

RAG-Based Volleyball Tactical Review from Play-by-Play Data: Serve-Receive-Attack Chain Analysis

Emily Jiang¹, Alan Feng²

¹Computer Science, The Ohio State University, Columbus, OH, USA

²Kinesiology, The Ohio State University, Columbus, OH, USA

e.jiang_tech@gmail.com

ABSTRACT

Play-by-play volleyball logs preserve the order of contacts that box scores collapse, but automated tactical review depends on accurate rally reconstruction and dependable evidence retrieval. This study evaluated an end-to-end retrieval-augmented generation pipeline on the 2025 NCAA Division I women's volleyball season. Canonicalization reduced 7,505,822 raw events to 5,654,956 by removing 1,684 verified duplicate full-match logs. A serve-anchored parser reconstructed 823,508 rallies, of which 821,923 (99.81%) had a valid sequence and point winner. Across the league, side-out was 57.13%, organized receive-set-attack occurred on 76.51% of receive opportunities, and the source-labeled first-ball kill rate was 29.26% (28.88% under the strict complete-chain funnel). Match winners exceeded losers by 12.02 percentage points in side-out and 7.56 points in first-ball kills. Pressure rallies showed a 1.17-point unadjusted side-out increase, but the team-, rotation-, and score-stratified odds ratio was 1.012 ($p = 0.198$). The pipeline created 60,852 match-team-rotation evidence cards and evaluated 270 retrieval queries. Structured hybrid retrieval achieved 99.63% Recall@1 and 100% Recall@5, compared with 30.37% and 61.48% for TF-IDF. Across 36 tactical reviews, chain-and-rotation retrieval produced 53.33% evidence coverage, 88.33% tactical specificity, a 68.10 composite quality score, and 100% numeric precision. In the championship final, Texas A&M's 71.67% side-out rate was driven by free points and extended-rally conversion rather than superior first-ball finishing. The workflow converts volleyball event streams into traceable, rotation-specific tactical narratives.

Keywords: *NCAA volleyball; play-by-play; side-out; first-ball kill; rotation analysis; retrieval-augmented generation; sports analytics; tactical review*

Introduction

Volleyball is a sequential sport. The serve constrains first contact, reception quality shapes setter access, the set determines tempo and attack lane, and the first attack either ends the rally or moves the point into transition. Match totals preserve kills, attempts, aces, errors, digs, and blocks, but they do not preserve the dependency structure that connects those actions. The `ncaavolleyballr` package provides programmatic access to NCAA records, schedules, player statistics, and play-by-play data [1], while the NCAA statistician's manual defines the event vocabulary used to distinguish attacks, kills, receptions, service errors, digs, blocks, and related actions [2]. These sources make season-scale analysis possible, but tactical meaning still has to be reconstructed from ordered contacts.

A useful tactical review must locate the stage at which a receiving possession succeeds or fails. A serve may end immediately as an opponent service error or ace; a live serve may be received without producing a complete offensive sequence; an organized first attack may fail to terminate; and a team may still recover through block-defense or transition play. Treating side-out as a single percentage hides these pathways. A serve-receive-set-attack chain instead separates access, organization, first-swing finishing, and the final rally outcome.

Season-scale reporting introduces a second problem. Once thousands of matches and rotation summaries have been created, a coach typically asks about one team, opponent, date, rotation, or pressure state. The reporting system must retrieve the correct evidence, preserve the numerical relationships among chain stages, and translate the evidence into a concise tactical diagnosis. Lexical similarity alone is unlikely to resolve near-duplicate records that share the same volleyball vocabulary but differ in team, opponent, date, and rotation.

This study evaluates a complete workflow for that task. It first canonicalizes complete match logs and reconstructs each rally from the initiating serve to the recorded point. It then measures league, team, rotation, pressure, and winner-loser differences; validates event counts against player-match box scores; converts the rally table into compact tactical evidence cards; compares four retrieval strategies; and evaluates tactical reviews produced without retrieval, with box-score retrieval, and with chain-and-rotation retrieval. The study asks whether ordered event logs can be reconstructed reliably, whether structured metadata materially improves tactical retrieval, and whether chain-level evidence yields more specific and numerically faithful reviews than conventional summaries.

Literature Review

Quantitative volleyball research increasingly treats the point as a process rather than a collection of isolated skills. Powers, Stancil, and Consiglio modeled NCAA women's volleyball contacts as a changing point-win probability and attributed each change to the responsible players [3]. Side-out research has likewise shown that the receiving phase is strongly associated with competitive success and that the way points are obtained matters alongside the final rate [4]. Setter position changes the available side-out structure [5], while jump-setting alters the timing and options of the first attack [6]. Together, these studies support a possession-chain view in which reception quality, setter access, attack creation, and rally completion are analytically distinct.

Research on elite volleyball further explains why the chain should be modeled explicitly. Setter decisions in high-performance women's volleyball depend on reception conditions and the opposing block [7], and setting zone in elite men's volleyball is linked to the quality of preceding contacts [8]. Attack type and efficacy also depend on the conditions under which the setter organizes the offense [9]. Serve, reception, and setting characteristics jointly determine setting efficacy [10], while Olympic match analysis shows a close relationship between reception and attack in both women's and men's competition [11]. These findings imply that an attack percentage without its preceding contacts provides only a partial tactical explanation.

The match-outcome literature reinforces that distinction. Direct scoring, error control, and terminal attacking are among the skills that separate winners from losers [12]. Network analysis extends the same logic by representing volleyball as connected game complexes whose central roles change between side-out and transition phases [13]. A tactical review should therefore identify whether a poor receiving result begins with serve pressure, reception access, offensive organization, first-ball conversion, or the team's inability to win the extended rally.

Retrieval-augmented generation offers a practical framework for converting those measurements into case-specific reports. RAG conditions generated text on selected external evidence [14], and dense passage retrieval provides learned semantic matching for open-domain questions [15]. Exact sports entities also suit classical vector-space retrieval [16] and BM25 [17], especially when team names, dates, opponents, and rotations are explicitly available. Ranked retrieval should be evaluated with measures such as normalized discounted cumulative gain [18] and reciprocal rank [19], while answer evaluation should separate retrieval quality, grounding, and generated-text quality [20].

Recent applied RAG work emphasizes that retrieval accuracy alone is not enough. Enterprise systems have combined evidence consistency, self-checking, and lightweight hallucination detection [21]. Financial applications have added XBRL-based numeric verification for SEC filing summaries [22] and claim-level checks designed to reduce numerical hallucination in corporate risk memos [23]. Although these domains differ from volleyball, they address the same reporting risk: a fluent narrative can become misleading when a number is detached from the evidence record that supports it.

Other studies show how feedback, abstention, and provenance can strengthen the reporting loop. Retrieval-augmented failure narratives for HDFS logs pair evidence with selective refusal when the record remains ambiguous [24]. Conversational text-to-SQL systems use schema-aware retrieval together with execution feedback to repair unsupported or invalid outputs [25]. Structured disclosure question answering demonstrates the value of metadata-constrained evidence over specialized financial records [26], while LLM reranking has been used to improve relevance and preserve faithful natural-language explanations [27]. Test-in-the-loop program repair adds an external verification cycle after generation [28], and evidence-chain RAG links claims to word-level attribution and provenance explanations [29].

Operational RAG systems also highlight the value of domain routing. Evidence-grounded DevOps question answering over private documents benefits from hybrid retrieval and evidence-chain filtering [30], while bilingual AIOps incident triage combines retrieval with root-cause summarization [31]. Retrieval-summary integration in code intelligence shows that finding the correct item and explaining it are separate but connected objectives [32]. Long-document clause analysis similarly depends on isolating the right document segment before generating a concise interpretation [33]. These results are especially relevant to tactical sports corpora, where thousands of records share common terminology and differ mainly in structured entities.

The remaining gap is a unified sports workflow that begins with raw play-by-play logs, reconstructs possession chains, retrieves the correct match-rotation evidence, and evaluates the resulting tactical narrative for both numerical fidelity and coaching specificity. Existing volleyball studies explain the importance of sequential contacts, while applied RAG studies explain how to retrieve, verify, and present evidence. The present study connects those two lines of work in a single auditable pipeline.

Materials and Methods

The corpus comprised the 2025 NCAA Division I women's volleyball collection and the companion `ncaavolleyballr` source. Package version 0.5.1 exposes NCAA women's and men's records, schedules, and statistics for the 2020-2025 seasons [1], while the materialized event table covered 2025 women's Division I. Play-by-play was linked with player-match, team-match, player-season, and team-season tables from August 22 through December 21. The raw event table contained

7,505,822 rows and 5,093 contest identifiers. Team-season membership defined ranking eligibility, while scheduled opponents remained in match and rally summaries.

