Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts

Sara Ghaboura[1†] Ketan More[1†] Ritesh Thawkar[1] Wafa Alghallabi [1] Omkar Thawakar[1]
Fahad Shahbaz Khan[1,2] Hisham Cholakkal[1] Salman Khan[1,3] Rao Muhammad Anwer[1,4]
[1]Mohamed bin Zayed University of AI, [2]Linköping University, [3]Australian National University,
[4]Aalto University
{sara.ghaboura, ketan.more, omkar.thawakar}@mbzuai.ac.ae

https://mbzuai-oryx.github.io/TimeTravel/

Abstract

Understanding historical and cultural artifacts demands human expertise and advanced computational techniques, yet the process remains complex and time-intensive. While large multimodal models offer promising support, their evaluation and improvement require a standardized benchmark. To address this, we introduce *TimeTravel*, a benchmark of 10,250 expert-verified samples spanning 266 distinct cultures across 10 major historical regions. Designed for AI-driven analysis of manuscripts, artworks, inscriptions, and archaeological discoveries, TimeTravel provides a structured dataset and robust evaluation framework to assess AI models' capabilities in classification, interpretation, and historical comprehension. By integrating AI with historical research, TimeTravel fosters AI-powered tools for historians, archaeologists, researchers, and cultural tourists to extract valuable insights while ensuring technology contributes meaningfully to historical discovery and cultural heritage preservation. We evaluate contemporary AI models on TimeTravel, highlighting their strengths and identifying areas for improvement. Our goal is to establish AI as a reliable partner in preserving cultural heritage, ensuring that technological advancements contribute meaningfully to historical discovery. We release the *TimeTravel* dataset[1] and evaluation suite[2] as open-source resources for culturally and historically informed research.

1 Introduction

In recent years, Large Multimodal Models (LMMs) have made significant strides in visual reasoning, perception, and multimodal understanding. Models such as GPT-4V (OpenAI, 2024) and LLaVA (Liu et al., 2023) have excelled in image captioning, visual question answering (VQA), and complex visual reasoning, driving the development of benchmarks (Chiu et al., 2024; Nayak et al., 2024; Alwajih et al., 2024) to assess their capabilities. These benchmarks predominantly focus on modern objects, cultural landmarks, and textual sources, extending multimodal AI applications to domains such as medical imaging, remote sensing, and real-world scene understanding (Ghaboura et al., 2025). However, a critical gap remains: LMMs fail to address the historical dimension of visual data, particularly artifacts that have shaped human civilization.

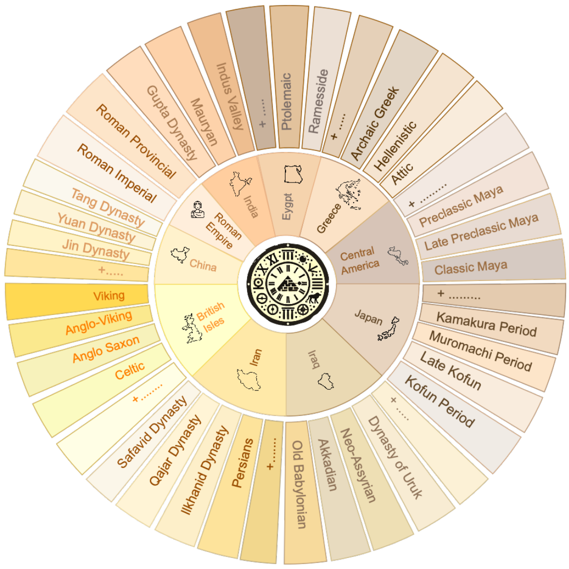

Figure 1: TimeTravel Taxonomy categorizes artifacts from 10 major civilizations, representing diverse historical and prehistoric periods. It encompasses 266 distinct cultures and over 10k manually verified historical artifact samples, providing a structured framework for comprehensive AI-driven analysis.

[1]https://huggingface.co/datasets/MBZUAI/TimeTravel
[2]https://github.com/mbzuai-oryx/TimeTravel
[†]Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2025, pages 23627–23641
July 27 - August 1, 2025

Figure 2: **TimeTravel Samples.** The figure shows diverse cultural representations from various regions across the globe. These examples span multiple artifact categories, including coins, accessories, tools, and statues from ancient civilizations. Each artifact is accompanied by a detailed description, providing valuable contextual and historical insights. Additional TimeTravel examples can be found in the Appendix (Figure 7 and Figure 8).

Historical artifacts, from ancient manuscripts and inscriptions to architectural ruins and cultural symbols, offer invaluable insights into the evolution of societies, artistic expression, and technological advancements. These artifacts preserve cultural heritage and serve as primary sources for understanding belief systems, trade networks, and sociopolitical structures of past civilizations. However, interpreting them requires deep contextual knowledge, which current LMMs struggle to achieve, particularly in non-English and non-Western historical contexts. While some models have been extended to low-resource languages to bridge cultural gaps (Heakl et al., 2025), they lack systematic capabilities to analyze artifacts from diverse civilizations. This limitation highlights the urgent need for a specialized benchmark that evaluates the ability of AI to process and understand historical artifacts with cultural and temporal awareness.
To address this challenge, we introduce TimeTravel, an open-source comprehensive benchmark (see Table 1) for evaluating LMM performance in historical artifact analysis across diverse civilizations. TimeTravel encompasses several major ancient and prehistoric civilizations across 10 distinct regions, spanning 266 cultural groups. It offers a structured taxonomy tailored for machine-assisted historical research (see Figure 1). Unlike existing benchmarks that focus on generic object recognition, TimeTravel prioritizes historical knowledge, contextual reasoning, and cultural preservation, making it a pioneering effort in multimodal computational evaluation. The benchmark consists of over 10k curated samples, each accompanied by high-quality images of manuscripts, inscriptions, sculptures, paintings, and archaeological discoveries. These samples assess key aspects of multimodal understanding, including visual perception, contextual reasoning, and cross-civilizational knowledge. Having been thoroughly verified by historians and archaeologists, the dataset ensures accuracy, cultural relevance, and historical integrity. By evaluating both closed- and open-source LMMs on TimeTravel, we aim to identify their strengths and limitations in handling historically significant artifacts, paving the way for intelligent models that contribute meaningfully to cultural heritage preservation and historical analysis.

