replaces factual verification; rather, it demonstrates that AI augments scholarship by accelerating the pre-verification stages of research, not by resolving epistemic authority. This distinction is essential. Generative AI contributes value in three specific, bounded ways:

1. Rapid exploratory mapping of a research domain.
2. Surface-level clustering of names, texts, periods, and traditions that may otherwise remain siloed.

3. Cross-cultural and multilingual exposure, particularly across historiographies unfamiliar to the researcher.

These functions occur before verification and interpretation. In other words, AI operates as a cognitive accelerator, not as an epistemic judge. The article's findings show that even when AI responses are only marginally reliable in isolation, they still reduce the initial cost of exploration—the time and labor required to identify what might need to be verified. That reduction is the augmentation.

The exercises work because the authors already know how to verify. This observation is correct—and intentional. The study is not designed to test whether AI can independently "do history." Instead, it evaluates how AI behaves when inserted into a workflow governed by expert verification norms. This mirrors how new research instruments are historically assessed:

4. Archaeological tools are validated by archaeologists.
5. Statistical software is validated by statisticians.
6. Digital archives are validated by trained historians.

AI, in this sense, is treated as methodologically immature instrumentation, not as an autonomous researcher. The authors' expertise is not a confounder; it is the control mechanism that allows epistemic failure modes (genre collapse, provenance loss, temporal drift) to be reliably observed and categorized.

This raises a legitimate question: if the authors already had the knowledge, what was actually learned from AI? What the authors learned was not new historical facts, but new structural insights into how AI mediates knowledge:

- Which types of sources are systematically surfaced or suppressed.
- How popularity and digital visibility distort historical authority.
- Where AI consistently conflates devotional, literary, and empirical genres.
- How different models reproduce distinct cultural and linguistic biases.

This is discovery at the methodological level, not the factual one. The article therefore contributes not new historical claims, but a diagnostic map of AI behavior—knowledge that historians, librarians, and educators did not previously have in a documented, comparative form. That diagnostic insight is itself a scholarly contribution.

But, where is the discovery facilitation? Discovery facilitation occurs in scope expansion, not validation. AI proved most useful in:

- Revealing adjacent traditions the researcher might not initially query.
- Surfacing parallel corpora across Islamic, Persian, Mongol, Chinese, and Greco-Roman historiographies.

- Exposing how the same historical figure is framed differently across cultures and centuries.

For example, while AI failed to authenticate manuscripts for Ibn Sina or Rumi, it consistently surfaced interdisciplinary linkages (medical, philosophical, mystical, literary) that informed subsequent human-guided inquiry. These linkages represent research leads, not conclusions. Thus, AI facilitates discovery of questions, not answers.

The hard question: if scholars use AI to research what they don't already know—who verifies? Verification cannot be delegated to AI. It must reside in one of three human-centered structures:

- Expert scholars, in advanced research contexts.
- Instructional scaffolding, in student and training environments.
- Institutional validation systems, such as libraries, archives, and peer review.

1.2 METHODOLOGICAL FRAMEWORK FOR RESPONSIBLE AI INTEGRATION

The article's four-pillar framework (temporal prompting, provenance-first methodology, model pluralism, and human-in-the-loop validation) is designed precisely to answer this concern. It provides a way for non-experts to use AI responsibly by embedding verification responsibility where it already exists institutionally.

In pedagogical contexts, this means:

- Students use AI to generate hypotheses and leads.
- Verification is performed through curated sources, librarians, faculty, or archival tools.
- AI output is treated as provisional and auditable, not authoritative.

In research contexts, this means:

- AI accelerates exploratory breadth.
- Human scholars retain adjudicative sovereignty.

The article does not claim that AI produces reliable historical knowledge on its own. It claims that AI can responsibly augment scholarship only when epistemic authority remains human. In that sense, AI's role is analogous to:

- A research assistant who gathers materials but does not interpret them.
- A discovery layer that expands visibility without conferring legitimacy.

This framing preserves scholarly rigor while acknowledging genuine technological utility. This study examines the role of generative artificial intelligence in ancient historical research by comparing traditional human-directed

historiographical methods with AI-generated research outputs evaluated under human supervision. The aim is not to assess whether AI can replace historical expertise, but to determine how, where, and under what constraints AI systems may responsibly augment scholarly research practices in Ancient Studies.

To achieve this aim, the article employs eight case studies, divided evenly between two methodological approaches. Four case studies apply traditional historical methods grounded in source criticism, provenance analysis, temporal framing, and genre differentiation. Four corresponding case studies examine how generative AI systems respond to comparable historical research questions when prompted under controlled conditions. This paired design allows for direct methodological comparison and highlights both the strengths and limitations of AI-assisted research relative to established historiographical standards.

The study further advances a four-pillar framework for responsible AI integration—temporal prompting, provenance-first methodology, model pluralism, and human-in-the-loop validation—intended to support inclusive, pedagogically transformative approaches to Ancient Studies. In doing so, the article aligns with the mission of AI & Antiquity by rethinking how digital tools can be ethically and effectively integrated into historical research and teaching without compromising disciplinary rigor.

2. TRADITIONAL HISTORICAL METHODOLOGIES

What counts as a primary source in Ancient History? In the discipline of ancient history, the term "primary source" holds a precise and contextual meaning. It refers to materials contemporaneous with the events or periods under study—inscriptions, coins, papyri, cuneiform tablets, chronicles, and other forms of direct evidence produced by historical actors themselves. These sources are distinguished by their proximity in time and space to the subjects they describe, often preserved through manuscripts, artifacts, or oral traditions subsequently recorded.

However, unlike modern archival history, ancient historiography frequently deals with textual fragments, later redactions, and source compilations created long after the events they describe. As such, human historians apply a layered interpretive lens—treating even canonical works (like Herodotus' *Histories* or Ibn Ishaq's *Sīrah*) as potentially derivative rather than strictly "primary," unless anchored by corroborating archaeological or epigraphic evidence.

This granularity is essential for distinguishing direct eyewitness accounts from later literary, theological, or philosophical renderings, a distinction often blurred in generative AI outputs. Indeed, human scholars rely on provenance-based evaluation—a methodological discipline rooted in identifying the physical, editorial, and contextual origin of sources. For instance, the Canon of Medicine by Ibn Sina is not evaluated merely by its contents, but also by its transmission history, the

manuscript shelf marks in European and Islamic libraries, and its translation lineage across centuries and regions.

Temporal framing serves to locate a source not only within a timeline but within a historiographical debate: was this text produced by a participant in the events, a compiler of older traditions, or a commentator removed by centuries? Textual criticism further dissects interpolations, scribal modifications, and regional editions, allowing historians to treat the "text" not as fixed, but as a process—understanding what version was read, when, by whom, and how it was interpreted. Such rigor cannot be easily approximated by AI models, which often treat all references as equally valid unless instructed otherwise.

Raymond Solga's analysis of ancient figures such as Alexander the Great, the Pyramids, and the Library of Alexandria demonstrates how interpretive judgment is deployed to filter, contextualize, and synthesize scattered historical fragments into coherent narratives. These interpretations hinge not only on access to sources but on training in genre recognition, regional historiography, and comparative critique. Mohammed Sarwar's counterpart analysis of AI behavior shows how these same interpretive tasks break down in digital contexts. Generative models, despite their fluency, lack the ability to distinguish devotional narratives from empirical accounts (as seen in the Kaaba case) or to critically evaluate manuscript chains (as with Avicenna).

