NRITYAM: Language Models Meet Art and Heritage of Dance

Punit Kumar Singh[1,5], Niladri Ghosh[2], Advait Joshi[3], Shailee Choudhary[4], Michael Färber[6], and Haiqin Yang (✉)[1,7]

[1] Shenzhen Technology University, 518118 Shenzhen, China
yanghaiqin@sztu.edu.cn
[2] Ramakrishna Mission Vivekananda Educational and Research Institute, India
[3] Swami Vivekananda Institute of Technology, India
[4] New Delhi Institute of Management, New Delhi, India
[5] Indian Institute of Technology, Patna, India
[6] Technische Universität Dresden, 01069 Dresden, Germany
[7] GuangDong Engineering Technology Research Center of Edge Intelligence

Abstract. Language models have become essential tools in shaping modern workflows. However, their global effectiveness hinges on a nuanced understanding of local socio-cultural contexts. To address this gap, we present NRITYAM, a comprehensive benchmark for evaluating the cultural comprehension capabilities of language models in the context of global dance traditions. NRITYAM comprises 9,260 carefully curated question-answer pairs spanning 12 languages, making it the largest dataset dedicated to evaluating cultural knowledge in dance. The dataset has been developed from the ground up through close collaboration with native dance artists and native speakers of the languages, who authored and validated culturally relevant questions specific to their regions. We evaluate a broad set of models, including large language models, small language models, multimodal large language models, and small multimodal language models. As a multilingual and multicultural benchmark, NRITYAM sets a new standard for evaluating the ability of AI systems to understand and reason about traditional performing arts. Detailed dataset samples are available at https://github.com/niladrighosh03/NRITYAM.

Keywords: Large Language Model · Multimodal and Multilingual Dataset

1 Introduction

Dance serves as a powerful medium for cultural and emotional expression, bringing together individuals from varied backgrounds and traditions [8]. Exploring different traditional dance forms offers meaningful perspectives on the values, historical experiences, and social frameworks of the communities that create and uphold them [36]. Moreover, dance holds significant societal importance, serving as a medium for preserving cultural knowledge and shaping collective identity [3]. The language, rituals, and evolving forms of dance serve as rich expressions

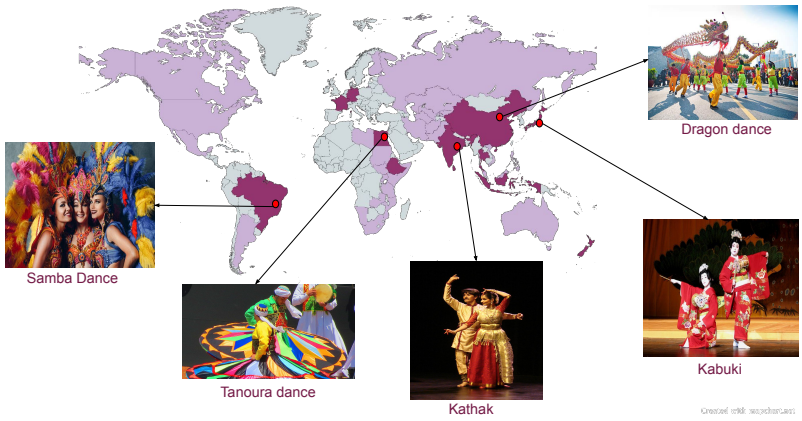


Fig. 1. NRITYAM is a diverse benchmark featuring 12 languages, with questions manually created and verified by native language speakers and dance experts. It spans 8 key aspects of traditional dance across two modalities, text and image, emphasizing mid- to low-resource languages. The benchmark features dances originating from 12 countries across 5 continents that are now performed in 100 countries across 6 continents, which are visualized with dark blue for their origins and light blue for their current reach. NRITYAM offers a wide range of question formats, including multiple-choice questions (MCQs) and both short and long visual question-answering (VQA) tasks.

of a community's historical experiences, cultural transformations, and collective identity encompassing elements such as traditional attire, religious associations, embodied movement, and social heritage [14].

Researchers have increasingly turned to traditional dance as a rich lens for analyzing cultural dynamics, providing a robust framework for understanding how embodied practices evolve across regional and societal boundaries [29]. Although many traditional dance forms share foundational elements such as expressive movement, symbolic gestures, and rhythmic structure, they manifest uniquely within different cultural contexts. These divergences are evident in variations in movement style, facial expression, ritual significance, performance settings, and the terminology used to describe them. For instance, Bharatanatyam, a classical Indian dance rooted in Tamil Nadu, has been adapted in Sri Lanka through a distinctive fusion with Kandyan dance aesthetics, reflecting localized narratives and performance traditions. Such adaptations highlight how traditional dance not only preserves the heritage but also evolves through intercultural exchange and regional reinterpretation. Language Models (LMs) have revolutionized natural language understanding, content generation, and decision-making, becoming indispensable across industries such as education, governance, and entertainment [11]. From Large Language Models (LLMs) to Multimodal Language Mod-

els (MLMs) and Small Language Models (SLMs) [1], these advancements have enabled seamless communication and efficient problem-solving [35]. However, a persistent challenge remains: ensuring that these models effectively recognize and reason about diverse linguistic and cultural contexts, particularly in underrepresented domains such as traditional dance [9].

