

ExhibitionBench: A Multi-Source, Multi-Task Benchmark for Evaluating LLMs as Exhibition Curation Assistants

Mengze Hong^{1,2}, Yanjie Sun¹, Mengyao Zhu¹, Chen Jason Zhang¹, Di Jiang¹
Guoqiong Ivanka Huang¹, Haoyang Li¹, Zhiqiang Kou¹, Shuning Zhang³
Wangze Ni⁴, Enhao Huang⁴, Zeyang Lei², Xia Zeng², Sheng Wang², Di Liang²

¹Hong Kong Polytechnic University, ²Yuanbao Team, Tencent

³Chinese University of Hong Kong, Shenzhen, ⁴Zhejiang University, Hangzhou

Abstract

Cultural reasoning lies at the core of many emerging AI applications, yet existing benchmarks provide limited insight into how well models understand and contextualize cultural artifacts. We introduce EXHIBITIONBENCH, the first benchmark that frames exhibition curation as a high-level evaluation testbed, comprising 300 institution-derived exhibition and thematic collection records, 23,658 artifacts, and three interconnected tasks that evaluate complementary aspects of cultural understanding and reasoning, translating them into user-perceivable utilities. Extensive experiments across 29 systems reveal a significant gap between frontier LLMs and expert-level competence, particularly on tasks that require collection-level coherence and thematic understanding. Performance gains from scaling, prompting, and multimodal inputs are limited and highly model-dependent, whereas tailored model optimizations can equip open-weight models with better capabilities that support on-premise deployment. These findings highlight the difficulty of reasoning over cultural collections and provide a practical benchmark for developing more culturally grounded and reliable AI systems. The full project is released at <https://github.com/mengze-hong/ExhibitionBench>.

1 Introduction

In conjunction with International Museum Day 2026, AI assistants such as Yuanbao and Doubao introduced museum-oriented features to help tourists better understand cultural artifacts. Unlike conventional tourism applications, which primarily focus on attraction search, itinerary planning, and travel question answering (Chen et al., 2026; Huang et al., 2026), museum assistance poses a fundamentally different challenge centered on cultural understanding and reasoning (Nayak et al., 2024). Effective support requires more than retrieving isolated facts:

AI systems must connect artifacts through historical context, stylistic influences, thematic relationships, and cultural significance, while understanding the underlying curatorial narratives. Yet it remains unclear whether LLMs possess the cultural understanding and reasoning capabilities required for museum assistant applications (Li et al., 2025).

Evaluating cultural understanding is challenging due to the scarcity of suitable datasets. Compared to generic knowledge benchmarks (Hong et al., 2025a) or task-oriented benchmarks (Budzianowski et al., 2018), cultural QA queries are rare in dialogue corpora, while large-scale evaluation requires expert-annotated ground truth (Ghaboura et al., 2025). In this paper, we address this gap by shifting to a higher-level perspective, **exhibition curation**, as a natural and demanding testbed. Our underlying hypothesis is that **an AI capable of designing coherent exhibitions can also coherently explain artifacts to users**, thereby serving as an effective museum assistant. Curation requires art-historical knowledge, cultural sensitivity, and narrative logic to assemble objects that inform and engage audiences, posing challenges beyond simple QA tasks and enabling a comprehensive assessment of modern LLMs as knowledgeable museum guides and tools for exhibition design. Our contributions are as follows:

1. **Benchmark Dataset:** We release EXHIBITIONBENCH, a multi-task benchmark for evaluating cultural reasoning in museum artifacts, comprising 300 exhibition and thematic collection records and 23,658 objects drawn from five public cultural-heritage sources. The benchmark defines three tasks: MEIP (object selection), TES (exhibition ranking), and ECD (coherence discrimination), which jointly capture object-level retrieval, thematic collection retrieval, and fine-grained coherence understanding over cultural artifacts.

Dataset / System	Curation	Exhibition	Cult. Bias	Multi-Task	Multi-Src	Open LLM	Open Data
MS MARCO (Bajaj et al., 2018)	✗	✗	✗	✗	✗	✗	✓
Spotify MPD (Chen et al., 2018)	✓	✗	✗	✗	✗	✗	✓
MIND (Wu et al., 2020)	✗	✗	✗	✗	✗	✗	✓
MealRec+ (Li et al., 2024)	✓	✗	✗	✗	✗	✗	✓
CulturalBench (Chiu et al., 2025)	✗	✗	✓	✗	✗	✓	✓
EXHIBITIONBENCH (ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison with five representative retrieval and recommendation benchmarks across seven dimensions. **Curation** denotes evaluation of a coherent set or bundle construction; **Exhibition** denotes exhibition-level evaluation; **Cult. Bias** denotes explicit quantification of cultural bias; **Multi-Src** denotes multiple data sources.

- Systematic Evaluations:** We evaluate retrieval baselines, frontier LLMs, and open-weight LMs under unified zero- and few-shot settings, along with prompting and training-based methods. Results are validated against human curator judgments to ensure alignment with real-world deployment.
- Practical Insights:** Results reveal significant performance gaps. Frontier LLMs underperform strong retrieval baselines on MEIP and exhibit limited thematic coherence, while gains from prompting remain modest and model-dependent. DPO fine-tuning consistently improves smaller models, suggesting promising paths for efficient model selection and tailored optimization.

2 Related Work

Information retrieval. Retrieval and recommendation have primarily focused on ranking individual items by relevance. Benchmarks such as DPR (Karpukhin et al., 2020) and MS-MARCO (Bajaj et al., 2018) evaluate passage retrieval, while playlist continuation (Schedl et al., 2018) and recipe completion (Bosselut et al., 2018) study sequential coherence. Fact-checking has been extensively studied to support tasks requiring strict accuracy and reliability (Sun et al., 2026), while knowledge base augmentation has emerged as another promising direction for improving retrieval accuracy (Hong et al., 2025b). However, existing evaluations and methodologies emphasize item-level relevance or next-item prediction rather than the coherence of an entire collection (see Table 1). In contrast, exhibition curation requires selecting and organizing multiple artifacts into a coherent cultural narrative, emphasizing holistic cultural reasoning over individual item prediction.