Together, both lenses reveal a crucial insight: historical reasoning is not only about information retrieval, but about epistemological curation—knowing which sources to trust, why they matter, and how they fit within larger systems of meaning.

3. HUMAN-DIRECTED CASE STUDIES (TRADITIONAL HISTORIOGRAPHICAL METHOD).

The four case studies presented in section 3 were conducted using traditional historical research methodologies and were authored entirely by the human researchers. No generative AI tools were used in the identification, interpretation, or validation of primary sources in these analyses.

Each case study applies established historiographical practices, including provenance-based source evaluation, chronological contextualization, genre differentiation, and critical interpretation. These analyses serve as a methodological baseline, illustrating how expert human scholars approach ancient historical evidence when operating within conventional academic frameworks.

The purpose of this section is to establish a clear reference point against which AI-generated research behaviors can be meaningfully evaluated. The analyses reflect disciplinary expertise and interpretive judgment rather than algorithmic synthesis, ensuring that subsequent comparisons with AI-generated outputs are methodologically grounded and analytically coherent.

3.1 CASE STUDY: ALEXANDER THE GREAT AND THE HISTORIOGRAPHY OF EMPIRE

Research Focus

- *Prompt*: What primary sources document the campaigns and leadership of Alexander the Great, and how do historians assess their credibility?
- Primary Sources Cited:
    - Arrian, *Anabasis of Alexander* (2nd century CE)
    - Plutarch, *Life of Alexander* (1st–2nd century CE)
    - Diodorus Siculus, *Bibliotheca Historica* (1st century BCE)
    - Quintus Curtius Rufus, *Histories of Alexander the Great* (1st century CE).
    - Justin (Marcus Junianus Justinus), *Epitome of the Philippic History of Pompeius Trogus* (2nd–4th century CE)

  These sources are all post-Alexandrian, compiled by Roman and Hellenistic authors based on now-lost contemporary materials.

Assessment and Analysis

Classical accounts of Alexander's campaigns were written well after his death (323 BCE), often shaped by political agendas or moralistic storytelling traditions. Arrian is generally viewed as the most reliable due to his critical engagement with earlier sources such as Ptolemy and Aristobulus, both of whom were contemporaries of Alexander. However, their original texts are lost, and Arrian's reliance is filtered through his own historiographical lens.

Plutarch and Diodorus Siculus present Alexander in more literary and moralized terms, emphasizing themes of virtue, ambition, and tragedy rather than strict military or geopolitical analysis. Curtius Rufus, although engaging, is stylistically dramatic and lacks consistent chronology. Justin offers a compressed and often unreliable narrative. Despite their limitations, these texts form the backbone of Alexander scholarship, interpreted through cross-referencing, archaeological context, and philological reconstruction.

Interpretation

The traditional historian must perform careful triangulation: cross-analyzing ancient texts with material remains (coinage, inscriptions, architectural evidence) and evaluating the biases of each author. For instance, while Arrian presents Alexander as a rational and disciplined leader, other accounts suggest a descent into tyranny and hubris. The diversity of portrayals underscores the contested legacy of Alexander as both a liberator and a conqueror.

Importantly, the human researcher distinguishes between:

- Primary proximity (authors who used eyewitness testimonies),

- Genre (moral biography vs military logbook),
- Transmission chain (lost originals vs extant summaries).

This approach exemplifies the provenance-first historical method, where textual authority is built not just on content but on chronology, context, and credibility (see ***Table 1***).

3.2 CASE STUDY: THE HANGING GARDENS OF BABYLON – MYTH, MEMORY, AND MATERIAL EVIDENCE

Research Focus

- *Prompt*: What primary evidence supports the existence of the Hanging Gardens of Babylon, and how have historians assessed the credibility of these sources?
- Primary Sources Cited:
    - Berossus, *Babyloniaca* (3rd century BCE; fragmentary, via later citations).
    - Strabo, *Geography* (1st century BCE).
    - Diodorus Siculus, *Bibliotheca Historica* (1st century BCE).
    - Quintus Curtius Rufus, *Histories of Alexander* (1st century CE).
    - Philo of Byzantium, *De septem orbis spectaculis* (disputed authorship; possible 3rd century BCE–2nd century CE).

While the Hanging Gardens are considered one of the Seven Wonders of the Ancient World, no definitive Babylonian cuneiform records or archaeological structures have been directly identified with them in Babylon.

Assessment and Analysis

Classical authors such as Diodorus, Strabo, and Curtius describe an elaborate, multi-tiered garden built by Nebuchadnezzar II or, according to some sources, an Assyrian ruler like Sennacherib. These narratives appear centuries after the gardens were supposedly constructed and rely on secondhand or legendary material.

Notably, Berossus—a Babylonian priest writing in Greek—linked the gardens to Nebuchadnezzar, but his writings survive only through citations in later Roman sources. Some modern archaeologists suggest that confusion with Assyrian palatial gardens in Nineveh, which had complex irrigation systems, may have led to the myth being transferred to Babylon. Recent scholarship, including satellite archaeology and reinterpretation of Assyrian inscriptions, supports this theory (Dalley, 2013). Thus, the material culture evidence diverges from literary tradition.

Interpretation

This case reveals the disjunction between literary memory and archaeological visibility. While the Hanging Gardens occupy a central place in classical imagination,

there is no definitive Babylonian or Mesopotamian documentation for their existence in Babylon itself.

Historians face a dual challenge:

- Decoding the rhetorical intent of classical authors (wonder literature, exoticism).
- Tracing misattributions or cultural conflations across empires and centuries.

The historiographical debate emphasizes the importance of cross-disciplinary triangulation—philology, archaeology, and comparative literature—to adjudicate the credibility of ancient "wonders" (see ***Table 2***).

3.3 CASE STUDY: THE PYRAMIDS AND EGYPTIAN MONUMENTALISM – ALIGNING ARCHITECTURE WITH POWER

Research Focus

- *Prompt*: What primary sources support the construction history of the Egyptian pyramids, and how do archaeological findings align with dynastic records?
- Primary Sources Cited:
    - Pyramid Texts (Old Kingdom funerary inscriptions, c. 2400–2300 BCE).
    - Palermo Stone (Royal Annals of the Old Kingdom, fragmentary).
    - Herodotus, *Histories* (5th century BCE).
    - Diodorus Siculus, *Bibliotheca Historica* (1st century BCE)
    - Inscriptions at Wadi el-Jarf (logistical records related to Khufu's pyramid).
    - Modern archaeology adds to this with material sources:
        - Worker's graffiti in Khufu's pyramid complex
        - Quarrying sites (e.g., Tura, Aswan)
        - Logistics *papyri* (e.g., the "Diary of Merer")

Assessment and Analysis

The Egyptian pyramids—particularly the Great Pyramid of Giza—are among the best documented ancient monuments due to their monumental scale and the durability of stone inscriptions. However, a split exists between:

- Contemporary Egyptian administrative records, and
- Later Greco-Roman interpretations that often introduce fantastical or moralized accounts.