Traditional and indigenous dance forms are deeply embedded in local histories, societal values, and cultural identities [20]. Despite their significance, current language models are predominantly trained and evaluated on global popular culture and mainstream dance styles, such as hip-hop, often neglecting heritage dance traditions and culturally distinctive practices. This bias can perpetuate inaccuracies, stereotypes, and the marginalization of underrepresented communities. Conversely, models capable of understanding and respecting cultural nuances can enhance performance while promoting greater inclusivity and equity in AI applications.

Motivation for NRITYAM[2] Dataset. Existing dance-related benchmarks are largely monolingual, English-centric [5,31], and focused primarily on motion recognition [19]. For example, PopDanceSet [30] is a monolingual dataset. To date, no comprehensive benchmark captures the rich cultural nuances of traditional dance reasoning across multiple languages, diverse cultural contexts, and visual question answering (VQA). To address this gap, we introduce NRITYAM, the **largest multicultural and multilingual traditional dance benchmark** to date. It consists of approximately 9,260 dance-related questions covering traditional dances originating from 12 countries across 5 continents, which are now performed in over 100 countries spanning 6 continents. The benchmark evaluates the capabilities of LLMs, SLMs, and MLMs across twelve languages. Questions are organized into two modalities, text-based and image-based, and systematically categorized into three key types: history-based, rule-based, scenario-based. [3]

This work is guided by the following research questions:

(1) How do different categories of models, i.e., LLMs, SLMs, MLMs, and Small Multimodal Language Models (SMLMs), perform on the NRITYAM dataset?
(2) What trends and patterns emerge in model performance across the various question types, including history-based, rule-based, scenario-based, and image-based questions, in the NRITYAM dataset?
(3) What are the performance trends of language models across different countries or languages in Asia, Africa, South America, Oceania, and Europe?

Key contributions. Our main contributions are as follows:

1. The NRITYAM Dataset: We present the first and the most comprehensive QA dataset on traditional dance, covering 12 countries across 5 continents (with global reach across 100+ countries on 6 continents). The dataset is available in 12 native languages as well as in English.

[1] Any model with 7B parameters or fewer is considered a Small Language Model (SLM) in this work.
[2] The Sanskrit term "NRITYAM" holds profound significance in classical Indian dance.
[3] MLMs are used exclusively for evaluating the image-based questions.

2. Diverse Question Types: The dataset includes 9,260 questions spanning two modalities (text and image) and three categories, challenging AI models to reason over textual, visual, multilingual, and culturally grounded inputs.

3. Comprehensive Benchmarking: We rigorously evaluate 13 state-of-the-art language models, including LLMs, SLMs, MLMs, and SMLMs, uncovering critical gaps in their cultural and contextual reasoning abilities regarding traditional dance.

By addressing cultural under-representation in AI, **NRITYAM** establishes a robust benchmark for evaluating and improving AI systems. This research advances the intersection of NLP and culturally rich domains, contributing to greater inclusivity and equity in AI applications worldwide.

2 Related Work

Prior Cultural VQA Benchmarks. Pioneering efforts in culturally aware Vision-Question Answering (VQA) have led to datasets spanning various regional and thematic dimensions, including FM-IQA [17], MCVQA [21], xGQA [37], MaXM [12], MTVQA [40], MABL [23], MAPS [26], and MaRVL [27]. Parallel lines of work, such as CVQA [33], CulturalVQA [34], ALM-bench [43], and CultSportQA [39], offer broader resources covering distinct cultural themes, with CVQA introducing multilingual queries paired with English translations. Other benchmarks adopt granular geographic or thematic scopes; for instance, SEA-VQA [42] focuses strictly on Southeast Asia, while FoodieQA [25], World Wide Dishes [32], and WORLDCUISINES [45] focus on specific culinary traditions. While NRITYAM shares the foundational objective of using a specific cultural lens, i.e., traditional dance, it distinctly expands upon prior art through a substantially larger dataset and broader multilingual coverage.

Sociocultural Reasoning in LLMs. Complementing visual benchmarks, recent NLP research scrutinizes the behavioral and alignment gaps of LLMs through established sociological frameworks like the World Values Survey and Hofstede's cultural dimensions [4]. These studies frequently reveal critical alignment deficiencies when adapting systems to user-specific or non-Western contexts [22]. Although interventions like synthetic personas and targeted fine-tuning show promise in elevating cultural adaptability and safe text moderation (e.g., cross-lingual hate speech detection), regional and low-resource language performance persistently lags behind English baselines [24,16,13]. These systemic limitations underscore the critical necessity for robust, multilingual evaluation suites to genuinely foster and measure the cultural competence of large foundation models.

3 Construction of NRITYAM

3.1 Manual Data Collection

The creation of NRITYAM follows a carefully structured, multi-phase process to ensure comprehensive coverage and high-quality standards. Domain experts

Table 1. Comparison of our dataset with other dance-related datasets. The metadata compared includes the number of samples (questions), the number of dances, whether cultural aspects are considered, the number of languages, modalities (i.e., whether the data includes multimodal questions), and question type. "—" indicates the value is not specified.