Within each contest, rows were sorted by source order and contiguous runs were treated as candidate complete logs. A repeated run was removed only when its length, the sum of its 64-bit row hashes, and its bitwise XOR hash matched another run. Retaining one copy removed 1,850,866 rows from 1,684 repeated logs and yielded 5,654,956 canonical events. Date suffixes were normalized. Adjacent identical actions were retained because legitimate repeated contacts can occur within a rally.

Each Serve event opened a rally that ended immediately before the next Serve in the same contest and set. Scores were parsed in away-home order. The first legal post-serve score was compared with the pre-serve score, and the side that increased by one was assigned the point. Non-score labels, including media timeouts, inherited the preceding legal score; the next legal score still determined the winner. A rally was excluded when the point winner remained unresolved or when a noncanonical sequence prevented deterministic interpretation. This procedure identified 823,508 rallies, of which 821,923 were valid.

Event labels followed the NCAA statistician's manual [2]. A receive opportunity was every valid rally faced by a named receiving team, including rallies ended by an opponent service error. Serve in play excluded service errors. Reception required a live serve and a Reception event credited to the receiving team. An organized chain required a reception, set, and attack by that team in the correct order. First-ball kill was measured both as a rally-level source label and as a strict funnel outcome. The strict measure required the source label First ball kill to follow an organized reception-set-attack chain. Side-out meant that the receiving team won the rally, whether through the first attack, an opponent error, or later transition play.

Because official lineup positions were not available, rotations were inferred from server order. A team's first server in a set was labeled R1; the label remained unchanged while the team continued to serve and advanced when the team regained service. The initial receiving team was assigned R6 so that its first side-out advanced to R1. These labels describe inferred service order rather than official front-row assignments or substitution states. Pressure was defined before the serve as a maximum score of at least 20 with an absolute score gap no greater than two.

Event reconstruction was checked against player-match box scores. Player records were mapped to contest teams through season rosters, and play-by-play events with direct statistical counterparts were aggregated to the team-match level. Six measures were compared: kills, total attacks, aces, service errors, digs, and ball-handling errors. Agreement was reported across 6,144 team-matches using exact-match percentage, mean absolute error, median absolute error, percentage within one event, and Pearson correlation. Kill counts were expected to differ more because the event stream separates First ball kill from the general Kill label, whereas the box score combines terminal attacks.

League proportions used 95% Wilson intervals [34]. Pressure differences were estimated with 2,000 contest-cluster bootstrap samples, and winner-loser paired differences used 3,000 match-level bootstrap samples [35]. Team rankings required Division I membership and at least 1,000 receive opportunities. Mantel-Haenszel odds ratios [36] compared rotations after stratification by team, pressure, and receiving score difference clipped to -5 through +5; pressure comparisons were stratified by team, rotation, and the same score bins. A 0.5 correction was applied only to zero cells. Tests were two-sided, and p values below 0.001 are reported as $p < 0.001$.

The evidence store contained one card for each contest, receiving team, and inferred rotation. Each card stored match metadata, sample size, side-out rate, reception rate, organized-attack rate, first-ball kill rate, conditional first-ball conversion, opponent service errors, aces conceded, and pressure performance. The resulting 60,852 cards were indexed as compact records. Ninety target cards with at least eight rallies each generated three query forms - explicit, natural-language, and tactical paraphrase - for a total of 270 test queries.

Four retrievers were evaluated. TF-IDF used unigram-bigram cosine similarity. BM25 used $k_1 = 1.5$ and $b = 0.75$. Team-filtered BM25 restricted candidates to the requested team. The structured hybrid retained the team filter and BM25 score, then added exact-match bonuses of 4 for opponent, 4 for date, and 3 for rotation. Performance was measured with Recall@1, Recall@5, MRR@10, nDCG@5, and median rank [18], [19].

Generation used 12 NCAA tournament team-match cases under three evidence conditions. The no-retrieval condition received only team, opponent, and outcome. Box-score RAG also received kills, attack errors, attempts, aces, service errors, digs, and reception errors. Chain-and-rotation RAG added overall chain rates, pressure performance, and six rotation summaries. GPT-5.5 Pro [37] produced one 130-180 word diagnosis and training review for each condition, yielding 36 reviews. The complete prompt set and generated reviews were retained for audit and scoring.

The evaluation rubric followed established RAG principles [20]. Numeric precision was the share of numerical claims that appeared in the supplied evidence. Coverage checked whether the review addressed box-score performance, the overall chain, the best rotation, and the weakest rotation. Specificity counted side-out, first-ball attack, rotation, reception, organization, pressure, serving, attacking, transition, and an explicit recommendation. Actionability required evidence-

linked recommendations across reception, setting or attack, serving, pressure, transition, and block-defense themes. Readability used Flesch Reading Ease [38]. The composite score weighted coverage at 35%, specificity at 25%, actionability at 20%, numeric precision at 10%, and length adherence at 10%; it was designed as an evidence rubric rather than a substitute for expert coaching judgment.

The final case examined Texas A&M's 3-0 victory over Kentucky in the NCAA championship match on December 21, 2025 [39]. The case used the same rally definitions, inferred rotations, and evidence-card construction as the season-wide analysis. Figure 1 summarizes the complete workflow from raw match records to tactical text.

End-to-end tactical review pipeline

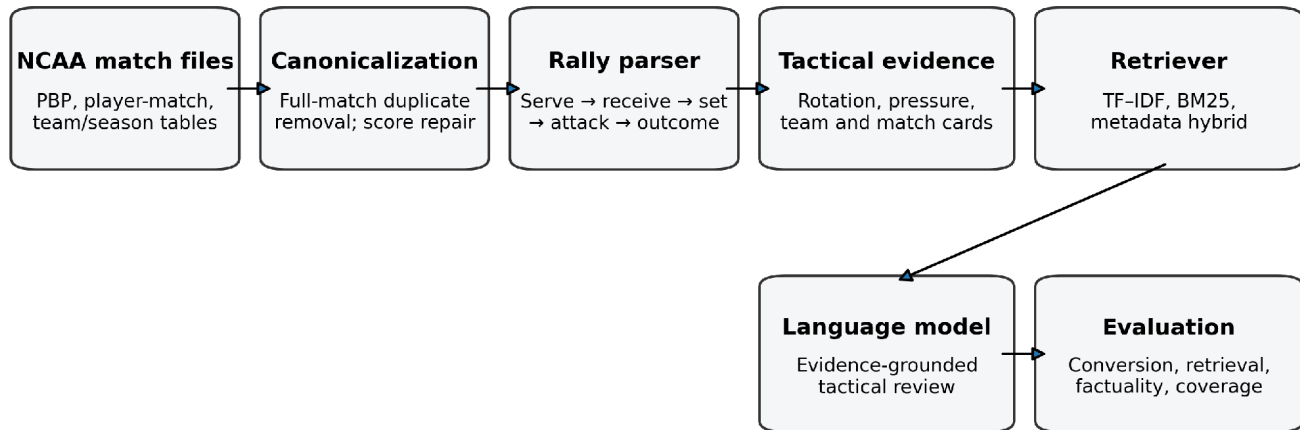


Figure 1. End-to-end workflow for canonicalizing NCAA events, reconstructing rallies, creating tactical evidence, retrieving match-rotation records, and evaluating generated reviews.

Results

Canonicalization and rally extraction produced a high-coverage season corpus. Table 1 reports the full audit trail. The source contained 7.51 million event rows; removing repeated complete logs reduced the corpus to 5.65 million canonical events. The parser identified 823,508 serves and resolved 821,923 valid tactical rallies, a success rate of 99.81%. Twenty-two contests had no rally that satisfied every validity rule, leaving 5,071 contests for match-level tactical aggregation. The valid corpus included 360 receiving teams because schedules contained Division I members and their opponents.

Table 1. Corpus audit and rally reconstruction coverage.

Measure	Value	Interpretation
Source event rows	7,505,822	Before repeated full-match runs were removed
Canonical event rows	5,654,956	One ordered log retained per contest
Repeated full-match runs removed	1,684	All copies passed length and dual-hash verification
Unique contests	5,093	Season match identifiers
Contest-sets	19,031	Observed set identifiers
Extracted rallies	823,508	One rally per Serve event
Valid tactical rallies	821,923	99.81% of extracted rallies
Contests with ≥ 1 valid rally	5,071	Used in tactical aggregation
Teams in valid rally corpus	360	Includes Division I teams and scheduled opponents
Analysis dates	2025-08-22 to 2025-12-21	2025 women's Division I season

A contest can remain in the source audit while contributing no rally that passed every tactical validity rule.

