From Metrics to Meaning: Rule-Grounded LLM Explanations for Data Literacy in the Case of Youth Football

Tomasz Piłka and Tomasz Kuczyński and Mateusz Czajka
Adam Mickiewicz University, Poznań
Center for Artificial Intelligence
Faculty of Mathematics and Computer Science
Uniwersytetu Poznańskiego 4, 61-614 Poznań, Poland

Abstract

Young athletes, parents, and coaches are increasingly exposed to training metrics from wearable technology, yet such metrics are difficult to interpret without contextual explanation. We present a rule-grounded data-to-text framework for supporting data literacy in youth football through concise, stakeholder-specific summaries of training sessions. A rule layer maps duration-normalised indicators to structured facts about session profile, internal intensity, speed exposure, and movement dynamics, which are then verbalised by a large language model for coaches, parents, or players. We compare direct generation from raw metrics, generation from rule-derived facts, and an augmented rule-grounded configuration, `ENRICHED`, that supplements validated facts with raw metrics and explicit threshold definitions. In this setting, selected open-weight models are additionally adapted using LoRA. The framework is developed using 122 anonymised player-session records from a U15 environment and evaluated on a held-out subset of ten sessions with stakeholder-oriented reference summaries. The results indicate that rule grounding improves reliability and audience adaptation compared with direct generation from raw metrics, particularly by reducing unsupported or overly strong interpretations. A school-based expert evaluation with physical education teachers further suggests that player-facing explanations in the evaluated `ENRICHED` setting can remain accurate, comprehensible, and practically useful. We position the framework as an interpretable data-literacy support interface for youth sport analytics.

1 Introduction

Wearable sensors and training analytics platforms are becoming increasingly common in youth sports. As a result, coaches, parents, and young athletes are exposed to growing amounts of numerical information about training load, movement dynamics, and physiological effort. However, such metrics require context to become understandable and practically meaningful. Stakeholders often encounter dashboards and indicators without sufficient guidance on how to interpret these measurements in relation to a specific training session. This creates a data literacy challenge: translating quantitative signals into concise, context-sensitive explanations.

This challenge is particularly evident in youth football, where the same session may need to be explained differently to different audiences. Coaches require concise technical interpretations that remain faithful to the data, whereas parents and players need clearer explanations that avoid specialist terminology and communicate the practical meaning of the metrics. In this setting, explanation functions as an educational interface that helps stakeholders read and reason about performance data.

This perspective aligns with recent educational Natural Language Processing (NLP) research on stakeholder needs, pedagogical alignment, and human-centred evaluation in Large Language Model (LLM)-based educational systems (Pal Chowdhury et al., 2025a; Galletti and Cesaroni, 2025). Recent work also suggests that pedagogical generation benefits from explicit control over communicative intent, while pedagogical quality remains difficult to assess using automatic methods alone (Petukhova and Kochmar, 2025; Kochmar et al., 2025). These concerns are directly relevant in our setting: when an LLM interprets raw training metrics without additional structure, it may produce explanations that are fluent but unsupported, overly strong, or inappropriate for the intended audience.

To address this issue, we propose a rule-grounded data-to-text framework for supporting data literacy in youth football. A rule layer converts normalised session metrics into interpretable facts about session profile, effort characteristics, and movement dynamics. An LLM then verbalises

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 492–502

these facts for a selected audience—coach, parent, or player—using audience-specific framing and terminology. We compare three generation settings: direct generation from raw metrics, strictly rule-grounded generation, and an augmented rule-grounded configuration, `ENRICHED`, that supplements validated rule-derived facts with raw metrics and explicit band definitions. In the open-weight experiments, LoRA adaptation is applied only to the selected models in `ENRICHED`; we therefore treat this condition as an applied configuration rather than as an isolated ablation of the added metric and threshold information.

The framework is developed using 122 anonymised player-session records from a U15 football environment. The controlled evaluation is conducted on a held-out subset of ten sessions for which aligned coach-, parent-, and player-facing reference summaries were prepared. The study addresses three questions: whether explicit rule grounding improves interpretive reliability compared with direct generation from raw metrics; whether the same rule-derived session interpretation can support multiple stakeholder groups while preserving a common session meaning; and whether `ENRICHED` can provide player-facing explanations that remain educationally useful and practically understandable. Our main contribution is to operationalise automated session summarisation as a stakeholder-oriented data literacy support problem rather than as a generic data-to-text task. By combining domain-informed rule grounding with persona-specific LLM verbalisation, the framework connects wearable-sensor metrics with explanations that are faithful to the data, adapted to the audience, and meaningful in practice.