Domain	British Museum	MMMU	Oracle-MNIST	Ithaca	Kao Kore	HUST-OBS	TimeTravel (ours)
Hist. Artifact Recog.	✓	✗	✗	✗	✓	✗	✓
Geographic Region	✓	✗	✗	✓	✓	✗	✓
Ancient Artifacts	✓	✗	✗	✗	✗	✗	✓
Contextual History	✗	✗	✗	✗	✗	✗	✓
Image-Text Pairs	✓	✓	✗	✗	✓	✓	✓
Open-Source	✗	✓	✓	✗	✓	✓	✓

Table 1: The comparison of datasets and benchmarks for historical and cultural artifacts, evaluating features like **artifact recognition**, **geographic coverage**, **multimodal understanding**, and **metadata inclusion** with existing data such as British Museum (Tully, 2020), MMMU (Yue et al., 2024), Oracle-MNIST (Wang and Deng, 2022), Ithaca (Assael et al., 2022), KaoKore (Tian et al., 2020), HUST-OBS (Wang et al., 2024). TimeTravel stands out as the most comprehensive benchmark, uniquely integrating multimodal data, historical context, and a dedicated focus on ancient artifacts to support AI-driven cultural heritage research.

2 The TimeTravel Dataset

2.1 Data Collection

Our research is based on a well-structured and carefully curated dataset sourced from museum collections, which houses an extensive collection of artifacts from diverse civilizations. From this vast repository, we compiled a dataset encompassing 266 cultural groups, allowing the analysis of cultural, technological, and social developments over a broad historical timeline.

To ensure the integrity of our benchmark, we followed a systematic data collection process. We first identified key civilizations and historical periods relevant to our study, then collaborated closely with

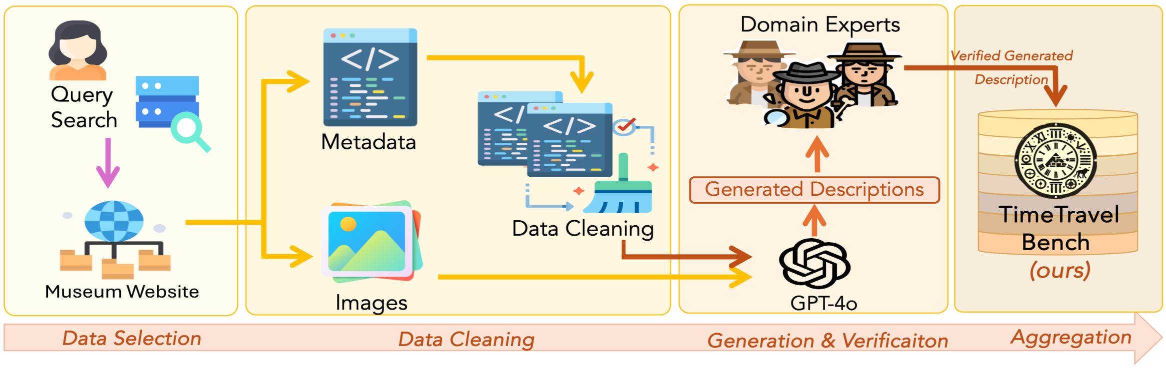

Figure 3: **TimeTravel Data Pipeline.** A structured workflow that collects image and text data from museum websites, cleans metadata, and integrates it with visual content. The GPT-4o model generates detailed, context-aware descriptions, which are refined by experts for accuracy before forming the TimeTravel Benchmark.

experts to validate the authenticity and completeness of each record. The resulting dataset comprises 10,250 carefully curated samples (see Fig 2). Each entry, ranging from artifacts and inscriptions to ancient manuscripts, was meticulously verified by historians and archaeologists, ensuring precision and reliability. By incorporating data from multiple civilizations, our benchmark provides a diverse and comprehensive perspective, avoiding the limitations of a single historical narrative while preserving the historical context for in-depth analysis. This approach enables the study of significant patterns in the evolution of human societies. Further details on data composition and distribution are provided in the Appendix (Sec C).

2.2 Image-Text pair Generation

The dataset features a diverse range of historical objects, ensuring comprehensive documentation and contextual understanding. However, many metadata fields, such as title, iconography, and date, were missing or incomplete. To address this, we used GPT-4o to generate detailed textual descriptions with context-sensitive information based on the available metadata (see Figure 5 and 6). To further enhance usability, we structured these descriptions into image-text pairs, ensuring that each artifact is not only visually documented but also enriched with contextual and cultural insights. By improving multimodal model compatibility and supporting digital archiving, this approach strengthens research in cultural heritage preservation while bridging gaps in existing records. More samples are presented in the Appendix (Sec. E).

2.3 Data Filtering and Verification

To guarantee the accuracy and reliability of our dataset, we implemented a rigorous data filtering and verification process (Figure 3). This process combined manual expert validation with automated techniques to eliminate inconsistencies, fill in missing details where possible, and authenticate historical records. During data cleaning, we addressed missing or incomplete metadata, such as titles, dates, and iconography, by cross-referencing museum archives, academic sources, and expert insights. Unavailable key information was transparently documented. Additionally, automated checks identified formatting inconsistencies, metadata mapping errors, and numerical anomalies, ensuring a structured and standardized dataset. For verification, we collaborated with historians, archaeologists, and museum curators to review each artifact's description, cultural attribution, and historical significance. Expert validation ensured that generated textual descriptions were accurate, contextually relevant, and aligned with historical records. This rigorous process enhances the dataset's credibility, making it a valuable resource for historical research, machine learning, and cultural heritage preservation while ensuring reliable insights into human history. More details on annotation and verification process can be found in the Appendix (Sec. D).

3 TimeTravel Benchmark Evaluation

Evaluation Metric: To assess the quality, accuracy and relevance of our generated textual descriptions, we employed a combination of traditional and advanced metrics. BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004) evaluate linguistic fluency and structural similarity, ensuring syntactic alignment with reference texts. METEOR (Banerjee and Lavie, 2005) enhances this by incorporating synonym matching and paraphrasing, improving the adaptability to human variations. SPICE (Anderson et al., 2016) assesses semantic accuracy

	Model	BLEU	METEOR	ROUGE-L	SPICE	BERTScore	LLM-Judge
Closed	GPT-4o-0806 (OpenAI, 2024)	**0.0190**	0.2399	**0.1740**	**0.0960**	**0.8482**	**0.3007**
	Gemini-2.0-Flash (Reid et al., 2024)	0.0120	0.2424	0.1470	0.0650	0.8180	0.2726
	Gemini-1.5-Pro (Reid et al., 2024)	0.0110	0.2469	0.1570	0.0770	0.8311	0.2303
	GPT-4o-mini-0718 (OpenAI, 2024)	0.0150	**0.2664**	0.1690	0.0930	0.8432	0.2500
Open	Llama-3.2-Vision-Inst (Meta AI, 2024)	0.0100	0.2069	0.1610	0.0590	0.8152	0.1226
	Qwen-2.5-VL (Team, 2025)	0.0140	0.2649	0.1630	0.0770	0.8379	0.1749
	Llava-Next (Liu et al., 2024)	0.0120	0.2353	0.1620	0.0650	0.8357	0.1156

Table 2: Performance comparison of various closed and open-source models on our proposed TimeTravel benchmark.

