When More Metadata Misleads: Probing Field-Level Grounding in LLM-Generated Museum Descriptions

Clemencia Siro and Laura Hollink
Centrum Wiskunde & Informatica (CWI)
Amsterdam, The Netherlands
{c.n.siro,l.hollink}@cwi.nl

Abstract

Metadata-conditioned generation is often framed as a data-to-text task in which additional input fields should improve grounding. Museum description generation challenges this assumption because catalogue records are heterogeneous, partially visual, and unevenly documented across geographic regions. We study how metadata composition shapes the generation of artefact descriptions through a two-phase framework. Phase 1 analyses natural metadata variation in a large museum collection, where field count, field type, image availability, and *Place* co-vary. Phase 2 introduces controlled field probes that hold object identity fixed while varying metadata quantity, field type, and image access. Our results show that metadata quantity alone does not explain generation behaviour. Field type determines whether models ground, omit, or infer information, while images can both support visual grounding and increase unsupported interpretation. More broadly, our findings show that metadata-conditioned data-to-text evaluation should assess field- and claim-level support, rather than relying only on aggregate input coverage or reference similarity.

1 Introduction

Large language models (LLMs) are invreasingly used in cultural heritage settings to generate public-facing object descriptions from structured catalogue records (Bu et al., 2025; Trichopoulos, 2023). This is a metadata-conditioned data-to-text task: it requires fluent prose that stays faithful to a bounded, heterogeneous set of catalogue fields. Data-to-text benchmarks such as E2E, WebNLG, RotoWire, and ToTTo have driven progress on content selection, fluency, and faithfulness (Novikova et al., 2017; Gardent et al., 2017; Wiseman et al., 2017; Parikh et al., 2020), with evaluation focused on reference similarity, input coverage, and hallucination (Dušek and Kasner, 2020; Pagnoni et al., 2021). However, this framing leaves unexplored how the composition of the input itself shapes generation.

Cultural heritage collections make this question especially important. Museum records are structured, but they are not uniform: catalogue fields differ in their epistemic status. Artefacts may encode provenance facts, physical descriptors, production dates, curatorial classifications, subject annotations, inscriptional content, and named associations within a single attribute set. These fields differ in the extent to which a model can be expected to recover missing values from domain priors. Multimodal models exacerbate this effect: an image supports description of form, colour, appearance, and visible markings (Liu et al., 2023), but cannot ground production place, excavation context, date, or associated historical figure. A model that conflates these grounding sources generates fluent prose that is, on the surface, indistinguishable from faithful description (Maynez et al., 2020; Ji et al., 2023). Field coverage is, in addition, systematically non-uniform across collections and cultural-geographic regions, reflecting heterogeneous acquisition histories and cataloguing conventions (Srinivasan, 2017; van Hooland and Verborgh, 2015). As a result, metadata is not only an input representation for generation, but also a trace of prior documentation practice.

Prior work does not isolate these dimensions. Cultural-heritage evaluation often aggregates quality across examples or fields, and reports regional gaps observationally (Ghaboura et al., 2025), making it difficult to determine whether errors arise from metadata quantity, field identity, image availability, or domain-specific priors. Additionally, cultural heritage research has largely focused on improving access to collections,

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49666–49688
October 24-29, 2026 ©2026 Association for Computational Linguistics

producing engaging descriptions, or linking heterogeneous museum records (Hyvönen, 2012; Candela et al., 2023), but less often examines how different metadata configurations change what models include, omit, or infer.

This paper asks: *how do metadata quantity, field type, image availability, and place-specific documentation patterns affect what LLMs include, omit, and infer in museum description generation?* We address this question with a two-phase study on the TimeTravel benchmark (Ghaboura et al., 2025). Phase 1 provides a corpus-scale observational analysis over 10,250 historical and cultural artefacts generated under each record's full available metadata. It shows that reference-based quality varies with metadata composition, but that naturalistic records confound field count, output length, and field composition: output length increases with field count ($\eta^2 = 0.33$), while longer generations receive lower BERTScore across civilisations ($r = -0.34$ to -0.70). Phase 2 introduces a *controlled field-type probing framework* that independently varies field count, field type, and image availability within the same objects across ten experimental conditions. Two matched-count arm sequences extend from a common sparse baseline: the **context arm** (S_CTX) adds fields encoding the geographic, historical, and associative circumstances of the object (*Find spot*, *Place*, *Assoc name*), and the **content arm** (S_CONT) adds fields encoding properties directly perceivable from or readable on the object (*Subjects*, *Technique*, *Inscription*). We evaluate each generation using reference-based surface similarity, NLI-estimated field coverage, and claim-level factuality via nugget extraction and verification.

Our results show that more metadata does not simply make generations more grounded. Increasing field count improves coverage of provided fields but can also increase unsupported inference over withheld fields. Field identity determines the dominant failure mode: contextual fields tend toward omission, while content fields are incorporated more readily but at higher hallucination risk. Images provide visual grounding in sparse settings but also encourage inference beyond the catalogue record. Place-level patterns further show that metadata effects depend on documentation practices and cultural-geographic priors, making field-level evaluation necessary for understanding LLM-generated museum descriptions.

We make three contributions. **(1)** We introduce a *controlled field-type probing framework* isolating the effects of metadata quantity, field type, and image access on LLM-generated museum descriptions. **(2)** We show that more metadata does not improve faithfulness monotonically: increasing field count improves coverage of provided fields but increases unsupported inference over withheld fields, with contextual and content fields failing through distinct mechanisms. **(3)** We show that place-level variation reflects field–region interactions and documentation practices, motivating field- and claim-level evaluation for metadata-conditioned generation.

2 Related Work

Data-to-text generation Our work is situated in data-to-text generation, where benchmarks such as E2E and WebNLG study generation from meaning representations and RDF triples, emphasising content selection, lexicalisation, aggregation, and surface realisation (Novikova et al., 2017; Gardent et al., 2017). Longer-form data-to-document tasks such as RotoWire show that generation from structured records also requires maintaining fidelity to the input database (Wiseman et al., 2017), while controlled table-to-text settings such as ToTTo restrict the input evidence through selected table cells to support more controlled evaluation of grounded generation (Parikh et al., 2020). Museum description generation is a metadata-conditioned form of data-to-text generation, but with heterogeneous and incomplete records whose fields differ in availability and interpretive status.

Faithfulness in grounded generation A central challenge in grounded generation is that fluent outputs may include unsupported content. This problem has been widely studied in summarization, where models may generate hallucinated or factually inconsistent claims despite high surface quality (Maynez et al., 2020; Pagnoni et al., 2021). Atomic claim-level verification has been proposed to improve factuality assessment beyond binary judgements (Min et al., 2023). In data-to-text generation, factual grounding is similarly central: models may omit input facts, hallucinate unsupported content, or generate claims that cannot be traced to the source record (Wiseman et al., 2017; Wang, 2019). We build on this line of

work by analysing faithfulness at the metadata-field level.

Generation from cultural heritage data The use of LLMs on and for cultural heritage data is extensive. For example, Bu et al. (2025) study how various open source LLMs answer questions about a museum exhibition; Trichopoulos (2023) finetune an LLM with collection data to serve as a chatbot; Arnold and Tilton (2024) use LLMs to generate captions and explanations for image archives; Reusens et al. (2025) employ LLMs to generate keywords from written descriptions of museum objects. Generating natural language from museum collection databases has been studied in the past as a structured NLG problem (Isard et al., 2018), demonstrating that metadata records can support adaptive visitor-facing texts. In line with this, we generate descriptions based on structured records consisting of both images and metadata.

Metadata completeness is a central dimension of metadata quality (Bruce and Hillmann, 2004; Matoni et al., 2026). However, cultural heritage records are often incomplete. In Cultura Italia, for example, even mandatory fields are frequently missing (Lorenzini et al., 2021). These gaps are not random. Metadata creation and aggregation reflect judgements about what information is valuable, which makes documentation practices subjective and culturally situated (Lewis et al., 2025). As a result, completeness can vary systematically across subcollections, communities, or domains. This phenomenon is captured by Metadata Completeness Disparity (Battaner and Spence, 2026). To account for this, our work studies cultural heritage artefacts from various regions. By investigating how metadata completeness, image availability, and region interact and shape generation, we provide insights into whether models respect the boundary between verified catalogue information and model-supplied contextualisation.

3 Experimental Setup

To understand how museum description generation is affected by the count and type of fields and whether visual support is available, we propose a two-phase controlled field-type probing framework: Phase 1 measures generation under natural metadata variation, while Phase 2 applies controlled field probes that hold object identity fixed and vary metadata composition and image access.

3.1 Task and Dataset

Task. The task is structured metadata-conditioned museum description generation. A model G_m receives a prompt $P(c, M_i, I_i)$ constructed from condition c, object metadata M_i, and optional image I_i, and generates a description: $\hat{y}_{i,m,c} = G_m(P(c, M_i, I_i))$. $\hat{y}_{i,m,c}$ is assessed against a reference description y_i. Fields provided in the prompt constitute *primary fields* F_{cond}; fields present in the object record but withheld from the prompt are *bonus fields* $F_{\text{all}} \setminus F_{\text{cond}}$. Primary-field analysis measures use of provided metadata; bonus-field analysis measures parametric inference over catalogue information absent from the prompt.

Dataset. We use the *TimeTravel* benchmark (Ghaboura et al., 2025), comprising 10,250 museum artefact records across ten geographic regions: Iraq, Egypt, Greece, Roman Empire, British Isles, China, Japan, Central America, India, and Iran. Each record contains structured catalogue metadata, an image, and a reference description. We analyse nine metadata fields: *Materials*, *Culture*, *Production date*, *Find spot*, *Place*, *Associated name*, *Subjects*, *Technique*, and *Inscription*. The *Place* field encodes which cultural-geographic region the artefact belongs to, and we use this to compare generations across regions. The mean number of populated fields per object is $|\bar{F}| = 6.8$ (range 3–9); documentation varies substantially across regions (Table 1).

3.2 Two-Phase Experimental Design

Phase 1: natural metadata variation. Phase 1 evaluates generation under the metadata naturally available for each object. We prompt each model with the object's populated catalogue fields and image, generating descriptions for 10,250 objects with four LLMs. This phase captures realistic collection-level variation, where field count, field type, image availability, and place co-vary and cannot be independently attributed to observed differences in generated descriptions, thus our motivating phase 2.