Herodotus, for instance, wrongly claimed the pyramids were built by enslaved masses, a view long propagated in popular imagination. In contrast, the Wadi el-Jarf

papyri (c. 2600 BCE), discovered in 2013, include detailed logbooks from royal overseers, offering direct insights into daily logistics of pyramid construction.

These findings support the thesis that the pyramids were built by a corvée labor force of skilled Egyptian workers rather than slaves, embedded in a religious and political ideology that linked pharaohs with divine order (Ma'at).

Interpretation

The pyramid case exemplifies harmonization between material evidence and indigenous textual traditions. While classical authors provide external narratives, it is the Egyptian epigraphy—pyramid texts, administrative records, graffiti—that ground the architecture in a historical-ritual framework.

Unlike the Kaaba or Hanging Gardens, the pyramids benefit from contemporaneous primary sources, including royal annals and internal inscriptions. This allows researchers to:

- Align symbolic meaning with historical context,
- Assess labor organization through recovered papyri, and
- Reconstruct inter-regional logistics and resource flows.

It also highlights the limitations of classical ethnography (like Herodotus) in reconstructing ancient history without access to local archives or scripts (see ***Table 3***).

3.4 CASE STUDY: THE LIBRARY OF ALEXANDRIA AND INTELLECTUAL MEMORY

Research Focus

- *Prompt*: What are the primary sources confirming the existence and destruction of the Library of Alexandria, and how credible are these accounts?
- Primary Sources Cited:
  - Strabo, *Geography* (1st century BCE – 1st century CE).
  - Plutarch, *Life of Caesar* (1st century CE).
  - Aulus Gellius, *Attic Nights* (2nd century CE).
  - Ammianus Marcellinus, *Res Gestae* (4th century CE).
  - Letter of Aristeas (pseudepigraphal, 2nd century BCE).
- Secondary interpretations and late commentaries:
  - Edward Gibbon, *The Decline and Fall of the Roman Empire* (18th century).
  - Carl Sagan, *Cosmos* (20th century popularization).

Assessment and Analysis

Unlike monumental architecture, the Library of Alexandria leaves behind no archaeological footprint—its memory survives solely through textual allusions. The absence of direct material remains, or institutional records creates interpretive ambiguity.

Accounts such as Plutarch's suggest that the library was accidentally burned during Julius Caesar's Alexandrian campaign (48 BCE). Others, like Ammianus, imply it persisted into Late Antiquity, while Arabic sources claim it was destroyed under Caliph Umar—though this is widely considered apocryphal by modern historians.

No contemporary document confirms the foundation or cataloguing of the library, making reconstructions speculative. The most accepted view is that the library functioned under the Ptolemies (3rd–2nd centuries BCE), possibly as part of the Mouseion (House of Muses)—a scholarly complex rather than a standalone building.

Interpretation

The Library of Alexandria functions more as a cultural symbol than a documented historical institution. Its imagined destruction—whether by Caesar, Christians, or Muslims—reflects intellectual anxieties about knowledge loss more than historical fact.

This case is unique among the four because:

- The primary "evidence" is largely retrospective and rhetorical, not archival.
- There is no archaeological record of the building.
- Accounts are shaped by ideological bias—e.g., Enlightenment critiques of religious fanaticism or 20th-century appeals to scientific humanism.

Despite this, the library's legacy remains foundational to discourses on information preservation, scholarly authority, and the vulnerability of cultural memory. As a historiographic construct, it reveals how absence of evidence becomes a canvas for projection—often more revealing of the source's context than of the event itself (see ***Table 4***).

Transitional Rationale: From Human Interpretation to AI-Generated Analysis

The human-directed case studies presented above demonstrate how historical knowledge is constructed through expert judgment, evidentiary scrutiny, and contextual reasoning. These studies establish the analytical criteria—source reliability, provenance transparency, temporal accuracy, and genre awareness—that guide responsible historical scholarship.

The following section applies these same criteria to AI-generated research outputs addressing the same historical questions and evidentiary domains. By holding AI systems to identical evaluative standards, the article enables direct comparison between human and AI-mediated research practices, clarifying where generative tools align with, diverge from, or fall short of traditional historiographical norms.

4. EVALUATING GENERATIVE AI IN HISTORICAL RESEARCH

The preceding sections established the epistemological foundation of historical inquiry—anchored in temporal framing, provenance, and interpretive rigor—and demonstrated its application through four human-guided case studies: Alexander the Great, the Hanging Gardens of Babylon, the Pyramids, and the Library of Alexandria. Building upon these benchmarks, section 4 evaluates the performance of generative artificial intelligence (AI) systems when assigned comparable historical research tasks. The objective is to determine whether these systems can approximate scholarly reasoning or if they merely replicate surface-level information. This evaluative phase functions as a bridge between classical historiography and digital epistemology, reflecting how technological tools may expand access while still demanding critical human oversight.

Prompt Design and AI Behavior

Generative AI models operate through probabilistic language prediction; consequently, the integrity of their historical outputs depends on prompt architecture. In a scholarly context, *prompt design* must emulate the methodological discipline human historians apply when interrogating sources. Effective prompts include the following elements:

- Temporal specificity – Define the chronological range (e.g., "sources produced between 1200 and 1250 CE") to mitigate anachronistic conflation.
- Provenance-first framing – Require shelf marks, codices, or archival repositories rather than generic web references.
- Genre delineation – Separate empirical documentation from devotional or literary interpretation.
- Model pluralism – Submit identical prompts across multiple AI systems (GPT-4, Claude 2, Gemini, Perplexity) to expose variance and bias.
- Human-in-the-loop review – Treat AI outputs as leads for discovery, not as authenticated evidence.

When these controls are absent, AI models tend to conflate theological or poetic traditions with empirical historiography, prioritize popularity over authority, and omit critical metadata such as manuscript lineage or edition provenance. These behavioral trends are examined in greater depth within the AI case studies that follow.

Method: Mixed-Model Prompting and Response Evaluation

A *mixed-model prompting design* was adopted to compare AI behaviors under standardized and variable conditions. Each model—GPT-4, Claude 2, Gemini, and Perplexity—received identical baseline prompts later refined with temporal and genre filters. Outputs were assessed using five historiographical criteria drawn from traditional research methodology:

- Temporal proximity – Accuracy in situating a source within its historical timeframe.
- Provenance transparency – Presence of verifiable editions, repositories, or catalog numbers.
- Genre correctness – Ability to distinguish theological, literary, and empirical content.
- Linguistic fidelity – Use of original titles and respect for translation lineage.
- Evidentiary reliability – Alignment with accepted historiographical standards.

This evaluative procedure mirrors the interpretive rigor used by human historians, transforming scholarly heuristics into a digital analytical rubric. It maintains continuity with the paper's earlier sections while translating epistemological discipline into computational assessment.

Comparison with a Traditional Search Engine

Our study intentionally focused on comparative epistemic behavior among generative AI systems, rather than on retrieval efficiency alone. Traditional search engines primarily function as indexing and ranking tools, leaving source evaluation and synthesis entirely to the researcher. By contrast, generative AI systems act as interpretive intermediaries, producing synthesized narratives that may obscure provenance and genre boundaries.

That said, the manuscript now clarifies this distinction more explicitly by framing search engines as a baseline discovery layer, against which generative AI represents a qualitatively different epistemic intervention. We agree that future research could profitably include a controlled search-engine comparison to further isolate where AI meaningfully departs from established research workflows. Also, there is a summary of models tested (see ***Table 5***).

