

# Let's Play Across Cultures: A Large Multilingual, Multicultural Benchmark for Assessing Language Models' Understanding of Sports

Punit Kumar Singh<sup>1</sup>, Nishant Kumar<sup>1</sup>, Akash Ghosh<sup>1</sup>, Kunal Pasad<sup>2</sup>,  
Khushi Soni<sup>2</sup>, Manisha Jaishwal<sup>1</sup>, Sriparna Saha<sup>1</sup>, Syukron Abu Ishaq Alfarozi<sup>3</sup>,  
Asres Temam Abagissa<sup>1</sup>, Kitsuchart Pasupa<sup>4</sup>, Haiqin Yang<sup>5\*</sup>, Jose G Moreno<sup>6\*</sup>

<sup>1</sup>Indian Institute of Technology Patna, India <sup>2</sup>Sardar Patel Institute of Technology, Mumbai

<sup>3</sup>Universitas Gadjah Mada, Indonesia <sup>4</sup>King Mongkut's Institute of Technology Ladkrabang, Thailand

<sup>5</sup>Shenzhen Technology University, China <sup>6</sup>Université de Toulouse, France

## Abstract

Language Models (LMs) are primarily evaluated on globally popular sports, often overlooking regional and indigenous sporting traditions. To address this gap, we introduce *CultSportQA*, a benchmark designed to assess LMs' understanding of traditional sports across 60 countries and 6 continents, encompassing four distinct cultural categories. The dataset features 33,000 multiple-choice questions (MCQs) across text and image modalities, each of which is categorized into three key types: history-based, rule-based, and scenario-based. To evaluate model performance, we employ zero-shot, few-shot, and chain-of-thought (CoT) prompting across a diverse set of Large Language Models (LLMs), Small Language Models (SLMs), and Multimodal Large Language Models (MLMs). By providing a comprehensive multilingual and multicultural sports benchmark, *CultSportQA* establishes a new standard for assessing AI's ability to understand and reason about traditional sports.

## 1 Introduction

Sports serve as a powerful medium for cultural exchange, uniting people across diverse backgrounds and traditions (Coakley, 2021). The study of various sports and athletic practices provides valuable insights into societal values, historical narratives, and social structures of the communities that develop and embrace them (Guttmann, 2004). Furthermore, sports play a crucial role in shaping language, acting as a conduit for cultural knowledge and identity formation (Maguire, 2011). Sports' terminology, rituals, and adaptations showcase community history, societal change, and cultural identity (Dyck, 2012).

Researchers have long utilized sports as a lens to analyze cultural dynamics, providing a framework

for quantifying differences in athletic traditions across regions (Bairner, 2015). Many sports share fundamental principles but have evolved uniquely in different societies, leading to variations in rules, playing styles, and even terminologies (Eichberg, 2010). Different nations have adapted bat-and-ball games uniquely, such as baseball, cricket, and pesäpallo, with distinct rules. Similarly, "football" varies in meaning across regions, referring to American football, soccer, or Australian rules football (Mangan, 1996).

Language Models (LMs) have revolutionized natural language understanding, content generation, and decision-making, becoming indispensable across industries such as education, governance, and entertainment, healthcare (Jain et al., 2022; Brown et al., 2020; Devlin et al., 2019; Ghosh et al., 2024a,c, 2025; Ghosal et al., 2025). From Large Language Models (LLMs) to Multimodal Language Models (MLMs) and Small Language Models (SLMs) \*, these advancements have enabled seamless communication and efficient problem-solving (Ouyang et al., 2022; Ghosh et al., 2024b). 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 sports (Bender et al., 2021).

Traditional and indigenous sports are deeply intertwined with local histories, societal values, and cultural identities (Blodgett et al., 2020). Despite their significance, LMs are predominantly trained and evaluated on globally popular sports, often overlooking regional variations and culturally unique athletic traditions. This oversight risks reinforcing biases, inaccuracies, and stereotypes, further marginalizing underrepresented communities. Conversely, models capable of understanding cultural contexts not only enhance performance but

\*Corresponding authors: yanghaiqin@sztu.edu.cn and jose.moreno@irit.fr.

\*Any model with less than or equal to 7B parameters are considered as Small Language Models (SLMs)

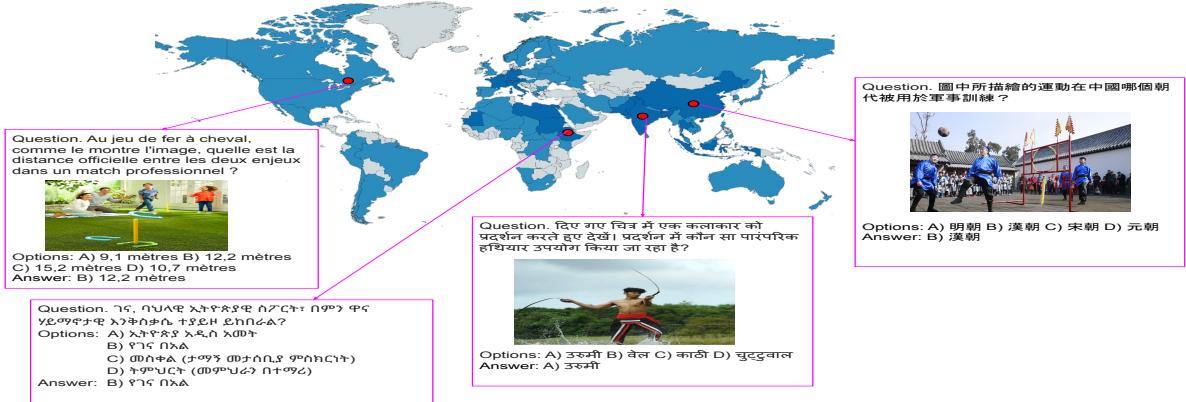


Figure 1: **CultSportQA** is a diverse benchmark featuring 11 languages, with questions manually created and verified by native language experts. It covers three key aspects of traditional sports across two modalities, text and image, emphasizing mid to low-resource languages and sports originating from 11 countries across 3 continents. These sports, now played in 60 countries across 6 continents, are depicted with dark blue for their origins and light blue for their current reach. **CultSportQA** offers a wide range of question formats, including multiple-choice questions (MCQs) and both short and long visual question-answering (VQA) tasks.

also promote inclusivity and equity in AI applications.

**Motivation for *CultSportQA* Dataset:** Existing benchmarks in sports understanding and reasoning primarily focus on globally well-known sports and are often limited in scope. For instance, *SportQA* (Xia et al., 2024a) is an unimodal dataset that only supports English and covers widely recognized sports. Similarly, *SportU* (Xia et al., 2024b) is the first dataset benchmarking multimodal large language models (MLLMs), but it includes only seven globally popular sports, all in English. However, no existing benchmark comprehensively captures the cultural nuances of sports reasoning across multiple languages, diverse cultural contexts, and visual question answering (VQA). To bridge this gap, we introduce the **largest multicultural and multilingual sports benchmark** to date, namely, ***CultSportQA***, featuring approximately 33,000 sports-related questions spanning countries originating from 3 continents and 11 countries and now being expanded in 6 continents and across 60 countries<sup>†</sup>. Our benchmark evaluates Large Language Models (LLMs), Small Language Models (SLMs)<sup>‡</sup>, and Multimodal Language Models (MLLMs) across eleven languages. The questions are systematically categorized across two modalities and each modality is subdivided into three key types of questions:

history-based, rule-based, and scenario-based<sup>§</sup>. By offering a comprehensive, culturally diverse, and multilingual benchmark, ***CultSportQA*** establishes a new standard for assessing the ability of language models to understand and reason about sports in a more inclusive, global, and culturally aware manner. **Research Questions:** This research aims to address the following key questions:

- *How do different categories of models—Large Language Models (LLMs), Small Language Models (SLMs), and Multimodal Large Language Models (MLLMs)—perform on the *CultSportQA* dataset?*
- *What trends and patterns emerge in model performance across the various question types in the *CultSportQA* dataset, including two modalities, text and image, and each modality covering three key types of questions: history-based, rule-based, and scenario-based*
- *What are the performance trends of language models across different countries and languages in Asia, Africa, and Europe?*

Our **key contributions** in this research are summarized as follows:

1. ***CultSportQA* Dataset:** We introduce the first and most comprehensive QA dataset focusing on

<sup>†</sup>The complete list is present in the Appendix

<sup>‡</sup>Any model below 7B parameters is considered a Small Language Model in this work.

<sup>§</sup>The MLLMs are used only for benchmarking image-based questions.

traditional sports, covering games played across 60 countries, 6 continents, and 11 languages.

**2. Diverse Question Types:** The dataset includes 33,000 questions spanning two modalities of data and each modality covers three categories, challenging AI models to reason through textual and visual input while incorporating multilingual and cultural contexts.

**3. Comprehensive Benchmarking:** We evaluate 8 state-of-the-art LLMs and five SLMs alongside four MLLMs, identifying critical gaps in their ability to reason about traditional and culturally sports-nuanced queries.

**4. Insights on AI Performance:** Using zero-shot, few-shot learning, and chain-of-thought (CoT) prompting, we analyze model strengths and limitations, advancing the understanding of AI performance in culturally rich domains.

**5. Expanding NLP in Sports:** Our work explores new applications of NLP in preserving cultural heritage, enriching sports journalism, and enhancing communication between athletes and coaches, particularly in regional and traditional sports contexts.

**6. Public Availability:** The CultSportQA dataset is available at: <https://github.com/M-Groot7/CultSportQA>.

By addressing the challenges of cultural underrepresentation in AI, *CultSportQA* establishes itself as a robust benchmark for evaluating and improving AI systems. This research contributes to fostering inclusivity and equity in AI applications while advancing the intersection of NLP and culturally rich domains like traditional sports.

## 2 Related Work

### 2.1 Sports Datasets and Benchmarks

Sports datasets are rapidly expanding, enabling diverse applications, such as sentiment analysis (Baca et al., 2023; Ljajić et al., 2015), game prediction, and video enhancement using computer vision (Beal et al., 2021; Oved et al., 2020). While datasets like **SportQA** (Xia et al., 2024a) and **BoolQ** (Clark et al., 2019) have significantly advanced sports-related question answering (QA), many existing datasets primarily focus on historical events and overlook critical aspects such as **rules, strategies, and complex situational analysis** (Oved et al., 2020; Huang et al., 2020). For instance, the **Sports Understanding** subtask in **BIG-bench (2023)** assesses **athlete recognition and**

**action identification** but lacks depth in **situational comprehension**. Among existing benchmarks, Xia et al. (2024a) introduced one of the largest **unimodal text-based** datasets, covering approximately **70,000 questions in English**. Meanwhile, Xia et al. (2024b) developed a **multimodal sports dataset** with **12,048 questions**, benchmarked on leading **Multimodal Large Language Models (MLLMs)**. Additionally, Yang et al. (2024) explored **multimodal sports understanding** by benchmarking various **video-language models** for **sports-related tasks**.

### 2.2 Cultural Benchmarks for MLLMs and LLMs

Several previous studies have focused on developing culturally relevant VQA benchmarks, including FM-IQA (Gao et al., 2015), MCVQA (Gupta et al., 2020), xGQA (Pfeiffer et al., 2021), MaXM (Changpinyo et al., 2022), MTVQA (Tang et al., 2024), MABL (Kabra et al., 2023), MAPS (Liu et al., 2024), and MaRVL (Liu et al., 2021). Additionally, datasets such as CVQA (Romero et al., 2024), CulturalVQA (Nayak et al., 2024) and ALM-bench (Vayani et al., 2024) provide VQA resources that encompass a wide range of regions and cultural themes, including food, with CVQA offering multilingual questions alongside English translations. SEA-VQA (Urailetrprasert et al., 2024) specifically benchmarks the Southeast Asian region, while FoodieQA (Li et al., 2024c), World Wide Dishes (Magomere et al., 2024) and WORLD-CUISINES (Winata et al., 2025) focus exclusively on food-related benchmarks. Our work is driven by a similar objective, using traditional sports as a cultural lens; however, it distinguishes itself with a significantly larger dataset and broader coverage of languages. Recent research has assessed LLMs' sociocultural reasoning using frameworks like the World Values Survey and Hofstede's dimensions, highlighting gaps in adapting to user-specific and non-Western cultural contexts (Johnson et al., 2022; Atari et al., 2023; Masoud et al., 2023; Seth et al., 2024; Li et al., 2024b; AlKhamissi et al., 2024; Durmus et al., 2023). While synthetic personas and fine-tuning have improved cultural adaptability and performance in tasks like hate speech detection, regional language evaluations still lag behind English benchmarks (Dwivedi and Patel, 2024; Shen et al., 2017). These findings emphasize the need for robust multilingual strategies to enhance LLMs' cultural competence.

### 3 Construction of *CultSportQA*

This section outlines the various stages of the creation of the benchmark *CultSportQA*.

#### 3.1 Manual Data Collection

The creation of the *CultSportQA* traditional sports dataset followed a carefully structured, multi-phase process to ensure comprehensive coverage and high-quality standards. Domain experts and country-specific annotators contributed 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 was curated using information from six carefully selected and credible sources to ensure comprehensive coverage of traditional sports across India, Pakistan, Bangladesh, Italy, France, China, Thailand, Indonesia, Sudan, Ethiopia, and Germany. These sources include Wikipedia, National Heritage and Sports Boards, Local Sports Blogs, Cultural Journals, News Outlets, and Academic Publications. Wikipedia served as a foundational resource for historical and rule-based information, while academic publications added scholarly depth with technical analyses. National heritage and sports boards contributed authentic cultural context and historical relevance. Cultural journals offered insights into the societal impact and evolution of these sports. Local blogs provided region-specific practices and community perspectives, and news outlets highlighted current events and preservation efforts. The resulting questions span historical facts, gameplay rules, scenario-based reasoning, and image-based understanding, capturing both the depth and diversity of traditional sports.

**Annotators Background:** The *CultSportQA* dataset was created with contributions from native speakers and cultural experts from 11 countries covering 11 languages across three continents. The annotators were selected such that they are fluent in the language of their respective country with most having over 10 years of residency in their respective regions. They were required to be fluent in their local languages and be aware of their cultural nuances. Contributors who provided significant input, such as validated question-answer pairs, were credited as co-authors. The team followed detailed guidelines and underwent training to ensure the questions reflected cultural relevance and diversity.

Additionally, a peer-validation process ensured the accuracy and consistency of the annotations, resulting in a culturally rich and multilingual VQA benchmark. *The detailed guidelines are discussed by showing an example in the Appendix.*

**Dataset Organization:** The *CultSportQA* dataset is organized into four question types: history-based, rule-based, scenario-based, and image-based. The dataset follows a multiple-choice question (MCQ) format with four options (A, B, C, D), where one is correct. Each question-answer (QA) pair includes metadata such as continent, country, sport name, and question type. The questions are divided into text-based, evaluated using Large Language Models (LLMs) and Small Language Models (SLMs), and image-based, assessed by Multimodal Large Language Models (MLLMs). **History-based** questions test the model's knowledge of a sport's origins and cultural significance. **Scenario-based** questions assess the model's ability to determine the best move in a game situation to score maximum points. **Rule-based** questions evaluate the model's understanding of the fundamental rules of the sport depicted in the text or image.

#### 3.2 Annotation Process

We outline four main steps in annotation below:

**1. Team Structure.** The annotation team consisted of experienced experts with deep cultural knowledge and fluency in their respective countries' languages, ensuring both linguistic clarity and cultural authenticity. To promote diversity and thoroughness, we aimed to hire at least three annotators from each country. Within each team, two-thirds of the annotators were responsible for creating questions based on provided guidelines, leveraging their knowledge of traditional sports. The remaining annotator was tasked with validating and filtering out questions that failed to meet quality standards.

**2. Question Formation.** For each selected textual passage or image, the annotator's first task was to verify whether the content aligned with the cultural sport associated with the country. Content unrelated to the regional sport was immediately rejected. If the content was relevant, the annotator created questions focusing on rules, history, and scenarios. Each question had to be complete, self-contained, and understandable without additional context. The questions followed a multiple-choice format, consisting of four options, with only one correct answer. The final annotated format included

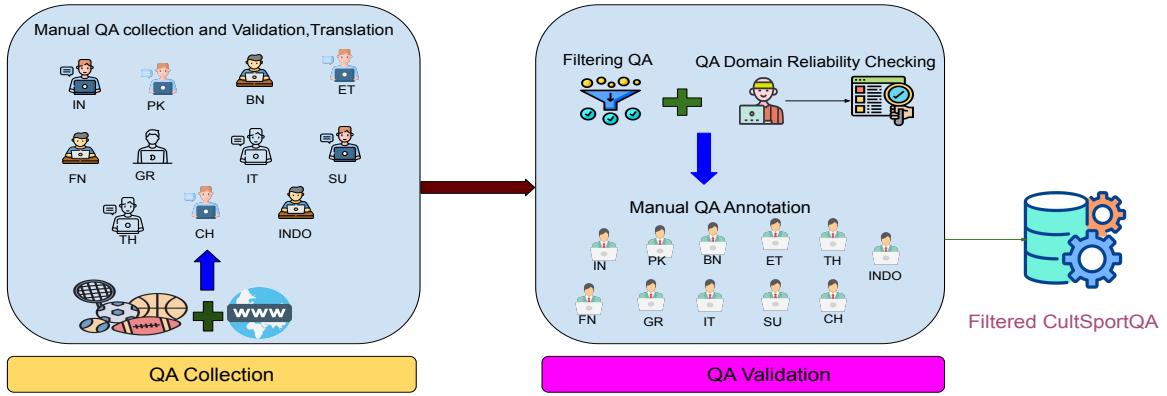


Figure 2: **CultSportQA** Manual Data Collection Pipeline: The data collection process involved two key stages. In the first stage, annotators gathered data sources and generated questions, drawing from their respective cultural backgrounds and languages. In the second stage, annotators reviewed and verified the questions to ensure cultural authenticity and maintain high translation quality.

