

Article

Sports Intelligence: Assessing the Sports Understanding Capabilities of Language Models Through Question Answering from Text to Video

Zhengbang Yang ^{1,†}, Haotian Xia ^{2,†}, Jingxi Li ², Zezhi Chen ², Zhuangdi Zhu ^{1,*}  and Weining Shen ^{2,*} 

¹ Department of Cyber Security Engineering, George Mason University, Fairfax, VA 22030, USA; zyang30@gmu.edu

² Department of Statistics, University of California, Irvine, CA 92697, USA; xiah6@uci.edu (H.X.)

* Correspondence: zzhu24@gmu.edu (Z.Z.); weinings@uci.edu (W.S.)

† These authors contributed equally to this work.

Abstract: Understanding sports presents a fascinating challenge for Natural Language Processing (NLP) due to its intricate and ever-changing nature. Current NLP technologies struggle with the advanced cognitive demands required to reason over complex sports scenarios. To explore the current boundaries of this field, we extensively evaluated main-stream and emerging large models on various sports tasks and addressed the limitations of previous benchmarks. Our study ranges from answering simple queries about basic rules and historical facts to engaging in complex, context-specific reasoning using strategies like few-shot learning and chain-of-thought techniques. Beyond text-based analysis, we also explored the sports reasoning capabilities of mainstream video language models to bridge the gap in benchmarking multimodal sports understanding. Based on a comprehensive overview of main-stream large models on diverse sports understanding tasks, we presented a new benchmark, which highlighted the critical challenges of sports understanding for NLP and the varying capabilities of state-of-the-art large models on sports understanding. We also provided an extensive set of error analyses that pointed to detailed reasoning defects of large model reasoning which model-based error analysis failed to reveal. We hope the benchmark and the error analysis set will help identify future research priorities in this field.

Keywords: benchmarking; evaluation; large language models



Academic Editors: Zhiqian Liu and Arkaitz Zubiaga

Received: 18 December 2024

Revised: 17 January 2025

Accepted: 21 January 2025

Published: 23 January 2025

Citation: Yang, Z.; Xia, H.; Li, J.; Chen, Z.; Zhu, Z.; Shen, W. Sports Intelligence: Assessing the Sports Understanding Capabilities of Language Models Through Question Answering from Text to Video. *Electronics* **2025**, *14*, 461. <https://doi.org/10.3390/electronics14030461>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Automated sports officiating marks a revolutionary advancement in enhancing fairness, accuracy, and efficiency in game management. In the form of Video Assistant Referees (VAR), Artificial Intelligence has already been implemented in various sports including football [1] and tennis ATP [2]. Notably, the ATP has announced that electronic line calling (ELC) will be applied to all ATP tournaments by 2025 [3], indicating a trend towards automation of tennis officiating. Spitz et al. [4] have shown that the use of VAR has significantly increased the likelihood of making correct decisions in football matches. Tamir and Bar-Eli [5] pointed out that the VAR system has promoted fairness and more accurate decisions, bringing ethical transformations in the football world.

Meanwhile, athletes are achieving enhanced physical conditions [6–9] and notably faster speeds [10] with the progress of nutritional science and systematic training, which has significantly increased the complexity and difficulty of the tasks of human referees.

This calls for robust, real-time decision-making assistive tools. Language models, specifically Large Language Models (LLMs) and Video Language Models (VLMs), have offered promising solutions by processing complex contextual data of textual and visual modalities from sports events to assist in impartial decision-making.

Our research focuses on the Sports Understanding ability of LLMs and VLMs, an under-explored area yet is crucial for their potential applications in automated refereeing and related domains. Previous benchmarks have fallen short by either focusing on datasets containing limited sports understanding [11], relying on a single modality [12], or lacking detailed error analysis [12,13]. Moreover, no prior work has addressed the capabilities of the latest LLMs, especially in light of recent rapid advancements in LLMs and VLMs.

To assess the sports understanding capabilities of LLMs and VLMs, we introduced a new benchmark dubbed as *Sport Intelligence*. The prevailing interaction method with LLMs and VLMs is based on multi-round dialogues. To evaluate their sports understanding capabilities, we need to pose questions to the models and subsequently analyze their responses. Given the congruence between the question-answer format and real-world sports scenarios, we established our benchmarks on QA datasets to facilitate a comprehensive evaluation. We conducted an comprehensive investigation of leading LLMs, including Llama3.1 [14], Llama3 [15], the GPT4 series [16,17], the Gemini 1.5 series [18,19], Claude3.5 Sonnet [20], and Claude3 Opus [21], as well as the latest VLMs including Minigpt-4 [22], Chat-UniVi [23], PLLaVA [24], and Video-LLaVA [25,26].

Our main contributions are as follows:

1. We consolidated existing sports-related QA datasets and introduced the most up-to-date benchmark to address the gaps in multimodal sports understanding benchmarks. Our work serves as the first benchmark that incorporates both text-based and video-based QA categorized by varying levels of complexity, which offers the most current and thorough assessment of the latest cutting-edge LLMs and VLMs in the realm of sports understanding.
2. Through a detailed evaluation, we have identified the limitations of existing foundation models in sports reasoning under complex and multi-modal scenarios, the challenges of multihop sports questions answering with long sequence prompts, and demonstrated the promising potential in integrating techniques such as chain-of-thought and multi-agent voting to enhance performance. Our findings emphasize the necessity of incorporating domain-specific training methods to enhance the reasoning capabilities of LLMs/VLMs in sports fields.
3. We performed an in-depth error analysis to uncover the primary causes of errors and detailed reasoning flaws in large models that were not evident through model-based error analysis alone. We hope that the error analysis results from this work can be readily used as in-context data resources to enhance the performance of LLMs and VLMs in sports understanding.

2. Related Work

QA in Sports: Sports in NLP has gained increasing recognition. The SportQA dataset [12] has introduced original and high-quality sports understanding questions and has integrated sports-related questions from popular QA datasets, including KQA Pro [27], BoolQ [28], HotpotQA [29], QUASAR [30], and Trivia QA [31]. Additionally, Sports-QA [13] presented the first video QA dataset targeted at sports understanding. Most QA datasets contained only a few questions related to sports understanding, primarily focusing on historical events rather than rules, strategies, or complex situational analysis, which limited the depth of LLMs' understanding of sports. For example, BIG-bench [11] includes a Sports Understanding subtask that differentiates whether a sports-related state-

ment is credible, requiring LLMs to recognize athletes' names, the sports they engage in, and the actions involved in specific sports. While related to sports understanding, the Sports Understanding subtask of BIG-bench lacks complex situational analysis and is relatively simple. QASports [32] is also related to sports which essentially serves as a context extraction QA dataset and does not require an understanding of sports. LiveQA [33] is derived from live broadcast segments and aims to test models' ability to reason over time series. Given that these are real matches, the answers to its questions might be retrievable from pre-trained corpora. To the best of our knowledge, we are the first to establish a benchmark that integrates both text-based QA and video-based QA with refined complexity level categorizations. Our investigation serves as the most up-to-date and comprehensive evaluation of the latest state-of-the-art LLMs and VLMs on sports understanding.

LLM Paradigms: Represented by BERT [34], pretrained models form the foundation of current NLP. GPT-3 has ushered in a new era of large-scale, general-purpose, multi-task handling, and high-quality generation for language models. GPT-3 excels in few-shot and zero-shot learning [35], allowing it to handle new tasks with minimal or no training data. This shift has spurred the development of advanced prompting techniques that enhance understanding and reasoning capabilities, including Chain of Thought (CoT) prompting [36] and zero-shot CoT prompting [37]. The emergence of GPT-4 [16] further enhances model generation quality and the capability to solve complex tasks. The subsequent introduction of the Gemini family (Gemini [38], Gemini 1.5 [18], and Gemini 1.5 flash [19]), Claude3 [21], and Llama3 [15] has intensified the competition among LLMs. The availability of the free GPT-4o [17] has brought state-of-the-art LLMs into everyday life, demonstrating the broad impact and accessibility of these advanced technologies.