2 Related Work

Recent work in educational NLP emphasises that LLM-based systems should be designed and evaluated with respect to stakeholder needs, pedagogical goals, and authentic educational workflows, rather than fluency alone (Pal Chowdhury et al., 2025a; Galletti and Cesaroni, 2025; Pal Chowdhury et al., 2025b). This view is consistent with research on data literacy, which shows that indicators and dashboards require contextualised support to become meaningful for users (Konold et al., 2015; Mandinach and Gummer, 2016; Michos et al., 2023; Valle et al., 2021). In youth football, this implies that training metrics should be communicated in ways that are understandable and appropriate for coaches, parents, and players.

From a generation perspective, our work builds on data-to-text conversion and controlled generation from structured inputs. Classical NLG research has emphasised content determination, content selection, planning, and surface realisation when generating text from data (Reiter and Dale, 2000; Gatt and Krahmer, 2018). These challenges remain central in neural and LLM-based generation, particularly where factual accuracy and numerical consistency are required (Wiseman et al., 2017; Puduppully et al., 2019; Chen et al., 2020; Maynez et al., 2020). Our framework follows this tradition by introducing an explicit interpretive layer between numerical session metrics and natural-language realisation.

Related work also highlights the importance of controlling communicative intent, audience, and readability in generated text (Petukhova and Kochmar, 2025; Moorjani et al., 2022; Tran et al., 2025). This is especially relevant in youth football, where movement and load indicators vary across age, role, and match context (Harley et al., 2010; Atan et al., 2016; Algroy et al., 2021; Palucci Vieira et al., 2019; Dalen and Loras, 2019; Hannon et al., 2021). We extend these perspectives by treating session summarisation as stakeholder-oriented explanatory generation for data literacy support. Rule grounding makes the interpretation traceable and auditable, while LLM realisation adapts the same underlying session meaning to coaches, parents, and players.

3 Rule-Grounded Formulation

Let $x_{i,t}$ denote a player-session record for player i in session t, where t indexes a training or match session rather than continuous time. Each record is represented as a vector of wearable-derived numerical indicators, including measures of external load, internal intensity, speed exposure, and movement dynamics. The goal is to generate a short natural-language explanation y for a target stakeholder $p \in \{\texttt{coach}, \texttt{parent}, \texttt{player}\}$.

We consider three generation settings that differ in the amount of interpretive structure provided to the language model. In all cases, a feature mapping ϕ preprocesses the raw player-session record and applies duration normalisation to selected indicators, yielding $\phi(x_{i,t})$. A banding function B then maps selected normalised indicators to ordinal lev-

els such as `low`, `medium`, `high`, and `elite`, using metric-specific threshold definitions Θ:

$$z_{i,t} = B(\phi(x_{i,t}); \Theta) \tag{1}$$

The banded representation $z_{i,t}$ is passed to a rule base R, which produces a set of structured explanatory facts:

$$F_{i,t} = R(z_{i,t}) = \{(f_k, s_k, \pi_k)\}_{k=1}^{K} \tag{2}$$

Here, f_k is an interpreted claim about the session, s_k is the supporting evidence associated with that claim, such as the relevant metric value and band, and π_k is a priority value used for content selection when multiple facts are available.

Each fact is represented as a compact structured object containing a fact type, supporting metric, interpreted value, and priority level. The rule base encodes domain knowledge about session characteristics, including overall load profile, internal intensity, speed exposure, and movement dynamics. For example, `HighDist` is activated when distance per minute is assigned to a high band and emits the following load-profile fact: "Distance per minute is high, so this session involved a large amount of running in volume terms. This may support endurance development and sustained lower-limb work, even if there were not many short bursts." The example shows how the rule layer converts a metric band into an interpretable statement that can later be verbalised differently for coaches, parents, or players. When multiple rules are activated simultaneously, priority-based selection resolves overlap and determines which facts are passed to the generation stage.

