SportsIQ: Tool-Augmented Reasoning for Sports Tactical Understanding

Cady He
Carnegie Mellon University
cadyh@andrew.cmu.edu

Siyani Vengatagiri
Carnegie Mellon University
svengata@andrew.cmu.edu

Yash Lothe
Carnegie Mellon University
ylothe@andrew.cmu.edu

Abstract

Complex sports scenario reasoning demands simultaneous mastery of precise rules, tactical strategy, and live game situational analysis, a challenge that exposes fundamental limits in large language models. Error analysis on the SportQA benchmark (Xia et al., 2024) identifies three distinct failure modes at the expert Level-3 tier: knowledge gaps (55%), contextual misreading (30%), and reasoning chain errors (15%). We hypothesize that the first two are addressable by domain-targeted retrieval, and the third by iterative agentic reasoning. We present two systems built on three purpose-built corpora (rules, tactics, game-situation analysis) indexed with hybrid dense-sparse retrieval and cross-encoder reranking: SportsRAG, a fixed decompose-then-retrieve pipeline, and SportsIQ, a ReAct agent with sequential dynamic tool selection. Hop-aware SportsRAG beats the no-RAG GPT-5.4 baseline (64.53% vs. 64.10% EM), confirming that targeted retrieval helps. A confidence-gated, hop-type ensemble combining the two approaches achieves 67.09% EM (+2.99pp over baseline), with each layer addressing a distinct class of expert-reasoning failure. These findings generalize to any rule-governed domain where LLM knowledge gaps are systematic and domain-partitioned.

1 Introduction

Sports has long served as a fruitful domain for NLP research, with substantial prior work in game summarization, commentary generation, and sports analytics. With the advent of large language models, the theoretical scope of sports NLP has expanded considerably. However, existing applications have remained largely surface-level: generating news articles, retrieving historical facts, and customizing commentary style. The SportQA benchmark (Xia et al., NAACL 2024) provides the most formidable evaluation of LLM capabilities in sports reasoning to date. On Level-3 expert scenario questions, defined as complex multi-select scenarios requiring simultaneous rule interpretation, tactical judgment, and game-situation reasoning, GPT-4 achieves a 47% exact match with the ground-truth answers while human experts score 93%, a gap of nearly 45 percentage points. Our own reproduction with GPT-5.4, a substantially stronger model, reaches only 64%, confirming that the gap persists even with state-of-the-art models. Xia et al.'s error analysis reveals that this gap is systematic and can be decomposed into three main error groups: knowledge gaps account for 55% of errors, contextual misreading for 30%, and reasoning chain errors for 15%. These failures are not arbitrary; they point to specific, targetable weaknesses in knowledge retrieval and multi-step reasoning that cannot be addressed by standard prompting strategies.

Prior research has not fully closed these gaps in the sports domain. The original SportQA paper applies no retrieval augmentation or reasoning angle. ReAct (Yao et al., ICLR 2023) applies agentic reasoning with a single undifferentiated Wikipedia index. IRCoT (Trivedi et al., ACL 2023) successfully interleaves retrieval with chain-of-thought reasoning but without the integration of any domain-specialized tools. RAG² (Sohn et al., NAACL 2025) demonstrates that domain-partitioned retrieval produces notable accuracy gains on expert medical QA, but no equivalent work exists for sports. Standard chain-of-thought reasoning alone is also insufficient for this task: it relies entirely on the model's parametric knowledge with no mechanism to verify or steer beliefs mid-reasoning, leaving it especially vulnerable to hallucination and error propagation on multi-step questions. No prior work combines domain-partitioned multi-corpus retrieval with iterative agentic reasoning for sports expert question answering.

We address these gaps with two complementary systems built on three independent purpose-built corpora covering rules, tactics, and game-situation analysis, all indexed with hybrid dense-sparse re-

trieval and cross-encoder reranking. SportsRAG is a fixed decompose-then-retrieve pipeline that routes each question into typed sub-queries against each of the three corpora. SportsIQ is a ReAct agent with sequential dynamic tool selection, allowing the model to reason iteratively and decide which corpus to query at each step. Building further, we apply confidence gating, self-verification, and disagreement-based routing to produce a hop-type ensemble that combines the strongest single-hop and multi-hop strategies.

Each layer produces incremental gains over vanilla prompting, with our hop-type ensemble reaching 67.09% exact match, a gain of 2.99 percentage points over our no-RAG GPT-5.4 baseline, and a gain of 19.95 percentage points over the GPT-4 results reported in the original SportQA paper. Three independent findings stand out. First, domain-partitioned retrieval outperforms single-index approaches on the whole because different question types draw on fundamentally different knowledge sources. Second, retrieval quality is not the bottleneck; knowing when not to retrieve matters as much as retrieval precision. Third, retrieval and reasoning interventions are consistently asymmetric across question types: techniques that improve multi-hop performance consistently hurt single-hop performance and vice versa, making hop-type-aware routing a necessary design principle rather than an optional optimization. More broadly, these findings suggest that structured knowledge access is a feasible path toward expert-level performance in other rule-governed domains.

2 Related Work

There has been a wide range of research done to improve language models' ability to comprehend complex real world scenarios by enabling reasoning and reducing hallucinations. To enable reasoning and factuality, researchers have explored techniques and paradigms such as chain-of-thought (CoT) prompting (Wei et al., 2023), verification (Cobbe et al., 2021), self reflection (Madaan et al., 2023), reinforcement learning with human feedback (RLHF) (Ziegler et al., 2020), and specialized fine-tuning (Rajani et al., 2019; Zelikman et al., 2022). To further reduce hallucinations and errors, approaches such as external retrieval through tools, retrieval augmented generation (RAG) (Lewis et al., 2021), and CoT with self-consistency (Wang et al., 2023) have been introduced where significant performance boost was observed compared to the benchmark language models.