VLM Paradigms: Currently, most VLMs are constructed using LLMs as decoders. Taking Video-LLaVa as an example, it employs the LanguageBind [39] encoder to map video or image data into a textual feature space, serving as a unified visual representation. After encoding by a shared projection layer, the unified visual representation is then fed into a large language model with tokenized text queries to generate the corresponding response. Models such as LLaMA-Adapter [40,41] and ImageBind-LLM [42] have demonstrated that a unified feature space enhances the multimodal reasoning capabilities of LLMs. This unified approach facilitates better integration of visual and textual data, enabling LLMs to perform more effectively across diverse multimodal tasks. Other VLMs that have been evaluated, such as Minigpt-4 [22], Chat-UniVi [23], and PLLaVA [24], are also implemented in a similar manner.

3. Sports Understanding Dataset and Benchmark

3.1. Dataset Construction

As discussed in Section 2, QASports [32] and LiveQA [33] are not suitable for the Sports Understanding tasks. Therefore, we carefully consolidated dataset from SportQA [12], Sports-QA [13], and the Sports Understanding subtask of BIG-bench [11] to construct the dataset for our benchmark.

SportQA [12] is the latest textual dataset focused on multiple-choice sports question answering (QA), comprising three levels of complexity. Level-1 encompasses 21,385 questions based on existing sports-related QA datasets, including KQA Pro [27], BoolQ [28], HotpotQA [29], QUASAR [30], and Trivia QA [31]. Level-2 contains 45,685 questions involving 35 types of sports, primarily focusing on rules, tactical understanding, and complex sports history inquiries. Level-3 consists of manually designed, scenario-based advanced sports comprehension questions, encompassing 3522 questions and six popular sports events, including football, basketball, volleyball, tennis, table tennis, and American football. Each type of sport in Level-3 involves both single-hop and multi-hop questions which

feature multiple-choice formats (one to four correct answers), and come with difficulty markings. Consequently, performing well on Level-3 requires LLMs to have a profound understanding of each sport type, including comprehending detailed sports rules, e.g., how penalties are assessed, reasoning over context-specific information, and making tactical choices. The comprehensive scope and high quality of SportQA make it particularly well-suited for evaluating the sports understanding capabilities of LLMs and we use its testing sets for our experiments.

Sports-QA [13] is the first video QA dataset dedicated to sports activities. It includes a total of 6000 videos and 94,000 questions covering 8 different sports. Sourced from MultiSports [43] and FineGym [44], this dataset features sports videos and professional action annotations, making it a high-quality resource. Sports-QA includes four types of questions: Descriptive, Temporal, Causal, and Counterfactual, which collectively assess video-based sports understanding from multiple perspectives. Moreover, questions in Sports-QA related to specific terms and actions require models to perform reasoning across various interactive scenarios, making it highly suitable for evaluating the sports understanding capabilities of VLMs and we use its testing sets for our experiments.

Big-bench [11] is a collaborative benchmark designed to explore the future capabilities of LLMs. It contains a Sports Understanding subtask which consists of 986 2-way multiple-choice questions that primarily aim to test LLMs' general understanding of sports activities. Specifically, it assesses the models' ability to distinguish whether sports-related statements are plausible. The 986 question statements primarily consist of athlete names and their actions, requiring the model to determine whether the athletes' actions align with the sport they are engaged in. Therefore, we consider the Sports Understanding subtask in BIG-bench appropriate for evaluating the basic sports comprehension capabilities of LLMs and we use its full dataset for our experiments.

3.2. Tasks Categorization

Considering the diversity of the datasets we selected, we have categorized the tasks into the following three subtasks: (1) Basic Sports Understanding, (2) Advanced Scenario Analysis, and (3) Video-based Sports Understanding. These subtasks are designed to provide a comprehensive evaluation of various aspects of sports understanding, ranging from fundamental knowledge of sports to complex situational analysis, and understanding of actions within sports videos, as shown in Figure 1.

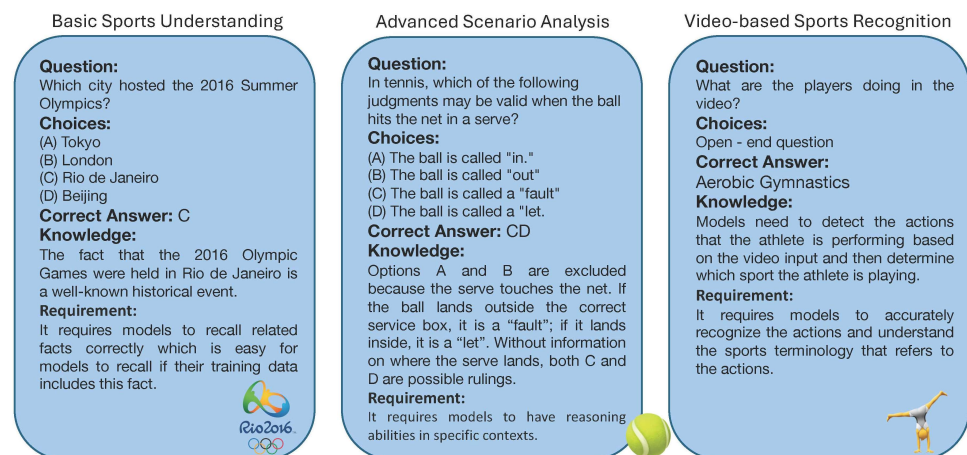


Figure 1. Sample questions from the proposed sports understanding benchmark.

Basic Sports Understanding: This subtask examines the primary understanding of sports, including historical events and facts about sports, basic rules, and the roles of basic tactics. It requires LLMs to comprehend relevant knowledge without the need for complex reasoning. The Level-1 of SportQA [12], which focuses on historical events and facts about sports, and Level-2, which covers basic sports rules, basic tactics, and a broader range of historical and factual knowledge, are well-suited for this task. The Sports Understanding subtask of BIG-bench [11], which mainly tests whether names and actions match, does not involve complex reasoning; therefore, we also categorize it under Basic Sports Understanding.

Advanced Scenario Analysis: This subtask evaluates LLMs' advanced sports understanding capabilities, particularly their ability to reason over complex situations, such as making penalty decisions or tactical choices in specific scenarios. It requires LLMs to have advanced capabilities to understand sports contexts and perform complex reasoning. The Level-3 of SportQA [12] mainly consists of complex scenario analysis questions that replicate real-world sports situations, which is highly suitable for this task. In contrast to prior work [24] that utilized GPT-3.5 for evaluation, we employed GPT-4 to score the generated answers of video-based questions as an evaluation metric.

Video-based Sports Recognition: This subtask aims to assess the fundamental understanding of sports through video content. This task examines whether models can accurately identify specific actions and the number of participants in various sports events. The questions in Sports-QA [13] focus on specific terms and actions related to sports, requiring models to perform simple reasoning across various interactive scenarios. Therefore, this dataset is highly suitable for this task.

Table 1 shows the overview of datasets and tasks.

Table 1. Overview of tasks included in this paper. The “Size” refers to the total number of data in both the training, development, and test sets. “*n*-Way MC” and “*n*-Way MS” signify multiple-choice response format and multiple-select choice (one or multiple correct answers) format respectively with *n* options. “Acc.” represents accuracy. “Sco.” represents GPT-4’s score for the generated answers (with a maximum score of 5).

Dataset	Size	Metrics	Answer Type
Basic Sports Understanding			
SportQA Level-1	21,385	Acc.	4-Way or 2-Way MC
SportQA Level-2	45,685	Acc.	4-Way MC
BIG-bench Sports Understanding	986	Acc.	2-Way MC
Advanced Scenario Analysis			
SportQA Level-3 Easy Multi-hop Tasks	915	Acc.	Multi-hop 4-Way MS
SportQA Level-3 Hard Multi-hop Tasks	808	Acc.	Multi-hop 4-Way MS
SportQA Level-3 Easy Single-hop Tasks	903	Acc.	Single-hop 4-Way MS
SportQA Level-3 Hard Single-hop Tasks	896	Acc.	Single-hop 4-Way MS
Video-based Sports Recognition			
Sports-QA Descriptive	48,268	Acc. & Sco.	Open end
Sports-QA Temporal	39,643	Acc. & Sco.	Open end
Sports-QA Causal	4676	Acc. & Sco.	Open end
Sports-QA Counterfactual	1486	Acc. & Sco.	Open end

4. Experiment

We evaluated the performance of prominent LLMs and VLMs on the above tasks to assess their ability to comprehend complex sports scenes across various tasks.