	Model	India	Roman Empire	China	British Isles	Iran	Iraq	Japan	Central America	Greece	Egypt
Closed	GPT-4o-0806	**0.2376**	**0.4486**	**0.2455**	**0.1862**	**0.3552**	**0.3560**	**0.2223**	**0.3187**	**0.2756**	0.3665
	Gemini-2.0-Flash	0.1876	0.3521	0.2447	0.1671	0.3409	0.3190	0.2123	0.2966	0.2369	**0.3849**
	Gemini-1.5-Pro	0.1407	0.2871	0.2372	0.1713	0.3078	0.2640	0.1727	0.2385	0.2042	0.2822
	GPT-4o-mini-0718	0.2308	0.3636	0.2151	0.1874	0.3072	0.2656	0.2131	0.3147	0.2070	0.2552
Open	Llama-3.2-Vision-Inst	0.0722	0.1429	0.1195	0.0779	0.1984	0.1107	0.1059	0.1549	0.1311	0.1131
	Qwen-2.5-VL	0.0859	0.1664	0.2149	0.1190	0.2344	0.2127	0.1607	0.2125	0.1417	0.2315
	Llava-Next	0.0796	0.1062	0.1332	0.1141	0.1624	0.1039	0.1129	0.1799	0.1220	0.0662

Table 3: Analysis of LLM-Judge evaluation of various models in describing archaeological artifacts across civilizations from different geographical locations.

	Model	India	Roman Empire	China	British Isles	Iran	Iraq	Japan	Central America	Greece	Egypt
Closed	GPT-4o-0806	0.2566	0.2713	0.2324	0.2175	0.2486	0.2428	0.2269	0.2384	0.2441	0.2567
	Gemini-2.0-Flash	0.2494	0.2644	0.2203	0.2202	0.2471	0.2413	0.2239	0.2251	0.2526	0.2605
	Gemini-1.5-Pro	0.2596	0.2635	0.2219	0.2237	0.2547	0.2516	0.2247	0.2253	0.2569	0.2656
	GPT-4o-mini-0718	**0.2762**	0.2731	**0.2570**	**0.2531**	**0.2660**	**0.2640**	**0.2611**	**0.2741**	0.2649	**0.2742**
Open	Llama-3.2-Vision-Inst	0.2116	0.2264	0.1894	0.1930	0.2132	0.2083	0.1955	0.2000	0.2139	0.2181
	Qwen-2.5-VL	0.2742	**0.2845**	0.2520	0.2456	0.2638	0.2621	0.2547	0.2659	**0.2695**	0.2731
	Llava-Next	0.2487	0.2512	0.2181	0.2189	0.2409	0.2344	0.2208	0.2247	0.2411	0.2440

Table 4: Analysis of METEOR Evaluation of various models in describing archaeological artifacts across civilizations from different geographical regions.

through scene graph analysis, preserving object relationships and cultural context. Furthermore, BERTScore (Zhang et al., 2019) offers a deep learning-based evaluation of semantic similarity, capturing contextual meaning beyond simple word overlap. LLM-Judge further enhances assessment by evaluating coherence, factual accuracy, and contextual appropriateness.

Results and Analysis: Our evaluation of closed- and open-source models in the TimeTravel dataset highlights clear disparities in their ability to produce historically rich descriptions (see Table 2). Among closed-source models, GPT-4o-0806 consistently led in most metrics, BLEU (0.0190), ROUGE-L (0.1740), SPICE (0.0960), BERTScore (0.8482), and LLM-Judge (0.3007), demonstrating strong semantic fidelity, structural precision, and contextual relevance. Its METEOR score (0.2399), although not the highest, reflects competent lexical variation. GPT-4o-mini-0718 achieved the best METEOR score (0.2664) and performed competitively in ROUGE-L (0.1690), BERTScore (0.8432), and SPICE (0.0930), indicating its strength in producing more fluent and lexically diverse descriptions. Gemini-2.0-Flash and Gemini-1.5-Pro, while achieving moderate performance across all metrics, demonstrated weaker lexical alignment (BLEU: 0.0120, 0.0110) and object-level alignment (SPICE: 0.0650, 0.0770), suggesting limitations in capturing fine-grained detail and historical specificity. Among open models, Qwen-2.5-VL was the top performer, leading in METEOR (0.2649), BLEU (0.0140), SPICE (0.0770), and showing strong alignment in BERTScore (0.8379) and LLM-Judge (0.1749), suggesting a robust balance between fluency and historical accuracy. In contrast, Llama-3.2-Vision-Inst and Llava-Next recorded lower scores, especially in SPICE (0.0590 and 0.0650) and LLM-Judge (0.1226 and 0.1156), highlighting difficulties in generating contextually rich and semantically faithful descriptions.

Table 3 presents the LLM-Judge evaluation of models to describe archaeological artifacts between civilizations in different geographic regions. GPT-4o-0806 consistently ranked highest in nearly all regions, with top scores in India, the Roman Empire, China, Iran, Iraq, the British Isles, Japan, Central America, and Greece, reflecting a strong contextual understanding. Gemini-2.0-Flash followed as the second-best performer in most areas, including Iran, Iraq, China, and Central America, and achieved the highest score in Egypt. Among open-source models, Qwen-2.5-VL led the group in Iran, Iraq, and Egypt, although the performance gap between closed and open models remained substantial in generating historically accurate descriptions.

In addition, Table 4 presents an analysis of the METEOR scores, which evaluates the performance of the model in describing archaeological artifacts from different civilizations and regions. GPT-4o-mini-0718 and Qwen-2.5-VL consistently achieve the highest scores in most regions, with GPT-4o-mini-0718 leading in India, China, Central America, the British Isles, Iran, Iraq, Japan, and Egypt, while Qwen-2.5-VL performs best in the Roman Empire and Greece.