Phase 2: controlled field probes. We construct Phase 2 as a within-object design over 500 objects balanced across values of the *Place* field. Objects are sampled from place-relative high-completeness tiers, rather than a global completeness threshold, to avoid making metadata quantity a proxy for civilisation (Appendix A.1; Table 6). Each object

Place	Assoc. name	Culture	Find spot	Inscription	Materials	Prod. date	Subjects	Technique
Overall	44.0	100.0	68.2	41.1	99.8	86.7	54.3	61.7
British Isles	1.0	100.0	98.0	1.0	100.0	93.0	6.0	66.0
Central America	1.0	100.0	100.0	0.0	100.0	22.0	3.0	25.0
China	0.0	100.0	100.0	0.0	100.0	100.0	12.0	100.0
Egypt	100.0	100.0	100.0	100.0	100.0	16.0	8.0	100.0
Greece	61.0	100.0	86.0	16.0	100.0	100.0	100.0	89.0
India	84.0	100.0	21.0	85.0	100.0	100.0	87.0	10.0
Iran	18.0	100.0	21.0	51.0	98.0	99.0	59.0	88.0
Iraq	28.0	100.0	94.0	30.0	100.0	100.0	87.0	41.0
Japan	8.0	100.0	47.0	28.0	99.0	100.0	17.0	43.0
Roman Empire	99.0	100.0	7.0	100.0	100.0	100.0	100.0	3.0

Table 1: Percentage of metadata available per place and overall; overall availability is computed across all 10,250 objects.

appears under multiple prompt conditions (Table 2), so comparisons are paired within object. Object identity, and prompt template remain fixed, while field composition and image access vary.

Controlled generation conditions. We define ten generation conditions (Table 2) that vary field count, field type, and image availability. Condition S is a three-field baseline comprising *Materials*, *Culture*, and *Production date*, selected because each is populated for nearly all of the benchmark objects. Six of the ten conditions form a 2 (field type) $\times$ 3 (field count) design over this baseline: each arm extends S by one field per level, so that level k is a superset of level $k-1$. The two arms partition the added fields along the content–context distinction (Gilliland, 2016): content comprises properties intrinsic to the object, including its subject matter, means of production, and any text it bears; context comprises properties extrinsic to the object, recording the circumstances of its creation, provenance, and custodial history.

The **context arm** (S_CTX4–6) adds *Find spot*, then *Place*, then *Assoc name*, which link the object to external documentary and associative records. The **content arm** (S_CONT4–6) adds *Subjects*, then *Technique*, then *Inscription*, which encode properties observable in the object or inscribed on it, although assigning catalogue values to them requires expert interpretation. Field order within each arm is fixed across all objects and models, so level k denotes an identical field set for every generation. The arms are matched on field count at levels 4, 5, and 6, and object identity, prompt template, and image availability are held fixed across matched pairs, so differences at a given level are attributable to field type rather than field count. The remaining four conditions: R and T bound the field range, and S_NOI and R_NOI withhold the image.

Cond.	Fields provided	$\lvert F \rvert$	Img.
R	All available	6–9	✓
T	All available	6–9	×
S	Materials, Culture, Prod. date	3	✓
S_NOI	Materials, Culture, Prod. date	3	×
S_CTX4	S + Find spot	4	✓
S_CTX5	S + Find spot, Place	5	✓
S_CTX6	S + Find spot, Place, Assoc. name	6	✓
S_CONT4	S + Subjects	4	✓
S_CONT5	S + Subjects, Technique	5	✓
S_CONT6	S + Subjects, Technique, Inscr.	6	✓

Table 2: Phase 2 controlled prompt conditions. Matched-count arm conditions compare field type while holding object identity, prompt template, and image access fixed.

3.3 Evaluation

We evaluate the generations on three dimensions:

Reference similarity. We compute BLEU-1 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and BERTScore (`roberta-large`) (Zhang et al., 2020) against y_i.

Field-value support. We measure whether $\hat{y}_{i,m,c}$ supports catalogue field values using DeBERTa-large-MNLI (He et al., 2021). For each $f \in F^{\text{eff}}_{i,c}$ with value v_f, we construct a value-specific hypothesis h_f and classify $\hat{y}$ via argmax over `entailment`, `neutral`, and `contradiction` [1]. Primary coverage is the entailment rate over the provided fields. For bonus

[1]We list the hypothesis for each field category in Appendix A.3 (Table 8)

fields, entailment indicates model recovery of a withheld catalogue value; contradiction indicates incorrect inference.

Claim-level factuality. NLI provides one entailment verdict per field and cannot identify which claims in the generation are responsible for that verdict. We therefore apply a two-stage nugget pipeline to verify factual accuracy at the claim level. Stage 1 decomposes g into atomic claims, attributing each to a field $f \in F_{\text{all}}$ or classifying it as *extra content* when no schema field accounts for the claim.

Stage 2 verifies each attributed claim against F_{all}, assigning `supported`, `hallucinated`, or `not_verifiable` with error-type classification (Table 9). Both labels are defined relative to the catalogue record rather than to external fact: a claim is `hallucinated` when the field to which it is attributed has a recorded value and the claim is not supported by that value, and `not_verifiable` when the field has no recorded value against which to assess it. A claim may therefore be historically accurate or visually correct and still be labelled `hallucinated`. Pipeline validation is reported in Appendix A.4.

The primary hallucination rate $\text{HallP}(i, c)$ is the proportion of claims in $\mathcal{N}^{P}_{i,c}$ assigned `hallucinated` by Stage 2. Claim verdicts are aggregated to the field level: a primary field $f \in F_{\text{cond}}$ is `hallucinated` if any attributed claim is hallucinated, `covered` if at least one is supported and none are hallucinated, and `silent` if no claims are attributed to it. Bonus fields $f \in F_{\text{all}} \setminus F_{\text{cond}}$ follow the same logic with labels `correct`, `wrong`, and `silent`. The extra-content rate is the proportion of claims unattributable to any field in F_{all}.

Implementation details, including model versions, prompt templates, and significance tests, are provided in Appendix A.3.

Human calibration. We assess the reliability of the claim-level pipeline against human annotation. Two annotators independently labeled a stratified random sample of 350 claims, drawn proportionally across conditions, following an annotation protocol derived from the Stage 2 prompt. Inter-annotator reliability is substantial (Cohen's $\kappa = 0.78$), with category-specific agreement of 96% for `supported` and 87% for `hallucinated`, showing that human raters apply the verdict criterion consistently without reference to the pipeline. Human–pipeline reliability is $\kappa_w = 0.67$ with category-specific agreement of 89% for `supported` and 73% for `hallucinated`. This shows that our nugget pipeline is able to correctly attribute most nuggets to their specific fields.

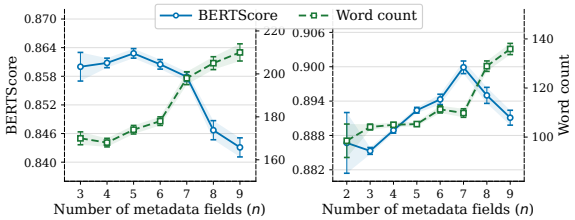

Figure 1: Mean BERTScore and generated word count as a function of the number of populated metadata fields n, across Phase 1 (left) and Phase 2 (right).

4 Results

We report results across the two phases.

4.1 Phase 1: Surface metrics invert documentation completeness

Phase 1 generates descriptions under each artefact's full available metadata. Figure 1(a) shows that mean description length increases monotonically with field count, from 160 to 213 words across the range of 3–9 fields, while mean BERTScore-P decreases from 0.863 to 0.839 over the same range. Across regions, BERTScore is consistently and negatively associated with description length, with correlations ranging from $r = -0.34$ to $r = -0.70$s ($\eta^2 = 0.33$ for the length–field-count relationship Table 13): indicating that more fields correspond to longer generations, which receive lower BERTScore against the per-object reference description.

Surface-metric rankings almost perfectly invert documentation completeness (Kendall's $W = 0.990$), with well-documented regions scoring lower despite richer records (Table 11), replicating the regional difficulty pattern reported by Ghaboura et al. (2025).[2] As field count and region co-vary in naturalistic data, their individual contributions to quality variation cannot be disentangled from observational results. Below, we report our results from phase 2, which disentangles these dimensions.

[2]Per-model results across all surface metrics are reported in Appendix B.1.

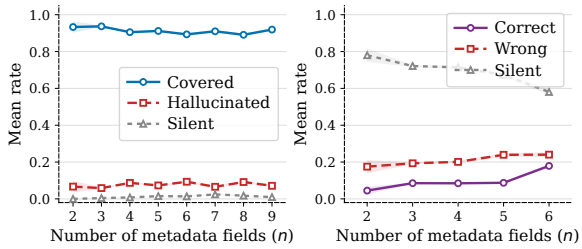

Figure 2: Mean primary fields' (coverage, hallucination, silence; left) and bonus fields' (correct, wrong, silent; right) as a function of observed metadata field count n.

	Primary fields			Bonus fields		
Cond.	Cov	Hall	Sil	Cor	Wro	Sil_b
S	0.933	0.058	0.009	0.092	0.223	0.685
S_NOI	0.885	0.091	0.024	0.055	0.146	0.799
S_CTX4	**0.949**	**0.040**	0.011	0.116	0.308	0.577
S_CTX5	0.941	0.044	0.016	0.128	0.361	0.511
S_CTX6	0.937	0.042	0.021	**0.218**	0.421	0.360
S_CONT4	0.893	0.102	0.005	0.062	0.153	0.786
S_CONT5	0.867	0.129	**0.004**	0.032	**0.031**	**0.937**
S_CONT6	0.876	0.113	0.011	0.060	0.076	0.864
R	0.885	0.092	0.023	—	—	—
T	0.879	0.082	0.039	—	—	—

Table 3: Primary fields rates (Cov, Hall, Sil) and bonus completion rates (Cor, Wro, Sil_b) by condition, averaged across all models; Bold indicates best value per column: highest for Cov, Cor, Sil_b; lowest for Hall, Sil, Wro. R and T contain no withheld fields by design.

4.2 Metadata composition creates grounding trade-offs

Field count scales bonus inference without improving primary grounding. Figure 1(b) shows that in Phase 2, BERTScore increases from 0.885 at $n = 3$ to 0.901 at $n = 7$, before decreasing to 0.890 at $n = 9$, while word count increases monotonically from 100 to 135 words. Unlike Phase 1, the inverse relationship between description length and BERTScore does not hold uniformly once object identity is held fixed. Table 16 confirms this at the regional level: within-place field-count–BERTScore correlations shift from negative across all regions in Phase 1 to positive across all ten regions in Phase 2 ($\rho =$ 0.200 to 0.421).