The event distribution in Table 2 is consistent with a contact-by-contact representation rather than a score-only feed. Every rally contained one initiating serve. The corpus included 1,306,183 Attack events and 1,294,174 Set events, equivalent to approximately 159 and 157 events per 100 rallies because longer rallies contained multiple offensive sequences. Reception appeared 83.35 times per 100 rallies. The source also recorded 240,757 First ball kill events, 56,138 aces, and

80,810 service errors. Challenge requests and error events remained in the canonical stream but did not replace score-based point assignment.

Table 2. Distribution of canonical play-by-play event labels.

Event	Count	Per 100 rallies
Serve	823,508	100.00
Reception	686,424	83.35
Set	1,294,174	157.15
Attack	1,306,183	158.61
First ball kill	240,757	29.24
Kill	220,703	26.80
Attack error	118,900	14.44
Dig	536,180	65.11
Block	231,631	28.13
Ace	56,138	6.82
Service error	80,810	9.81
Set error	26,604	3.23
Block error	13,235	1.61
Challenge request	12,788	1.55
Ball handling error	3,249	0.39

Multiple sets, attacks, digs, and blocks can occur in one rally; therefore their rates can exceed 100 per 100 rallies.

Agreement with player-match statistics was strong for events that shared a direct definition (Table 3). Pearson correlations ranged from 0.963 to 0.999. Aces matched exactly in 98.65% of team-matches, service errors in 97.92%, digs in 97.41%, and total attacks in 96.60%. Total attacks differed by only 0.80 actions per team-match on average. Kills had lower exact agreement (35.79%) and a mean absolute error of 1.58 while retaining $r = 0.965$; the small directional gap is consistent with the event stream's separate terminal labels and the combined box-score kill field. Ball-handling errors showed $r = 0.999$ despite a higher aggregate play-by-play total, reflecting sparse team-match variation around a small count.

Table 3. Team-match agreement between play-by-play counts and player-match box scores.

Metric	Team-matches	Box total	PBP total	Exact %	MAE	Median AE	r	Within 1 %
Kills	6,144	285,191	275,583	35.79	1.578	1.0	0.965	64.26
Total Attacks	6,144	787,907	783,475	96.60	0.801	0.0	0.963	98.36
Aces	6,144	33,982	33,797	98.65	0.031	0.0	0.992	99.48
SErr	6,144	48,964	48,625	97.92	0.058	0.0	0.982	99.27
Digs	6,144	323,368	321,492	97.41	0.326	0.0	0.976	98.26
BHE	6,144	975	1,943	87.73	0.158	0.0	0.999	97.33

PBP = play-by-play; MAE = mean absolute error; AE = absolute error. The Kill comparison does not merge the source's separate First ball kill label into the general Kill event.

The season conversion funnel is reported in Table 4 and Figure 2. Of 821,923 receive opportunities, 741,251 serves remained in play (90.18%). A reception was recorded on 684,688 rallies, or 92.37% of live serves. The receiving team completed an ordered reception-set-attack chain on 628,859 rallies, representing 76.51% of all opportunities and 91.85% of rallies with a recorded reception. A strict immediate kill followed 237,350 complete chains, so 37.74% of organized first attacks ended the rally. The broader side-out rate was 57.13%. The 28.26-point gap between strict first-ball kills and all side-outs captures opponent service errors, other opponent faults, and extended-rally wins.

Table 4. Serve-receive-attack conversion funnel.

Stage	Rallies	% of receive opportunities	% of previous stage
Valid receive opportunities	821,923	100.00	100.00
Serve in play	741,251	90.18	90.18
Reception recorded	684,688	83.30	92.37

Stage	Rallies	% of receive opportunities	% of previous stage
Organized receive-set-attack	628,859	76.51	91.85
Immediate kill after complete chain	237,350	28.88	37.74
All side-outs (parallel outcome)	469,584	57.13	-

All side-outs are a parallel final outcome and include free points and extended-rally wins, not only immediate first-ball kills.

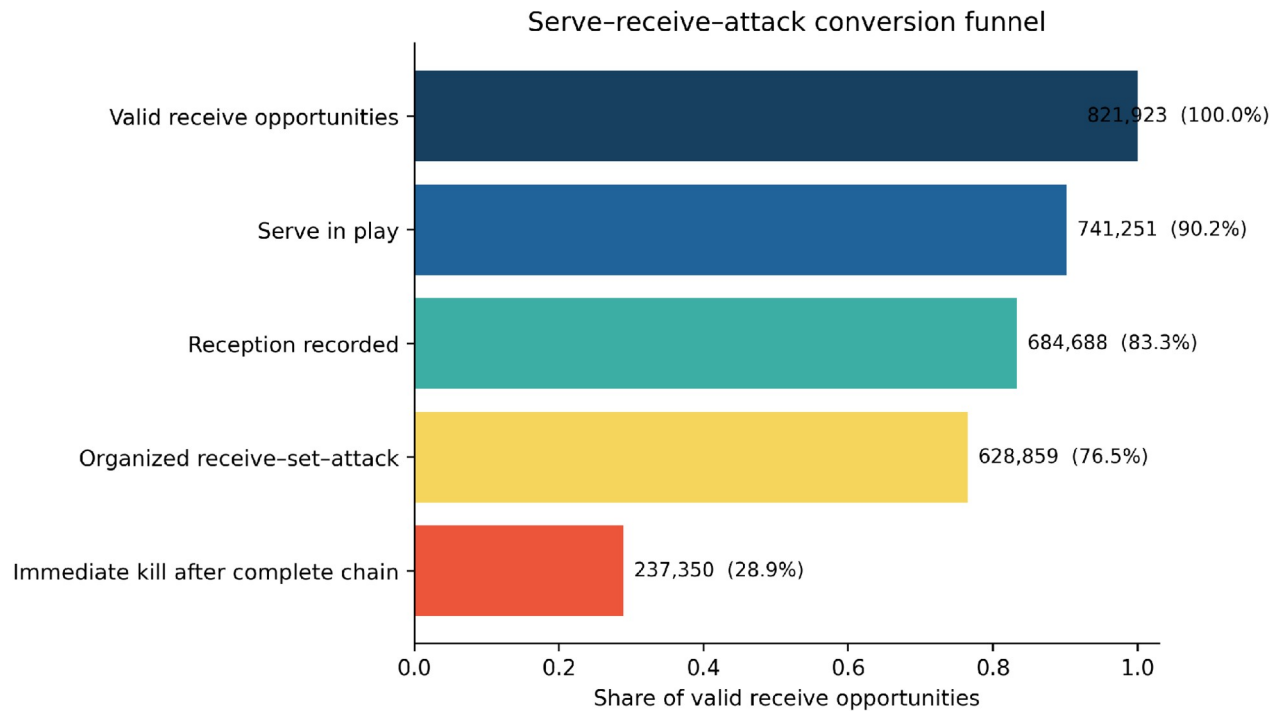


Figure 2. League conversion funnel from valid receive opportunity to organized first attack and immediate kill. Side-out is shown as the broader outcome.

Rotation summaries revealed different access and finishing profiles (Table 5 and Figure 3). R6 had the highest serve-in-play rate (91.09%), reception rate (84.17%), and organized-attack rate (77.43%), yet it had the lowest side-out rate at 56.15%. R2 recorded the highest side-out rate (57.64%) and first-ball kill rate (29.98%). R4 had the weakest reception (82.42%) and organized-attack (75.21%) rates, but its 57.46% side-out rate exceeded R6. Conditional first-ball conversion varied by less than two percentage points across rotations. The pattern shows that a rotation can create first attacks reliably without converting the full rally at the same rate.

Table 5. League performance by inferred receiving rotation.

Rotation	Rallies	Serve in play %	Reception %	Organized %	FBK %	FBK organized %	Side-out % (95% CI)	Pressure n
R1	159,714	90.20	83.16	76.66	29.26	37.77	56.86 (56.62-57.11)	11,645
R2	145,552	90.27	83.58	77.12	29.98	38.45	57.64 (57.38-57.89)	9,838
R3	135,296	89.86	82.89	76.03	29.05	37.73	57.39 (57.13-57.66)	7,884
R4	125,109	89.41	82.42	75.21	28.75	37.59	57.46 (57.19-57.73)	7,201
R5	115,520	90.17	83.52	76.38	29.37	37.86	57.41 (57.13-57.70)	8,582
R6	140,732	91.09	84.17	77.43	29.05	37.04	56.15 (55.89-56.40)	11,148

Rotation labels are inferred from service order; FBK = first-ball kill. Side-out confidence intervals are Wilson intervals.