Overall, closed-source models outperform open-source models in generating context-aware descriptions, but ongoing improvements in open-source models highlight opportunities for fine-tuning and dataset expansion. These findings will guide further model enhancements, advancing AI-driven historical analysis and cultural heritage preservation.

4 Conclusion

We present the TimeTravel dataset, a curated collection of historical artifacts from 10 cultural regions, extensively curated by domain experts. We developed a rigorous data collection, filtering, and verification process to ensure accuracy and completeness. Using GPT-4o, we generated detailed textual descriptions, making the dataset more accessible and valuable for computational historical research. Our evaluation, using BLEU, METEOR, ROUGE-L, SPICE, BERTScore, and LLM-as-Judge, showed that closed-source models outperformed open-source alternatives, although open models are rapidly improving. Our analysis highlights the potential of LMMs in bridging historical records gaps while maintaining academic integrity. Using AI-driven methodologies, this work lays the foundation for advancing cultural heritage preservation and enhancing digital humanities research, thereby ensuring greater accessibility and precision in historical documentation.

5 Limitations and Societal Impact

While this research demonstrates the potential of LMMs to enhance historical documentation, the quality of the descriptions generated depends on the completeness and precision of the input data. In cases where historical records are fragmented or ambiguous, synthetic text may lack full contextual depth. Furthermore, biases present in training data can influence how models interpret and describe cultural artifacts, necessitating continuous evaluation and expert validation to ensure historical accuracy and cultural sensitivity. Despite these challenges, this research contributes to the preservation of cultural heritage, educational accessibility, and AI-driven humanities research. By digitizing and enriching historical records, it enables broader public engagement with history, supports museum digitization efforts, and provides a foundation for future advancements in AI-assisted historical analysis, bridging the gap between technology and human expertise in understanding our collective past.

Acknowledgment

The computations were enabled by resources provided by NAISS at Alvis partially funded by Swedish Research Council through grant agreement no. 2022-06725, LUMI hosted by CSC (Finland) and LUMI consortium, and by Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the NSC.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Badr AlKhamissi, Muhammad N. ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *ArXiv*, abs/2402.13231.

Fakhraddin Alwajih, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. 2024. Peacock: A family of arabic multimodal large language models and benchmarks. *arXiv preprint arXiv:2403.01031*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Marita

Chatzipanagiotou, Ion Androutsopoulos, J. Prag, and Nando de Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280 – 283.

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.

Fan Bu, Zheng Wang, Siyi Wang, and Ziyao Liu. 2025. An investigation into value misalignment in llm-generated texts for cultural heritage. *arXiv preprint arXiv:2501.02039*.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2024. Culturalbench: a robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of llms. *arXiv preprint arXiv:2410.02677*.

Sara Ghaboura, Ahmed Heakl, Omkar Thawakar, Ali Husain Salem Abdulla Alharthi, Ines Riahi, Abduljalil Saif, Jorma Laaksonen, Fahad Shahbaz Khan, Salman H Khan, and Rao Muhammad Anwer. 2025. Camel-bench: A comprehensive arabic lmm benchmark. *NAACL*.

Ahmed Heakl, Sara Ghaboura, Omkar Thawkar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman Khan. 2025. Ain: The arabic inclusive large multimodal model.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. Culturevlm: Characterizing and improving cultural understanding of vision-language models for over 100 countries. *arXiv preprint arXiv:2501.01282*.

Meta AI. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *arXiv preprint arXiv:2407.10920*.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. *ArXiv*, abs/2306.01857.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, et al. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *arXiv preprint arXiv:2406.05967*.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Qwen Team. 2025. Qwen2.5-vl.

Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, and Asanobu Kitamoto. 2020. Kaokore: A pre-modern japanese art facial expression dataset. *arXiv preprint arXiv:2002.08595*.

Caroline Tully. 2020. British museum. In *Encyclopedia of Global Archaeology*, pages 1618–1620. Springer.

Michael EW Varnum, Nicolas Baumard, Mohammad Atari, and Kurt Gray. 2024. Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences*, 121(42):e2407639121.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, et al. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Mei Wang and Weihong Deng. 2022. Oracle-mnist: a realistic image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:2205.09442*.

Pengjie Wang, Kaile Zhang, Xinyu Wang, Shengwei Han, Yongge Liu, Jinpeng Wan, Haisu Guan, Zhebin Kuang, Lianwen Jin, Xiang Bai, et al. 2024. An open dataset for oracle bone script recognition and decipherment. *arXiv preprint arXiv:2401.15365*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

In this appendix, we present supporting materials for our research, including related work, dataset statistics with regional coverage, annotation and verification details, and examples of archaeological samples from diverse cultures and time periods, accompanied by cross-model qualitative comparisons. The related work section contextualizes our contributions within ongoing efforts in AI-driven historical text generation. The data section highlights the balanced regional distribution of the samples. The annotation and verification process details the steps taken to ensure sample accuracy and historical integrity. Finally, we include qualitative examples illustrating the performance of both open- and closed-source models on selected artifacts.

B Related Work

Recent years have seen significant progress in studying cultural representation in AI, particularly in behavioral patterns, food, landmarks, and historical knowledge. However, most works focus on misalignment and biases in AI models or modern cultural trends, rather than positioning artifacts within their historical context and era across ancient civilizations. Meanwhile, studies on cultural inclusion in LLMs highlight the challenges of capturing the contextual and multifaceted nature of culture, emphasizing the limitations of text-based models in highlighting underrepresented cultures and the need for more robust evaluation methods (Adilazuarda et al., 2024).

Research on cultural influences in AI has increasingly focused on biases and misalignment in language models, particularly how they reflect and perpetuate dominant cultural norms. Early research on cultural biases in LLMs revealed their alignment with Western norms, particularly in moral reasoning, historical narratives, and societal values. Ramezani et al. (2023) analyze how monolingual English language models tend to reflect Western moral norms more strongly than diverse cultural perspectives, limiting their applicability in cross-cultural ethical contexts (Ramezani and Xu, 2023). Tao et al. (2024) further highlight the overrepresentation of Anglo-American and Protestant European values in AI-generated content, often underrepresenting non-Western traditions and belief systems (Tao et al., 2024). Similarly, Bu et al. (2025) explore value misalignment in cultural heritage-related text generation, warning of historical inaccuracies, cultural identity erosion, and oversimplification of complex narratives, with 65% of the generated content showing significant misalignment (Bu et al., 2025).