Field-count effects in Phase 2 show that increasing metadata quantity does not uniformly improve grounding, as shown in Figure 2. For primary fields, coverage remains stable across all field counts, ranging from 0.90 to 0.93, while hallucination rates vary between 0.07 and 0.10. Silence rates remain near zero throughout, indicating that models verbalise provided fields regardless of how many are provided.

Bonus field rates show greater variation with field count. Bonus silence decreases from 0.79 at $n = 2$ to 0.65 at $n = 6$, while both correct and wrong bonus rates increase over the same range. Bonus wrong exceeds bonus correct at every observed field count, with wrong rates stable at 0.18–0.24 against correct rates rising from 0.05 to 0.19. Higher field counts are therefore associated with both greater parametric recovery of withheld fields and greater metadata-boundary overreach.

Field type changes the dominant failure mode. The claim-level pipeline reveals systematic differences between matched-count conditions. S_CTX conditions are more conservative and more faithful on primary fields: S_CTX4–6 achieve primary coverage of 0.937–0.949 with hallucination rates of 0.040–0.044 (Table 3). S_CONT conditions produce lower primary coverage (0.867–0.893) and higher hallucination (0.102–0.129). The largest contrast occurs in S_CONT5, which adds *Subjects* and *Technique*: it has the highest primary hallucination rate (0.129) while maintaining near-zero silence (0.004). S_CONT fields therefore show high primary coverage rates while also having high primary hallucination rates. Paired permutation tests over matched field counts (10,000 resamples, BH–FDR corrected) confirm that this split holds at every step of the arm: S_CTX exceeds S_CONT in primary coverage by 5.6–7.4 pp and in bonus wrong by 15.5–34.5 pp, and falls below it in primary hallucination by 6.2–8.5 pp (all $p < 0.001$).

Table 4 shows that, *Subjects* and *Technique* have the highest primary hallucination rates (24.2% and 21.3%), far above fields such as *Culture* (1.3%) and *Production date* (1.7%). S_CTX fields have lower hallucination but higher omission: *Find spot* and *Place* show hallucination rates of 3.9% and 5.6%, with silence rates of 5.0% and 8.1%. S_CONT fields therefore fail primarily through unsupported elaboration, whereas S_CTX fields fail primarily through omission. *Inscription* is the only field for which bonus correct exceeds bonus wrong (16.7% vs. 10.6%), alongside a moderate primary hallucination rate (9.1%), indicating that *Inscription* content is partially recoverable when withheld.

Withheld-field follows the same field-type split.

	Primary			Bonus		
Field	Cov	Hall	Sil	Cor	Wro	Sil_b
Subjects	74.9	24.2	0.9	14.6	41.8	43.6
Technique	78.3	21.3	**0.3**	9.8	51.7	38.5
Assoc name	87.1	11.9	1.0	5.5	6.5	88.0
Materials	88.7	10.9	0.4	—	—	—
Inscription	87.5	9.1	3.4	**16.7**	10.6	72.6
Place	86.3	5.6	8.1	6.9	6.3	86.8
Find spot	91.2	3.9	5.0	0.4	**0.4**	**99.2**
Prod. date	96.2	1.7	2.1	—	—	—
Culture	**98.2**	**1.3**	0.5	—	—	—

Table 4: Primary fields rates (Cov, Hall, Sil) and bonus completion rates (Cor, Wro, Sil_b) by metadata field, averaged across all conditions and models.

Wrong bonus mentions exceed correct ones overall, indicating that unprompted field recovery is more often speculative than reliable. The highest bonus wrong rates occur for *Technique* (51.7%) and *Subjects* (41.8%). In the S_CTX arm, where S_CONT fields remain withheld, bonus wrong increases from 0.308 in S_CTX4 to 0.421 in S_CTX6, while bonus silence decreases from 0.577 to 0.360 (Table 3). Additional context from S_CTX increases completion of the withheld record, but much of this completion is unsupported. In contrast, withheld S_CTX and association fields are usually omitted: *Find spot* has 99.2% bonus silence and 0.4% bonus wrong, while *Place* and *Assoc. name* show bonus silence rates of 86.8% and 88.0%. Field type, therefore, determines whether missing information is omitted or speculatively completed, making field count alone insufficient for explaining claim-level faithfulness. The gap widens along the S_CTX arm, from 15.5 pp at count 4 to 34.5 pp at count 6, while bonus silence diverges in the opposite direction, from 0.577 against 0.786 to 0.360 against 0.864 (all $p < 0.001$).

4.3 Grounding reliability varies by region

Regions differ in dominant failure mode, not in overall reliability. BERTScore varies significantly across regions in all conditions (Kruskal–Wallis $H > 100$, $p \ll 0.001$; Table 15), but over a narrow range of means, from 0.877 (Greece) to 0.896 (Central America). A negative length–BERTScore correlation holds within every region ($r = -0.38$ to -0.71, all $p \ll 0.001$; Figure 3). Claim-level rates separate the regions more sharply than BERTScore does (Table 5): the British Isles records the highest primary coverage (0.936) and lowest hallucination (0.054), Central America the highest hallucination (0.112), and China the lowest coverage (0.859) with the highest silence (0.047). Bonus wrong exceeds bonus correct in every region; India records the lowest bonus wrong rate (0.133) and the Roman Empire the highest bonus correct (0.170).

These aggregate differences arise from field–region interactions (Figure 4). *Subjects* and *Technique* record the highest primary hallucination rates in every region, but the magnitude varies by a factor of 2–4: among regions where the field is populated, *Technique* ranges from 0 (Roman Empire) to 0.46 (Central America), and *Subjects* spans a narrower 0.14–0.35. China and Iran record higher primary silence on S_CTX fields than other regions (*Find spot* 0.26 and *Place* 0.24 for China; *Find spot* 0.19 for Iran). For Greece, models generate bonus *Technique* content in 71% of cases, of which 0.69 are wrong against 0.04 correct.

Part of the apparent regional variation reflects catalogue composition rather than generation behaviour. The near-zero *Technique* hallucination rate for the Roman Empire follows from the field being absent from its records, so no primary claims are exposed to verification against it. Regional rates are therefore conditional on which fields each collection populates, and comparisons across regions hold only at matched field type and field count.

Generation failure is not uniform across regions. The British Isles exhibits high faithfulness to the provided metadata; Central America produces unsupported elaboration on *Subjects* and *Technique*; China omits provided field content; Greece incorrectly infers withheld *Technique*. These modes carry different consequences for museum description: hallucination introduces false content, silence omits documented content, and unsupported bonus inference fabricates attributes absent from the record. Deployment in cultural heritage settings therefore requires region-specific quality assurance rather than a single global accuracy threshold.

4.4 Image helps sparse grounding but increases extrapolation

Visual input affects grounding only under sparse metadata. Image ablation under sparse metadata (S_NOI versus S) reduces BERTScore across all models, lowers primary coverage from 93.3% to 88.5%, and raises primary hallucination from

	Primary fields			Bonus fields		
region	Cov	Hall	Sil	Cor	Wro	Sil_b
Central A.	0.877	0.112	0.010	0.050	0.304	0.646
China	0.859	0.094	**0.047**	0.062	0.268	0.669
British I.	**0.936**	**0.054**	0.010	0.059	0.232	0.710
Japan	0.918	0.072	0.010	0.114	0.185	0.701
India	0.901	0.082	0.017	**0.137**	**0.133**	0.730
Iraq	0.897	0.093	0.010	0.113	0.187	0.700
Iran	0.913	0.068	0.020	0.102	0.199	0.699
Egypt	0.922	0.063	0.015	0.082	0.145	**0.774**
Roman E.	0.912	0.071	0.017	0.170	0.236	0.593
Greece	0.898	0.086	0.016	0.051	0.234	0.715

Table 5: Primary and bonus fields results from the claim-level factuality.

5.8% to 9.1% (Table 3). Paired permutation tests find each difference significant (4.8 pp coverage, 3.3 pp hallucination, 7.7 pp bonus wrong; all $p < 0.001$). The same ablation under full metadata yields no significant difference: R and T differ by 0.6 pp in primary coverage and 1.0 pp in primary hallucination, neither significant after correction. The contribution of visual input to primary grounding is therefore conditional on metadata density, and is not detectable once the prompt specifies all available fields.

The effect is concentrated in withheld rather than provided fields. Ablation raises bonus silence from 68.5% to 79.9% and lowers bonus wrong from 22.3% to 14.6%, with the wrong-to-correct ratio increasing from 2.4:1 to 2.7:1. Image-inclusive conditions thus produce more attempts at withheld fields and a higher absolute count of incorrect completions, but a lower proportion of incorrect attempts. Visual input extends generation beyond the provided fields without a corresponding gain in accuracy, reproducing the metadata-boundary overreach observed along the S_CTX arm (Section 4.2) from a visual rather than a textual source.

Model-level effects vary under full metadata. LLAMA-VISION records the largest R-to-T primary coverage difference of any model (0.882 to 0.796; $\Delta = 0.086$), against an aggregate difference of 0.6 pp. Per-model and per-region results across field count, type, place, and image ablation are reported in Appendix B.

5 Discussion

Our results show that metadata-conditioned generation is not only a question of input coverage, but also of field–claim alignment. Adding metadata changes what models mention, but also how they redistribute information across catalogue fields, visual evidence, and domain priors. This challenges the prevailing data-to-text assumption that richer inputs consistently improve generation quality.

First, metadata quantity is not a reliable indicator of grounding. In Phase 2, increasing field count improves coverage of provided fields but also increases hallucination over withheld fields. Better grounding of visible metadata can therefore coincide with weaker respect for the metadata boundary. Additional fields change the distribution of claims in the output: models mention more catalogue-supported information, but they also produce longer descriptions with more opportunities for unsupported elaboration. This distinction is visible in the metric results. Field-count correlations with BERTScore vary across regions, whereas output length is consistently and negatively associated with BERTScore. Surface metrics, therefore, risk penalizing grounded elaboration while rewarding shorter descriptions that are closer to the reference but less complete.