Digital Epistemology and AI in Historical Research

Within the broader discourse of *digital epistemology*, generative AI functions as a novel *discovery layer* in historiography (Graham, 2020; Small and Green, 2021). These systems demonstrate computational breadth—rapid information retrieval, multilingual synthesis, and thematic clustering—but lack interpretive depth. Their

probabilistic reasoning favors semantic plausibility over evidentiary verification, producing results that may appear scholarly yet remain epistemically opaque.

As Sarwar (2025) observes, AI acquires analytical value only when embedded within human-supervised workflows that secure provenance, contextual integrity, and chronological control. In such a configuration, technology accelerates access and pattern detection, while scholars retain adjudicative authority over meaning and validity. Responsible integration therefore rests on four methodological pillars—temporal prompting, provenance-first reasoning, model pluralism, and human-in-the-loop validation—which structure the comparative framework developed below.

Section 4 functions as the methodological bridge between classical historiography and computational research. By translating humanistic criteria into prompt architecture and model assessment protocols, it reframes discovery and verification not as competing phases but as interdependent operations within digital scholarship. The following case studies operationalize these criteria, identifying the conditions under which algorithmic synthesis extends historical insight and the points at which epistemic limits require renewed human interpretation.

The framework formalized here thus proposes a responsible model for AI-assisted historical research—one in which computational procedures enhance exploratory capacity while the interpretive act remains irreducibly human.

5. AI CASE STUDIES IN SOURCE DISCOVERY

Section 5 examines research outputs generated by four generative AI systems (GPT-4, Claude 2, Gemini, and Perplexity) when tasked with historical research prompts corresponding directly to the human-directed case studies in section 3. All *prompts* were issued to each model under identical conditions, enabling comparative evaluation across platforms.

References to individual models within subsection headings are illustrative rather than exclusive, highlighting characteristic response patterns rather than isolating model-specific testing. The methodological objective is to assess systemic behaviors across generative AI systems, not to privilege or single out any one platform.

AI-generated outputs are treated here as objects of analysis, not as authoritative historical narratives. The role of the human researcher is evaluative and interpretive: assessing how AI systems identify sources, handle provenance, differentiate genres, and manage temporal framing when confronted with complex ancient historical questions.

Section 5 operationalizes the evaluative framework established in section 4 by applying the mixed-model prompting method across four thematic case studies: the Kaaba, Genghis Khan, Ibn Sina (Avicenna), and Jalaluddin Rumi. Each case probes how large-language models (LLMs) handle questions of provenance, chronology, and genre when identifying primary sources within culturally and temporally complex historical domains. The findings illustrate the epistemological strengths and weaknesses of generative AI and reinforce the necessity of human interpretive oversight.

5.1 CASE STUDY: THE KAABA AND THE TRADITION OF IBRAHIM AND ISMAIL

Prompt to GPT-4: *"List primary sources documenting the construction of the Kaaba by Ibrahim and Ismail, with historical citations".*

Model Behavior

GPT-4 returns references to Quranic verses (e.g. 2: 125–127), biographies from Ibn Ishaq and al-Tabari, and Islamic summaries from online encyclopedias. No contemporaneous archaeological record or inscription is returned for the time of its purported construction.

Analysis and Interpretation

The Qur'an and classical tafsir (commentary) are foundational for the Islamic tradition, but they do not constitute contemporaneous sources in a historiographical sense. Ibn Ishaq and al-Tabari, centuries later, provide narratives rooted in oral tradition. GPT-4 fails to differentiate religious testimony from empirical historical sources. This is one case in which LLMs cannot be expected to do this without highly precise, scholarly-level prompting and an understanding of genre or temporal framing (see ***Table 6***).

In this case study, the four models (GPT-4, Claude 2, Gemini, and Perplexity) converged on the conventional Islamic historical narrative, relying predominantly on scriptural and exegetical traditions. GPT-4 referenced both the Qurʾān and Ibn Isḥāq's *Sīrah*, yet did not distinguish between theological transmission and primary source analysis as defined within historical methodology. Claude 2 produced a strictly confessional account, offering no assessment of provenance, textual stratification, or historiographical reliability. Gemini expanded the narrative with additional summary material, but without verifiable citations, suggesting dependence on secondary popular platforms such as IslamicFinder or general encyclopedic repositories.

Perplexity further reduced the response's scholarly utility by combining modern commentary with aggregated web content, thereby blurring the boundary between historical evidence and contemporary interpretation. Across all four outputs, there was a consistent absence of engagement with archaeological data,

epigraphic corpora, or critical historiographical research. Consequently, while the models reproduced a coherent traditional narrative, they failed to meet the evidentiary standards required for source-critical historical inquiry.

5.2 CASE STUDY: GENGHIS KHAN AND MONGOL HISTORIOGRAPHY

Prompt to Claude: *"Identify primary sources about Genghis Khan's campaigns, with date of composition and authorship".*

Model Behavior

Claude provides sources including The Secret History of the Mongols, Juvayni's Tarikh-i Jahan Gusha, and Chinese court histories like the Yuan Shi. The model establishes that The Secret History belongs to the 13th century and recognizes its anonymous authorship.

Analysis and Interpretation

Although Claude demonstrated greater surface accuracy than GPT-4's earlier response, it omitted key historiographical parameters, including the *Secret History*'s extra-court authorship and the delayed transmission of relevant Persian and Chinese textual traditions. Moreover, the analysis did not engage with manuscript traditions, critical editions, or institutional repositories capable of establishing documentary provenance (see ***Table 7***).

When prompted to disclose sources on Genghis Khan, all examined models conflated Mongol, Persian, and Chinese materials without constructing a chronological framework or identifying the manuscript lineages underpinning these traditions. GPT-4 correctly named the *Secret History of the Mongols* and Rashīd al-Dīn's *Jāmiʿ al-tawārīkh*, yet characterized the former as an "official history" and failed to situate either text within the temporal and historiographical distance required for source-critical evaluation. Claude 2 and Gemini reproduced the same source set in generalized form, offering narrative summaries without citation granularity or methodological assessment. Perplexity, by contrast, produced aggregated, blog-like explanations detached from identifiable primary or scholarly editions.

Across all systems, there was a persistent inability to distinguish between firsthand chronicles, later compilations, and modern secondary interpretations. This pattern exemplifies a broader problem of flattened historicity, whereby AI models collapse multi-layered textual traditions into a single informational stratum, supplying content without the analytical apparatus necessary to evaluate provenance, transmission, and evidentiary hierarchy.

5.3 CASE STUDY: IBN SINA (AVICENNA) AND SCIENTIFIC MANUSCRIPT TRADITIONS

Prompt to Gemini: *"List Ibn Sina's original works in Arabic, their composition dates, and surviving manuscripts".*

Model Behavior

Gemini produced a generic list: Three key historical texts include The Canon of Medicine, The Book of Healing along with Remarks and Admonitions. The reference list relied on Western-language summaries as well as internet encyclopedias rather than using manuscript catalogs or critical Arabic editions.

