# SENTAKU: A Web-Based Tournament Management System with Zero-Shot LLM Referee Evaluation for Shorinji Kempo

**Troy*[1], Kathryn Widiyanti[2], Agni Saraswati[3]**

[1-2]Animation Study Program, Institut Seni Indonesia Yogyakarta, Indonesia

[3]Winchester School of Art, University of Southampton, United Kingdom

E-mail: troy@isi.ac.id[1], kathryn@isi.ac.id[2], a.saraswati@southampton.ac.uk[3]

**Abstract.** This paper presents SENTAKU (*Sistem Elektronik Pertandingan Kenshi Unggul*), a referee analytics platform for Shorinji Kempo built on a PHP-based competition management system. The core contribution is a structured referee analytics framework that (i) computes seven per-referee statistical indicators from the raw embu score matrix, (ii) submits them with a structured prompt to Llama 3.3 70B via the Groq API, and (iii) audits the generated report through a rule-based reproducible content audit of coverage, numerical grounding, and section completeness. Two research questions are addressed: whether the framework surfaces referee patterns invisible to manual workflows, and whether a zero-shot LLM with structured context produces faithful text. Deployment at POPDA DIY 2026 (5 contingents, 53 randori athletes, 10 judges, 111 matches) surfaced component-level calibration differences and same-regency patterns unattainable through paper workflows. The LLM report named all ten judges but cited no numerical value from the prompt and omitted two of five requested sections, while introducing no fabricated entities, events, or contradicted statistics. These findings suggest that absence of hallucination alone may not guarantee faithful LLM reporting: zero-shot prompting can fail through systematic under-utilisation of numerical input even when explicitly required. Schema-enforced decoding is motivated as the next step.

**Keywords:** tournament management system; referee evaluation; large language model; Shorinji Kempo; prompt engineering.

## 1. Introduction

Sports competition management requires coordinating parallel matches, multi-judge scoring panels, athlete verification, and bracket progression under time pressure. When these depend on manual workflows, transcription errors accumulate and the data generated by judged competitions remain unanalysed, particularly in martial arts using subjective scoring panels where digital infrastructure for referee accountability is largely absent [1].

Shorinji Kempo is a Japanese martial art practised competitively in Indonesia through PERKEMI (*Persaudaraan Shorinji Kempo Indonesia*). Its competitive format includes embu, a performed discipline evaluated by five independent judges on ten scoring components, and randori, a sparring discipline with point-based scoring. Paper score sheets are filed after each event without subsequent analysis of inter-judge consistency, scoring tendencies, or affiliation bias.

Referee bias in sports is extensively documented. Geographic or club affiliations can influence scoring through patterns invisible without digital data collection [2]; football research shows that systematic biases persist across competition cycles without technological accountability [3], and technological interventions such as VAR can reduce home-team bias [4]. In basketball, geographic affiliations between referees and teams have been shown to influence discretionary decisions in ways that aggregate evaluations never detect [5]. In subjective-scoring disciplines this is compounded by the absence of an objective ground truth; statistical divergence from panel consensus remains the only measurable indicator of potential bias [6]. Yet existing martial arts tournament software provides no mechanism for collecting or analysing this data.

AI adoption in sports officiating has accelerated. AI-assisted tools have demonstrated improvements in decision accuracy in visually verifiable disciplines [7], and Large Language Models (LLMs) have shown capabilities relevant to sports analytics, particularly structured data interpretation and narrative generation [8], [9]. Connor and O'Neill identified opportunities for LLMs in sport science, while cautioning that reliability and context specificity require careful system design [9]. Commercial platforms for martial arts tournaments such as Kihapp, Smoothcomp, and Scoring Wi-Fi Pro provide bracket management but, to the authors' knowledge based on public documentation, no referee analytics. Academic AI officiating research has focused on team sports using computer vision requiring specialised hardware incompatible with community-level events [10].

This gap motivates SENTAKU (*Sistem Elektronik Pertandingan Kenshi Unggul*), deployed at https://sentaku.shorinjikempo.or.id. The primary contribution is a structured referee analytics framework for a subjective-scoring martial art: it collects per-referee scoring records, computes seven performance indicators, generates a narrative evaluation through an LLM integration, and audits the report against its numerical input. The competition management layer (registration, bracket, scoring, scheduling, publication) is the enabling infrastructure. The research questions addressed are:

**RQ1**. Can a structured referee analytics framework that combines per-referee statistical indicators with a zero-shot Large Language Model (Llama 3.3 70B) surface scoring patterns that are invisible to manual workflows, and does the LLM output faithfully reflect the numerical findings supplied to it through prompt engineering alone?

**RQ2**. Can the underlying PHP-based, commodity-hardware web system reliably digitise the complete competition lifecycle of a community-level Shorinji Kempo event, including multi-device judge scoring and conditional bracket logic, to produce the data stream that the framework requires?

Section 2 describes the methodology. Section 3 presents results, including a quantitative content audit of the LLM output. Section 4 discusses the findings, separated into limitations, implications, and future work. Section 5 states conclusions.

The contributions of this paper are: (i) a structured referee analytics framework for a subjective-scoring martial art that integrates seven statistical indicators with a zero-shot LLM narrative layer and a rule-based reproducible content audit; (ii) a documented operational deployment of the framework at the 2026 Shorinji Kempo competition between students from all over Daerah Istimewa Yogyakarta province (POPDA DIY 2026, 111 matches across 12 categories, two days, zero data loss) that produces a reproducible dataset for officiating analysis; (iii) the identification and empirical characterisation of a failure pattern in zero-shot LLM reporting on structured data, which we label ignored-context under-utilisation, framed as a specific manifestation of grounding failure in which the source is fully supplied in structured form yet systematically under-cited; the pattern is quantified using three reproducible, set-membership audit metrics; and (iv) three concrete design implications for subsequent statistics-to-narrative LLM systems, namely schema-enforced decoding, automatic post-generation content checks, and dashboard-primary architectures with the narrative treated as an assistive generative layer.

## 2. Method

This research followerd a Design Science Research (DSR) approach[11], comprising the design and development of the SENTAKU artefact followed by operational deployment for evaluation. The research

was conducted in four sequential phases.

*2.1. Phase 1: Requirements Analysis*

Requirements were gathered through structured observation of prior PERKEMI events and semi-structured interviews with seven PERKEMI coordinators and officiating supervisors. Core requirements included: (a) a complete digital workflow from registration to result publication; (b) multi-device judge scoring via personal smartphones without dedicated hardware; (c) automatic bracket management including conditional Grand Final Rematch (GF2) logic; and (d) structured post-event referee performance reporting for federation administrators. A specific requirement from officiating supervisors was the ability to identify judges whose scoring patterns diverge consistently from the panel consensus, a need unmet by any tool in current federation use.