The three generation settings differ in the conditioning input provided to the language model:

$$C_m(x_{i,t}) = \begin{cases} \phi(x_{i,t}), & m = \texttt{RAW}, \\ F_{i,t}, & m = \texttt{RULES}, \\ \big(F_{i,t}, \phi(x_{i,t}), \Theta\big), & m = \texttt{ENRICHED} \end{cases} \tag{3}$$

The final explanation is produced by a language generation module conditioned on the selected input and the target persona:

$$y = G(C_m(x_{i,t}), p) \tag{4}$$

The generation setting is indexed by $m \in \{\texttt{RAW}, \texttt{RULES}, \texttt{ENRICHED}\}$.

This formulation separates analytical interpretation from linguistic realisation. In the `RAW` setting, the language model receives normalised metrics and must infer their meaning directly. In the `RULES` setting, the rule layer determines the structured interpretation to be verbalised for a given audience. In `ENRICHED`, rule-derived facts are supplemented with the underlying normalised metrics and threshold definitions. In the implementation evaluated here, the selected open-weight models in this setting are additionally adapted using LoRA.

4 System Overview

4.1 Architecture

Figure 1 shows the proposed pipeline for generating short, stakeholder-oriented explanations from structured training-session data. The system preprocesses wearable-derived metrics into normalised features, maps them to ordinal interpretive bands, and passes them to a rule layer that produces structured explanatory facts about the session profile. A persona-conditioned LLM then generates audience-specific summaries under three input settings: direct generation from normalised raw metrics (`RAW`), generation from validated rule-derived facts only (`RULES`), and generation from rule-derived facts supplemented with raw metrics and explicit band definitions (`ENRICHED`).

In the implementation evaluated here, the selected open-weight models used in the `ENRICHED` setting are additionally adapted using LoRA. The architecture supports both the `RAW`–`RULES` comparison and the applied evaluation of `ENRICHED`.

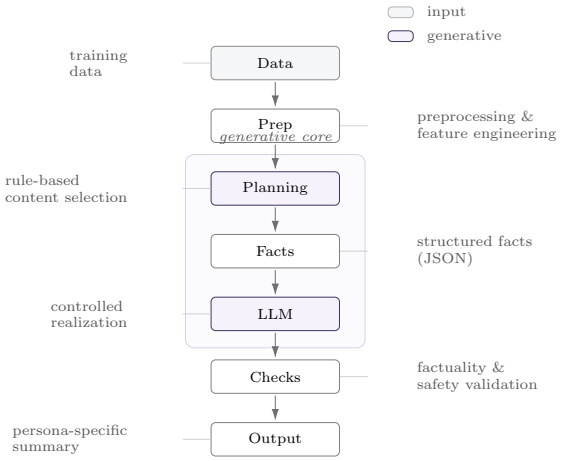

Figure 1: Overview of the proposed persona-specific summary generation pipeline. The central flow represents the main processing stages, while side labels indicate the functional role of each stage. Violet highlights the generative core.

4.2 Rule Layer

The rule layer transforms normalised session features into structured explanatory facts and performs content selection before generation. Selected metrics are mapped to ordinal bands (*low*, *medium*, *high*, *elite*) using metric-specific threshold definitions. These bands provide a traceable intermediate representation for deterministic reasoning about the session profile.

The rule base applies domain-informed conditions to derive compact facts about overall load, internal intensity, speed exposure, and movement dynamics. Each activated rule is represented by a rule identifier, a short explanatory fact, the metric context from which it was triggered, and a priority value used to resolve overlap between simultaneously activated rules. For example, a high distance-per-minute band can activate a load-profile fact indicating substantial running volume. Priority-based selection determines which facts are passed to the generation stage.

In the RULES setting, only the selected rule-derived facts are passed to the generator. In the ENRICHED setting, these facts are accompanied by the normalised raw metrics and the explicit threshold definitions used to derive the bands.

4.3 LLM Realisation

The language model verbalises the selected session interpretation for the target audience under persona-specific style constraints. Generation is prompt-based in all settings: the model receives a structured input for the active condition and an instruction specifying the target persona, expected level of technical detail, and output length. Inputs were serialised as compact JSON-style records containing either the eight metric values, the selected rule-derived facts, or the selected facts together with metric values and threshold definitions.

For coach-facing summaries, the realisation emphasises concise technical interpretation and metric-aware phrasing. For parents, the model uses accessible language and practical framing while avoiding unnecessary technical detail. For players, the summaries use simpler, experience-oriented language adapted to the athlete's perspective.