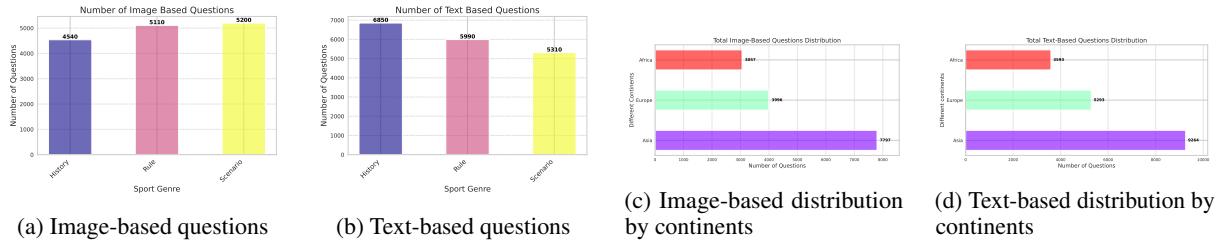


Figure 3: Distribution of image-based and text-based questions across different question types and continents.

the source passage, a relationship attribute indicating the question’s context, the type of question (e.g., history-based, rule-based, or scenario-based), and the four answer options. After a question is constructed it is translated to the regional language that the annotator is familiar with.<sup>¶</sup>

**3. Training and Guidelines.** Annotators underwent comprehensive training that included objectives of the **CultSportQA** dataset, definitions and examples of question types, and best practices for maintaining consistency and cultural sensitivity. Detailed guideline documents provided templates, metadata tagging standards, and examples of appropriate cultural representation. Additional language and cultural training sessions emphasized the use of local terminologies and traditions. *The complete process has been shown in the Appendix.*

**4. Quality Assurance and Cross-Validation.** A rigorous quality assurance process was conducted through multi-step validations. Each question-answer pair underwent cross-validation by at least

one annotator, who understood the basic requirement to be qualified to be included in the dataset, and also the translation quality. Image-based questions were reviewed for proper alignment between visual elements and textual prompts. Spot checks and random sampling were performed by quality analysts to maintain clarity and consistency. Bias mitigation measures ensured a balanced representation of sports across regions and question types, with cultural sensitivity reviews eliminating stereotypes or offensive content. *The complete validation guidelines are added in the Appendix.*

## 4 Statistical Analysis of **CultSportQA**

The **CultSportQA** dataset shows a balanced mix of text-based and image-based questions, with a slight dominance of text-based questions comprising 18,150 questions over the visual ones comprising 14,850. Image-based questions focus more on sports scenarios, testing practical understanding, while text-based questions lean toward sports history, highlighting knowledge recall. Regionally, Asia has the highest question count in both types, reflecting a strong focus on Asian sports themes.

<sup>¶</sup>The annotators were paid at the rate between 0.10 dollar to 0.50 dollar per example, depending on country exchange rate and difficulty of annotation.

Figure-3 shows the distribution of text- and image-based questions across question categories and continents. *More statistical analysis is shown in the Appendix.*

## 5 Experimental Setup

### 5.1 Models

To conduct a comprehensive evaluation of our benchmark, *CultSportQA*, we carried out an extensive assessment across a diverse range of language models across different modalities of text and image. For text, our evaluation encompassed leading LLMs, including Llama2-13B (Dubey et al., 2024), Llama3-8B (Dubey et al., 2024), Llama3.1-70B-Instruct (Dubey et al., 2024), and GPT-3.5. Additionally, we tested several SLMs such as Mistral-7B (Jiang et al., 2023), Gemma-7B (Team et al., 2024), Phi-3-medium-4k (Abdin et al., 2024), Llama3.2-3B (Dubey et al., 2024) (Touvron et al., 2023), BLOOMZ-3B (Muennighoff et al., 2022) Qwen-2 (1.5B) (Bai et al., 2023), and FLAN-T5-780M (Chung et al., 2024) and BART (Lewis, 2019). Beyond text-based models, we also evaluated a range of MLLMs to test the reasoning of sports across multilingual settings. In this category, we assessed InstructBLIP (Panagopoulou et al., 2023) and mBLIP—a BLIP-2-based model (Geigle et al., 2023) supporting 96 languages—where we tested two variations: PaliGemma-3B (Beyer et al., 2024) and LLaVA-7B (Liu et al., 2023). Finally, for a more holistic comparison, we incorporated the proprietary model GPT-4o into our evaluation.

### 5.2 Evaluation Setup

We conducted a comprehensive evaluation of the *CultSportQA* dataset, which includes text and image-based **Multiple-Choice Questions (MCQs)** grouped into three categories: **1. Cultural and Historical Knowledge**, **2. Rule Comprehension**, and **3. Scenario-Based Reasoning**. To assess model performance across languages and modalities, we used three prompting techniques: **zero-shot**, **few-shot (3-shot)**, and **Chain of Thought (CoT)**, with the **temperature parameter set to 0** for consistency. **Accuracy** was the sole evaluation metric. **Open-source models** used **16-bit floating-point precision** and **greedy decoding**, while **proprietary models** were accessed via APIs. Predictions were based on the highest output probability, ensuring a standardized evaluation process.

## 6 Experimental Results

### 6.1 Main Results

The overall performances on the *CultSportQA* dataset across various LLMs, SLMs, and MLLMs are presented in Table-2 and Table-3

**Performance of LLMs:** LLMs consistently outperformed other model types, with GPT-4o .87 and Llama-3.1-70B .84 leading across all evaluation categories. These models demonstrated superior accuracy in language-based tasks, question-type performance, and continent-based variations. GPT-3.5 .81, while slightly behind the top two, remained highly competitive, particularly excelling in history-based (83.1%) and rule-based (82.7%) questions.

The performance comparison shows **GPT-4o** as the strongest model, excelling in few-shot setting due to its superior ability in-context learning. GPT-3.5 performs well but lags behind GPT-4o, especially with more shots. Among LLaMA models, **LLaMA-3.1-70B** leads, benefiting from its larger size, LLaMA2-13B performs moderately, while Pretrained Language Models (PLMs) **BART** ranks lowest, highlighting its limited capacity. Overall, larger models and few-shot learning drive the highest performance.

**Performance of SLMs:** Among SLMs, **Mistral-7B** achieves the highest performance, particularly excelling with few-shot learning, highlighting its strong generalization capabilities. **Gemma-7B** and **Phi-3-medium** follow closely, showing competitive results with steady improvement across settings. **LLaMA-3.2-3B** performs moderately, benefiting from few-shot examples but trailing behind larger models. **BLOOMZ-3B, Qwen-2(1.5B), and FLAN-T5-780M** exhibit the lowest performance, with FLAN-T5 performing weakest, reflecting its limited capacity for complex reasoning. Overall, performance scales with model size and improves with few-shot learning, with Mistral and Gemma emerging as the strongest contenders among SLMs.

**Performance of MLLMs:** The performance comparison of MLLMs shows that **InstructBLIP** outperforms all other models, demonstrating strong reasoning and adaptability, especially in the few-shot setting. **mBLIP** performs well but trails behind **InstructBLIP**, indicating slightly weaker multimodal integration. **PaliGemma-3B** and **LLaVA-7B** show lower performance, with **LLaVA-7B** performing the weakest, highlighting its limitations in complex tasks. All models benefit from few-shot

Dataset	Number of Samples	Number of Sports	Cultural Aspects	Number of Language	Modalities	Type Of Questions
SportQA (Xia et al., 2024a)	70,000	36	No	1	Text	MCQ
SPORTU (Xia et al., 2024b)	12,948	7	No	1	Text+video	MCQ
BoolQ (Clark et al., 2019)	15,942	Not specified	No	1	No	YES/NO
Sports-QA (Li et al., 2024a)	<b>94,000</b>	4	No	1	No	Descriptive
<i>CultSportQA (ours)</i>	33,000	<b>84</b>	<b>Yes</b>	<b>11</b>	Both Text and Text+Image	<b>MCQ</b>

Table 1: Comparison of our dataset with other Sports datasets. The metadata includes Number of Samples (number of questions), Number of Sports, Cultural Aspects (whether the data considers cultural nuances), Number of Languages, and Modalities (whether the data includes multimodal questions), and Type Of Questions.

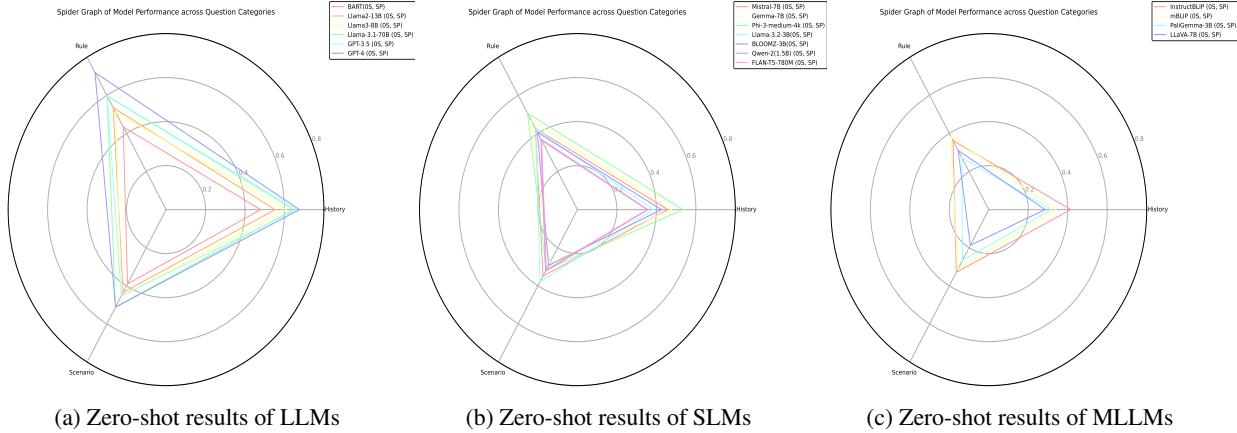


Figure 4: Zero-shot results of language models across different question types.

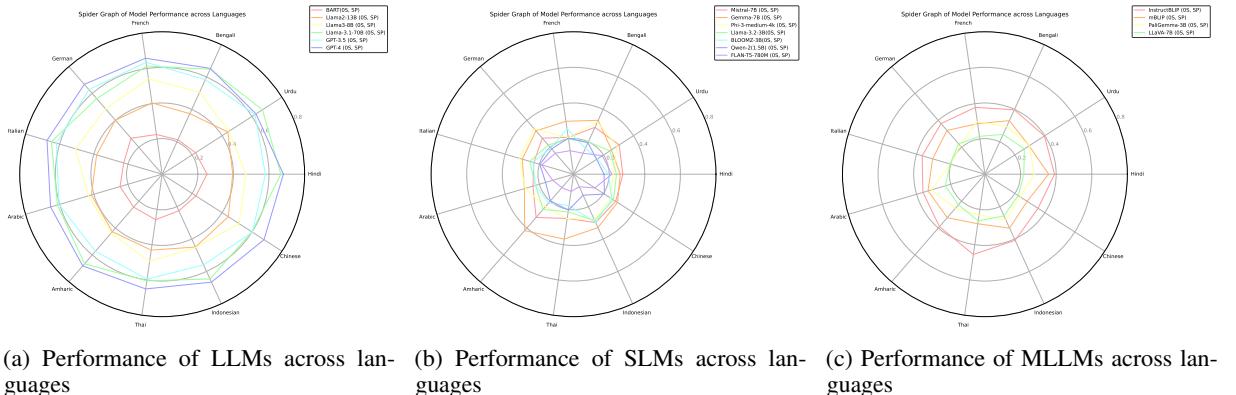


Figure 5: Average results of language models on the *CultSportQA* dataset classified on the basis of languages.

learning, with **InstructBLIP** showing the greatest improvement, underscoring its superior in-context learning capability.

**Performance across Languages:** In the case of LLMs, GPT-4o leads in performance across all languages, closely followed by GPT-3.5, while LLaMA-3B (70B) outperforms LLaMA-2-13B in most languages; BART performs the weakest, showing significant gaps, especially in non-Latin scripts like Amharic and Thai. For SLMs, Mistral-7B and Gemma-7B lead across most languages, with Mistral-7B excelling particularly in Arabic and Italian; Phi-3-medium and LLaMA-3B

show moderate performance, while BLOOMZ-3B, Qwen-2.5B, and FLAN-T5-780M lag, especially in non-Latin languages like Amharic and Thai. And finally, concerning MLLMs, InstructBLIP leads overall, excelling in Hindi, Chinese, and Arabic, while mBLIP performs well but falls behind in Urdu and German; PoliGemma-3B shows moderate performance, outperforming LLaVA-7B, which struggles across most languages.

**Performance across Question types:** Figure-4 analyzes the performance of LLMs, SLMs, and MLLMs in a zero-shot setup across various kinds of questions related to history, rule-based, and

Model	BART	Llama2-13B	Llama3-8B	Llama-3.1-70B	GPT-3.5	GPT-4o	Mistral-7B	Gemma-7B	Phi-3-medium	Llama-3.2-3B	BLOOMZ-3B	Qwen-2(1.5B)	FLAN-T5-780M
Zero-shot	24.24	41.47	48.09	62.07	59.99	<b>66.29</b>	23.71	30.56	31.45	24.67	21.67	18.60	15.45
Few-shot	27.24	43.72	51.87	64.54	63.93	<b>69.53</b>	33.73	33.78	34.67	27.08	25.11	21.98	18.90
CoT	31.93	46.24	54.90	69.34	67.18	<b>74.51</b>	40.98	38.98	39.56	30.23	28.79	26.29	23.76

Table 2: Performance comparison of various LLMs and SLMs in the text-based questions of *CultSportQA*

Qu. भारतीय इतिहास में लगोरी का पहली बार उल्लेख कब किया गया था? A. प्राचीन B. मध्यकालीन C. आधुनिक D. पार्वीतिहासिक Correct:- A. प्राचीन, Predicted :- A. प्राचीन	Qu. Wenn ein Quoit außerhalb des Ziels landet, wie wird das genannt? A. Fehler B. Verfehren C. Überwurf D. Null Correct:- B. Verfehren, Predicted :- D. Null
Qu. Quel est le nombre maximum de joueurs par équipe à la pétanque ? A. Deux B. Trois C. Quatre D. Six Correct:- B. Trois, Predicted :- B. Trois	Qu. গোলা ছুটে "রামান"-এর ভূমিকা কী? A. লুকানো B. তাঁতা করা C. প্রতিরক্ষা করা D. পালানো Correct:- D. পালানো, Predicted :- C. প্রতিরক্ষা করা
Qu. जीवे दिखाए गए वित्र में क्या प्रदर्शित किया गया है, जो एक पारंपरिक भारतीय युद्धकला का प्रतिनिधित्व करता है? A. कुशली B. तलवारबाजी C. मुक्केबाजी D. मुश्ती युद्ध Correct:- D. मुश्ती युद्ध, Predicted:- D. मुश्ती युद्ध	Qu. Was ist das Hauptziel im Radball, wie im Bild gezeigt? A. Tore mit einem Fahrrad schießen B. Fahrräder rennen C. Den Ball mit den Händen schlagen D. Den Ball vermeiden Correct:- A. Tore mit einem Fahrrad schießen, Predicted:- B. Fahrräder rennen

Figure 6: The LHS displays the correctly answered questions, while the RHS highlights the incorrectly answered ones by the language models on the *CultSportQA* dataset.

Model	InstructBLIP	mBLIP	PoliGemma-3B	LLaVA-7B
Zero-shot	<b>38.90</b>	32.90	27.33	24.95
Few-shot	<b>44.83</b>	36.22	31.67	29.02
CoT	<b>49.45</b>	40.85	37.90	35.37

Table 3: Performance comparison of various MLLMs in the image-based questions of *CultSportQA*

scenario-based of *CultSportQA* dataset. With respect to LLMs, GPT-4o leads across all categories, with GPT-3.5 closely following, while LLaMA models show moderate performance, with the 70B variant surpassing other versions; BART performs the lowest across all categories. Among SLMs, Mistral-7B leads across all categories, especially in History and Rule, with Gemma-7B and LLaMA-3B showing competitive performance, while BLOOMZ-3B and FLAN-T5-780M lag behind, particularly in Scenario-based questions. Among MLLMs, InstructBLIP outperforms all models, particularly excelling in History and Scenario categories, while mBLIP and PoliGemma-3B show moderate performance across all categories, with LLaVA-7B trailing, especially in Scenario-based questions. *Appendix contains the results of language models in COT and few-shot across languages.*

## 6.2 Error Analysis

To evaluate the strengths and limitations of the best-performing models on the *CultSportQA* dataset, we conducted an error analysis, grouping questions into correctly and incorrectly answered sets, as shown in Figure 6. The analysis highlights key patterns of success and failure. On the left (LHS), correct predictions stem from strong keyword associations (e.g., “Lagori” and “Petanque”) and well-structured questions with distinct answer choices. On the right (RHS), errors are driven by limited knowledge of culturally nuanced sports (e.g., “Quoits”) and confusion caused by ambiguous or overlapping answer options (e.g., “Game of Skill” vs. “Strategic Team Game”). These issues point to gaps in cultural coverage and underscore the need for more diverse training data to improve model performance on sports-related queries.

## 7 Conclusion

In this work, we introduced *CultSportQA*, a comprehensive benchmark designed to evaluate language models’ understanding of Asian, African, and European traditional sports. The dataset, consisting of 33,000 curated question-answer pairs from 11 countries, covers key aspects such as rules, cultural significance, and historical context. Evalu-

ations with leading models revealed notable gaps in answering traditional sport-specific questions, highlighting biases likely caused by training data limitations. *CultSportQA*, built for quality and cultural sensitivity, advances inclusive AI research. Future expansions will add more languages and traditional sports to enhance its impact.

## 8 Limitations

While this study represents one of the most comprehensive evaluations of language models in the context of traditional sports and cultural knowledge, several notable limitations must be acknowledged:

**(1) Limited Geographic Scope:** The dataset and analysis are focused solely on the regional sports of 11 countries spanning across 3 continents. While these regions provide valuable insights, the dataset can still be extended to other countries across different continents. In the future, we will expand the dataset to include regional sports from additional countries, which could offer a broader understanding and uncover more diverse trends.

**(2) Limited Representation of Traditional Sports:** Although the study covers 84 traditional sports (46 from Asia, 25 from Europe, and 13 from Africa), which is the largest sports cultural data set, the data set may not fully represent the rich tapestry of traditional sports across these continents. Future iterations could expand the dataset to include a wider range of sports and introduce diverse question-answering tasks, such as True/False questions, adversarial questions, and scenario-based reasoning.

**(3) Limited Language and Cultural Coverage:** The *CultSportQA* dataset spans 11 languages from 11 different countries, providing a valuable initial benchmark for evaluating language models. However, expanding the dataset to include more low-resource languages would enhance its diversity and inclusivity. Such an expansion would not only promote traditional sports at the grassroots level but also enable more comprehensive assessments of language models across diverse linguistic and cultural contexts.