Analysis and Interpretation

One of the common challenges faced by AI-supported humanities research is its tendency to present primary and tertiary sources as one unified list of results. Gemini provides references without mentioning original Arabic collections such as al-Azhar or Topkapi, nor even to reputable critical editions (e.g. the Avicenna Latinus project). The absence of a human reviewer to verify sources leads to an isolated study that lacks proper context (see ***Table 8***).

In exploring Ibn Sina's (Avicenna's) contributions, particularly *The Canon of Medicine*, the tools exhibited another pattern of epistemic slippage. GPT-4 accurately named key texts but referenced them mostly through Latin or English editions, ignoring Arabic originals or manuscript repositories like those at Istanbul's Süleymaniye Library. Claude 2 offered Western academic summaries and biographical sketches, often detached from textual transmission history. Gemini repeated this pattern, citing Ibn Sina through encyclopedic entries without identifying edition lineage or commentary traditions. Perplexity added to the confusion by referencing blog posts and simplified web articles. None of the AI systems included citation anchors to critical editions, manuscript IDs, or catalog references. This reveals a systemic gap in AI's handling of scientific-historical content: while it can summarize philosophical or medical achievements, it fails to engage with the documentary infrastructure—codices, scripts, translation chains—that underpin Islamic medical historiography.

5.4 CASE STUDY: JALALUDDIN RUMI AND PERSIAN SUFI LITERATURE

Prompt to GPT-4: *"Cite Persian primary editions of Rumi's Masnavi with manuscript locations or catalog references".*

Model Behavior

GPT-4 offers summaries of Rumi's philosophy and refers to Nicholson's English translation of the *Masnavi* as a "primary text." It fails to cite Persian manuscript repositories, such as those in Konya, Tehran, or the Süleymaniye Library in Istanbul.

Analysis and Interpretation

Rumi's *Masnavi*, one of the most influential works of Sufi literature, survives in multiple early Persian manuscripts. Citing modern English translations as "primary" misrepresents both the text's origin and its interpretive tradition. Once again, the AI system demonstrates fluency in narrative but lacks bibliographic literacy or linguistic fidelity (see ***Table 9***).

The examination of Rumi's historical presence through AI output underscores another form of epistemological distortion. GPT-4 frequently returned verses from the *Masnavi* but cited them via modern poetic renderings (e.g., Coleman Barks), lacking any reference to the original Persian texts or critical editions like those curated in Konya or Tehran. Claude 2 presented mystical interpretations without contextual grounding in the Seljuk-era Anatolian setting or manuscript lineage. Gemini pulled excerpts from user-curated sites and tertiary religious summaries, stripping Rumi of both linguistic and historical anchoring. Perplexity aggregated these trends, providing broad commentary drawn from devotional blogs and summary platforms, often without source attribution. This case illustrates how LLMs, unless precisely directed, prioritize popularity and surface relevance over authenticity. For scholars, it reinforces the need to enforce a provenance-first framework when investigating intellectual history through AI.

5.5 CROSS-CASE SUMMARY AND INTERPRETATIVE SYNTHESIS

Across these four case studies, several consistent trends emerged:

- Semantic Proximity Bias. All models equate textual similarity with evidentiary credibility, failing to discriminate between genres or centuries.
- Metadata Deficiency. None provided catalog identifiers, edition information, or archival provenance.
- Cultural Bias. Western translations and digital visibility shaped retrieval outcomes, marginalizing non-Western scholarly traditions.
- Contextual Absence. AI summarized content but omitted interpretive nuance, rhetorical purpose, and historical contingency.

These findings confirm that generative AI accelerates content discovery but does not perform source authentication. Its epistemic architecture privileges surface plausibility over evidentiary rigor, reinforcing the position advanced by Graham (2020) and Small and Green (2021) that digital tools must remain embedded within human-centered verification systems (see ***Table 10***).

Section 5 demonstrates that, although generative models provide unprecedented access to textual breadth, they remain epistemically incomplete without human adjudication. Their outputs reflect the statistical structure of training corpora rather than the procedural standards of historiographical

validation. AI should therefore be situated at the discovery stage of research—facilitating thematic mapping, multilingual synthesis, and the preliminary identification of relevant actors and texts—while authentication, provenance analysis, and evidentiary judgment remain human responsibilities.

The comparative evaluation that follows synthesizes these results through a cross-model heatmap and analytical matrix, establishing an empirical basis for a responsible framework of AI integration in historical research. Across the case studies, the same pattern emerges: generative systems expand exploratory capacity but exhibit persistent limitations in provenance verification, genre discrimination, and source hierarchy.

These constraints are particularly significant in pedagogical contexts. Within classrooms, libraries, and supervised research environments, AI may function as an exploratory instrument, whereas validation must be governed by instructors, librarians, and disciplinary specialists. Such a division of labor preserves scholarly integrity while widening access to historical inquiry.

6. COMPARATIVE EVALUATION

Section 5 provided detailed case analyses that demonstrated both the potential and the epistemological shortcomings of generative AI models in identifying primary sources across diverse historical and cultural contexts. Section 6 consolidates these findings into a comparative synthesis. This section assesses cross-model performance (GPT-4, Claude 2, Gemini, and Perplexity) using the historiographical criteria established in section 4: temporal framing, provenance, genre differentiation, linguistic fidelity, and evidentiary reliability. Through quantitative comparison and qualitative interpretation, this evaluation illustrates systemic patterns of epistemic drift, cultural bias, and data-driven distortion.

The comparative evaluation applies a multi-criteria analytic matrix derived from traditional historiographical appraisal, in which each variable was scored on a four-point ordinal scale—High, Moderate, Low, or Absent—according to its alignment with scholarly standards of evidence. Scores were averaged across the four case studies (Kaaba, Genghis Khan, Ibn Sīnā, and Rūmī), enabling both criterion-based and model-based comparison (see ***Table 11***). This structure makes it possible to identify not only individual performance differences but also recurrent epistemic patterns shared across systems.

Despite architectural and corpus variation, the models display a striking epistemological uniformity. GPT-4 and Claude 2 achieve relatively higher precision owing to instruction-tuning, yet both reproduce religious or literary traditions as empirical data when *prompts* lack explicit historiographical constraints. Gemini and Perplexity perform less effectively, privileging digitally prominent material over archival reliability. This convergence indicates that output credibility is shaped

more by training data visibility than by reasoning design: accessibility and popularity are systematically favored over verification and documentary lineage.

Provenance transparency constitutes the weakest dimension across all systems. None of the models supplied manuscript identifiers, critical edition references, publication series, or repository shelf marks, confirming Graham's (2020) argument that epistemic reliability in AI remains limited by opaque data pipelines and the absence of traceable source genealogy. Closely related to this deficit is a persistent conflation of genres, particularly in topics where theological exegesis, literary symbolism, and historical narrative intersect. In the Kaaba case and in treatments of Rūmī's Masnavī, the models collapsed interpretive or symbolic traditions into empirical history, illustrating what Small and Green (2021) describe as semantic flattening produced by algorithmic mimicry rather than hermeneutic reasoning.