Across all settings, the LLM expresses the session meaning in audience-appropriate language. The settings differ in how much of this meaning is provided explicitly: in RAW, the model must infer the interpretation from normalised metrics; in RULES, it receives selected rule-derived facts; and in ENRICHED, these facts are supplemented with supporting numerical metrics and threshold definitions. In the ENRICHED experiments reported here, the selected open-weight realisation models are additionally adapted using LoRA.

5 Data and Preprocessing

The dataset consists of 122 anonymised player-session records collected from 24 U15 players aged 14–15 in a youth football environment. Each record corresponds to one player observed in one training or match session and contains session metadata together with wearable-derived indicators of external load, internal intensity, speed exposure, and movement dynamics. The data were processed at the player-session level, so each instance represents one athlete's session profile rather than an aggregate team-level session.

The analytical vector used by the rule layer contained eight core metrics: distance per minute, high-metabolic-load distance per minute, high-speed-running distance per minute, maximal velocity, accelerations per minute, decelerations per minute, time in heart-rate zones 4–5, and metabolic power per minute. Session duration was retained as a separate metadata field, while selected exposure variables were expressed per minute to support comparison across sessions of different length.

For rule-based interpretation, selected normalised indicators were mapped to ordinal bands (*low*, *medium*, *high*, *elite*). Higher-intensity thresholds were based on youth-football reference values, while lower bands were calibrated using the empirical distribution of the collected data and coach input. The resulting banded representation was used to derive structured explanatory facts.

The processed dataset was split according to the experimental design. A held-out subset of ten player-session records was used for the controlled RAW–RULES comparison and for the evaluation of the ENRICHED setting. The remaining 112 records were used as adaptation data for the LoRA-based implementation of ENRICHED. In this setting, the generator also received normalised raw metrics and corresponding threshold definitions together with the rule-derived facts. Records with incomplete or inconsistent data were excluded from subsequent rule-based analysis. All data were processed in anonymised form.

6 Experimental Design

Rather than evaluating the framework as a broad model benchmark, we assess it as a stakeholder-oriented educational communication system. The evaluation has two complementary stages. First, we compare RAW and RULES generation for coach-facing summaries, where unsupported claims and scale-related errors are particularly salient because the summaries require technical interpretation. Second, we evaluate ENRICHED for player-facing summaries through a school-based expert study. Because these stages address different parts of the framework and use different evaluation procedures, their results are interpreted as complementary rather than directly interchangeable.

The study addresses three questions: whether explicit rule grounding improves reliability compared with direct generation from raw metrics; whether the same rule-derived session interpretation can support multiple stakeholder groups; and whether ENRICHED can produce player-facing explanations that remain educationally useful in the evaluated setting.

6.1 Compared Conditions

We compare three generation conditions.

RAW. In the baseline condition, duration-normalised numerical session indicators are provided directly to the language model, which must infer their meaning during generation.

RULES. In the strictly rule-grounded condition, the rule layer converts the normalised metrics into structured explanatory facts. The language model then verbalises only these selected facts for the target audience.

ENRICHED. In this augmented rule-grounded condition, the model receives the selected rule-derived facts together with normalised raw metrics and the explicit threshold definitions used to assign ordinal bands (*low*, *medium*, *high*, *elite*). The additional information therefore refers specifically to supporting metric values and threshold definitions, rather than to broader contextual variables such as individual condition or training history. In the implementation studied here, the selected open-weight models are additionally adapted using LoRA. This condition therefore combines richer input information, threshold exposure, and model adaptation, and is not treated as an ablation of any single factor.

In all settings, the output is a short natural-language session summary intended as an explanatory note rather than a full analytical report. Aligned coach–parent–player reference triplets are used as a qualitative anchor for stakeholder alignment rather than as a separate quantitative benchmark.

6.2 Evaluation Data and Reference Set

The full dataset comprises 122 player-session records collected in a U15 football setting. For the controlled comparison, we used a held-out subset of ten sessions with manually prepared, stakeholder-oriented reference summaries for coaches, parents, and players. The selected sessions are as follows: 3, 9, 10, 12, 14, 23, 24, 25, 50, and 89.

The three reference sets provide the same underlying session interpretation in audience-specific formats. Coach summaries explicitly mention performance indicators and their interpretation, parent summaries retain the main message while simplifying terminology, and player summaries use second-person, experience-oriented language. This aligned triplet design supports qualitative assessment of whether generated summaries preserve the same semantic core while adapting to different stakeholder needs.