Dataset	#Samples	#Dances	Cultural Aspects	#Languages	Modalities	Question Type
CULTURALVQA [34]	2,378	—	No	20	Image	MCQ
CVQA [33]	10,000	—	No	31	Text + Image	MCQ
NRITYAM (ours)	**9,260**	**30**	**Yes**	**12**	Text + Image	**MCQ**

and country-specific annotators contribute at every stage, from data collection to question formulation and manual translation across multiple languages, incorporating their cultural knowledge and expertise

Data Sources: The dataset is built by sourcing information from multiple diverse and reliable platforms, including Wikipedia, government culture websites, local culture dance blogs, dance culture journals, and news outlets. These sources are carefully selected to authentically capture the essence of traditional dance from 12 countries: Brazil, China, Egypt, Ethiopia, France, Germany, India, Indonesia, Japan, New Zealand, Sri Lanka, Thailand, with emphasis on their cultural and regional significance. The questions are designed to focus on historical, rule-based, and image-based aspects of the dance.

1. Wikipedia: As a foundational resource, Wikipedia offers well-documented, detailed insights into the history, origins, and rules of traditional dance, serving as a cornerstone for verified information across diverse dance traditions.

2. Government Culture Website: The government culture websites of each country contribute unique perspectives on the historical and cultural relevance of traditional dance forms, ensuring authenticity and depth.

3. Local Culture Dance Blogs: Dedicated blogs and websites focusing on indigenous and traditional dance offer community-driven perspectives, unique insights, and region-specific practices associated with each dance form.

4. Dance Cultural Journals: Scholarly journals and publications specializing in cultural studies enrich the dataset with detailed articles on the evolution and societal impact of traditional dance across different regions.

5. News outlets: Renowned news outlets from each country provide contemporary coverage, including recent trends, key events, and ongoing efforts to preserve or revive traditional dances.

Dataset Organization

The NRITYAM dataset is divided into two categories: text-based and image-based. Each category is further organized into three question types: history-based, rule-based, and scenario-based questions.

The dataset follows a multiple-choice question (MCQ) format with four options (A, B, C, D), out of which only one is correct. Each question-answer (QA) pair includes metadata such as continent, country, posture, religion, and other question types.

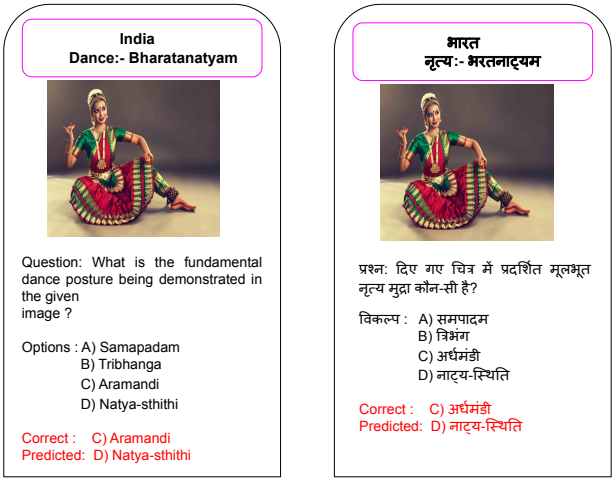


Fig. 2. Example illustration of Indian traditional dance wrong prediction by language model (in Hindi and English)



Fig. 3. Example illustration of Ethiopia's traditional dance correct prediction by language model (in Amharic and English)

The text-based questions are evaluated using LLMs and SLMs, while the image-based questions are assessed using MLMs and SMLMs.

History-based questions test the model's knowledge of a dance's origins and cultural significance. Scenario-based questions assess the model's ability to determine the most appropriate move in a given dance situation for a good performance. Rule-based questions evaluate the model's understanding of the fundamental rules of the dance depicted in the text or image. Religion-based questions assess the model's understanding of dances associated with particular religions, and so on. Examples from the NRITYAM dataset are illustrated in Figures 2 and 3.

3.2 Annotation Process

We outline the main steps in the annotation below:

1. Team Structure and Annotators Background. We hire 36 expert workers from 12 countries, with 3 representatives from each country, based on the following eligibility criteria: (1) The individual must be a native speaker. (2) They must have lived in the country for at least 10 years. (3) They must possess a strong understanding of local culture and traditional dance. (4) Their parents must also be from the country and currently reside there. (5) They must hold at least a degree from a nationally accredited dance institution. Two-thirds of the annotators are responsible for creating questions based on the provided guidelines, leveraging their knowledge of traditional dance. The remaining annotator was tasked with validating and filtering out questions that fail to meet quality standards.

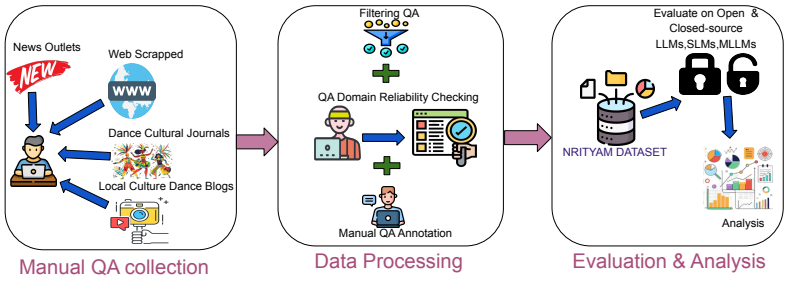


Fig. 4. Manual dataset construction pipeline for NRITYAM: The data collection process involved two key stages: (1) annotators gather data sources and generate questions, drawing from their respective cultural backgrounds and languages; (2) annotators review and verify the questions to ensure cultural authenticity and maintain high translation quality.