**(4) Limited scope of Modalities:** The dataset includes only text and image modalities, lacking other potential modalities. It specifically focuses on multimodal combinations that require reasoning across multiple modalities simultaneously to answer queries effectively. The complexity involved in creating multimodal questions is high, but we

remain committed to continually updating and expanding the dataset to enhance its scope and depth.

## 9 Ethics Statement

**Data Collection and Bias Mitigation:** The data used in the development of *CultSportQA* was collected from publicly accessible platforms, as outlined in Section 3.1. These platforms were carefully selected to ensure authenticity, making *CultSportQA* a significant milestone in establishing a standardized and inclusive benchmark for evaluating Asian, European, and African traditional sports. The dataset sources were thoroughly verified by annotators through multiple rounds of group discussions. Following the collection process, annotators curated the dataset by extracting portions suitable for question generation and discarding irrelevant metadata. To prevent language bias, the dataset comprises 3,000 data points for each of the 11 selected languages, ensuring balanced representation.

**Human Annotation:** Human annotators were key in creating, checking, and translating questions to make sure the dataset truly reflects cultural and sports contexts. The team included 42 experts from 11 countries, with backgrounds in sports, linguistics, and related fields. Most were native or bilingual speakers with over 15 years of sports experience, aged between 30 and 50. They received training on the dataset's goals, question types, and sports-specific guidelines. To maintain quality, a separate sub-team cross-checked the work. Throughout the process, fairness and inclusivity were emphasized, avoiding stereotypes and ensuring cultural diversity was respected.

## Acknowledgments

This work was partially funded by the Geo-R2LLM CHIST-ERA project. The experiments presented were partially conducted using the OCCIDATA platform administered by IRIT (CNRS/University of Toulouse).

## References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating

cultural alignment of large language models. *arXiv preprint arXiv:2402.13231*.

Mohammad Atari, Jonathan Haidt, Jesse Graham, Sena Koleva, Sean T Stevens, and Morteza Dehghani. 2023. Morality beyond the weird: How the nomological network of morality varies across cultures. *Journal of Personality and Social Psychology*.

Luis Baca, Nátila Ardiles, Jose Cruz, Wilson Mamani, and John Capcha. 2023. Deep learning model based on a transformers network for sentiment analysis using nlp in sports worldwide. In *International Conference on Advances in Computing and Data Sciences*, pages 328–339. Springer.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Alan Bairner. 2015. *Assessing the sociology of sport: On race and ethnicity*. *International Review for the Sociology of Sport*, 50(4-5):379–384.

Ryan Beal, Stuart E. Middleton, Timothy J. Norman, and Sarvapali D. Ramchurn. 2021. Combining machine learning and human experts to predict match outcomes in football: A baseline model. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 15447–15451. AAAI Press.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Su Lin Blodgett, Solon Barocas, et al. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *ACL 2020*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Soravit Changpinyo, Linting Xue, Michal Yarom, Ashish V Thapliyal, Idan Szpektor, Julien Amelot, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Jay Coakley. 2021. *Sports in Society: Issues and Controversies*, 13th edition. McGraw-Hill Education.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Sanjay Kumar Dwivedi and Rahul Patel. 2024. Exploring the intersections: Anthropological insights into studying language and culture. *State Institute of Education, Allahabad*, 30:171–182.

Noel Dyck. 2012. *Sport, Anthropology, and Culture*. University of Toronto Press.

Henning Eichberg. 2010. *Bodily Democracy: Towards a Philosophy of Sport for All*. Routledge.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

Soumya Suvra Ghosal, Vaibhav Singh, Akash Ghosh, Soumyabrata Pal, Subhadip Baidya, Sriparna Saha, and Dinesh Manocha. 2025. Relic: Enhancing reward model generalization for low-resource indic languages with few-shot examples. *arXiv preprint arXiv:2506.16502*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024b. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Gaurav Pandey, Dinesh Raghu, and Setu Sinha. 2024c. Healthalignsumm: Utilizing alignment for multimodal summarization of code-mixed healthcare dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11546–11560.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating complex statistical infographics from documents. *arXiv preprint arXiv:2507.20046*.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. 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*, pages 900–913.

Allen Guttmann. 2004. *Sports: The First Five Millennia*. University of Massachusetts Press.

Kuan-Hao Huang, Chen Li, and Kai-Wei Chang. 2020. Generating sports news from live commentary: A chinese dataset for sports game summarization. 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*, pages 609–615.

Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. A survey on medical document summarization. *arXiv preprint arXiv:2212.01669*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Indra Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. *arXiv preprint arXiv:2305.16171*.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Haopeng Li, Andong Deng, QiuHong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. 2024a. Sports-QA: A large-scale video question answering benchmark for complex and professional sports. *arXiv preprint arXiv:2401.01505*.

Huihan Li, Liwei Jiang, Jena D Hwang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024b. Culture-gen: Revealing global cultural perception in language models through natural language prompting. *arXiv preprint arXiv:2404.10199*.

Wenyan Li, Xinyu Zhang, Jiaang Li, Qiwei Peng, Raphael Tang, Li Zhou, Weijia Zhang, Guimin Hu, Yifei Yuan, Anders Søgaard, et al. 2024c. Foodieqa: A multimodal dataset for fine-grained understanding of chinese food culture. *arXiv preprint arXiv:2406.11030*.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. *Visual instruction tuning*. In *Advances in Neural Information Processing Systems*. NeurIPS 2023 Oral.

Adela Ljajić, Ertan Ljajić, Petar Spalević, Branko Arsić, and Darko Vučković. 2015. Sentiment analysis of textual comments in field of sport. In *24nd International Electrotechnical and Computer Science Conference (ERK 2015), IEEE, Slovenia*.

Jabez Magomere, Shu Ishida, Tejumade Afonja, Aya Salama, Daniel Kochin, Foutse Yuehgooh, Imane Hamzaoui, Raesetje Sefala, Aisha Alaagib, Elizaveta Semenova, et al. 2024. You are what you eat? feeding foundation models a regionally diverse food dataset of world wide dishes. *arXiv preprint arXiv:2406.09496*.

Joseph Maguire. 2011. *Sport and Globalization: A Critical Introduction*. Routledge.

James A. Mangan. 1996. *Tribal Identities: Nationalism, Europe, Sport*. Frank Cass Publishers.

Reem I Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2023. Cultural alignment in large language models: An explanatory analysis based on Hofstede's cultural dimensions. *arXiv preprint arXiv:2309.12342*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Nadav Oved, Amir Feder, and Roi Reichart. 2020. Predicting in-game actions from interviews of NBA players. *Computational Linguistics*, 46(3):667–712.

Artemis Panagopoulou, Le Xue, Ning Yu, Junnan Li, Dongxu Li, Shafiq Joty, Ran Xu, Silvio Savarese, Caiming Xiong, and Juan Carlos Niebles. 2023. X-instructblip: A framework for aligning x-modal instruction-aware representations to LMs and emergent cross-modal reasoning. *arXiv preprint arXiv:2311.18799*.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

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*.

Agrima Seth, Sanchit Ahuja, Kalika Bali, and Sunayana Sitaram. 2024. *DOSA: A dataset of social artifacts from different Indian geographical subcultures*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5323–5337, Torino, Italia. ELRA and ICCL.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2017. *Style transfer from non-parallel text by cross-alignment*. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6830–6841.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. *Llama: Open and efficient foundation language models*. *Preprint, arXiv:2302.13971*.

Norawit Urailertprasert, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. 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)*, pages 173–185.

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

Genta Winata, Frederikus Hudi, Patrick Amadeus Irawan, David Anugraha, Rifki Afina Putri, Yutong Wang, Adam Nohejl, Ubaidillah Prathama, Nedjma Ousidhoum, Afifa Amriani, et al. 2025. Worldcuisines: A massive-scale benchmark for multilingual and multicultural visual question answering on global cuisines. In *2025 Annual Conference of the Nations of the Americas Chapter of the Association*

for Computational Linguistics. Association for Computational Linguistics.

Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zehzi Chen, Yan Zhu, Yuan-fang Wang, and Weining Shen. 2024a. Sportqa: A benchmark for sports understanding in large language models. *arXiv preprint arXiv:2402.15862*.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. 2024b. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. *arXiv preprint arXiv:2410.08474*.

Zhengbang Yang, Haotian Xia, Jingxi Li, Zehzi Chen, Zhuangdi Zhu, and Weining Shen. 2024. Sports intelligence: Assessing the sports understanding capabilities of language models through question answering from text to video. *arXiv preprint arXiv:2406.14877*.

## A Appendix

The Appendix includes information about discussion about models, information about annotators wage annotators distribution across countries, an annotation example of a data point, prompts for evaluation, additional statistical analysis of the dataset *CultSportQA*, further results on languages (in few and CoT questions), four categories of questions (in few-shot and CoT questions), performance across continents, and many more qualitative examples from our benchmark *CultSportQA*.

## B Discussion about Models

To ensure a holistic evaluation, our study includes models from diverse categories—Large Language Models (LLMs,  $>7B$  parameters), Small Language Models (SLMs,  $<=7B$  parameters), and Multimodal Large Language Models (MLLMs), the latter being essential for addressing the visual question answering component of our benchmark.

### B.1 Large Language Models (LLMs)

- **Meta’s LLaMA Series:**

- **LLaMA 2 13B and LLaMA 3 8B:** Open-source models optimized for general-purpose language tasks.
- **LLaMA 3.1 70B Instruct:** Supports a 128K token context window and multilingual capabilities across eight languages, suitable for complex reasoning and enterprise applications.

### B.2 Small Language Models (SLMs)

- **Mistral 7B (Mistral AI):** Employs grouped-query and sliding window attention mechanisms, offering efficient inference and handling of longer sequences, ideal for deployment in resource-constrained environments.
- **Gemma 7B (Google DeepMind):** Demonstrates strong performance in code generation and mathematical problem-solving tasks, outperforming similar-sized models in these domains.
- **Phi-3 Medium (Microsoft):** With 14 billion parameters and a 128K token context window, Phi-3 Medium is designed for demanding computational tasks, offering a balance between performance and efficiency.
- **BART (Facebook AI):** A denoising autoencoder combining bidirectional and autoregressive transformers, excelling in text generation and comprehension tasks such as summarization and translation.

### B.3 Multimodal Large Language Models (MLLMs)

- **InstructBLIP (Salesforce):** An instruction-tuned vision-language model built upon BLIP-2, excelling in zero-shot performance across various multimodal tasks, including image captioning and visual question answering.
- **mBLIP:** A multilingual extension of BLIP, designed to handle vision-language tasks across multiple languages, enhancing cross-cultural understanding and accessibility.
- **LLaVA 7B:** Integrates visual and textual information, enabling tasks that require understanding and generating content from both modalities.
- **GPT-4o (OpenAI):** A multimodal model capable of processing and generating text, images, and audio, offering advanced capabilities in tasks that require integrating information across different modalities.

## C Sources for Dataset Collection

**India:** The dataset for India was compiled from publicly accessible sources including <https://www.wikipedia.org>,

<https://www.traditionalsports.org/>,  
<https://indianexpress.com/section/sports/>, <https://sports.ndtv.com/>,  
<https://www.cnbc.com/sport/>, and  
<https://www.traditionalsportsgames.org/>.

**Pakistan:** Relevant data for Pakistan was gathered from <https://www.wikipedia.org>,  
<https://www.traditionalsports.org/>,  
<https://www.cnbc.com/sport/>, <https://www.pakistantoday.com.pk/>,  
<https://tsgpakistan.com>, and <https://www.traditionalsportsgames.org/>.

**Bangladesh:** Sources for Bangladesh include <https://www.wikipedia.org>,  
<https://www.traditionalsports.org/>,  
<https://www.cnbc.com/sport/>, and  
<https://www.traditionalsportsgames.org/>.

**Thailand:** The Thai dataset was created using data from <https://www.wikipedia.org>,  
<https://www.traditionalsports.org/>,  
<https://www.cnbc.com/sport/>, and  
<https://www.traditionalsportsgames.org/>.

**Indonesia:** For Indonesia, sources include <https://www.wikipedia.org>,  
<https://www.traditionalsports.org/>,  
<https://www.cnbc.com/sport/>, and  
<https://www.traditionalsportsgames.org/>.

**China:** The dataset covering China was built using information from <https://www.wikipedia.org>,  
<https://www.traditionalsports.org/>,  
<https://www.cnbc.com/sport/>, <https://www.traditionalsportsgames.org/>, and  
<https://chcp.org/Games>.

**France:** Data for France was collected from <https://www.wikipedia.org> and <https://www.traditionalsports.org/>.

[traditionalsports.org/](https://www.traditionalsports.org/).

**Germany:** The German dataset is based on content from <https://www.wikipedia.org> and <https://www.traditionalsports.org/>.

**Italy:** Italy's dataset draws from <https://www.wikipedia.org> and <https://www.traditionalsports.org/>.

**Ethiopia:** Relevant materials for Ethiopia were obtained from <https://www.wikipedia.org> and <https://www.traditionalsports.org/>.

**Sudan:** For Sudan it was sourced from <https://www.wikipedia.org>

## Source text for Kho-Kho“Sports” From “India”

खो-खो मैदानी खेलों के सबसे प्राचीनतम रूपों में से एक है जिसका उद्भव प्रागैतिहासिक भारत में माना जा सकता है। मुख्य रूप से आत्मरक्षा, आक्रमण व प्रत्याक्रमण के कौशल को विकसित करने के लिए इसकी खोज हुई थी।

खो-खो का जन्मस्थान पुण कहा जाता है। यह खेल महाराष्ट्र, गुजरात और मध्य प्रदेश आदि प्रदेशों में अधिक खेला जाता है, किंतु भारत के अन्य प्रदेशों में भी इसका प्रचार अब बढ़ रहा है। यह खेल सरल है और इसमें कोई खतरा नहीं है। पुरुष और महिलाएँ दोनों समान रूप से इस खेल को खेल सकते हैं। खो-खो खेल में न किसी गेंद की आवश्यकता होती है, न बल्ले की। इसके लिये केवल ११ फुट लंबे और ७१ फुट चौड़े मैदान की आवश्यकता होती है। दोनों और दस-दस फुट स्थान छोड़कर चार चार फुट ऊँचे, लकड़ी के दो खंभे गाड़ दिए जाते हैं और इन खंभों के बीच की दूरी आठ बार भागों में इस प्रकार विभाजित कर दी जाती है कि दोनों दलों के खिलाड़ी एक दूसरे की विरुद्ध दिशाओं की ओर मुँह करके अपने अपने नियत स्थान पर बैठ जाते हैं। प्रत्येक दल को एक-एक पारी के लिए सात सात मिनट दिए जाते हैं और नियत समय में उस दल को अपनी पारी समाप्त करनी पड़ती है। दोनों दलों में से एक-एक खिलाड़ी खड़ा होता है, पीछा करने वाले दल का खिलाड़ी विपक्षी दल के खिलाड़ी को पकड़ने के लिए सीटी बचाते ही दौड़ता है। विपक्षी दल का खिलाड़ी पंक्ति में बैठे हए खिलाड़ियों का चक्कर लगाता है। जब पीछा करने वाला खिलाड़ी उस भागने वाले खिलाड़ी के निकट आ जाता है, तब वह अपने ही दल के खिलाड़ी के पीछे जाकर 'खो' शब्द का उच्चारण करता है तो वह उठकर भागने लगता है और पीछा करने वाला खिलाड़ी पहले को छोड़कर दूसरे का पीछा करने लगता है। पहले इस खेल का कोई व्यवस्थित नियम न था। खेल की लोकप्रियता के साथ इसके नियम बनते-बिंगड़ते रहे। १९१४ ई. में पहली बार पुना के डकन जिमखाना ने अनेक मैदानी खेलों के नियम लिपिबद्ध किए और उनमें खो-खो भी था। तब से उसके बनाए नियम के अनुसार, थोड़े स्थानीय हेर-फेर के साथ यह खेल खेला जाता है। खो-खो की पहली प्रतियोगिता पूना के जिमखाने में १९१८ ई. में हुई। फिर सन् १९१९ में बड़ौदा के जिमखाने में भारतीय स्तर पर प्रतियोगिता का आयोजन हुआ। तब से समय-समय पर इस खेल की अखिल भारतीय स्तर पर प्रतियोगिताएँ होती रहती हैं। खेल का मैदान खो-खो का क्रीड़ा क्षेत्र आयताकार होता है। यह 27 X 16 मीटर होता है। मैदान के अंत में दो मुक्त आयताकार क्षेत्र होते हैं। आयताकार की भुजा 16 मीटर और दूसरी भुजा 1.50 मी. होती है। इन दोनों आयताकारों के मध्य में दो लकड़ी के स्तम्भ होते हैं। केन्द्रीय गलौ 24 मी. लम्बी और 30 सेंटीमीटर चौड़ी होती है।

### Questions formulated using above source text and the rationale behind it

Question	Option A	Option B	Option c	Option D
खो-खो किस उद्देश्य के लिए	केवल मनोरंजन के	आत्मरक्षा, आक्रमण और	शारीरिक सौंदर्य बढ़ाने के लिए	मानसिक विकास के लिए

विकसित किया गया था?	लिए	प्रत्याक्रमण के कौशल विकसित करने के लिए		
खो-खो का जन्मस्थान कहाँ माना जाता है?	पुणे	मुंबई	दिल्ली	कोलकाता
निम्नलिखित में से कौन से प्रदेशों में खो-खो खेल का प्रमुख रूप से आयोजन होता है?	पंजाब, हरियाणा और राजस्थान	महाराष्ट्र, गुजरात और मध्य प्रदेश	उत्तराखण्ड, हिमाचल प्रदेश और जम्मू-कश्मीर	पश्चिम बंगाल, बिहार और ओडिशा
खो-खो खेलने के लिए किस वस्तु की आवश्यकता नहीं होती?	गेंद	बल्ला	दोनों गेंद और बल्ला	सीटी
खो-खो के मैदान का आकार (मीटर में) क्या है?	30 x 20 मीटर	27 x 16 मीटर	25 x 15 मीटर	20 x 10 मीटर
भारतीय स्तर पर खो-खो प्रतियोगिता का पहला आयोजन कब किया गया?	1914	1918	1919	1920

## Prompt Templates for CultSportQA

### Text-Based Prompts

#### 1. Zero-Shot Text-Based Prompt:

You are a multilingual sports expert. Answer the following multiple-choice question:

Question: [Insert text-based question here]

Options:

- A) [Option A]
- B) [Option B]
- C) [Option C]
- D) [Option D]

Select the correct option from A, B, C, or D.

---

#### 2. Few-Shot (3-Shot) Text-Based Prompt:

You are a multilingual sports expert. Answer the following multiple-choice question based on the provided examples.