Cultural understanding. Prior work has applied deep learning to art analysis (Cetinic et al., 2019)

Source	Records	Objects	Cultural focus
Met	43	1,221	Western, East Asian
Europeana	87	6,283	Cross-European
AIC	91	7,270	American, Western
CMA	33	4,339	East Asian, African
V&A	46	4,545	Japanese, Islamic, South Asian
Total	300	23,658	

Table 2: Data sources and exhibition/thematic-collection record statistics.

and multimodal models to cultural artifacts (Radford et al., 2021). More recently, researchers have examined how LLMs acquire, represent, and reason about cultural knowledge, as well as the biases that emerge across cultural groups (Balaucă et al., 2025; Adilazuarda et al., 2024; Chiu et al., 2025). While these studies primarily focus on the recognition and interpretation of individual cultural artifacts, less attention has been paid to how models reason about relationships among artifacts.

3 EXHIBITIONBENCH

3.1 Data Collection

Figure 1 gives an overview of the benchmark. EXHIBITIONBENCH integrates five public museum and cultural-heritage collections released under source-specific open-access terms, including CC0, CC BY, and Europeana rights statements. For each source, we fetch exhibition records via the public API, normalize them to a shared schema, and retain only exhibitions with at least 5 remaining objects after basic quality checks. We unify per-object metadata fields (such as title, date, culture, department), assign coarse cultural groups by keyword matching, and deduplicate objects by (source, id) across exhibitions. This pipeline yields 300 exhibitions and 23,658 unique objects (Table 2); detailed API usage and filtering rules are documented in Appendix A.

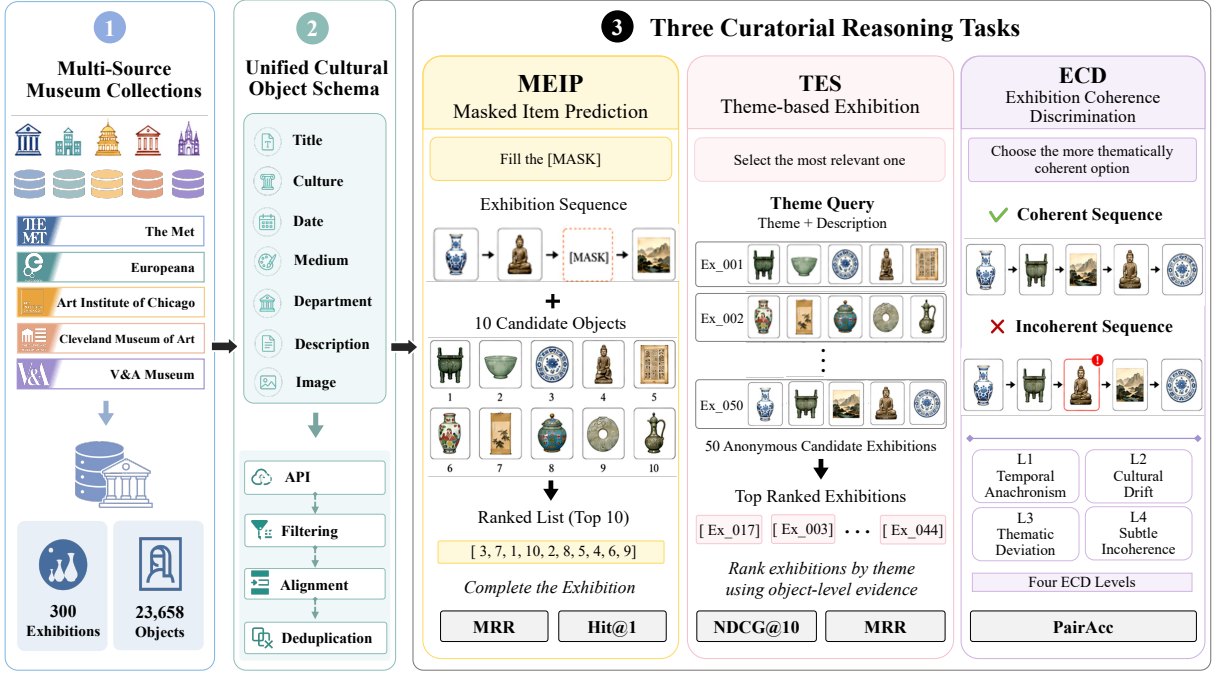


Figure 1: Overview of EXHIBITIONBENCH: five museum APIs are unified into a shared object schema supporting three evaluation tasks (MEIP, TES, ECD).

3.2 Task Definitions

We design EXHIBITIONBENCH to decompose curatorial competence across object, exhibition, and collection-coherence decisions, allowing us to identify *where* systems fail and the implications for museum assistants. Each task is anchored in documented decisions from real exhibitions. Automatic metrics, therefore, measure agreement with professional curator choices as the ground-truth.

Masked Exhibition Item Prediction (MEIP).

Given an exhibition with one object removed, the task is to identify the missing object from 10 candidates. Let $\mathbf{c} = (o_1, \dots, o_{j-1}, o_{j+1}, \dots, o_n)$ denote the remaining context, and $\mathcal{C} = o^* \cup \mathcal{N}$ the candidate set, where o^* is the curator choice and $\mathcal{N}$ contains nine hard negatives. The negatives are top BM25-retrieved objects from other records with respect to theme τ_e , using $\text{title} \oplus \text{description}$. This design tests collection-level fit among topically plausible objects rather than simple lexical relevance. We report Hit@1 and MRR over **1,409 test cases**; the released evaluator additionally computes Hits@ k for $k \in \{3, 5\}$. These metrics quantify recovery of the recorded choice; an unmatched alternative may remain curatorially plausible.

Theme-based Exhibition Selection (TES).

Given a thematic query τ and 50 candidates containing one ground-truth e^* and 49 hard

negatives, systems rank exhibitions by thematic fit. Titles, themes, and descriptions are hidden to prevent direct lexical matching, requiring relevance to be inferred from the exhibited objects. For candidate pool $\mathcal{P} = e_1, \dots, e_{50}$, we report NDCG@10 and MRR over **283 test cases**.

Exhibition Coherence Discrimination (ECD).