2. Question Formation. For each selected textual passage or image, the annotator's first task is to verify whether the content aligned with the traditional dance associated with the country. Content unrelated to the traditional dance is immediately rejected. If the content is relevant, the annotator creates questions focusing on rules, history, location, attire, religion, scenario, artist, and posture. Each question has to be complete, self-contained, and understandable without additional context. The questions follow a multiple-choice format, consisting of four options, with only one correct answer. The final annotated format includes the source passage, a relationship attribute indicating the question's context, the type of question (e.g., history-based, rule-based, scenario-based), and the four answer options. After a question is constructed, it is translated into the regional language that the annotator is familiar with. The annotators are paid at a rate between $0.10 and $0.50 per example, depending on the country exchange rate and difficulty of annotation.

3. Training and Guidelines. Annotators are receiving comprehensive training that covers the objectives of the NRITYAM dataset, clear definitions and examples of various question types, and best practices for ensuring consistency and cultural sensitivity. Detailed guideline documents are being provided, including templates, metadata tagging standards, and examples of culturally appropriate representations. Additional sessions focus on language and cultural training, emphasizing the correct use of local terminologies and traditions. Each participant is required to attend a live one-hour online workshop or watch a recorded version. These sessions offer an overview of the project, explaining the task instructions in detail and addressing any potential questions. To ensure a thorough understanding of the assignment, a pilot study is being conducted before the main annotation phase.

4. Quality Assurance and Cross-Validation. A rigorous quality assurance process is being conducted through multi-step validations. Each question-answer pair is undergoing cross-validation by at least one annotator, who understands the basic requirements to be qualified for inclusion in the dataset, as well as the translation quality. Image-based questions are being reviewed for proper alignment between visual elements and textual prompts. Spot checks and random sampling are being performed by quality analysts to maintain clarity and consistency. Bias mitigation measures are ensuring a balanced representation of dance across regions and question types, while cultural sensitivity reviews are eliminating stereotypes or offensive content.

4 Statistical Analysis of NRITYAM

The NRITYAM dataset shows a balanced mix of text-based and image-based questions, with a slight dominance of text-based ones comprising 5,110 questions over the visual ones, which comprise of 4,150. Image-based questions mainly focuses on dance rules, scenarios, and history. Figures 5 and 6 show the distribution of text and image-based questions across languages.

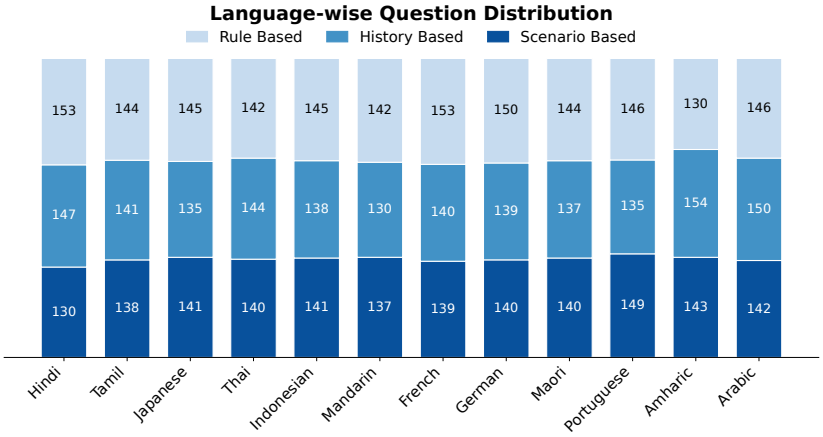


Fig. 5. Distribution of text-based questions across language.

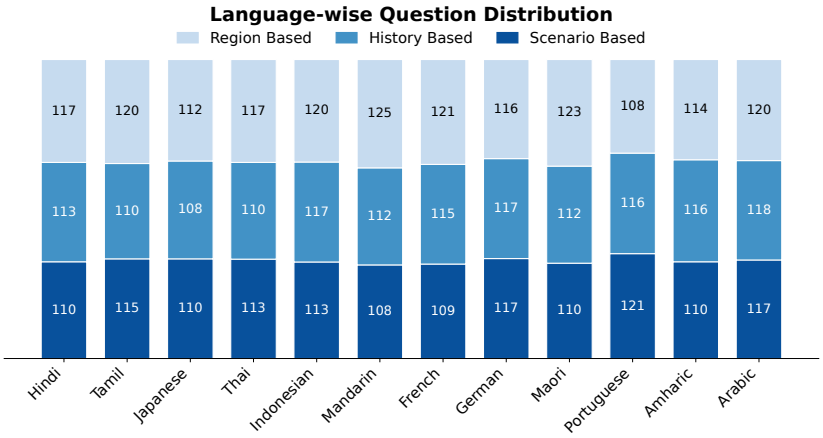


Fig. 6. Distribution of image-based questions across language.

5 Experiments

5.1 Models

To comprehensively evaluate our proposed benchmark, NRITYAM, we conduct an extensive assessment across a diverse suite of foundation models spanning both textual and visual modalities. For text-based evaluation, our suite encompasses leading LLMs, including Llama-3.1 [15], GPT-4 [2], GPT-5 [38], Gemma-7B [41], Phi-4 [1], Qwen-3 [6], and Claude-4.5. Beyond text-based models, we evaluate a wide range of MLMs to assess dance reasoning capabilities in multilingual settings. Specifically, this visual category includes mBLIP (a BLIP-2-based model) [18], PaliGemma-2 [10], LLaVA [28], Qwen3-VL [7], and DeepSeek-OCR [44].