Second, field type shapes the primary error distribution. Context fields (*Find spot*, *Place*, *Assoc name*), are reliably grounded when present and are rarely inferred when absent. Whereas content fields such as *Technique* and *Subjects* encourage domain-specific elaboration and visual inference, resulting in higher hallucination rates even when provided. *Materials* occupies an intermediate position: it is usually grounded when present, but can trigger unsupported object-type or quality inferences. The context arm (S_CTX), therefore, improves primary fidelity, while the content arm (S_CONT) increases primary coverage at higher hallucination cost. These patterns show that many errors are not arbitrary fabrications. Instead, models transfer values across neighbouring fields, narrow broad catalogue labels using domain priors, or convert visual observations into field-specific claims. The error often lies not in whether a claim is plausible, but in whether the correct field supports it.

This field–claim mismatch is the central failure mode our analysis reveals. A generated statement may be visually plausible, historically reasonable, or supported by another metadata field, while still being unsupported by the target field being evaluated. For example, a model

may use a *Find spot* value as a *Place* of origin, convert a broad technique label into a specific art-historical sub-type, or add subject details visible in the image but absent from the catalogue subject vocabulary. Such cases are difficult to capture with evaluation methods that only check whether a claim is plausible or whether it appears somewhere in the input. Metadata-conditioned generation requires verifying the relation between claims and the specific fields that support them. Omission and hallucination also carry different consequences: omission reduces completeness, whereas unsupported elaboration presents unverified claims as catalogue-grounded facts.

Images further expose this issue, and their effect depends on metadata density. In sparse conditions, images provide grounding evidence that reduces hallucination; under richer metadata, they can amplify content-field elaboration beyond the catalogue record (Rohrbach et al., 2018). Our evaluation treats catalogue metadata as the reference schema, not as an exhaustive description of the object. Image-grounded claims may therefore be visually valid while remaining unsupported by the metadata field under evaluation. This distinction is important for cultural heritage generation: a description can be visually plausible while going beyond what the catalogue schema supports.

These findings have implications for evaluating metadata-conditioned generation. Reference similarity metrics, field-level NLI, and claim-level verification capture different aspects of quality and can produce conflicting condition rankings. NLI-based coverage measures semantic consistency rather than value accuracy: a description can receive an entailment verdict without correctly reproducing the catalogue value (Falke et al., 2019), and no single evaluation layer separates omission from hallucination. Evaluation should therefore be field-aware: not only whether a claim is supported, but which field supports it, whether that field was provided, and whether the claim stays within its intended scope. Treating all metadata fields as equivalent hides important differences in coverage rate, hallucination risk, and omission. More broadly, metadata-conditioned generation should be treated as a schema-alignment problem: models must learn not only what information is available, but what each field permits them to assert. In cultural heritage settings, unsupported elaboration can present speculative interpretation as authoritative descriptions.

6 Conclusion

We presented a controlled study of metadata-conditioned description generation for museum objects with multimodal LLMs, varying the number of metadata fields, their type, and image availability over a fixed object set. Increasing the number of provided fields does not monotonically improve faithfulness: entailment-based coverage of the provided fields rises, and so does the rate of unsupported claims about fields withheld from the prompt. Field type affects both quantities at a fixed number of fields. Context fields are either expressed or omitted and are rarely misstated, whereas content fields such as *Technique* and *Subjects* are expressed whenever provided but yield the highest rate of unsupported claims and encourage visual inference, domain-specific elaboration, and schema-boundary errors. Reference-based similarity and entailment-based coverage do not distinguish these behaviours: they order the matched conditions differently, and a difference in error type is read as a difference in quality. Faithfulness requires attributing each generated claim to the metadata field it addresses and verifying it against that field's value, which separates support, contradiction, and omission. For cultural heritage applications, this matters because fluent descriptions can present plausible but unsupported interpretation as authoritative object information. Systems deployed in this setting should therefore expose the provenance of each claim, so that catalogue-supported content can be reviewed separately from content generated from parametric knowledge.

Acknowledgments

This research is supported by the Dutch Science Foundation (NWO) project HAICu with project number NWA.1518.22.105. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Limitations

We distinguish contextual and content fields, but this distinction may not be absolute: some fields can behave differently depending on object

type, record format, and catalogue practice. For example, *Materials* is usually grounded reliably when present, but can still trigger unsupported object-type or quality inferences. Future work should examine whether field categories should be defined dynamically based on both schema role and object-specific usage.

Our claim-level factuality is evaluated with an automatic nugget-verification pipeline and two crowdworkers rather than expert curator annotation. We validate the pipeline through span checks and agreement with NLI-based field signals, but automated verification may still miss subtle curatorial distinctions. Future work should include expert review, especially for visually plausible claims that are not recorded in the catalogue metadata.

Finally, expert references are selective descriptions rather than exhaustive records, which partly explains the divergence between reference similarity and field-level support. Our evaluation treats catalogue metadata as the authoritative source of support, which is appropriate for studying schema-faithful generation, but catalogue records may omit details visible in the image or known from external sources. Future work should examine how humans and domain experts balance descriptiveness, visual plausibility, and catalogue fidelity, and should distinguish catalogue-supported, image-supported, and externally supported claims rather than treating all unsupported content as equivalent.

Ethical Considerations

LLM-generated cultural heritage descriptions can shape how public audiences understand objects, places, and cultural histories. Our results show that fluent descriptions may introduce plausible but unsupported interpretations, especially for content fields such as *Technique* and *Subjects*. In public-facing settings, such errors risk presenting speculative claims as authoritative information.

This risk is particularly important for collections whose documentation is uneven across regions, periods, or communities. If models elaborate more confidently for some cultural-geographic contexts than others, generated descriptions may reinforce existing documentation imbalances rather than correct them. We therefore do not recommend using these systems as standalone curatorial tools. Instead, domain experts should review generated descriptions, and systems should show which claims are supported by catalogue metadata, which are inferred from images, and which remain uncertain.

Our work uses existing museum catalogue records and object images. We do not collect personal data or generate descriptions about living individuals. The main ethical concern is representational: ensuring that automated description generation does not overstate, distort, or homogenise cultural knowledge.

References

Taylor B. Arnold and Lauren Tilton. 2024. Explainable search and discovery of visual cultural heritage collections with multimodal large language models. In *Workshop on Computational Humanities Research*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL technical report. *ArXiv*, abs/2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Elena Battaner and Paul Spence. 2026. The linguistic asymmetry index (LAI): Benchmarking equity in multilingual research infrastructures. *Journal of Open Humanities Data*.

Thomas R. Bruce and Diane I. Hillmann. 2004. The continuum of metadata quality: Defining, expressing, exploiting. In Diane I. Hillmann and Elaine L. Westbrooks, editors, *Metadata in Practice*, pages 238–256. ALA Editions, Chicago.

Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. An investigation into value misalignment in LLM-generated texts for cultural heritage. *ArXiv*, abs/2501.02039.

Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2023. An ontological approach for unlocking the colonial archive. *J. Comput. Cult. Herit.*, 16(4).

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with

natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261.

Sara Ghaboura, Ketan Pravin More, Ritesh Thawkar, Wafa Al Ghallabi, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Salman Khan, and Rao Muhammad Anwer. 2025. Time travel: A comprehensive benchmark to evaluate LMMs on historical and cultural artifacts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641, Vienna, Austria. Association for Computational Linguistics.

Anne J. Gilliland. 2016. Setting the stage. In Murtha Baca, editor, *Introduction to Metadata*, 3rd edition. Getty Research Institute, Los Angeles.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Eero Hyvönen. 2012. *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool.

Amy Isard, Jon Oberlander, and Claire Grover. 2018. Up-cycling data for natural language generation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Michael Lewis, Eljas Oksanen, Frida Ehrnsten, Heikki Rantala, Jouni Tuominen, and Eero Hyvönen. 2025. The impact of human decision-making on the research value of archaeological data. *J. Comput. Cult. Herit.*, 18(3).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *ArXiv*, abs/2407.21783.

Matteo Lorenzini, Marco Rospocher, and Sara Tonelli. 2021. On assessing metadata completeness in digital cultural heritage repositories. *Digital Scholarship in the Humanities*, 36(Supplement 2):ii182–ii188.

Markus Matoni, Arno Kesper, and Gabriele Taentzer. 2026. Quality of descriptive information on cultural heritage objects: Definition and empirical evaluation. *Preprint*, arXiv:2602.21249.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Meta. 2025. Llama 4 model cards and prompt formats.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrücken, Germany. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 4812–4829, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Manon Reusens, Amy Adams, and Bart Baesens. 2025. Large language models to make museum archive collections more accessible. *AI Soc.*, 40(6):4485–4497.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Ramesh Srinivasan. 2017. *Whose Global Village? Rethinking How Technology Shapes Our World*. New York University Press, New York.

Georgios Trichopoulos. 2023. Large language models for cultural heritage. In *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*, CHIGREECE '23, New York, NY, USA. Association for Computing Machinery.

Seth van Hooland and Ruben Verborgh. 2015. *Linked Data for Libraries, Archives and Museums: How to clean, Link and Publish your Metadata*. Facet.

Hongmin Wang. 2019. Revisiting challenges in data-to-text generation with fact grounding. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 311–322, Tokyo, Japan. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Experimental Setup Details

This appendix provides additional detail on metadata field availability and the place-relative sampling procedure used to construct the Phase 2 controlled sample. We also discuss how we handled field collapse across conditions and list the prompts used in generation and evaluation.

A.1 Metadata Availability and Place-Relative Sampling

Table 1 reports field availability by region, defined as the proportion of objects within each region for which a given field is populated. *Culture*, *Place*, and *Materials* are near-universally populated across all regions. Some fields vary substantially: *Subjects* is fully populated for Greece and the Roman Empire but below 10% for British Isles and Central America; *Find spot* is fully populated for Central America and China but below 25% for India and the Roman Empire; *Inscription* ranges from 0% in Central America and China to 100% in Egypt. These disparities reflect differences in institutional documentation practice across collections and motivate the place-relative sampling for phase 2.

In phase 2, we apply a place-relative criterion: for each region, we sample 50 objects from the highest available field-count tiers within that category, defined as the one or two tiers with the largest object counts at the upper end of each region's completeness distribution. We refrain from using a global completeness threshold as this would disproportionately exclude regions with lower field counts, biasing the controlled sample toward better-documented regions. This ensures all ten regions are represented in phase 2 and that the ceiling conditions (R and T) use the richest metadata available within each region rather than a globally imposed minimum.

Table 6 shows the distribution of objects by region and field count. Bold cells indicate the tiers from which Phase 2 objects are drawn. The minimum selected tier ranges from six fields for Central America, where nine-field objects are absent or negligible, to seven or more for the remaining regions.