2.1 Sports Understanding in NLP

Sports NLP is an emerging field with many research directions. A majority of research in sports NLP looks into sports analytics, but there's a lack of research on improving deep comprehension of sports using LLMs, especially with nuanced game scenarios and rules. In 2024, Xia et al. released a comprehensive sports QA dataset containing 3 levels, with level 3 testing complex sports scenario analysis (Xia et al., 2024). GPT-4 was found to consistently outperform other language models at the time across all three levels, but shows a significant drop in performance on level 3 questions, especially multi-hop, falling roughly 45 percentage points below human expert performance. Through error analysis, the researchers pointed out several areas where significant work needs to be done, including the correct use of known information, accurate recalling of facts, precise contextual understanding and interpretation, and effective logical reasoning. Overall, LLM reasoning limitations in sports can be categorized into two broad areas: poor reasoning for complex scenarios and inaccurate factual recall due to knowledge gaps, which our work directly targets.

2.2 Reasoning Strategies and Subsequent Advances

Chain of thought (CoT) was first introduced by Wei et al. in 2023 as a prompting technique that asks language models to explicitly generate intermediate reasoning steps during response generation, significantly improving performance on complex reasoning tasks (Wei et al., 2023). Wang et al. presented that self-consistency improves CoT by proposing a "sample and marginalize" decoding approach that generates a variety of reasoning paths and determines the final answer by marginalizing out the sampled paths to arrive at the most consistent answer (Wang et al., 2023).

2.3 Knowledge Retrieval

Beyond CoT reasoning, Yao et al. equipped language models with capabilities to acquire external knowledge through a paradigm called ReAct (Yao et al., 2022). By combining CoT and tool calling, the model strategically interleaves verbal reasoning traces with action plans, allowing it to create

and adjust plans and leverage the external environment to inform its decision making. ReAct is described as easily generalizable across distinct action spaces and reasoning needs, and we adapt similar thought patterns to our sports QA setting. Trivedi et al. built on this by proposing IRCoT, a RAG-style method that interleaves CoT with retrieval to improve multi-step reasoning without requiring finetuning (Trivedi et al., 2023). It was found that interleaving multiple steps of reasoning and retrieval reduced factual errors in CoT and improved performance over both no-retrieval and single-step retrieval baselines.

2.4 Research Trends and Gaps

Across the existing literature, chain-of-thought reasoning has developed into iterative processes where models can revise reasoning steps mid-generation, and knowledge retrieval is increasingly integrated into these interleaved processes. However, there has been little application of retrieval augmentation to sports reasoning benchmarks and no agentic approaches tested on SportQA where structured domain knowledge can be beneficial. Our work intends to bridge this gap by combining CoT, tool calling, and domain-partitioned RAG to improve language model performance on expert-level sports scenario reasoning.

3 Data

We evaluate on SportQA (Xia et al., 2024), a publicly available benchmark for sports reasoning tasks with large language models (https://github.com/haotianxia/SportQA). The dataset comprises 70,592 multiple-choice questions offering varying levels of difficulty, from straightforward historical facts to complex, scenario-based reasoning tasks that necessitate extensive sports knowledge and experience. The questions are divided into three precise levels of difficulty: factual and historical knowledge (Level-1, 21,385 questions), rules and tactics comprehension (Level-2, 45,685 questions), and complex sports scenario analysis designed for sports specialists with years of experience (Level-3, 3,522 questions). Our primary evaluation target for this work is Level-3: 3,522 expert scenario questions across six sports (soccer, basketball, volleyball, tennis, table tennis, and American football). Each Level-3 question is multi-select, requiring the model to identify between one and four correct answers from four to five options, accurately reflecting the multi-faceted nature of such tactical decisions. Due to compute constraints, we sampled 234 questions from the Level-3 test set and used this same sample consistently across all ablations, with 119 single-hop and 115 multi-hop questions. Single-hop questions consist of a single main question with one set of answer options. Multi-hop questions consist of a main question and several sub-questions, each with its own set of answer options; a multi-hop question is scored correct only if the main question and all sub-questions are answered correctly. We evaluate using exact match (EM) and token-level F1 as defined in Section 4.2.

For the retrieval phase, both SportsRAG and SportsIQ share the same knowledge corpus created via an expansive web crawl and document extraction. The corpus is split into three indices to consult the model's reasoning skills from three different yet equally important angles: rules, strategy, and game context. The `check_rules` index contains official rulebooks published by major sports leagues and governing bodies, adding up to more than 800 pages of official rule-based text. The `lookup_tactics` index draws from coaching handbooks, strategy guides, and curated Wikipedia articles balanced across sports. Finally, the `analyze_situation` index is built from play-by-play recaps and post-game analyses sourced from various internet sources. We hand-curated a seed corpus by prompting the LLM to paraphrase and synthesize specific situational topics (i.e. two-minute drill, late-game fouling, penalty shootouts, etc.) from official sources and asked a second LLM to adversarially audit the content to correct for factual mistakes, then supplemented with data from Wikipedia, BBC Sport, NCAA.com, and similar platforms. The final corpus contains 6,833 chunks: 1,414 rulebook chunks, 1,947 tactics chunks, and 3,472 game situation chunks.

The SportQA dataset was the natural choice for our task given it is the only currently existing benchmark that tests tactical reasoning in sports, and is thus best suited to serve as the judge for the extent to which retrieval augmented generation and multi-step agentic reasoning close the gap between language model and human expert performance. The data sources we chose for retrieval are suited for our task because the three indices map directly onto the three dimensions of sports tactical reasoning that the SportQA error analysis identified as bottlenecks: rule knowledge, tactical principles, and situ-

ational context. We also deliberately chose sources to balance across sports to address the accuracy gap we observed in underrepresented sports, and chose exclusively from official rulebooks, popular coaching materials, and reputable analysis sources to ensure the retrieved evidence is factually grounded.