5.2 Evaluation Setup

We conducted a comprehensive evaluation of the NRITYAM dataset, which comprises text and image-based Multiple Choice Questions (MCQs) categorized into three distinct dimensions: 1. Cultural and Historical Knowledge, 2. Rule Comprehension, and 3. Scenario-Based Reasoning. To systematically benchmark model performance across diverse languages and modalities, all evaluations are executed under a **zero-shot** prompting paradigm. For reproducibility and consistency, the temperature is set to 0, and **accuracy** serves as the primary evaluation metric. For deployment, open-source models are initialized using 16-bit floating-point precision (FP16) and evaluated via greedy decoding, whereas proprietary models are queried through their official developer APIs. Final model predictions are extracted based on the highest output token probability, establishing a standardized and deterministic evaluation pipeline.

6 Discussion on Results

6.1 Main Results

Performance of LLMs and SLMs. Across our multilingual evaluation, frontier LLMs consistently outperform compact models. GPT-5 emerges as the top performer (61.73%), closely followed by Claude Opus 4.5 (61.32%). Both display robust stability across categories: GPT-5 scores 62.51% (Rule), 62.18% (History), and 60.51% (Scenario), while Claude-4.5 Opus yields 62.21%, 61.29%, and 60.46%, respectively. GPT-4 occupies a distinct secondary tier at 53.08%. Among mid-range models, Qwen-3 32B (43.76%) outperforms Phi-4 (40.36%), LLaMA-3.1-8B (36.99%), and Gemma-7B (35.32%). Universally, models excel in European and East Asian languages (e.g., German, French, Mandarin) but degrade on low-resource languages (e.g., Māori, Amharic, Arabic). This highlights a persistent cross-lingual disparity, demonstrating that model scale correlates with better generalization on cultural QA tasks. Granular breakdowns are provided in Figures 7, 8, and 9.

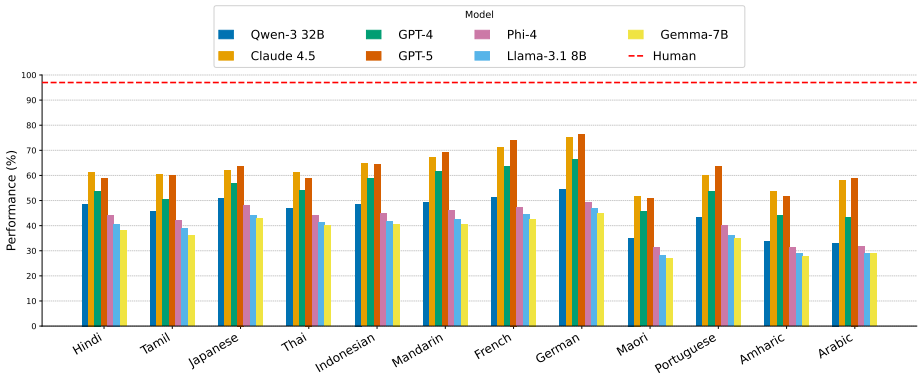


Fig. 7. Results of LLMs and SLMs on rule-based questions across languages (text modality).

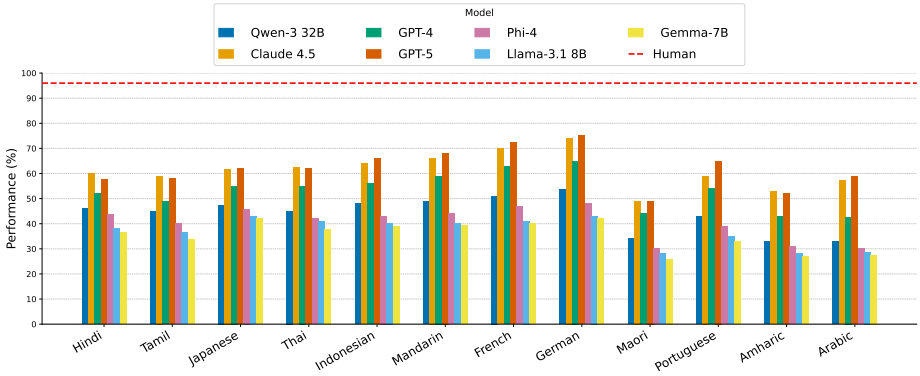


Fig. 8. Results of LLMs and SLMs on history-based questions across languages (text modality).

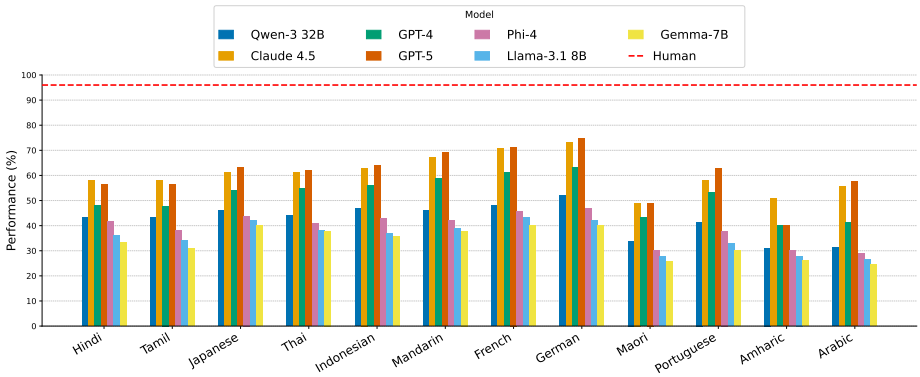


Fig. 9. Results of LLMs and SLMs on scenario-based questions across languages (text modality).