Example 1:

Q: [Example question 1]

A: [Correct answer]

Example 2:

Q: [Example question 2]

A: [Correct answer]

Example 3:

Q: [Example question 3]

A: [Correct answer]

Now, answer this question:

Question: [Insert text-based question here]

Options:

A) [Option A]

B) [Option B]

C) [Option C]

D) [Option D]

Select the correct option from A, B, C, or D.

---

### 3. Chain-of-Thought (CoT) Text-Based Prompt:

You are a multilingual sports expert. Answer the following question step-by-step:

Question: [Insert text-based question here]

Options:

A) [Option A]

B) [Option B]

C) [Option C]

D) [Option D]

Think carefully step-by-step before selecting the final answer:

1. Analyze the question.

2. Recall relevant sports knowledge.
3. Eliminate incorrect options.
4. Choose the most suitable answer.

Final Answer: [Provide answer as A, B, C, or D]

---

### **Image-Based Prompts**

#### **4. Zero-Shot Image-Based Prompt:**

You are a multilingual sports expert. Analyze the following image and answer the multiple-choice question:

[Insert Image Here]

Question: [Insert image-based question here]

Options:

- A) [Option A]
- B) [Option B]
- C) [Option C]
- D) [Option D]

Select the correct option from A, B, C, or D.

---

#### **5. Few-Shot (3-Shot) Image-Based Prompt:**

You are a multilingual sports expert. Analyze the following examples and answer the multiple-choice question.

Example 1:

[Image 1]

Q: [Example question 1]

A: [Correct answer]

Example 2:

[Image 2]

Q: [Example question 2]

A: [Correct answer]

Example 3:

[Image 3]

Q: [Example question 3]

A: [Correct answer]

Now answer this:

[Insert Image Here]

Question: [Insert image-based question here]

Options:

A) [Option A]

B) [Option B]

C) [Option C]

D) [Option D]

Select the correct option from A, B, C, or D.

---

## **6. Chain-of-Thought (CoT) Image-Based Prompt:**

You are a multilingual sports expert. Analyze the following image and answer the question with reasoning.

[Insert Image Here]

Question: [Insert image-based question here]

Options:

- A) [Option A]
- B) [Option B]
- C) [Option C]
- D) [Option D]

Think carefully step-by-step before selecting the final answer:

1. Describe what you observe in the image.
2. Analyze any sports-related cues (e.g., equipment, uniforms).
3. Eliminate unlikely options.
4. Select the most suitable answer.

Final Answer: [Provide answer as A, B, C, or D]

## Traditional Games by Continent

### Games in Asia

1. Kho-Kho	24. Panjat Pinang
2. Mallakhamb	25. Karapan Ayam
3. Mushti Yuddha	26. Lari Balok
4. Lagori	27. Chon Wua
5. Kalarippayattu	28. Kaeng Ruer
6. Nadan Panthu Kali	29. Len Wow
7. Kabbadi	30. Muay Thai
8. Gilli Danda	31. Dern Kala
9. Pambaran	32. Boli Khela
10. Buffalo Race	33. Ha-Du-Du
11. Carrom	34. Lathi Khela
12. Malakhra	35. Nouka-Baich
13. Buzkashi	36. Golla-Chut
14. Gatka	37. Latim
15. Tent Pegging	38. Kanamachi
16. Chaturanga	39. Cuju
17. Oonch Neech	40. Dragon Boat Racing
18. Karapan Sapi	41. Go
19. Pencak Silat	42. Jianzi
20. Peresean	43. Tai Chi
21. Sepak Takraw	44. Liubo
22. Tarung Derajat	45. Dariabandha (Bangladesh)
23. Egrang	46. Pasha Khela

## **Games in Europe**

1. Joutes	13. Quoits
2. Tambourin	14. Eisstockschießen
3. Pelota	15. Fingerhakeln (Finger Wrestling)
4. Camargue	16. Morra
5. Petanque	17. Calcio Storico
6. Boule Lyonnaise	18. Ruzzola
7. Course Camarguaise	19. Pallanuoto
8. Balle au Poing	20. Pallone col Bracciale
9. Boules Carrées	21. Palla Tamburello
10. Bosseln	22. Corsa con Sacco (Sack Race)
11. Klootschießen	23. Schleuderball (Germany)
12. Radball	24. Water Jousting (France)

## **Games in Africa**

1. Nuba Wrestling	8. Yeferas Guks
2. Camel Racing	9. Sidama
3. Hyena	10. Afalula
4. Boruboru	11. Alemungula
5. Cow Fighting	12. Dala (Herding the Cows)
6. Genna	13. Nubian Stick Fighting
7. Gebeta	

## Annotation Guidelines for Evaluating Questions

### Objective

The purpose of this guideline is to ensure consistent and accurate evaluation of questions in terms of their correct area, their relation to sports, and their relation to a specific country. Annotators must assess whether each question aligns with these criteria based on the provided definitions and examples.

### Annotation Criteria

#### 1. Evaluating Whether the Question is Related to Traditional Sports of their country and its origin is from that country

- The question must be directly related to a sport, sporting event, or activity. It should involve aspects such as rules, history, equipment, famous players, strategies, or tournaments.

##### **Valid Example :**

*"Which country hosts the annual Sepak Takraw World Championship?"*

( The question clearly relates to a specific sport.)

##### **Invalid Example :**

*"What is the capital of Ethiopia?"*

( This question is about geography, not sports.)

#### 2. Evaluating Whether the Question is Related to a Country

- The question should explicitly or implicitly refer to a specific country. If the country is not directly mentioned, it should still be clear from the context. Questions asking about sports in general or in multiple countries should not be marked as country-specific.

##### **Valid Example :**

*"Which traditional Ethiopian game is played with wooden sticks and involves tactical movements?"*

( The question is sports-related and explicitly mentions Ethiopia.)

##### **Invalid Example :**

*"What is the national dish of Sudan?"*

( The question is country-related but not about sports.)

## Question Types & Their Annotation Guidelines

### 1. Evaluating History-Based Questions :

These questions focus on the origins, historical significance, or evolution of a sport. They must be fact-based and grounded in verifiable history.

#### **Valid Example:**

*"When was Kho-Kho first included in the National Games of India?"*

( This question relates to the historical development of the sport.)

#### **Invalid Example:**

*"Who was the greatest Kabaddi player of all time?"*

( This question is subjective and lacks historical specificity.)

### 2. Evaluating Rule-Based Questions

These questions focus on the rules, gameplay mechanics, or official regulations of a sport. They should be objective and specific to a sport's gameplay.

#### **Valid Example:**

*"How many players are allowed on a Kabaddi team during a match?"*

( The question asks about official gameplay rules.)

#### **Invalid Example:**

*"Which is the best strategy to win in Kho-Kho?"*

( The question is subjective and not strictly about rules.)

### 3. Evaluating Image-Based Questions

These questions are accompanied by an image and require the annotator to check if the visual elements are correctly aligned with the textual question. They should involve identifying a sport, equipment, movement, or player role.

#### **Valid Example:**

*"Based on the image, which traditional African sport involves the use of long sticks in combat?"*

( The question asks for an identification based on the image.)

#### **Invalid Example:**

*"How many goals did a player score in this match?"* (without a scoreboard or necessary context in the image)

( The question cannot be answered using the given image alone.)

### 4. Evaluating Scenario-Based Questions

These questions assess the ability to analyze a situation within a sport and determine the correct decision or action. The scenario should be realistic and sport-specific.

#### **Valid Example:**

*"In a Kabaddi match, if a player successfully touches an opponent and returns to their side without being tackled, what is the outcome?"*

( The question describes an in-game scenario and requires an understanding of the sport's rules.)

### **Invalid Example:**

*"What is the best way to win a Kho-Kho game?"*

( This question is too broad and lacks a structured scenario.)

## Annotation Labels Guidelines

**Each question should be assigned a label based on the following**

### **categories:**

#### **1. Area Correctness**

Correct (): The question belongs to the relevant dataset domain area.

Incorrect (): The question does not align with the intended.

#### **2. Sports Relevance**

Sports-related (): The question is directly about a sport or sporting event.

Not sports-related (): The question is about unrelated topics.

#### **3. Country Relevance**

Country-specific (): The question is explicitly tied to a country.

Not country-specific (): The question does not refer to a specific country.

#### **4. Question Type** **Each question should be categorized into one of the following types:**

History-Based (H)

Rule-Based (R)

Scenario-Based (S)

## Final Notes for Annotators

If a question is incorrectly categorized, mark it appropriately and provide a suggested correction. If a question is ambiguous, flag it for review by a senior annotator. Ensure that culturally sensitive and respectful language is maintained in all annotations.

## Validation Guidelines

The purpose of this guideline is to ensure that annotated questions meet the defined quality standards for correct area classification, sports relevance, country relevance, and question type categorization. This validation process helps maintain accuracy, consistency, and cultural sensitivity across the dataset.

### **Validation Criteria :**

#### **1. Verifying Whether the Question is Related to Traditional Sports of its Country**

Ensure that the question directly relates to a traditional sport that originates from the specified country.

2. The sport should be recognized as a culturally significant activity and should not be confused with globally popular sports. The question should cover historical aspects, rules, equipment, tournaments, or well-known figures in the sport.

**✓ Valid Example:**

*"Which country hosts the annual Sepak Takraw World Championship?"*

(✓ The question correctly relates to a traditional sport and references a country.)

**✗ Invalid Example:**

*"Which country won the last FIFA World Cup?"*

(✗ FIFA World Cup is not a traditional or indigenous sport.)

**Validation Process:**

Check authenticity: Verify the sport's origin and cultural relevance using reliable sources. Reject incorrect sports: If the question refers to modern or globally commercialized sports, mark it as incorrect.

**2. Verifying Whether the Question is Related to a Country**

The question must explicitly or implicitly mention a country related to the sport. If the country is not mentioned, the context should still make it clear. If the sport is played across multiple countries, it should not be marked as country-specific.

**✓ Valid Example:**

*"Which traditional Ethiopian game is played with wooden sticks and involves tactical movements?"*

(✓ The question is clearly related to Ethiopia and mentions a traditional game.)

**✗ Invalid Example:**

*"What is the national dish of Sudan?"*

(✗ The question is country-related but not about sports.)

**Validation Process:**

Verify country relevance using credible sources. Ensure specificity: The question should be linked to one country unless multiple origins are explicitly referenced.

**Validation by Question Type**

Each question type has specific validation steps to ensure clarity, accuracy, and relevance.

### 3. History-Based Questions Validation

These questions should focus on the historical significance, origin, or evolution of a

traditional sport. They must be fact-based, verifiable, and free from subjective opinions.

Ensure the historical reference is accurate and properly framed.

 **Valid Example:**

*"When was Kho-Kho first included in the National Games of India?"*

( Verifiable historical event.)  **Invalid Example:**

*"Who was the greatest Kabaddi player of all time?"*

( Subjective question, not grounded in history.)

**Validation Process:**

**Cross-check the historical claim** with multiple sources. **Reject vague or opinion-based questions** that lack factual grounding.

### 4. Rule-Based Questions Validation

Questions must focus on specific rules, regulations, or gameplay mechanics of a sport. They should be objective, measurable, and universally recognized within the sport.

Avoid opinion-based or strategy-based questions.

 **Valid Example:**

*"How many players are allowed on a Kabaddi team during a match?"*

( The question correctly asks about an official rule.)

 **Invalid Example:**

*"Which is the best strategy to win in Kho-Kho?"*

( The question is subjective and not strictly about rules.)

**Validation Process:**

Check the accuracy of the rule from an official governing body or sports authority. Ensure the

question refers to an actual rule rather than strategy or subjective gameplay aspects.

**Image-Based Questions Validation** The image should be clear, relevant, and accurately represent the sport being referenced in the question. The question must require the image for answering—if it can be answered without the image, it may not be valid. The image should be culturally appropriate and avoid offensive or misleading depictions.

 **Valid Example:**

*"Based on the image, which traditional African sport involves the use of long sticks in combat?"*

( The image is essential for answering the question.)

### **Invalid Example:**

*"How many goals did a player score in this match?"*

( If the image lacks a scoreboard or relevant context, the question is invalid.)

### **Validation Process:**

**Check whether the image is necessary** for answering the question.

**Verify image authenticity**:-nsure it represents the correct sport.

**Ensure visual clarity**—blurry or unclear images should be flagged for replacement.

## **6. Scenario-Based Questions Validation**

- These questions should describe a realistic in-game situation and ask for the appropriate action or rule-based outcome.
- They should be clear, structured, and have only one correct answer.
- Avoid speculative or open-ended questions.

### **Valid Example:**

*"In a Kabaddi match, if a player successfully touches an opponent and returns to their side without being tackled, what is the outcome?"*

( The scenario is realistic and directly tied to a rule.)

### **Invalid Example:**

*"What is the best way to win a Kho-Kho game?"*

( The question is too broad and lacks a structured scenario.)

### **Validation Process:**

- Ensure the scenario is plausible and follows the actual gameplay mechanics. Check whether the answer is well-defined and not open to multiple interpretations.

### **Validation Labels**

Each validated question should be categorized with the following labels:

#### **1. Area Correctness**

- **Correct (✓)**: The question belongs to the relevant dataset domain.
- **Incorrect (X)**: The question does not align with the intended area.

#### **2. Sports Relevance**

- **Sports-related (✓)**: The question is directly about a sport or sporting event.
- **Not sports-related (X)**: The question is about unrelated topics such as history, geography, or politics.

#### **3. Country Relevance**

- **Country-specific (✓)**: The question is explicitly tied to a country.
- **Not country-specific (X)**: The question does not refer to a specific country.

#### **4. Question Type**

- **History-Based (H)**

- **Rule-Based (R)**
- **Scenario-Based (S)**
- **Image-Based (I)**

#### **Final Notes for Validators**

**Provide feedback** if a question needs revision or clarification. **Flag questions** that contain **unclear language, incorrect information, or cultural insensitivity**. **Maintain consistency** in labeling and categorization across the dataset. **Consult subject matter experts** if uncertain about a question's accuracy or relevance. By following this validation guideline, we ensure that all questions meet **high-quality standards** for accuracy, clarity, and cultural sensitivity

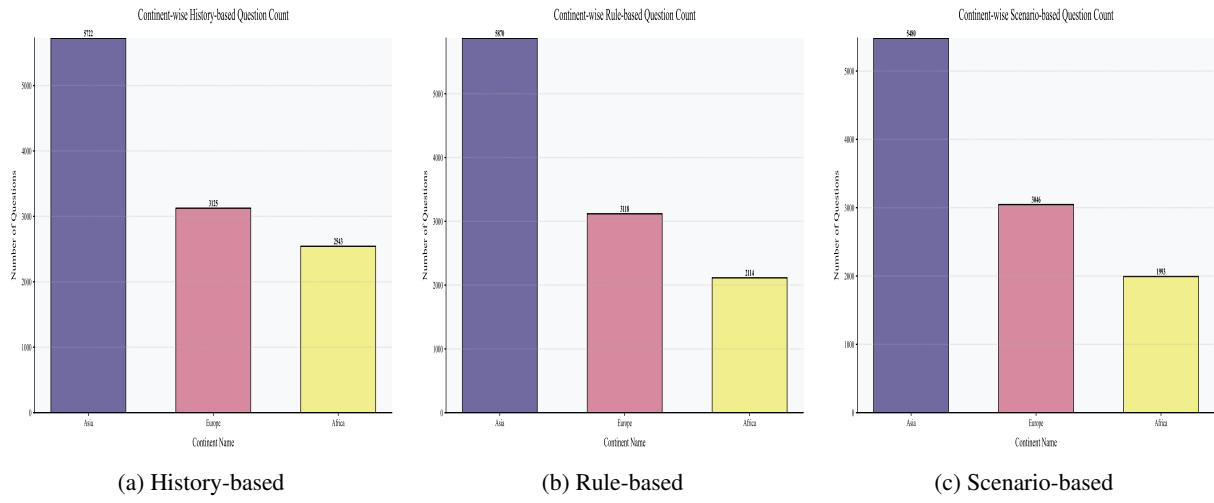


Figure 7: Statistics of history-based, rule-based, and scenario-based questions across continents