For data preprocessing, all three corpora share a common chunk schema (text, sport, source, section, page number, token count) and are indexed via dense BAAI/bge-base-en-v1.5 embeddings in a FAISS index alongside a BM25 sparse index with stopword removal. For the `check_rules` index, we extracted text from each rulebook PDF using pypdf, then applied a section-aware sliding-window chunking method with sport-specific regex patterns that detect rule headers, targeting 300-500 tokens with 50 token overlap, splitting oversized paragraphs at sentence boundaries and merging undersized ones. The `lookup_tactics` chunks were manually curated and cleaned and chunked similarly. For the `analyze_situation` index, we kept only paragraphs that mentioned both a specific game state (i.e. a score pattern) and a named entity (team or player), and excluded sections like biographies, broadcasting notes, team rosters, etc. SportQA questions were used as-is, with sampling from 980 to 234 questions.

4 Method

4.1 Baseline Reproduction

For baseline reproduction, we replicate results from the original SportQA paper (Xia et al., 2024), which introduced a three-level benchmark for evaluating sports reasoning in language models. The best results at the time of publication were produced by the GPT-4 series, with the authors noting that budgetary constraints prevented evaluation of stronger models. We address this directly by evaluating two models: GPT-4.1-mini, which closely approximates GPT-4 in capability, and GPT-5.4, one of the most recent models available to us. We replicate the three prompting conditions from the original paper: zero-shot chain-of-thought (0s_CoT), five-shot simple prompting (5s_SP), and five-shot chain-of-thought (5s_CoT). Since our SportsIQ system targets Level-3 questions exclusively, we focus our analysis there and adopt GPT-5.4 under zero-shot CoT as the primary baseline for all subsequent experiments, as it achieves the strongest Level-3 performance without requiring hand-crafted examples.

GPT-4.1-mini meets or exceeds GPT-4 across all three prompt conditions on Level-3 (see Appendix B), validating that our experimental setup is consistent with the original paper.

Table 1 shows overall Level-3 performance for both models under 0s-CoT. GPT-5.4 achieves 64.10% EM and 87.99% F1, roughly 12 percentage points above GPT-4.1-mini (51.71% EM). Despite this improvement, a substantial gap relative to human performance remains: human accuracy exceeds 90% across all Level-3 subtasks (Xia et al., 2024). Table 2 reveals the primary bottleneck: GPT-5.4 achieves 83.19% single-hop EM but only 44.35% multi-hop EM, a gap of roughly 39 percentage points. The high multi-hop F1 (84.95%) relative to EM indicates that the model correctly answers most individual sub-questions but rarely gets all of them right simultaneously.

Model	EM	F1
GPT-4.1-mini	51.71	83.07
GPT-5.4	64.10	87.99

Table 1: Level-3 performance under zero-shot CoT prompting. GPT-5.4 serves as the primary baseline for all SportsRAG experiments.

Model	Multi EM	Multi F1	Single EM	Single F1
4.1-mini	30.43	79.00	72.27	87.01
5.4	44.35	84.95	83.19	90.92
Human	>93	–	>96	–

*4.1-mini: GPT-4.1-mini, *5.4: GPT-5.4
Human EM averaged across easy/hard subsets (Xia et al., 2024)

Table 2: Level-3 breakdown by hop under zero-shot CoT. Human F1 is not reported in the original paper.

4.2 Evaluation Setup

We evaluate all methods using two metrics: Exact Match (EM) and token-level F1. EM checks whether the predicted answer set exactly matches the gold answer set, and is the primary metric for all comparisons. F1 measures token-level precision and recall overlap between the prediction and the gold answer, giving partial credit when a subset of correct answers is identified. We report both because Level-3 questions follow a multiple-select format where one to four answer choices may be correct, making EM strict but necessary for a fair assessment of full reasoning correctness.

4.3 Retrieval Infrastructure

All three approaches in this paper share a common retrieval layer built on top of the three corpora described in Section 3. We briefly describe the retrieval components here; corpus construction details are covered in Section 3.

Hybrid Retrieval. At query time, each retrieval query is issued against all three indices in parallel using hybrid retrieval combining dense and sparse signals. Dense retrieval retrieves by cosine similarity in a shared embedding space and sparse retrieval uses BM25, scoring chunks by term frequency and inverse document frequency. Results from both retrievers are fused using Reciprocal Rank Fusion before being passed to the reranker.

Reranking. Retrieved candidates from all three indices are merged, deduplicated, and scored by bge-reranker-v2-m3, a cross-encoder reranker that jointly encodes the query and each candidate chunk to produce a relevance score. Chunks scoring below a threshold of 0.5 are discarded. If all chunks fall below the threshold, the model answers from parametric knowledge alone. The top k surviving chunks are passed as context to the language model, where $k = 3$ in the final configuration.

Query Decomposition. Rather than issuing the full question text as a single retrieval query, we prompt the model to decompose the question into short, focused retrieval queries targeting specific rules, tactics, or situational conditions. For multi-hop questions, each sub-question is decomposed independently, so each part of the question retrieves its own relevant evidence. This decomposition step is applied consistently across all three methods.

4.4 SportsRAG

SportsRAG is a retrieval-augmented generation pipeline that augments GPT-5.4 inference on Level-3 questions with evidence retrieved from the three domain-specific corpora. Given a question, the system retrieves relevant chunks using the shared infrastructure described in Section 4.3, constructs a context window, and prompts the model to answer using both the retrieved evidence and its parametric knowledge.

Ablation Experiments. We developed SportsRAG through 13 ablations, each targeting a failure mode identified in the previous configuration. Naive retrieval with no score threshold introduced noise that hurt overall performance relative to the no-retrieval baseline; adding a conditional prompt and a reranker score threshold of 0.5 recovered most of the gap, and restricting retrieved chunks to the question's sport brought single-hop EM to within 1 percentage point of the baseline. Multi-hop questions remained the key bottleneck across all standard retrieval configurations, motivating the HyDE approach described below.