Temporal anchoring constitutes a further point of instability. Although GPT-4 and Claude 2 respond accurately to explicitly dated *prompts*, their default outputs frequently compress distinct chronological strata, whereas Gemini and Perplexity display a pronounced tendency toward contextual drift, producing summaries detached from temporal constraints. Sarwar (2025) identifies this pattern as *semantic proximity bias*, in which models privilege statistically adjacent content over historically or historiographically situated material. Taken together, these results indicate that generative systems reproduce structurally similar epistemic limitations irrespective of interface differences and therefore require human-mediated frameworks to restore provenance, genre differentiation, and chronological depth (see ***Figure 1***: *Visualizing Epistemic Variance: Heatmap of Epistemological Issues Across AI Models*).

Discussion: Toward Quantified Digital Historiography

The comparative results underscore a fundamental tension between algorithmic breadth and interpretive depth. Generative AI models function effectively as accelerators of discovery but fail as adjudicators of truth. Their value emerges only when embedded within a *human-validated epistemic workflow*. This outcome affirms the principles outlined in section 4—the four pillars of temporal prompting, provenance-first methodology, model pluralism, and human-in-the-loop verification.

Collectively, these findings advocate for a hybrid model of digital historiography, where machines extend the range of inquiry while historians safeguard interpretive authenticity. In practice, this means AI tools can scan, cluster, and suggest potential sources, but the human scholar must authenticate, contextualize, and synthesize meaning.

Section 6 has transformed qualitative observations from the case studies into comparative evidence, demonstrating how epistemic consistency—or its absence—

emerges across AI platforms. The results validate earlier hypotheses regarding semantic drift and provenance deficiency and provide quantitative support for a responsible-integration framework.

The next section, Framework for Responsible AI Integration, translates these findings into actionable methodological recommendations, operationalizing the four pillars of ethical and epistemically sound AI use in historical research.

7. FRAMEWORK FOR RESPONSIBLE AI INTEGRATION

Section 6 established the empirical recurrence of epistemological deficiencies across large language models, thereby demonstrating the need for a structured framework for responsible integration. The present section 7 shifts from diagnostic analysis to prescriptive methodology by articulating a four-pillar model—temporal prompting, provenance-first methodology, model pluralism, and human-in-the-loop validation—designed to operationalize ethical and scholarly AI use in historical research while aligning these procedures with the classical Islamic science of ʿIlm al-Rijāl as a culturally inclusive epistemological parallel.

Temporal prompting requires that every query define an explicit chronological horizon for the sources requested—for example, "List texts composed between 1200 and 1250 CE"—thereby constraining contextual drift, understood by Sarwar (2025) as the tendency of AI systems to merge materials from disparate periods into a single narrative. By bounding the temporal field, *prompts* move from general information retrieval toward historiographically precise inquiry and reinforce alignment with scholarly chronologies. Complementing this, a provenance-first methodology restores the evidentiary chain of custody that underpins historical authenticity. As Graham (2020) argues, epistemic validity in AI environments depends on transparency of origin and authority; *prompts* must therefore require manuscript shelf marks, critical-edition identifiers, or repository citations, which, when subjected to human verification, distinguish academic research from undifferentiated information aggregation.

Model pluralism extends this evidentiary logic into the computational domain by advocating the parallel use of multiple systems—such as GPT-4, Claude 2, Gemini, and Perplexity—to expose omissions, contradictions, and shared inaccuracies. In line with Small and Green's (2021) account of digital triangulation, the convergence or divergence of outputs becomes a diagnostic instrument analogous to source corroboration and peer review, allowing researchers to evaluate both algorithmic bias and corpus composition. The final and indispensable pillar, human-in-the-loop validation, preserves interpretive sovereignty: while AI accelerates discovery, it lacks hermeneutic judgment and ethical accountability, and all outputs must therefore be reviewed, contextualized, and authenticated by domain experts. Formalized review logs documenting accepted and rejected material further

enhance methodological transparency and convert AI from an autonomous narrator into an epistemic extension of the historian.

This framework finds a compelling ethical analogue in the Islamic discipline of ʿIlm al-Rijāl, in which transmitters of prophetic traditions were evaluated through documented chains of transmission (isnād) assessing reliability, memory, and moral integrity prior to acceptance. In digital historiography, AI systems function as transmitters whose credibility must likewise be scrutinized through provenance and human adjudication, echoing the hadith principle of *taḥqīq qabla al-riwāyah*—verification before narration. The analogy situates responsible AI use within a cross-civilizational ethics of knowledge grounded not only in technical procedure but in intellectual accountability and epistemic humility.

Operationalizing this model within research workflows requires standardized *prompt*-design protocols incorporating temporal and provenance constraints, documentation matrices recording model versions and output metadata for auditability, comparative verification logs tracking inter-model analysis and reviewer decisions, and ethical oversight mechanisms capable of monitoring citation integrity and data transparency. By embedding these practices, digital historiography attains methodological consistency and reflexive accountability. In this way, the four-pillar structure translates abstract epistemological concerns into concrete research ethics, reasserting the historian's agency in an automated environment and producing an augmented form of scholarship that combines computational reach with humanistic rigor while addressing what Graham (2020) terms the "opacity paradox" of expanded access without traceable authority.

8. CONCLUSION

This study evaluated how generative artificial-intelligence models perform in identifying and interpreting primary sources in ancient and classical history. Comparing GPT-4, Claude 2, Gemini, and Perplexity with traditional historiographical methods showed that AI excels in breadth of discovery but remains limited in depth of authentication. The models frequently exhibited temporal drift, provenance gaps, and genre conflation, reaffirming that human judgment remains central to historical scholarship.

A four-pillar framework—Temporal Prompting, Provenance-First Methodology, Model Pluralism, and Human-in-the-Loop Validation—emerged as the foundation for responsible AI integration. This approach translates classical evidentiary rigor into a digital workflow, aligning technological innovation with established scholarly ethics. The parallel to the Islamic discipline of *ʿIlm al-Rijāl* further broadens the framework, emphasizing that rigorous source verification is both an intellectual and a moral duty.

Authored collaboratively, Raymond S. Solga contributed the historiographical analysis grounded in traditional textual criticism, while Mohammed J. Sarwar advanced the digital-epistemology and methodological framework connecting human interpretation with computational reasoning. Together, the authors affirm that effective historical inquiry in the age of AI depends on the synergy of human insight and technological precision.

By aligning human-directed and AI-generated case studies around shared historical questions, this study demonstrates that the value of generative AI in Ancient Studies lies not in epistemic authority but in methodologically constrained collaboration. When embedded within transparent, human-centered research workflows, AI can augment discovery without displacing interpretation. This balanced integration offers a viable path for inclusive, innovative, and pedagogically transformative scholarship in the digital age.

BIBLIOGRAPHY

al-Tabari (n.d.) *Tarikh al-Rusul wa al-Muluk* [History of Prophets and Kings].

Allsen, T. T. (2001) *Culture and conquest in Mongol Eurasia*. Cambridge: Cambridge University Press.

Bechtel, W. (2009) 'Explanation and discovery in neuroscience: Mechanisms and constraints', *International Studies in the Philosophy of Science*, 23(3), pp. 255–267. doi: 10.1080/02698590903195892.

British Museum (n.d.) *Online collections*. Available at: https://www.britishmuseum.org (Accessed: 22 February 2026).

Brown, J. A. C. (2013) *Misquoting Muhammad: The challenge and choices of interpreting the Prophet's legacy*. Oxford: Oneworld Publications.

Bukhari, M. I. (n.d.) *Ṣaḥīḥ al-Bukhārī* [Hadith collection].