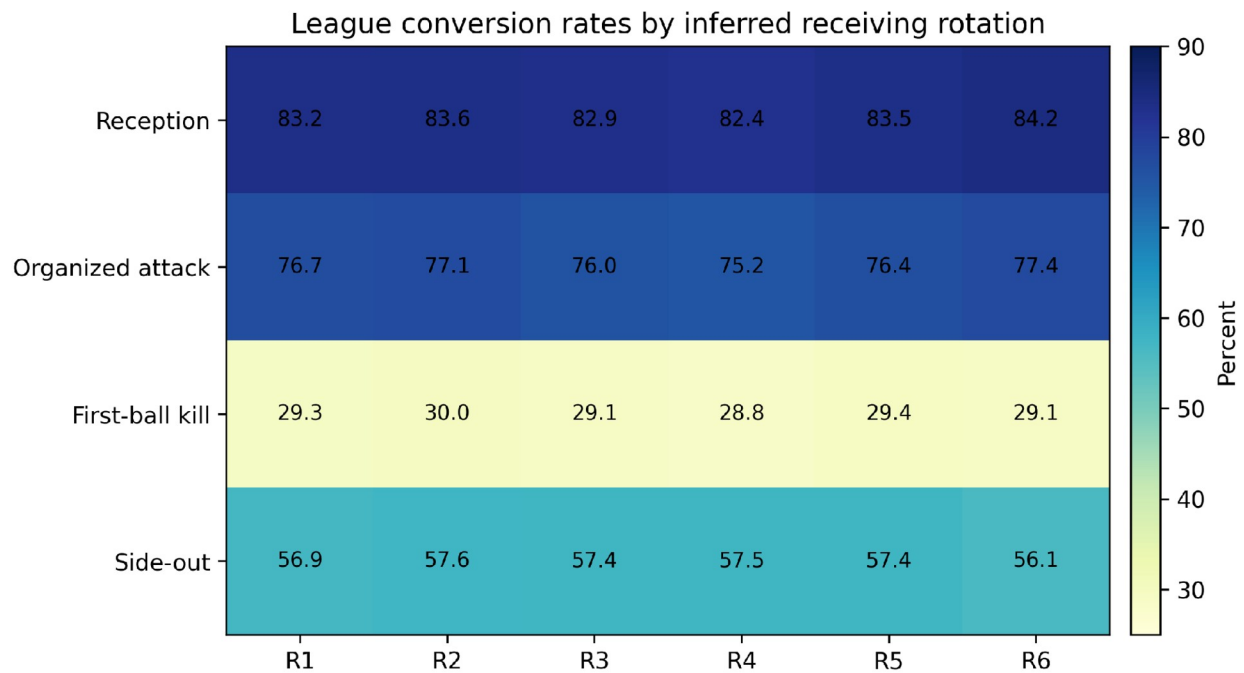


Figure 3. Reception, organized-attack, first-ball kill, and side-out rates by inferred receiving rotation.

The stratified analysis in Table 6 preserved the rotation contrast. Relative to R6, R4 had lower organized-attack odds (OR = 0.890, 95% CI 0.874-0.906) but higher side-out odds (OR = 1.056, 95% CI 1.040-1.073). R2 had similar organization odds (OR = 0.985, $p = 0.102$), yet it showed the strongest conditional first-ball finishing (OR = 1.060) and side-out odds (OR = 1.060). R1, R3, and R5 also had higher adjusted side-out odds than R6, indicating that team mix, pressure, and score differential did not account for R6's lower conversion.

Table 6. Mantel-Haenszel odds ratios for rotation and pressure comparisons.

Outcome	Predictor	OR	95% CI	p	Strata	Rallies
Organized attack	R1 vs R6	0.963	0.947-0.980	<0.001	5,617	300,366
Organized attack	R2 vs R6	0.985	0.968-1.003	0.102	5,595	286,119
Organized attack	R3 vs R6	0.931	0.915-0.948	<0.001	5,563	275,677
Organized attack	R4 vs R6	0.890	0.874-0.906	<0.001	5,571	265,557
Organized attack	R5 vs R6	0.948	0.931-0.966	<0.001	5,578	256,084
Organized attack	Pressure vs normal	1.104	1.081-1.127	<0.001	10,316	427,754
First-ball kill organized	R1 vs R6	1.026	1.009-1.044	0.003	5,579	231,272
First-ball kill organized	R2 vs R6	1.060	1.042-1.079	<0.001	5,553	220,957
First-ball kill organized	R3 vs R6	1.032	1.013-1.050	<0.001	5,491	211,308
First-ball kill organized	R4 vs R6	1.028	1.009-1.047	0.004	5,499	202,630
First-ball kill organized	R5 vs R6	1.038	1.018-1.057	<0.001	5,529	196,932
First-ball kill organized	Pressure vs normal	1.044	1.023-1.065	<0.001	10,155	326,369
Side-out	R1 vs R6	1.021	1.007-1.037	0.005	5,617	300,366
Side-out	R2 vs R6	1.060	1.044-1.076	<0.001	5,595	286,119
Side-out	R3 vs R6	1.051	1.035-1.067	<0.001	5,563	275,677
Side-out	R4 vs R6	1.056	1.040-1.073	<0.001	5,571	265,557
Side-out	R5 vs R6	1.057	1.040-1.074	<0.001	5,578	256,084
Side-out	Pressure vs normal	1.012	0.994-1.030	0.198	10,316	427,754

R6 is the reference rotation. Rotation comparisons are stratified by team, pressure, and score bin; pressure comparisons are stratified by team, rotation, and score bin.

Team-level results were strongly ordered by first-ball conversion (Table 7 and Figure 4). Among 348 eligible Division I teams, Nebraska ranked first in side-out at 70.31% and also led the listed teams in first-ball kills at 38.57%. SMU, Wisconsin, Stanford, Kentucky, Western Kentucky, and Texas A&M each exceeded 65.7% side-out. Kentucky recorded the highest reception rate among the top 12 teams (87.24%) and the highest organized-attack rate (81.20%). Texas A&M paired a 65.78% side-out rate with a 38.14% first-ball kill rate. The scatterplot shows a strong positive season-level relationship, although teams with similar first-ball kill rates still differed in total side-out, leaving room for free-point management and extended-rally conversion.

Table 7. Top 12 Division I teams by season side-out rate (minimum 1,000 receive opportunities).

Rank	Team	Conf.	Matches	Rallies	Side-out %	FBK %	Organized %	Reception %	Pressure n	Pressure SO %
1	Nebraska	Big Ten	34	1,960	70.31	38.57	75.05	80.56	90	74.44
2	SMU	ACC	33	2,425	66.89	38.10	77.69	82.76	188	59.04
3	Wisconsin	Big Ten	33	2,202	66.39	37.92	77.84	84.88	167	71.26
4	Stanford	ACC	34	2,424	66.25	37.46	79.08	84.53	137	67.88
5	Kentucky	SEC	33	2,383	66.22	37.05	81.20	87.24	199	66.33
6	Western Ky.	CUSA	33	2,156	66.00	37.20	77.09	84.23	142	73.94
7	Texas A&M	SEC	33	2,373	65.78	38.14	79.22	84.66	176	64.20
8	Pittsburgh	ACC	35	2,468	65.64	35.33	76.50	81.89	172	60.47
9	Texas	SEC	30	2,199	65.48	36.97	78.67	83.81	176	67.61
10	Creighton	Big East	34	2,287	64.76	35.85	78.44	84.48	159	61.01
11	Arizona St.	Big 12	32	2,393	64.69	36.15	77.52	82.20	179	62.57
12	Cal Poly	Big West	35	2,556	64.28	33.53	75.98	82.55	165	66.06

Team eligibility uses the Division I team-season table. Opponent service errors remain part of side-out because they end a receive opportunity with a receiving-team point.