*2.2. Phase 2: System Design and Implementation*

SENTAKU was implemented as a three-tier web application using PHP 8.x, MariaDB 10.4, and standard browser clients. The system is designed for local-area network (LAN) deployment to ensure low-latency operation during live events, with a public-facing deployment at https://sentaku.shorinjikempo.or.id for result access. Five roles interact with the system: Coordinator (full event lifecycle management), Scorekeeper (real-time court operation), Judge (score submission via QR-coded time-limited tokens), Public (read-only access to results and live court displays), and Super Administrator (user management).

All categories use a Double Elimination (DE) bracket, guaranteeing each participant a minimum of two matches. A Grand Final (GF) is contested between the Winner Bracket (WB) and Loser Bracket (LB) survivors, with a Grand Final Rematch (GF2) triggered automatically only if the LB survivor wins the GF, ensuring the champion has lost at most once. This conditional logic is encoded procedurally in the scheduling module and was activated live in three of the ten randori categories during the POPDA DIY 2026 deployment (Section 3.1). The bracket draw uses constraint-based randomisation that prevents same-contingent first-round matchups, using PHP's *random_int()* cryptographically secure random generation.

The embu scoring model collects scores from five independent judges on ten components: six technical (K1 through K6, covering specific executed techniques) and four presentation components (P1 through P4, covering rhythm and flow, posture, energy and kiai, and zanshin), each scored on a discrete 0/8/9/10 scale yielding a maximum of 100 points per judge. The panel score uses a trimmed mean: the highest and lowest individual totals are discarded and the remaining three averaged, a standard approach in artistic judging disciplines to mitigate outlier influence [12]. Figure 1 illustrates the trimmed-mean computation from Game 8 of the paired embu of a female category (Gunungkidul vs Bantul, POPDA DIY 2026). The Bantul (Pita Putih) panel shows a four-point spread (83–87), with the highest score of 87 and lowest score of 83 discarded, leaving three scores of 84 and a trimmed mean of 84.00.

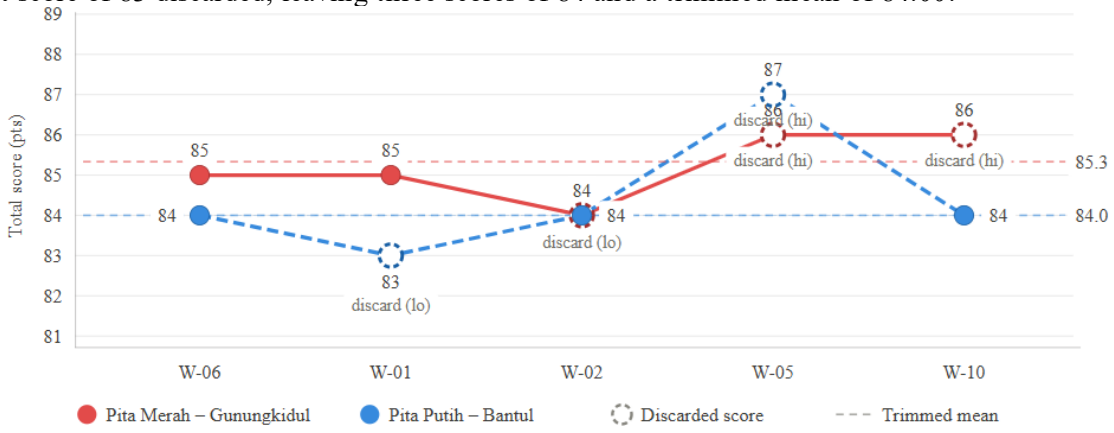


**Figure 1.** Per-judge total score from Game 8 of NOM-002 paired embu (Gunungkidul vs Bantul, POPDA DIY 2026)

When trimmed panel scores are equal, a five-level tiebreaking hierarchy is applied automatically: (a) panel total after deductions, (b) sum of the three middle technical scores, (c) total from Judge 1 (chief judge), (d) technical score from Judge 1, and (e) designation of the winner by the chief judge.

*2.3. Phase 3: Statistical and LLM Integration*

The referee evaluation module addresses the specific gap identified by federation supervisors: the inability to systematically assess whether individual referees score consistently and fairly within their panel. The module operates in two stages: a statistical computation stage that runs entirely in PHP, and an LLM narrative generation stage.

In the statistical stage, the system aggregates the complete score matrix from all embu matches to compute seven performance indicators for each judge. Because subjective-scoring disciplines have no objective ground truth, the panel median is used as a proxy reference: each judge is compared against the median of his or her own five-judge panel for the same performance, not against an external expert rating. This is standard in the referee-analytics literature [6] but is an assumption rather than a truth, with implications discussed in Section 4. Mean score establishes baseline scoring level. Standard deviation measures intra-judge consistency. Median deviation is the average signed difference between a judge's panel total and the panel median (positive indicates inflation). Discard rate is the proportion of a judge's scores that are panel maximum or minimum and excluded from the trimmed mean. Agreement rate is the proportion of matches in which the judge's ordinal preference aligned with the panel majority. Component profile is a ten-dimension vector of per-component median deviations across K1–K6 and P1–P4. Contingent-affiliation pattern cross-tabulates each judge's average deviation against the regency of the competing contingent, flagged when judge and contingent share the same regency and the deviation exceeds |0.50| points over at least three panels. Threshold values for the five flags were set through historical calibration combined with expert judgment, not chosen freely. Historical scoring data from prior PERKEMI events was retrospectively digitised and used to compute the same seven indicators, producing reference distributions for each metric. The PERKEMI DIY officiating board then reviewed these distributions and set threshold values, based on their officiating experience, that would catch the upper tail of observed divergence while leaving the bulk of the panel unflagged. The resulting values are |0.50| points for median and affiliation deviation, 1.80 for standard deviation, 75% for agreement (set as a lower bound because lower agreement is more divergent), and 70% for discard rate. These thresholds are starting values to be revised after each subsequent deployment using the growing longitudinal dataset. As Section 3.2 reports, the 70% discard-rate threshold in particular proved miscalibrated against the POPDA DIY 2026 data and is the first scheduled recalibration target. All metrics are rendered in an interactive coordinator dashboard with heat maps using a red-green diverging colour scale.

