heritage

MDPI

Article

Narrative Licence: An Evidence-Bounded Protocol for Multivocal Museum Storytelling from Reused Object Records

M. Izani Zainal Abidin [1,*] and Aishah Abdul Razak [2]

[1] Applied Media Department, Higher Colleges of Technology, Abu Dhabi P.O. Box 25026, United Arab Emirates
[2] Faculty of Engineering, Multimedia University, Cyberjaya 63100, Malaysia; aishah.abdulrazak@mmu.edu.my
* Correspondence: izainal@hct.ac.ae

Abstract

Generative systems can produce fluent museum prose, but fluency does not show whether a sentence is warranted by an object record. This article proposes Narrative Licence, a protocol that classifies each candidate sentence as ATTESTED, INFERRED or NOT LICENSED against named metadata fields and withholds unlicensed sentences before public release. The framework comprises an ordered decision procedure, a field-to-licence map, three evidence-partitioned institutional voices, and a matched comparison of an unconstrained prompt with a licence-bounded pipeline. A completed pre-generation audit of the public TimeTravel release (10,250 records) found quoted inscriptions in 14.9%, recorded findspots in 68.2% and all ten licence-bearing fields in 2.9% of records, and showed that 64% of quoted inscriptions come from coin records. A deterministic working sample of 48 records was drawn in three strata. A worked example traces one record from source fields to released gallery text. The planned confirmatory evaluation will estimate the paired difference in released unlicensed-sentence rate, with coherence and readability as secondary outcomes. No effectiveness result is reported. Whether the protocol reduces unlicensed claims without degrading narrative quality remains to be tested.

Keywords: cultural heritage; museum interpretation; generative artificial intelligence; epistemic provenance; multivocality; TimeTravel

1. Introduction

Large multimodal and language models can now move quickly from catalogue fields and images to plausible descriptions, labels, audio scripts, and conversational answers. TimeTravel formalised an important part of this development by providing a structured benchmark for historical and cultural artefact understanding [1,2]. Yet a benchmark question—whether a model can produce a plausible account of an artefact—is not identical to an institutional question: what may a museum responsibly say when the available record is partial, uneven, or silent?

That distinction matters because museum speech is not merely generated content. It is public interpretation issued under institutional authority. The current ICOM museum definition places ethical and professional communication alongside accessibility, inclusion, participation, and knowledge sharing [3]. In digital heritage, the London Charter similarly connects intellectual integrity to structured source evaluation and documentation of interpretive processes [4]. Generative systems intensify this longstanding problem: a sentence can be rhetorically convincing while its warrant is absent, indirect, or stronger than the source permits.

Check for updates

Academic Editors: Matina Kiourexidou, Sofia Stamou and Evangelia Kavakli

Received: 5 September 2026
Revised: 10 October 2026
Accepted: 13 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

We therefore define narrative licence as an epistemic rule, not a copyright licence and not a caption-quality score. The object record is the sole warrant for public prose in this experiment. Every independently interpretable sentence or gallery fragment must point to a named field or to a conservatively described visible element. Sentences that cannot do so are recorded and withheld. This approach does not eliminate interpretation; it separates source-attested statements, marked visual inferences, and unlicensed narrative additions.

The present study asks whether that rule changes generated museum-facing material under a matched design. The primary comparison holds the model, input record, output genres, approximate lengths, and decoding settings constant while changing only the licence requirement and its ledger/refusal mechanism. The published TimeTravel description is retained unchanged as Voice 0, but only as a descriptive external baseline; it is not treated as ground truth, as a licensing source, or as project prose. The evaluation is restricted to output-level epistemic performance. Curator deployment, visitor response, student recall, and trail implementation are outside the present study because they require participant-based methods, distinct outcome measures, and separate ethical procedures.

Contributions

The article makes five contributions:

(i) A task shift from plausible captioning to evidence-bounded institutional narration.
(ii) A frozen protocol that binds each candidate sentence to named metadata or a marked visual inference.
(iii) An institutionally multivocal output format that separates epigraphic/object speech, provenance/making, and contemporary reception without inventing historical or community voices.
(iv) A matched, human-coded B/C evaluation design centred on released unlicensed claims and heritage-relevant errors rather than n-gram similarity.
(v) An ordered decision procedure and a complete worked example intended to make sentence classification reproducible across coders.

2. Related Work

2.1. From Artefact Description to Licensed Institutional Speech

TimeTravel contains structured object metadata, images, and generated-and-verified descriptions intended to evaluate multimodal models on historical and cultural artefacts [1,2]. The present study does not reproduce that benchmark, compare model families, fine-tune on its captions, or reuse its description field as target prose. It treats each selected record as a bounded source packet. The analytical object is therefore not semantic similarity to a caption but the warrant for each sentence released under an institutional voice.

This change of unit is consequential. Record-level fluency can mask sentence-level strengthening: a named culture can become a claim about a people; an imperial legend can become public sentiment; a findspot can become an imagined journey. Narrative Licence is designed to make those transitions auditable by requiring a field-to-claim link before prose is released.

2.2. Museum Chatbots, Audio Guides, and Curatorial Control

Museum chatbots have been built around structured information models and tested in cultural venues, demonstrating the practical value of conversational access to collections [5]. More recent work also shows why engagement alone is insufficient. Survey evidence on conversational agents records concerns about distraction, depersonalisation, and the authenticity of heritage encounters [6]. Research on personalised LLM-generated audio

guides likewise concludes that models can assist drafting but require curator-governed preparation, review, and release gates [7].

Narrative Licence complements this curator-in-the-loop orientation by specifying the object of review. Rather than asking only whether a text sounds appropriate or contains obvious factual errors, annotators ask whether each proposition is licensed by a named field, licensed as a marked visual inference, or not licensed and therefore withheld. The protocol is designed to make refusal visible as a positive curatorial act.

2.3. Multivocal and Emotionally Engaging Interpretation

Digital storytelling research has shown that affective engagement and historical accuracy need not be treated as opposites, but that the balance requires explicit design and evaluation [8]. In museum and heritage studies, multivocality usually refers to the inclusion of perspectives held by different people or groups, often communities whose knowledge has been excluded from institutional interpretation; multivocal digital exhibition research accordingly challenges the single authoritative voice by enabling culturally situated and user-curated interpretations [9].