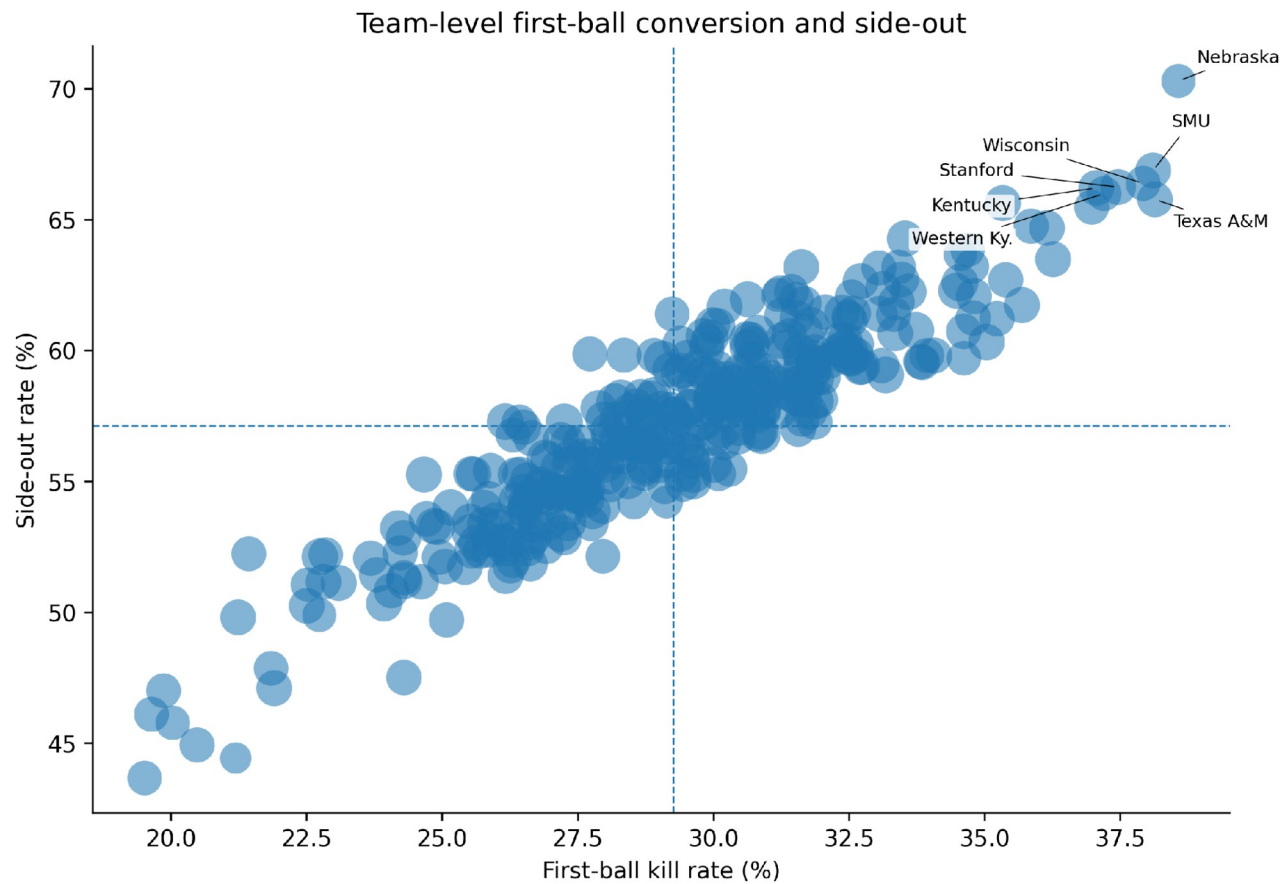


Figure 4. Team-level first-ball kill and side-out rates for 348 eligible Division I teams. Bubble area reflects receive opportunities; dashed lines show unweighted team means.

Pressure rallies produced a clear unadjusted pattern (Table 8 and Figure 5). Side-out increased from 57.05% in normal rallies to 58.22% under the pressure definition, a difference of 1.17 percentage points (95% contest-cluster bootstrap CI 0.74-1.61). Organized attack rose by 3.82 points and first-ball kills rose by 2.05 points. Opponent service errors fell by 1.89 points and aces conceded fell by 1.16 points, indicating more controlled late-score serving. The stratified model changed the interpretation of the final outcome: after receiving team, rotation, and score bin were held constant, pressure had OR = 1.012 (95% CI 0.994-1.030, $p = 0.198$) for side-out. Organized attack and conditional first-ball conversion retained positive adjusted effects, with OR = 1.104 and OR = 1.044, respectively.

Table 8. Unadjusted pressure-rally differences with contest-cluster bootstrap intervals.

Metric	Normal %	Pressure %	Difference, pp (95% CI)	p
Side-out	57.05	58.22	1.17 (0.74 to 1.61)	<0.001
Organized attack	76.25	80.07	3.82 (3.43 to 4.19)	<0.001
First-ball kill	29.12	31.17	2.05 (1.66 to 2.45)	<0.001
Opponent service error	9.94	8.05	-1.89 (-2.12 to -1.66)	<0.001
Ace conceded	6.90	5.74	-1.16 (-1.37 to -0.95)	<0.001

Pressure rallies begin with a maximum score of at least 20 and a score gap no larger than two points. pp = percentage points.

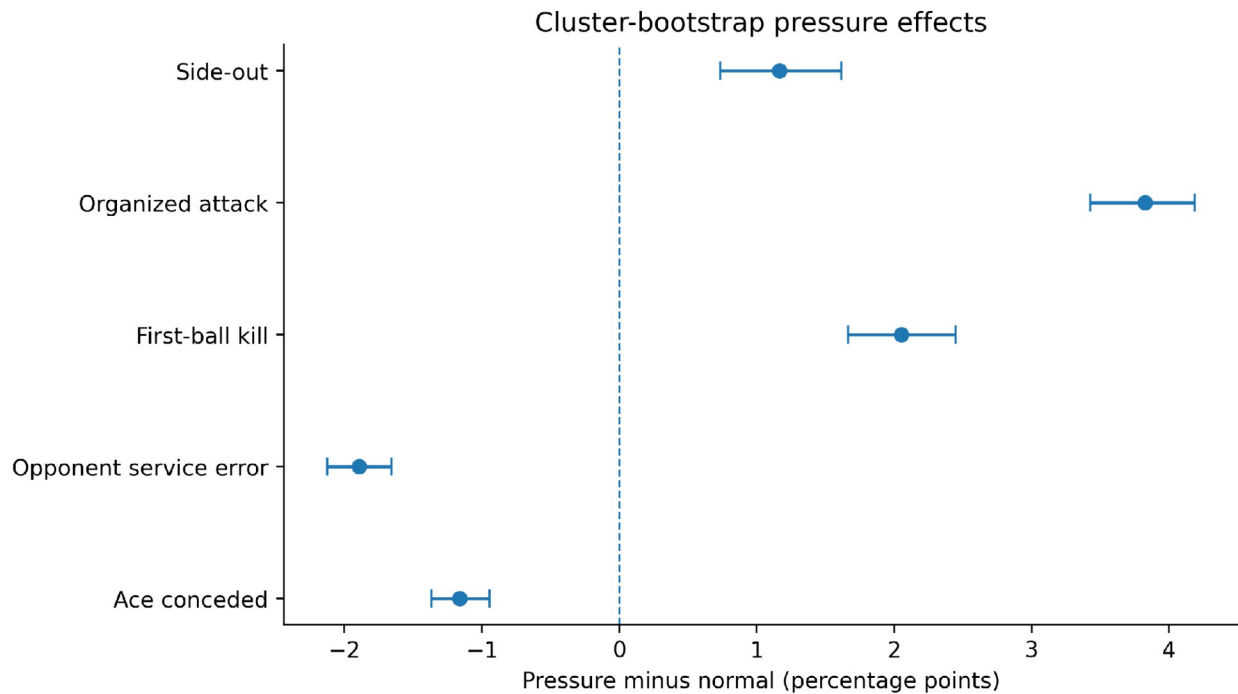


Figure 5. Pressure-minus-normal differences with 95% contest-cluster bootstrap intervals.

Match winners separated from losers most strongly in conversion rather than reception access (Table 9). Across 5,071 paired matches, winners averaged 63.84% side-out and losers 51.82%, a 12.02-point difference (95% CI 11.77-12.27). The first-ball kill difference was 7.56 points, compared with 2.18 points for organized attack and 0.82 points for recorded reception. Winners also received 1.02 points more service-error help and conceded 1.85 points fewer aces. All paired bootstrap tests were $p < 0.001$. The widening gaps across the chain locate the largest competitive separation after reception has already been achieved.

Table 9. Paired winner-loser differences across valid matches.

Metric	Matches	Winner %	Loser %	Paired difference, pp (95% CI)	p
Side-out	5,071	63.84	51.82	12.02 (11.77 to 12.27)	<0.001
Organized attack	5,071	77.59	75.40	2.18 (1.95 to 2.40)	<0.001
First-ball kill	5,071	33.44	25.88	7.56 (7.31 to 7.80)	<0.001
Reception recorded	5,071	83.63	82.82	0.82 (0.62 to 1.02)	<0.001
Opponent service error	5,071	10.46	9.44	1.02 (0.87 to 1.17)	<0.001
Ace conceded	5,071	5.84	7.69	-1.85 (-1.98 to -1.73)	<0.001

Intervals and p values are based on 3,000 paired match bootstrap samples.