To mitigate these biases, several approaches have been proposed. AlKhamissi et al. (2024) introduce Anthropological Prompting, a method that encourages LLMs to reason like cultural anthropologists by incorporating both emic (insider) and etic (outsider) perspectives (AlKhamissi et al., 2024). Similarly, Li et al. (2024) propose CultureLLM, a fine-tuning approach designed to integrate cultural knowledge into LLMs, particularly for low-resource cultures (Li et al., 2024). While these techniques improve cultural alignment, their focus remains on modern cultural settings, leaving gaps in historical artifact contextualization across different time periods.

With the rise of Vision-Language Models (VLMs), cultural research has expanded into multimodal AI, revealing persistent patterns of bias. Liu et al. (2025) introduce CultureVLM to improve cultural understanding in VLMs by addressing limitations in recognizing non-Western symbols, artifacts, and gestures (Liu et al., 2025). They also present CultureVerse, a large-scale dataset evaluating cultural reasoning, though it primarily focuses on modern traditions and daily life. Similarly, Romero et al. (2024) develop CVQA, a multilingual VQA benchmark showing that state-of-the-art VLMs struggle with culturally grounded reasoning in non-Western settings (Romero et al., 2024). Extending this direction, ALM-Bench evaluates LMMs across 100 culturally and linguistically diverse languages, offering a large-scale framework for assessing multimodal inclusion (Vayani et al., 2024). Despite these contributions, most datasets focus on present-day cultures; even when historical artifacts appear, they are often framed through modern national lenses rather than within their original civilizations and time periods. This indicates a significant gap in the representation of artifacts in their authentic temporal and cultural contexts.

Efforts to bridge AI research with historical studies have led to the development of Historical Large Language Models (HLLMs), trained on historical texts to simulate past societies' psychology and value systems (Varnum et al., 2024). These models aim to provide insight into long-term cultural evolution, but their reliance on text-only representations limits their application in multimodal historical studies. Similarly, Assael et al. (2022) introduce

Ithaca, a deep learning model designed to assist historians in restoring, geographically attributing, and dating ancient Greek inscriptions, significantly improving accuracy over traditional methods (Assael et al., 2022). While these works contribute to historical AI, they primarily focus on text-based reconstruction rather than multimodal representations of historical artifacts across civilizations.

TimeTravel fills this gap by providing an open-source dataset of over 10k historical artifacts spanning 10 ancient world regions, both prehistoric and historic, offering the first benchmark to evaluate LMMs on temporal-cultural understanding, supported by expert verification. Unlike prior datasets focused on contemporary cultural knowledge, TimeTravel enables models to contextualize artifacts within their historical era, ensuring a more accurate representation of civilizations and their material culture. Supported by domain specialists, the dataset enhances reliability and authenticity, mitigating potential biases and inaccuracies in model-generated interpretations. By integrating both textual and multimodal perspectives, TimeTravel advances historical-cultural research, enabling systems to better understand and reason about artifacts within their original context.

C TimeTravel Samples Regional Distribution

Figure 4 illustrates the balanced regional distribution of dataset samples based on archaeological provenance. Greece holds the largest share at 18%, followed by multiple regions, including the Roman Empire, China, British Isles, Egypt, Iraq, and Iran, each at 10%. Japan (9%), India (8%), and Central America (5%) contribute smaller yet significant portions. Overall, the dataset ensures diverse cultural representation without dominance by any single region.

Tables 5 to 14 present further details about sample counts categorized by region of discovery, section, and cultural affiliation.

The covered areas in our study are ordered as follows:
Table 5 → "China"; Table 6→ "Central America"; Table 7 → "Iran"; Table 8 → "India"; Table 9 → "British Isles"; Table 10 → "Roman Empire", Table 11 → "Greece"; Table 12 → "Iraq", Tab. 13 →"Japan", and Table 14 → "Egypt".

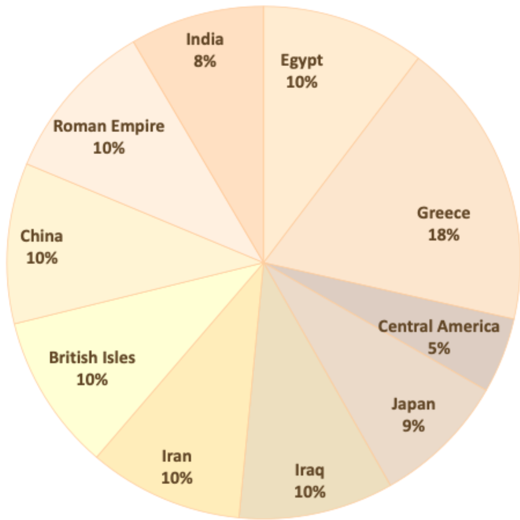

Figure 4: Regional distribution of dataset samples based on their archaeological provenance. Greece holds the largest share at 18%, with a balance-like distribution over regions.

D Annotation and Verification Process

The TimeTravel dataset underwent a rigorous human-in-the-loop annotation and verification pipeline to ensure high-quality, culturally grounded content:

- **Initial Descriptions:** GPT-4o (version 0806) was used to generate initial draft descriptions based on expert-curated metadata from museum websites. These drafts served solely as a starting point for further expert refinement.
- **Expert Annotation:** Ten expert annotators, specializing in history and archaeology, were organized into 5 pairs. Each pair annotated approximately 2,500 artifacts over a two-week period.
- **Senior Review:** Two senior controller annotators manually reviewed the full dataset across three dimensions: (1) informational accuracy and adequacy; (2) linguistic clarity; and (3) overall coherence.
- **Consensus-Based Acceptance:** Approximately 35% of samples were accepted directly when both annotators approved. The remaining 65% were revised or re-annotated based on expert feedback or via re-annotation using GPT-4o suggestions, which were subsequently validated and finalized by human experts.
- **Final Ground Truth:** All dataset entries included in the benchmark reflect human-

verified and expert-edited descriptions, ensuring the ground truth is not based on raw model output.

- **Evaluation Fairness:** The same version of GPT-4o used in the initial draft generation was also evaluated in the finalized data set. This ensured a consistent and fair comparison between all models.

E TimeTravel Benchmark Examples

Figures 5 to 8 present examples from the *TimeTravel* dataset. Figure 5 and 6 present artifact samples from India and Central America (Maya) with its descriptions. Figure 7 highlights the cultural and material diversity of the dataset in historical periods and regions, while Figure 8 compares model-generated outputs.