A.2 Collapsed Conditions

A condition is *collapsed* for object i when $|F_{i,c}^{\text{eff}}| < |F_c^{\text{tar}}|$, occurring when one or more target fields are unpopulated in the object catalogue record. Collapse is most frequent in S_CONT6 (*Inscription*:

Place	3	4	5	6	7	8	9
British Isles	17	45	276	616	**54**	**5**	–
Iran	15	18	155	352	332	**125**	**2**
Japan	5	239	225	200	**180**	**20**	–
Central America	–	304	132	**48**	**3**	**1**	–
India	–	7	77	50	621	**112**	1
Iraq	–	–	34	400	355	159	**52**
Greece	–	–	14	100	861	709	**185**
China	–	–	–	918	**117**	**3**	**1**
Roman Empire	–	–	–	12	943	**88**	**6**
Egypt	–	–	–	–	861	140	**55**

Table 6: Distribution of objects by *Place* and number of available metadata fields. Bold cells indicate place-relative high-completeness tiers used for Phase 2 sampling.

41% collection-wide availability) and S_STR6 (*Associated name*: 44%). All field-level analyses use $F^{\text{eff}}_{i,c}$ throughout; collapsed instances are retained in aggregate metrics.

To verify robustness, we recompute CovP and HallP on the non-collapsed subset for the three conditions with the highest collapse rates. Mean absolute deviation from full-sample estimates is below 0.01 in all cases, confirming that collapsed instances do not drive reported patterns.

A.3 Models and Implementation Details

Generation models. Table 7 lists the exact model versions used in the study. All generation models are run with temperature 0 and a maximum output length of 512 tokens. No output reached this limit. The fixed generation prompt is shown in Prompt A.1. In Phase 2, LLaMA-3.2-Vision is run only on the image ablation conditions (R, S, S_NOI, and T). The other three models are run on all ten conditions, yielding 17,000 attempted generations. In Phase 1, all four models generate descriptions for all 10,250 objects under full metadata, giving $n = 41{,}000$ generations. GPT-4o is accessed via OpenAI, while the remaining generation models are accessed via DeepInfra. Fewer than five failed or empty Phase 2 responses are excluded.

Image input. For image-enabled conditions, the artefact photograph is appended to the user prompt as a base64 JPEG vision block. We apply no client-side resizing or preprocessing. Provider APIs handle any internal normalisation. No image is included for S_NOI or T.

Surface metrics. BLEU-1 is computed with `nltk` using smoothing method 1. ROUGE-L is computed with `rouge-score` and stemming. METEOR is computed with `nltk.translate.meteor_score`. BERTScore is computed with `roberta-large`, no IDF weighting, and no baseline rescaling. BERTScore precision (P_{BERT}) is the primary surface metric.

NLI field coverage. Field coverage is computed with `microsoft/deberta-large-mnli` (He et al., 2021). Each field-value pair is converted into a template hypothesis (Table 8) and classified against the generated description by argmax over entailment, neutral, and contradiction. Entailment is counted as coverage. Contradiction is counted as active hallucination. Neutral is counted as silence.

Nugget verification. Claim-level factuality is evaluated with Gemini 2.5 Flash (Gemini Team, 2025) at temperature 0. Stage 1 extracts atomic claims using `max_tokens`=1,024. Stage 2 verifies each claim against the relevant catalogue value using `max_tokens`=200. Prompts are shown in Prompts A.2 and A.3. Primary fields are shown with ground-truth values, while bonus fields are shown only as field concepts. Malformed JSON responses are retried twice. Persistent failures are marked `not_verifiable` and account for fewer than 1% of calls. Due to API budget constraints, nugget verification is run on 6,120 descriptions stratified by region, model, and condition.

Model	Version	Task	API
GPT-4o (OpenAI, 2023)	`gpt-4o-2024-08-06`	Generation	OpenAI
Llama-4-Maverick (Meta, 2025)	`Llama-4-Maverick-17B-128E-Instruct-FP8`	Generation	DeepInfra
Llama-3.2-Vision (Llama Team, AI @ Meta, 2024)	`Llama-3.2-11B-Vision-Instruct`	Generation	DeepInfra
Qwen2.5-VL (Bai et al., 2025)	`Qwen2.5-VL-32B-Instruct`	Generation	DeepInfra
Gemini 2.5 Flash (Gemini Team, 2025)	`gemini-2.5-flash`	Evaluation	Google

Table 7: Model versions and APIs. LLaMA-3.2-Vision is restricted to the four base Phase 2 conditions. Gemini 2.5 Flash is used only for nugget extraction and verification.

A.4 Claim-Level Factuality Pipeline Validation

We run two sanity checks on the nugget pipeline. First, extracted spans must occur verbatim in the generated text. Span validity is 98% over the 6,120 evaluated descriptions. Second, NLI predictions are compared with nugget verdicts for primary fields. Entailment should align with `supported`, and contradiction should align with `hallucinated`. Agreement over primary field–generation pairs ($n \approx 30{,}000$) exceeds Cohen's $\kappa > 0.70$. Most disagreements are cases where NLI returns `neutral` but Stage 2 finds span-level support.

Prompt A.1: Generation Prompt

[System]
You are an expert museum curator writing object labels for a public gallery. Produce clear, professional descriptions accessible to a general audience.

[User]
Write a museum label for the artefact described by the following metadata.
Your response must:
- consists of continuous prose;
- contextualise the object culturally, historically, or artistically where the metadata supports it;
- not reproduce or reference the field label names;
- not introduce any information not grounded in the provided metadata;
- not use bullet points, headers, or lists.

Metadata
`{FIELD}: {VALUE}`
`{FIELD}: {VALUE}`
`...`

`{IMAGE appended for image-enabled conditions}`

Prompt A.2: Stage 1: Nugget Extraction Prompt

[System]
You are a specialist fact-checker for museum artefact descriptions. Extract every factual claim from the provided description as atomic nuggets grounded in verbatim text spans.

[User]
Decompose the following museum description into atomic factual claims (nuggets). For each nugget, provide:
- **span**: exact verbatim substring from the description, which must appear character-for-character;
- **text**: normalised paraphrase of the claim;
- **field**: the metadata field this claim is attributable to. Choose from: *Culture*, *Production date*, *Materials*, *Find spot*, *Place*, *Associated name*, *Subjects*, *Technique*, or *Inscription*. If the claim cannot be attributed to any schema field, assign `extra_content`.

Provided fields (ground-truth values shown):
`{FIELD}: {VALUE}`
`{FIELD}: {VALUE}`
`...`

Withheld field concepts (values not provided):
`{FIELD}, {FIELD}, . . .`

Description:
`{GENERATION}`

Return a JSON array of nugget objects. Each nugget should have the following form:
`{"span": "...", "text": "...", "field": "..."}`

Prompt A.3: Stage 2: Nugget Verification Prompt

[System]
You are an expert museum curator verifying the factual accuracy of claims about museum artefacts against authoritative catalogue records.

[User]
Verify the following claim against the provided ground-truth metadata value.
Assign one of the following verdicts:
- **supported**: the claim is factually consistent with the ground-truth value;
- **hallucinated**: the claim contradicts or is unsupported by the ground-truth value;
- **not_verifiable**: the claim cannot be confirmed or refuted from the ground-truth value alone.

If the verdict is `hallucinated`, additionally assign:
- **error_type**: one of `wrong_material`, `wrong_date`, `wrong_culture`, `wrong_place`, `wrong_subject`, `wrong_technique`, or `fabricated_detail`;
- **severity**: `minor`, for a factual deviation with limited impact, or `major`, for an error that would materially mislead a museum visitor.

Field: `{FIELD}`
Ground-truth value: `{VALUE}`
Claim (verbatim span): `{SPAN}`
Claim (normalised): `{TEXT}`

Field	Template	Example h_f
Materials	This object is made of [v].	...*made of bronze.*
Culture	This object belongs to the [v] culture.	...*Roman culture.*
Production date	This object was produced [v].	...*produced 100 BC.*
Find spot	This object was found at [v].	...*found at Pompeii.*
Place	This object originates from [v].	...*from Roman Empire.*
Associated name	This object is associated with [v].	...*associated with Augustus.*
Subjects	This object depicts [v].	...*depicts a battle scene.*
Technique	This object was produced using [v].	...*using lost-wax casting.*
Inscription	This object bears the inscription: [v].	...*inscription: DEO MARTI.*

Table 8: NLI hypothesis templates and instantiated examples. [v] is substituted with ground-truth value v_f.

Error type	Definition
`wrong_material`	Incorrect material stated (e.g., *iron* for ground-truth *bronze*)
`wrong_date`	Incorrect production date, century, or period
`wrong_culture`	Incorrect cultural or regional attribution
`wrong_place`	Incorrect geographic or archaeological provenance
`wrong_subject`	Incorrect iconographic or figurative content
`wrong_technique`	Incorrect production or manufacturing technique
`fabricated_detail`	Plausible elaboration with no anchor in any metadata field

Table 9: Hallucination error taxonomy. Each hallucinated nugget receives one error type and a severity label.

Return JSON in the following format:
```
{"verdict": "...", "error_type": "...",
"severity": "...", "reasoning": "max 20
words"}
```

B Additional Results Details

B.1 Phase 1 Results

Overall performance. Table 10 reports the overall automatic evaluation scores. LLaMA-Maverick achieves the highest mean scores across all reported metrics. Qwen2.5-VL-32B performs worst overall. Differences between LLaMA-Maverick and the remaining models are significant under paired Wilcoxon signed-rank tests computed at the object level. All comparisons have $p < 0.05$ and are marked with † in Table 10.

Place predicts score. We test whether metric scores vary by geographic origin using Kruskal–Wallis tests, with effect sizes reported as η^2. Place has a significant effect for every model–metric combination ($p < 0.001$ throughout). The effect is largest for BERTScore ($\eta^2 = 0.33$–0.43), indicating that place explains up to 43% of score variance. Effects are smaller for METEOR ($\eta^2 \approx 0.19$–0.21), ROUGE-L ($\eta^2 \approx 0.10$–0.15), and BLEU ($\eta^2 \approx 0.07$–0.11).