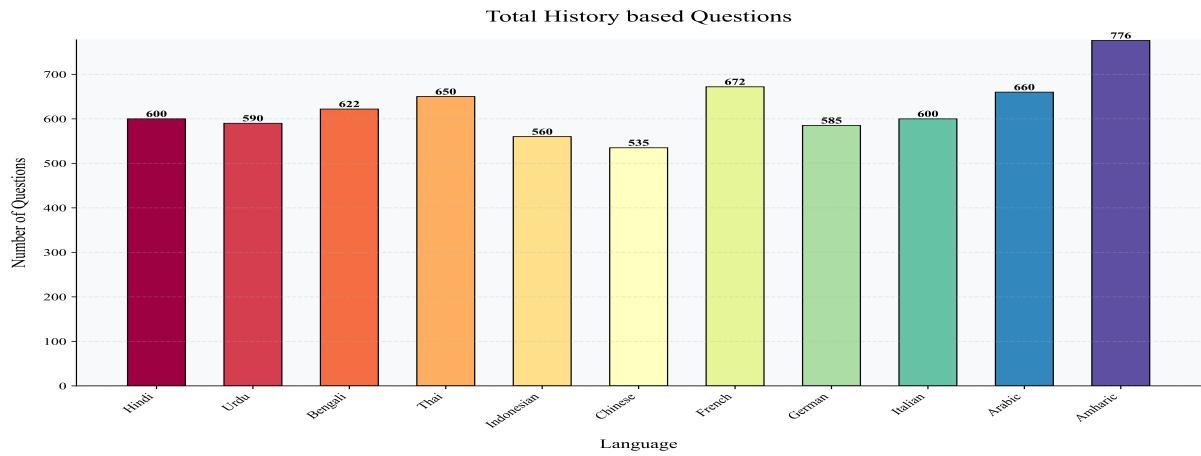


Figure 8: Statistics of history-based questions across languages

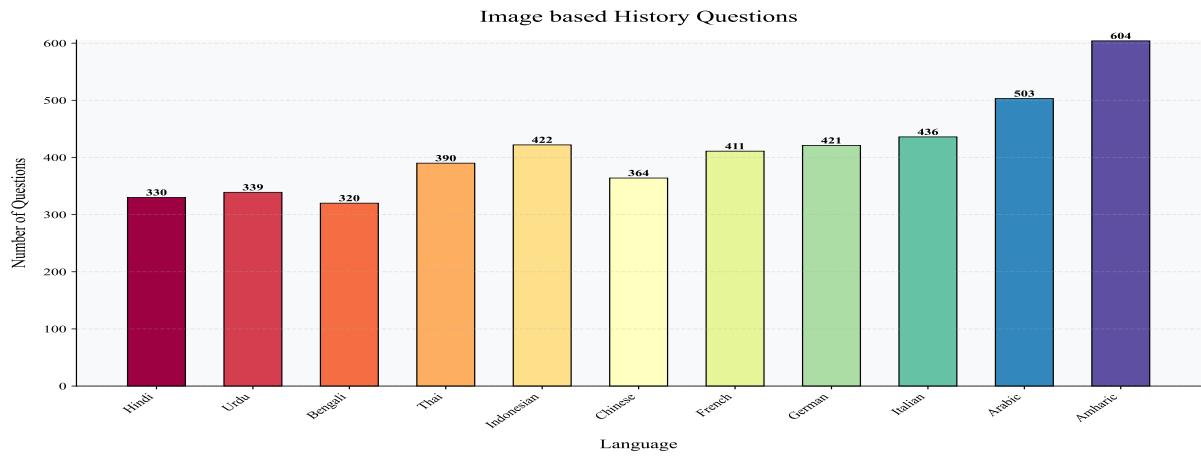


Figure 9: Statistics of image-based history questions across languages

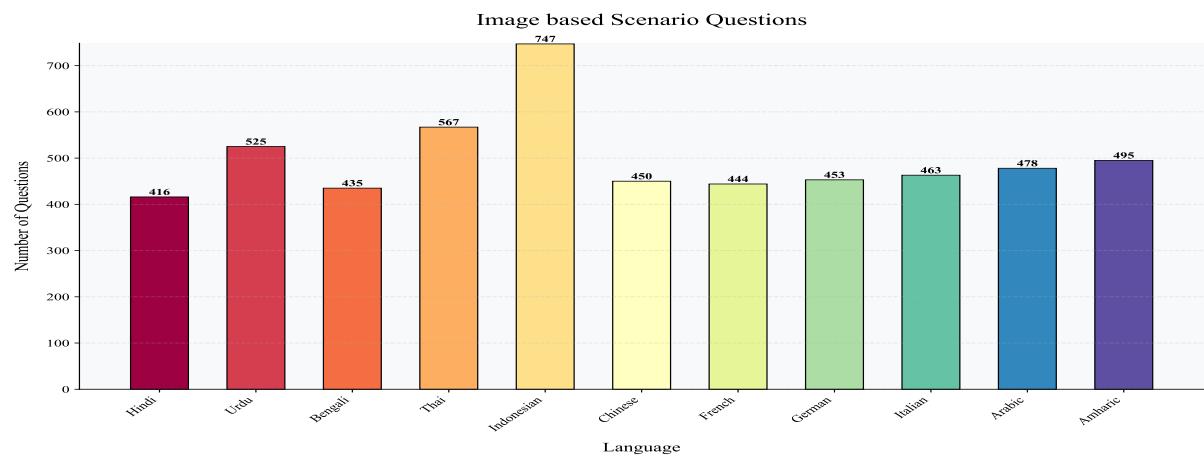


Figure 10: Statistics of image-based scenario questions across languages

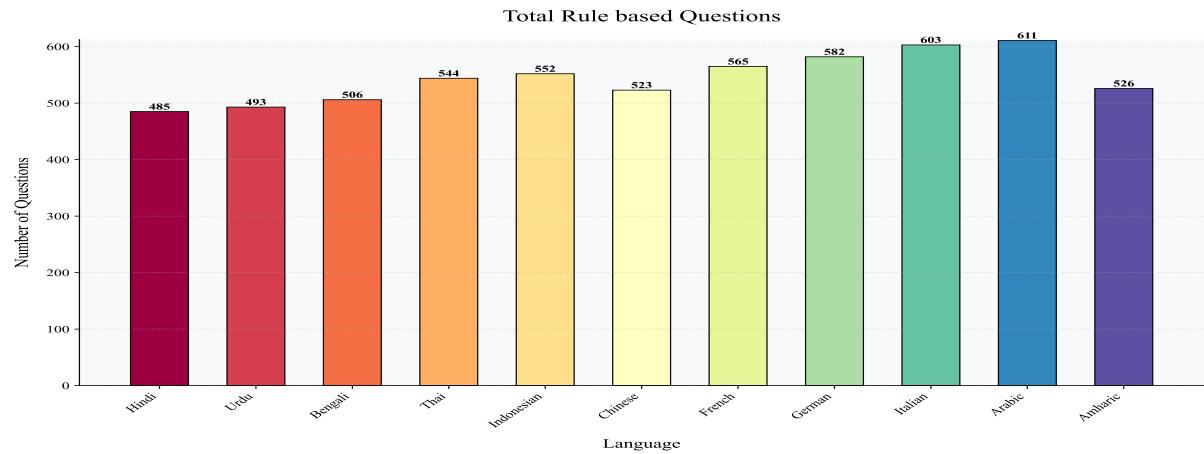


Figure 11: Statistics of rule-based questions across languages

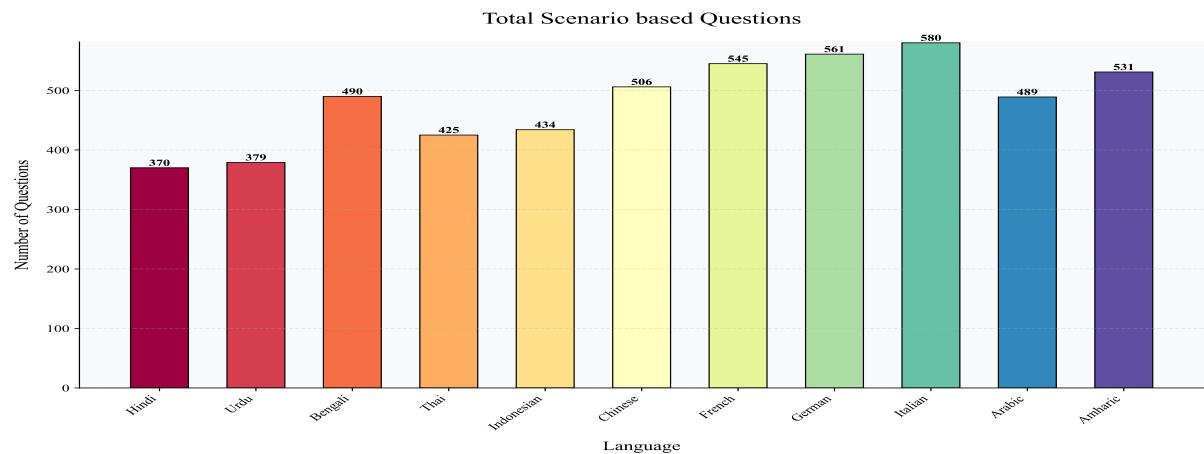


Figure 12: Statistics of scenario-based questions across languages

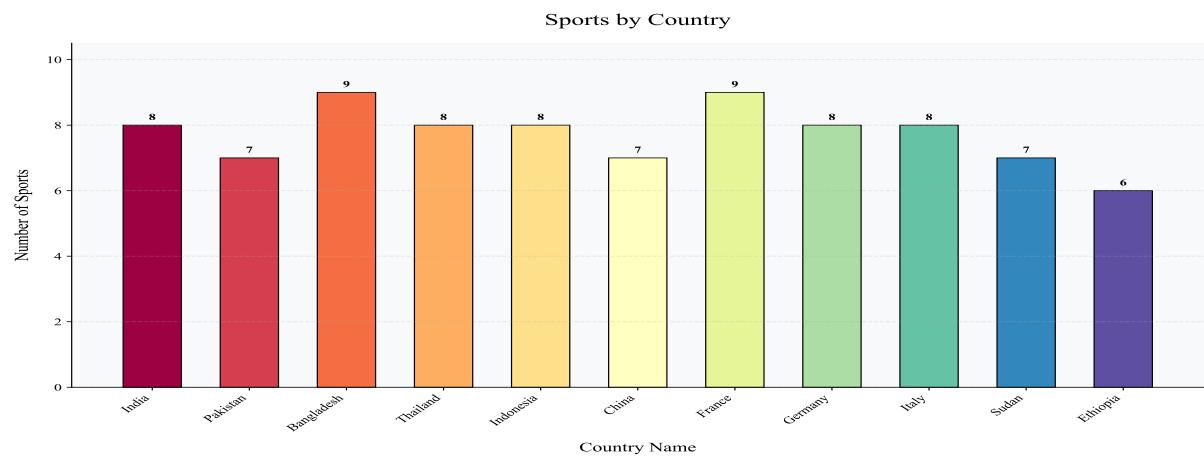


Figure 13: Statistics of sports across countries

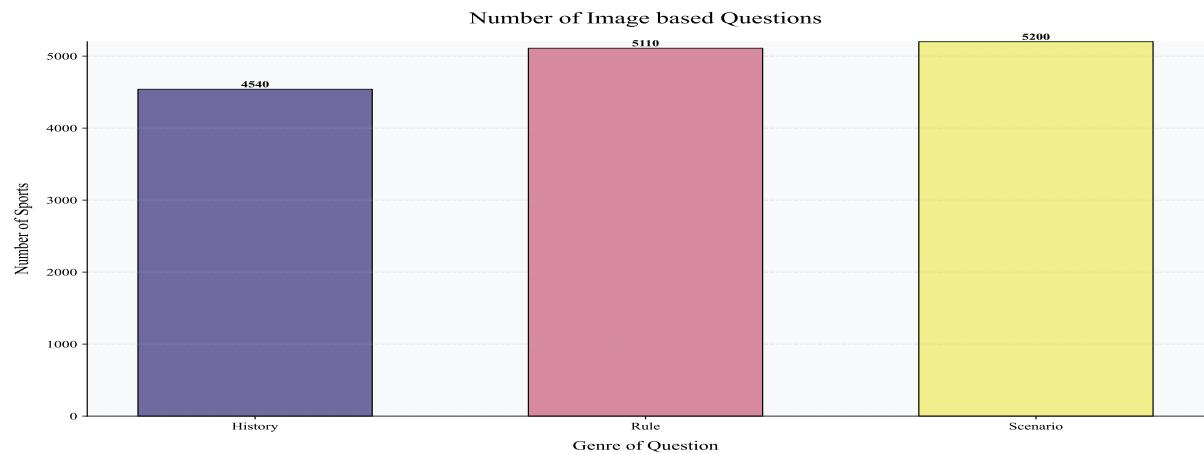


Figure 14: Statistics of image-based questions across types

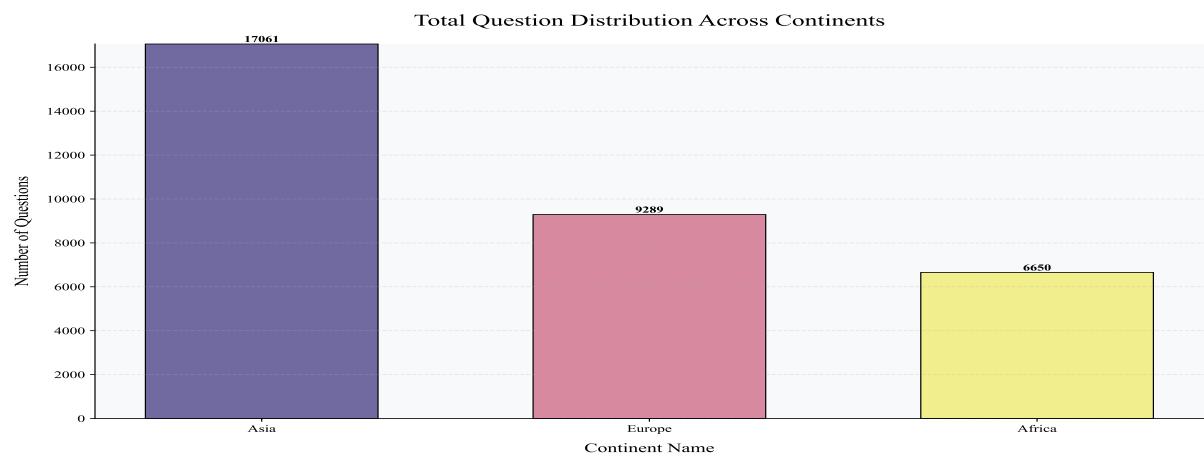
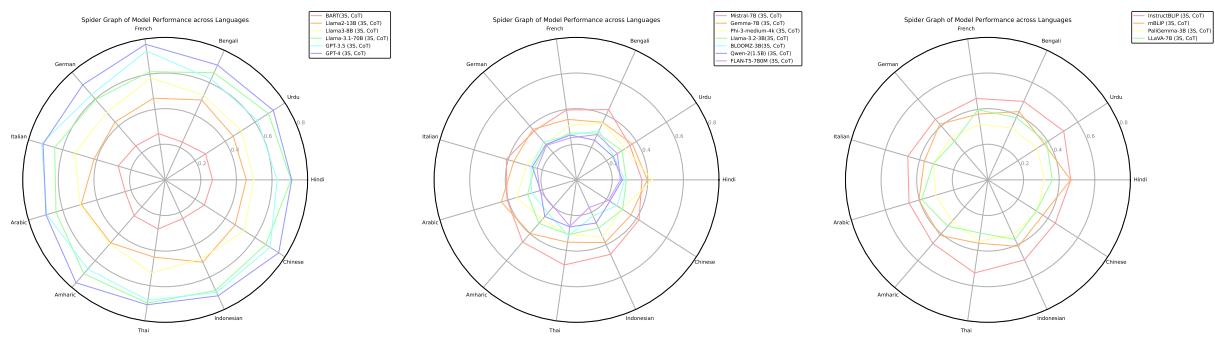


Figure 15: Statistics of questions across continents

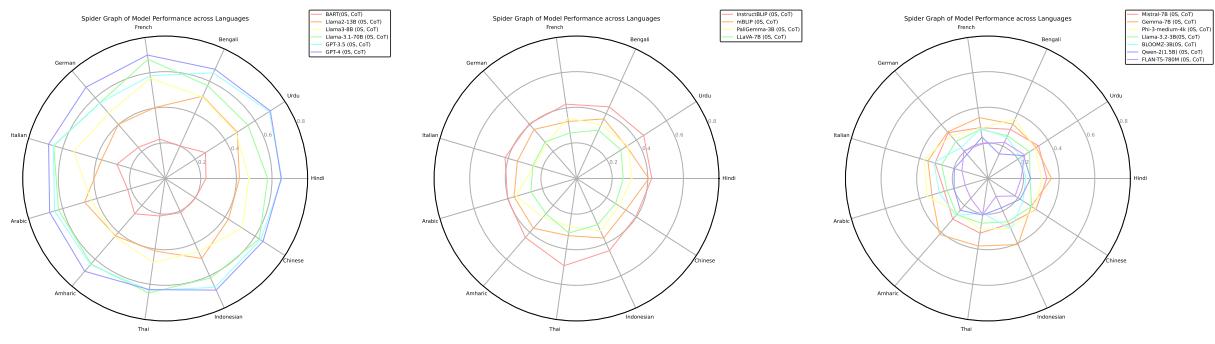


(a) CoT-based results of LLMs across languages

(b) CoT-based results of SLMs across languages

(c) CoT-based results of MLLMs across languages

Figure 16: CoT-based results of models across languages



(a) Few-shot-based results of LLMs across languages

(b) Few-shot-based results of MLLMs across languages

(c) Few-shot-based Results of SLMs across languages

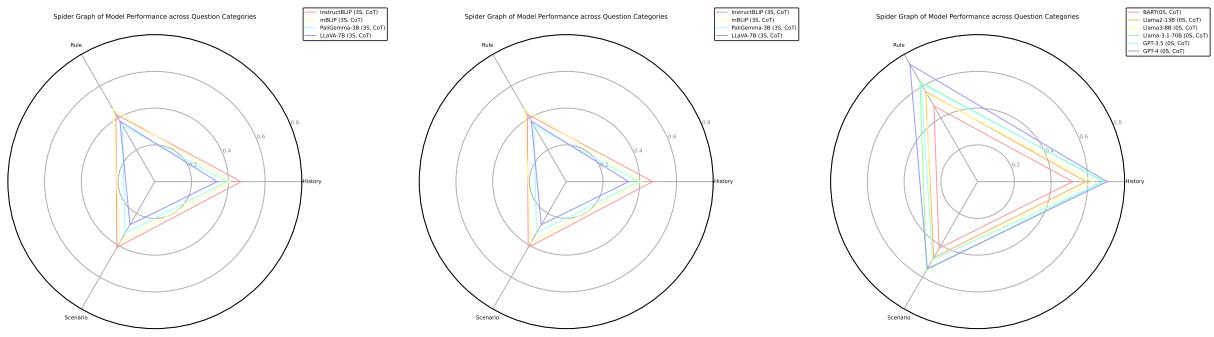
Figure 17: Few-shot-based results of models across languages