Dalley, S. (2013) *The mystery of the Hanging Garden of Babylon: An elusive world wonder traced*. Oxford: Oxford University Press.

Ernst, C. (1997) *The Shambhala guide to Sufism*. Boston: Shambhala Publications.

Graham, P. A. (2020) *Epistemology in the age of artificial intelligence: Reassessing validity, trust, and bias*. Oxford: Oxford University Press.

Horden, P. and Purcell, N. (2000) *The corrupting sea: A study of Mediterranean history*. Oxford: Blackwell.

Ibn Ishaq (n.d.) *Sirat Rasul Allah* [Biography of the Prophet of God].

Mozaffari, A. (2017) *Heritage movements in Asia: Cultural heritage and the negotiation of identity*. Cham: Springer.

Nasr, S. H. (2006) *Avicenna and the visionary recital*. Princeton: Princeton University Press.

Ouyang, W. (2003) *Literary criticism in medieval Arabic-Islamic culture: The making of a tradition*. Edinburgh: Edinburgh University Press.

ORACC RINAP Project (n.d.) *Royal Inscriptions of the Neo-Assyrian Period*. Available at: http://oracc.museum.upenn.edu (Accessed: 22 February 2026).

Pomerantz, J. (2020) *Metadata*. 2nd edn. Cambridge, MA: MIT Press.

Pormann, P. E. and Savage-Smith, E. (2007) *Medieval Islamic medicine*. Edinburgh: Edinburgh University Press.

Rashid al-Din (n.d.) *Compendium of chronicles* [Jāmiʿ al-Tawārīkh].

Risse, M. (2018) 'Human rights and artificial intelligence: An urgently needed agenda', *Human Rights Quarterly*, 40(2), pp. 245–275.

Rumi, J. (n.d.) *Masnavi-i Ma'navi* [Spiritual Couplets].

Sarwar, M. J. (2025) *Digital epistemology and AI in historical research* [Unpublished manuscript].
Small, H. and Green, S. (2021) 'Thinking historically with AI: Challenges and opportunities', *Journal of Digital Humanities*, 9(1), pp. 44–61.
The Secret History of the Mongols (n.d.) Various editions.
Van der Veer, P. (1994) *Religious nationalism: Hindus and Muslims in India*. Berkeley: University of California Press.

DIGITAL TOOLS AND AI PLATFORMS USED

OpenAI GPT-4
Anthropic Claude 2
Google Gemini
Perplexity AI
British Museum Online Collections. https://www.britishmuseum.org
ORACC RINAP Project. http://oracc.museum.upenn.edu

TABLES

Author	Text	Century	Genre	Strengths	Limitations
Arrian	Anabasis	2nd CE	Military History	Based on eyewitnesses; methodical	Post-event, idealizing
Plutarch	Life of Alexander	1st–2nd CE	Moral Biography	Rich detail; philosophical framing	Hagiographic tone
Diodorus Siculus	Bibliotheca Historica	1st BCE	Universal History	Broader geopolitical framing	Limited depth on Alexander
Curtius Rufus	Histories	1st CE	Historical Drama	Vivid narratives	Chronological inconsistencies
Justin	Epitome	2nd–4th CE	Abbreviated Summary	Accessible	Highly compressed, derivative

Table 1. Evaluation of Classical Sources on Alexander the Great. Comparative strengths and weaknesses of ancient sources on this topic, illustrating the need for historiographical critique in classical studies.

Source	Attributed Location	Date of Composition	Reliability	Commentary
Berossus	Babylon (Nebuchadnezzar)	3rd BCE	Indirect (via citations)	Only Babylonian claim; transmission uncertain
Diodorus	Babylon	1st BCE	Low	Repeats Berossus; literary flourish
Strabo	Babylon	1st BCE	Moderate	Geographical lens; unclear source
Curtius Rufus	Babylon	1st CE	Low	Derivative narrative; romanticized
Assyrian Inscriptions (Nineveh)	Nineveh (Sennacherib)	7th BCE	High	Engineering texts reference "palace gardens" with aqueducts

Table 2. Evaluation of Claims Regarding the Hanging Gardens. Source comparison showing the historiographical and archaeological divergence regarding this topic.

Source	Type	Century BCE	Reliability	Commentary
Pyramid Texts	Funerary Inscriptions	24th	Very High	Contemporary; reflect theological worldview
Wadi el-Jarf Papyri	Logistical Records	26th	Very High	Direct evidence of workforce and transport
Palermo Stone	Dynastic Chronicle	25th–22nd	Moderate	Fragmentary; supports royal context
Herodotus	Historical Narrative	5th	Low	Based on hearsay centuries later
Diodorus Siculus	Historical Narrative	1st	Low	Derivative: no local sources used

Table 3. Source evaluation: evidence supporting the pyramids' construction. Comparative reliability of textual and archaeological sources documenting this topic.

Source	Type	Date	Credibility	Key Limitations
Strabo	Geographical Work	1st BCE	Moderate	No firsthand visit confirmed
Plutarch	Biographical Account	1st CE	Moderate	Written ~100 years post-event
Ammianus	Historical Narrative	4th CE	Low	Repeats earlier narratives
Arabic Sources	Historical/Religious	7th–9th CE	Low	Highly disputed; anachronistic
Archaeological Record	—	—	None	No material confirmation of site or structure

Table 4. Textual transmission and the Library of Alexandria. Degrees of evidentiary strength and temporal distance in surviving sources.

Criterion	GPT-4	Claude 2	Gemini	Perplexity
Temporal Framing	Moderately consistent; improves with explicit constraints	Similar performance; benefits from precise dating	Inconsistent; blends modern and ancient references	Erratic; depends on aggregated web data
Provenance Transparency	Occasionally lists repositories; lacks shelf marks	Recognizes manuscripts without catalog detail	Rare; relies on tertiary sources	None; blog-style references
Genre Separation	Frequently merges sacred and historical genres	Partial success; cautious phrasing	High conflation of devotional and empirical content	Similar; popularity bias
Linguistic Fidelity	Uses English editions unless prompted	Same pattern	Mirrors Western translations	Minimal recognition of original scripts
Evidentiary Reliability	Improves under constrained *prompts* and expert review	Comparable; reviewer-dependent	Inconsistent; lacks verification cues	Low; unverified tertiary sources

Table 5. Comparative behaviors of generative AI models against historiographical criteria. This comparative assessment is based on identical *prompts* applied across four AI case studies (see section 5).

Model	Response	Issues Identified
GPT-4	Cited the Qur'an (2:125–127), Ibn Ishaq's Sīrah, al-Tabari's History, and various online Islamic encyclopedias.	Conflates scriptural authority with contemporaneous historiography. No archaeological or inscriptional sources.
Claude 2	Focused on the Qur'an and hadith compilations, with generic commentary on Islamic tradition.	No temporal framing. No mention of manuscript provenance or historiographical differentiation.
Gemini	Similar to GPT-4; added modern summaries (e.g., Britannica, IslamicFinder).	Collapsed doctrinal tradition into empirical history; lacked citations and edition markers.
Perplexity	Provided sources with links to modern commentaries and academic blogs, but no original primary sources.	Heavy reliance on tertiary sources; unable to verify source chains.