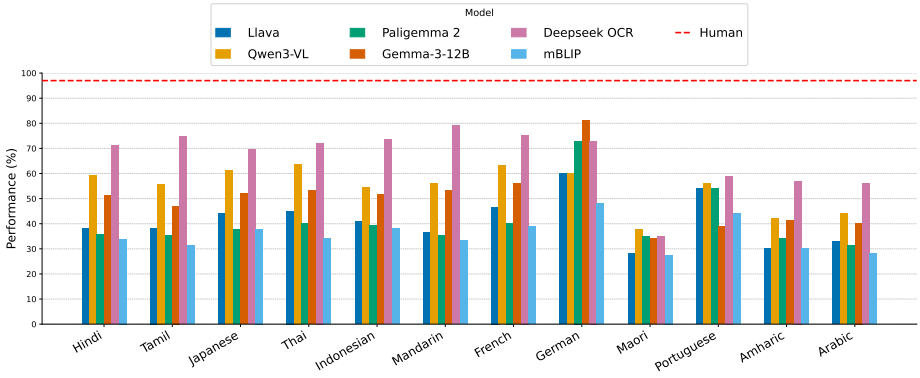


Fig. 10. Results of MLMs and SMLMs on rule-based questions across languages (image modality).

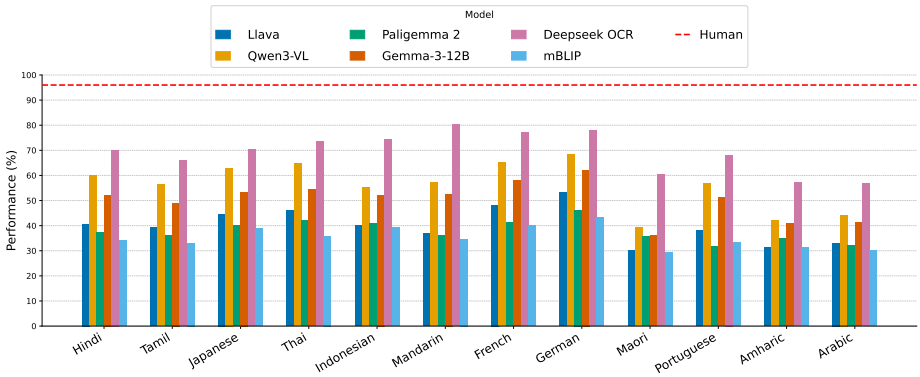


Fig. 11. Results of MLMs and SMLMs on history-based questions across languages (image modality).

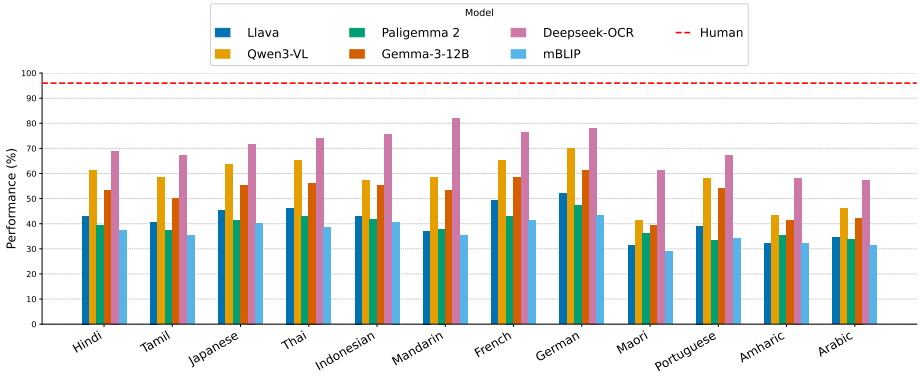


Fig. 12. Results of MLMs and SMLMs on scenario-based questions across languages (image modality).

Performance of MLMs and SMLMs. Across the multilingual multimodal evaluation, DeepSeek-OCR emerges as the top-performing system with 68.64% overall average, consistently leading across Scenario-based (69.80%), History (69.32%), and Rule-based (66.30%) categories. Qwen3-VL forms a distinct secondary tier, achieving a solid overall average of 55.77% (Scenario: 57.32%, History: 56.02%, Rules: 54.47%). Among smaller and mid-range architectures, Gemma-3-12B performs competitively at 50.73%, outperforming LLaVA (40.95%), PaliGemma-2 (40.26%), and mBLIP (35.83%). Universally, models excel on scenario-based reasoning but encounter performance degradation on historical and rule-oriented cultural tasks, particularly under low-resource language settings. A shared cross-lingual bottleneck is observed: while architectures generalize effectively in high-resource languages (e.g., German, French, Mandarin, Japanese, and Thai), low-resource targets like Māori, Amharic, and Arabic present persistent challenges. These results confirm that advanced vision-language scales substantially improve multimodal cultural comprehension, whereas compact architectures continue to struggle with historically grounded and rule-driven reasoning. Figures 10, 11, and 12 provide granular performance breakdowns.