In the LLM narrative generation stage the system submits the computed statistics to the Groq inference API with Llama 3.3 70B, selected for sub-second inference latency on large open-weight models suitable for competition-day workflows [13]. The PHP backend assembles a JSON payload containing all seven indicator values for each judge and constructs a system prompt (~1,200 tokens) that instructs the model to function as an experienced officiating supervisor. The prompt specifies scoring conventions (0/8/9/10 scale, trimmed mean, tiebreaking), defines each metric with its threshold interpretation, and requests output covering five named sections: overall panel consistency, individual referee tendencies, component-level calibration, contingent-affiliation observations, and development recommendations. The prompt underwent three refinement cycles during pre-event testing against synthetic data generated from the historical scoring data used for threshold calibration: establishing persona, metric glossary, and five-section schema; adding explicit instructions to cite specific numerical values and to address every judge by name; and adding scoring-instrument vocabulary in Bahasa Indonesia (kihon, kiai, zanshin, embu berpasangan). The prompt is structured as four blocks: role and audience, JSON payload of all indicators with threshold dictionary, reporting instructions enumerating the five required sections, and tone and language constraints (Bahasa Indonesia, professional register, no speculation beyond the data). The complete prompt template is provided as supplementary material. No fine-tuning was applied; the

approach relies on zero-shot prompting [8]. The generated report is stored in the database (*ai_analisis_wasit* table) and rendered to the coordinator interface alongside the statistical dashboard.

*2.4. Phase 4: Validation*

System validation was conducted through operational deployment at POPDA DIY 2026, held on 2–3 April 2026 at Sasana Among Raga Yogyakarta, Indonesia. The event comprised 12 competition categories (2 embu and 10 randori), 5 participating contingents representing regencies and cities of Daerah Istimewa Yogyakarta province (Sleman, Kulon Progo, Gunungkidul, Bantul, and Kota Yogyakarta), 53 randori athletes, 26 embu athletes organised as 13 paired teams (7 teams in NOM-001 *Embu Berpasangan Putra Kyukenshi* and 6 teams in NOM-002 *Embu Berpasangan Putri Kyukenshi*, each team comprising two athletes), with several athletes competing in both disciplines (67 unique athletes overall), 10 active embu judges, and a total of 111 completed matches across two parallel court stations.

Validation was structured around four criteria. Data integrity required zero loss or corruption of any match record. Bracket correctness required that all 111 match results propagated correctly through the DE bracket logic across all 12 nomor pertandingan without manual override. Operational continuity required that no system downtime or unplanned manual intervention occurred during the event. LLM report faithfulness was assessed post-hoc through verbatim comparison of the model's output against the statistical metrics that had been supplied to it in the prompt. Three structured checks were applied to the LLM output: (i) coverage, namely the proportion of the ten judges explicitly named; (ii) numerical grounding, namely the presence of any cited metric value (e.g., median deviation, discard rate); and (iii) section completeness, namely which of the five requested output sections were actually produced. These three checks operationalise the input coverage, numerical grounding, and schema compliance dimensions most commonly cited as necessary for faithful LLM reporting on structured data [8], [14], and correspond to the faithfulness and completeness dimensions in surveys of hallucination and grounding failure [13], [15]. Each check executes a pre-specified deterministic rule rather than a subjective judgment: coverage is a set-membership test against the ten judge names in the input payload, numerical grounding is a regular-expression search for any indicator value, and section completeness is a match against the five mandatory headings in the prompt schema. Each yields an integer count that any third party can recompute from the stored LLM output and the prompt payload, without reference to the research team's interpretation. This makes the audit reproducible in the strict sense (re-running it on the supplementary data must return the same numbers) and removes the inter-rater variance that typically dominates expert Likert evaluations of generated text. The selection of these checks reflects prior design decisions; what is claimed is the rule-based reproducibility of audit execution given these choices. Figure 2 illustrates the framework.

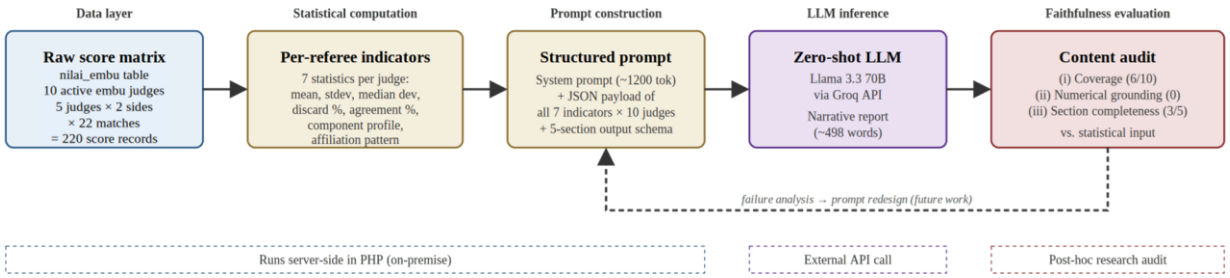


**Figure 2.** SENTAKU structured referee analytics framework

The post-hoc faithfulness assessment was performed by the research team against the underlying *nilai_embu* table; this represents an internal, verifiable check of model behaviour and is not a substitute for expert review of actionability. As noted in Section 4, expert actionability review is a planned follow-up not yet completed and is therefore not claimed in this paper.

## 3. Result

### *3.1. System Operation*

SENTAKU processed all 111 matches across the two-day event without data loss or system interruption. Figure 3 shows the coordinator dashboard during the event, displaying real-time bracket progression, match status, and court activity across all 12 categories.

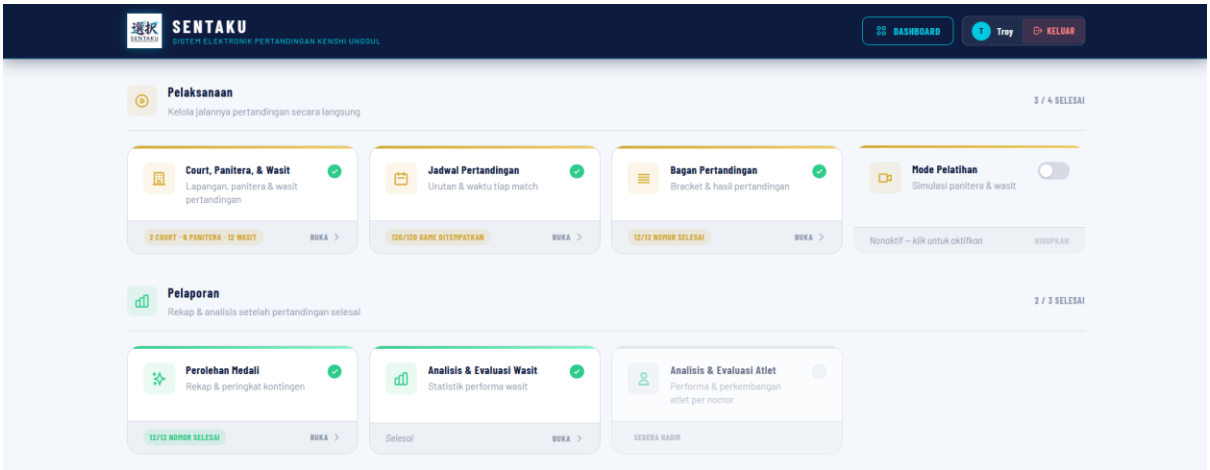


**Figure 3.** Coordinator dashboard on SENTAKU

Result publication latency from scorekeeper confirmation to public court screen averaged under 15 seconds, consistent with the 3-second polling cycle. No score correction was required. This contrasts with the preceding year's event in which three post-event disputes arose from paper transcription errors.