The three voices used here are narrower and of a different kind, which we describe as institutional (evidential) multivocality. They are generated within one institutional framework and differ by the class of evidence that licenses each sentence: the quoted inscription (Voice A), provenance and making (Voice B), and present-day questioning (Voice C). They do not differ by whose perspective they carry; authority remains with the institution and is audited through the ledger; and they are designed to keep evidence types visibly apart, so that one smooth paragraph does not merge inscription, findspot and imagery into a single implied history. The protocol does not claim to reproduce community testimony or historical consciousness. A community-authored statement could enter the output only as a fourth, separately licensed source class, attributed to its author and released with consent; the present schema contains no such class and therefore generates no community voice.

2.4. Inscription Models and the Boundary of the Task

Systems such as Ithaca restore missing characters and estimate the geographical or chronological attribution of ancient inscriptions while supporting expert decision-making [10]. That is a different epistemic task. Narrative Licence does not restore, translate, or complete an inscription. It permits exact quotation and literal legend reading only when the record supplies the text. Any wider historical motive, emotion, or public response requires a separate source and is outside this experiment.

3. Materials and Methods

3.1. Study Design and Scope

The present study uses a prospectively specified, within-record comparison of Conditions B and C across 48 object records. Each record receives both prompts under the same generation system and recorded run settings. Condition order is randomised within record, and outputs are stored separately before coding. Condition A is the unchanged TimeTravel description and is used only to describe external-baseline errors such as invention, laundering, or erasure. No causal A-versus-C claim is planned because A differs in provenance, genre, and production process.

The analytical scope is deliberately limited to protocol validity and the matched B/C comparison. Practice-facing outcomes, including curator or educator judgements, visitor or student recall, and evaluation of an accessible interpretive trail, are outside the present study because they require participant-based methods and distinct outcome measures. This

boundary prevents public value from being inferred from text-level performance alone. Figure 1 summarises the resulting evidence-bounded generation and evaluation workflow.

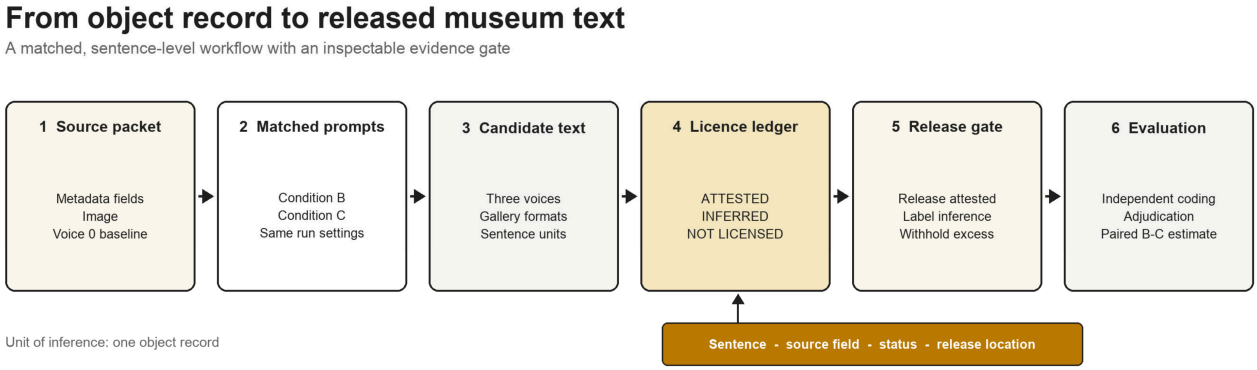


Figure 1. Evidence-bounded generation and evaluation workflow. The same source packet and run settings are used for Conditions B and C. Condition C adds the licence rule, evidence ledger, and refusal mechanism. Every candidate sentence is classified before public release; two annotators then code matched outputs independently.

3.2. Materials and Reuse Scope

Object records and images are reused from TimeTravel [1], deposited by the Mohamed Bin Zayed University of Artificial Intelligence at MBZUAI/TimeTravel. We use the records as source constraints only. We do not claim authorship of the dataset, the original captions, or the benchmark protocol. Underlying artefacts remain the property of their holding museums; no reuse terms are attached to the records or the dataset release, so images are used only for local coding and are not redistributed. Our contribution is limited to the narrative-licence protocol, the multivocal outputs, and the human evaluation.

The analysis uses the downloaded Parquet release as observed on 5 September 2026. A SHA-256 checksum was recorded for the source file (97876104c6d3bcb9ada24703c54cfe19ed 57e98da73faa34ee67fae3987ceb9c). Local inspection is separated from publication rights: the dataset page did not expose an explicit licence field during the audit. Images may be used for local coding, but no TimeTravel will be redistributed in the article or supplementary package because no dataset-level or image-level terms are documented.

3.3. Source Audit and Working Sample

The release contains 10,250 rows and 13 columns, including the image. IDs were unique, and no exact duplicate metadata rows were found. The physical Parquet type for id was integer, although the dataset card declared a string. The audit identified 1526 records with non-placeholder quoted inscriptions (records whose Inscription field contains at least one quoted text; 4213 records have a non-empty Inscription field, most recording only type, position, language or script), 6991 with recorded findspots, and 302 with all ten licence-bearing metadata fields populated. Completeness was an indicator, not an exclusion criterion.

Sampling used deterministic SHA-256 ranking with seed ‘narrative-licence-2026-09-05-v1’ and mutually exclusive allocation in preregistered order. Sixteen inscription-rich records were selected first; 16 remaining findspot records were then selected with available forgery, plating, and hoard markers seeded before balanced fill; and 16 remaining non-Roman records were balanced across canonical place groups. The value ‘Inida’ occurred in 114 source rows and was normalised to India for grouping only; exported records retain the source value. The third stratum is non-Roman rather than non-coin because the release has no independent object-type field. The sample is a working analytical sample, not a

taxonomy of the full release. Figure 2 places corpus-wide evidence availability beside the feature profile of the three sampling strata.