The same held-out ten-session subset was used for the RAW–RULES comparison and for the evaluation of ENRICHED. The remaining 112 player-session records were used as adaptation data for the LoRA-based implementation described in Section 6.3. This allocation reflects the constraints of the available dataset and keeps the evaluation subset separate from adaptation. Given the size of the held-out subset, the quantitative scores are used primarily as descriptive indicators within this evaluation setting.

6.3 Models and Generation Setup

The study follows a staged evaluation setup. In the first stage, we replicate the RAW–RULES comparison on the ten-session evaluation subset using LLM-as-a-judge and qualitative expert review. This stage is used to assess the effect of explicit rule grounding and to identify models suitable for further adaptation.

On this basis, we selected two models for the ENRICHED stage: **Mistral-3B-Instruct** and **Llama-8B-Instruct**. Mistral-3B-Instruct was selected as the stronger variant within the Mistral family in the replicated comparison, while Llama-8B-Instruct

was retained as a representative of a different open-weight model family. Hosted commercial models were not included in this stage because LoRA adaptation was only feasible for open-weight models in our implementation.

Generation parameters were held constant within each model across the RAW and RULES conditions. All outputs were generated from the same ten-session subset using matched prompts and identical output-length constraints. For the ENRICHED condition, the selected models were adapted with LoRA on the remaining 112 sessions and evaluated on the held-out ten-session subset.

For LoRA adaptation, we constructed persona-specific prompt–completion pairs from the non-evaluation records only. Each input used the same ENRICHED representation as at inference time: a standardised session header, normalised raw metrics, rule-derived observations, threshold definitions, and the persona instruction. Each target completion was a short Polish summary for the selected stakeholder. In our implementation, these target summaries were obtained through teacher-model distillation using the strongest prompt-based system available in the pipeline. The authors then screened them for formatting consistency and obvious factual issues. The held-out evaluation subset was excluded from this adaptation process.

Given the limited adaptation set and the absence of a dedicated development split, we used one conservative LoRA configuration for both adapted open-weight models rather than performing a broad hyperparameter search. We used QLoRA with 4-bit NF4 quantization and bf16 compute, LoRA rank $r = 8$, $\alpha = 16$, dropout $= 0.05$, and attention-projection target modules (q_proj, k_proj, v_proj, o_proj). Training used paged AdamW 8-bit, learning rate 1×10^{-4}, per-device batch size 1, gradient accumulation 8, 5 epochs, maximum sequence length 2048, and random seed 42.

Generation remained prompt-based in all conditions. Within each summary-generating model, the decoding setup was kept fixed across the compared prompting conditions. For the selected open-weight models carried forward to the ENRICHED stage, we used low-temperature decoding with fixed output-length limits: temperature 0.05 for Mistral-3B-Instruct and 0.07 for Llama-8B-Instruct, with a maximum of 400 generated tokens. Persona prompts and output-length constraints were kept unchanged within each model family.

Prompt format example. In the ENRICHED setting, the model input was serialised as a compact JSON-style structure containing: (i) a standardised session header, (ii) normalised raw metrics, (iii) rule-derived textual observations, and (iv) threshold definitions. This structured input was followed by a persona-specific instruction controlling audience, expected technicality, and target length.

6.4 Evaluation Dimensions

The evaluation combines quantitative and qualitative perspectives to capture interpretive reliability and educational usefulness across the studied conditions.

Interpretive reliability. First, we assess whether the generated summaries remain faithful to the source session profile. This dimension focuses on factual faithfulness and numerical consistency, with special attention to unsupported inferences, incorrect magnitude interpretation, and misleading summary framing.

LLM-as-a-judge assessment. Second, in the RAW–RULES comparison, we evaluate coach-facing summaries using a fixed LLM-as-a-judge rubric. GPT-4o mini was used as the judge model for all candidate summaries. Summaries are assessed with respect to faithfulness, coach value, narrative quality, output hygiene, conciseness, and language quality, together with an aggregate *Overall* score. Output hygiene refers to whether the summary avoids malformed structure, irrelevant boilerplate, contradictory statements, excessive repetition, and formatting artefacts. The resulting scores are interpreted descriptively, with emphasis on within-generator differences between RAW and RULES.