Judge QR-code token delivery was effective: all five judges connected to the scoring interface within 90 seconds of match initiation in 106 of 111 matches (95.5%). In the remaining five, one judge experienced network delay and submitted within an extended window; no match result was affected.

Bracket progression was correct for all 12 nomor pertandingan, all of which were conducted under the Double Elimination scheme. For the 10 randori categories, the GF2 rematch condition was triggered live in three categories (NOM-006 *Putra > 55–60 Kg*, NOM-008 *Putri 35–40 Kg*, and NOM-010 *Putri > 45–50 Kg*) where the Loser Bracket survivor won the initial Grand Final; in each case the scheduler auto-generated the rematch slot without manual intervention and the rematch completed within the same court allocation. The remaining seven randori categories concluded at the GF stage because the Winner Bracket survivor won the initial GF. The two embu categories also followed the DE scheme. NOM-001 (*Embu Berpasangan Putra Kyukenshi*) with 7 paired teams produced 12 matches (2N−2 under DE without GF2), and NOM-002 (*Embu Berpasangan Putri Kyukenshi*) with 6 paired teams produced 10 matches; both categories completed at the GF stage without dispute. Table 1 summarises system performance.

**Table 1.** SENTAKU operational performance at POPDA DIY 2026

| Performance indicator | Result | Benchmark / target |
|---|---|---|
| Total matches processed | 111 | 111 (zero loss) |
| Data integrity | 100% | 100% |
| Randori Grand Finals completed | 10 of 10 | 100% |
| Randori GF2 trigger events | 3 of 10 | Dependent on outcomes |
| Judge connection within 90 s | 106 of 111 (95.5%) | > 90% |
| Average result publication latency | < 15 seconds | < 20 seconds |
| Post-event score disputes | 0 | 0 |

| Performance indicator | Result | Benchmark / target |
|---|---|---|
| System downtime during event | 0 minutes | 0 minutes |

*3.2. Referee Statistical Analytics*

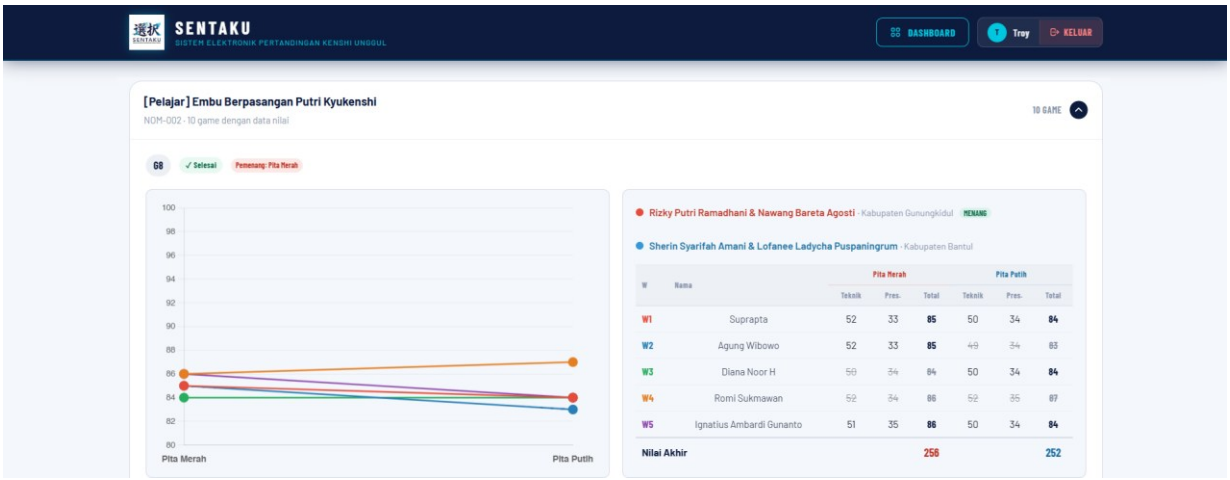


**Figure 4.** Coordinator interface for embu match results

Post-competition analysis covered 10 active judges across the two embu categories, producing 220 individual score records (all 22 embu matches involved two actively scored sides; 22 matches × 5 judges × 2 sides, comprising 12 matches from NOM-001 and 10 matches from NOM-002). Figure 4 shows the embu match result in the coordinator interface, presenting the five judge totals alongside discarded scores and the computed trimmed mean.

Table 2 presents referee performance metrics aggregated across all embu matches, computed directly from the *nilai_embu* table and can be reproduced from the supplementary SQL dump.

**Table 2.** Summary of referee performance metrics (POPDA DIY 2026, n=10 judges)

| Metric | Min | Max | Mean | Flag threshold | Judges flagged |
|---|---|---|---|---|---|
| Mean panel total (points) | 85.04 | 86.17 | 85.62 | n/a | n/a |
| Median deviation (points) | −0.67 | +0.60 | −0.01 | \|0.50\| | 3 (30%) |
| Standard deviation | 0.88 | 2.13 | 1.44 | > 1.80 | 2 (20%) |
| Discard rate (%) | 15.0 | 65.0 | 40 | > 70% | 0 (0%) |
| Agreement rate (%) | 50.0 | 100.0 | 86.8 | < 75% | 1 (10%) |

Mean panel deviation ranged from −0.67 to +0.60 points, reflecting a panel that is largely calibrated but with meaningful individual divergence. Three judges (W-02 at −0.500, W-03 at −0.667, and W-05 at +0.600) showed median deviations at or beyond the |0.50| threshold, triggering the scoring tendency flag. Standard deviation ranged from 0.88 to 2.13; two judges (W-04 at 2.13 and W-08 at 2.03) exceeded the 1.80 threshold, indicating less consistent application of scoring standards across matches of similar quality. Discard rates ranged from 15.0% to 65.0% with a mean of 40.0%, and no judge exceeded the pre-registered 70% threshold; the threshold should be recalibrated using POPDA DIY 2026 as the first longitudinal baseline. Agreement rate ranged from 50.0% to 100.0%; W-02 fell below the 75% threshold at exactly 50.0%, the most divergent ordinal-ranking profile.