Given an authentic exhibition e^+ and a perturbed version e^- , the task is to identify the more curatorially coherent collection. Perturbations have four diagnostic levels: **L1 Temporal Anachronism** (objects over 500 years from the exhibition mean), **L2 Cultural Drift** (different cultural groups), **L3 Thematic Deviation** (topically related but curatorially inconsistent BM25 negatives), and **L4 Subtle Incoherence** (SBERT neighbors with stylistic or provenance inconsistencies). L1, L2, and L4 retain metadata or similarity cues, while L3 controls topical overlap and more directly tests joint curatorial intent. We report per-level and macro pairwise accuracy on **500 test cases (125 per level)**.

Overall, MEIP tests whether an object completes a collection, TES whether a collection expresses a theme, and ECD whether it remains coherent under controlled perturbations. Together, the three tasks assess complementary aspects of collection-level reasoning.

Model	MEIP		TES		ECD				
	Hit@1 ↑	MRR ↑	NDCG@10 ↑	MRR ↑	L1 ↑	L2 ↑	L3 ↑	L4 ↑	Macro ↑
BM25	0.281	0.449	0.347	0.328	0.96	0.98	0.48	0.94	0.838
SBERT	0.780	0.859	0.282	0.255	0.60	0.67	0.47	0.62	0.588
RAG+KG (GPT-5.2)	0.853	0.911	0.282	0.255	0.62	0.67	0.48	0.60	0.593
GPT-5	0.483	0.599	0.077	0.082	0.84	0.82	0.52	0.85	0.756
GPT-5.1	0.502	0.615	0.406	0.255	0.93	0.95	0.58	0.88	0.836
GPT-5.2	0.510	0.619	0.387	0.231	0.84	0.87	0.58	0.81	0.774
Claude Sonnet 4.5	0.453	0.571	0.418	0.266	0.90	0.89	0.54	0.88	0.802
Claude Opus 4.5	0.498	0.609	0.409	0.257	0.92	0.88	0.56	0.83	0.798
Claude Opus 4.6	0.510	0.621	0.437	0.290	0.93	0.94	0.58	0.90	0.836
Gemini 2.5 Pro	0.503	0.615	0.407	0.254	0.94	0.98	0.58	0.94	0.860
Gemini 2.5 Flash	0.429	0.552	0.379	0.334	0.79	0.81	0.54	0.79	0.732
Gemini 3 Flash	0.542	0.645	0.352	0.212	0.65	0.63	0.50	0.68	0.616
Gemini 3.1 Pro	0.589	0.685	0.408	0.283	0.98	0.98	0.58	0.98	0.876
DeepSeek-R1	0.391	0.529	0.343	0.226	0.49	0.49	0.47	0.50	0.486
DeepSeek-V3	0.484	0.598	0.390	0.240	0.82	0.90	0.63	0.85	0.800
DeepSeek-V3.2	0.475	0.594	0.382	0.225	0.62	0.78	0.59	0.75	0.686
Doubao-Seed-1.6	0.506	0.617	0.376	0.225	0.96	0.97	0.57	0.92	0.854
Doubao-Seed-2.0-Lite	0.517	0.626	0.399	0.241	0.96	0.97	0.56	0.94	0.856
Doubao-Seed-2.0-Pro	0.539	0.642	0.410	0.256	0.97	0.98	0.53	0.94	0.852
Kimi-K2.5	0.369	0.506	0.119	0.125	0.49	0.53	0.45	0.53	0.498
GLM-5	0.227	0.392	0.076	0.081	0.64	0.62	0.46	0.64	0.588
MiniMax-M2.5	0.319	0.469	0.169	0.167	0.55	0.53	0.48	0.58	0.536
<i>Open-weight models (local deployment)</i>									
Llama-3.1-8B	0.380	0.498	0.322	0.230	0.50	0.52	0.46	0.54	0.504
Llama-3.1-70B	0.582	0.663	0.437	0.306	0.62	0.66	0.50	0.68	0.618
Llama-3.3-70B	0.574	0.657	0.415	0.273	0.66	0.67	0.52	0.70	0.638
Qwen2.5-7B	0.226	0.375	0.156	0.113	0.42	0.45	0.47	0.60	0.484
Qwen3-8B	0.498	0.595	0.382	0.272	0.55	0.58	0.46	0.61	0.550
Qwen3-14B	0.482	0.581	0.376	0.253	0.59	0.53	0.48	0.59	0.548
Qwen2.5-72B	0.668	0.733	0.391	0.241	0.74	0.74	0.50	0.70	0.674

Table 3: Zero-shot results on all three tasks. **Bold+light amber shading** denote the best proprietary LLM; **bold+light blue shading** denote the best open-weight model. Gray rows indicate non-LLM retrieval baselines.

4 Experimental Setup

We evaluate 29 systems spanning retrieval baselines, proprietary LLMs accessed via API, and locally deployed open-weight LLMs on a server with 8×H20 GPUs. SBERT (Reimers and Gurevych, 2019) uses cosine similarity over contextual embeddings, while RAG+KG augments prompts with CIDOC-CRM triples that encode structured museum knowledge (e.g., artwork–period, artwork–location) for entity grounding (Hu et al., 2025), with GPT-5.2 serving as the inference model. Direct zero-shot LLM evaluations use the same task-specific templates (Appendix C), while RAG+KG adds its documented retrieval context. Temperature is set to 0 where supported. In TES, candidate exhibitions are anonymized as EX_001–EX_050, with titles, themes, and descriptions masked to prevent lexical leakage. In ECD, we mitigate position bias by randomly shuffling the order of positive and negative options.

5 Experimental Results

5.1 MEIP Results

Table 3 shows that MEIP is challenging for LLMs. Among frontier models, Gemini 3.1 Pro performs best (MRR 0.685, Hit@1 0.589), followed by Doubao-Seed-2.0-Pro (MRR 0.642). The Doubao-Seed family consistently outperforms several GPT and Claude variants, suggesting that strong reasoning and coding abilities on general benchmarks do not necessarily translate into better performance on cultural reasoning tasks. Although many LLMs outperform BM25, all evaluated proprietary LLMs remain below the stronger SBERT and RAG+KG baselines on this task, indicating that pretrained knowledge alone is insufficient to reliably select the best-aligned item.