4.1. Experimental Setup, Models, and Baselines

Models: We evaluated leading LLMs including Llama3 [15], the GPT4 series [16,17], the Gemini 1.5 series [18,19], and Claude [21] on text-based tasks. Access to these models was facilitated through their respective APIs. Additionally, we evaluated the latest VLMs including Minigpt-4 [22], Chat-UniVi [23], PLLaVA [24], and Video-LLaVA [25,26] on video-based sports recognition. LLMs were accessed via their respective APIs, while VLMs were deployed locally on the cloud server.

Baseline Methods:

Chain-of-Thought (CoT): Wei et al. [36] highlighted the effectiveness of few-shot CoT in sports understanding contexts. Hence we primarily focused on the CoT prompting method for model evaluation, which involves a step-by-step reasoning process suitable for complex sports understanding tasks.

Few-shot Standard Prompting (SP): Brown et al. [35] noted that few-shot SP can enhance model performance. Thus, for textual tasks, we considered the zero-shot CoT method [37] and few-shot SP as additional prompting baselines. We used a 5-shot setup for the few-shot SP. In both few-shot settings (CoT or SP), we employed 5-shot prompts annotated by human experts provided by the SportQA dataset; for BIG-bench, we used the first five examples provided by CoT. For Basic Sports Understanding and Advanced Scenario Analysis, the temperature parameter was set to zero to ensure consistent responses. The prompt setting refers to Appendix A.1.

For Video-based Sports Understanding, we only employed the zero-shot CoT [37] as an additional prompting baseline. This choice was primarily due to technical limitations, as current VLMs can only process a single video as input. Due to the large volume of Video-based QA data, we used GPT-4 to assist in determining the accuracy of the VLMs' responses. The temperature parameter was set to zero to ensure consistent responses. The prompt setting refers to Appendix A.2.

Multi-Agent Voting: Chan et al. [45] proposes a method to enhance the performance of LLMs in text evaluation by LLMs collaboration through Multi-Agent debate. To investigate whether collaborative knowledge among SOTA LLMs can yield superior performance we formed a voting committee with evaluated LLMs to investigate whether collective intelligence yields superior performance, considering the outputs from LLMs might be too brief in few-shot SP scenarios, making which unsuitable for debate. Each LLM submits an answer, and the response with the most votes is selected as the final answer.

4.2. Performance Evaluation

4.2.1. Performance Overview

For text-based tasks, the performance of LLMs on Basic Sports Understanding and Advanced Scenario Analysis are presented in Tables 2 and 3, respectively, which show that LLMs generally performed better in the Basic Sports Understanding task but faltered in Advanced Scenario Analysis. The results of VLMs on *video-based sports recognition* tasks are shown in Table 4. The Auto-Focus Transformer (AFT), proposed by SportsQA [13], achieved the highest accuracy, while the overall scores of VLMs remain low, indicating significant room for improvement. This highlights the complexity of multimodal sports understanding tasks and the need for VLMs to strengthen their capabilities in recognizing specific sports and actions.

Table 2. Model performance comparison (accuracy \uparrow) on the *Basic Sports Understanding* task. “0S” represents zero-shot setup; “0S,CoT” represents zero-shot CoT setup; “5S,SP” represents 5-shot SP setup; “5S,CoT” represents 5-shot CoT setup. We conducted χ^2 to support difference exists between performances under different models and settings. The p -values of χ^2 are nearly 0 on Level-1, Level-2, and Big-bench. On each sub-task, we can reject H_0 : There is no difference between models’ performance. \sim means nearly 0 ($<1 \times 10^{-10}$). Bolded number highlights the best performance in a sub-task.

Model	Level-1	Level-2	Big-Bench
Llama3-70B(0S)	75.65	68.54	74.40
Llama3-70B(0S,CoT)	75.20	68.59	84.70
Llama3-70B(5S,SP)	76.85	72.91	87.20
Llama3-70B(5S,CoT)	78.15	71.89	71.40
Llama3.1-405B(0S)	80.15	81.26	81.50
Llama3.1-405B(0S,CoT)	81.40	74.92	82.80
Llama3.1-405B(5S,SP)	75.25	80.50	88.50
Llama3.1-405B(5S,CoT)	82.65	78.40	96.90
Gemini 1.5 Pro(0S)	80.45	72.29	75.50
Gemini 1.5 Pro(0S,CoT)	75.90	68.41	83.30
Gemini 1.5 Pro(5S,SP)	79.15	71.84	73.40
Gemini 1.5 Pro(5S,CoT)	67.45	71.44	95.40
Gemini 1.5 Flash(0S)	66.75	59.97	73.70
Gemini 1.5 Flash(0S,CoT)	65.30	58.23	83.60
Gemini 1.5 Flash(5S,SP)	48.95	66.89	79.60
Gemini 1.5 Flash(5S,CoT)	65.40	62.16	94.60
Claude 3 Opus(0S)	79.15	72.16	82.00
Claude 3 Opus(0S,CoT)	78.75	68.90	81.20
Claude 3 Opus(5S,SP)	78.55	77.38	91.30
Claude 3 Opus(5S,CoT)	79.85	76.17	93.30
Claude 3.5 Sonnet(0S)	81.20	78.09	83.00
Claude 3.5 Sonnet(0S,CoT)	86.25	77.51	84.20
Claude 3.5 Sonnet(5S,SP)	83.50	81.08	91.90
Claude 3.5 Sonnet(5S,CoT)	74.30	75.68	94.44
GPT-4(0S)	83.70	74.48	80.80
GPT-4(0S,CoT)	84.00	74.39	82.50
GPT-4(5S,SP)	83.40	75.90	89.40
GPT-4(5S,CoT)	84.05	75.15	94.70
GPT-4o(0S)	81.70	73.98	78.50
GPT-4o(0S,CoT)	85.35	66.00	84.30
GPT-4o(5S,SP)	72.85	69.50	90.40
GPT-4o(5S,CoT)	81.90	77.33	94.60
p -value of χ^2	0.00 \sim	0.00 \sim	0.00 \sim

Table 3. Comparison of model performance (accuracy \uparrow) for the *Advanced Scenario Analysis* task. GPT-4 family consistently outperforms other models under both zero-shot (0S) and 5-shot (5S) settings across all subtasks. Human performance (provided by SportQA [12]) serves as an upper bound, illustrating that there still exists room for improvement in LLMs on sports understanding tasks. We conducted χ^2 to support difference exists between performances under different models and settings. The p -value of χ^2 are close to 0 on Easy Single-hop, Hard Single-hop, Easy Multi-hop, and Hard Multi-hop. We can reject H_0 : There is no difference between models’ performance. \sim means nearly 0 ($<1 \times 10^{-10}$). Bolded number number highlights the best performance in a sub-task.