The component-level heat map revealed patterns that are invisible to aggregate analysis. Judges W-03 and W-06 showed systematic deflation specifically on P3 (energy and kiai) at −0.417 and −0.350 respectively, and on P4 (zanshin), while maintaining closer alignment with the panel on K-series technical

components. Judge W-02 showed a contrasting pattern: deflation on K1 (−0.350), K2 (−0.450), and P1 (−0.450), combined with inflation on P4 (+0.550), indicating component-specific calibration rather than uniform tendency. Judge W-05 showed consistent inflation of +0.200 across K3, K4, K5, and P2. These divergences are calibration differences that aggregate scoring cannot surface; because individual component scores are on a coarse discrete scale (0/8/9/10), only cross-match aggregation makes them visible. Figure 5 illustrates the component heat map.


**Figure 5**. Component-level median-deviation heat map per judge

Contingent-affiliation analysis examined, for each judge, whether his or her mean median deviation differed systematically when scoring a contingent from the same regency as the judge. Eight of the ten judges had at least one same-regency panel, yielding eight judge–contingent same-regency pairs. Four pairs showed absolute deviations of 0.40 points or more, of which three exceeded the formal |0.50| flag threshold and one met it exactly: W-02 toward Sleman at +1.200 (n = 5 panels), W-03 toward Kulon Progo at −0.750 (n = 4 panels), W-10 toward Gunungkidul at +0.500 (n = 8 panels, meeting the threshold exactly), and W-08 toward Sleman at +0.444 (n = 9 panels, descriptively notable but below the formal flag). These patterns are heterogeneous: W-02, W-08, and W-10 inflated toward their own regency, whereas W-03 deflated toward his own regency. In every flagged case, final match outcomes were unaffected because winning margins exceeded the observed bias magnitude. Nevertheless the patterns are surfaced explicitly in the statistical dashboard as a basis for officiating-assignment policy. The heterogeneity itself (inflation in three judges, deflation in one) is a substantive finding and is consistent with the observation in team-sport literature that affiliation effects vary across individuals rather than being a uniform directional pull [2], [5].

### *3.3. LLM Narrative Report: Content Analysis*

LLM report is the second analytical output of the system. Figure 6 shows the report as rendered in the coordinator interface. The report was generated at 09:27:01 on 3 April 2026 (final day of the competition) using Llama 3.3 70B via the Groq API. End-to-end inference latency from dashboard trigger to stored response was under three seconds. The stored report contains 498 words.

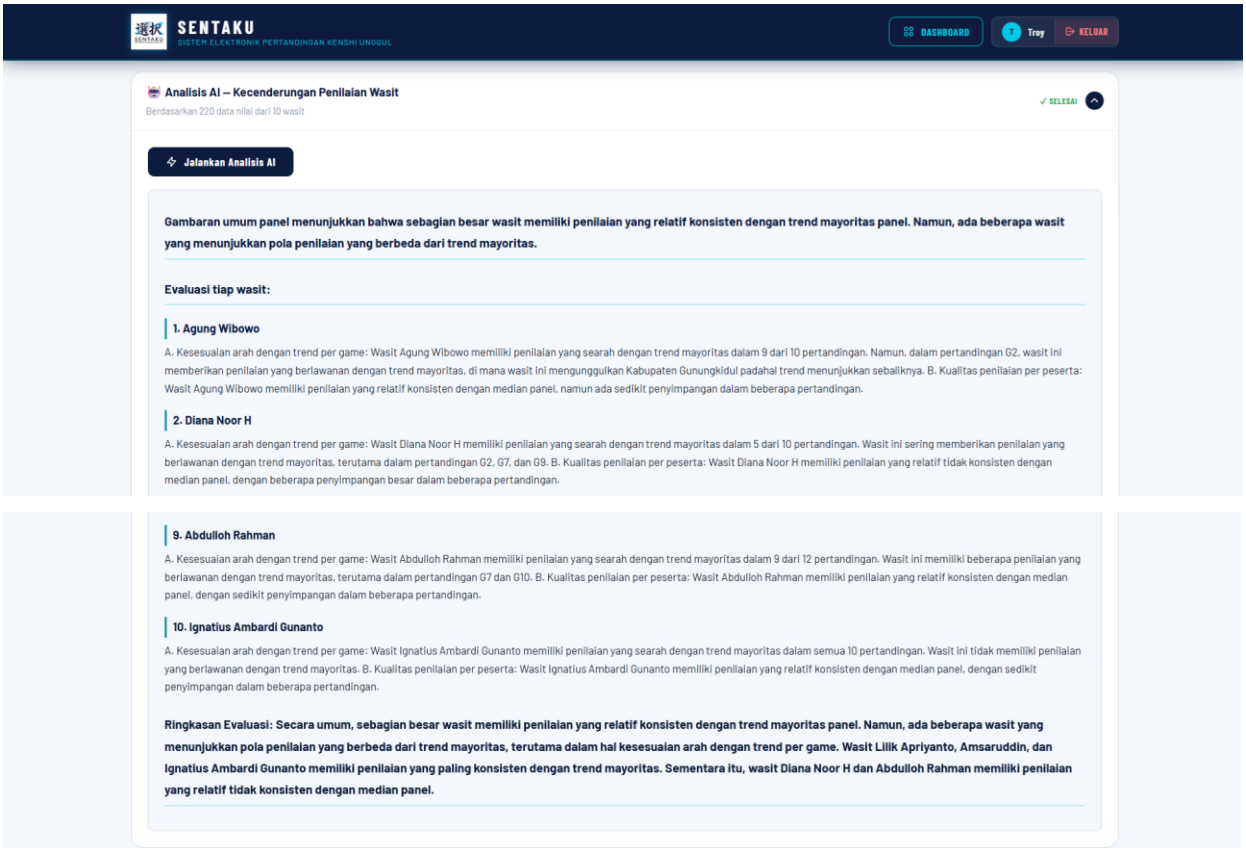


**Figure 6.** Narrative report produced by the Llama 3.3 70B, showing individualised commentary for all ten judges (W-01 through W-10)

A structured content audit of the report against the statistical input yielded the following findings.

*Coverage*. The report names all 10 judges (W-01 through W-10), addressing the coverage instruction in full. Each judge receives an individualised commentary paragraph identifying perceived alignment or divergence from the panel median.

*Numerical grounding*. No numerical value from the statistical input appears in the report text. The report does not cite a single median deviation, discard rate, agreement rate, standard deviation, or component score. Its language is qualitative throughout (e.g., "*searah dengan trend mayoritas pada 9 dari 10 pertandingan*", "*konsisten dengan median panel*", "*beberapa kasus di mana ia memberikan nilai yang lebih tinggi atau lebih rendah*"). The "9 of 10" and "5 of 10" phrasings visible in the report appear to be model-internal estimates rather than values from the structured input, as the input payload contained no such ratio.