Manual inspections show that most MEIP errors are systematic and fall into two categories. First, models select objects that are semantically similar to the ground truth but lack curatorial coherence

Model	Temp.	Cult.	Style	Sem.O.	Pop.B.
BM25	17.6%	28.0%	1.1%	32.4%	20.9%
SBERT	15.0%	23.3%	0.0%	41.2%	20.5%
GPT-5.2 0-shot	7.2%	18.6%	1.0%	46.0%	27.2%
GPT-5.2 few-shot	9.0%	23.8%	0.8%	40.5%	25.9%
Gemini-2.5-Pro	6.8%	17.9%	0.9%	47.1%	27.3%
RAG+KG	10.6%	24.0%	0.0%	39.8%	25.6%

Table 4: Error type distribution (normalised per model). Sem.O. = Semantic Overlap; Pop.B. = Popularity Bias.

with the exhibition narrative, failing to support its structure, contrasts, or progression. Second, a *popularity bias* emerges, with models favoring famous or visually iconic works over less prominent yet better-fitting alternatives. The error distribution in Table 4 confirms these patterns: semantic overlap is the dominant LLM failure mode (40.5%–47.1%), followed by popularity bias (25.9%–27.3%), while retrieval baselines exhibit more temporal and cultural mismatches. This contrast suggests that LLM errors often arise after identifying a topically plausible object, at the point where collection-level fit must be distinguished from item-level similarity, producing recommendations that appear plausible in isolation but undermine the overall coherence.

5.2 TES Results

On TES, where models must identify the correct exhibition from a thematic query, performance drops sharply relative to MEIP, revealing a major weakness in exhibition-level reasoning. No system exceeds 0.50 on either NDCG@10 or MRR, and several score below 0.20. These results indicate that retrieving individually relevant objects is substantially easier than inferring the unifying curatorial theme of an exhibition. When explicit lexical cues are removed, models struggle to connect objects into a coherent narrative, suggesting limited ability to reason over collections as structured cultural entities rather than sets of related artifacts. This gap has practical implications for recommendation and discovery scenarios, such as “I like this exhibition, can you recommend similar ones”, where users expect systems to preserve thematic intent and curatorial logic.

5.3 ECD Results

For ECD, which distinguishes an original exhibition from a perturbed variant, Gemini 3.1 Pro achieves the strongest overall performance. However, the macro score combines qualitatively different levels. At L1, L2, and L4, lexical and struc-

Model	Western	Asian	Islamic	Δ_{W-NW}
Gemini 2.5 Flash	0.511	0.689	0.511	−0.089
Gemini 2.5 Pro	0.467	0.667	0.556	−0.145
Gemini 3 Flash	0.533	0.533	0.400	+0.067
Gemini 3.1 Pro	0.578	0.711	0.556	−0.055
GPT-5.2	0.489	0.578	0.556	−0.078
Doubao-2.0-Lite	0.511	0.622	0.578	−0.089

Table 5: Cultural variation: Hit@1 by group on MEIP.

tured metadata provide informative signals, allowing BM25 to reach 0.94–0.98 accuracy. L3 instead uses intra-theme distractors that share surface vocabulary with the target exhibition. Here, BM25 obtains 0.48; many proprietary LLMs score only 0.52–0.63, while several others remain at or below the 0.50 random baseline. Thus, ECD diagnoses when surface evidence is sufficient and isolates a harder collection-level coherence problem that remains unsolved.

5.4 Open-weight Models

Among open-weight models, Qwen2.5-72B achieves the strongest overall performance, narrowing the gap to frontier systems on object-level retrieval, though a performance gap remains in coherence-sensitive tasks. On TES, Llama-3.1-70B matches Claude Opus 4.6 with NDCG@10 0.437. In contrast, smaller open-weight models such as Llama-3.1-8B and Qwen2.5-7B perform significantly worse across tasks. An exception is Qwen3-8B, which outperforms several larger proprietary models, suggesting that smaller, deployable models can potentially serve as viable components in on-premise museum applications, particularly when fine-tuned with domain-specific training data (see Section 6).

5.5 Cultural Performance Variation

We measure performance variation across cultural groups using per-group Hit@1 on MEIP, and define Δ_{W-NW} as the difference between Western performance and the weighted mean of Asian and Islamic performance. To ensure comparability, we sample a balanced subset with $n=45$ queries per group. As shown in Table 5, five of the six evaluated models have negative Δ_{W-NW} on this subset, indicating higher average performance on the sampled Asian and Islamic groups than on the Western group. This limited comparison measures performance variation and should not be interpreted as evidence for or against broader cultural bias.

Model	0-shot	1-shot	3-shot	Δ_{best}
GPT-5.2	0.619	0.623	0.624	+0.005
Claude Opus 4.6	0.621	0.629	0.631	+0.010
Gemini 2.5 Pro	0.615	0.635	0.640	+0.025
DeepSeek-V3.2	0.594	0.608	0.601	+0.014
Gemini 3.1 Pro	0.685	0.694	0.685	+0.009

Table 6: MEIP MRR under different few-shot prompt settings. Δ_{best} is the improvement over 0-shot.

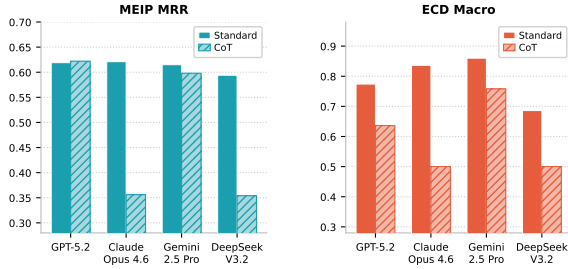


Figure 2: Zero-shot vs. CoT prompting (hatched bar).

Model	Hit@1	MRR	NDCG@10	ECD Macro
Qwen2.5-7B	0.226	0.375	0.156	0.484
+ DPO	0.312	0.448	0.213	0.518
Δ	+0.086	+0.073	+0.057	+0.034
Llama-3.1-8B	0.380	0.498	0.322	0.504
+ DPO	0.461	0.564	0.375	0.536
Δ	+0.081	+0.066	+0.053	+0.032

Table 7: Effect of DPO fine-tuning with LoRA adapters on two open-weight models.