Table 11 shows per-place BERTScore scores. We measure agreement in regional rankings across models using Kendall's rank concordance coefficient (W). BERTScore shows near-perfect agreement across models, with $W = 0.990$, $\chi^2_9 = 44.6$, and $p < 0.0001$. This indicates that models rank places similarly by surface similarity. METEOR also shows high agreement, with $W = 0.877$ and $p < 0.0001$. ROUGE-L shows moderate agreement, with $W = 0.691$ and $p < 0.01$. BLEU shows weak and non-significant agreement, with $W = 0.310$ and $p = 0.125$. This suggests that BLEU place rankings are inconsistent across models, consistent with BLEU's higher sensitivity to output length variation.

Field count. Table 12 and Table 13 show that the relationship between metadata quantity and BERTScore differs by region. For Greece ($\rho = -0.350$), India ($\rho = -0.411$), and Iran ($\rho =$

Model	BLEU	ROUGE-L	METEOR	BERTScore
GPT-4o	0.240†	0.159†	0.353†	0.856†
LLaMA-Maverick	**0.309**	**0.196**	**0.389**	**0.860**
LLaMA-3.2-Vision	0.229†	0.159†	0.334†	0.848†
Qwen2.5-32B	0.152†	0.105†	0.277†	0.841†

Table 10: Phase 1 surface metrics ecaluation of generated descriptions. Bold = best score. † = significantly worse than best model under paired bootstrap and Wilcoxon tests (p<0.05).

Place	GPT-4o	LLaMA-3.2-Vision	LLaMA-Maverick	Qwen2.5-32B	Mean
British Isles	0.862	0.850	**0.867**	0.847	0.857
Central America	0.872	0.857	**0.875**	0.858	**0.866**
China	0.870	0.855	**0.872**	0.853	0.863
Egypt	0.850	0.845	**0.855**	0.837	0.847
Greece	0.846	0.842	**0.851**	0.831	0.843
India	0.852	0.849	**0.859**	0.839	0.850
Iran	0.859	0.850	**0.862**	0.841	0.853
Iraq	0.855	0.849	**0.862**	0.840	0.852
Japan	0.865	0.853	**0.868**	0.850	0.859
Roman Empire	0.844	0.844	**0.850**	0.833	0.843

Table 11: BERTScore by place and model. Bold indicates the best model per place; gray shading indicates the lowest score per place. The mean column summarizes overall difficulty per region. Differences across places are significant for all models ($p < 0.001$, Kruskal–Wallis and permutation tests).

-0.337), higher field counts are associated with lower BERTScore within the same region. These regions also show the largest increases in output length per additional field, ranging from 14.9 to 21.0 words per field. This suggests that richer metadata may induce longer or more specialised descriptions that are less similar to the reference under surface metrics. Roman Empire is the main exception ($\rho = 0.040$). In this region, BERTScore increases from 0.840 at six fields to 0.847 at nine fields. Roman Empire is also the only region where output length does not increase with field count ($\Delta = -0.8$ words per field). This is consistent with its positive BERTScore–field-count relationship and suggests that additional metadata is more closely aligned with reference wording in this region.

B.2 Phase 2 Results

Surface metrics performance Table 14 reports surface-metric performance for the three models evaluated across all ten Phase 2 conditions; LLaMA-Vision is excluded because it was evaluated on four conditions only and is therefore not directly comparable. GPT-4o achieves the highest BERTScore (0.884) and METEOR (0.251), while Qwen2.5-32B achieves the highest BLEU-1 score (0.398). BERTScore is compressed across models (range 0.009), consistent with the length confound identified in Phase 1.

Table 15 reports mean BERTScore by region and model. Among the three models evaluated across all ten conditions, GPT-4o obtains the highest regional means in most regions, while Qwen2.5-VL-32B is generally lower. Figure 3 shows per-region field-count–BERTScore curves. Peak field counts differ across regions: British Isles peaks early ($n = 6$), Japan and Iraq peak at intermediate counts ($n = 7$ and $n = 6$, respectively), while China and India show broadly increasing curves peaking at $n = 9$. Greece peaks at $n = 4$–5 and declines at higher field counts, consistent with its weak positive field-count correlation ($r = +0.081$) and stronger negative word-count correlation ($r = -0.514$), suggesting that richer metadata is associated with longer outputs that are less similar to the reference under surface metrics (Table 16. Central America

Region	3	4	5	6	7	8	9
British Isles*	0.863	**0.864**	0.857	0.856	0.852	0.854	–
Central America†	–	0.865	0.865	**0.865**	0.858	–	–
China†	–	–	–	**0.863**	0.860	0.858	–
Egypt†	–	–	–	–	**0.846**	0.846	0.841
Greece*	–	–	**0.853**	0.851	0.845	0.840	0.835
India*	–	0.861	**0.868**	0.862	0.847	0.845	–
Iran*	0.858	**0.859**	0.856	0.856	0.849	0.847	0.850
Iraq*	–	–	**0.859**	0.852	0.852	0.848	0.848
Japan*	0.863	0.860	**0.863**	0.858	0.852	0.849	–
Roman Empire	–	–	–	0.840	0.842	0.843	**0.847**

Table 12: BERTScore by region and number of available metadata fields, averaged across all models. Bold indicates the highest score per region. * scores differ significantly across field counts for all models (Kruskal–Wallis, $p < 0.05$). † significant for at least one model. Roman Empire shows no significant variation across field counts.

Region	Spearman's ρ	Slope
India	-0.411	21.0
Greece	-0.350	20.6
Iran	-0.337	14.9
Egypt	-0.089	14.1
Iraq	-0.155	13.8
British Isles	-0.164	8.8
Japan	-0.295	8.3
Central America	$-0.031^{\dagger}$	6.3
China	-0.092	2.9
Roman Empire	$0.040^{\dagger}$	-0.8

Table 13: Within-region relationships between metadata field count and BERTScore in Phase 1. The correlation column reports Spearman's ρ between field count and BERTScore. The slope column reports the average increase in output length, measured as additional words per metadata field. † not significant at $p < 0.05$; all other correlations are significant.

shows a non-monotonic curve with no clear peak. Word-count correlations are uniformly negative and larger in magnitude than field-count correlations across all regions (overall $r = -0.543$ versus $r = +0.201$), with China ($\rho = -0.711$) and Central America ($r = -0.697$) showing the strongest length–BERTScore associations, suggesting that output length is a stronger predictor of BERTScore variation than metadata quantity alone.

B.2.1 Per-field Analysis

Figure 4 shows primary coverage and bonus correct panels. Primary coverage is high across most field–region combinations (> 0.90), with the lowest values concentrated in *Technique* and *Subjects* for Central America and China. Bonus correct rates are uniformly low, with *Inscription* showing the highest correct recovery across regions, reaching 0.30 for Roman Empire, 0.28 for India, and 0.25 for British Isles, consistent with the aggregate field-level finding that *Inscription* is the only withheld field for which bonus correct exceeds bonus wrong. *Find spot* shows near-zero bonus correct across all regions. Table 17 shows a per-field breakdown across the two field type ladders.

C Qualitative Analysis

C.1 Qualitative Error Patterns

We complement the aggregate claim-level analysis with qualitative inspection of hallucinated and silent field outcomes. The dominant pattern is schema-boundary overreach: models generate claims that are plausible in the artefact domain yet unsupported by the specific catalogue field being evaluated. Table 18 summarises the main error mechanisms.

Place and Find spot. Errors involving *Place* and *Find spot* arise from substitution, over-specification, or redundancy. Models sometimes assign a *Find spot* value to *Place*; for example, an object catalogued with *Place*: *Central America* is described as *Meso-America*, a value present only in *Find spot*. In other cases, models narrow broad place labels using information from *Culture* or *Production date*, such as describing a Greek object as *ancient Greece*. Models also add temporal or contextual qualifiers not supported by the target field: objects catalogued with broad

Model	BLEU-1	ROUGE-L	METEOR	BERTScore
GPT-4o	0.381	0.271	0.251	0.884
LLaMA-Maverick	0.327	0.244	0.213	0.877
Qwen2.5-32B	0.398	0.246	0.247	0.875

Table 14: Phase 2 model-level surface-metric performance averaged across Phase 2 conditions. Values are condition-level means averaged over all probing conditions.

region	GPT-4o	LLaMA-Mav.	LLaMA-Vis.	Qwen2.5	Overall
Central America	**0.9091**	**0.9033**	**0.8769**	0.8823	**0.8957**
China	0.9084	0.8952	0.8762	0.8767	0.8914
British Isles	0.9001	0.8967	0.8749	0.8821	0.8908
Japan	0.8991	0.8947	0.8761	0.8802	0.8895
India	0.8964	0.8924	0.8760	**0.8832**	0.8890
Iraq	0.8978	0.8915	0.8739	0.8800	0.8879
Iran	0.8978	0.8884	0.8724	0.8785	0.8864
Egypt	0.8936	0.8897	0.8727	0.8794	0.8858
Roman Empire	0.8949	0.8906	0.8716	0.8776	0.8858
Greece	0.8864	0.8809	0.8637	0.8690	0.8770

Table 15: Phase 2 mean BERTScore by region and model. Place differences are significant for all models (Kruskal–Wallis, all $p < 0.001$; Bold indicates highest value per column. LLaMA-Vision ran on conditions R, S, S_noi, T only.