*Section completeness*. The report contains three free-form sections (an opening overall assessment of two sentences, a per-judge evaluation body with an A/B sub-structure, and a closing summary) rather than the five named sections requested in the prompt (overall panel consistency, individual tendencies, component-level patterns, contingent-affiliation observations, development recommendations). Two of the five requested sections are absent outright: there is no component-level calibration discussion (K1–P4 are not mentioned at any point), and there is no explicit development recommendation section. Contingent-affiliation content is present only as a single qualitative clause in the closing summary ("*terutama pada pertandingan yang melibatkan kontingen dari daerah yang sama dengan wasit*"), without naming specific judge–contingent pairs.

*Factual errors*. Among the judges discussed, the report makes several direction-of-trend statements that match the data (e.g., W-02 is correctly described as having less consistent alignment with the panel

median, consistent with her 50% agreement rate and −0.500 median deviation). No statement was found to contradict the statistical input. The report does not, however, provide the specificity requested by the prompt, and its correctness therefore rests on generic, direction-only language that cannot be wrong in any sharp sense.

*3.3.1. Audit Results*

In aggregate numerical terms, the audit yields the following three rates against the input payload: coverage rate = 10/10 = 1.00, numerical grounding rate = 0/N = 0.00 where N is the count of numerical claims that the prompt instructs to be cited; under a maximal interpretation of the prompt instruction that each of the seven indicators should be explicitly referenced for each of the ten judges, N ≥ 70 expected citations, and section completeness rate = 3/5 = 0.60. The composite Faithfulness Index, defined as the arithmetic mean of the three rates, is therefore 0.53 against a target of 1.00. Equal weighting across the three rates was adopted for interpretability rather than empirical optimisation; the index is therefore heuristic and intended as a comparative baseline for benchmarking against alternative LLM integration designs, not as a validated psychometric measure of faithfulness. These three rates and the composite index are the primary quantitative outputs of the LLM evaluation, summarised in Table 3.

**Table 3.** Quantitative content audit of the LLM report against the structured input payload

| Metric | Operationalisation | Observed | Target | Gap |
|---|---|---|---|---|
| Coverage | Judges named in output / judges in input | 10/10 = 1.00 | 1.00 | 0.00 |
| Numerical grounding | Indicator values cited in output / expected (N ≥ 70) | 0/70+ = 0.00 | 1.00 | 1.00 |
| Section completeness | Output sections matching schema / sections requested | 3/5 = 0.60 | 1.00 | 0.40 |
| Faithfulness Index | Arithmetic mean of the three rates | 0.53 | 1.00 | 0.47 |

*3.3.2. Interpretation of Failure Pattern*

The audit distinguishes three categories of content: (a) direction-of-trend statements consistent with the input statistics (directionally consistent), (b) unsupported numerical assertions such as "9 of 10" and "5 of 10" with no corresponding ratio in the payload (locally fabricated), and (c) omitted content, namely all numerical indicator values and two requested sections (component-level calibration and development recommendations). The failure mode is dominated by the absence of numerical grounding: the model addressed every judge by name and partially followed the requested schema (3 of 5 sections), yet declined to cite a single indicator value from the payload despite all values being present in structured form. The pattern does not match hallucination in the narrow sense [16], [17]: no fabricated official metric value, official indicator value, or contradicted statistic appears in the report. Within Huang et al.'s taxonomy [17], it most closely resembles instruction inconsistency within the faithfulness-hallucination family, with the distinguishing property that the ignored content is the structured numerical payload rather than the prose instruction; the localised "9 of 10" phrasings correspond to a minor fabrication subcategory but as subordinate phenomenon. The pattern can also be viewed as a specific manifestation of grounding failure contrasting with the canonical open-domain QA case where the model lacks a specific source: here the model was given the full numerical context in structured form and still systematically declined to cite it. This research label this pattern ignored-context under-utilisation.

*3.3.3. Scope and Caution*

The label is used descriptively for the pattern observed in this research and does not claim a new taxonomic category on the basis of a single instance. Alternative interpretations, such as weak instruction following on instruction-tuned models with high-density numerical context, remain consistent with the present evidence and are not excluded. The audit is descriptive rather than comparative: a single report

from a single event, with no matched control under schema-constrained decoding or tool-calling and no repeated sampling. The composite index of 0.53 should be interpreted as a baseline against which subsequent design changes (schema-enforced decoding, automatic content checks) can be benchmarked, not as a population estimate. In this deployment, absence of hallucination alone was not sufficient to guarantee faithful LLM reporting; this is a negative result with respect to the faithfulness half of RQ1 with direct design implications (Section 4).

## 4. Discussion

### *4.1. RQ1: Analytics framework and LLM faithfulness*

The content audit in Section 3.3 addresses RQ1 in two parts. On surfacing patterns invisible to manual workflows, the answer is affirmative: component-level calibration differences (e.g., P3 deflation at −0.417 and −0.350 in two judges while K-series scores remained near panel consensus) and same-regency patterns (e.g., +1.200 over 5 panels) were produced from a single event dataset, not practically observable through conventional paper workflows. On LLM narrative faithfulness, the answer is negative and more instructive. The core design assumption (that a well-engineered prompt with numerical context and a five-section schema would suffice) is violated in two ways simultaneously: none of the supplied numerical values cited, and two of five requested sections missing. This is not a hallucination problem in the narrow sense, since the report does not invent judge-level indicator values or contradict the input statistics, but rather a grounding and completeness problem combined with localised unsupported quantitative assertions ("9 of 10" and "5 of 10" ratios in the prose with no corresponding ratio in the input payload): the model produced generic qualitative prose that under-uses the specific structured information available to it. The pattern is also consistent with weak instruction following on high-density numerical context, an interpretation that the present single-instance evidence does not rule out. Future work should compare this behaviour against schema-constrained decoding; the present deployment suggests that prompt-only control may be insufficient for reliable structured statistical reporting, an observation that future comparative work would either confirm or qualify. Cook and Karakuş noted that output structure compliance under prompt engineering is fragile [8]; our results extend that observation to an analytics domain where the failure is systematic rather than incidental. Anh-Hoang et al. report that prompting-induced failures persist even with carefully engineered prompts on instruction-tuned models [18], consistent with the prompting-induced category that the present pattern most closely resembles, further supporting the design recommendation to move control from the prompt to the decoding step.