Table 10 and Figure 6 compare retrieval performance. TF-IDF reached 30.37% Recall@1 and 61.48% Recall@5. BM25 reached 27.41% and 70.74%, with a median rank of two. Team filtering increased Recall@1 to 66.30%, Recall@5 to 87.04%, and MRR@10 to 0.754. The structured hybrid retrieved 269 of 270 targets at rank one (99.63%) and every target by rank five, with nDCG@5 = 0.999. Its only rank-one miss occurred for a tactical paraphrase. The result shows that exact opponent, date, and rotation routing resolved ambiguity that shared volleyball vocabulary could not.

Table 10. Retrieval benchmark on 270 match-team-rotation queries.

Method	Queries	Recall@1 %	Recall@5 %	MRR@10	nDCG@5	Median rank
TF-IDF	270	30.37	61.48	0.447	0.471	3
BM25	270	27.41	70.74	0.476	0.530	2
Team-filtered BM25	270	66.30	87.04	0.754	0.771	1
Structured hybrid	270	99.63	100.00	0.998	0.999	1

The structured hybrid combines team filtering, BM25, and exact opponent, date, and rotation bonuses.

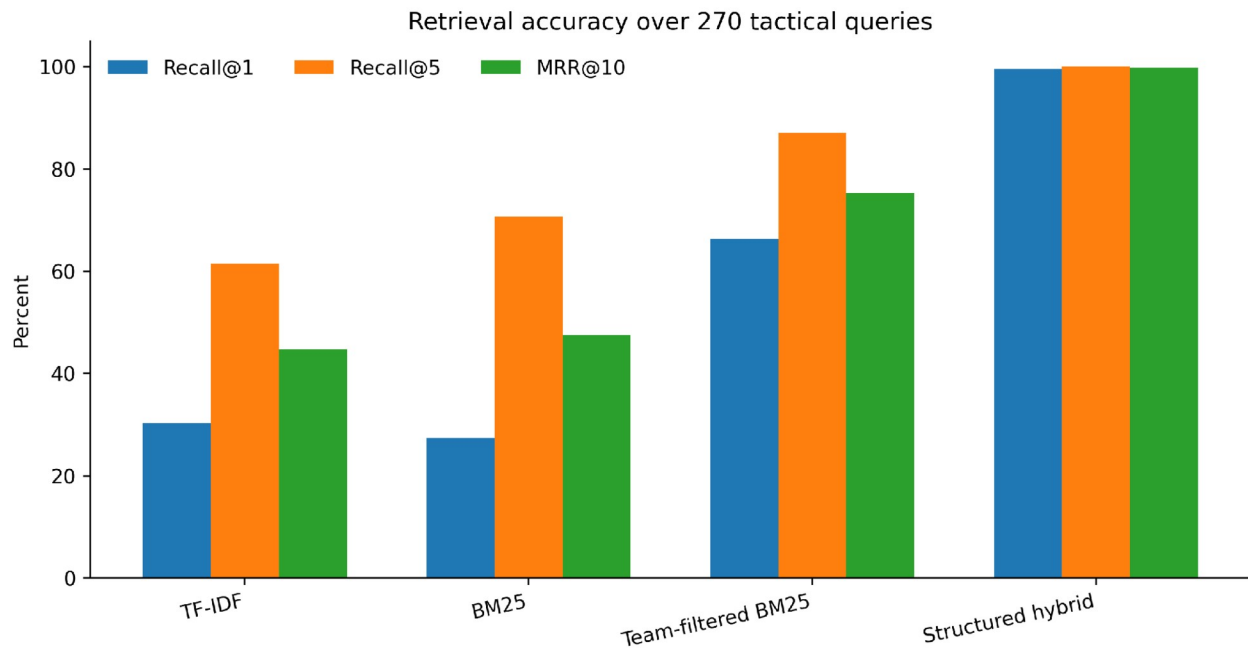


Figure 6. Ranked-retrieval performance for lexical, team-filtered, and structured hybrid methods.

The 36 tactical reviews showed distinct evidence profiles (Table 11 and Figure 7). No-retrieval reviews averaged 152.5 words, made no numerical claims, and covered no case-specific evidence, although generic coaching language produced 78.33% tactical specificity. Box-score RAG made 5.08 numerical claims per review with 100% precision and 27.78% evidence coverage, but only 61.67% specificity. Chain-and-rotation RAG made 15.08 numerical claims, retained 100% precision, and reached 53.33% coverage, 88.33% specificity, 36.76% grounded actionability, and a 68.10 composite score. Its Flesch score increased to 51.56 because the reviews were shorter and organized around evidence-based diagnoses.

Table 11. Automated quality evaluation of 36 tactical reviews.

Condition	Reviews	Words	Numeric claims	Numeric precision %	Evidence coverage %	Tactical specificity %	Grounded actionability %	Flesch	Composite
No retrieval	12	152.50	0.00	-	0.00	78.33	28.10	40.09	35.20
Box-score RAG	12	148.25	5.08	100.00	27.78	61.67	28.33	39.49	50.81
Chain+rotation RAG	12	139.42	15.08	100.00	53.33	88.33	36.76	51.56	68.10

No-retrieval reviews made no numerical claims, so numeric precision is not defined. The composite score ranges from 0 to 100.

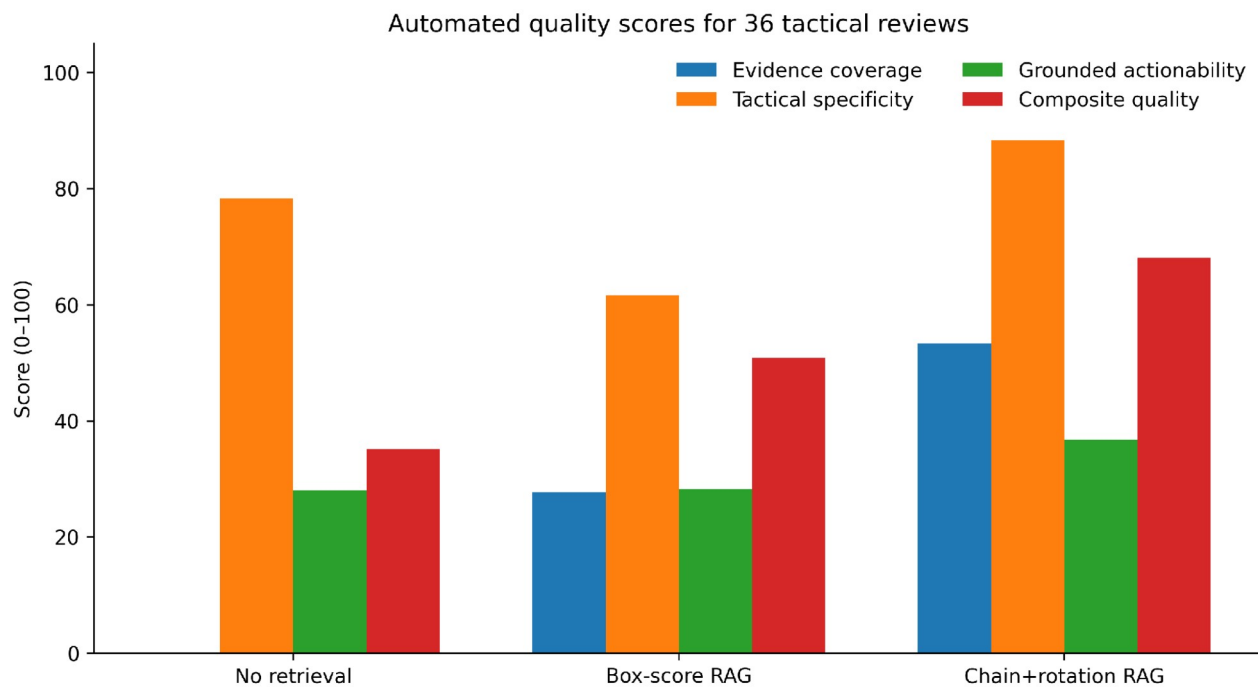


Figure 7. Evidence coverage, tactical specificity, grounded actionability, and composite quality across the three generation conditions.

The championship final illustrates why chain stages and final outcomes must be interpreted together (Table 12 and Figure 8). Texas A&M side-outed on 43 of 60 opportunities (71.67%), while Kentucky won 42 of 75 receiving rallies (56.00%). Their organized-attack rates were nearly identical at 78.33% and 77.33%. Kentucky recorded the higher first-ball kill rate, 38.67% versus 23.33%. Texas A&M, however, received nine Kentucky service errors, conceded only one ace, and converted side-out well beyond its first-ball kills. Kentucky received five service errors and conceded four aces. Rotation detail further localized the result: Kentucky reached 77.78% side-out in R3 and R4 but only 43.75% in R6, while Texas A&M reached at least 70% in five of six rotations. Similar organized access therefore became a 15.67-point side-out advantage through the serve-error exchange and extended-rally conversion.