Hypothetical Document Expansion (HyDE). To improve multi-hop retrieval, we applied Hypothetical Document Expansion (HyDE) (Yoon et al., 2025). Before query decomposition, the model is prompted to generate a short hypothetical passage written in the declarative style of an official rulebook or coaching manual, directly addressing the core concept in the question. This hypothesis is used in place of the raw question text for retrieval. Because rulebook and coaching content is declarative in style, a hypothesis written in that style embeds closer to real corpus chunks than the raw question does, improving retrieval precision. Applying HyDE to multi-hop questions produced the first configuration to exceed the no-retrieval baseline on multi-hop EM (v9: 46.09% vs 44.35%).

Final System. The final SportsRAG configuration (v13) routes questions by hop type, applying the optimal retrieval strategy for each. Single-hop questions use top-k=3 retrieval with a score threshold of 0.5 and sport filtering, without HyDE. Multi-hop questions use HyDE followed by top-k=3 retrieval with a score threshold of 0.5, without sport filtering, as multi-hop sub-questions benefit from broader cross-sport context. This configuration achieves 64.53% overall EM and 86.11% F1, exceeding the no-retrieval baseline of 64.10% EM, with single-hop EM of 82.35% and multi-hop EM of 46.09%.

4.5 SportsIQ

While SportsRAG retrieves evidence through a fixed pipeline regardless of question content, our second approach, named SportsIQ, lets the model decide dynamically which corpus to query, in what order, and whether additional retrieval is warranted after seeing the initial results via an agentic reasoning workflow. We implement this using a ReAct (Reason + Act) loop (Yao et al., 2022), where the model interleaves reasoning steps (the *Reason* step) with tool calls (the *Action* step) and observes each result before deciding the next ac-

tion. The agent has access to three tools corresponding to the three corpora: `check_rules`, `lookup_tactics`, and `analyze_situation`, each taking a free-form query string and returning the top-k reranked chunks to ground the model. Tool selection and query formulation are left entirely to the model. A critical implementation detail is that parallel tool calling is disabled: forcing strictly sequential tool calls is what enables organic iterative reasoning, as opposed to batch retrieval followed by a single generation step.

For multi-hop questions, we noticed that early experiments using separate ReAct sub-loops for each individual sub-question suffered from error propagation: a single incorrect sub-answer could poison the context for subsequent sub-questions. We addressed this by running a single ReAct loop over the full question, with all sub-questions presented simultaneously. The model gathers evidence holistically and produces all answers in one generation, eliminating the cascade failure. This single architectural change recovered 6.09 percentage points on multi-hop EM (38.26% $\rightarrow$ 44.35%), bringing the agent to parity with the no-retrieval baseline on multi-hop. The best standalone agent configuration achieves 62.39% overall EM and 87.71% F1, establishing dynamic tool routing as a viable alternative to fixed retrieval pipelines. The following experiments build on this foundation by adding confidence gating, conditional verification, and a multi-hop disagreement gate to address the remaining failure modes.

Ablation Experiments. Our baseline error analysis showed that 93% of wrong predictions are close misses ($EM = 0$ but $F1 >= 0.5$) and that single-hop errors are dominated by over-selecting or under-selecting (42% over-select, 39% under-select), not missing knowledge. To act on this, we experimented with several techniques on both SportsRAG and SportsIQ, including prompt engineering for option calibration, sub-question patching for off-by-one multi-hop errors, decomposed self-consistency voting, confidence gating, conditional verification, and inter-sample disagreement detection, and found the following three techniques to be the most helpful.

Confidence gate. Before retrieval, the model answers the question directly and self-reports its confidence as HIGH, MEDIUM, or LOW. HIGH-confidence questions skip retrieval entirely and use the original answer from the base model. MEDIUM and LOW questions proceed through the full retrieval pipeline. This addresses the failure mode we observed in early SportsRAG ablations where naive RAG hurt easy questions since the model already knew the answers, and retrieval added noise.

Conditional verification. After an answer is generated, a second LLM call reviews each option using the top retrieved chunks as evidence and revises its answer if needed. We discovered through systematic ablation that verification is strongly hop-dependent. It helps single-hop questions where the model needs to decide which options to include or exclude, but it affected multi-hop questions negatively because reconsidering one sub-answer can cascade into flipping correct sub-answers in other parts. Our final design applies verification only to single-hop questions.

Multi-hop disagreement gate. For multi-hop HIGH-confidence questions, where the gate's accuracy drops, we replace self-reported confidence with inter-sample agreement, which is a second independent generation at temperature=0.3 either agrees with the first answer (in which case we keep the gated parametric answer) or disagrees (in which case we override the HIGH classification and route the question to the agent loop). This addresses the multi-hop overconfidence problem directly. The model's own confidence judgment can be inaccurate, but two independent samples diverging is a more reliable signal that the answer is uncertain.

Final System. Our best configuration combines two variants of SportsIQ via hop-based routing. Single-hop questions are handled by the agent with confidence gating and verification, which leverages adaptive retrieval on the single-hop questions the gate sends to RAG and benefits from verification's calibration boost on single-answer selection. Multi-hop questions are handled by the same agent setup with the disagreement gate added, and disagreements override HIGH and route the question to the agent's retrieval loop. Full implementation details including frameworks, compute, and inference setup are provided in Appendix E.

5 Results

We report results for SportsRAG in Section 5.1. Results for the agentic approaches are reported in Section 5.2. Statistical significance testing is not reported, as all methods are evaluated on the same fixed test set of 234 Level-3 questions with

no repeated runs.

5.1 SportsRAG

Table 3 shows overall and per-hop EM across the key SportsRAG configurations. Naive retrieval (v1) hurts performance substantially relative to the no-retrieval baseline (54.27% vs 64.10% overall EM), as retrieved chunks introduce noise for questions the model can already answer correctly. Introducing a conditional generation prompt and score threshold (v2) recovers most of this loss, with multi-hop EM nearly matching the baseline (43.48% vs 44.35%). Applying HyDE to multi-hop questions (v9) produces the first configuration to exceed the no-retrieval baseline on multi-hop EM (46.09% vs 44.35%). The final configuration (v13) routes by hop type, combining v3 single-hop retrieval with v9 multi-hop retrieval, and achieves 64.53% overall EM, 82.35% single-hop EM, and 46.09% multi-hop EM, exceeding the no-retrieval baseline on all three measures.