Reference-based stakeholder alignment analysis. Third, we use the manually prepared coach–parent–player reference triplets as a qualitative interpretive anchor for examining whether generated summaries preserve the same underlying session meaning while adapting terminology, explicitness, and communicative stance across audiences. This component supports qualitative analysis of stakeholder alignment.

School-based expert evaluation. Finally, ENRICHED is evaluated in a school-based expert study focused on player-facing summaries. Nine

physical education teachers from primary, technical secondary, and general secondary schools evaluated both selected models on all ten held-out sessions using a 0–5 scale. The evaluation was blind with respect to model identity, and raters had access to the source session data. The assessed dimensions were: (i) content correctness, (ii) clarity and understandability, and (iii) practical usefulness. An aggregate *Overall* score is additionally reported as the arithmetic mean of these three ratings.

7 Results

7.1 Reference-Based Stakeholder Alignment

The aligned coach, parent, and player reference summaries show that stakeholder adaptation in this task involves more than stylistic variation. Across the ten shared sessions, the three variants preserve the same underlying interpretation while varying terminology, explicitness, and communicative stance. Coach-facing references name relevant indicators such as distance per minute, metabolic power, time in high heart-rate zones, high-speed running (HSR), high-metabolic-load distance (HMLD), and acceleration/deceleration counts. Parent-facing references express the same session meaning through accessible descriptions of effort and rhythm, while player-facing references use second-person, experience-oriented language linked to how the athlete may have experienced the session. Sessions dominated by continuous running volume, internal load, stop–start movement, low load, or speed exposure are therefore consistently described across audiences without changing the underlying interpretation. For example, internally demanding sessions remain framed as heart-rate-intensive rather than speed-driven, whereas stop-start sessions are consistently described as mechanically demanding across the three audience variants. The reference triplets serve as a qualitative anchor for audience alignment, not merely as stylistic examples.

Coach-facing LLM-as-a-judge evaluation. Table 1 reports the coach-facing comparison between the RAW and RULES conditions across all evaluated summary-generating models. Scores are averaged over the ten-session evaluation subset and are interpreted descriptively. *Faith.* refers to *Faithfulness*, or semantic consistency with the source session profile; *Coach* denotes practical coach value; and *Narr.*, *Hyg.*, *Conc.*, *Lang.*, and *Overall* denote narrative quality, output hygiene, conciseness, language quality, and the aggregate judge score, respectively.

Table 1: LLM-as-a-judge scores assigned by GPT-4o mini for coach-facing summaries, averaged over the ten-session evaluation subset. Rows denote summary-generating models under the RAW and RULES conditions. Scores are on a 0–10 scale. Higher is better on all dimensions.

Model	Variant	Faith.	Coach	Narr.	Hyg.	Conc.	Lang.	Overall
GPT-5.2	RAW	5.2	6.0	5.3	6.7	6.2	7.8	5.1
GPT-5.2	RULES	7.7	7.1	6.8	9.2	7.5	8.8	7.1
Mistral-3B	RAW	2.6	4.1	4.5	2.8	5.2	6.8	2.8
Mistral-3B	RULES	7.7	7.2	6.9	9.3	7.4	9.0	7.1
Mistral-14B	RAW	3.2	4.4	4.5	3.5	5.4	7.0	3.3
Mistral-14B	RULES	6.7	6.8	6.4	7.9	6.9	8.5	6.4
Bielik-1.5B	RAW	0.5	1.3	1.0	0.7	1.9	3.5	0.7
Bielik-1.5B	RULES	4.0	5.0	4.5	4.6	5.4	6.8	3.8
Bielik-11B	RAW	2.3	3.9	3.1	2.3	4.3	6.3	2.3
Bielik-11B	RULES	5.6	5.8	6.0	6.7	6.5	7.8	5.2
Llama-1B	RAW	1.8	2.4	2.5	2.0	3.7	5.5	2.0
Llama-1B	RULES	4.1	5.1	4.7	5.1	5.1	6.5	4.1
Llama-8B	RAW	1.5	2.3	2.1	1.5	3.8	5.7	1.5
Llama-8B	RULES	6.1	6.2	5.7	8.1	6.4	8.1	5.9

Interpretation of Table 1. Across all evaluated summary-generating models, the RULES condition received higher mean *Overall* scores than RAW, with the largest gains in *Faithfulness*, *Coach Value*, and *Output Hygiene*. Improvements in *Language Quality* were smaller, suggesting that rule grounding primarily improved semantic faithfulness, communicative discipline, and practical usefulness rather than surface fluency alone.