Place	China
Section	Tang Dynasty
Culture	**Samples**
Tang Dynasty; Sui Dynasty	1
Tang Dynasty; Ming Dynasty	3
Tang Dynasty; Ming Dynasty; Jin Dynasty; Yuan Dynasty	1
Tang Dynasty; Song Dynasty	1
Song Dynasty; Tang Dynasty	1
Liao Dynasty; Tang Dynasty	2
Tang Dynasty; Northern Wei Dynasty	1
Six Dynasties; Tang Dynasty	5
Tang Dynasty	1
Northern Qi Dynasty; Sui Dynasty; Tang Dynasty	1
Tang Dynasty; Liao Dynasty	3
Six Dynasties; Sui Dynasty; Tang Dynasty	1
Tang Dynasty; Five Dynasties; Northern Song Dynasty	381
Five Dynasties; Tang Dynasty	4
Tang Dynasty	628
Sui Dynasty; Tang Dynasty	5
Total	**1039**

Table 5: Culture Sample Counts from China (Tang Dynasty Section).

Place	Central America
Section	Maya
Culture	**Samples**
Classic Maya; Classic	3
Classic Maya; Late Preclassic Maya	64
Formative (Pre-Classic); Early Classic Maya	8
Late Classic Maya	23
Olmec; Maya	1
Classic Maya	275
Preclassic Maya	10
Classic Maya; Late Classic	2
Classic Maya; Olmec	1
Preclassic Maya; Classic Maya	2
Maya	95
Late Classic Maya; Late Classic	4
Total	**488**

Table 6: Culture Sample Counts from Central America (Maya Section).

Place	Iran
Section	Persian
Culture	**Samples**
Inju Dynasty	3
Middle Islamic; Seljuq Dynasty; Persian	1
Safavid Dynasty; Mughal Dynasty	1
Persian; Islamic	11
Persian; Late Islamic	3
Samanid Dynasty	27
Safavid Dynasty	395
Timurid Dynasty; Islamic	1
Safavid Dynasty; Post-Medieval	1
Mughal Dynasty; Persian	1
Ilkhanid Dynasty; Persian	3
Turkman Dynasty	3
Early Sasanian; Safavid Dynasty	1
Islamic; Safavid Dynasty	1
Ilkhanid Dynasty	192
Middle Islamic; Persian	6
Islamic; Qajar Dynasty	2
Persian; Safavid Dynasty	1
Safavid Dynasty; Persian; Islamic	2
Mughal Dynasty; Safavid Dynasty	1
Qajar Dynasty	193
Safavid Dynasty; Islamic	4
Persian; Mughal Dynasty	1
Islamic; Persian	2
Timurid Dynasty	35
Persian	108
Total	**999**

Table 7: Culture Sample Counts from Iran (Persian Section).

Place	India
Section	Mohenjo-Daro
Culture	**Sample Count**
Indus Valley Civil.	114
Section	Mauryan
Culture	**Sample Count**
Mauryan	17
Section	Gupta Dynasty
Culture	**Sample Count**
Gupta	737
Total	**868**

Table 8: Culture Sample Counts from India.

Place	British Isles
Section	Viking
Culture	**Samples**
Viking; Carolingian; Late Anglo-Saxon	1
Viking; Early Anglo-Saxon; Mid. Anglo-Saxon	1
Middle Anglo-Saxon Viking; Anglo-Saxon	1
Celtic; Viking	14
Viking; Late Anglo-Saxon	19
Viking; Finno-Ugrian	1
Anglo-Viking	52
Viking	895
Carolingian; Viking	1
Viking; Medieval	1
Late Anglo-Saxon; Viking	1
Viking; Celtic	26
Total	**1013**

Table 9: Culture Sample Counts from the British Isles (Viking Section).

Place	Roman Empire
Section	Roman
Culture	**Samples**
Roman Imperial	610
Roman	3
Roman Provincial	436
Total	**1049**

Table 10: Culture Sample Counts from the Roman Empire.

Place		Greece	
Section		Greek	
Culture	**Sample**	**Culture**	**Sample**
Greek; Hellenistic; Roman Imperial	4	Hellenistic; Roman Imperial	2
Attic	806	Middle Corinthian	5
Corinthian	41	East Greek; Classical Greek	1
Attic; Classical Greek	47	Transitional Corinthian	1
Middle Corinthian; Late Corinthian; Archaic Greek	7	Classical Greek; Attic	2
Proto-Corinthian	4	Classical Greek; Attic; Archaic Greek	1
Orientalising Period	14	East Greek Archaic II; Archaic Greek	1
Archaic Greek; Classical Greek	1	Attic; Western Greek	1
Archaic Greek	40	East Greek	23
Late Corinthian; Archaic Greek	11	Attic; Archaic Greek	318
Western Greek; Hellenistic	1	Attic; Archaic Greek; Classical Greek	12
Early Corinthian	8	Attic; Classical Greek; Archaic Greek	3
Laconian; Archaic Greek	10	Archaic Greek; East Greek	2
Classical Greek; Corinthian; Hellenistic	1	Rhodian	3
Late Helladic IIIB	2	Greek; Classical Greek	2
Transitional Corinthian; Archaic Greek	1	Early Corinthian; Archaic Greek	3
East Greek; Hellenistic	2	Middle Corinthian; Archaic Greek	11
Late Geometric IIA; Attic	1	East Greek; Orientalising Period	1
Archaic Greek; Attic	8	Late Minoan I; Late Minoan II	1
Late Minoan I	2	Archaic Greek; East Greek; North Ionian	1
Paestan	1	East Greek; Archaic Greek	237
Early Corinthian; Middle Corinthian; Archaic Greek	1	Greek; Hellenistic	2
Archaic Greek; East Dorian	1	Greek	3
Hellenistic	110	Western Greek	5
East Greek; Archaic Greek; Classical Greek	1	Roman; Hellenistic	3
East Dorian; Archaic Greek	2	Classical Greek	38
East Greek; East Dorian; Archaic Greek	11	Boeotian	25
Geometric Greek; Early Proto-Attic	1	Hellenistic; Classical Greek	2
East Greek; South Ionian	1	Geometric Greek	8
Greek; Classical Greek; Hellenistic	5	Hellenistic; Roman	4
Total			**1869**

Table 11: Culture Sample Counts from Greece (Greek Section).