System	Ovr. EM	F1	S-EM	M-EM
No retrieval	64.10	87.99	83.19	44.35
Naive RAG	54.27	82.64	72.27	35.65
Cond. prompt + threshold	58.55	83.07	73.11	43.48
Sport-filtered retrieval	61.54	85.43	82.35	40.00
HyDE retrieval	59.83	84.20	73.11	46.09
SportsRAG	**64.53**	**86.11**	**82.35**	**46.09**

Table 3: SportsRAG ablation results. S-EM = Single-hop EM, M-EM = Multi-hop EM.

Table 4 shows multi-hop EM broken down by sport for the baseline, v2, and v13. Basketball and Soccer both fall below the no-retrieval baseline in v13 (42.11% vs 47.37% and 37.50% vs 45.83% respectively). Basketball likely suffers because the model already has strong parametric knowledge of the rules, making retrieved chunks redundant or distracting. Soccer consistently underperforms across all retrieval configurations, suggesting that the complexity of FIFA regulations causes the model to generate inaccurate hypotheses that retrieve irrelevant chunks. Table Tennis shows a different pattern: the conditional prompt and threshold in v2 alone produced the highest single-sport multi-hop EM of any configuration (60.00%), but HyDE in v13 returned it to the baseline level (40.00%), indicating that precise threshold filtering benefited Table Tennis specifically in ways that hypothesis-based retrieval did not preserve.

Sport	Baseline	v2	v13
American Football	37.50	37.50	**46.88**
Basketball	47.37	47.37	42.11
Soccer	45.83	37.50	37.50
Table Tennis	40.00	**60.00**	40.00
Tennis	57.14	57.14	**64.29**
Volleyball	43.75	37.50	**50.00**
Overall multi-hop	44.35	43.48	**46.09**

Table 4: Multi-hop EM (%) by sport for the no-retrieval baseline, v2 (best pre-HyDE configuration), and v13 (final SportsRAG). Bold indicates best result per row.

5.2 SportsIQ

The base ReAct agent with parallel tool calls scores 59.40% overall EM, well below the GPT 5.4 vanilla no-retrieval baseline. Switching to sequential tool calls recovers most of this gap (61.54%), and tuned prompts that present all sub-questions in a single loop bring multi-hop EM to 44.35%, matching the no-retrieval baseline and confirming that output formatting rather than retrieval quality was the limiting factor at that stage. Expanding the situation corpus 20x further improves single-hop EM to 80.67%, yielding our best standalone configuration at 62.39% overall EM and 87.71% F1, exceeding SportsRAG on F1 (+1.60 pp). The consistently high multi-hop F1 (85.10%) relative to multi-hop EM (43.48%) indicates the agent resolves most individual sub-questions correctly but rarely gets all parts right simultaneously, motivating the extensions described below.

System	Ovr.	F1	S-EM	M-EM
No retrieval	64.10	87.99	83.19	44.35
ReAct, parallel tools	59.40	86.03	79.83	38.26
ReAct, sequential tools	61.54	86.91	79.83	42.61
+tuned prompts	62.39	87.48	79.83	44.35
+expanded corpus	62.39	87.71	80.67	43.48
+conf. gate + verification	64.96	89.36	85.71	43.48
+disagreement gate	65.81	88.71	83.19	47.83
Ensemble (hop-routed)	67.09	89.49	85.71	47.83

Table 5: SportsIQ ablation results. Ovr. = Overall EM, S-EM = Single-hop EM, M-EM = Multi-hop EM.

Best result. Our reproducible best is the hop-based ensemble combining SportsIQ (ReAct + gate + verify) on single-hop questions and SportsRAG (gate + verify-single) on multi-hop questions, achieving 66.24% overall EM (+2.14 pp over the no-retrieval baseline), with 85.71% on single-hop (+2.52 pp) and 46.09% on multi-hop (+1.74 pp). Replacing SportsRAG with SportsIQ + dis-

agreement gate on the multi-hop side raises the ensemble to 67.09% in our best run, but reproductions (Exp 9v4/v5) converge to 63.68%, suggesting that the 65.81% result of Exp 9v2 benefited from favorable single-hop gate stochasticity. However, the disagreement gate's multi-hop contribution is still positive and stable with 3 questions fixed across all runs.

Qualitative Analysis. An LLM-as-judge evaluation of 1,020 retrieved chunks (Exp 2) finds 99.8% precision@k, a 0.2% noise rate, and 95.2% has-relevant rate. Of 96 wrong predictions, 91 had relevant retrieved evidence and only 5 errors (5%) are attributable to retrieval failure. This indicates that performance gaps are on generation and reasoning rather than retrieval.

In terms of confidence gate calibration, the model reports HIGH confidence on 78-81% of questions. In AGV, HIGH-confidence single-hop questions reach 90.3% EM without retrieval, while multi-hop reaches only 51.7%, indicating overconfidence on multi-hop. MEDIUM/LOW questions achieve only 17.9% multi-hop EM even with retrieval, showing that difficult questions remain unresolved. Detailed error analysis as well as examples are covered in Appendix C.

6 Discussion

Interpretation of results. One core finding from SportsRAG is that retrieval helps when the model's parametric knowledge is uncertain and hurts when it is not. For single-hop questions, the conditional prompt and score threshold are effective because they allow the model to ignore low-confidence chunks rather than being forced to use all retrieved context. For multi-hop questions, standard retrieval fails because the semantic gap between a complex scenario question and a relevant rulebook chunk is too large for bi-encoder retrieval alone. HyDE closes this gap by rephrasing the query in corpus style before retrieval. The routing strategy in SportsRAG v13, applying different retrieval configurations to single-hop and multi-hop questions, is the key decision that allows SportsRAG to exceed the no-retrieval baseline on both subtasks simultaneously.