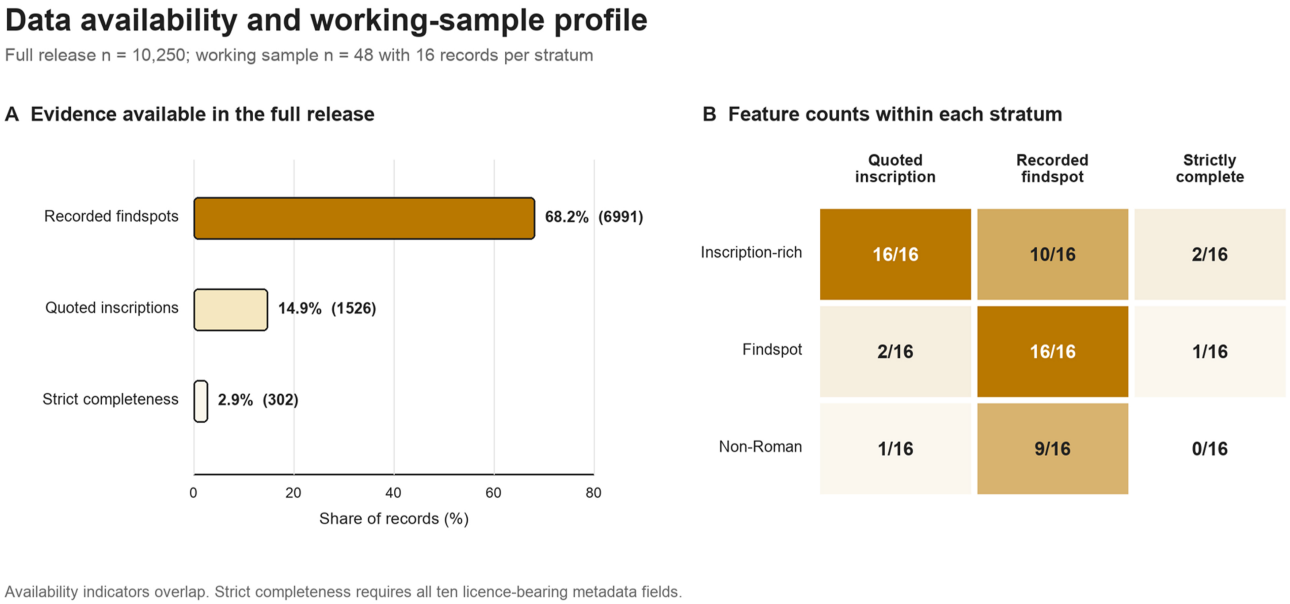


Figure 2. Dataset availability and working-sample profile. (**A**) Availability among all 10,250 records; categories overlap, and strict completeness requires all ten licence-bearing fields. (**B**) Feature counts within each mutually exclusive 16-record sampling stratum. Source: local audit of the MBZUAI/Time-Travel Parquet release observed 5 September 2026.

The sample size follows from the paired design rather than from a population-estimation target. Because the primary outcome is a within-record difference between Conditions C and B, the record is the unit of analysis. Under a paired t-test approximation, 48 paired records give 80% power at a two-sided $\alpha = 0.05$ for a standardised paired difference of approximately $d_z = 0.41$; within a 16-record stratum the corresponding value is approximately $d_z = 0.75$. These values indicate scale only: the confirmatory analysis uses a paired bootstrap, and because Condition C rates are expected to lie near zero, paired differences may be skewed. This is a sensitivity analysis, not a power calculation, because no prior estimate of the effect exists; the confirmatory estimate will be reported with its bootstrap interval whatever its size. Only the pooled paired estimate is confirmatory; stratum-level results, which allow each heritage-error type to be examined where it can occur, are descriptive. The sample also bounds the coding load: every sentence of 96 outputs (48 records × 2 conditions) is double coded and adjudicated. The sample supports a controlled test of the protocol, not generalisation to museum collections at large (Section Transferability).

3.4. Frozen Narrative-Licence Schema

Schema version 1.0 was frozen on 5 September 2026 before confirmatory generation. The coding unit is one independently interpretable proposition (rule 1, Section 3.4.1). Mixed-status sentences are split; if splitting changes their meaning, the least permissive status applies. Voice 0 is never a source of licence. Table 1 gives the operational map. Version 1.1 redefines the coding unit as one proposition (Section 3.4.1) and adds the decision procedure, worked examples, coherence rules, secondary annotation dimensions and outcomes listed in Section 3.9, without changing the field-to-licence map.

Table 1. Frozen field-to-licence map (version 1.0).

Source	Licensed Use	Boundary
Inscription quoted	Exact quotation and literal legend reading	No translation, completion, motive, sentiment, or historical expansion
Find spot	Recorded find location or discovery context	Empty field becomes 'Find not recorded'; no invented landscape or journey
Production date	Recorded date or range	No unrecorded event, reign, chronology, or causal claim
Materials	Named material or plating	No value, prestige, skill, authenticity, or sensory claim
Technique	Named technique	No workshop practice or maker intention unless recorded
Subjects	Subjects exactly as named	No added identity, action, symbolism, or emotion
Assoc name	Association as named, with qualifiers	Preserve 'possibly', 'named in inscription', and similar qualifiers
Culture	Culture label exactly as recorded	No generalisation to a people, civilisation, identity, or belief
Section	Collection/category label	Not an object-level historical claim
Place	Place label exactly as recorded	Not a substitute for production place, findspot, territory, or ownership
Image	Visible object or pictorial type	INFERRED and explicitly marked; no off-frame reconstruction or intention
description	Voice 0 baseline only	Never a warrant, ground truth, or reusable project prose

3.4.1. Decision Procedure

Coders apply rules 1 and 2 to every unit, then rules 3 to 7 in order, stopping at the first rule that decides the status. The full procedure, with operational definitions, is given in Supplementary File S1.

(1) **Segment.** One coding unit holds one proposition. Coordinated clauses, relative clauses and appositives are split. If splitting changes the meaning, the whole sentence is coded at its least permissive status.