### *4.2. RQ2: System reliability*

The operational results address RQ2 affirmatively. SENTAKU completed the full POPDA DIY 2026 event (111 matches, two courts, two days) with zero data loss, zero downtime, zero post-event disputes, and a sub-15-second publication latency. The DE bracket engine handled all 12 nomor pertandingan correctly without manual override, including three randori categories in which the GF2 rematch condition was triggered live; the scheduler auto-generated each rematch slot and the rematch completed within the same court allocation. The conditional GF2 logic was therefore exercised both in pre-event unit testing and in live operation. Wen and Wang proposed a comprehensive sports event management information system but without officiating analytics [1]; Zhang et al. developed sensor-based sports event optimisation but did not address multi-judge scoring or referee behaviour [12]. SENTAKU occupies a so-far underexplored intersection by combining commodity-hardware competition management with a statistical referee-behaviour layer; to the authors' knowledge, limited prior work has reported this combination in a deployed community-event setting, but the novelty claimed rests on the analytical layer rather than the management infrastructure.

Independently of its role as data source, the management infrastructure contributed one operational observation: the combination of QR-coded time-limited tokens and a 3-second polling display avoided two common failure modes of judge-based scoring (lost credentials and stale public-display data) across 111 matches without dedicated hardware.

Three concrete design lessons follow. First, for faithful statistics-to-narrative generation, zero-shot plain-text prompting should be replaced with schema-enforced decoding: JSON-mode constrained output with a per-judge object array, or tool-calling against a read-only statistics API the model must invoke for every named judge. Second, a pre-submission content check (coverage, grounding, completeness) should be computed automatically on the LLM output and the response regenerated on failure; such a check is cheap and would have detected the absence of numerical grounding and missing schema elements immediately. Third, the statistical dashboard must remain the primary decision artefact, with the narrative treated as a convenience layer rather than the evaluation of record [19].

*4.3. Substantive findings from the statistical layer*

Independent of the LLM result, the statistical layer surfaced patterns of governance relevance. The component-level analysis identified P3 (energy and kiai) deflation in W-03 and W-06 while their technical components remained near panel consensus; such component-specific divergence is structurally undetectable through paper workflows and aggregate-only scoring. Carvalho et al. observed an analogous limitation in football refereeing, where component-level divergence is missed by aggregate scores [6]. The contingent-affiliation pattern for W-02 toward Sleman (+1.200 across 5 panels) is the largest same-regency deviation observed; because Sleman-involving randori finals were decided by margins far exceeding this magnitude, no outcome was affected, but the federation now has a quantified basis to consider panel-composition policies that avoid same-regency assignments, consistent with bias-reduction evidence reported for VAR deployment in football [4].

*4.4. Limitations*

Several limitations qualify the contribution. (i) The statistical thresholds (|0.50|, 70%, 75%, 1.80) were set through expert judgment supported by historical scoring data and proved partially miscalibrated; no judge reached the 70% discard-rate flag, and the threshold should be recalibrated using POPDA DIY 2026 as the first longitudinal baseline. (ii) The panel median is used as a proxy reference rather than ground truth, since no objective criterion exists for embu performance; a judge flagged for high median deviation is flagged for divergence from the rest of the panel, not for being wrong, and a panel systematically biased in a consistent direction would not be detectable by median-based indicators. Cross-panel comparison and inter-panel calibration against external expert raters are needed to distinguish individual divergence from panel-level drift. (iii) The LLM faithfulness assessment is an internal, single-instance content audit performed by the research team; an independent expert actionability review was not completed for this paper, no matched control was generated under schema-constrained decoding, and no repeated sampling of the same prompt was performed to estimate run-to-run variance. The composite Faithfulness Index of 0.53 is therefore a single-instance baseline, not a population estimate, and any characterisation of the report as "actionable" beyond what the audit supports has been removed. (iv) The deployment covered a single regional event with 5 contingents, 10 judges, and 22 embu matches. Statistical patterns identified here are hypotheses to be tested across future events rather than established regularities. (v) The hash-data identifier stored alongside the report (value *6_NaN*) indicates a latent bug in the hash-generation routine when certain numeric values are absent; the bug did not affect model inputs but has been logged for patching. (vi) The Groq API dependency would preclude deployment in venues without reliable internet; a local inference backend (Ollama with a 70B-class quantised model) is planned. (vii) The label "ignored-context under-utilisation" was assigned descriptively to a single observed report; its status as a generalisable pattern rather than weak instruction following or a model-specific artefact can only be established by replication across models, prompts, and domains, and the label is offered as a candidate description rather than a validated taxonomic category.

*4.5. Implications*

The findings carry three distinct sets of implications. For LLM-based analytics design, the single-instance Faithfulness Index of 0.53 establishes an empirical baseline for zero-shot plain-text prompting on

structured statistical input; subsequent designs (schema-enforced decoding, tool-calling, retrieval-augmented generation) can be benchmarked against this baseline using the same three audit metrics, which makes the audit reusable as an evaluation protocol rather than a one-off measurement. For tournament software practice, the deployment demonstrates that commodity-hardware web infrastructure is sufficient to digitise a federation-level event and that the resulting data supports referee analytics that paper workflows cannot; the analytical value does not depend on the LLM working well, since the component-level and same-regency patterns surfaced at POPDA DIY 2026 are products of the statistical layer alone. For officiating governance in PERKEMI, the four same-regency pairs at or beyond the descriptive 0.40 threshold (W-02, W-08, W-10 inflation; W-03 deflation), three exceeding the formal |0.50| flag, provide a quantified evidence base for panel-composition policies; the heterogeneous direction cautions against simple home-bias assumptions in subjective panels.

### *4.6. Future Work*

Five extensions follow directly from the limitations. First, a controlled within-payload comparison should submit the same statistical payload under (a) the current plain-text zero-shot prompt, (b) JSON-mode constrained output with a per-judge object array, (c) tool-calling against a read-only statistics API, and (d) multiple independent samples per condition; the three audit metrics applied across all conditions will turn the present single-instance baseline into an effect-size estimate. Second, an independent expert actionability review by at least three external reviewers using a pre-registered rubric will complement audit reproducibility with content validity. Third, statistical thresholds should be recalibrated using POPDA DIY 2026 as the first longitudinal anchor, with intraclass correlation coefficients added to support cross-panel comparison. Fourth, multi-event longitudinal deployment will distinguish individual-match variance from persistent tendencies and detect scoring drift across seasons. Fifth, a local inference backend (Ollama with a 70B-class quantised model) should be integrated to remove the external-API dependency and its faithfulness compared with Groq-hosted inference using the same audit.

### *4.7. Ethics*

Referee evaluation with individual identification raises ethical obligations handled as follows. All participating judges were informed in writing that their scoring data would be aggregated and analysed post-event and gave written consent for statistical analysis. Judges are referred to by anonymised codes (W-01 through W-10); the name-to-code mapping is retained only by the research team. The LLM report naming specific judges was viewed only by the federation's officiating committee and the research team and is not reproduced verbatim. The study protocol was reviewed by the PERKEMI DIY officiating board; no separate institutional ethics review was conducted, a gap that future extensions will address.