Model	Easy Single-Hop	Hard Single-Hop	Easy Multi-Hop	Hard Multi-Hop
Llama3-70B(0S)	63.49	60.57	35.92	21.10
Llama3-70B(0S,CoT)	67.06	63.41	37.14	21.10
Llama3-70B(5S,SP)	61.11	61.79	26.94	16.46
Llama3-70B(5S,CoT)	58.33	58.13	26.12	17.30

Table 3. *Cont.*

Model	Easy Single-Hop	Hard Single-Hop	Easy Multi-Hop	Hard Multi-Hop
Llama3.1-405B(0S)	68.25	63.41	39.18	21.94
Llama3.1-405B(0S,CoT)	65.48	59.76	32.24	22.36
Llama3.1-405B(5S,SP)	70.24	63.41	40.82	27.85
Llama3.1-405B(5S,CoT)	69.05	66.67	42.45	28.27
Gemini 1.5 Pro(0S)	66.67	56.91	33.47	19.83
Gemini 1.5 Pro(0S,CoT)	61.51	53.25	33.06	20.25
Gemini 1.5 Pro(5S,SP)	66.27	58.13	40.82	24.47
Gemini 1.5 Pro(5S,CoT)	65.48	57.32	39.18	21.94
Gemini 1.5 Flash(0S)	57.94	54.47	35.92	21.94
Gemini 1.5 Flash(0S,CoT)	59.13	53.25	36.32	21.94
Gemini 1.5 Flash(5S,SP)	60.32	57.72	38.78	22.36
Gemini 1.5 Flash(5S,CoT)	64.68	58.13	35.51	19.83
Claude 3 Opus(0S)	66.67	60.16	40.41	27.00
Claude 3 Opus(0S,CoT)	58.73	59.75	42.86	28.27
Claude 3 Opus(5S,SP)	55.95	43.09	40.00	26.16
Claude 3 Opus(5S,CoT)	64.29	58.53	42.86	29.11
Claude 3.5 Sonnet(0S)	71.83	66.67	45.31	26.58
Claude 3.5 Sonnet(0S,CoT)	69.44	64.23	47.76	30.80
Claude 3.5 Sonnet(5S,SP)	72.22	66.67	44.90	29.96
Claude 3.5 Sonnet(5S,CoT)	73.81	71.14	45.71	32.07
GPT-4(0S)	71.83	65.45	38.78	26.58
GPT-4(0S,CoT)	73.02	67.48	42.04	31.65
GPT-4(5S,SP)	70.63	63.41	38.37	28.69
GPT-4(5S,CoT)	67.46	63.01	44.49	27.43
GPT-4o(0S)	76.98	69.51	38.78	22.78
GPT-4o(0S,CoT)	79.37	73.17	38.37	27.00
GPT-4o(5S,SP)	77.38	71.14	39.59	27.43
GPT-4o(5S,CoT)	78.17	72.36	42.45	29.11
p -value of χ^2	0.00~	0.00~	0.000006409	0.0001553
Human	96.63	96.02	94.90	91.84

Table 4. Model performance comparison (accuracy \uparrow) on the *Video-based Sports Recognition* (Sports-QA [13]). The results of Auto-Focus Transformer (AFT) come from Sports-QA [13]. The results suggest that video-based sports understanding poses a significant challenge for all VLMs. “Avg. Score” represents GPT-4’s score for the generated answers (with a maximum score of 5) [24]. We conducted χ^2 to support difference exists between performances under different models and settings. The p -value of χ^2 are nearly 0 on Descriptive, Temporal, Causal, Counterfactual, and Overall. On each sub-task, we can reject H_0 : There is no difference between models’ performance. ~ means nearly 0 ($<1 \times 10^{-10}$). Bolded number highlights the best performance in a sub-task.

Model	Descriptive	Temporal	Causal	Counterfactual	Overall	Avg. Score
Minigt-4-7B(0S)	36.85	14.10	23.20	36.51	26.70	1.68
Minigt-4-7B(0S,CoT)	36.38	14.15	19.79	34.54	26.26	1.68
Chat-UniVi-7B(0S)	45.31	16.23	24.23	6.58	31.52	1.89
Chat-UniVi-7B(0S,CoT)	45.04	14.71	23.09	13.82	30.81	1.92
PLLaVA-7B(0S)	32.98	8.82	19.18	17.76	21.99	1.60
PLLaVA-7B(0S,CoT)	28.11	9.82	9.48	10.86	19.28	1.49
Video-LLaVA-7B(0S)	42.80	13.52	36.80	42.76	30.33	1.87
Video-LLaVA-7B(0S,CoT)	43.46	13.50	35.57	39.47	30.55	1.90
AFT with GloVe(0S)	78.9	35.3	55.1	56.3	59.2	-
AFT with BERT(0S)	78.3	35.5	56.8	58.2	59.1	-
p -value of χ^2	0.00~	0.00~	0.00~	0.00~	0.00~	

We also delved into the model evaluation results and identified key findings below:

- *Open-source models demonstrated performance in Basic Sports Understanding comparable to that of closed-source models:* Llama3.1-405B achieved the best performance on Big-bench, while its performance in Level-1 and Level-2 tasks was not significantly behind the best-performing Claude 3.5 Sonnet. This is a positive signal for the open-source community, suggesting that it is feasible to fine-tune open-source models to achieve state-of-the-art performance.

- *Relative progress in advanced sports reasoning tasks:* Compared to the benchmarks set by SportQA [12], there is no significant improvement in the performance ceiling of LLMs in the *Basic Sports Understanding* task, which means that the domain of sports understanding did not receive sufficient attention in LLM training. Contrastively, we observed notable progress of LLMs on the *Advanced Scenario Analysis* task. Compared to the SportQA benchmarks, the performance ceiling of LLMs has improved by approximately 4.5%, 7.0%, 10%, and 7% in Easy Single-hop, Hard Single-hop, Easy Multi-hop, and Hard Multi-hop, respectively.
- *GPT-4 and Claude series are the frontrunners:* While LLM model performance varied and all have shown notable improvement in complex sports understanding, the GPT-4 and Claude series perform the best across all models. GPT-4o performs better in single-hop tasks while Claude Sonnet excels in multi-hop tasks. GPT-4o and Claude Sonnet perform better than other models in their respective series, aligning with their status as flagship models of their companies.
- *CoT benefits sports reasoning:* Specifically, regarding the effectiveness of hints, we have observed that the best results on the four subtasks of Advanced Scenario Analysis all include CoT. We conclude that step-by-step hints are effective in improving the performance of LLM in complex reasoning tasks, particularly in cases involving a small number of examples. In addition, CoT reasoning significantly improves performance on Level-1 in Basic Sports Understanding and on Big-bench, indicating that reasoning is beneficial even for simpler questions.
- *Few-shot setting effect:* Brown et al. [35] indicates that LLMs are few-shot learners, while Kojima et al. [37] points out that current advanced LLMs are zero-shot reasoners. Therefore, it is understandable that the few-shot setting may be ineffective on easier tasks like Basic Sports Understanding and on some models. However, in more complex tasks, Advanced Scenario Analysis, more sophisticated and powerful models, such as GPT-4o and Claude Sonnet, can still learn sports understanding-based reasoning patterns from few-shot prompts, thereby achieving better performance.
- *Multi-Agent LLM voting stabilized performance:* As shown in Table 5, the collaborative approach achieved performance comparable to the top-performing single model, while significantly outperforming lower-ranked models. Although it did not surpass the best individual result, the results of the voting committee are clearly superior to those of the lower-performing models, which indicates the voting committee demonstrated more robust outcomes than individual LLMs.

Table 5. Performance (accuracy \uparrow) of LLMs committee for the *Advanced Scenario Analysis* task. 0S stands for zero-shot and 5S stands for 5-shot.

Condition	Easy Single-Hop	Hard Single-Hop	Easy Multi-Hop	Hard Multi-Hop
0S	73.02	67.89	41.22	25.74
0S, CoT	71.83	66.26	40.82	28.69
5S	73.41	67.07	43.67	25.74
5S, CoT	73.02	68.29	45.71	26.58

4.2.2. Analysis on Multi-Hop Questions in Advanced Scenarios

To further illustrate how LLMs stumble in sports tasks for advanced scenario analysis, we break down models' performance on the multi-hop questions and summarize some interesting phenomena below. We defer the detailed performance reports on multi-hop questions and case study examples to Tables A1 and A2 in the Appendix, respectively.

Note that we adhere to strict criteria in our evaluation, answers to a multiple-select choice question are considered accurate only when all correct options are chosen and answers to a multi-hop question are considered accurate only when the main and all the subquestions have been correctly answered.