(2) **Map each content word.** Every noun, verb, adjective and qualifier must map to (a) a value in a named, non-empty field, (b) that field's category noun or verb (for example 'material', 'findspot', 'dates', 'made', 'attributed'), or (c) a visible element of the image. Function words, the meta-nouns object, record, image, legend and inscription, and a closed list of glue verbs (is, reads, names, shows, records, gives, lists, states, marks) need no warrant. A field holding a placeholder string such as 'None' counts as empty.

(3) **Absence.** A unit that reports an empty field in a fixed absence note (for example, 'How the object was made is not recorded') is ATTESTED by the empty field; it speaks about the record, not the object.

(4) **Inscriptions.** Expanding abbreviations, translating, or attributing meaning or purpose to a legend is NOT LICENSED, even where the expansion is standard in the relevant discipline. A literal reading, which reproduces the quoted characters with their recorded position, type and language, proceeds to rule 5.

(5) ATTESTED applies if every content word maps to (a) or (b), every recorded qualifier (for example 'about', 'probably', 'possibly', 'named in inscription') is kept verbatim, and no slash-joined value is reduced to one of its parts. A slash-joined value such as 'Found/Acquired' or 'Excavated/Findspot' records alternatives and must be kept whole; an ampersand-joined value such as 'Named in inscription & portrayed' records both facts and may be split, each part keeping its wording.

(6) INFERRED applies if at least one content word maps only to (c) and the unit (i) reports visible object or pictorial type, shape, colour, count, arrangement, posture or a visible motif, (ii) names no identity, date, function or meaning that the fields do not supply, and (iii) carries an explicit marker such as 'The image shows'. A visual unit without

the marker is coded INFERRED-unmarked; it is withheld at the gate and, if found in released text, counts in the numerator of the released unlicensed rate. Objects visible in the frame that are not the recorded object are never described.

(7) NOT LICENSED applies otherwise, including purpose, symbolism, significance, causation, emotion, audience response, unrecorded events, and any content whose only source is Voice 0.

(8) **Release.** ATTESTED and marked INFERRED units pass the release gate; all other units are entered on the withheld list with the field or source that would have been needed to license them. Public prose uses only released units but need not use all of them.

(9) **Replacement.** A withheld unit may be replaced by a new candidate unit that restates only licensed content. In Condition C, the model proposes replacements; coders classify them under rules 1 to 8, and each replacement is logged with a link to the unit it replaces.

3.4.2. Worked Example: Record 534

Record 534, the format-pilot record named in Section 3.5, is used to trace the procedure from source fields to released text. Its candidate sentences were written by the authors to exercise each rule; they are not model output, and the pilot remains excluded from analysis. The source packet is: Production date '324–325 (about)'; Find spot 'Excavated/Findspot: Warsop (Warsop hoard)'; Materials 'copper alloy'; Technique 'None' (empty under rule 2); Inscription: obverse, Latin, quoted 'FL IVL CONSTANTIVS NOB C'; reverse, Latin, quoted 'PROVIDENTIAE CAESS'; mintmark, reverse field and exergue, quoted 'PLON'; Subjects 'emperor/empress; fort/fortification'; Assoc name 'Named in inscription & portrayed: Constantius II'; Culture 'Roman Imperial'; Place 'roman empire'. The record has no object-type field, so the word 'coin' can be licensed only as a marked inference from the image (rule 6); Voice 0 licenses nothing (rule 7).

The image, checked by the authors, shows two round faces side by side. On the left is a draped bust in profile facing left, the head wearing a wreath, with legend letters around the edge. On the right is a rectangular structure of stacked courses with a central doorway, two turrets above and a star between them, with legend letters around the edge. Table 2 gives the ledger.

Table 2. Worked ledger for record 534. Candidate sentences were written by the authors to exercise the decision procedure. Letters mark units produced by splitting (rule 1) or replacement (rule 9).

#	Candidate Sentence	Unit	Warrant	Status	Released
1	This copper-alloy object dates to about 324–325.	As candidate	Materials; Production date, qualifier kept	ATTESTED	Yes
2a	It is a Roman coin.	It is Roman.	Culture label shortened; field reads 'Roman Imperial'	NOT LICENSED	No; replaced by 2b
2b		The record gives its culture as Roman Imperial.	Culture, exact label (rule 9)	ATTESTED	Yes
2c		It is a coin.	Visible object type, unmarked	INFERRED-unmarked	No; replaced by 2d
2d		The image shows a round coin with two faces.	Visible object type and count, marked (rule 9)	INFERRED	Yes
3a	The image shows a draped bust facing left on the obverse; the record states that Constantius II is portrayed.	The image shows a draped bust facing left on the obverse.	Visible posture, marked; 'obverse' from Inscription position	INFERRED	Yes
3b		The record states that Constantius II is portrayed.	Assoc name 'portrayed' (ampersand part, rule 5)	ATTESTED	Yes

Table 2. *Cont.*

#	Candidate Sentence	Unit	Warrant	Status	Released
4a	The obverse legend reads FL IVL CONSTANTIVS NOB C, naming Constantius II as Caesar.	The obverse legend reads FL IVL CONSTANTIVS NOB C.	Inscription, obverse, quoted	ATTESTED	Yes
4b		The record states that Constantius II is named in the inscription.	Assoc name 'Named in inscription' (ampersand part, rule 5)	ATTESTED	Yes
4c		… as Caesar.	Expansion of 'NOB C' (rule 4)	NOT LICENSED	No
5a	The reverse legend PROVIDENTIAE CAESS celebrates the foresight of the Caesars.	The reverse legend reads PROVIDENTIAE CAESS.	Inscription, reverse, quoted	ATTESTED	Yes
5b		… celebrates the foresight of the Caesars.	Translation and purpose (rule 4)	NOT LICENSED	No
6a	The mintmark PLON shows it was struck in London.	A mintmark reads PLON.	Inscription type 'mintmark', quoted	ATTESTED	Yes
6b		… shows it was struck in London.	Expansion; Place gives only 'roman empire' (rule 4)	NOT LICENSED	No
7a	The reverse shows a fortified gate with two turrets and a star.	The image shows a structure with a central doorway, two turrets and a star above.	Visible arrangement, marked	INFERRED	Yes
7b		… a fortified gate.	Identification; Subjects gives 'fortification' only as a subject label	NOT LICENSED	No; replaced by 7c
7c		The record lists fortification among its subjects.	Subjects (rule 9)	ATTESTED	Yes
8	The design symbolised security and imperial protection.	As candidate	Symbolism; wording traces to Voice 0	NOT LICENSED	No
9a	It was excavated in the Warsop hoard.	As candidate	Field label 'Excavated/Findspot' reduced to 'excavated'	NOT LICENSED	No; replaced by 9b
9b		The record gives its findspot as 'Excavated/Findspot: Warsop (Warsop hoard)'.	Find spot, verbatim (rule 9)	ATTESTED	Yes
10	Its owner buried it for safekeeping in troubled times.	As candidate	Unrecorded event and motive	NOT LICENSED	No
11	How the object was made is not recorded.	As candidate	Technique empty (rule 3)	ATTESTED	Yes
12	It offers valuable insights into hoarding practices.	As candidate	Significance claim; Voice 0 wording	NOT LICENSED	No