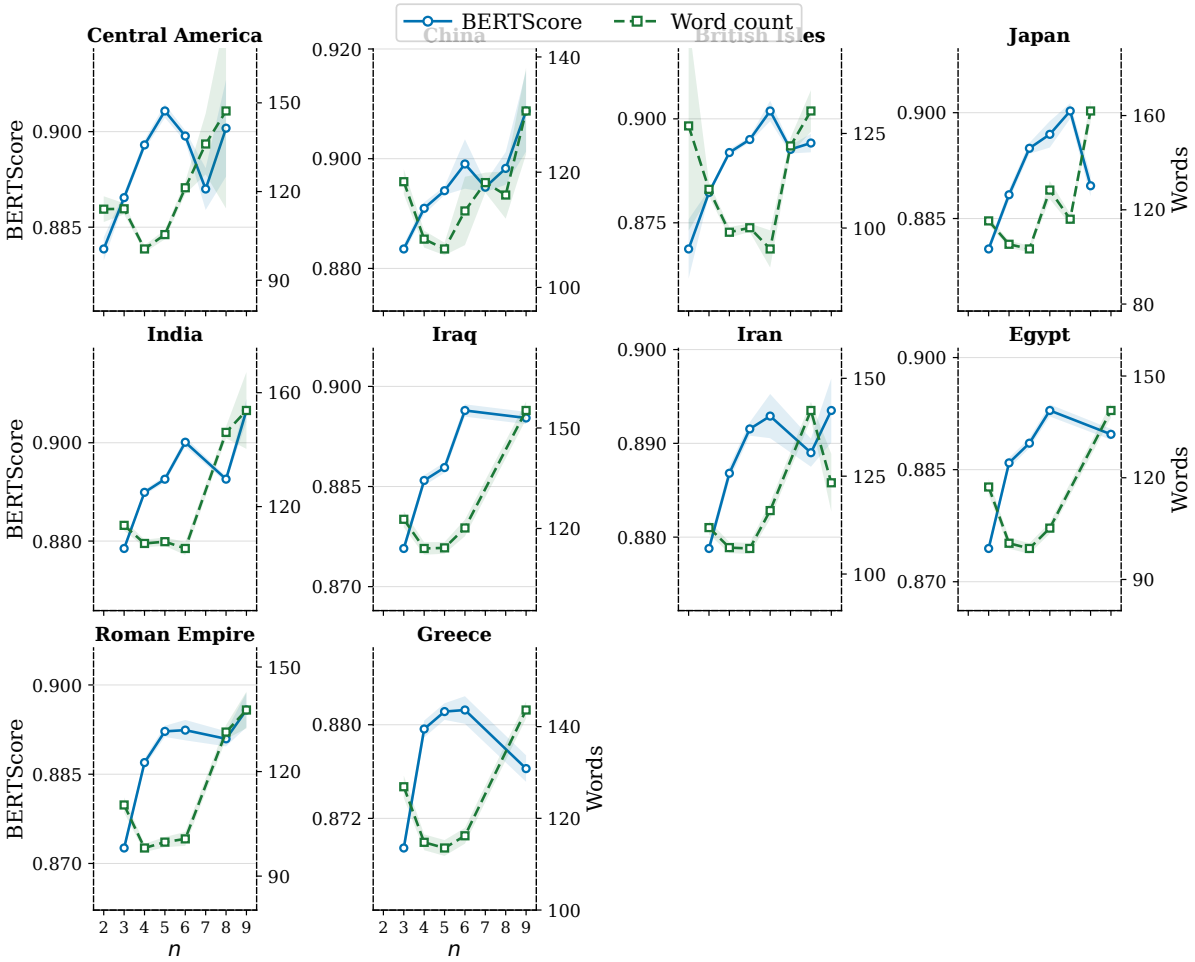

Figure 3: Mean BERTScore and generated word count as a function of observed metadata field count n for each of the ten regions in Phase 2, averaged across all models and conditions.

	Field count (n)		Word count	
region	r***	ρ***	r***	ρ***
Central America	$+0.248$	$+0.257$	-0.697	-0.704
China	$+0.216$	$+0.223$	-0.700	$\mathbf{-0.711}$
British Isles	$+0.223$	$+0.263$	-0.647	-0.618
Japan	$+0.320$	$+0.363$	-0.592	-0.578
India	$+0.359$	$+0.417$	-0.424	-0.398
Iraq	$\mathbf{+0.381}$	$+0.421$	-0.531	-0.512
Iran	$+0.182$	$+0.359$	-0.504	-0.479
Egypt	$+0.320$	$+0.384$	-0.521	-0.486
Roman Empire	$+0.301$	$+0.394$	-0.421	-0.400
Greece	$+0.081$	$+0.200$	-0.514	-0.547
Overall	$+0.201$	$+0.285$	$\mathbf{-0.543}$	-0.544

Table 16: Phase 2 Pearson (r) and Spearman (ρ) correlations between BERTScore and number of metadata fields n (left panel) and generated word count (right panel), computed per region across all conditions and models. All correlations are significant at $p < 0.001$ (***). Bold marks the strongest correlation per column.

	S_CTX arm			S_CONT arm		
Field	Cor	Wro	Sil	Cor	Wro	Sil
Technique	10.6	51.1	38.3	9.8	56.8	33.4
Subjects	17.9	49.7	32.4	—	—	—
Inscription	**27.2**	13.3	59.5	**10.6**	9.4	80.1
Assoc name	2.9	**4.3**	**92.9**	8.1	7.7	84.2
Place	11.2	20.2	68.5	3.7	2.5	93.8
Find spot	—	—	—	0.1	**0.6**	**99.4**

Table 17: Phase 2 bonus field rates split by arm type: the S_CTX arm and S_CONT arm . All values are percentages. Bold indicates best value per column within each arm: highest for Cor and Sil; lowest for Wro. Fields absent from a arm are shown as —.

Place values such as *Egypt* or *India* receive descriptors such as *ancient*, and *Production date* ranges are occasionally extended beyond the recorded value. Conversely, some apparent *Find spot* silence reflects redundant documentation rather than omission: several Chinese records use values such as *Found/Acquired: China*, which add no geographic information beyond the broader *Place* value. These cases illustrate how geographic claims may remain semantically plausible while violating the intended distinction between neighbouring catalogue fields.

Technique. Technique errors reflect over-specification when the field is present and unsupported curatorial elaboration when it is withheld. When *Technique* is provided, models sometimes replace broad catalogue values with more specialised art-historical subtypes; Greek objects catalogued simply as *painted* are described using terms such as *black-figure technique*. When *Technique* is withheld, models generate aesthetic or interpretive descriptions rather than recoverable manufacturing processes: objects with withheld values such as *painted*, *carved*, or *gilded* receive claims about naturalism, craftsmanship, or calligraphic style. These outputs are domain-plausible but not supported by the metadata field being evaluated, explaining the high bonus wrong rates for *Technique*: absent process information invites stylistic elaboration rather than recovery of catalogue-vocabulary manufacturing terms.

Subjects. For *Subjects*, models frequently elaborate beyond the controlled subject label through visual and compositional inference. Controlled-vocabulary subject labels are often sparse relative to the visual complexity of the depicted object; models expand the label using visual priors and domain knowledge drawn from pretraining, generating content that is image-consistent but not catalogue-consistent. Records labelled only as *bird* may receive additional claims about foliage, vine motifs, or decorative scenery, while a record labelled

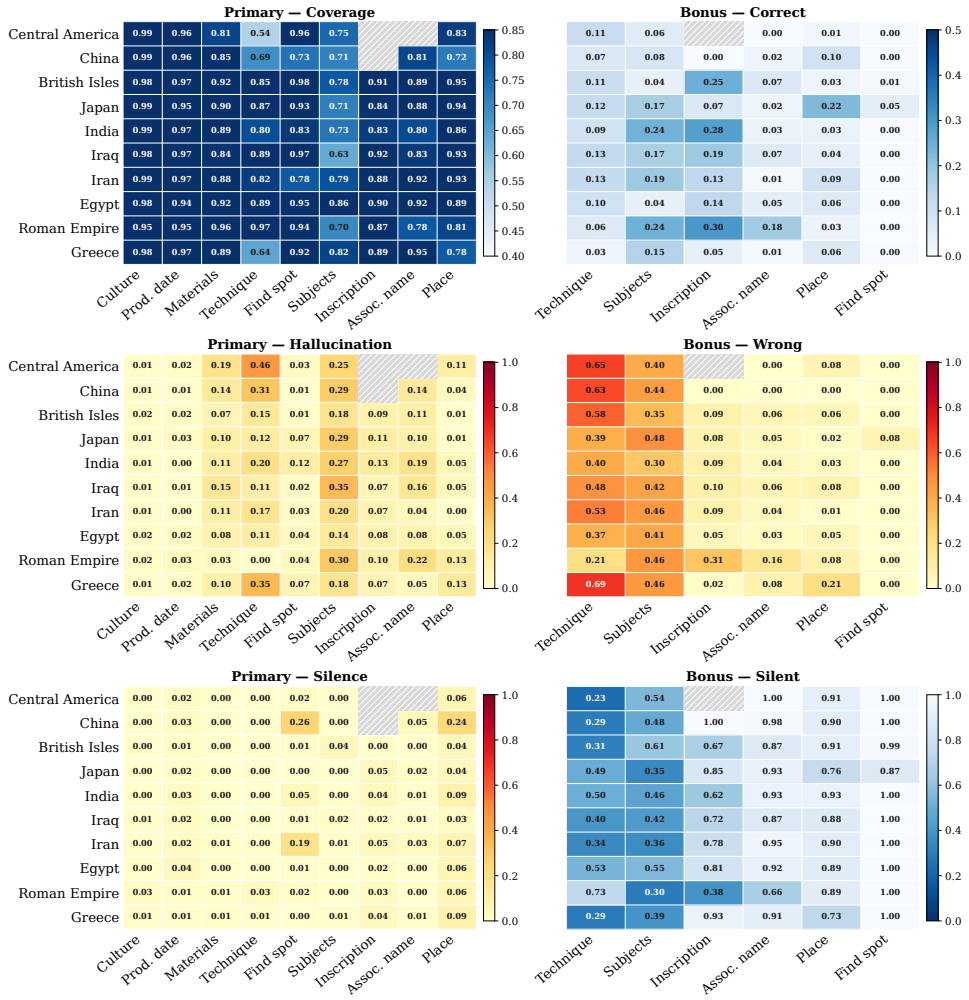

Figure 4: Nugget-verified field outcomes by region and field, averaged across all models and conditions. Each row pairs a primary outcome (field provided, left) with its corresponding bonus outcome (field withheld, right): coverage with bonus correct (top), hallucination with bonus wrong (middle), and silence with bonus silent (bottom). Blue panels indicate outcomes where higher values are better; yellow–red panels indicate outcomes where higher values are worse. Hatched cells indicate fewer than 15 observations.

horse/ass may be described as including a detailed saddle. The resulting claims are not supported by the catalogue subject value but reflect legitimate visual observations of content not captured by the controlled vocabulary.

Materials. Material mentions are comparatively well grounded, consistent with the low primary hallucination rates for *Materials* in the quantitative results; errors here are edge cases rather than a dominant failure mode. Most arise not from incorrect material identification but from unsupported inference triggered by the material value: *pottery* may be realised as *bowl*, and *jadeite* as *high-quality jadeite*. A related pattern is field-boundary confusion: models occasionally assign a value to a neighbouring field, treating *glaze* or *velvet* as a material when the catalogue records them under *Technique*. These cases show how correctly grounded material mentions can induce additional assumptions about object type or quality, and how schema proximity between fields creates an additional source of boundary error.

The dominant failure mode is therefore not arbitrary hallucination, but boundary overreach across neighbouring schema fields, visual priors, and domain-plausible curatorial conventions. This explains why fluent descriptions can appear plausible to a general reader while violating the intended relation between catalogue field, field value, and generated claim.