- *Lacking contextual understanding capability*: All evaluated models show higher accuracy in answering main questions compared to answering sub-questions, while a sub-question often require reasoning within the context provided by the main question. This highlights a critical need for LLMs to effectively manage longer prompts and complex contexts, especially in the sports understanding field where nuanced relationships between context and query are crucial.
- *Prompt length and order of questions matter*: Most LLMs perform better on sub-questions posed at the beginning of a multi-hop prompt sequence compared to those at the end, except for the Gemni-family models which show an opposite or fluctuating trend. This suggests that the prompt length and sequence order of sub-questions significantly impact the reasoning ability of LLMs.
- *Unstable Reasoning*: Even when all sub-questions are answered correctly, the main question still has an error rate ranging from 24% to 44%. This indicates that current LLMs lack a solid understanding of sports knowledge and have unstable reasoning abilities in the sports domain.

5. Error Analysis

To investigate the performance limitation of LLMs and VLMs, we conducted a detailed analysis across all three tasks. For each task, we sampled incorrectly answered questions and their responses from all models. We sampled 30 incorrectly answered questions for Advanced Scenario Analysis and 40 incorrectly answered for Video-based Sports Recognition. Each response was categorized into a distinct error type and further analyzed by human judges. A total of four researchers participated in the error categorization.

5.1. Error Types

For the *text-based* tasks that include both Basic Sports Understanding and Advanced Scenario Analysis, we identified four primary error types:

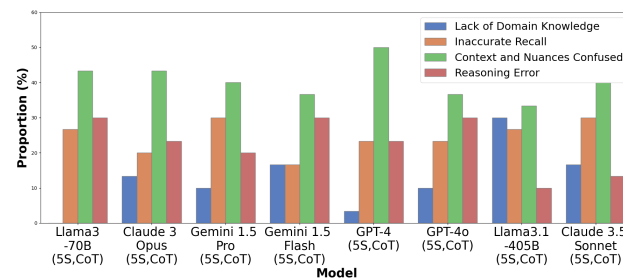
- A. *Lack of Domain Knowledge* includes misunderstandings of rules, concepts, or terms, and failing to understand specific tactics.
- B. *Inaccurate Recall* indicate errors where models incorrectly remember facts or details.
- C. *Context and Nuances Confused* are misinterpretations or oversimplifications of complex scenarios.
- D. *Reasoning Error* reflects failures in logical processing or connecting relevant pieces of information correctly.

For the *Video-based* Sports Recognition, we identified three primary types of errors:

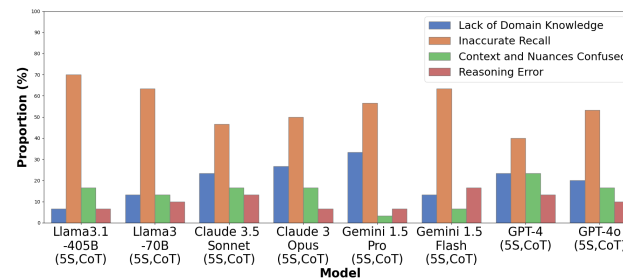
- A. *Number Recognition Error* encompasses errors such as incorrectly recognizing the number of people, the number of actions, the number of action types, or similar recognition mistakes.
- B. *Action Recognition Error* where actions recognized by VLMs are not relevant to the correct answer.
- C. *Lack of domain knowledge* relates to not understanding specific actions depicted in the video.

5.2. Case Study and Error Distribution Analysis

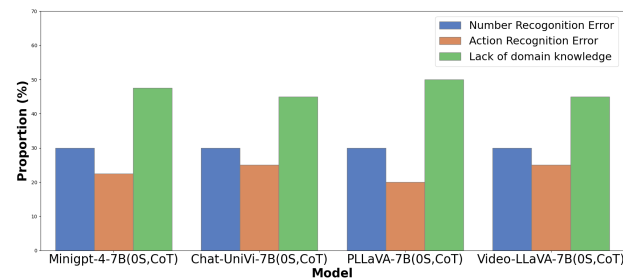
Figure 2a demonstrated the error distribution in the Advanced Scenario Analysis task, where errors of *Context and Nuances Confused* dominated. This disparity highlights the need for LLMs to improve their ability to discern subtle context differences and reason from a sports expert’s perspective. For instance, in tennis, while LLMs often discourage net play due to its high risks, professionals frequently use it as an aggressive yet effective strategy. This suggests that LLMs struggle to accurately assess scenarios in the context of professional-level sports and fail to think from the athlete’s perspective. For future efforts, we believe training data should focus on more granular content, such as sports commentary.



(a) Advanced Scenario Analysis by Human Judges.



(b) Advanced Scenario Analysis by GPT-4.



(c) Video-based Sports Recognition by Human Judges.

Figure 2. Type of Error Distribution.

Incongruence of model-based error analysis: We s attempted to use GPT-4 for more scalable error analysis by incorporating the error category descriptions and case studies into the prompt for few-shot learning. Comparing results in Figure 2a with Figure 2b, we observed that GPT-4 tends to inaccurately categorize errors into the type of *Inaccurate Recall*, which highlights the limitations of model-based error analysis for sports understanding tasks and underscores the necessity of human-centric error analysis at the current stage.

Video encoder limits the multimodal sports understanding: Figure 2c illustrates the error distribution in the video-based sports recognition task for VLMs, with the primary error type being *Lack of Domain Knowledge*. However, our analysis of text-based tasks

indicates that the underlying LLMs do possess domain knowledge. We believe this issue stems from the video encoder not being optimized for the sports domain, leading to difficulties in capturing and understanding sports-related actions. To mitigate this, the video encoder should be tailored for the sports domain. Training video encoder with videos featuring higher frame rates or slow motion is a potential solution to enhance the recognition of complex sports movements.

6. Conclusions and Future Work

We benchmarked the performance of current state-of-the-art LLMs and VLMs in the sports domain using a consolidated dataset that includes both text-based and video-based QA, categorized by varying complexity levels. This benchmark offers valuable reference points for researchers and practitioners in the field. Our assessment revealed key limitations in current foundation models, especially in handling complex sports reasoning and multi-hop question answering. Our assessment also shows that CoT reasoning and multi-agent voting can potentially improve performance in the domain of sports understanding. Our assessment in Advanced Scenario Analysis indicates that there is still significant room for model performance improvement, highlighting the importance of domain-specific training.

Additionally, we provided in-depth error analysis and identified critical reasoning flaws not captured by model-based evaluations, providing insights to refine LLMs and VLMs for sports applications.

Moving forward, we plan to explore multimodal fine-tuning methods for VLMs on the sports understanding domain, by further leveraging the sports dataset overview in this paper. Inspired by our error analysis, we aim to prioritize interpretability in model decision-making, with a focus on improving cross-modality alignment. Our goal is to contribute to more reliable and transparent automated systems in sports officiating empowered by VLMs.

7. Limitations

Financial and computational resource constraints limited our ability to evaluate certain high-capacity models, such as PLLaVA-34B [24]. In future work, we aim to broaden the spectrum of models we assess and incorporate models with larger parameter scales to ensure a more comprehensive analysis.

Author Contributions: Methodology, Z.Y. and H.X.; Software, Z.Y., J.L. and Z.C.; Formal analysis, H.X.; Writing—original draft, Z.Y.; Writing—review & editing, Z.Z. and W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data is included in the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This section provides the prompt we used in our experiments. Our prompts are inspired by LangChain.

Appendix A.1. Prompt for Basic Sports Understanding and Advanced Scenario Analysis

In this part, {question} refers to the question in the dataset, {five-shot prompt} refers to 5-shot prompts in Section 4.1.

Zero-shot (0S):

You are an assistant for question-answering tasks. If you don't know the answer, just say that you don't know. Please indicate the correct answer(s) clearly with letter(s).

Question: {question}

Answer:

Zero-shot CoT (0S, CoT):

You are an assistant for question-answering tasks. Answer the question(s) step by step. If you don't know the answer, just say that you don't know. Please indicate the correct answer(s) clearly with letter(s).