The released gallery stop, assembled only from released units (113 words), reads: 'This copper-alloy object dates to about 324–325. The record gives its culture as Roman Imperial. The image shows a round coin with two faces. The image shows a draped bust facing left on the obverse; the record states that Constantius II is portrayed. The obverse legend reads FL IVL CONSTANTIVS NOB C. The record states that Constantius II is named in the inscription. The reverse legend reads PROVIDENTIAE CAESS. A mintmark reads PLON. The image shows a structure with a central doorway, two turrets and a star above. The record lists fortification among its subjects. The record gives its findspot as "Excavated/Findspot: Warsop (Warsop hoard)". How the object was made is not recorded.'

Units 2a, 2c, 4c, 5b, 6b, 7b, 8, 9a, 10 and 12 enter the withheld list, each with the field or source that would be needed to license it (for 6b, a mint attribution field; for 8 and 12, a cited interpretive source); 2a, 2c, 7b and 9a are replaced under rule 9. The example covers

the mixed cases raised in review: a documented fact combined with an expansion (units 4 and 6), with a translation (unit 5) and with a visual identification (units 2, 3 and 7), as well as a shortened culture label (unit 2) and a reduced field label (unit 9). It also illustrates rule 7: the two units whose wording traces to Voice 0 (units 8 and 12) are withheld. A second worked record from a non-coin, non-Roman section is given in Supplementary File S1.

3.5. Conditions and Output Controls

Condition A is the TimeTravel description unchanged. Condition B uses a matched museum-material prompt with the same three voices, gallery-stop length, audio format, storyboard limit, Easy Read text, and image description as Condition C, but omits the licence rule, ledger, and refusal mechanism. Condition C requires the ledger before public prose, retains only ATTESTED and explicitly marked INFERRED statements, and records candidate NOT LICENSED claims in a withheld list.

For both B and C, the model provider, exact model identifier, access date, system prompt, temperature, top-p, seed where supported, maximum tokens, retry policy, and failure handling will be frozen before the first confirmatory record. The same values will be used for each matched pair. The single existing Condition C output for record 534 is a format pilot; it will not enter the analysis. Record 534 will be regenerated in both conditions after the controls are locked, and only those confirmatory runs will be coded.

3.6. Multivocal Output

The public-facing form contains three bounded voices. Voice A reports the quoted inscription or its official absence. Voice B reports provenance and making information, with plating, forgery, and hoard information obligatory when recorded. Voice C is marked contemporary reception: it may ask questions of present-day viewers but may not supply invented historical or visitor answers. The same licensed content is then adapted into a 90–150-word gallery stop, a one-speaker audio script, a maximum-six-card still-image storyboard, Easy Read text, and an image description. Accessible design is treated as an evidential transformation, not a permission to add context; this extends established attention to audio description in museums [11]. The three voices are institutional: they differ by the evidence class that licenses each sentence, not by the standpoint of the speaker (Section 2.3).

Four design choices are intended to protect narrative coherence. First, in Condition C, classification and withholding precede composition: public prose is assembled only from released units, so no finished paragraph has sentences removed from it. Second, connectives that assert a relation (because, therefore, after, as a result) are treated as propositions that require a warrant, whereas neutral connectives (and, also) are not; coherence cannot therefore be restored by adding unlicensed causal or temporal links. Third, where a withheld claim would have filled a gap, the fixed absence note fills it, so that the text remains continuous and the silence is stated. Fourth, if a coder flags a coherence break in a Condition C output, the system may regenerate that output once from the same ledger without adding units; repairs are logged, and a repair that adds a unit is coded as a violation. When a record licenses fewer words than the 90–150-word gallery-stop target, the stop is released short rather than padded.

3.7. Sentence Coding and Agreement

Two annotators will independently segment and code every sentence in Conditions B and C against the frozen schema. Each code records the source field or visible element, status, release location, and any heritage-relevant error. Training uses records outside the 48-record analysis set. The worked examples in Section 3.4.2 and Supplementary File S1 are an exception that coders will have seen; records used as worked examples are therefore

coded last, and the primary estimate is reported with and without them. Independent labels are locked before discussion. A third pass resolves disagreements for the analysis dataset while preserving both original labels.

Agreement reporting includes raw percentage agreement, Cohen's kappa, a three-class confusion matrix, and class-specific agreement, because a dominant ATTESTED class can make a single summary statistic misleading. Condition A is coded only for invention, slogan laundering, findspot erasure, forgery overwrite, and culture-label collapse; it is not recoded into the project's voices.

In addition to the release status, coders record three secondary dimensions for each unit, which do not change what is released: uncertainty (whether a qualifier recorded in the field, such as 'about', 'probably' or 'possibly', is kept, dropped or added); transformation (whether a field value is preserved on the surface but given a historical, cultural or narrative function that the record does not state, for example a findspot rendered as a journey, or moved to another category); and interpretive dependency (whether the unit's meaning depends on another unit or on an external source). Agreement is reported separately for each dimension. The dimensions test whether the three release statuses are expressive enough for complex records: if units that pass the release gate frequently carry transformations or dependencies, or if coders disagree mainly on these dimensions, the status set is too coarse.