6 Discussion

6.1 Effect of Prompting and Model Training

Few-shot prompting yields modest and highly model-dependent gains on MEIP: Gemini-2.5-Pro benefits most, while GPT-5.2 shows nearly no change. Chain-of-thought prompting provides no consistent benefit, substantially reducing MEIP performance for several models while changing GPT-5.2 only slightly, suggesting that unguided, verbose reasoning is not reliably helpful for cultural retrieval and decision-making in this setting. We further apply DPO with LoRA adapters (rank 64, $\alpha=128$, learning rate 2×10^{-5} , 3 epochs) using 2,400 preference pairs constructed from MEIP. Table 7 shows consistent gains for both Qwen2.5-7B and Llama-3.1-8B across all three tasks, with the Llama model improving by 0.081 Hit@1, 0.066 MRR, 0.053 NDCG@10, and 0.032 ECD Macro. These results provide initial cross-architecture evidence that preference optimization can improve compact open-weight models, although broader validation across model families remains necessary.

Model	MRR			Hit@1		
	Text	+Image	Δ	Text	+Image	Δ
GPT-5.2	0.634	0.617	-0.018	0.528	0.509	-0.018
Gemini 2.5 Pro	0.632	0.648	+0.016	0.524	0.545	+0.021
Gemini 2.5 Flash	0.552	0.592	+0.039	0.429	0.479	+0.050
Doubao Seed 2.0 Pro	0.642	0.639	-0.003	0.539	0.538	-0.002
GPT-5.1	0.615	0.619	+0.004	0.502	0.509	+0.007
DeepSeek-V3.2	0.594	0.571	-0.022	0.475	0.449	-0.027

Table 8: Effect of image inputs on MEIP performance.

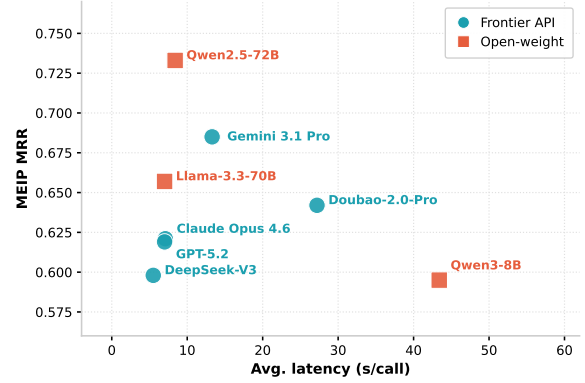


Figure 3: Accuracy–latency trade-off.

6.2 Effect of Image Input

We evaluate whether artwork images improve MEIP performance by comparing paired text-only and text+image inputs for six frontier models on a subset of test instances. As shown in Table 8, image augmentation yields negligible overall gains (average MRR +0.003), with only Gemini 2.5 Flash showing a notable improvement (+0.039). Structured records already encode necessary information such as titles, dates, and cultural origins, leaving images largely complementary when metadata is complete. This suggests that museums with well-maintained catalogs may obtain comparable performance from text-only systems while avoiding the additional cost of vision-enabled inference.

6.3 Deployment Efficiency

Figure 3 reports latency and performance on MEIP. Latency is measured as the average client-side wall-clock time per request, using API endpoints for proprietary models and local inference for open-weight models. The results highlight a clear trade-off between speed and accuracy, with locally deployed Qwen-2.5-72B offering the strongest accuracy–latency trade-off among the evaluated open models. For hosted/API models in the released deployment summary, estimated costs range from USD 0.33 to 28.75 per 1,000 decisions; across the released zero-shot MEIP runs, average request latency ranges from 3.0 to 97.7 seconds. At 16-bit precision, model weights for 7B–8B models

Model	MEIP		TES		ECD	
	Plaus.	Acc.%	Plaus.	Acc.%	Plaus.	Acc.%
GPT-5.2	3.4	58	3.1	51	3.5	61
Claude Opus 4.6	3.7	63	3.4	56	3.8	65
Gemini 3.1 Pro	3.8	67	3.2	53	4.0	69
Human upper bound	4.6	91	4.3	85	4.4	88

Table 9: Human evaluation: mean curatorial plausibility (1 to 5) and majority-vote acceptability (%) per task.

require approximately 14–16 GB of GPU memory, compared with approximately 140 GB for 70B models, excluding runtime overhead. These trade-offs are particularly important for institutions that require on-premise inference to meet data-sovereignty requirements.

6.4 Human Alignment

Three expert museum curators independently evaluated LLM outputs on 50-sample subsets per task, rating curatorial plausibility on a 1–5 Likert scale and deriving binary acceptability via majority vote. Inter-annotator agreement is moderate ($\alpha = 0.61$). On MEIP, the ordering of the three evaluated models by acceptability (58%, 63%, and 67%) matches their ordering by MRR (0.619, 0.621, and 0.685). This supports the reference-based protocol while recognizing that alternatives to the original curator’s choice can remain plausible.

Multiple plausible choices. MEIP and TES use documented curator decisions as references because the primary objective is to measure whether models recover real institutional choices. The metrics, therefore, measure agreement with those decisions rather than the validity of every possible alternative. This distinction is especially important for TES: NDCG@10 evaluates the full ranking against the curator reference, whereas the human-validated subset assesses top-1 plausibility. Frontier models achieve 0.94–1.00 Hit@1 on this subset despite lower full-benchmark NDCG@10, indicating that part of the mismatch arises from the stricter reference-ranking objective rather than uniformly implausible predictions.

To further interpret these results, two additional expert curators completed all three tasks, which we use to estimate an expert upper bound. This upper bound provides a reference point for human-level performance, against which LLM outputs still lag. The gap suggests that, despite progress, current models still fall short of fully replacing expert curatorial judgment in exhibition curation.