Given the ten-session evaluation subset, Table 1 is best interpreted as a within-generator comparison between RAW and RULES, not as a definitive ranking of generators. The highest *Overall* means in the RULES condition were observed for GPT-5.2 and Mistral-3B. These results motivated the selection of Mistral-3B for the augmented ENRICHED stage, with Llama-8B retained as a representative model from a different open-weight family.

School-based expert evaluation of player-facing summaries. Table 2 summarizes the school-based expert evaluation of player-facing summaries in the augmented ENRICHED setting. Nine physical education teachers rated the outputs on a 0–5 scale for *Content Correctness*, *Clarity*, and *Practical Usefulness*. Inter-rater agreement was estimated using ICC(A,9) and indicated limited agreement across dimensions. For *Content Correctness*, *Clarity*, and *Practical Usefulness*, the ICC(A,9) values were 0.434, 0.158, and 0.497 for Mistral-3B, and 0.000, 0.084, and 0.144 for Llama-8B, respectively.

Table 2: School-based expert evaluation of player-facing summaries in the augmented ENRICHED setting. Ratings are on a 0–5 scale. The *Overall* column reports the mean of Content Corr., Clarity, and Pract. Usef. Higher is better on all dimensions.

Model	Content Corr.	Clarity	Pract. Usef.	Overall
Mistral-3B	4.6	3.9	4.5	4.3
Llama-8B	4.4	3.4	4.5	4.1

Interpretation of Table 2. Both models received high teacher ratings for *Content Correctness* and *Practical Usefulness*, suggesting that the player-facing summaries produced in the evaluated augmented setting were factually controlled and practically meaningful. Mistral-3B scored slightly higher on *Content Correctness*, *Clarity*, and *Overall*, while both models obtained the same *Practical Usefulness* score. The comparatively lower *Clarity* ratings indicate that simplifying session meaning for young athletes remains an important design challenge, consistent with treating player-facing generation as both factual summarisation and educational communication.

7.2 Qualitative Error Analysis

The qualitative analysis revealed three recurring error patterns in RAW generation. First, some summaries overstated conclusions not fully supported by the data, especially when high overall cost was verbalised as sustained maximal intensity. Second, some outputs confused the dominant session dimension, for example by describing a session as speed-oriented when the reference interpretation emphasised internal load or continuous volume. Third, persona adaptation sometimes remained superficial, especially when parent- or player-facing outputs retained overly technical phrasing.

The RULES condition reduced these issues by separating interpretation from realisation. Because the dominant session profile was established before generation, the resulting summaries more consistently preserved audience-invariant meaning while still allowing audience-specific wording. This distinction is important for educational communication: a useful stakeholder explanation should differ in form across coach, parent, and player, but not in its underlying interpretation of the session.

In ENRICHED, raw metrics and explicit band definitions provided a richer explanatory input, but also allowed more phrasing variation than the strictly rule-grounded condition. Because this implementation combined enriched conditioning with LoRA-based adaptation, the results should be interpreted as feasibility evidence for a more flexible applied setup rather than as an isolated estimate of the contribution of enriched conditioning alone. The school-based evaluation suggests that, in this configuration, player-facing summaries can remain factually controlled and practically useful.

8 Discussion

The results support the central design choice of the paper within the evaluated setting: separating analytical interpretation from linguistic realisation can improve the reliability of stakeholder-facing explanations derived from structured training metrics. When language models generate directly from raw numerical inputs, they may produce fluent summaries that overstate conclusions, confuse the dominant session profile, or adapt style without preserving the intended meaning. By contrast, the rule-grounded pipeline constrains interpretation before generation, making the resulting summaries more stable and easier to control.

This effect is especially relevant in the present educational setting, where the goal is not only to produce natural-sounding text, but to help stakeholders understand what recorded metrics mean in practice. The system therefore functions as an explanatory interface between sensor-based analytics and stakeholder understanding. The aligned coach–parent–player reference set further shows that useful adaptation should change wording, level of detail, and communicative stance while preserving the same semantic core across audiences.

The ENRICHED setting adds an applied perspective. It combines validated rule-derived facts, supporting raw metrics, explicit band definitions, and LoRA-adapted open-weight models; the present study therefore cannot isolate the contribution of any single component. In the evaluated setting, the school-based expert evaluation suggests that ENRICHED can produce player-facing summaries that remain correct, understandable, and practically useful.