Question: {question}

Answer:

5-shot (5S): You are an assistant for question-answering tasks. If you don't know the answer, just say that you don't know. Please indicate the correct answer(s) clearly with letter(s). Some examples are given below.

{five-shot prompt}

Question: {question}

Answer:

5-shot CoT (5S, CoT):

You are an assistant for question-answering tasks. Answer the question(s) step by step. If you don't know the answer, just say that you don't know. Please indicate the correct answer(s) clearly with letter(s). Some examples are given below.

{five-shot prompt}

Question: {question}

Answer:

Appendix A.2. Prompt for Video-Based Sports Understanding and Advanced Scenario Analysis

In this part, {question} refers to the question in the dataset, {answer} is the ground truth answer, {pred} is the answer provided by the model.

System prompt for zero-shot (0S) and zero-shot CoT (0S,CoT):

You are an intelligent chatbot designed for evaluating the correctness of generative outputs for question-answer pairs. Your task is to compare the predicted answer with the correct answer and determine if they match meaningfully. Here's how you can accomplish the task:—##INSTRUCTIONS: - Focus on the meaningful match between the predicted answer and the correct answer.

- Consider synonyms or paraphrases as valid matches.
- Evaluate the correctness of the prediction compared to the answer.

Prompt for zero-shot (0S):

{question}

Prompt for zero-shot CoT (0S,CoT):

Let's think step by step. {question}

GPT-4 judge and score:

Please evaluate the following video-based question-answer pair:

Question: {question}

Correct Answer: {answer}

Predicted Answer: {pred}

Provide your evaluation only as a yes/no and score.

The score is an integer value between 0 and 5, with 5 indicating the highest meaningful match. Respond in JSON format like this: {"pred": "yes", "score": 4.8}.

Appendix B. Multi-Hop Question Analysis and Case Study

Table A1. Multi-hop question analysis.

Model	easy_multi	easy_multi_main	easy_multi_sub	hard_multi	hard_multi_main	hard_multi_sub	P(main=0 sub=1)	P(sub1=1)	P(sub2=1)	P(sub3=1)
Llama3-70B (0S)	35.92	57.55	46.53	21.10	42.62	32.91	37.18	54.24	52.27	40.00
Llama3-70B (0S,CoT)	37.14	55.51	48.57	21.10	41.77	33.33	37.97	54.66	54.09	42.86
Llama3-70B (5S)	26.94	54.69	35.92	16.46	39.24	27.43	41.54	48.31	46.82	40.00
Llama3-70B (5S,CoT)	26.12	53.88	34.29	17.30	40.93	22.78	25.93	44.07	42.27	42.86
llama3.1-405B (0S)	39.18	53.47	48.57	21.94	37.13	34.18	37.04	49.15	48.64	28.57
llama3.1-405B (0S,CoT)	32.24	49.8	37.96	22.36	39.66	32.07	31.58	46.61	46.82	48.57
llama3.1-405B (5S)	40.82	59.18	52.24	27.85	43.88	41.77	34.34	57.2	53.64	48.57
llama3.1-405B (5S,CoT)	42.45	58.37	55.51	28.27	45.15	39.66	29.79	54.24	54.09	45.71
Claude 3 Opus (0S)	40.41	61.22	51.43	27.00	50.21	40.51	34.38	63.98	58.64	42.86
Claude 3 Opus (0S,CoT)	42.86	64.08	53.88	28.27	51.05	41.35	32.65	63.14	60.91	54.29
Claude 3 Opus (5S)	40.00	60.41	48.57	26.16	47.68	35.86	27.06	53.81	50.91	37.14
Claude 3 Opus (5S,CoT)	42.86	63.27	52.24	29.11	50.21	42.62	32.67	64.41	60.45	51.43
Claude 3.5 Sonnet (0S)	45.31	64.9	56.73	26.58	44.3	38.82	32.61	55.51	58.18	45.71
Claude 3.5 Sonnet (0S,CoT)	47.76	64.9	57.55	30.8	46.84	43.04	29.41	55.93	64.55	42.86
Claude 3.5 Sonnet (5S)	44.9	65.71	55.92	29.96	50.21	42.62	30.69	60.59	59.55	48.57
Claude 3.5 Sonnet (5S,CoT)	45.71	66.53	57.14	32.07	53.59	41.77	24.24	62.29	58.64	51.43
Gemini 1.5 Pro (0S)	33.47	45.71	37.96	19.83	32.07	29.96	40.85	44.07	47.73	51.43
Gemini 1.5 Pro (0S,CoT)	33.06	46.12	36.33	20.25	33.33	29.11	37.68	44.49	49.09	57.14
Gemini 1.5 Pro (5S)	40.82	64.49	46.94	24.47	46.41	35.86	32.94	55.93	59.09	45.71
Gemini 1.5 Pro (5S,CoT)	39.18	64.90	47.35	21.94	45.99	34.60	37.80	56.36	56.36	57.14
Gemini 1.5 Flash (0S)	35.92	52.65	46.12	21.94	41.77	33.76	37.50	52.54	55.91	45.71
Gemini 1.5 Flash (0S,CoT)	36.33	49.80	44.49	21.94	34.60	32.07	36.84	47.88	47.73	48.57
Gemini 1.5 Flash (5S)	38.78	62.86	49.80	22.36	45.99	34.18	34.57	55.51	54.55	45.71
Gemini 1.5 Flash (5S,CoT)	35.51	56.73	46.94	19.83	40.08	31.65	38.67	52.12	52.73	51.43
GPT-4 (0S)	38.78	55.92	50.20	26.58	41.35	37.97	34.44	57.63	57.73	40.00
GPT-4 (0S,CoT)	42.04	62.45	51.84	31.65	49.79	46.41	32.73	63.56	63.18	51.43
GPT-4 (5S)	38.37	65.31	51.02	28.69	53.59	41.77	32.32	61.86	59.09	45.71
GPT-4 (5S,CoT)	44.49	65.31	53.88	27.43	48.95	39.66	32.98	58.05	56.36	45.71
GPT-4o (0S)	38.78	53.47	45.31	22.78	38.40	35.44	44.05	51.69	50.91	48.57
GPT-4o (0S,CoT)	38.37	63.27	51.02	27.00	47.26	43.46	38.83	65.68	65.91	51.43
GPT-4o (5S)	39.59	62.86	46.53	27.43	51.48	37.97	33.33	60.17	58.64	48.57
GPT-4o (5S,CoT)	42.45	56.73	46.12	29.11	45.99	40.51	31.25	60.59	61.36	48.57

Table A2. More examples of case study on model response errors.

Question: During the final minutes of a one-point off football game, how should the team with the ball, which is one-point off, best decide their strategy?

(A) Attempt a risky long pass to potentially score quickly. (B) Continue with their standard gameplay to maintain possession (C) Play defensively to prevent the opponent from scoring. (D) Try to slow down the game. **Answer:** A

Model Output: The correct answer is B. The team with the ball should continue with their standard gameplay to maintain possession and prevent the opponent from scoring. This will give them the best chance of winning the game. **Explanation:** In a one-point off football game, where both teams are tied at 0-1, the team with the ball has two main goals: to score and to prevent the opponent from scoring. Attempting a risky long pass may result in a turnover, which could lead to the opponent scoring and winning the game. Playing defensively may also be ineffective as it allows the opponent more time to score. Therefore, continuing with their standard gameplay to maintain possession and prevent the opponent from scoring is the most effective strategy for this situation.

Error Classification: "0-1 tie" is a clear logical error, and the model's tactical choice is based on this erroneous premise. Therefore, this error should be classified as a Reasoning Error.

Question: What is the main purpose of a double block strategy in volleyball defensive schemes?

A. To cover large areas of the court B. To focus on one particular player C. To increase the chance of blocking a spike D. To allow the libero to serve **Answer:** AC

Model Output: The main purpose of a double block strategy in volleyball defensive schemes is to increase the chance of blocking a spike. By having two players jump and attempt to block the ball at the net, the team can create a more formidable barrier against the opposing hitter, making it more difficult for the spike to get through. Hence, the correct answer is: C. To increase the chance of blocking a spike.