Place		Iraq	
Section		Mesopotamian	
Culture	**Samples**	**Culture**	**Samples**
Neo-Assyrian; Late Babylonian	9	Late Babylonian; Assyrian	1
Elamite; Third Dynasty Of Ur	1	Early Dynastic (Middle East)	1
Old Assyrian; Early Bronze Age III	1	Late Uruk	26
Isin-Larsa	3	Neo-Assyrian	406
Uruk	3	Late Uruk; Chalcolithic	1
Middle Babylonian; Neo-Babylonian Dynasty	1	Old Babylonian; Cypriot	1
Late Babylonian	20	Babylonian; Neo-Assyrian	1
Neo-Assyrian; Babylonian	1	Assyrian; Late Babylonian	2
Jemdet Nasr; Proto-Elamite	1	Halaf	38
Assyrian	7	Middle Assyrian	11
Jemdet Nasr	27	Third Dynasty Of Ur; Ubaid	1
Old Babylonian	41	Kassite	4
Babylonian	3	Neo-Babylonian Dynasty	2
Babylonian; Akkadian	1	Old Assyrian	2
Old Babylonian; Third Dynasty Of Ur	1	Ubaid	15
Early Dynastic (Middle East); Akkadian	2	Early Dynastic II	2
Isin-Larsa; Old Babylonian	1	Jemdet Nasr; Akkadian	1
Old Babylonian; Assyrian	1	Akkadian	102
Early Dynastic III; Akkadian	10	Old Babylonian; Old Assyrian	1
Isin-Larsa; Old Babylonian; Kassite	1	Uruk; Jemdet Nasr	3
Early Dynastic II; Early Dynastic III	1	Early Dynastic III	81
Mesopotamian	1	Late Babylonian; Neo-Assyrian	3
Assyrian; Ubaid	1	Third Dynasty Of Ur; Old Babylonian	2
Third Dynasty Of Ur	137	Neo-Assyrian; Phoenician	2
Middle Babylonian	11	Lagash II	1
Third Dynasty Of Ur; Isin-Larsa	4		
Total			**1000**

Table 12: Culture Sample Counts from Iraq (Mesopotamian Section).

Place	Japan
Section	Japanese
Culture	**Samples**
Momoyama Period	6
Genroku Era; Hoei Era	1
Asuka Period	1
Muromachi Period; Momoyama Period	2
Late Kofun; Nara Period	1
Nara Period	12
Middle Kofun	13
Yayoi Period	5
Middle Kofun; Late Kofun	34
Edo Period; Kamakura Period	1
Oei Era	2
Kyowa Era; Oei Era	1
Edo Period; Momoyama Period	1
Jomon Period	16
Kyowa Era	1
Bunka Era	1
Bun'An Era; Bunsei Era	1
Muromachi Period	40
Asuka Period; Nara Period	1
Heian Period	9
Muromachi Period; Momoyama Period; Edo Period	1
Muromachi Period; Buddhist	1
Meiji Era	1
Hakuho Period	1
Showa Era	13
Early Kofun; Middle Kofun	26
Nanbokucho Period	2
Kofun Period; Edo Period	1
Edo Period	24
Kamakura Period; Meiji Era	1
Kofun Period	419
Early Kofun	7
Wado Era	1
Late Kofun	179
Kofun Period; Asuka Period	5
Kamakura Period	26
Nara Period; Edo Period	1
Kofun Period; Nara Period	1
Kamakura Period; Muromachi Period	9
Heian Period; Kamakura Period	1
Total	**869**

Table 13: Culture Sample Counts from Japan (Japanese Section).

Place	Egypt
Section	Ancient Egyptian
Culture	**Samples**
6th Dynasty	1
Late Cypriot; 18th Dynasty	1
26th Dynasty; Archaic Greek; Punic	1
Late Period; 30th Dynasty	1
30th Dynasty; Ptolemaic	15
22nd Dynasty	69
18th Dynasty; 19th Dynasty	2
New Kingdom; 19th Dynasty; 20th Dynasty	1
12th Dynasty	1
26th Dynasty; Archaic Greek; East Greek; Hellenistic	1
New Kingdom	21
Late Predynastic; 1st Dynasty	2
25th Dynasty	7
30th Dynasty	128
Middle Kingdom	1
Late Period	96
18th Dynasty; 21St Dynasty	1
21st Dynasty	171
19th Dynasty; 20th Dynasty	3
20th Dynasty; 21St Dynasty	2
26th Dynasty	257
19th Dynasty	40
18th Dynasty	95
1st Dynasty	7
Ramesside	21
23rd Dynasty	1
22nd Dynasty; $23R^{rd}$ Dynasty	3
26th Dynasty; 27th Dynasty	1
20th Dynasty	25
Late Period; 26th Dynasty	3
25th Dynasty; Kushite	1
26th Dynasty; Punic; Archaic Greek	1
27th Dynasty	13
25th Dynasty; 26th Dynasty	1
Third Intermediate	47
Late Period; Archaic Greek	2
Late Period; Ptolemaic	5
29th Dynasty	1
New Kingdom; Third Intermediate	1
Ancient Egypt	5
26th Dynasty; 30th Dynasty	1
Total	**1056**

Table 14: Culture Sample Counts from Egypt (Ancient Egyptian Section).

{
    "id": 4621,
    "Image": "https://media.britishmuseum.org/media/Repository/Documents/2014_11/5_17/e572b5a1_adcb_4691_871c_a3da0123a3fa/preview_01222265_001.jpg",
    "Description": "Silver coin.; Bust of Skandagupta, facing to the right, wearing a decorated cap and collar.; Peacock, with outspread wings and tail, standing facing with head to left. Surrounded by a border of dots.",
    "Production date": "456-467 (about)",
    "Find spot": null,
    "Materials": "silver",
    Technique": null,
    "Inscription": "Inscription type: inscription Inscription position: reverse Inscription language: Sanskrit Inscription script: Brahmi",
    "Subjects": "king/queen; peacock",
    "Assoc name": "Named in inscription & portrayed: Skandagupta",
    "Culture": "Gupta",
    "Section": "Gupta dynasty",
    "Place": "India"
}

for this artifact, we generated a description such as:

{
"description": "This silver coin from the Gupta dynasty, produced approximately between 456-467 CE, features a distinguished portrait of Skandagupta on the obverse. Wearing an ornate cap and collar, Skandagupta is depicted facing right, encapsulating the regal aesthetic of the period. The reverse displays a majestic peacock, wings and tail fully spread, elegantly poised with its head turned to the left, bordered by a circle of dots adding to the coin's decorative appeal. An inscription in Sanskrit, using the Brahmi script, adorns the reverse, providing linguistic and cultural context. This artifact, emblematic of Gupta artistry and power, underscores the dynastic symbolism and reverence for nature characteristic of this era in Indian history."
}

Figure 5: This entry represents a silver coin from the Gupta dynastyfrom India, featuring a distinguished portrait of Skandagupta on the obverse. GPT-4o generated a detailed, context-aware description based on the available metadata, highlighting its craftsmanship, ceremonial significance, and cultural context.