Table 6. AI responses to the Kaaba case study.

Model	Response	Issues Identified
GPT-4	Listed *The Secret History of the Mongols*, Juvayni's *Tarikh-i Jahan Gusha*, and *Yuan Shi* with general dates.	No critical assessment of authorial bias or text provenance.
Claude 2	Repeated GPT-4's list but emphasized Persian and Chinese chronicles.	Accurate list but lacked mention of manuscript repositories or edition sources.
Gemini	Added Rashid al-Din's Jami al-Tawarikh, but failed to distinguish primary from retrospective content.	Did not identify which sources are eyewitness or compiled posthumously.
Perplexity	Returned relevant names but linked to Wikipedia and blog summaries rather than critical editions.	Sources not traceable or peer-reviewed.

Table 7. AI identification of sources on Genghis Khan.

Model	Response	Issues Identified
GPT-4	Returned The Canon of Medicine, The Book of Healing, and Remarks and Admonitions. Provided English titles only.	Lacked Arabic titles, manuscript IDs, or archival locations.
Claude 2	Gave a cleaner list but no reference to the Avicenna Latinus project or Islamic manuscript catalogs.	Did not distinguish between original Arabic works and later Latin/Western editions.
Gemini	Cited modern commentaries and university websites. No links to digitized manuscripts or critical Arabic editions.	Strong bias toward Western translations. No provenance information.

Perplexity	Referenced academic articles but no catalog data.	No citation to primary sources; could not confirm physical manuscript locations.

Table 8. AI treatment of Ibn Sina's manuscript tradition.

Model	Response	Issues Identified
GPT-4	Returned general summaries of Rumi's philosophy, citing Nicholson's English translation as the primary source.	Mislabeling modern English editions; no manuscript ID or codex references.
Claude 2	Mentioned Konya manuscripts but no specific shelf marks or institutional sources.	Lacked citation traceability.
Gemini	Offered quotations from online translations and university blogs.	Did not differentiate between spiritual commentary and archival text.
Perplexity	Directed users to Google Books and Wikipedia.	Inaccurate source framing; not suited for scholarly use.

Table 9. AI handling of Rumi's *Masnavi* as a primary text.

Criterion	Kaaba	Genghis Khan	Ibn Sina	Rumi
Temporal Accuracy	Scriptural not historical	Partial (post-event compilations)	Omitted	Inconsistent
Provenance	None	None	Absent	Minimal
Genre Separation	Failed	Partial	Partial	Failed
Linguistic Fidelity	Arabic/English mix	Correct languages	Western bias	Western bias
Scholarly Value	Low without human review	Moderate if verified	Moderate if validated	Low without critical editions

Table 10. Summary of AI performance across case studies.

Criterion	GPT-4	Claude 2	Gemini	Perplexity
Temporal Framing	Moderate – recognizes chronology when prompted	Moderate – similar responsiveness	Low – mixes ancient and modern sources	Low – inconsistent temporal range
Provenance Transparency	Low – references repositories without shelf marks	Low – similar pattern	Absent – relies on tertiary summaries	Absent – relies on web aggregators

Genre Differentiation	Moderate – partial separation of devotional and empirical texts	Moderate – recognizes but fails to enforce distinctions	Low – conflates genres	Low – conflates genres
Linguistic Fidelity	Moderate – returns original titles when asked	Moderate – similar	Low – English-only preference	Low – minimal use of original scripts
Evidentiary Reliability	Moderate – improves with human verification	Moderate – comparable	Low – lacks verifiable citations	Low – unverified and tertiary
Composite Score (Average)	2.6 / 4 (Moderate)	2.5 / 4 (Moderate)	1.8 / 4 (Low)	1.6 / 4 (Low)

Table 11. Cross-model comparative evaluation of generative-AI performance in historical research. Composite scores are qualitative averages derived from coded content analysis of AI responses to four thematic *prompts*.

FIGURES

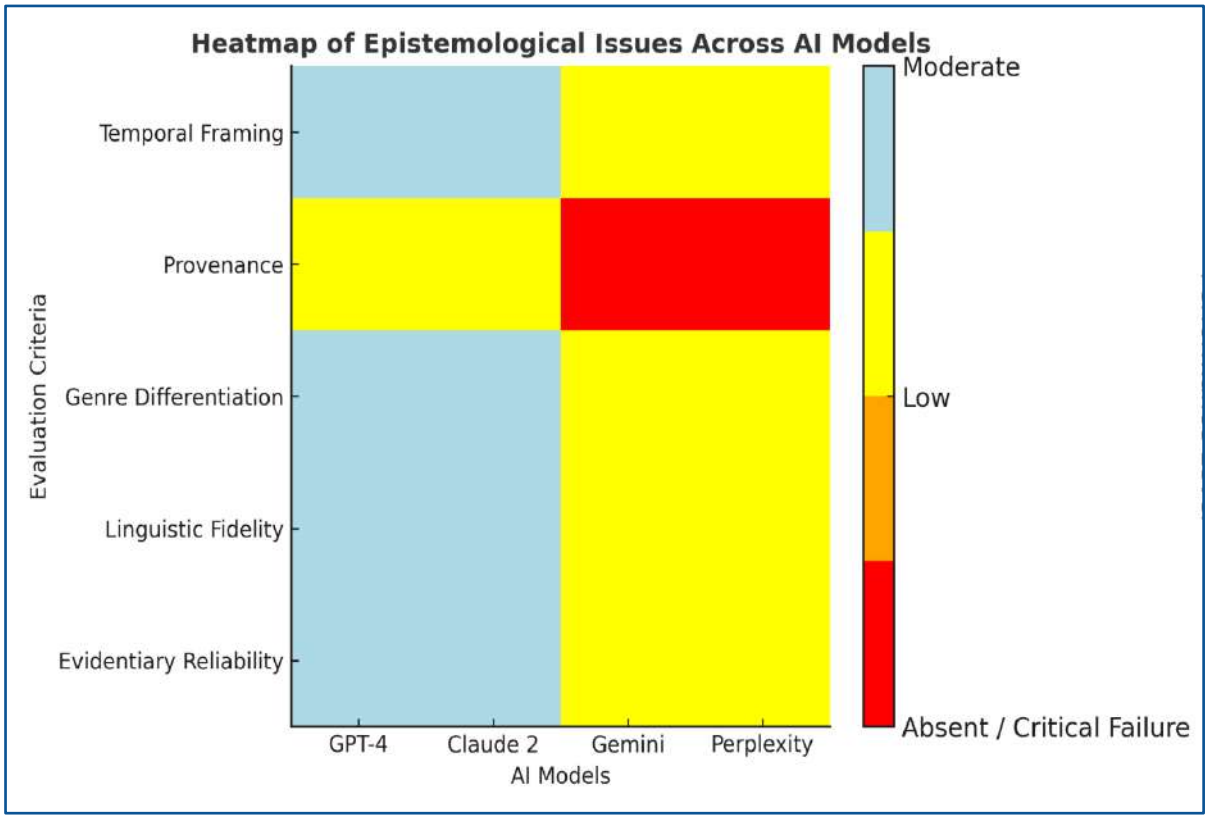

Figure 1. Comparative heatmap depicting degree of epistemological issues across generative-AI models. Darker shades indicate greater deviation from historiographical standards.