Error Classification: Though the model considers double block strategy can create a more formidable barrier (covering larger areas), it lacks the knowledge that the primary purpose of a double block is to cover large areas of the court. Therefore, this error should be classified as a Lack of Domain Knowledge.

Table A2. Cont.

Question: Why would a tennis player, during a game, choose to consistently serve to their opponent's backhand, employ a topspin-heavy style of play, and opt for a more defensive baseline position?

(A) The player has identified their opponent's backhand as a weaker shot, the topspin can generate higher bounces making it harder for the opponent to hit aggressively, and the defensive baseline position allows the player to cover the court more effectively. (B) The player is trying to tire out their opponent quickly by making them run across the court more frequently. (C) The player is not confident in their own forehand and is trying to avoid it. (D) The player's coach has instructed them to stick to this game plan regardless of the opponent's strengths and weaknesses. **Answer:** AB

Model Output: (A) The player has identified their opponent's backhand as a weaker shot, the topspin can generate higher bounces making it harder for the opponent to hit aggressively, and the defensive baseline position allows the player to cover the court more effectively.

This is the correct answer. In tennis, players often adapt their game strategy based on their opponent's strengths and weaknesses. If a player identifies their opponent's backhand as a weaker shot, consistently serving to that side can put the opponent at a disadvantage. Using topspin-heavy shots can create higher bounces, making it harder for the opponent to hit aggressively. Additionally, adopting a more defensive baseline position allows the player to cover the court more effectively and return shots from a variety of angles.

Error Classification: The model incorrectly recalls that tiring out their opponent quickly by making them run across the court more frequently is not a primary reason. Therefore, it is an example of Inaccurate Recall.

Question: Question: How might a tennis player effectively use tactics to gain an advantage in a match?

(A) By consistently serving to the opponent's weak side (B) By hitting deep balls to limit the opponent's shot options (C) By always aiming for the lines to score direct points (D) By varying the pace and spin of the ball to disrupt the opponent's rhythm **Answer:** ABCD

Model Output: The correct answers are A, B, and D.

Here's why:

(A) Serving to the opponent's weak side can be an effective tactic, as it can make it more difficult for them to return the serve and create opportunities for the server to take control of the point.

(B) Hitting deep balls can limit the opponent's shot options and make it harder for them to take control of the point. This can be especially effective if the opponent is not comfortable hitting from the back of the court.

(C) Aiming for the lines to score direct points is not a reliable tactic, as it's a high-risk strategy that can result in errors and unforced mistakes.

(D) Varying the pace and spin of the ball can disrupt the opponent's rhythm and make it harder for them to anticipate and react to the shot. This can be an effective way to create uncertainty and gain an advantage.

Overall, a tennis player can effectively use tactics to gain an advantage in a match by exploiting their opponent's weaknesses, controlling the pace and direction of the ball, and disrupting their opponent's rhythm and timing.

Error Classification: Here are the reasons why it is classified as Context and Nuances Confused:

1. Misinterpretation of effectiveness: The model labeled aiming for the lines as a "high-risk strategy" that could lead to errors and unforced mistakes, suggesting a misunderstanding of the context in which this tactic can be effective. While it is true that aiming for the lines is riskier, it is also a common and effective tactic used in tennis to win points outright.

2. Overemphasis on negative: The model focused solely on the negative aspects of aiming for the lines, failing to recognize or acknowledge that this tactic, when used judiciously, can be a significant advantage in a match. This shows a nuanced misunderstanding of when and how this tactic should be applied.

3. Ignoring professional usage in real world: At higher levels of play, aiming for the lines is a tactic frequently used to apply pressure and secure points, which is crucial for understanding the full context of this tactic's effectiveness.

By not appreciating these nuances, the model failed to see the full strategic value of aiming for the lines, leading to the exclusion of this option from their correct answers. This indicates a confusion of context and nuances surrounding the application of tennis tactics.

Table A2. Cont.

Question: Main Question: How does the number of players in a basketball team and the number of substitutes impact the overall game strategy, considering that a standard basketball team comprises 5 players and a professional team can have up to 13 players (including substitutes)?

A. The number of players and substitutes has no impact on the game strategy. B. Having more substitutes allows a team to maintain an aggressive playing style without worrying about player fatigue. C. The number of players in a team determines the number of substitutes. D. A team with more players has a disadvantage as it is harder to coordinate.

Sub-Question 1: What is the standard number of players in a basketball team during a game?

A. 10 B. 13 C. 5 D. 7

Sub-Question 2: Considering a professional basketball team, how many substitutes can a team have?

A. 5 B. 8 C. 10 D. 13 **Answer:** Main Question: B; Sub-Question 1: B; Sub-Question 2: B

Model Output: Main Question: The correct answer is B. Having more substitutes allows a team to maintain an aggressive playing style without worrying about player fatigue.

Having a larger roster with more substitutes provides a team with several strategic advantages. It allows them to:

Maintain a high level of intensity and aggression throughout the game, without worrying about player fatigue. Make tactical substitutions to counter the opponent's strategy or to take advantage of mismatches. Provide rest for key players, allowing them to recover and perform at a higher level. Experiment with different lineups and rotations to find the most effective combinations.

Sub-Question 1: The correct answer is C. 5.

In a standard basketball game, each team has 5 players on the court at a time.

Sub-Question 2: The correct answer is D. 13.

In professional basketball, a team can have up to 13 players on its active roster, including substitutes. This allows teams to have a deep bench and make strategic substitutions throughout the game.

Error Classification: For Sub-Question 1, the model incorrectly recalled the number of people in a professional basketball team can have. The correct number is 13, but the model answered 5, which is the number of starting players. For Sub-Question 2, the model incorrectly recalled the number of substitutes a professional basketball team can have. The correct number is 8, but the model answered 13, which is actually the total number of players on a professional team, including the starting five. Therefore, the type of error should be Inaccurate Recall.

Question: Question: How can a player execute a successful attack hit in volleyball?

(A) By hitting the ball while it's high above the net. (B) By hitting the ball with force and accuracy towards the opponent's court. (C) By targeting the blockers' hands and hitting the ball out of bounds. (D) By hitting the ball over the net, aiming for the opponent's weak spots.

Model Output: A successful attack hit in volleyball involves a combination of power, accuracy, and strategic placement. Here's a breakdown of the options:

(A) Hitting the ball while it's high above the net can be part of a successful attack, as it allows the hitter to have a better angle and more options for placement. However, this alone does not guarantee success.

(B) Hitting the ball with force and accuracy towards the opponent's court is a key component of a successful attack. Force can make it difficult for the opponent to defend, and accuracy ensures the ball lands in bounds and in a challenging spot for the defense.

(C) Targeting the blockers' hands and hitting the ball out of bounds can be a strategic move known as a "tool" or "wipe". This can be effective if done correctly, as it results in a point for the attacking team.

(D) Hitting the ball over the net and aiming for the opponent's weak spots is another crucial aspect of a successful attack. Identifying and targeting areas where the opponent is vulnerable can increase the chances of scoring.

Given these considerations, the most comprehensive answer is:

(B) By hitting the ball with force and accuracy towards the opponent's court. (D) By hitting the ball over the net, aiming for the opponent's weak spots.

Both B and D together encompass the essential elements of a successful attack hit in volleyball.

Error Classification: The model focused primarily on options B and D, emphasizing force, accuracy, and targeting weak spots. While these are indeed components of a successful attack, the model did not fully acknowledge the importance of the other options (A and C) which are also part of the correct answer. This shows a misunderstanding of the full context of what constitutes a successful attack hit in volleyball. Therefore, this error is classified as Context and Nuances Confused.