We identified failure patterns by breaking down results per sport and comparing F1 against EM, rather than relying on aggregate metrics alone. This revealed that retrieval consistently hurts sports where parametric knowledge is strong (Basketball) or where rulebook complexity causes inaccurate hypothesis generation (Soccer), as discussed in Section 5.1. The high multi-hop F1 relative to EM across all configurations also shows that the model gets most sub-questions right individually but fails to get all of them right at once. More broadly, these results suggest that retrieval augmentation for sports reasoning is most useful when targeted carefully by question type and sport, rather than applied uniformly.

Strengths of our method compared to baselines. The same selective-application principle that drives SportsRAG v13 also applies to the agentic setup, and the pattern transfers between the two systems. SportsIQ's baseline already outperforms SportsRAG v3 on every aggregate metric (+3.9 pp overall), confirming that adaptive retrieval, allowing the model to decide which tool to call and with what query, outperforms a fixed three-way decomposition. However, the same single-hop noise problem persists: the agent over-retrieves on questions the model already knows, reducing single-hop EM below the no-retrieval baseline. The confidence gate and conditional verification address this directly.

Layering the confidence gate and conditional verification on top of the agent (AGV) raises single-hop performance to 85.71%, the highest single-hop EM of any configuration we tested and +2.52 pp above the no-retrieval baseline. The fact that the same gate and conditional-verify recipe is effective for both fixed-pipeline RAG and ReAct suggests that this is a general property of selective retrieval rather than an artifact of either architecture.

The strongest single-system result is the AGV configuration at 64.96% overall EM. However, the most reliable result is the hop-based ensemble at 66.24% (+2.14 pp), combining SportsIQ on single-hop questions with SportsRAG v13 on multi-hop. These systems exhibit complementary strengths: the agent's adaptive retrieval excels on single-hop, while SportsRAG's HyDE-based routing performs better on structurally complex multi-hop questions. Pipelines specialized for different question structures are complementary rather than competing.

Surprising findings. Several results were surprising. First, verification, which we expected to provide general calibration, turns out to be strongly hop-dependent. It yields a +4 net improvement on single-hop but a -8 net degradation on multi-hop. This arises from a structural issue: multi-hop EM

requires all sub-answers to be correct, so verification's tendency to reconsider individual options can flip a correct sub-answer to incorrect.

Second, sub-question patching and decomposed self-consistency fail to address the off-by-one bottleneck. The model's confidence for each sub-question is poorly calibrated (MEDIUM-confidence sub-answers are roughly equally likely to be correct or incorrect), and decomposing multi-hop questions into independent retrieval steps degrades performance because SportQA sub-questions share contextual dependencies that are lost under decomposition.

Third, the disagreement gate consistently improves multi-hop performance (+3 fixed, 0 broken across four reproduction runs), but its single-system EM does not fully reproduce due to stochasticity. Across all three observations, the pattern is consistent that components that appear beneficial in isolation can show failure modes depending on hop types.

Finally, HyDE, which was the key intervention that pushed SportsRAG past the no-retrieval baseline on multi-hop, consistently hurt the ReAct agent across two configurations. The option-blind hypothesis is generated without seeing the answer choices, and in the agentic setting where the model also controls query formulation dynamically, the pre-generated hypothesis appears to anchor retrieval on the wrong aspect of the question.

Practical Implications. Practically, this means further pipeline engineering on a single base model has diminishing returns past where we are. Breaking through the multi-hop bottleneck would require either model diversity (a second voter from a different model family like Claude or Gemini to reduce correlated errors) or a fundamentally different approach to multi-hop reasoning that does not treat sub-questions as independently verifiable. These results also suggest that the SportQA Level-3 multi-hop bottleneck is not a knowledge problem (99.8% retrieval precision rules that out) and not a prompting problem, but a calibration problem at the edge of what the model can do. Selective component/technique application is the most helpful available option without changing the underlying model. This points to a broader finding: for expert-level sports QA, the bottleneck is not knowledge access but reasoning calibration, and future work should prioritize confidence modeling and incorporating model diversity over retrieval architecture.

7 Conclusion and Future Work

To summarize, our work aims to bridge the performance gap between baseline LLM and human experts on Level-3 SportQA scenario reasoning questions, which is an unsolved bottleneck identified in Xia et al. (2024) (Xia et al., 2024). In our baseline error analysis, we found that even GPT-5.4 sits well below human expert performance, largely driven by knowledge gaps and multi-hop reasoning failures. To target these two failure modes, we built a sport-specialized retrieval corpus with three partitions: rules, tactics, and game situations. We implemented two retrieval strategies: SportsRAG, a fixed three-way decomposition RAG approach, and SportsIQ, a ReAct agent with dynamic tool selection over the same indices. On top of these two frameworks, we performed extensions and experiments using a variety of techniques, including confidence gating, conditional verification, inter-sample disagreement gate, HyDE, per-option retrieval, sequential processing, and forced looping. The sharpest pattern across both systems is that nothing worked for both question types at once: every technique that improved multi-hop hurt single-hop, and vice versa. Hop-type routing turned out to be the single most important architectural decision we made.

The ensemble reaches 67.09% EM, up from 64.10% with no retrieval and 47% with GPT-4 in the original paper. Human experts still score above 93%. The error analysis suggests that closing the remaining gap is not primarily a retrieval problem: 91 of 96 wrong predictions had relevant evidence retrieved, and 66% of shared failures were HIGH-confidence errors.