6.2 Human Evaluation

Human raters evaluated the rule-based, history-based, and scenario-based categories to assess cultural reasoning. In the text modality, accuracies were 97%, 96%, and 96%, while in the image modality, they were 97%, 96%, and 96%, respectively. Despite strong human consistency, a partial misalignment with model predictions highlights that language models still fall short of replicating human-level cultural interpretation and contextual reasoning.

7 Conclusion

In this work, we construct NRITYAM, a comprehensive benchmark designed to evaluate language models' understanding of Asian, African, South American, Oceania, and European traditional dance. The dataset, consisting of 9,260 curated question-answer pairs from 12 countries, covers eight key aspects such as rules, history, location, attire, and other cultural significance. Evaluations with leading models reveal notable gaps in answering traditional dance-specific questions, highlighting biases likely caused by training data limitations. NRITYAM, built for quality and cultural sensitivity, advances inclusive AI research. Future expansions will add more languages and traditional dance to enhance its impact.

Limitations

Despite being one of the most comprehensive evaluations of language models on traditional dance and cultural knowledge, this study has several limitations:

(1) **Limited Geographic, Language, and Cultural Scope:** The NRITYAM dataset currently covers 12 languages from 12 countries across 5 continents. While valuable, its scope remains limited. Expanding to include more countries, especially those with underrepresented regions and low-resource languages, would improve cultural diversity and inclusivity, enabling fairer and broader assessments of language models.
(2) **Limited Representation of Traditional Dances:** Although the dataset covers 30 traditional dances—the largest multilingual and multimodal cultural dance dataset—it may still not fully capture the diversity of traditional dances across continents. Future versions could include more dance forms and additional question types, such as True/False, adversarial, and scenario-based reasoning tasks.

Ethics Statement

Data Collection and Bias Mitigation. Data for NRITYAM are collected from publicly accessible platforms (see Sec. 3.1), carefully selected to ensure authenticity. NRITYAM thus represents a significant step toward a standardized, inclusive benchmark for traditional dances from Asia, Europe, South America, Oceania, and Africa. All sources are verified by annotators through multiple group discussions, and irrelevant metadata was discarded.

Human Annotation. A diverse team of 36 annotators—experts in Asian, Oceanian, African, South American, and European dance, linguistics, and related fields—crafted, verified, and translated questions. The team included native and bilingual speakers from 12 countries, all with deep regional and dance knowledge. Annotators underwent comprehensive training on dataset objectives, question categories, and dance-specific guidelines. Nearly all were native speakers with approximately 15 years of experience in dance, ensuring linguistic and cultural accuracy. The team, aged 28–50, provided a balanced generational perspective. Annotation was collaborative, with cross-validation by a separate subteam to ensure consistency and address biases. Ethical considerations—avoiding stereotypes and promoting inclusivity—remained central, ensuring the dataset fairly reflects diverse cultural values.

Acknowledgments. This work was supported by the Stable Support Program of Universities in Shenzhen (20231129211559001). We further acknowledge financial support from BMFTR and SMWK within the Center of Excellence for AI Research "ScaDS.AI Dresden/Leipzig".

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Afolaranmi, V.B., Afolaranmi, A.O.: Cultural revitalization through dance as a panacea for peacebuilding. Advanced Journal of Theatre and Film Studies **2**(1), 39–45 (2024)
4. AlKhamissi, B., ElNokrashy, M.N., Alkhamissi, M., Diab, M.T.: Investigating cultural alignment of large language models. In: ACL. pp. 12404–12422 (2024)
5. Aristidou, A., Chalmers, A., Chrysanthou, Y., Loscos, C., Multon, F., Parkins, J., Sarupuri, B., Stavrakis, E.: Safeguarding our dance cultural heritage. In: Eurographics 2022-43nd Annual Conference of the European Association for Computer Graphics. pp. 1–6 (2022)
6. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
7. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
8. Bannerman, H.: Is dance a language? movement, meaning and communication. Dance Research **32**(1), 65–80 (2014)
9. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
10. Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al.: Paligemma: A versatile 3b vlm for transfer. arXiv preprint arXiv:2407.07726 (2024)
11. Brown, T., Mann, B., Ryder, N., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
12. Changpinyo, S., Xue, L., Yarom, M., Thapliyal, A., Szpektor, I., Amelot, J., Chen, X., Soricut, R.: Maxm: Towards multilingual visual question answering. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 2667–2682 (2023)
13. Deng, Y., Zhang, W., Pan, S.J., Bing, L.: Multilingual jailbreak challenges in large language models. In: International Conference on Learning Representations. vol. 2024, pp. 24634–24651 (2024)
14. Desmond, J.C.: Embodying difference: Issues in dance and cultural studies. Cultural Critique (26), 33–63 (1993)
15. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
16. Dwivedi, S.K., Patel, R.: Exploring the intersections: Anthropological insights into studying language and culture. State Institute of Education, Allahabad **30**, 171–182 (2024)
17. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? dataset and methods for multilingual image question. Advances in neural information processing systems **28** (2015)