References

1. Araújo, D.; Couceiro, M.; Seifert, L.; Sarmiento, H.; Davids, K. *Artificial Intelligence in Sport Performance Analysis*; Routledge: London, UK, 2021.
2. ATP Tour, Inc. *The 2023 ATP Official Rulebook*; ATP Tour, Inc.: London, UK, 2023.
3. Electronic Line Calling Live To Be Adopted Across The ATP Tour. *ATP Tour*, 28 April 2023.
4. Spitz, J.; Wagemans, J.; Memmert, D.; Williams, A.M.; Helsen, W.F. Video assistant referees (VAR): The impact of technology on decision making in association football referees. *J. Sports Sci.* **2021**, *39*, 147–153. [[CrossRef](#)]
5. Tamir, I.; Bar-Eli, M. The moral gatekeeper: Soccer and technology, the case of video assistant referee (VAR). *Front. Psychol.* **2020**, *11*, 613469. [[CrossRef](#)] [[PubMed](#)]
6. Guest, N.S.; Horne, J.; Vanderhout, S.M.; El-Sohemy, A. Sport nutrigenomics: Personalized nutrition for athletic performance. *Front. Nutr.* **2019**, *6*, 8. [[CrossRef](#)] [[PubMed](#)]
7. Bonilla, D.A.; Boullousa, D.; Del Coso, J. Advances in nutrition, dietary supplements and ergogenic aids for athletic performance: Trends and future prospects. *Nutrients* **2023**, *15*, 2246. [[CrossRef](#)] [[PubMed](#)]
8. Zhang, X.; Feng, S.; Peng, R.; Li, H. The role of velocity-based training (VBT) in enhancing athletic performance in trained individuals: A meta-analysis of controlled trials. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9252. [[CrossRef](#)] [[PubMed](#)]
9. Haugen, T.; Seiler, S.; Sandbakk, Ø.; Tønnessen, E. The training and development of elite sprint performance: An integration of scientific and best practice literature. *Sports Med. Open* **2019**, *5*, 44. [[CrossRef](#)]
10. Xia, H.; Tracy, R.; Zhao, Y.; Wang, Y.; Wang, Y.F.; Shen, W. Advanced Volleyball Stats for All Levels: Automatic Setting Tactic Detection and Classification with a Single Camera. In Proceedings of the 2023 IEEE International Conference on Data Mining Workshops (ICDMW), Shanghai, China, 1–4 December 2023; pp. 1407–1416. [[CrossRef](#)]
11. BIG-Bench Authors. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Trans. Mach. Learn. Res.* **2023**, 1–95. [[CrossRef](#)]
12. Xia, H.; Yang, Z.; Wang, Y.; Tracy, R.; Zhao, Y.; Huang, D.; Chen, Z.; Zhu, Y.; Fang Wang, Y.; Shen, W. SportQA: A Benchmark for Sports Understanding in Large Language Models. *arXiv* **2024**, arXiv:2402.15862.
13. Li, H.; Deng, A.; Ke, Q.; Liu, J.; Rahmani, H.; Guo, Y.; Schiele, B.; Chen, C. Sports-QA: A Large-Scale Video Question Answering Benchmark for Complex and Professional Sports. *arXiv* **2024**, arXiv:2401.01505.
14. AI@Meta. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
15. AI@Meta. Llama 3 Model Card. Available online: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (accessed on 18 April 2024).
16. OpenAI. GPT-4 Technical Report. *arXiv* **2024**, arXiv:2303.08774.
17. OpenAI. Hello GPT-4o. Available online: <https://openai.com/index/hello-gpt-4o/> (accessed on 13 May 2024).
18. Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* **2024**, arXiv:2403.05530.
19. Gemini Team. Gemini Flash. 2024. Available online: <https://deepmind.google/technologies/gemini/flash/> (accessed on 18 December 2024).
20. Anthropic. Claude 3.5 Sonnet Model Card Addendum. 2024. Available online: https://www-cdn.anthropic.com/fed9cc193a14b84131812372d8d5857f8f304c52/Model_Card_Claude_3_Addendum.pdf (accessed on 18 December 2024).
21. Anthropic. Introducing the Next Generation of Claude. Available online: <https://www.anthropic.com/news/claude-3-family> (accessed on 4 March 2024).
22. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv* **2023**, arXiv:2304.10592.
23. Jin, P.; Takanobu, R.; Zhang, C.; Cao, X.; Yuan, L. Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding. *arXiv* **2023**, arXiv:2311.08046.
24. Xu, L.; Zhao, Y.; Zhou, D.; Lin, Z.; Ng, S.K.; Feng, J. PLLaVA: Parameter-free LLaVA Extension from Images to Videos for Video Dense Captioning. *arXiv* **2024**, arXiv:2404.16994.
25. Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv* **2023**, arXiv:2311.10122.
26. Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; Wang, H.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; et al. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv* **2023**, arXiv:2310.01852.
27. Cao, S.; Shi, J.; Pan, L.; Nie, L.; Xiang, Y.; Hou, L.; Li, J.; He, B.; Zhang, H. KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 22–27 May 2022; Muresan, S., Nakov, P., Villavicencio, A., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2022; Volume 1: Long Papers; pp. 6101–6119. [[CrossRef](#)]

28. Clark, C.; Lee, K.; Chang, M.W.; Kwiatkowski, T.; Collins, M.; Toutanova, K. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 (Long and Short Papers); pp. 2924–2936. [[CrossRef](#)]
29. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; Brussels, Belgium, 31 October–4 November 2018; Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 2369–2380. [[CrossRef](#)]
30. Dhingra, B.; Mazaitis, K.; Cohen, W.W. Quasar: Datasets for Question Answering by Search and Reading. *arXiv* **2017**, arXiv:1707.03904.
31. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017; Barzilay, R., Kan, M.Y., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; Volume 1: Long Papers; pp. 1601–1611. [[CrossRef](#)]
32. Jardim, P.C.; Moraes, L.M.P.; Aguiar, C.D. QASports: A Question Answering Dataset about Sports. In Proceedings of the Brazilian Symposium on Databases: Dataset Showcase Workshop, Belo Horizonte, MG, Brazil, 25–29 September 2023; pp. 1–12.
33. Liu, Q.; Jiang, S.; Wang, Y.; Li, S. LiveQA: A Question Answering Dataset over Sports Live. In Proceedings of the 19th Chinese National Conference on Computational Linguistics, Haikou, China, 30 October–1 November 2020; Sun, M., Li, S., Zhang, Y., Liu, Y., Eds.; Springer Nature: Cham, Switzerland, 2020; pp. 1057–1067.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Burstein, J., Doran, C., Solorio, T., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1 (Long and Short Papers); pp. 4171–4186. [[CrossRef](#)]
35. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
36. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
37. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large language models are zero-shot reasoners. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 22199–22213.
38. Gemini Team. Gemini: A Family of Highly Capable Multimodal Models. *arXiv* **2024**, arXiv:2312.11805.
39. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. ImageBind One Embedding Space to Bind Them All. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 15180–15190.
40. Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model. *arXiv* **2023**, arXiv:2304.15010.
41. Zhang, R.; Han, J.; Liu, C.; Gao, P.; Zhou, A.; Hu, X.; Yan, S.; Lu, P.; Li, H.; Qiao, Y. LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention. *arXiv* **2023**, arXiv:2303.16199.
42. Han, J.; Zhang, R.; Shao, W.; Gao, P.; Xu, P.; Xiao, H.; Zhang, K.; Liu, C.; Wen, S.; Guo, Z.; et al. ImageBind-LLM: Multi-modality Instruction Tuning. *arXiv* **2023**, arXiv:2309.03905.
43. Li, Y.; Chen, L.; He, R.; Wang, Z.; Wu, G.; Wang, L. MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions. *arXiv* **2021**, arXiv:2105.07404.
44. Shao, D.; Zhao, Y.; Dai, B.; Lin, D. FineGym: A Hierarchical Video Dataset for Fine-grained Action Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020.
45. Chan, C.M.; Chen, W.; Su, Y.; Yu, J.; Xue, W.; Zhang, S.; Fu, J.; Liu, Z. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. *arXiv* **2023**, arXiv:2308.07201.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.