## 5. Conclusion

This paper presented SENTAKU, a referee analytics platform for Shorinji Kempo built on a PHP-based competition management system. Deployment at POPDA DIY 2026 processed 111 matches across 12 categories with zero data loss, zero disputes, and 95.5% judge connection within 90 seconds, affirmatively answering RQ2.

RQ1 is answered in two parts. The statistical layer surfaced officiating patterns unreachable through paper workflows: component-level calibration differences and same-regency deviations from +0.444 to +1.200 (inflation) and a maximum deflation of −0.750, providing a quantified basis for officiating-assignment policies. The LLM layer, by contrast, named all 10 judges and produced part of the requested schema (3 of 5 sections), but cited no numerical value from the structured input, while introducing no fabricated entities, events, or contradicted statistics; localised unsupported ratio phrasings ("9 of 10", "5 of 10") appear as subordinate phenomena within a grounding-dominated failure pattern (Section 3.3).

The scientific contribution is the empirical identification of a specific failure pattern, labelled ignored-context under-utilisation, framed as a specific manifestation of the broader grounding/faithfulness failure family rather than a new taxonomic category; the present deployment indicates that absence of

hallucination alone may not guarantee faithful LLM reporting on structured data. The characterisation is inferential from a single event without a matched comparative baseline, and is offered as a phenomenon to be confirmed and quantified by future work.

The most consequential next step (Section 4.6) is a within-payload comparative evaluation under schema-enforced decoding and tool-calling, which would turn the single-instance Faithfulness Index of 0.53 into an effect-size estimate and establish whether the pattern is eliminated, attenuated, or replaced under stricter output control.

## 6. References

[1] Y. Wen and F. Wang, "Design and Application of Major Sports Events Management Information System Based on Integration Algorithm," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–10, May 2022, doi: 10.1155/2022/6480522.

[2] K. Pelechrinis, "Quantifying implicit biases in refereeing using NBA referees as a testbed," *Sci. Rep.*, vol. 13, no. 1, p. 4664, Mar. 2023, doi: 10.1038/s41598-023-31799-y.

[3] T. Gasparetto and K. Loktionov, "Does the Video Assistant Referee (VAR) mitigate referee bias on professional football?," *PLoS One*, vol. 18, no. 11, p. e0294507, Nov. 2023, doi: 10.1371/journal.pone.0294507.

[4] C. H. Kim, K. Y. Lee, and Y. Kwon, "Does VAR reduce referee bias? Causal evidence from global professional football leagues," *Int. J. Sports Sci. Coach.*, Nov. 2025, doi: 10.1177/17479541251391395.

[5] F. Shu, X. Xu, and J. Mao, "Exploring Referee Bias in Basketball: A Case Study of the Chinese Basketball Association," in *Proceedings of the 2024 International Conference on Sports Technology and Performance Analysis*, New York, NY, USA: ACM, Dec. 2024, pp. 391–399. doi: 10.1145/3723936.3723997.

[6] V. Carvalho, P. T. Esteves, C. Nunes, W. F. Helsen, and B. Travassos, "The assessment of the match performance of association football referees: Identification of key variables," *PLoS One*, vol. 18, no. 9, p. e0291917, Sep. 2023, doi: 10.1371/journal.pone.0291917.

[7] M. Li, X. Wang, and S. Zhang, "The effect of video assistant referee (VAR) on match performance in elite football: A systematic review with meta-analysis," *Proc. Inst. Mech. Eng. P J. Sport. Eng. Technol.*, Jun. 2024, doi: 10.1177/17543371241254596.

[8] A. Cook and O. Karakuş, "LLM-Commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data," *Knowl. Based. Syst.*, vol. 300, p. 112219, Sep. 2024, doi: 10.1016/j.knosys.2024.112219.

[9] M. Connor and M. O'Neill, "Large Language Models in Sport Science & Medicine: Opportunities, Risks and Considerations," May 2023.

[10] Y. Zhang, R. Qu, and O. Girard, "Faster, more accurate? A feasibility study on replacing human judges with artificial intelligence in video review for the Paris Olympics Taekwondo competition," *Front. Sports Act. Living*, vol. 7, Aug. 2025, doi: 10.3389/fspor.2025.1632326.

[11] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research1," *MIS Quarterly*, vol. 28, no. 1, pp. 75–106, Mar. 2004, doi: 10.2307/25148625.

[12] Y. Zhang, X. Zhao, J. Shen, K. Shi, and Y. Yu, "Optimization of Sports Event Management System Based on Wireless Sensor Network," *J. Sens.*, vol. 2021, no. 1, Jan. 2021, doi: 10.1155/2021/1174351.

[13] A. Passerini, A. Gema, P. Minervini, B. Sayin, and K. Tentori, "Fostering effective hybrid human-LLM reasoning and decision making," *Front. Artif. Intell.*, vol. 7, Jan. 2025, doi: 10.3389/frai.2024.1464690.

[14] Z. Yang, H. Xia, J. Li, Z. Chen, Z. Zhu, and W. Shen, "Sports Intelligence: Assessing the Sports Understanding Capabilities of Language Models Through Question Answering from Text to Video," *Electronics (Basel).*, vol. 14, no. 3, p. 461, Jan. 2025, doi: 10.3390/electronics14030461.

[15] A. R. Doshi, J. J. Bell, E. Mirzayev, and B. S. Vanneste, "Generative artificial intelligence and evaluating strategic decisions," *Strategic Management Journal*, vol. 46, no. 3, pp. 583–610, Mar. 2025, doi: 10.1002/smj.3677.

[16] Z. Ji *et al.*, "Survey of Hallucination in Natural Language Generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Dec. 2023, doi: 10.1145/3571730.

[17] L. Huang *et al.*, "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions," *ACM Trans. Inf. Syst.*, vol. 43, no. 2, pp. 1–55, Mar. 2025, doi: 10.1145/3703155.

[18] D. Anh-Hoang, V. Tran, and L.-M. Nguyen, "Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior," *Front. Artif. Intell.*, vol. 8, Sep. 2025, doi: 10.3389/frai.2025.1622292.

[19] V. C. Storey, W. T. Yue, J. L. Zhao, and R. Lukyanenko, "Generative Artificial Intelligence: Evolving Technology, Growing Societal Impact, and Opportunities for Information Systems Research," *Information Systems Frontiers*, vol. 27, no. 5, pp. 2081–2102, Oct. 2025, doi: 10.1007/s10796-025-10581-7.