18. Geigle, G., Jain, A., Timofte, R., Glavaš, G.: mblip: Efficient bootstrapping of multilingual vision-llms. arXiv preprint arXiv:2307.06930 (2023)
19. Grammalidis, N., Dimitropoulos, K., Tsalakanidou, F., Kitsikidis, A., Roussel, P., Denby, B., Chawah, P., Buchman, L., Dupont, S., Laraba, S., et al.: The i-treasures intangible cultural heritage dataset. In: Proceedings of the 3rd International Symposium on Movement and Computing. pp. 1–8 (2016)
20. Guo, C., Li, Z.: The impact of dance culture learning on students' cultural values. Cultura: International Journal of Philosophy of Culture and Axiology **22**(2), 422–440 (2025)
21. Gupta, D., Lenka, P., Ekbal, A., Bhattacharyya, P.: A unified framework for multilingual and code-mixed visual question answering. In: Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing. pp. 900–913 (2020)
22. Johnson, R.L., Pistilli, G., Menédez-González, N., Duran, L.D.D., Panai, E., Kalpokiene, J., Bertulfo, D.J.: The ghost in the machine has an american accent: value conflict in gpt-3. arXiv preprint arXiv:2203.07785 (2022)
23. Kabra, A., Liu, E., Khanuja, S., Aji, A.F., Winata, G.I., Cahyawijaya, S., Aremu, A., Ogayo, P., Neubig, G.: Multi-lingual and multi-cultural figurative language understanding. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 8269–8284 (2023)
24. Kwok, L., Bravansky, M., Griffin, L.: Evaluating cultural adaptability of a large language model via simulation of synthetic personas. In: The First Conference on Language Modeling (2024)
25. Li, W., Zhang, C., Li, J., Peng, Q., Tang, R., Zhou, L., Zhang, W., Hu, G., Yuan, Y., Søgaard, A., et al.: Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 19077–19095 (2024)
26. Liu, C., Koto, F., Baldwin, T., Gurevych, I.: Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In: NAACL. pp. 2016–2039 (2024)
27. Liu, F., Bugliarello, E., Ponti, E.M., Reddy, S., Collier, N., Elliott, D.: Visually grounded reasoning across languages and cultures. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10467–10485 (2021)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in Neural Information Processing Systems **36** (2023)
29. Lucchi Basili, L., Sacco, P.L.: Dance and the embodied social cognition of mating: Carlos saura's tango in the perspective of the tie-up theory. Integrative Psychological and Behavioral Science **59**(1), 31 (2025)
30. Luo, Z., Ren, M., Hu, X., et al.: Popdg: Popular 3d dance generation with popdanceset. arXiv preprint arXiv:2405.03178 (2024)
31. Ma, T., Ngo, T., Tran, H.N., Ton, M.L., Ha, N.L., Phan, B.C., Do, T.N.: Vimva: Innovative multimodal recognition in vietnamese folk dance video analysis. In: International Conference on Future Data and Security Engineering. pp. 283–298. Springer (2024)
32. Magomere, J., Ishida, S., Afonja, T., Salama, A., Kochin, D., Yuehgoh, F., Hamzaoui, I., Sefala, R., Alaagib, A., Semenova, E., et al.: You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes. arXiv preprint arXiv:2406.09496 (2024)

33. Mogrovejo, D.O.R., Lyu, C., Wibowo, H.A., Góngora, S., Mandal, A., Purkayastha, S., Ortiz-Barajas, J.G., Cueva, E.V., Baek, J., Jeong, S., et al.: Cvqa: Culturally-diverse multilingual visual question answering benchmark. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)
34. Nayak, S., Jain, K., Awal, R., Reddy, S., Van Steenkiste, S., Hendricks, L.A., Stańczak, K., Agrawal, A.: Benchmarking vision language models for cultural understanding. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 5769–5790 (2024)
35. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
36. Peng, X.: Historical development and cross-cultural influence of dance creation: Evolution of body language. Herança **7**(1), 88–99 (2024)
37. Pfeiffer, J., Geigle, G., Kamath, A., Steitz, J.M.O., Roth, S., Vulić, I., Gurevych, I.: xgqa: Cross-lingual visual question answering. In: Findings of the association for computational linguistics: ACL 2022. pp. 2497–2511 (2022)
38. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
39. Singh, P.K., Kumar, N., Ghosh, A., Pasad, K., Soni, K., Jaishwal, M., Saha, S., Alfarozi, S.A.I., Abagissa, A.T., Pasupa, K., et al.: Let's play across cultures: A large multilingual, multicultural benchmark for assessing language models' understanding of sports. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 15205–15252 (2025)
40. Tang, J., Liu, Q., Ye, Y., Lu, J., Wei, S., Wang, A.L., Lin, C., Feng, H., Zhao, Z., Wang, Y., et al.: Mtvqa: Benchmarking multilingual text-centric visual question answering. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 7748–7763 (2025)
41. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)
42. Urailertprasert, N., Limkonchotiwat, P., Suwajanakorn, S., Nutanong, S.: Sea-vqa: Southeast asian cultural context dataset for visual question answering. In: Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR). pp. 173–185 (2024)
43. Vayani, A., Dissanayake, D., Watawana, H., Ahsan, N., Sasikumar, N., Thawakar, O., Ademtew, H.B., Hmaiti, Y., Kumar, A., Kuckreja, K., et al.: All languages matter: Evaluating lmms on culturally diverse 100 languages. arXiv preprint arXiv:2411.16508 (2024)
44. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
45. Winata, G.I., Hudi, F., Irawan, P.A., Anugraha, D., Putri, R.A., Yutong, W., Nohejl, A., Prathama, U.A., Ousidhoum, N., Amriani, A., et al.: Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 3242–3264 (2025)