3.8. Outcomes and Statistical Analysis

The primary outcome is released_unlicensed_rate: NOT LICENSED or INFERRED-unmarked units that survive into public output divided by all released public-output units. Secondary unit outcomes are draft_licensed_unit_rate = (ATTESTED + marked INFERRED)/all candidate public-prose units before withholding; draft_invention_rate = (NOT LICENSED + INFERRED-unmarked)/all candidate public-prose units before withholding; and withholding_capture_rate = NOT LICENSED and INFERRED-unmarked units withheld/all such candidate units.

Rates are calculated first within record to avoid treating sentences from the same object as independent observations. The primary estimate is the paired C-minus-B difference in record-level released_unlicensed_rate, reported with medians, interquartile ranges, and a paired bootstrap 95% confidence interval. Binary record-output errors—slogan laundering, findspot erasure, forgery overwrite, and culture-label collapse—are reported as matched risk differences with exact McNemar tests where cell counts permit. Effect estimates and uncertainty remain primary; *p*-values are secondary. No BLEU, METEOR, ROUGE, SPICE, BERTScore, or LLM-as-judge score is a primary outcome.

All rates are computed over coding units as defined in rule 1 (Section 3.4.1); a sentence containing several units contributes each unit separately. Two further secondary outcomes address narrative quality. Coherence breaks per output (dangling references, unsupported topic jumps and orphaned connectives, defined operationally in Supplementary File S1) are coded by the same annotators, blind to condition where the output format allows. Readability and retention are reported descriptively as word count, Flesch reading-ease score and the proportion of candidate units released. If Condition C outputs are measurably less coherent or less readable than Condition B outputs, this will be reported as the cost of the licence. Interpretive quality as experienced by curators and visitors is a different outcome that requires participant-based evaluation and is outside the present study.

3.9. Reproducibility, Governance, and Human Oversight

The frozen schema, sample identifiers, source checksum, matched prompts, raw outputs, independent codes, adjudication decisions, and analysis code will be versioned. Any post-generation schema clarification will be issued as a dated protocol amendment rather

than silently incorporated. The pipeline is designed so that no sentence is released without an inspectable warrant and so that a curator or researcher can review the full ledger [12]; whether it achieves this is part of the planned evaluation.

Protocol amendment (schema version 1.1, 10 October 2026). In response to editorial and peer review, and before any confirmatory generation, version 1.0 was amended as follows: the coding unit was redefined as one proposition, and the ordered decision procedure, including a replacement rule, was added (Section 3.4.1); worked examples were added (Section 3.4.2 and Supplementary File S1); the conditions under which a community-authored source class could be added were specified (Section 2.3); the pooled paired estimate was specified as the only confirmatory estimate, with stratum-level results descriptive (Section 3.3); coherence-protecting composition rules were specified (Section 3.6); three secondary annotation dimensions were added, and records used as worked examples are coded last with the primary estimate reported with and without them (Section 3.7); and the outcome definitions were restated over coding units, with INFERRED-unmarked units counted with NOT LICENSED units, and coherence breaks, readability and retention were added as secondary outcomes (Section 3.8). The field-to-licence map (Table 1), the sample, the conditions and the primary outcome measure are otherwise unchanged.

4. Results of the Completed Audit

4.1. Completed Data Audit

The completed pre-generation audit established an analysis frame of 10,250 unique records. Quoted-inscription coverage was 14.9% (1526/10,250), findspot coverage was 68.2% (6991/10,250), and strict completeness across the ten licence-bearing metadata fields was 2.9% (302/10,250). These values describe source availability, not artefact quality or cultural coverage. A re-audit during revision reproduced these values from the same release and added two structural observations. Coin types are named in 1888 of the 10,250 descriptions (18.4%), but 972 of the 1526 quoted inscriptions (63.7%) come from coin records. Findspot coverage ranges from 6.9% (72/1049) in the Roman section to 100% in the Ancient Egyptian, Tang dynasty, Maya and Mohenjo-daro sections (Section Transferability).

4.2. Working-Sample Composition

The 48-record working sample met the three mutually exclusive quotas. The inscription-rich stratum contains 16/16 records with quoted inscriptions and 10/16 with findspots. The findspot stratum contains 16/16 records with findspots, including two forgery-or-plating cases and five hoard cases. The non-Roman stratum covers 16 recorded culture labels across nine canonical place groups. Table 3 reports the exact sample counts.

Table 3. Composition of the 48-record working sample.

Stratum	n	Places	Cultures	Inscription	Findspot	Complete
Inscription-rich	16	8	11	16	10	2
Findspot	16	10	12	2	16	1
Non-Roman	16	9	16	1	9	0
Total	48	—	—	19	35	3

4.3. Status of the Planned Confirmatory Evaluation

Confirmatory B/C generation, independent coding, and agreement estimation have not begun. Consequently, the present study does not yet report a licence-rate difference, error-rate difference, or wall-readiness claim. This boundary is intentional: the matched settings and retry policy must be frozen before output inspection, and the pilot for record

534 is excluded from analysis. The worked examples in Section 3.4.2 and Supplementary File S1 illustrate the procedure; their candidate sentences were written by the authors and are not outcome data.

5. Discussion

Narrative Licence treats authenticity as partly a capacity to withhold. In this account, an institution demonstrates epistemic responsibility not only by saying correct things, but by showing why a sentence may be said and by refusing rhetorically attractive additions that exceed the record. The ledger is therefore not an explanation added after generation. It is designed as a release gate intended to change the relationship between source, draft, and public text.

The protocol also reframes Voice 0. A polished dataset description can occupy the interpretive field as if it were the object's natural voice. Retaining it as an unchanged baseline while denying it licensing authority is intended to make its role analytically visible. The comparison of interest is not whether Condition C resembles Voice 0, but whether matched public outputs expose and respect the limits of the record more reliably than Condition B.