Table 12. Serve-receive-attack summary for the 2025 NCAA championship final.

Receiving team	Rallies	Side-outs	Side-out %	Organized n	Organized %	FBK n	FBK %	Opp. SE	Aces conceded	Pressure n
Kentucky	75	42	56.00	58	77.33	29	38.67	5	4	7
Texas A&M	60	43	71.67	47	78.33	14	23.33	9	1	5

Opp. SE = opponent service errors. Texas A&M defeated Kentucky 3-0 on December 21, 2025.

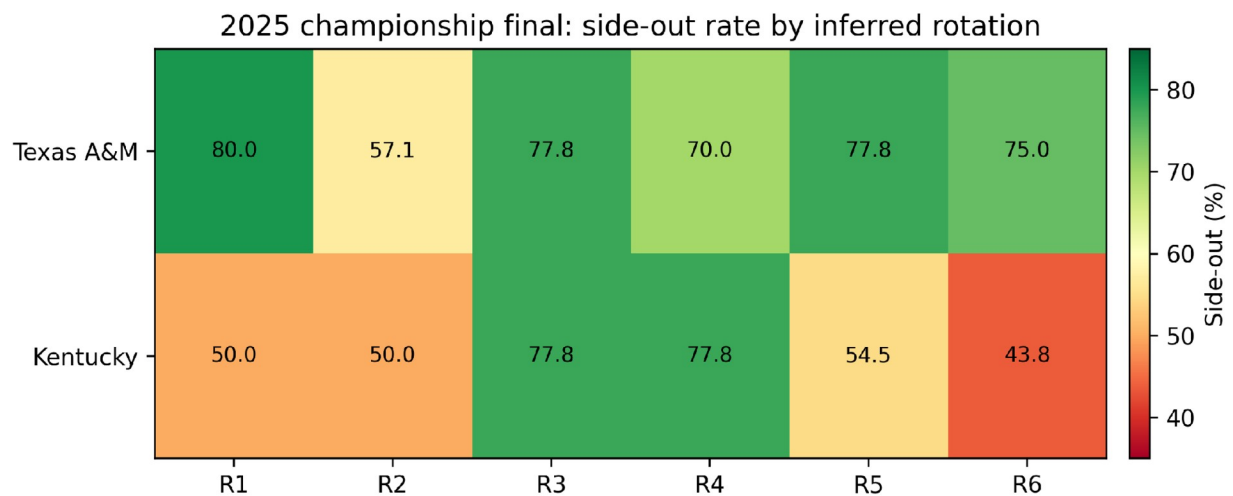


Figure 8. Side-out rate by inferred receiving rotation in the 2025 NCAA championship final.

Discussion

The season-scale results show that serve-receive-attack reconstruction adds information that conventional match summaries cannot preserve. The overall side-out rate was 57.13%, whereas strict immediate kills after a complete recorded chain accounted for 28.88% of receive opportunities. More than half of successful side-outs therefore came from opponent service errors, other opponent faults, or rallies that continued beyond the first attack. A box score can count these events, but it cannot reliably assign them to the receiving rotation and preceding contact sequence. The reconstructed funnel identifies whether the offense failed to gain access, failed to organize, failed to terminate, or recovered later in the rally, which is consistent with prior work on reception, setting, and attack interactions [7]–[11].

The winner-loser comparison places first-ball offense and complete side-out at the center of match separation. Winners held only a 0.82-point advantage in recorded reception, but the difference widened to 2.18 points in organized attack, 7.56 points in first-ball kills, and 12.02 points in side-out. Merely recording a reception was not the decisive threshold. Teams separated themselves by converting controlled contacts into terminal or recoverable attacks and by winning the rallies that continued. This process interpretation complements earlier evidence that victory depends on scoring skills and error control [12].

Rotation results demonstrate why tactical review should not rank rotations by a single percentage. R6 led the league in serve-in-play exposure, reception, and organized offense but finished last in side-out. R4 showed the opposite profile: weaker access and organization but better complete conversion. The adjusted models preserved that contrast. For coaching purposes, R6 is not accurately described as a passing failure; the evidence points toward first-swing finishing or extended-rally vulnerability after a frequently organized attack. R4 calls for a different intervention because the deficit appears before the attack. Although inferred service order cannot identify every player matchup, it narrows video review to the rotation and chain stage most likely to explain the result.

Pressure analysis also shows the importance of separating pooled rates from adjusted comparisons. The unadjusted pressure subset had better side-out, organization, and first-ball kill rates, along with fewer service errors and aces. After stratification by team, rotation, and score difference, the side-out advantage was no longer statistically significant. The raw increase therefore reflected which teams and rotations reached the defined late-score state as well as the state itself. Organized attack and conditional first-ball conversion remained positive, suggesting more controlled execution without a uniform increase in the final probability of side-out.

Retrieval performance was primarily an entity-resolution problem. TF-IDF and BM25 recognized volleyball vocabulary, but thousands of evidence cards shared the same terms. Team filtering more than doubled rank-one accuracy, and exact opponent, date, and rotation routing increased Recall@1 to 99.63%. This result supports a general design principle found in structured disclosure QA, reranking, code intelligence, and long-document analysis: explicit fields should constrain the candidate set before semantic similarity resolves the remaining ambiguity [26], [27], [32], [33]. Dense retrieval remains useful for broad or underspecified questions, but it should not replace exact metadata already available in the evidence store.

Richer evidence also changed the substance of generation. No-retrieval reviews sounded plausible but could not establish a match-specific diagnosis. Box scores added accurate numbers yet remained limited to aggregate outcomes. Chain-and-rotation evidence restored stage and rotation specificity and supported different prescriptions for disrupted access, weak first-swing conversion, serving exchange, and transition play. The 100% numeric precision observed in both retrieval

conditions is consistent with the value of explicit evidence checks and provenance controls reported in enterprise, financial, and evidence-chain RAG systems [21]-[23], [29]-[31]. The evaluation rubric measures claim matching and tactical breadth; expert coaches must still judge whether a recommendation fits roster roles, opponent tendencies, and video context.

The championship match provides a concrete test of the pipeline's tactical logic. Kentucky led Texas A&M in first-ball kill rate and produced nearly the same organized-attack rate, so a first-contact-only explanation does not fit the match. Texas A&M received four more service-error points, conceded three fewer aces, and won 29 receiving rallies that were not first-ball kills, compared with 13 for Kentucky. The rotation heatmap adds stability information: Texas A&M reached at least 70% side-out in five rotations, while Kentucky's R6 remained below 44%. The resulting diagnosis prioritizes reducing free points, stabilizing the weakest exit rotation, and improving extended-rally conversion rather than rebuilding an offense that already created and finished many first attacks.

Four limitations bound the conclusions. First, the complete event experiment covers one Division I women's season. Second, the event labels omit ball trajectories, reception grades, set tempo, attack zone, block formation, and official lineups, so rotation labels are inferred from service order. Third, the associations are observational rather than causal. Fourth, the retrieval benchmark targets known match-team-rotation cards, and the text evaluation uses an evidence rubric rather than blinded coach ratings. Broader seasons, official lineup data, attack coordinates, video timestamps, semantic query sets, and human inter-rater studies would extend the validation.

Within those limits, the workflow provides a practical architecture for NCAA tactical review. Canonicalization prevents repeated match logs from inflating totals. Score-based rally resolution supplies a stable point outcome. Stage-specific indicators connect access, organization, termination, and final side-out. Structured retrieval selects the correct evidence at season scale, and the language model converts that evidence into a concise diagnosis whose numerical claims can be traced to a match-rotation card. The same architecture can incorporate reception grades, official lineups, attack coordinates, and linked video when those fields are available.

Conclusions

A serve-anchored parser reconstructed 821,923 valid NCAA Division I women's volleyball rallies and converted them into a measured serve-serve-attack chain. League side-out was 57.13%, organized attack was 76.51%, and source-labeled first-ball kills occurred on 29.26% of receive opportunities; 28.88% met the strict complete-chain definition. Match winners held a 12.02-point side-out advantage and a 7.56-point first-ball kill advantage over losers. Rotation and pressure analyses showed that access, organization, finishing, and final rally outcome do not move in parallel: R6 organized offense most often but side-outed least, and the pooled pressure advantage disappeared after team, rotation, and score stratification.