Model	C1: Institution split			C2: Title masking		
	Met	Non-M.	Gap	Full	Masked	Drop%
GPT-5.2	0.723	0.847	−0.124	0.669	0.478	28.6%
Claude Opus 4.6	0.744	0.889	−0.145	0.730	0.530	27.4%
Gemini 2.5 Pro	0.741	0.826	−0.085	0.692	0.505	27.0%
Doubao-Seed-2.0	0.736	0.856	−0.120	0.674	0.500	25.9%
DeepSeek-V3.2	0.775	0.728	+0.047	0.640	0.449	29.9%
Gemini 2.5 Flash	0.669	0.649	+0.021	0.576	0.415	27.9%

Table 10: Contamination ablations. **C1:** Met vs. non-Met MRR gap. **C2:** MRR with vs. without object titles.

6.5 Memorization and Surface Cues

Because EXHIBITIONBENCH is constructed from public museum records, we probe potential memorization using an institution split and title masking. Table 10 shows MRR drops of 25.9%–29.9% when object names are masked, indicating substantial reliance on names or associated lexical cues. This analysis does not rule out memorization, but the remaining performance shows that titles do not fully account for model behavior. We therefore treat title masking as a conservative supplementary protocol rather than a definitive contamination test.

7 Conclusion

We introduced EXHIBITIONBENCH, a large-scale benchmark that evaluates cultural understanding through exhibition curation, spanning object retrieval, exhibition selection, and coherence discrimination. Extensive evaluation verifies the performance gaps and motivates continued model improvement and task-specific optimization for downstream museum applications. Frontier LLMs remain substantially below expert-level capability on collection-level reasoning and curatorial coherence. Scaling, prompting, multimodal inputs, and knowledge grounding yield heterogeneous gains, while lightweight adaptation consistently improves two compact open-weight models. Practically, these findings support benchmark-guided model selection that jointly considers capability, latency, cost, and data-sovereignty requirements. They also indicate that current systems are better suited to assisting, rather than replacing, professional curators in cultural design. Because reference-based scores do not capture every plausible choice and title cues explain part of the model performance, EXHIBITIONBENCH should be used as a diagnostic capability test rather than deployment certification. Future work should add institutions and cultural traditions, develop curator-adjudicated evaluation, test multimodal models with sparse metadata, and conduct visitor-facing user studies.

Limitations

The paper has several limitations. First, although preference fine-tuning improves two model architectures, broader validation across model families and scales is needed. Second, reference-based metrics do not exhaustively represent all plausible curator choices, particularly for MEIP and TES. Third, the benchmark is largely text-centric and only indirectly captures visual aspects such as iconography, color, and texture. Fourth, the five public museums and cultural heritage sources do not establish generalizations to institutions with different cataloging practices or to cultural domains outside museums. Finally, the benchmark does not measure visitor experience or downstream behavioral outcomes. Visitor-facing validation requires institutional partnerships, ethics approval, participant recruitment, and in-situ evaluation.

Ethics Statement

All data used in this work are constructed from public museum and cultural heritage sources under source-specific rights statements, including CC0, CC BY, and Europeana item-level terms. Users of the released records remain responsible for complying with the applicable source rights statements. The benchmark compilation is released under CC BY 4.0, and the code under the MIT License. The use of AI assistants was limited to proofreading and language refinement only.

Acknowledgments

This work was partially supported by the NSFC/RGC Joint Research Scheme (N_PolyU5179/25), the Hong Kong RGC (PolyU25600624), the Innovation and Technology Fund (ITS/052/23MX, PRP/009/22FX), industrial sponsors, RMGS (P0045948, P0048183, P0048191, P0046453, P0060272), the HK PolyU Start-up Fund (No. P0059983), the HK PolyU RCEGA Fund (No. P0064421), the HK PolyU RIFL Fund (No. P0061940), and the support of the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (No. 2026C02A1236).

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O’Neill, Ashutosh Modi, and Monojit Choudhury. 2024. [Towards measuring and](#)

[modeling “culture” in LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. [Ms marco: A human generated machine reading comprehension dataset](#). *Preprint*, arXiv:1611.09268.

Ada-Astrid Balauca, Sanjana Garai, Stefan Balauca, Rasesh Udayakumar Shetty, Naitik Agrawal, Dhwanil Subhashbhai Shah, Yuqian Fu, Xi Wang, Kristina Toutanova, Danda Pani Paudel, and Luc Van Gool. 2025. [Understanding museum exhibits using vision-language reasoning](#). In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2227–2238.

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. [Simulating action dynamics with neural process networks](#). *Preprint*, arXiv:1711.05313.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. [MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Eva Cetinic, Tomislav Lipic, and Sonja Grgic. 2019. [A deep learning perspective on beauty, sentiment, and remembrance of art](#). *IEEE Access*, 7:73694–73710.

Ching-Wei Chen, Paul Lamere, Markus Schedl, and Hamed Zamani. 2018. [Recsys challenge 2018: automatic music playlist continuation](#). In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys ’18, page 527–528, New York, NY, USA. Association for Computing Machinery.

Junle Chen, Wei Chen, Yehong Xu, Zhengjun Huang, Yuqian Wu, Zhoujin Tian, Kai Wang, Lei Wang, and Xiaofang Zhou. 2026. [Trip+: Benchmarking agents in personalized interactive travel planning](#). *Preprint*, arXiv:2606.21169.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2025. [CulturalBench: A robust, diverse and challenging benchmark for measuring LMs’ cultural knowledge through human-AI red-teaming](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701, Vienna, Austria. Association for Computational Linguistics.