C.2 Generation Samples

Figure 5 and Figure 6 shows generations samples from GPT4o across condtions and regions.

Error pattern	Main fields	Mechanism	Example pattern
Field substitution	*Place*, *Find spot*	The model transfers a value from one metadata field into another, especially using an excavation site, city, or acquisition location as the broader *Place* value.	*Find spot*: Babylon / Nineveh / Marcianopolis → *Place*.
Geographic over-specification	*Place*	The model narrows a broad catalogue place label into a more specific region, city, or modern location that is not supported by the target field value.	*Central America* → northern Peten / Guatemala; *Roman Empire* → Rome.
Material-to-object inference	*Materials*	The model uses a material label to infer an object type or function, rather than restricting the claim to material composition.	*Earthenware* → vessel; *stone* → stele; *gold* → coin / jewellery.
Field-boundary confusion	*Materials*, *Technique*	The model assigns a value to a neighbouring field, producing a plausible claim that violates the catalogue schema.	*Glaze* or *velvet* treated as material when recorded as technique.
Art-historical over-specification	*Technique*	The model refines a broad catalogue technique into a more specific art-historical technique using domain priors or visual cues.	*Painted* → red-figure / black-figure.
Visual elaboration beyond vocabulary	*Subjects*	The model adds visible motifs, attributes, posture, or iconographic details not present in the controlled subject vocabulary.	*Bird; phoenix* → foliage, vines, secondary motifs.
Bonus technique confabulation	*Technique*	When technique is withheld, the model produces art-critical or stylistic descriptions instead of recovering catalogue technique values.	*Plated* → intricate engraving; *drilled* → smooth and polished.
Temporal/contextual expansion	*Production date*, *Place*	The model adds broad temporal or contextual qualifiers that are plausible but not supported by the metadata field.	*Egypt* → ancient Egypt; date ranges expanded beyond the recorded value.
Documentation redundancy and silence	*Find spot*, *Place*	The model omits a populated field when its value duplicates another field or adds little distinct information.	*Find spot*: Found/Acquired: China omitted after mentioning *Place*: China.

Table 18: Qualitative schema-boundary error patterns identified from hallucinated and silent field-level outcomes. The examples illustrate recurring mechanisms rather than exhaustive error categories.

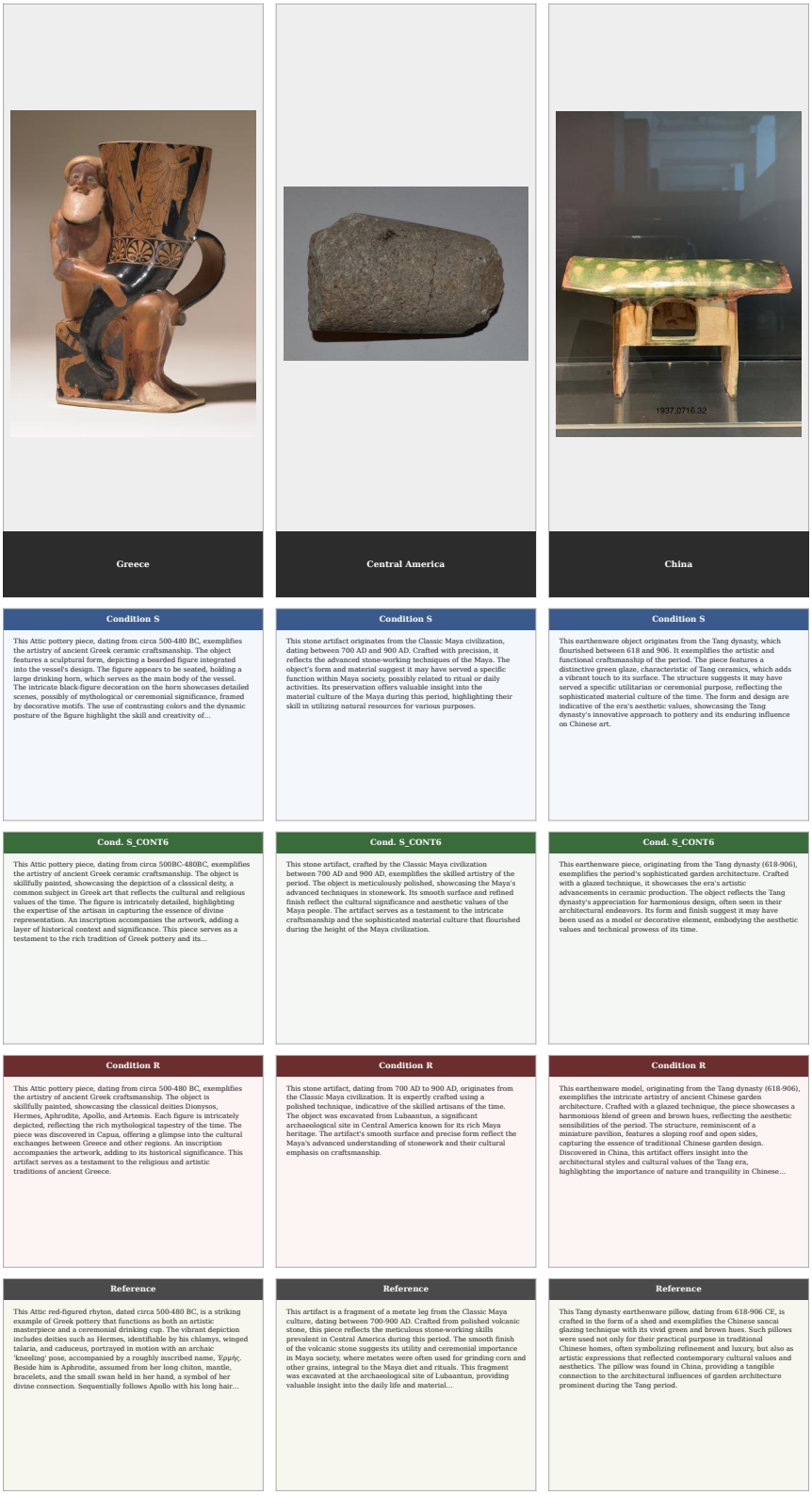

Figure 5: Generated descriptions for three objects across conditions S (sparse floor), S_CONT6 (sparse floor plus *Subjects*, *Technique*, *Inscription*), and R (all available fields with image), alongside the museum reference description. Each row shows the same object; descriptions are truncated to 100 words.

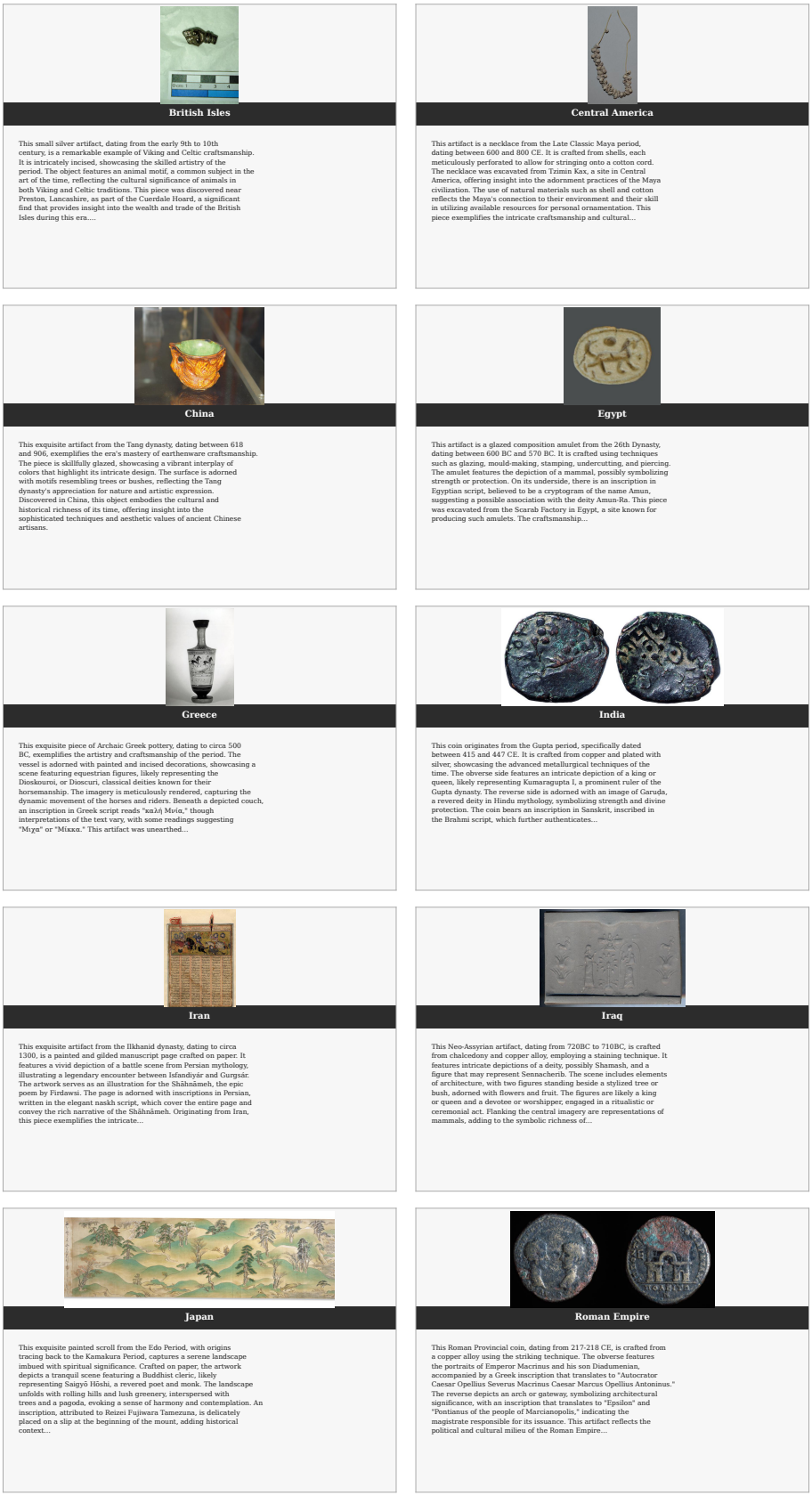

Figure 6: Representative GPT-4o generated descriptions under condition R (all available fields with image) for one object per civilisation. Each cell shows the museum object and the generated description.