Beyond the SportQA setup, our results point to a general principle: for any retrieval-augmented system built on top of a strong base model, the dominant failure mode of RAG is interference, not missing knowledge. The practical question for practitioners is less about "how to retrieve more relevant context" and more "when should the model not retrieve at all." For future directions, the most helpful next step would be introducing model diversity through a third voter from a different model family (e.g., Claude or Gemini), since our analysis showed that 70% of residual errors are identical across both systems, indicating the ceiling we face is due to the model not the pipeline. Additionally, making the multi-hop confidence gate adaptive per sport (e.g., always routing American

Football penalty questions to retrieval, where overconfidence is most severe) and exploring CoT distillation to improve the base model's multi-hop calibration without retrieval overhead are promising directions for closing the remaining gap to human expert performance.

8 Limitations

Several limitations constrain the scope and generalizability of our findings. All reported results are based on a fixed 234-question sample of the 980-question local L3 test set, fixed before any system was evaluated to ensure fair comparison. That being said, absolute numbers may shift on the full set. L3 SportQA also covers only 6 sports, while the full benchmark spans 35 across Levels 1 and 2, leaving generalization to other sport types untested at this stage. On the method side, all experiments use a single model (GPT-5.4 via the CMU LiteLLM gateway), and performance may differ for other LLMs, particularly smaller models with weaker parametric sports knowledge. HyDE is additionally option-blind: hypotheses are generated from the question stem alone, without seeing the answer choices, which may cause retrieval to target the wrong aspect of a question and hurt overall performance. Regarding evaluation, exact match proves to be a strict metric for multi-select questions, likely underestimating practical system quality across the board, but the inclusion of F1 score as a secondary metric supplements this well. Finally, many failures happen to be shared across all three systems, suggesting a model-level knowledge ceiling that retrieval and agentic routing alone cannot address.

9 Ethical Considerations

This work raises a small number of ethical considerations. The most direct risk is deployment: at 67% overall exact match, no component of this system is accurate enough for unsupervised use in officiating assistance, coaching tools, or rules adjudication, and any real-world application would require a human in the loop setup for final reviews. LLM and RAG systems are occasionally inclined to produce confident-sounding incorrect answers even when retrieved evidence is correct, as the model may misinterpret or selectively use retrieved chunks. On the data side, all three corpora draw from English-language public sources, which skews coverage toward Western and Olympic sports with stronger online documentation. Regional sports in non-English speaking countries would receive weaker retrieval support, a gap our 6-sport evaluation does not surface. Researchers building on this work should prioritize non-English and non-Western sources when extending coverage, and treat model confidence scores as weak signals rather than reliable thresholds for automation. Those extending this work to higher-stakes rule-governed domains like law or medicine should treat the domain-partitioned corpus design as transferable but invest heavily in extensive auditing, as errors in rule retrieval carry significantly greater consequences in these domains than they would in sports.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *Preprint*, arXiv:1906.02361.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *Preprint*, arXiv:2212.10509.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and

Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-fang Wang, and Weining Shen. 2024. Sportqa: A benchmark for sports understanding in large language models. 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 5061–5081.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Yejun Yoon, Jaeyoon Jung, Seunghyun Yoon, and Kunwoo Park. 2025. Hypothetical documents or knowledge leakage? rethinking llm-based query expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19170–19187.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Preprint*, arXiv:2203.14465.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A SportsRAG Ablation Results

All runs use GPT-5.4 on 234 Level-3 questions (seed=42).

The key findings across ablations are: (1) score thresholding and conditional prompting are necessary to prevent retrieved noise from hurting parametric knowledge; (2) sport filtering helps single-hop questions but hurts multi-hop, which benefits from broader cross-sport context; (3) more complex retrieval strategies (grouped context, specific decomposition prompts, larger top-k) consistently underperformed simpler high-precision retrieval; and (4) HyDE is the only intervention that improved multi-hop EM above the no-retrieval baseline.

B GPT-4 vs. GPT-4.1-mini Comparison

Table 6 compares our GPT-4.1-mini reproduction against the original GPT-4 results reported in Xia et al. (2024) on Level-3. GPT-4.1-mini meets or exceeds GPT-4 on both single-hop and multi-hop EM across all three prompt conditions, validating that our experimental setup is consistent with the original paper.

C Error analysis.

We performed error analysis by comparing per-question predictions against gold answers across all configurations. For single-hop questions, we classified each error as over-selection (the model picked more options than correct), under-selection (the model picked fewer options than correct), or completely wrong (no overlap between predicted and gold answer letters). For multi-hop questions, we counted how many sub-question parts were wrong per question to identify the off-by-one pattern, and we aggregated errors with the confidence gate's routing decision (HIGH vs MEDIUM vs LOW) and with the retrieval quality judgments from our LLM-as-judge evaluation to distinguish generation failures (relevant chunks were retrieved but the answer was still wrong) from retrieval failures (no relevant chunks were found).

The consistently high F1 relative to EM on multi-hop questions (84.26% vs 46.09% in v13) confirms that the model resolves most individual sub-questions correctly but fails to get all of them right simultaneously. Soccer is the most persistent failure case: retrieval hurts it across every configuration, and HyDE does not recover the gap. This could point to a corpus coverage issue rather than a retrieval quality issue where the FIFA rulebook chunks may not align well with the scenario-based framing of Level-3 soccer questions. Basketball presents the opposite problem: the model's parametric knowledge is strong enough that retrieved chunks are more likely to distract than to help, and no retrieval configuration improves on the baseline.

Verification, which works cleanly on single-hop (+4 net), is destructive on multi-hop (-8 net in Exp 5b) because reconsidering one sub-answer often cascades into flipping correct sub-answers in other parts. Multi-hop EM requires every sub-answer to be exactly right, so any cascade reduces the score. This is why our final configuration disables verification on the multi-hop side, and it is also why the off-by-one error pattern is so persistent.