- 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 LLMs on historical and cultural artifacts](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23627–23641, Vienna, Austria. Association for Computational Linguistics.
- Mengze Hong, Wailing Ng, Chen Jason Zhang, and Di Jiang. 2025a. [QualBench: Benchmarking Chinese LLMs with localized professional qualifications for vertical domain evaluation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5938–5953, Suzhou, China. Association for Computational Linguistics.
- Mengze Hong, Chen Jason Zhang, Di Jiang, and Yuanqin He. 2025b. [Augmenting compliance-guaranteed customer service chatbots: Context-aware knowledge expansion with large language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 753–765, Suzhou (China). Association for Computational Linguistics.
- Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiao-Yong Wei, and Li Qing. 2025. [Removal of hallucination on hallucination: Debate-augmented RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.
- GuoQiong Ivanka Huang, Chen Jason Zhang, Wanyi Christina Zhou Torres, Xiling Xiong, and Huahang Li. 2026. [Responsible ai and human collaboration in tourism management: Ethical considerations and identity disclosure](#). *Tourism Management*, 113:105315.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- Ming Li, Lin Li, Xiaohui Tao, and Jimmy Xiangji Huang. 2024. [Mealrec+: A meal recommendation dataset with meal-course affiliation for personalization and healthiness](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, page 564–574, New York, NY, USA. Association for Computing Machinery.
- Yifan Li, Jun Meng, and Qingqiang Wu. 2025. [Talking to the mona lisa? llm-based role-playing agent enables artworks in virtual exhibitions to speak](#). In *2025 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.
- Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stanczak, and Aishwarya Agrawal. 2024. [Benchmarking vision language models for cultural understanding](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5769–5790, Miami, Florida, USA. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Markus Schedl, Hamed Zamani, Ching-Wei Chen, Yashar Deldjoo, and Mehdi Elahi. 2018. Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 7(2):95–116.
- Yuxi Sun, Wenbo Shang, Wei Gao, Xin Huang, and Jing Ma. 2026. [Diagnosing llm arbitration behavior over pre-evidence epistemic states in rag-based fact-checking](#). *Preprint*, arXiv:2606.01120.
- Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. [MIND: A large-scale dataset for news recommendation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online. Association for Computational Linguistics.

A Data Collection Algorithm

Algorithm 1 summarizes the data collection and task-specific filtering procedure used to construct the final aligned benchmark dataset from institutional exhibition records, thematic collection metadata, and associated object metadata.

B Evaluation Metrics

This appendix defines the evaluation metrics used across the three benchmark tasks.

Algorithm 1 Data collection and filtering

Require: Institutional exhibition records $\mathcal{E}$, thematic collection metadata $\mathcal{T}$, object metadata APIs/datasets

Ensure: Aligned benchmark dataset $\mathcal{D}$

```
1:  $\mathcal{D} \leftarrow \emptyset$ 
2: for all source  $s \in \{\text{Met, Europeana, AIC, CMA, V\&A}\}$  do
3:   Normalize institutional records  $\mathcal{E}_s$  and thematic records  $\mathcal{T}_s$ 
4:   for all record  $e \in \mathcal{E}_s \cup \mathcal{T}_s$  do
5:     Fetch or link object metadata for items associated with  $e$ 
6:     for all object  $o$  associated with  $e$  do
7:       if  $o$  lacks a usable identifier or title then
8:         Remove  $o$  from  $e$ 
9:       end if
10:    end for
11:    Add normalized record  $e$  and its valid objects to  $\mathcal{D}$ 
12:  end for
13: end for
14: Deduplicate objects by (source, id) across records
15: Apply task-specific eligibility and context-size filters when constructing MEIP, TES, and ECD
16: return  $\mathcal{D}$ 
```

MEIP Metrics. For each test case, the model selects one object from the 10 candidates. To preserve the released evaluation pipeline used for the reported results, the selected object is placed first, and the unselected candidates retain their released input order; let r_i denote the resulting position of the ground-truth object in case i . Thus, the prompt is a single-choice interface, while the metrics below are computed from the deterministic ranking induced by that choice. Successfully parsed predictions are scored without changing the released candidate order or any reported values.

1. **Hit@1:** proportion of cases where the ground-truth object is ranked first,

$$\text{Hit@1} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(r_i = 1).$$

2. **Hits@ k :** whether the ground-truth object appears within the top- k positions,

$$\text{Hits@}k = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(r_i \leq k), \quad k \in \{3, 5\}.$$

3. **MRR:** mean reciprocal rank of the ground-truth object,

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i}.$$

TES Metrics. For each thematic query, the model returns the top 10 anonymous exhibition IDs from a pool of 50 candidates. The released parser preserves the returned order and appends any unmentioned candidates in their released input order, yielding a deterministic complete ranking for scoring; unparseable responses are not scored. Each released case has one ground-truth exhibition. Let $\text{rel}_j \in \{0, 1\}$ indicate whether the exhibition at rank position j is that reference, and let r_i be its position in the completed ranking.

1. **Precision@ k ($\mathbf{P@}k$):**

$$\mathbf{P@}k = \frac{1}{k} \sum_{j=1}^k \text{rel}_j.$$

2. **Hits@1:** proportion of queries where the relevant exhibition is ranked first,

$$\text{Hits@1} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(r_i = 1),$$

where r_i is the rank of the relevant exhibition.

3. **NDCG@10:**

$$\text{DCG@10} = \sum_{j=1}^{10} \frac{\text{rel}_j}{\log_2(j+1)},$$
$$\text{NDCG@10} = \frac{\text{DCG@10}}{\text{IDCG@10}}.$$

4. **MRR** (reported by the released evaluator):

$$\text{MRR} = \frac{1}{N} \sum_{i=1}^N \frac{1}{r_i},$$

This documents the existing computation and does not alter any reported result.

ECD Metrics. The ECD task is a pairwise binary classification problem with randomized option order, ground-truth label y_i , and prediction $\hat{y}_i$; the released set contains 125 pairs at each difficulty level.

1. **Accuracy per difficulty level:**

$$\text{Acc}_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \mathbf{1}(y_i = \hat{y}_i).$$

2. **Macro Accuracy:**

$$\text{MacroAcc} = \frac{1}{4} \sum_{\ell=1}^4 \text{Acc}_\ell.$$

Macro averaging prevents performance on easier levels from dominating the overall score.

C Prompt Templates

C.1 MEIP Zero-Shot Prompt

You are assisting a museum curator. Given an exhibition theme and some context objects already selected, identify which ONE candidate object best fits the exhibition and should be added next.

Exhibition theme: {theme}
Context objects already in exhibition: {context_objects}
Candidate objects (choose the best fit): {candidate_list}
Reply with ONLY the ID of the best-fitting candidate object (e.g., met_123456).

C.2 TES Zero-Shot Prompt (anonymized, leak-free)

You are a museum curator selecting an exhibition. Given a thematic query, rank the provided exhibition candidates from most to least relevant.