(a) CoT-based results of LLMs across types

(b) CoT-based results of SLMs across types

Figure 18: CoT-based results of models across types

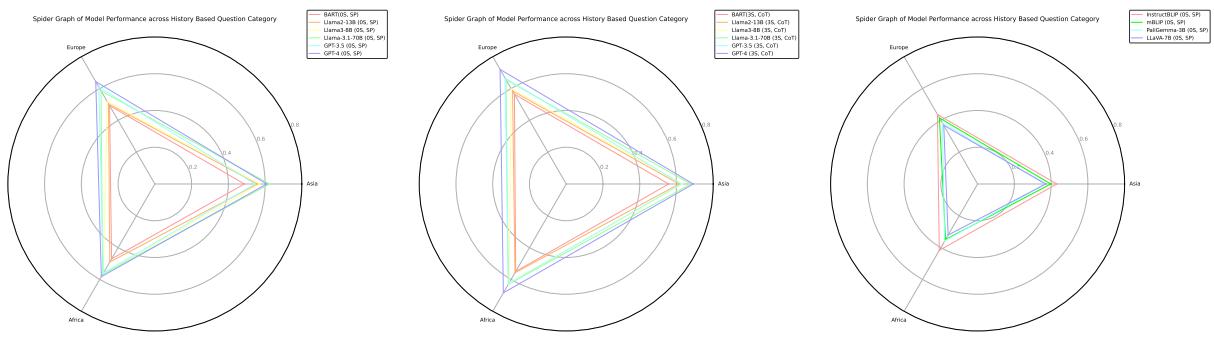


(a) CoT-based results of MLLMs across types

(b) CoT-based results of MLLMs across types

(c) Few-shot based results of LLMs across types

Figure 19: Results of models across types

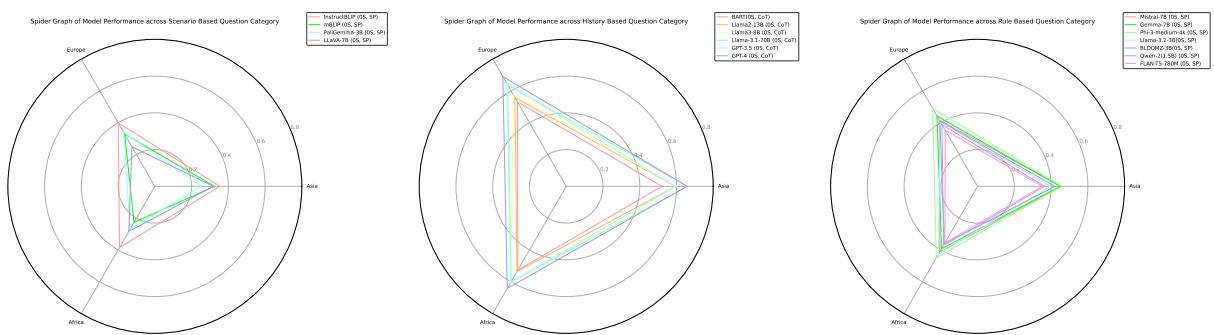


(a) Zero-shot-based results of LLMs across continents

(b) CoT-based results of LLMs across continents

(c) Zero-shot-based results of MLLMs across continents

Figure 20: Results across continents



(a) Zero-shot-based results of MLLMs across continents

(b) Few-shot-based results of LLMs across continents

(c) Zero-shot-based results of LLMs across continents

Figure 21: Results of models across continents

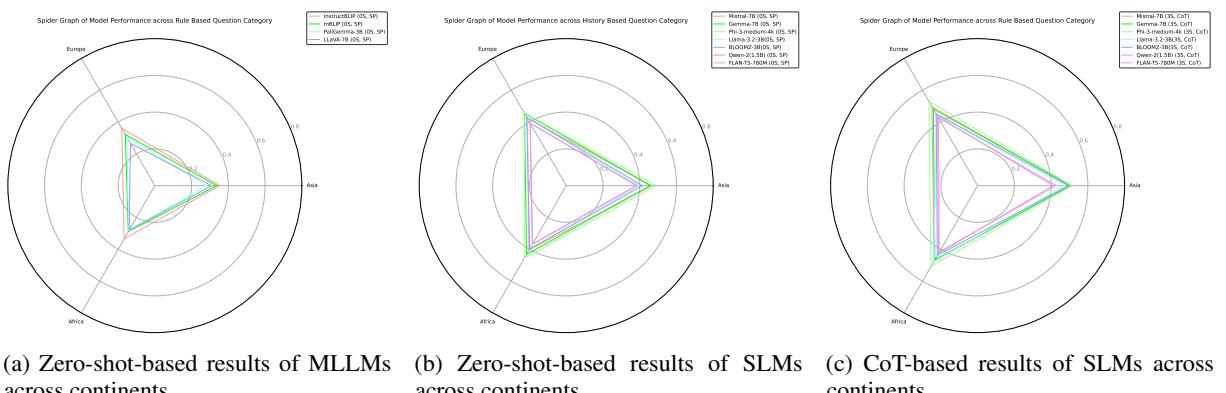


Figure 22: Results of models across continents

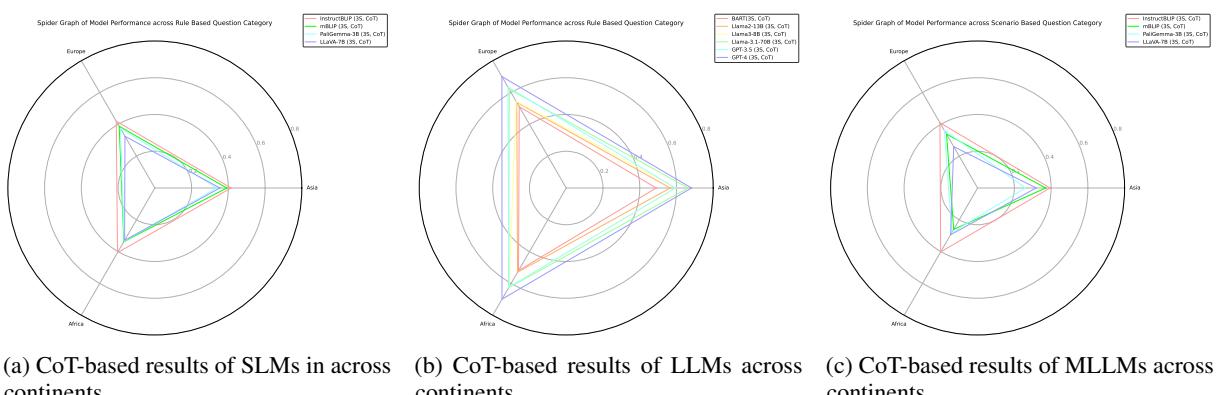


Figure 23: CoT-based results of models across continents

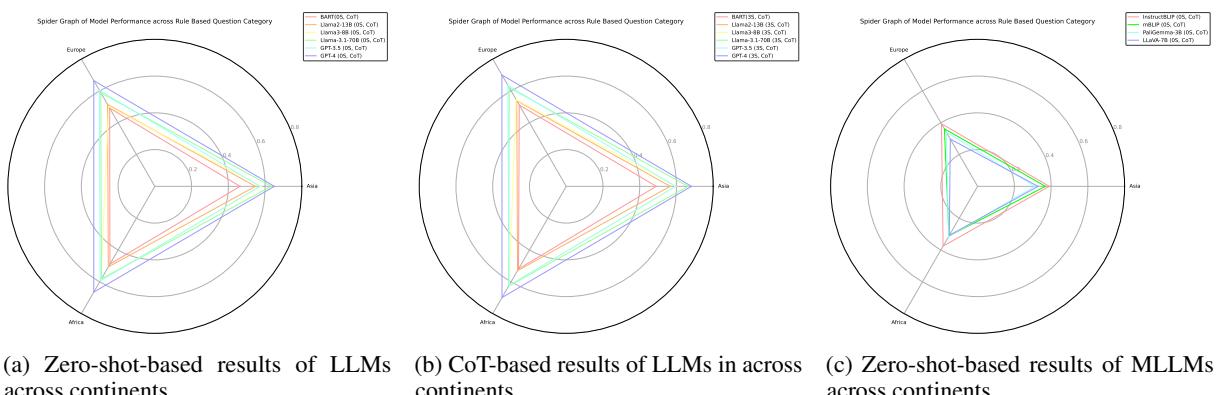


Figure 24

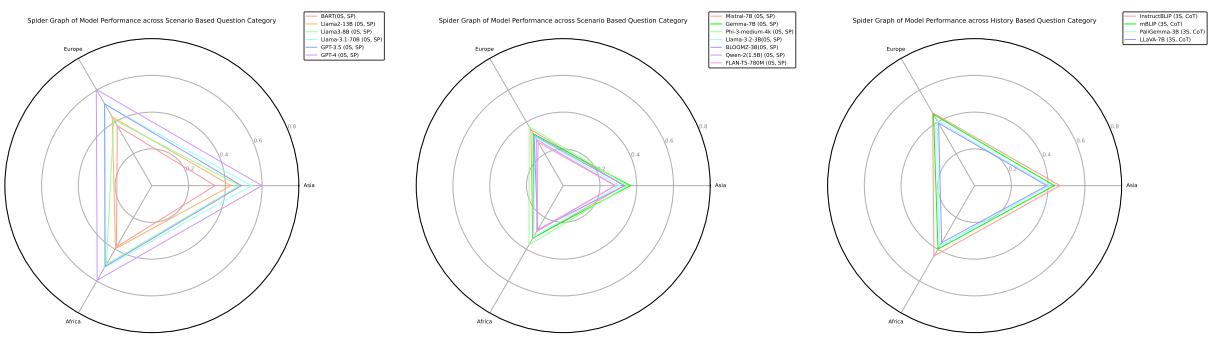


Figure 25

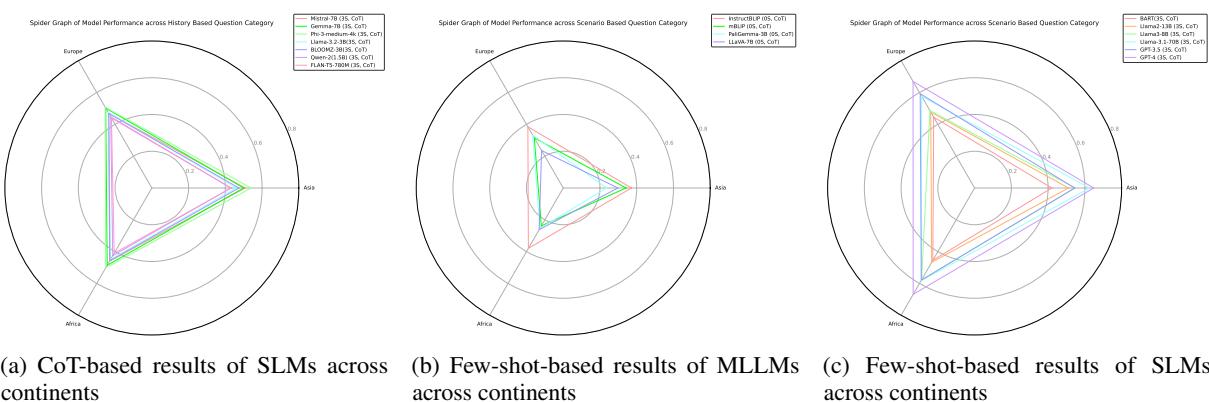


Figure 26



Figure 27

<p>Question: मुष्टि युद्ध, एक पारपरिक भारतीय कुश्ती खेल, भारत के किस भाग में उत्पन्न हुआ था?</p> <p>Options: A) पंजाब B) महाराष्ट्र C) केरल D) राजस्थान</p> <p>Correct :- B) महाराष्ट्र Predicted :- B) महाराष्ट्र</p>	<p>Question: मुष्टि युद्ध मैच के दौरान, एक पहलवान अपने प्रतिद्वंदी को एक मजबूत पकड़ से जमीन पर पटखनी देने में सक्षम होता है। इसकिं पकड़े हुए पहलवान ने गिनती तोन होने से पहले बच निकलने में सफलता पाई। रेसरी को अब क्या करना चाहिए?</p> <p>Options : A) पहले पहलवान को बिजेता घोषित करना B) मैच को पूरा शुरू करना, ताकि दोनों पहलवानों को पूरा ताकत मिल सके C) फिर से शिरांती करना, जब तक एक पहलवान पूरी तरह से पटका न जाए D) मैच को जारी रखना, क्योंकि प्रतिद्वंदी का बच निकलना बैध था</p> <p>Correct :- D) मैच को जारी रखना, क्योंकि प्रतिद्वंदी का बच निकलना बैध था Predicted :- D) मैच को जारी रखना, क्योंकि प्रतिद्वंदी का बच निकलना बैध था</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 28: Example Illustration of India Traditional Sports Correct Prediction.

<p>Question: Musti Yuddha, a traditional Indian wrestling game, was originated in which part of India?</p> <p>Options: A) Punjab B) Maharashtra C) Kerala D) Rajasthan</p> <p>Correct :- B) Maharashtra Predicted :- B) Maharashtra</p>	<p>Question: During a Musti Yuddha match, one wrestler successfully slams his opponent to the ground with a strong hold. However, the slammed wrestler manages to escape before the count of three. What should the referee do now?</p> <p>Options: A) Declare the first wrestler as the winner B) Restart the match to allow both wrestlers to regain their strength C) Start the count again until one wrestler is fully pinned D) Continue the match, as the opponent's escape was valid</p> <p>Correct: D) Continue the match, as the opponent's escape was valid Predicted: D) Continue the match, as the opponent's escape was valid</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 29: Example Illustration of India Traditional Sports Correct Prediction (In English)

<p>Question: मुष्टि युद्ध प्राचीन भारतीय युद्धकला में उत्पन्न हुआ था और यह किस प्रकार के सामाजिक या सारकृतिक संर्थक में अधिकतर अभ्यास किया जाता था?</p> <p>Options: A) राजसी दरबारों में सैन्य प्रशिक्षण के दौरान B) धार्मिक मंदिरों में आध्यात्मिक अभ्यास के रूप में C) गांवों गांवों में सामुदायिक सुखा खेल के रूप में D) शैक्षिक संस्थानों में शारीरिक शिक्षा के रूप में</p> <p>Correct :- A) राजसी दरबारों में सैन्य प्रशिक्षण के दौरान Predicted :- C) गांवों गांवों में सामुदायिक सुखा खेल के रूप में</p> <p>History Based</p>	<p>Question: मुष्टि युद्ध मैच के दौरान, एक पहलवान अपने लपेटे हुए मुट्ठी से प्रतिदंडी को प्रहर करता है, लेकिन गलती से वह प्रतिदंडी को कमर से नीचे हिट कर देता है। रेफरी इसे देखता है और मैच को रोकता है। रेफरी को अब क्या कदम उठाना चाहिए?</p> <p>Options : A) पहलवान को घेतावनी देना और मैच को जारी रखने देना B) पहलवान से अंक घटाना और मैच को जारी रखना C) पहलवान को टूरंट डिसक्वालिफाई करना क्योंकि उसने कमर से नीचे हिट किया D) मैच को बिना किसी टक्के के रिंग से केंद्र से फिर से शुरू करना</p> <p>Correct :- A) पहलवान को घेतावनी देना और मैच को जारी रखने देना Predicted :- C) पहलवान को टूरंट डिसक्वालिफाई करना क्योंकि उसने कमर से नीचे हिट किया</p> <p>Scenario Based</p>
<p>Question: मुष्टि युद्ध के पारंपरिक नियमों के अनुसार, निम्नलिखित में से कौन सा कृत्य मैच के दौरान अवैध माना जाता है?</p> <p>Options: A) प्रतिदंडी के घेहरे पर लपेटे हुए मुट्ठी से प्रहर करना B) प्रतिदंडी के पैरों को धोने के लिए पात का उपयोग करना C) लाभ प्राप्त करने के लिए प्रतिदंडी के कपड़े पकड़ना D) प्रतिदंडी को जमीन पर फेंकने का प्रयास करना</p> <p>Correct :- C) लाभ प्राप्त करने के लिए प्रतिदंडी के कपड़े पकड़ना Predicted :- D) प्रतिदंडी को जमीन पर फेंकने का प्रयास करना</p> <p>Rule Based</p>	<p>Question: नीचे दी गई छवि में, एक पहलवान अपने प्रतिदंडी को उठाने और फेंकने की कोशिश कर रहा है। मुष्टि युद्ध के नियमों के आधार पर, निम्नलिखित में से कौन सा क्रियावली इस स्थिति में अवैध मानी जाएगी?</p> <p>Options: A) प्रतिदंडी को कंधे पर पकड़कर उठाना B) पैरों का सहारा लेकर प्रतिदंडी को फेंकना C) प्रतिदंडी के कपड़ों का उपयोग उठाने समय लाभ के लिए करना D) प्रतिदंडी को नियंत्रित बल के साथ जमीन पर फेंकना</p> <p>Correct :- C) प्रतिदंडी के कपड़ों का उपयोग उठाने समय लाभ के लिए करना Predicted :- B) पैरों का सहारा लेकर प्रतिदंडी को फेंकना</p> <p>Image Based</p>

Figure 30: Example Illustration of India Traditional Sports Wrong Prediction.

<p>Question: Musti Yuddha originated in ancient Indian martial arts. In what kind of social or cultural context was it mostly practiced?</p> <p>Options: A) In royal courts during military training B) In religious temples as a form of spiritual practice C) In rural villages as a community defense sport D) In educational institutions as physical education</p> <p>Correct: A) In royal courts during military training Predicted: C) In rural villages as a community defense sport</p> <p>History Based</p>	<p>Question: During a Musti Yuddha match, a wrestler strikes the opponent with his wrapped fist, but accidentally hits below the waist. The referee sees this and stops the match. What should the referee do now?</p> <p>Options: A) Warn the wrestler and allow the match to continue B) Deduct points from the wrestler and continue the match C) Immediately disqualify the wrestler for hitting below the waist D) Restart the match from the center of the ring without any penalty</p> <p>Correct: A) Warn the wrestler and allow the match to continue Predicted: C) Immediately disqualify the wrestler for hitting below the waist</p> <p>Scenario Based</p>
<p>Question: According to the traditional rules of Mushti Yuddha, which of the following actions is considered illegal during the match?</p> <p>Options: A) Striking the opponent's face with a wrapped fist B) Using the leg to trap the opponent's feet C) Grabbing the opponent's clothes to gain an advantage D) Attempting to throw the opponent to the ground</p> <p>Correct: C) Grabbing the opponent's clothes to gain an advantage Predicted: D) Attempting to throw the opponent to the ground</p> <p>Rule Based</p>	<p>Question: In the image below, a wrestler is trying to lift and throw his opponent. Based on the rules of Musti Yuddha, which of the following actions would be considered illegal in this situation?</p> <p>Options: A) Lifting the opponent by holding them on the shoulder B) Using the legs to assist in throwing the opponent C) Using the opponent's clothes to gain an advantage while lifting D) Throwing the opponent to the ground with controlled force</p> <p>Correct: C) Using the opponent's clothes to gain an advantage while lifting Predicted: B) Using the legs to assist in throwing the opponent</p> <p>Image Based</p>

Figure 31: Example Illustration of India Traditional Sports Wrong Prediction (In English)