{
    "id": 7443,
    "Image": "https://media.britishmuseum.org/media/Repository/Documents/2015_9/4_10/b0f2152e_0ab5_4585_a7ca_a50900a611e8/preview_1997_q_1879__a_.jpg",
    "Description": "Votive object in the form of polished piece of jade with six holes drilled in it.",
    "Production date": null,
    "Find spot": "Excavated/Findspot: Tzimin Kax (?); Excavated/Findspot: Hatzcap Ceel (?); Excavated/Findspot: Cahal Pichik (?)",
    "Materials": "jade",
    "Technique": "polished; drilled",
    "Inscription": null,
    "Subjects": null,
    "Assoc name": null,
    "Culture": "Classic Maya; Late Preclassic Maya",
    "Section": "Maya",
    "Place": "Central America"
}

for this artifact, we generated a description such as:

{
"description": "This votive object, crafted by the Classic and Late Preclassic Maya cultures, is a testament to the artistry and spiritual practices of ancient Central America. Made from polished jade, a material revered for its symbolic and aesthetic value, this artifact features six precision-drilled holes, showcasing the sophisticated techniques employed by the Maya artisans. Although specific inscription details are absent, the object's polished surface and drilled holes suggest it may have had ceremonial or religious significance. Unearthed at find spots such as Tzimin Kax, Hatzcap Ceel, and Cahal Pichik, this piece provides valuable insights into the ritualistic and cultural engagements of the Maya civilization."
}

Figure 6: This particular entry represents a polished jade votive object from the Classic and Late Preclassic Maya features six precision-drilled holes, reflecting advanced craftsmanship and likely ceremonial significance. Unearthed at sites like Tzimin Kax, it offers insight into Maya rituals.

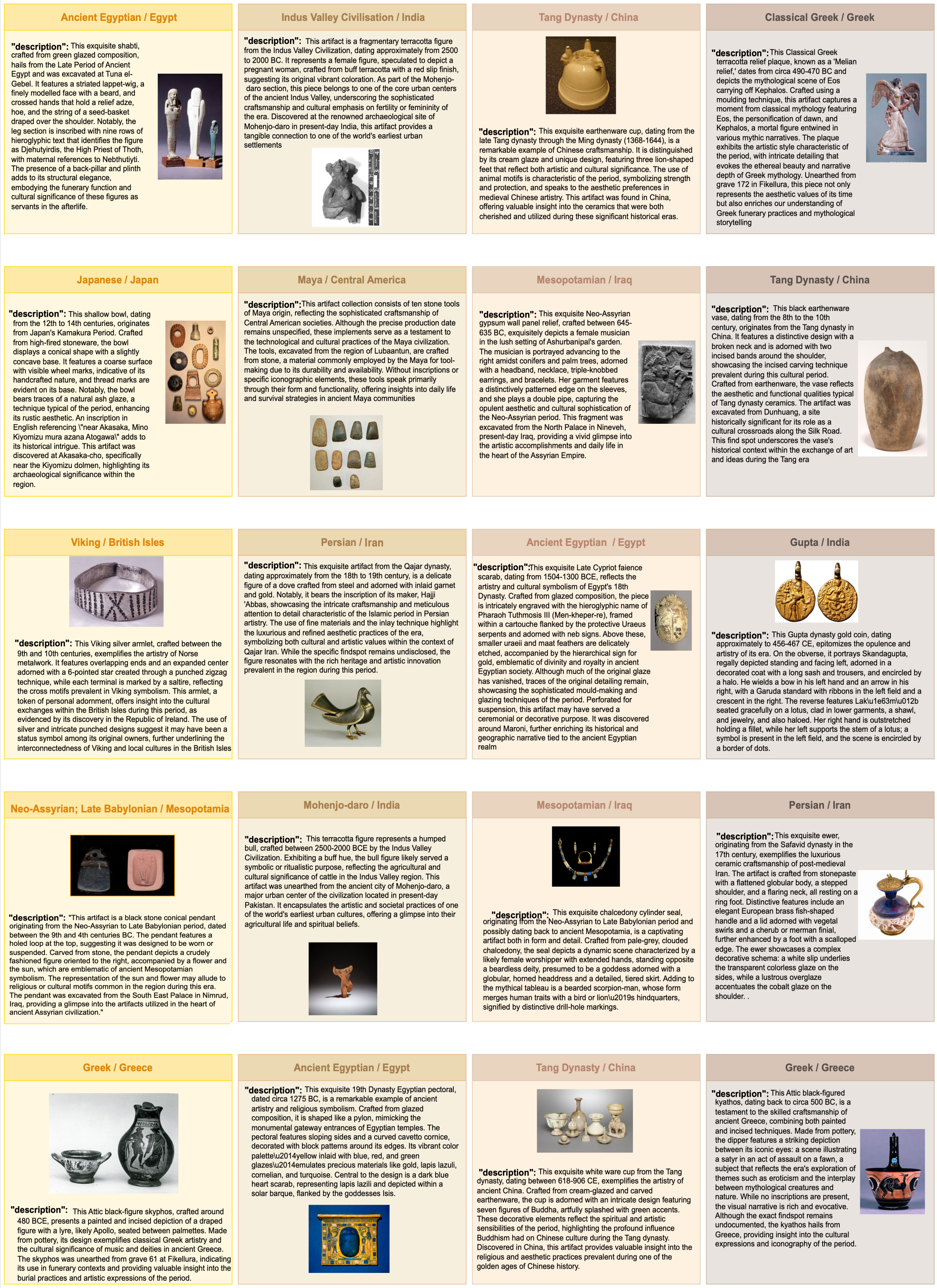

Figure 7: Cultural and material diversity of TimeTravel dataset samples across civilizations and historical periods. The dataset includes artifacts from Ancient Egypt, Greece, Mesopotamia, China, and Japan, spanning prehistoric to medieval times. A wide range of materials, including ceramics, metals, and stone, highlights artistic, technological, and societal influences, ensuring a comprehensive representation of historical craftsmanship and cultural heritage.

Figure 8: Cross-model comparison of generated descriptions for TimeTravel dataset samples, highlighting variations in detail and accuracy. It illustrates differences in descriptive depth across open- and closed-source models, emphasizing the diversity in interpretative approaches and alignment with the ground truth.