Query theme: {theme}
Query description: {description}
Exhibition candidates (each identified by an anonymous ID; judge solely by the artworks inside): {candidate_list}
Return ONLY the anonymous IDs of the top 10 most relevant exhibitions, in order from best to worst.
Format: EX_001, EX_002, EX_003, ... (comma-separated)

C.3 ECD Zero-Shot Prompt

You are a museum curator evaluating exhibition coherence.

Exhibition theme: {theme}
Two exhibition sequences are shown below. One is the ORIGINAL coherent sequence; the other has been DISRUPTED by swapping one item.
Sequence A: {sequence_a}
Sequence B: {sequence_b}
Which sequence is more coherent and fits the exhibition theme better?
Reply with ONLY "A" or "B".

D Additional Ablations

Metadata and input design. Ablations over six cumulative metadata configurations (see Figure 4) show that adding department and description fields produces the largest average incremental gains across the four evaluated models. Adding fields is not uniformly beneficial at every step: for example, medium slightly hurts some models. Using titles alone already recovers most of the performance achieved with full metadata for models such as GPT-5.2.

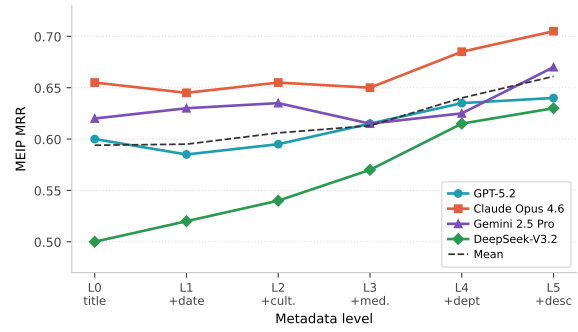


Figure 4: Effect of metadata richness on MEIP MRR. Each level adds one field: L0 title only, L1 adds date, L2 culture, L3 medium, L4 department, L5 description.

Model	0-shot	1-shot	3-shot
Claude Opus 4.6	0.437	0.438	0.436
Claude Sonnet 4.5	0.418	0.414	0.408
Claude Opus 4.5	0.409	0.409	0.409
Gemini 2.5 Flash	0.379	0.428	0.393
Gemini 2.5 Pro	0.407	0.417	0.413
Gemini 3 Pro Preview	0.405	0.415	0.396
Gemini 3 Flash Preview	0.352	0.352	0.353
Gemini 3.1 Pro Preview	0.408	0.402	0.404
GPT-5.1	0.406	0.399	0.406
GPT-5.2	0.387	0.388	0.389
GPT-5	0.077	0.078	0.076
DeepSeek-R1	0.343	0.394	0.387
DeepSeek-V3.2	0.382	0.391	0.385
DeepSeek-V3	0.390	0.373	0.390
Doubao-Seed-2.0-Pro	0.410	0.400	0.403
Doubao-Seed-2.0-Lite	0.399	0.390	0.398
Doubao-Seed-1.6	0.376	0.389	0.385
Kimi-K2.5	0.119	0.109	0.112
GLM-5	0.076	0.148	0.161
MiniMax-M2.5	0.169	0.197	0.181

Table 11: TES NDCG@10 (leak-free protocol) across shot counts. Bold = best per model.

E TES Few-Shot Results (20 Models with Available Runs)

Table 11 reports NDCG@10 for the 20 models with available 0-shot, 1-shot, and 3-shot runs under the anonymized, leak-free TES protocol. This auxiliary set is defined by few-shot result availability and is not identical to the zero-shot main-table model list. Effects are highly model-dependent: several models change by at most about 2 pp, whereas Gemini 2.5 Flash, DeepSeek-R1, GLM-5, and MiniMax-M2.5 improve by approximately 2.8–8.5 pp at their best setting. Kimi-K2.5 declines slightly, showing that examples do not provide consistent gains in the 50-way ranking format.

Task	Case	Test Input Snapshot	Model Input (Prompt Core)	Gold		Logged Model Output	Evaluation / What It Shows
MEIP	M1 (good) meip_0016	Theme: Dutch Masters Context starts with <i>Interior of the Sint-Pieterskerk..., A Woman Seated at a Window</i>	Single-step completion: choose 1 ID from 10 candidates	met_437877		status=ok, raw=met_437877, parsed=met_437877	Exact-match correct. Demonstrates strict ID-based scoring path.
MEIP	M2 (bad) meip_0018	Theme: Dutch Masters Context starts with <i>The Golden Age, Kitchen Scene</i>	Same 10-way completion format	met_436085		status=invalid_output, raw=met_436524, parsed=met_436524	Wrong ID + parser-level invalidation recorded. Shows output-format and ID-fidelity failure.
TES	T1 (good)	Theme: Impressionism 50 candidate exhibitions	Rank anonymous exhibitions EX_001–EX_050 from their artworks; return top 10	Reference exhibition: EX_002	exhibi-	Valid illustrative output begins: EX_002, EX_006, EX_017, ... (10 unique IDs total)	Reference ranked first. Shows the final exhibition-ranking interface and anonymous-ID scoring path.
TES	T2 (bad)	Theme: Dutch Masters 50 candidate exhibitions	Same anonymous top-10 exhibition-ranking format	Reference exhibition: EX_022	exhibi-	Illustrative output omits EX_022 from its top 10	Reference not recovered in the returned list. Shows a thematic-ranking miss under the final protocol.
ECD	E1 (good)	Theme: Indian Textiles and Decorative Arts Perturbation level: L3	Binary discrimination: choose A or B	A		status=ok, raw=A, parsed=A, is_correct=True	Correct pairwise decision. Shows a correct decision on a hard theme-matched perturbation.
ECD	E2 (bad)	Theme: Nordic Mythology Art Perturbation level: L2	Same A/B coherence choice	B		status=ok, raw=A, parsed=A, is_correct=False	Wrong binary choice. Shows local-coherence trap under cultural-drift perturbation.

Table 12: Case studies and protocol examples. MEIP and ECD rows report logged model inputs and outputs, while TES rows illustrate the released 50-candidate anonymous-exhibition ranking protocol without introducing additional experimental results.

F Case Studies

The MEIP and ECD cases below use real sample IDs and logged model outputs. The TES rows use released final-protocol samples to illustrate the anonymous 50-exhibition ranking interface and valid top-10 output format; they are protocol examples rather than additional quantitative results.