<p>Question: লাঠি খেল, একটি প্রতিহ্বাবাহী বাংলা মার্শাল আর্ট, কোন প্রতিহ্বাসিক সময়কালে এর উৎপত্তি হয়?</p> <p>Options: A) মুঘল যুগ B) বিট্টিন উপনিষদিক যুগ C) প্রাচীন বাংলা রাজা D) ইসলাম পূর্ব বাংলা</p> <p>Correct :- A) মুঘল যুগ Predicted :- A) মুঘল যুগ</p>	History Based	Scenario Based
<p>Question: লাঠি খেল-তে, প্রতিযোগিতার সময় অংশগ্রহকারীরা প্রধানত কোন উপকরণ ব্যবহার করেন?</p> <p>Options: A) তলামার B) লাঠি C) চাল D) চাকু</p> <p>Correct :- B) লাঠি Predicted :- B) লাঠি</p>	Rule Based	<p>Question: দ্বিতীয় নিচে, আপনি একজন খেলোয়াড়কে ম্যাচের মাঝখানে দেখতে পাচ্ছেন। খেলোয়াড়টি দ্বিতীয় কী করার প্রস্তুতি নিছে?</p> <p>Options: A) লাঠি দিয়ে প্রতিহ্বাবীকে সুরোপুরি আঘাত করা B) লাঠি দিয়ে প্রতিরক্ষা এবং রাঁক করা C) লাঠি প্রতিহ্বাবীর দিকে ঝুঁড়ে ফেলা D) প্রতিহ্বাবীর উপর একটি জয়েন্ট লক গ্রয়েগ করা</p> <p>Correct :- A) লাঠি দিয়ে প্রতিহ্বাবীকে সুরোপুরি আঘাত করা Predicted :- A) লাঠি দিয়ে প্রতিহ্বাবীকে সুরোপুরি আঘাত করা</p>

Figure 32: Example Illustration of Bangladesh Traditional Sports Correct Prediction

<p>Question: Lathi Khela, a traditional Bengali martial art, originated during which historical period?</p> <p>Options: A) Mughal era B) British colonial period C) Ancient Bengali kingdoms D) Pre-Islamic Bengal</p> <p>Correct: A) Mughal era Predicted: A) Mughal era</p>	<p>Question: In a Lathi Khela match, a player strikes the opponent's stick, and the opponent loses balance but does not fall. The match continues without interruption. What should the referee do in this situation?</p> <p>Options: A) Stop the match and issue a warning B) Continue the match if the opponent is able to defend themselves C) Award a point to the player who struck the stick D) Declare the match a draw</p> <p>Correct: B) Continue the match if the opponent is able to defend themselves Predicted: B) Continue the match if the opponent is able to defend themselves</p>
<p>Question: In Lathi Khela, what primary equipment do participants use during the competition?</p> <p>Options: A) Sword B) Stick C) Shield D) Knife</p> <p>Correct: B) Stick Predicted: B) Stick</p>	<p>Question: In the image below, you can see a player in the middle of a match. What is the player preparing to do in the image?</p> <p>Options: A) Strike the opponent forcefully with the stick B) Defend and block with the stick C) Throw the stick towards the opponent D) Apply a joint lock on the opponent</p> <p>Correct: A) Strike the opponent forcefully with the stick Predicted: A) Strike the opponent forcefully with the stick</p>

Figure 33: Example Illustration of Bangladesh Traditional Sports Correct Prediction (In English)

<p>Question: बांग्लादेश के नव अफलाटि प्रतिष्ठानी मार्शल आर्ट लाठि खेल र चढ़ा एवं संवर्क्षण विधेयकावे परिचित ?</p> <p>Options: A) निलाटे B) छड़ियाम C) राजशक्ति D) थलना</p> <p>Correct :- A) राजसी दरबारी में सैन्य प्रशिक्षण के दौरान Predicted :- C) गामोण गांवों में सामुदायिक सुरक्षा खेल के रूप में</p>	<p>Question: एकटि लाठि खेला याच, एकजन खेलायाडु तार प्रतिष्ठानीके लाठि दिये आयात करार पर, तार प्रतिष्ठानी प्रतिरक्षा कोल्पन दियेवे लाठि आक्रमणके बाबेल खेलायाडुया याच। विच समय गर, प्रतिष्ठानी आवार आक्रमण कराते दिये आसेब, किन्तु तिनि लिपिट्ट समयावे याच आक्रमण कराते वार्ष रन। एই परिस्थितिते रेखावि शी कराबेल?</p> <p>Options : A) याच अविलाय शिखित करने सक्किता देबेल B) गालिये याओया खेलायाडुके पाट्टे केटे देबेल एवं याच चालिये याबेल C) गालिये याओया खेलायाडुके डिस्कोयालिफाइ करने देबेल D) याचेर फलाफल घेराग कराबेल</p> <p>Correct :- B) गालिये याओया खेलायाडुके गलेटि केटे देबेल एवं याच चालिये याबेल Predicted :- C) गालिये याओया खेलायाडुके डिस्कोयालिफाइ करने देबेल</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 34: Example Illustration of Bangladesh Traditional Sports Wrong Prediction.

<p>Question: Balapan Karapan Sapi secara historis diadakan untuk merayakan acara budaya atau pertanian apa di Jawa Timur?</p> <p>Options: A) Festival panen B) Tahun Baru Islam C) Kedatangan musim hujan D) Penobatan kerajaan</p> <p>Correct :- A) Festival panen Predicted :- A) Festival panen</p>	<p>Question: Selama balapan Karapan Sapi, salah satu sapi tiba-tiba berhenti di tengah lintasan, menyebabkan pengemudi kehilangan kendali sejenak. Apa yang harus terjadi dalam situasi ini?</p> <p>Options : A) Balapan secara otomatis dihentikan, dan sapi didiskualifikasi B) Balapan dilanjutkan selama sapi masih berlari C) Pengemudi diberikan peringatan, dan balapan dilanjutkan D) Sapi diberikan kesempatan kedua untuk memulai lagi</p> <p>Correct :- C) Pengemudi diberikan peringatan, dan balapan dilanjutkan Predicted :- C) Pengemudi diberikan peringatan, dan balapan dilanjutkan</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 35: Example Illustration of Indonesia Traditional Sports Correct Prediction.

<p>Question: Secara historis, Karapan Sapi bukan hanya sebuah olahraga tetapi juga merupakan tontonan dari hal berikut ini ?</p> <p>Options: A) Pengabdian agama B) Kekayaan dan status peserta C) Keterampilan maritim D) Alat dan teknik pertanian</p> <p>Correct :- B) Kekayaan dan status peserta Predicted :- A) Pengabdian agama</p>	<p>Question: Di tengah balapan Karapan Sapi, salah satu sapi tiba-tiba mulai melambat karena kelelahan. Pengemudi berusaha untuk menyemangati sapi, tetapi sapi terus terlengkap, menyebabkan sapi lainnya dan pengemudi mengambil alih posisi terdepan. Apa yang terjadi dalam situasi ini?</p> <p>Options :A) Balapan dilanjutkan, dan sapi yang lebih lambat diberikan penalti waktu B) Balapan dihentikan, dan kedua sapi diberi kesempatan untuk beristirahat C) Pengemudi sapi yang lebih lambat didiskualifikasi karena gagal menjaga kecepatan D) Balapan dilanjutkan tanpa penalti, dan sapi dengan kecepatan lebih lambat mungkin tidak menang</p> <p>Correct :- D) Balapan dilanjutkan tanpa penalti, dan sapi dengan kecepatan lebih lambat mungkin tidak menang Predicted :- A) Balapan dilanjutkan, dan sapi yang lebih lambat diberikan penalti waktu</p>
<p>History Based</p> <p>Question: Dalam Karapan Sapi, manakah dari berikut ini yang dianggap sebagai pelanggaran selama balapan?</p> <p>Options: A) Sapi tidak berlari dengan kecepatan yang sama B) Pengemudi sapi tidak memegang tali kekang dengan benar C) Sapi melintasi garis finis dalam posisi apapun D) Sapi terlalu dekat satu sama lain selama balapan</p> <p>Correct :- B) Pengemudi sapi tidak memegang tali kekang dengan benar Predicted :- D) Sapi terlalu dekat satu sama lain selama balapan</p>	<p>Question: Pada gambar di bawah ini, sapi-sapi sedang berlomba menuju garis finish dalam acara Karapan Sapi. Salah satu sapi tampaknya kehilangan kecepatan. Apa yang menjadi kemungkinan penyebabnya dalam balapan Karapan Sapi ini?</p> <p>Options: A) Pengemudi gagal mengendalikan tali kekang dengan benar B) Sapi tersebut terluka atau kelelahan C) Lintasan tidak rata, menyebabkan sapi kehilangan keseimbangan D) Sapi tersebut sengaja memperlambat laju untuk mengganggu balapan</p> <p>Correct :- B) Sapi tersebut terluka atau kelelahan Predicted :- A) Pengemudi gagal mengendalikan tali kekang dengan benar</p>

Figure 36: Example Illustration of Indonesia Traditional Sports Wrong Prediction.

<p>Question: กีฬาแข่งเรือแบบตั้งเดิมของไทย มีต้นกำเนิดในช่วงยุคประวัติศาสตร์ใด?</p> <p>Options: A) ราชอาณาจักรสุโขทัย B) ราชอาณาจักรอยุธยา C) ยุคจักรวรรดิ D) ประเทศไทยในปัจจุบัน</p> <p>Correct:- B) ราชอาณาจักรอยุธยา Predicted:-B) ราชอาณาจักรอยุธยา</p>	<p>Question: ในการแข่งขันแข่งเรือ ทีมเรือหนึ่งกำลังตามหลังทีมอีกทีมในจังหวะการพายที่สูญเสียจังหวะการพาย ทีมนี้ทีมเรือที่มีอย่างมีโอกาสที่จะตามทัน แต่ทีมนี้อ่อนใจลังเลสืบเชิงแล้ว ลังที่ทีมนี้ควรจะทำเพื่อแก้สถานการณ์ดีอะไร?</p> <p>Options: A) นักพากษาทุกคนเพิ่มความเร็วในการพายและประسانจังหวะการพายเพื่อให้ได้ตาม B) หยุดพานและรอให้ทีมอีกทีมตามมา C) พยายามใช้การพายที่เร็วขึ้นโดยไม่ประسانจังหวะการพาย D) ยังคงเบี่ยงออกจากเส้นทางจนกว่าจะถึงเส้นชัยเพื่อหลีกเลี่ยงการลงโทษ</p> <p>Correct:- A) นักพากษาทุกคนเพิ่มความเร็วในการพายและประسانจังหวะการพายเพื่อให้ได้ตาม Predicted:-A) นักพากษาทุกคนเพิ่มความเร็วในการพายและประسانจังหวะการพายเพื่อให้ได้ตาม</p>
<p>History Based</p> <p>Question: ในการแข่งเรือ วัดถุประสังค์หลักของผู้เข้าร่วมการแข่งคืออะไร?</p> <p>Options: A) การแข่งขันเป็นทีมที่มีเรือสองลำ B) การท้าพื้นที่ราชศรีและหัวการแข่งขัน C) การเข้าสืบเชิญมาแล้วที่เริ่วที่สุด D) การแสดงรูปแบบการออกแนวเรือมากกว่าความเร็ว</p> <p>Correct:- C) การเข้าสืบเชิญมาแล้วที่เริ่วที่สุด Predicted:-C) การเข้าสืบเชิญมาแล้วที่เริ่วที่สุด</p>	<p>Question: ในภาคตัดส่วนนี้ คุณสามารถเห็นการแข่งขันแข่งเรือ ทีมกำลังตามหลังทีมอีกทีม กัน และอยู่ในตัวแทนที่ให้มา ๆ กัน ในระหว่างนี้ นักพากษาทุกคนให้ความสำคัญกับอะไรในระหว่างการแข่งขัน?</p> <p>Options: A) วักษากวนเรือและจังหวะการพายเพื่อให้อยู่ข้างหน้า B) เตรียมการท้าพื้นที่ราชศรีในระหว่างการแข่งขัน C) ลดความเร็วเพื่อหลีกเลี่ยงการชนหรือเสียหาย D) ปรับโครงสร้างเรือในนาทีสุดท้าย</p> <p>Correct:- A) วักษากวนเรือและจังหวะการพายเพื่อให้อยู่ข้างหน้า Predicted:-A) วักษากวนเรือและจังหวะการพายเพื่อให้อยู่ข้างหน้า</p>

Figure 37: Example Illustration of Thailand Traditional Sports Correct Prediction.

<p>Question: กษัตริย์ไทยพระองค์ใดที่ได้รับการยกย่องว่าเป็นผู้ส่งเสริมการแข่งเรือให้เป็นประเพณีหลวงประจำปี?</p> <p>Options: A) พ่อขุนรามคำแหง (สุโขทัย) B) สมเด็จพระนเรศวรมหาราช (อยุธยา) C) สมเด็จพระเจ้าตากสินมหาราช (ธนบุรี) D) พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช (รัตนโกสินทร์)</p> <p>Correct:- B) สมเด็จพระนเรศวรมหาราช (อยุธยา) Predicted:- D) พระบาทสมเด็จพระพุทธยอดฟ้าจุฬาโลกมหาราช (รัตนโกสินทร์)</p>	<p>Question: ในงานแข่งขันแข่งเรือ เรือลากลั่นแข่งกันอย่างสุ่ม ในช่วงมีสำคัญสืบสานเชื้อเมืองล้านนาที่เรียกว่าอะไร?</p> <p>Options: A) หันหน้าเรือกลับไปทางเดินทางที่ถูกต้องทันทีและป่วนเทคนิคการพยายามเพื่อเพิ่มความเร็ว B) หมุนเรือและรอให้เรือเดินทางมา C) พยายามใช้การพยายามที่รีบวันโดยไม่ประเมินสถานจังหวัดการทราบ D) เมืองอุบลราชธานีทางเดินทางน้ำกว่าจะสิ้นเชิงเพื่อหลีกเลี่ยงการลังบัง</p> <p>Correct:- A) หันหน้าเรือกลับไปทางเดินทางที่ถูกต้องทันทีและป่วนเทคนิคการพยายามเพื่อเพิ่มความเร็ว Predicted:- D) เมืองอุบลราชธานีทางเดินทางน้ำกว่าจะสิ้นเชิงเพื่อหลีกเลี่ยงการลังบัง</p>
<p>History Based</p> <p>Question: ในการแข่งพายเรือ หากทีมเรือชิงชนะเลิศ แต่พบว่าไม่เป็นมีน้ำพายคนหนึ่งในน้ำที่มีติดตามอยู่ในน้ำพายบนเรือถูกต้อง สิ่งที่มีภาระเกิดขึ้นต่อต่อจะเป็น?</p> <p>Options: A) ทีมจะได้รับการลงโทษ และเวลาของทีมจะถูกปรับ B) ทีมจะถูกตัดสิทธิ์จากการแข่งขัน C) การแข่งขันจะถูกยกเลิก และจะมีการจัดการแข่งขันใหม่ D) นักพายของทีมหนึ่งที่ต้องผ่านการแข่งขันจะถูกตัดออกไม่ต่อไป</p> <p>Correct:- B) ทีมจะได้รับการลงโทษ และเวลาของทีมจะถูกปรับ Predicted:- A) ทีมจะได้รับการลงโทษ และเวลาของทีมจะถูกปรับ</p>	<p>Scenario Based</p> <p>Question: จากภาพผู้เล่นล่างน้ำที่จะเห็นได้ว่า เรือลากลั่นที่เพื่อยืนในการแข่งขันมีการอึดเชิงลากลั่นอย่างมาก แนะนำพากลั่นลากลั่นที่จะดีที่สุดในการแข่งขัน?</p> <p>Options: A) ปั๊มน้ำหินเทคนิคการพยายามเพื่อเก็บไชสมุดและหลีกเลี่ยงการลังบังอัตต์ B) หมุนเรือบนการแข่งขันด้วยการอึดเชิงลากลั่นเพื่อเพิ่มความเร็วและลดความล่าช้า C) ตัดลั่นที่รีและตัวในในการแข่งขันที่มีปะการัง D) ใช้การอึดเชิงลากลั่นเพื่อเพิ่มความเร็วและทำให้คู่แข่งลับสน</p> <p>Correct:- A) ปั๊มน้ำหินเทคนิคการพยายามเพื่อเก็บไชสมุดและหลีกเลี่ยงการลังบังอัตต์ Predicted:- D) ใช้การอึดเชิงลากลั่นเพื่อเพิ่มความเร็วและทำให้คู่แข่งลับสน</p>
<p>Rule Based</p>	<p>Image Based</p> 

Figure 38: Example Illustration of Thailand Traditional Sports Wrong Prediction.

<p>Question: 跳鞠最早的记载出现在中国哪个朝代，显示了其长久的历史意义？</p> <p>Options: A) 唐朝 B) 宋朝 C) 秦朝 D) 汉朝</p> <p>Correct :- D) 汉朝 Predicted :- D) 汉朝</p>	<p>Question: 在蹴鞠比赛中，一名球员试图通过踢球进网来得分，但门将成功地挡住了这次射门。接下来，球员应该怎么做？</p> <p>Options : A) 比赛结束，比分保持不变 B) 比赛从球被挡住的位置恢复 C) 球被踢回场地中心重新开始 D) 门将因挡球被处罚</p> <p>Correct :- B) 比赛从球被挡住的位置恢复 Predicted :- B) 比赛从球被挡住的位置恢复</p>
<p>History Based</p> <p>Question: 在蹴鞠中，游戏的主要目标是什么？</p> <p>Options: A) 通过将球踢进球门得分 B) 将球带过整个场地 C) 将球投掷到对方的领土 D) 将球从对方手中打掉</p> <p>Correct :- A) 通过将球踢进球门得分 Predicted :- A) 通过将球踢进球门得分</p>	<p>Scenario Based</p> <p>Question: 在下图中，你可以看到一名蹴鞠球员准备踢球。球员最有可能想做什么？(提供一张球员准备踢球的图片。)</p> <p>Options: A) 通过将球踢进球门来得分 B) 将球传给队友 C) 阻止对方的进攻 D) 抓住球，防止它出界</p> <p>Correct :- A) 通过将球踢进球门来得分 Predicted :- A) 通过将球踢进球门来得分</p>

Figure 39: Example Illustration of China Traditional Sports Correct Prediction.

<p>Question: 蹴鞠, 一种古老的中国运动, 通常被认为是现代哪项运动的最早形式之一?</p> <p>Options: A) 足球 B) 橄榄球 C) 棒球 D) 篮球</p> <p>Correct :- A) 足球 Predicted :- C) 棒球</p>	<p>Question: 在蹴鞠比赛中, 一队的球员在进攻时被对方球员犯规, 裁判判定对方犯规, 并给进攻方一个罚球机会。若进攻方球员成功罚进球门, 以下哪项情况最有可能发生?</p> <p>Options : A) 进攻方得分, 并且比赛继续从中场开始 B) 进攻方得分, 但比赛暂停, 并记录犯规 C) 进攻方得分, 比赛继续, 但犯规方受到警告 D) 进攻方得分, 但比赛立即结束, 犯规方被判失格</p> <p>Correct :- A) 进攻方得分, 并且比赛继续从中场开始 Predicted :- C) 进攻方得分, 比赛继续, 但犯规方受到警告</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 40: Example Illustration of China Traditional Sports Wrong Prediction.