These dynamics show up clearly in the residual error breakdown of our best configuration. The AGV configuration leaves 82 errors: 17 single-hop (14.3%) and 65 multi-hop (56.5%). About 85% are close misses with F1>=0.5, and 49% of multi-hop errors miss exactly one sub-question. Under-selection dominates single-hop errors (65%). Among 77 shared failures between AGV and SportsRAG, 66% are HIGH-confidence errors, indicat-

Model	Prompt	L3 Single EM	L3 Multi EM
GPT-4	0s_cot	61.09	27.30
GPT-4.1-mini	0s_cot	72.27	30.43
GPT-4	5s_cot	68.83	28.71
GPT-4.1-mini	5s_cot	72.27	33.91
GPT-4	5s_sp	67.00	29.15
GPT-4.1-mini	5s_sp	72.27	32.17

Table 6: Direct comparison of GPT-4 (Xia et al., 2024) and our GPT-4.1-mini reproduction on Level-3. GPT-4 numbers are averaged across easy and hard subsets, as the original paper does not report aggregate hop-level scores.

Configuration	Overall EM	Single EM	Multi EM	Overall F1	Single F1	Multi F1
Baseline (no RAG)	64.10	83.19	44.35	87.99	90.92	84.95
SportsRAG v1 (naive)	54.27	72.27	35.65	82.64	85.29	79.90
SportsRAG v3 (+threshold, sport filter)	61.54	82.35	40.00	85.43	87.90	82.88
SportsRAG + gate (Exp 3)	61.97	78.15	45.22	85.33	87.23	83.36
SportsRAG + gate + verify-all (Exp 5b)	60.26	81.51	38.26	83.56	90.14	76.75
SportsRAG + gate + verify-single (Exp 5c)	**64.10**	81.51	**46.09**	87.07	89.16	84.92
SportsIQ baseline	62.40	79.80	44.40	86.20	88.70	86.20
SportsIQ + gate + verify-single (AGV)	**64.96**	**85.71**	43.48	**89.36**	**93.28**	85.30
SportsIQ + AGV + disagreement gate (Exp 9v2)	65.81	83.19	47.83	88.71	91.74	85.57
SportsIQ + AGV + disagreement gate (Exp 9v4/v5)*	63.68	80.67	46.09	88.34	—	—
Ensemble (AGV single + SportsRAG 5c multi)	**66.24**	**85.71**	46.09	**89.17**	**93.28**	84.92
Ensemble (AGV single + Exp 9v2 multi)	**67.09**	**85.71**	**47.83**	**89.49**	**93.28**	**85.57**

Table 7: Level-3 results across configurations on the 234-question sample. All metrics are percentages. AGV = Agent + gate + verify-single. * 9v4/v5 are reproductions of 9v2 with explicit seeds; the disagreement gate's multi-hop contribution (+3 fixed, 0 broken) is consistent across runs, while v2's overall EM benefited from favorable single-hop gate stochasticity.

ing systematic overconfidence of the LLM. Some failure examples are included in the appendix.

D Failure Case Analysis

D.1 Category 1: Multi-hop Off-by-One with Overconfident Gating

The model reports HIGH confidence and skips retrieval, but gets exactly one sub-question wrong. These are near-miss failures where EM is counted as 0 despite being one answer away from fully correct.

D.2 Category 2: Single-hop Under-selection

The model predicts a strict subset of the correct answers. Under-selection accounts for 65% of single-hop errors.

D.3 Category 3: Verification Cascade Failure (Multi-hop)

Verification can degrade correct answers by second-guessing. The third case is particularly severe: verification collapses a fully correct prediction into an empty answer. This motivates disabling verification for multi-hop.

D.4 Category 4: Systematic Wrong Predictions

Both systems produce identical incorrect outputs, indicating model-level limitations. Approximately 70% of shared failures follow this pattern, suggesting errors stem from parametric knowledge rather than retrieval or prompting.

D.5 Category 5: Disagreement Gate Success

A second generation (temperature = 0.3) disagrees with the initial answer, triggering retrieval that resolves the error.

E Implementation Details

All experiments run GPT-5.4 via the CMU LiteLLM gateway with temperature=0 unless otherwise noted. The SportsIQ ReAct loop is implemented in Python using the OpenAI function-calling API with tool definitions for each of the three corpora. Dense retrieval uses BAAI/bge-base-

Question	Sport	Part	Error	Gold	Pred
4th and long tactic...	Am. Football	sq2	Wrong choice	B	C
Face mask on QB...	Am. Football	sq1	Over-select	B	B,D
Sideline catch...	Am. Football	sq1	Under-select	A,B	B

Table 8: Multi-hop off-by-one errors under HIGH-confidence gating.

Question	Sport	Gold	Pred	Missing
Replay reviewable plays...	Am. Football	A,B,C	A	B,C
Replay assistance scenarios...	Am. Football	A,B,D	A,B	D
Equipment regulations...	Am. Football	A,B	B	A

Table 9: Single-hop under-selection errors.

en-v1.5 via the sentence-transformers library indexed in FAISS; sparse retrieval uses rank_bm25. Reranking uses bge-reranker-v2-m3 via HuggingFace. Each query generates on average 3 tool calls and roughly 2,000–3,000 tokens per question.

Question	Sport	Before	After	Change
4th and long tactic...	Am. Football	All correct	sq2: B→C	Flipped correct
2–0 game scorer...	Am. Football	All correct	main: C→D	Flipped correct
QB mental toughness...	Am. Football	All correct	Empty	Collapse

Table 10: Verification-induced failures.

Question	Sport	Gold	Both Predict	Wrong Parts
Receiver end zone case...	Am. Football	main=A, sq2=A	main=B, sq2=B	2/3
Hail Mary penalties...	Am. Football	main=C, sq2=A	main=D, sq2=B	2/3
Receiver contact...	Am. Football	sq1=C	sq1=A	1/3

Table 11: Shared failure cases across systems.

Question	Sport	LLM1	LLM2	Disagreement	Final
QB contact penalties...	Am. Football	B,C,D	B,D	Over-select C	B,D
Tennis in/out calls...	Tennis	A,B,D	A	Under-select	A,B,D
Umpire contested call...	Tennis	sq2=C	sq2=A,C	Missing option	A,C

Table 12: Disagreement gate corrections.