Coins and inscribed objects also complicate the common idea that an AI system adds narrative to otherwise mute things. Many objects are already historical media: legends name rulers, slogans assert values, images repeat types, and materials record production choices. The protocol attempts to preserve that mediated character. It is designed to allow quotation without turning propaganda into public feeling, and to allow visible-type inference without reconstructing lost environments or intention.

A null or modest B/C difference would remain informative. It could indicate that the matched control prompt already provides strong grounding, that the licence schema is too permissive, or that coders cannot apply the categories reliably. Conversely, a lower released unlicensed rate in C would not by itself establish public value. That claim requires participant-based evaluation of whether bounded multivocal outputs are usable, trustworthy, and memorable for heritage practitioners and visitors.

Sentence-level control raises an obvious objection: museum labels connect facts and supply background, and a text assembled from individually warranted units may read as a list. The protocol addresses this by composing from the ledger rather than deleting from a draft, by treating relational connectives as claims, and by stating absences in place of the claims they replace (Section 3.6). Whether these choices preserve enough continuity is an empirical question; the coherence and readability outcomes (Section 3.8) are designed to detect a licence that improves warrant only by degrading the text, which would not be adoptable. Background of the kind museums routinely supply is not excluded in principle; it requires its own cited source, which the ledger would then record as a separately licensed field.

The three release statuses are deliberately coarse because they govern a binary decision: release or withhold. Complex records put pressure on that coarseness. A field can be kept verbatim while its function changes, as when a findspot becomes a narrative of journey, and a qualifier such as 'probably' can be silently dropped. The secondary annotation dimensions (Section 3.7) record these cases without multiplying release categories, and their agreement statistics will indicate whether a finer scheme is needed. A further step would formalise the licence in a controlled semantic model, for example by aligning the field-to-licence map with the CIDOC Conceptual Reference Model [13] and expressing the record-to-sentence derivation as provenance in W3C PROV-O [14], so that ledgers could be exchanged with museum information systems.

Deployment Considerations

The present study evaluates static outputs. Conversational agents and live tours would change three parameters. First, the unit of release: in dialogue each turn is released separately, so the ledger must be built per turn and must carry forward what has already been released; otherwise a licensed answer can presuppose an unlicensed earlier one. Second, refusal wording: a withheld claim in a gallery text simply does not appear, whereas in dialogue a visitor may ask for it directly, so the system needs a refusal form that states what the record does not say and, where possible, refers the visitor to a curator or a cited source rather than answering. Third, latency and review: classifying every candidate sentence adds computation before release, and curator review of withheld lists cannot happen in real time, so a deployment would need precomputed ledgers for frequently requested objects and asynchronous review of logged refusals. Each adaptation creates new outcomes, including response time, refusal frequency, visitor response and staff workload, that require a live system and participant-based evaluation. They are outside the evidentiary scope of the present study and are identified as the next stage of work.

6. Limitations

First, n = 48 supports a controlled protocol test, not population-level claims about cultural-heritage records. The public corpus has uneven metadata, and its quoted-inscription evidence is dominated by coin legends (Section Transferability); only 2.9% of rows are complete under the strict definition used here. Second, the released columns do not independently identify object type, so the third stratum is non-Roman rather than non-coin. Third, data-card and physical ID types differ, and one place-label typo requires grouping-only normalisation.

Fourth, a frozen prompt cannot eliminate model drift, nondeterminism, or provider-side changes; exact access dates and repeated-run rules are therefore necessary. Fifth, sentence categories simplify propositions whose warrants may be layered or contested. Agreement statistics and preserved disagreement cases are essential. Sixth, Voice C remains an institutional construction; it does not substitute for community-authored interpretation. Section 2.3 specifies how a separately licensed class for community-authored statements would be added; the present schema contains none and generates no community voice. Finally, unclear image-reuse terms constrain publication of object stills. The present study therefore reports metadata-based analysis and will not redistribute images without documented permission.

Transferability

Table 4 shows that evidence in the release is unevenly distributed across its twelve section labels. Coin types are named in about 18% of descriptions, but the quoted-inscription pool is dominated by coin legends: 972 of the 1526 quoted inscriptions come from coin records, 968 of them from the Roman section. Results in the inscription-rich stratum will therefore mainly describe coin legends and should not be generalised to inscribed objects in other media, such as stelae, manuscripts or ceramics. Findspot coverage likewise depends on the section, from 6.9% of Roman records to 100% of Ancient Egyptian, Tang dynasty, Maya and Mohenjo-daro records. A protocol tested on one collection inherits that collection's cataloguing habits; this variation is one reason why stratification by evidence type is appropriate.

Table 4. Evidence availability by section label in the TimeTravel release (n = 10,250). Coin-type descriptions name a coin or coin denomination; the count is approximate because it is based on keyword matching of the descriptions, as the release has no object-type field. Source: re-audit of the Parquet release (SHA-256 97876104…ceb9c), 10 October 2026.

Section Label	Records	Coin-Type Descriptions	Quoted Inscriptions	Findspot Recorded
Greek	1869	1	141	1598
Ancient Egyptian	1056	0	4	1056
Roman	1049	1013	968	72
Tang dynasty	1039	0	1	1039
Viking	1013	66	11	992
Mesopotamian	1000	0	0	942
Persian	999	57	184	214
Japanese	869	3	216	405
Gupta dynasty	737	733	0	63
Maya	488	0	0	488
Mohenjo-daro	114	0	0	114
Mauryan	17	15	1	8
Total	**10,250**	**1888**	**1526**	**6991**

What transfers is the decision procedure and the field-type logic. The schema keys on field types that most collection documentation shares, such as production date, findspot, materials, technique, inscription, associated names and culture, so a collection with a different structure requires a new field-to-licence map rather than a new procedure. Adaptation is required for collections that transcribe rather than quote inscriptions, that hold provenance in free text, or that provide an object-type field, which TimeTravel lacks.