<p>Question: ገና, በኋላው አገልግሎት ለገኘት: በምን ብኩ የይማኖት አገልግሎት ተኋዳው ለከበሩል?</p> <p>Options: A) አገልግሎት አይደለም B) የገና በኋላ C) መስቀል (ሁምና ማተሚበት ምክበናት) D) ከሚሆነት (ሚምሮችን በተማሩ)</p> <p>Correct :- B) የገና በኋላ Predicted :- B) የገና በኋላ</p>	<p>Question: በገና ማወቃ አይደለም ተኋዳው የገናውን በይማኖት በይል አገልግሎት ምክበና አለበት?</p> <p>Options : A) በይልዎች ማስከል አይወጥናል B) ለሁሉም ሆነ ተገናው መሆኑ C) የይልዎች እና ለሁሉም የገናውን አርብ ስም መሆኑ D) አይደለም ማስከል አይደለም ምክበና</p> <p>Correct :- B) ለሁሉም ሆነ ተገናው መሆኑ<sup>1</sup> Predicted :- B) ለሁሉም ሆነ ተገናው መሆኑ</p>
<p>History Based</p>	<p>Scenario Based</p>

Figure 41: Example Illustration of Ethiopia Traditional Sports Correct Prediction.

<p>Question: Gena, a traditional Ethiopian sport, is celebrated in association with which major religious event?</p> <p>Options: A) Ethiopian New Year B) Genna (Ethiopian Christmas) C) Meskel (Finding of the True Cross) D) Education (Teachers with Students)</p> <p>Correct: B) Genna (Ethiopian Christmas) Predicted: B) Genna (Ethiopian Christmas)</p>	<p>Question: In the Gena game, how should a player respect their own team or the opposing team?</p> <p>Options: A) Show sportsmanship in the center of the game B) Be fair and abide by the rules for everyone C) Only listen to teammates, not the game referee D) Show respect only to one side and strictly follow it</p> <p>Correct: B) Be fair and abide by the rules for everyone Predicted: B) Be fair and abide by the rules for everyone</p>
<p><b>History Based</b></p> <p>Question: How does the traditional Gena game typically begin?</p> <p>Options: A) Teams walk to the center line B) The activity starts randomly anywhere C) A well-known person traditionally starts the game D) Players begin at the center</p> <p>Correct: D) Players begin at the center Predicted: D) Players begin at the center</p>	<p><b>Scenario Based</b></p> <p>Question: In the following image, a player on the line in the Gena game is participating as a team member. What is the correct behavior for the player? (Please refer to the provided image as an example of Gena game participation.)</p> <p>Options: A) As a team member, assist and support the runner B) Always protect the runner C) They are the referees shown in the image D) Remain silent and do not shout</p> <p>Correct: A) As a team member, assist and support the runner Predicted: A) As a team member, assist and support the runner</p> 

Figure 42: Example Illustration of Ethiopia Traditional Sports Correct Prediction (In English)

<p>Question: አንድ አልፎ በኋላ በአንድና የተወሰደ ገኑ በቅርቡ ወቻ የሚከበር በኋላ የቻ?</p> <p>Options: A) የአንድ መልኩ ተወስኝ B) አክሰጣም በኋላ C) አይደለም አይደለም D) የፋይነት በኋላ</p> <p>Correct :- A) የአንድ መልኩ ተወስኝ Predicted :- C) አይደለም አይደለም</p>	<p>Question: በኋና መቻዎች መካከላ እና ተወስኝ ከተኞች፣ መቻዎች ተወስኝ እና በወሰን ተወስኝ እና ተወስኝ እና ተወስኝ እና ተወስኝ?</p> <p>Options : A) የተወሰደው ተወስኝ እና በአንድና በወሰን እና ተወስኝ B) የተወሰደው ተወስኝ በላላ ተወስኝ ድንብናና እና ስለጠዋቱ C) የተወሰደው ተወስኝ መቻዎች እና የሚገኘው መቻዎች እና ስለጠዋቱ D) ተወስኝ በላላ እና የቻዎች እና ተወስኝ</p> <p>Correct :- B) የተወሰደው ተወስኝ በላላ ተወስኝ ድንብናና እና አካባቢና ስለጠዋቱ Predicted :- C) የተወሰደው ተወስኝ መቻዎች እና የሚገኘው መቻዎች እና ስለጠዋቱ</p>
<p>History Based</p> <p>Question: በኋና መቻዎች ላይ የተከለለ ገኑ የቻ?</p> <p>Options: A) ፊርማ ለመፈጸም የለ ወመሆኑ B) መቻዎች ለመፈጸም ከዚህ መቻዎች C) ተተወቃቁት መቻዎች ለማለፍ D) ላና መቻዎች ለማለፍ መስቀል</p> <p>Correct :- C) ተተወቃቁት መቻዎች Predicted :- B) መቻዎች ለመፈጸም ከዚህ መቻዎች</p>	<p>Question: በዚህ ማዣ መስጥ፣ በኋና መቻዎች በተወሰኑ ላይ ምን ምሳሌ የተሰጠት ተግባር ያተመጣው በኋና የቻ?</p> <p>Options: A) ተተወቃቁት ወለበት እንበበ ተወስኝ በወሰን ለመሆኑ ለመሆኑ የኋገናቸው B) ተተወቃቁት ወለበት የሚመለው ወለበት ተተናናው ለአሁን ለሁሉም ለሁሉም C) በኋና መቻዎች ለመሆኑ ለአሁን ለሁሉም ወለበት D) የሚመለው ወለበት በኋና የቻዎች ለመሆኑ ለሁሉም</p> <p>Correct :- A) ተተወቃቁት ወለበት እንበበ ተወስኝ በወሰን ለመሆኑ ለመሆኑ የኋገናቸው Predicted :- B) በኋና መቻዎች ለመሆኑ ለአሁን ለሁሉም ወለበት</p>
<p>Rule Based</p>	<p>Image Based</p> 

Figure 43: Example Illustration of Ethiopia Traditional Sports Wrong Prediction.

<p>Question: During the time when Gena, a major festival in Ethiopia, is celebrated, which other festival is also observed?</p> <p>Options: A) The commemoration of the Battle of Adwa B) Axum Festival C) New Year D) Easter Festival</p> <p>Correct: A) The commemoration of the Battle of Adwa Predicted: C) New Year</p>	<p>Question: During the Gena game, if a player gets injured, does the game stop or is the injured player expected to continue playing until the end? How do the team and the game proceed in this situation?</p> <p>Options: A) The injured player is taken away by a vehicle, like an ambulance B) The injured player is substituted by another player and receives medical treatment C) The injured player must continue playing the game as usual D) The player stays in one place</p> <p>Correct: B) The injured player is substituted by another player and receives medical treatment Predicted: C) The injured player must continue playing the game as usual</p>
History Based	Scenario Based

<p>Question: What is prohibited in the game of Gena?</p> <p>Options: A) Playing the ball aggressively with force B) Starting the game from the center C) Stepping outside the playing area D) Eating food during the Gena game</p> <p>Correct: C) Stepping outside the playing area Predicted: B) Starting the game from the center</p>	<p>Question: In this image, what kind of incorrect behavior is exemplified as a mistake by the opposing team during the Gena game?</p> <p>Options: A) The players wear very similar clothing, making it difficult to distinguish between teams B) The players' sticks are tangled together, so this is a mistake C) Playing with a curved stick is more traditional D) Holding the stick with both hands is a mistake</p> <p>Correct: A) The players wear very similar clothing, making it difficult to distinguish between teams Predicted: B) Playing with a curved stick is more traditional</p>
Rule Based	Image Based

Figure 44: Example Illustration of Ethiopia Traditional Sports Wrong Prediction (In English)

<p>Question: Dans quelle région de France le Tambourin était-il traditionnellement joué, et quelle était sa signification historique dans cette région ?</p> <p>Options: A) Provence, où il était utilisé pour célébrer les festivals de la récolte B) Normandie, en tant que divertissement à la cour royale C) Bretagne, où il servait d'outil d'entraînement pour les guerriers D) Alsace, principalement pour des cérémonies religieuses</p> <p>Correct :- A) Provence, où il était utilisé pour célébrer les festivals de la récolte Predicted :- A) Provence, où il était utilisé pour célébrer les festivals de la récolte</p>	<p>Question: Lors d'une partie de Tambourin, l'équipe A perd 10-8, et c'est le dernier round. Le joueur 1 de l'équipe A se prépare à frapper la balle. Cependant, le joueur 2 de l'équipe B commence soudainement à distraire le joueur 1 en agitant les bras. Que doit faire l'équipe A ?</p> <p>Options : A) Laisser le joueur 1 continuer sans rien dire B) Demander à l'arbitre d'arrêter le jeu et de traiter la distraction C) Ignorer la distraction et continuer à jouer D) Remplacer le joueur 1 par quelqu'un d'autre de leur équipe</p> <p>Correct :- B) Demander à l'arbitre d'arrêter le jeu et de traiter la distraction Predicted :- B) Demander à l'arbitre d'arrêter le jeu et de traiter la distraction</p>
History Based	Scenario Based

<p>Question: Quel est l'objectif principal dans le jeu de Tambourin ?</p> <p>Options: A) Marquer des points en frappant une balle à travers un cerceau B) Marquer des points en frappant une balle contre une cible C) Garder la balle en l'air sans qu'elle touche le sol D) Frapper la balle au-dessus du filet et dans le côté de l'adversaire</p> <p>Correct :- B) Marquer des points en frappant une balle contre une cible Predicted :- B) Marquer des points en frappant une balle contre une cible</p>	<p>Question: À partir de l'image donnée, quel est l'objectif principal de cette action dans le jeu de Tambourin ?</p> <p>Options: A) Passer la balle à un coéquipier B) Frapper la balle vers la zone cible et marquer des points C) Bloquer les mouvements de l'adversaire D) Défendre leur côté du terrain</p> <p>Correct :- B) Frapper la balle vers la zone cible et marquer des points Predicted :- B) Frapper la balle vers la zone cible et marquer des points</p>
Rule Based	Image Based

Figure 45: Example Illustration of France Traditional Sports Correct Prediction

<p>Question: In which region of France was the Tambourin traditionally played, and what was its historical significance in that region?</p> <p>Options: A) Provence, where it was used to celebrate harvest festivals B) Normandy, as entertainment at the royal court C) Brittany, where it served as a training tool for warriors D) Alsace, mainly for religious ceremonies</p> <p>Correct: A) Provence, where it was used to celebrate harvest festivals Predicted: A) Provence, where it was used to celebrate harvest festivals</p>	<p>Question: During a game of Tambourin, Team A is losing 10-8, and it's the last round. Player 1 from Team A is preparing to hit the ball. However, Player 2 from Team B suddenly starts distracting Player 1 by waving their arms. What should Team A do?</p> <p>Options: A) Let Player 1 continue without saying anything B) Ask the referee to stop the game and address the distraction C) Ignore the distraction and keep playing D) Substitute Player 1 with someone else from their team</p> <p>Correct: B) Ask the referee to stop the game and address the distraction Predicted: B) Ask the referee to stop the game and address the distraction</p>
<p>History Based</p>	<p>Scenario Based</p>

<p>Question: What is the main objective in the game of Tambourin?</p> <p>Options: A) Score points by hitting a ball through a hoop B) Score points by hitting a ball against a target C) Keep the ball in the air without letting it touch the ground D) Hit the ball over the net and into the opponent's side</p> <p>Correct: B) Score points by hitting a ball against a target Predicted: B) Score points by hitting a ball against a target</p>	<p>Question: Based on the given image, what is the main objective of this action in the game of Tambourin?</p> <p>Options: A) Pass the ball to a teammate B) Hit the ball toward the target area and score points C) Block the opponent's movements D) Defend their side of the court</p> <p>Correct: B) Hit the ball toward the target area and score points Predicted: B) Hit the ball toward the target area and score points</p>
<p>Rule Based</p>	<p>Image Based</p>

Figure 46: Example Illustration of France Traditional Sports Correct Prediction (In English)

<p>Question: Dans l'histoire du Tambourin, quel rôle important ce sport a-t-il joué dans les échanges culturels entre la France et l'Italie au XVIIe siècle ?</p> <p>Options: A) Il a influencé le développement du tennis moderne B) Il était un outil diplomatique utilisé pour renforcer les alliances à travers des événements sportifs partagés C) Il est devenu un sport compétitif dans les cours royales des deux pays D) Il était utilisé comme méthode d'entraînement militaire pendant les périodes de guerre</p> <p>Correct :- B) Il était un outil diplomatique utilisé pour renforcer les alliances à travers des événements sportifs partagés Predicted :- C) Il est devenu un sport compétitif dans les cours royales des deux pays</p>	<p>Question: Lors d'une partie de Tambourin, l'équipe X mène de 5 points. L'équipe Y parvient à marquer un point avec une superbe frappe, mais la balle touche le sol juste après avoir franchi la zone cible. L'équipe X affirme qu'il s'agit d'une faute, tandis que l'équipe Y pense que c'était un point valide. Que doit-il se passer ensuite ?</p> <p>Options : A) L'arbitre attribue le point à l'équipe Y, car la balle a franchi la zone cible B) Le point est attribué à l'équipe X en raison de la balle ayant touché le sol C) Un arbitre neutre ou un joueur senior doit revoir le jeu et prendre une décision D) Les deux équipes rejouent le point</p> <p>Correct :- A) L'arbitre attribue le point à l'équipe Y, car la balle a franchi la zone cible Predicted :- C) Un arbitre neutre ou un joueur senior doit revoir le jeu et prendre une décision</p>
<p>History Based</p>	<p>Scenario Based</p>

<p>Question: Dans le jeu de Tambourin, quelle est l'une des règles importantes concernant le mouvement des joueurs pendant la partie ?</p> <p>Options: A) Les joueurs peuvent se déplacer librement sur le terrain sans restrictions B) Les joueurs sont limités à leur côté du terrain et ne peuvent pas traverser la ligne centrale C) Les joueurs doivent rester derrière une ligne spécifique lorsqu'ils servent la balle D) Les joueurs doivent utiliser uniquement une main pour frapper la balle pendant le jeu</p> <p>Correct :- B) Les joueurs sont limités à leur côté du terrain et ne peuvent pas traverser la ligne centrale Predicted :- C) Les joueurs doivent rester derrière une ligne spécifique lorsqu'ils servent la balle</p>	<p>Question: Quelle situation décrit cette image ?</p> <p>Options: A) Le joueur avec le tambourin est sur le point d'exécuter une attaque puissante, tandis que l'autre joueur est positionné pour bloquer le coup B) Le joueur avec le tambourin prépare un mouvement défensif pour protéger son objectif C) Les deux joueurs attendent que le jeu redémarre après un temps mort D) Le joueur qui défend l'objectif essaie de distraire le frappeur</p> <p>Correct :- A) Le joueur avec le tambourin est sur le point d'exécuter une attaque puissante, tandis que l'autre joueur est positionné pour bloquer le coup Predicted :- B) Le joueur avec le tambourin prépare un mouvement défensif pour protéger son objectif</p>
<p>Rule Based</p>	<p>Image Based</p>

Figure 47: Example Illustration of France Traditional Sports Wrong Prediction

<p>Question: In the history of Tambourin, what important role did this sport play in cultural exchanges between France and Italy in the 18th century?</p> <p>Options: A) It influenced the development of modern tennis B) It was a diplomatic tool used to strengthen alliances through shared sporting events C) It became a competitive sport in the royal courts of both countries D) It was used as a military training method during times of war</p> <p>Correct: B) It was a diplomatic tool used to strengthen alliances through shared sporting events Predicted: C) It became a competitive sport in the royal courts of both countries</p>	<p>Question: During a game of Tambourin, Team X is leading by 5 points. Team Y manages to score a point with a great shot, but the ball touches the ground just after crossing the target area. Team X claims it's a fault, while Team Y believes it was a valid point. What should happen next?</p> <p>Options: A) The referee awards the point to Team Y because the ball crossed the target area B) The point is awarded to Team X because the ball touched the ground C) A neutral referee or a senior player should review the play and make a decision D) Both teams replay the point</p> <p>Correct: A) The referee awards the point to Team Y because the ball crossed the target area Predicted: C) A neutral referee or a senior player should review the play and make a decision</p>
<b>History Based</b>	<b>Scenario Based</b>

Figure 48: Example Illustration of France Traditional Sports Wrong Prediction (In English)

<p>Question: In welchem Jahr wurde Radball offiziell als Sportart vom Deutschen Radsportverband anerkannt?</p> <p>Options: A) 1885 B) 1900 C) 1920 D) 1930</p> <p>Correct :- A) 1885 Predicted :- A) 1885</p>	<p>Question: Während eines Radball-Spiels versucht Spieler A, ein Tor zu schießen, indem er den Ball in Richtung des gegnerischen Tors schießt, aber der Ball wird vom Torwart des Gegners blockiert. Der Ball trifft dann den Pfosten und prallt ins Spielfeld zurück. Was sollte Spieler A als Nächstes tun?</p> <p>ok Options : A) Spieler A muss sofort einen zweiten Schuss auf das Tor versuchen B) Das Spiel geht weiter, und Spieler A's Team muss sich für einen weiteren Versuch neu formieren C) Spieler A wird bestraft, weil er den Pfosten getroffen hat d) Das Spiel wird gestoppt und ein Freiwurf für Spieler A's Team wird vergeben</p> <p>Correct :- B) Das Spiel geht weiter, und Spieler A's Team muss sich für einen weiteren Versuch neu formieren Predicted :- B) Das Spiel geht weiter, und Spieler A's Team muss sich für einen weiteren Versuch neu formieren</p>
<b>History Based</b>	<b>Scenario Based</b>

<p>Question: Wie viele Spieler sind in einem Radball-Team?</p> <p>Options: A) 1 B) 2 C) 3 D) 4</p> <p>Correct :- B) 2 Predicted :- B) 2</p>	<p>Question: Was ist die wahrscheinlichste Aktion, die die Spieler auf diesem Bild ausführen?</p> <p>Options: A) Sie blockieren den Ball mit ihren Fahrrädern B) Sie versuchen, ein Tor zu schießen C) Sie versuchen, den Ball zu einem Mitspieler zu passen D) Sie verhindern, dass der Ball aus dem Spielfeld geht</p> <p>Correct :- B) Sie versuchen, ein Tor zu schießen Predicted :- B) Sie versuchen, ein Tor zu schießen</p>
---	---

Figure 49: Example Illustration of Germany Traditional Sports Correct Prediction