The RAG evaluation showed that exact tactical metadata is essential for retrieval. Structured hybrid routing achieved 99.63% Recall@1 and 100% Recall@5 across 270 queries. Chain-and-rotation evidence produced the strongest reviews, with 100% numeric precision, 53.33% evidence coverage, 88.33% tactical specificity, and a 68.10 composite score. In the 2025 championship final, the method separated Kentucky's stronger first-ball finishing from Texas A&M's superior total side-out and identified serving exchange and extended-rally conversion as the decisive layers. The resulting workflow supports traceable tactical review at the match, rotation, and possession-chain levels.

References

- [1] J. R. Stevens, "Extract Data from NCAA Women's and Men's Volleyball Website," R package version 0.5.1, 2026. [Online]. Available: <https://github.com/JeffreyRStevens/ncaavolleyballr>
- [2] National Collegiate Athletic Association, 2022 Volleyball Statisticians' Manual. Indianapolis, IN, USA: NCAA, 2022. [Online]. Available: https://fs.ncaa.org/Docs/stats/Stats_Manuals/VB/2022.pdf
- [3] S. Powers, L. Stancil, and N. Consiglio, "Estimating individual contributions to team success in women's college volleyball," *Journal of Quantitative Analysis in Sports*, vol. 21, no. 2, pp. 117-135, 2025, doi: 10.1515/jqas-2024-0038.
- [4] J. M. Palao, "Side-out success and ways that points are obtained in women's college volleyball," *Journal of Sports Analytics*, vol. 4, no. 4, pp. 243-250, 2018, doi: 10.3233/JSA-180153.
- [5] J. M. Palao and I. Ahrabi-Fard, "Side-out success in relation to setter's position on court in women's college volleyball," *International Journal of Applied Sports Sciences*, vol. 23, no. 1, pp. 155-167, 2011, doi: 10.24985/ijass.2011.23.1.155.
- [6] Jubin Zhang, "Interpretable Skill Prioritization for Volleyball Education via Team-Stat Modeling," *JACS*, vol. 3, no. 3, pp. 34-49, Mar. 2023, doi: 10.69987/JACS.2023.30304.
- [7] J. Afonso, I. Mesquita, R. Marcelino, and J. A. da Silva, "Analysis of the setter's tactical action in high-performance women's volleyball," *Kinesiology*, vol. 42, no. 1, pp. 82-89, 2010.
- [8] J. Afonso, F. Esteves, R. Araujo, L. Thomas, and I. Mesquita, "Tactical determinants of setting zone in elite men's volleyball," *Journal of Sports Science and Medicine*, vol. 11, no. 1, pp. 64-70, 2012.

- [9] G. C. T. Costa, J. Afonso, R. V. Barbosa, P. Coutinho, and I. Mesquita, "Predictors of attack efficacy and attack type in high-level Brazilian women's volleyball," *Kinesiology*, vol. 46, no. 2, pp. 242-248, 2014.
- [10] J. Gonzalez-Silva, C. Fernandez-Echeverria, M. Conejero, and M. P. Moreno, "Characteristics of serve, reception and set that determine the setting efficacy in men's volleyball," *Frontiers in Psychology*, vol. 11, Art. no. 222, 2020, doi: 10.3389/fpsyg.2020.00222.
- [11] R. M. Araujo, L. Tosini, A. B. Freire, G. C. T. Costa, and C. M. Meira Jr., "Reception-attack relation in men's and women's volleyball during the Rio 2016 Olympics," *Journal of Physical Education and Sport*, vol. 20, suppl. 3, pp. 2008-2012, 2020, doi: 10.7752/jpes.2020.s3271.
- [12] M. Silva, D. Lacerda, and P. V. Joao, "Game-related volleyball skills that influence victory," *Journal of Human Kinetics*, vol. 41, pp. 173-179, 2014, doi: 10.2478/hukin-2014-0045.
- [13] L. Laporta, J. Afonso, and I. Mesquita, "Interaction network analysis of the six game complexes in high-level volleyball through the use of Eigenvector Centrality," *PLOS ONE*, vol. 13, no. 9, Art. no. e0203348, 2018, doi: 10.1371/journal.pone.0203348.
- [14] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems* 33, 2020, pp. 9459-9474.
- [15] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing*, 2020, pp. 6769-6781, doi: 10.18653/v1/2020.emnlp-main.550.
- [16] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613-620, 1975, doi: 10.1145/361219.361220.
- [17] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333-389, 2009, doi: 10.1561/1500000019.
- [18] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422-446, 2002, doi: 10.1145/582415.582418.
- [19] Jubin Zhang, "Tactical Language + AI Tutoring from Structured Volleyball Rally Logs: Reproducible Experiments on NCAA Play-by-Play," *JACS*, vol. 4, no. 1, pp. 58-66, Jan. 2024, doi: 10.69987/JACS.2024.40105.
- [20] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, "RAGAs: Automated evaluation of retrieval augmented generation," in *Proc. 18th Conf. European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2024, pp. 150-158.
- [21] C. Li, W. Su, and E. Zhang, "Lightweight Hallucination Firewall for Enterprise LLM Applications: Evidence Consistency, Self-Checking, and Small-Model Detection on TruthfulQA," *Journal of Advanced Computing Systems*, vol. 3, no. 1, pp. 49-65, Jan. 2023, doi: 10.69987/JACS.2023.30104.
- [22] K. Zhang, S. Meng, and E. Zhou, "Evidence-Grounded Trading Desk Risk Memos over SEC Filings: Retrieval-Augmented Generation with XBRL Numeric Verification," *Journal of Advanced Computing Systems*, vol. 3, no. 2, pp. 60-76, Feb. 2023, doi: 10.69987/JACS.2023.30205.
- [23] Q. Wu, J. Bai, and X. Zhou, "Evidence-Grounded Financial RAG: Reducing Numerical Hallucination in LLM-Generated Corporate Risk Memos," *Journal of Advanced Computing Systems*, vol. 3, no. 3, pp. 65-84, Mar. 2023, doi: 10.69987/JACS.2023.30306.
- [24] J. Nie and D. Zheng, "Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal," *Journal of Advanced Computing Systems*, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.
- [25] Y. Li, "Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs," *Journal of Advanced Computing Systems*, vol. 3, no. 2, pp. 1-17, Feb. 2023, doi: 10.69987/JACS.2023.30201.
- [26] S. Zhou, Z. Li, and E. Wang, "Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards," *Journal of Advanced Computing Systems*, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.
- [27] X. Chang, T. Ye, and S. Luo, "LLM-as-Reranker for Personalized Recommendation: Popularity Bias Mitigation and Faithful Natural-Language Explanations on MovieLens 100K," *Journal of Advanced Computing Systems*, vol. 3, no. 8, pp. 61-78, Aug. 2023, doi: 10.69987/JACS.2023.30806.
- [28] Y. Li, "Test-in-the-loop LLM Repair: Verifiable Automated Program Repair on QuixBugs with a 'Failing Test → Patch → Regression Test' Loop," *Journal of Advanced Computing Systems*, vol. 4, no. 2, pp. 62-75, Feb. 2024, doi: 10.69987/JACS.2024.40206.
- [28] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," *Journal of Advanced Computing Systems*, vol. 4, no. 2, pp. 76-92, Feb. 2024, doi: 10.69987/JACS.2024.40207.
- [29] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," *Journal of Advanced Computing Systems*, vol. 4, no. 3, pp. 109-125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

- [30] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," *Journal of Advanced Computing Systems*, vol. 4, no. 4, pp. 97-111, Apr. 2024, doi: 10.69987/JACS.2024.40408.
- [31] Y. Li, "Findable then Explainable: Retrieval-Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," *Journal of Advanced Computing Systems*, vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.
- [32] S. Zhou, Z. Li, and E. Wang, "Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures," *Journal of Advanced Computing Systems*, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.
- [33] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," *Journal of the American Statistical Association*, vol. 22, no. 158, pp. 209-212, 1927, doi: 10.1080/01621459.1927.10502953.
- [34] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [35] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, vol. 22, no. 4, pp. 719-748, 1959.
- [36] OpenAI, "GPT-5.5 System Card," Apr. 24, 2026. [Online]. Available: <https://openai.com/index/gpt-5-5-system-card/>
- [37] R. Flesch, "A new readability yardstick," *Journal of Applied Psychology*, vol. 32, no. 3, pp. 221-233, 1948, doi: 10.1037/h0057532.
- [38] National Collegiate Athletic Association, "Texas A&M wins 2025 DI women's volleyball championship," Dec. 21, 2025. [Online]. Available: <https://www.ncaa.com/news/volleyball-women/article/2025-12-21/2025-ncaa-womens-volleyball-tournament-bracket-schedule-scores>