7. Conclusions

Narrative Licence specifies a testable boundary between fluent museum prose and institutionally licensed speech. The schema, sample, matched B/C design, sentence-level coding, and heritage-error analysis are fixed in schema version 1.1 before confirmatory generation. The protocol does not assume that constraint will improve every output; it creates the conditions under which that claim can be tested without confusing format, model choice, or Voice 0 with the intervention. Confirmatory inference will be reported only after matched outputs and independent codes are locked. Claims about curator judgement, visitor recall, or accessible-trail use remain outside the evidentiary scope of the present study and require participant-based evaluation.

No effectiveness result is reported here. The working sample of 48 records supports a controlled test of the protocol, not generalisation to museum collections or digital heritage datasets at large. Three extensions follow from the present design: alignment of the licence schema with controlled semantic models for interoperability [13,14]; a participatory pilot in which community-authored statements enter as a separately licensed source class under their own ethics approval; and evaluation of conversational and live-tour deployments (Section Deployment Considerations).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/heritage9110447/s1, File S1: Narrative Licence coder manual (schema version 1.1), including the decision procedure, secondary annotation dimensions, coherence-break definitions and two worked records.

Author Contributions: Conceptualization, methodology, software, formal analysis, data curation, visualisation, and writing—original draft preparation, M.I.Z.A.; validation and writing—review and editing, A.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The present study reports analysis of public dataset records and does not involve human participants.

Informed Consent Statement: Not applicable.

Data Availability Statement: Object records and images are reused from TimeTravel [1], deposited by the Mohamed Bin Zayed University of Artificial Intelligence at MBZUAI/TimeTravel. We use the records as source constraints only. We do not claim authorship of the dataset, the original captions, or the benchmark protocol. Underlying artefacts remain the property of their holding museums; no reuse terms are attached to the records or the dataset release, so images are used only for local coding and are not redistributed. Our contribution is limited to the narrative-licence protocol, the multivocal outputs, and the human evaluation. The sampled identifiers, derived text outputs, sentence codes, and analysis code will be shared with the completed article. Museum images will not be redistributed. The deposit page states no licence or terms of use (checked on 10 October 2026). The shared materials will contain record identifiers, field-level codes, model outputs and author-written text, but no images. The analysed file has SHA-256 97876104c6d3bcb9ada24703c54cfe19ed57e98da73faa34ee67fae3987ceb9c.

Acknowledgments: The authors acknowledge first the TimeTravel depositors—Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer—and the Mohamed Bin Zayed University of Artificial Intelligence for the MBZUAI/TimeTravel deposit. During preparation of this protocol-stage manuscript, the authors used the OpenAI 5.5 Codex desktop application (accessed 5 September 2026) to assist with data-audit scripting, document formatting, and manuscript drafting. The authors are responsible for verifying the source data, citations, analysis decisions, and final text. During revision, the authors used Claude Fable 5.1 (Anthropic) to assist with re-auditing the release, drafting revised text and preparing Supplementary File S1; the authors are responsible for verifying all figures against the data and for the final text.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ghaboura, S.; More, K.P.; Thawkar, R.; Al Ghallabi, W.; Thawakar, O.; Khan, F.S.; Cholakkal, H.; Khan, S.; Anwer, R.M. Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Vienna, Austria, 2025; pp. 23627–23641. [CrossRef]
2. Mohamed Bin Zayed University of Artificial Intelligence. TimeTravel Dataset Deposit. Available online: https://huggingface.co/datasets/MBZUAI/TimeTravel (accessed on 5 September 2026).
3. International Council of Museums. ICOM Approves a New Museum Definition. 24 August 2022. Available online: https://icom.museum/en/news/icom-approves-a-new-museum-definition/ (accessed on 5 September 2026).
4. The London Charter. The London Charter for the Computer-Based Visualisation of Cultural Heritage, Version 2.1. 2009. Available online: https://londoncharter.org/principles.html (accessed on 5 September 2026).
5. Bouras, V.; Spiliotopoulos, D.; Margaris, D.; Vassilakis, C.; Kotis, K.; Antoniou, A.; Lepouras, G.; Wallace, M.; Poulopoulos, V. Chatbots for Cultural Venues: A Topic-Based Approach. *Algorithms* **2023**, *16*, 339. [CrossRef]
6. Kist, C.; Moshfeghi, Y. To Bot or Not? Conversational Agents and Authentic Heritage Experiences. In *CHI EA '26: Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*; Article 674; ACM: New York, NY, USA, 2026; pp. 1–5. [CrossRef]
7. Dibitonto, M.; Ferrato, A.; Limongelli, C.; Patroni, O.C. Museum Audio Guides Generation Using Visitor Categories and Large Language Models. *Multimed. Syst.* **2026**, *32*, 439. [CrossRef]
8. Economou, M.; Young, H.; Sosnowska, E. Digital Storytelling for Emotional Engagement in Museums: Design and Evaluation of the Hunterian Antonine Wall EMOTIVE Experiences. *Int. J. Herit. Stud.* **2025**, *32*, 324–355. [CrossRef]
9. Sartini, B.; Shoilee, S.B.A.; Libbi, C.A.; de Boer, V. Multivocal Exhibition: Exploring Cultural Perspectives through User-Curated Art Exhibitions. *ACM J. Comput. Cult. Herit.* **2024**, *17*, 73. [CrossRef]
10. Assael, Y.; Sommerschield, T.; Shillingford, B.; Bordbar, M.; Pavlopoulos, J.; Chatzipanagiotou, M.; Androutsopoulos, I.; Prag, J.; de Freitas, N. Restoring and Attributing Ancient Texts Using Deep Neural Networks. *Nature* **2022**, *603*, 280–283. [CrossRef] [PubMed]
11. Hutchinson, R.S.; Eardley, A.F. The Accessible Museum: Towards an Understanding of International Audio Description Practices in Museums. *J. Vis. Impair. Blind.* **2020**, *114*, 475–487. [CrossRef]

12. Shneiderman, B. *Human-Centered AI*; Oxford University Press: Oxford, UK, 2022.
13. Doerr, M. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Mag*. **2003**, *24*, 75–92. [CrossRef]
14. PROV-O: The PROV Ontology. W3C Recommendation. 30 April 2013. Available online: https://www.w3.org/TR/2013/REC-prov-o-20130430/ (accessed on 10 October 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.