More broadly, the study supports a methodological point relevant to educational NLP: systems that explain structured numerical data should not be evaluated solely on fluency or stylistic quality. In this context, faithfulness, audience appropriateness, communicative restraint, and educational usefulness are equally important, especially when explanations are intended for young stakeholders

and must remain understandable without becoming overstated, diagnostic, or misleading.

9 Limitations

Several limitations should be acknowledged. First, the dataset comes from a single youth football environment and contains 122 player-session records from a relatively small group of U15 players. The controlled evaluation used a held-out subset of ten sessions, while the remaining 112 records were used for LoRA adaptation. This split reflects the constraints of the available dataset, but limits robust model assessment; we did not conduct cross-validation or repeated-split evaluation. For the coach-facing LLM-as-a-judge comparison, the small ten-session subset precludes strong statistical claims, so the results are interpreted descriptively.

Second, the system relies on a manually designed rule base. While this improves interpretability and control, it requires domain expertise and may hinder transfer to new settings. Duration-normalised indicators and ordinal bands support comparability across sessions, but may reduce information relevant to practical interpretation, such as absolute session duration or individually calibrated physiological zones.

Third, ENRICHED combines selected rule-derived facts, supporting raw metrics, explicit threshold definitions, and LoRA-adapted open-weight models. The present design therefore cannot isolate whether the observed quality of player-facing summaries is attributable to richer input information, threshold exposure, model adaptation, or their combination.

We did not compare LoRA adaptation against few-shot in-context prompting with the same open-weight models. Given the relatively small size of the adaptation set, such a comparison would be important before making stronger claims about the necessity of parameter-efficient fine-tuning. We therefore interpret the LoRA results as evidence of feasibility for a low-resource-adapted configuration rather than as proof that LoRA is superior to few-shot prompting.

Fourth, the school-based evaluation involved nine physical education teachers and focused only on player-facing summaries. Although this provides an educationally meaningful expert perspective, it is not a large-scale end-user study with players themselves and does not provide equally strong evidence for parent- or coach-facing outputs in ENRICHED.

Finally, the LLM-as-a-judge evaluation was used as a scalable diagnostic assessment, not as a replacement for human evaluation. Although the same GPT-4o mini judge and fixed rubric were used for all candidate summaries, we did not assess judge reliability through repeated judging runs or calibration against a large independent human-rated benchmark. The RAW–RULES comparison and the school-based expert study should therefore be interpreted as complementary rather than directly comparable stages of evaluation.

10 Conclusion

This paper introduced a rule-grounded framework for generating short, audience-appropriate explanations of youth football training data for coaches, parents, and players. By separating validated interpretation from linguistic realisation, the approach can improve control over what is said while allowing explanations to be adapted to different stakeholders. In the evaluated setting, rule grounding increased reliability relative to direct generation from raw metrics, and the evaluated ENRICHED variant produced player-facing summaries that teachers judged to be correct and practically useful. We therefore position the framework not simply as a generation pipeline but as an interface that supports data literacy in educational communication about sports analytics. Future work should test the approach on larger datasets and evaluate its effects on stakeholder comprehension, trust, and learning.

Ethical and Privacy Considerations

The study used anonymised training session data provided by an external partner. No data that could directly identify individual players were used in the generation or evaluation process. The system was designed to facilitate communication and interpretation, rather than for diagnosis or decision-making.

Acknowledgements

The authors would like to thank JUSTWIN sp. z o.o. for their collaboration and support throughout the research process. The authors remain solely responsible for the interpretation of the results and the conclusions presented in this paper.

References

Erling Algroy, Halvard Grendstad, Amund Riiser, Tone Nybakken, Atle Hole Saeterbakken, Vidar Andersen, and Hilde Stokvold Gundersen. 2021. Motion analysis of match play in U14 male soccer players and the influence of position, competitive level, and contextual variables. *International Journal of Environmental Research and Public Health*, 18(14):7287.

Siti A. Atan, Andrew Foskett, and Ajmol Ali. 2016. Motion analysis of match play in new zealand U13 to U15 age-group soccer players. *Journal of Strength and Conditioning Research*, 30(9):2416–2423.

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logicnlg: Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7929–7942, Online. Association for Computational Linguistics.

Terje Dalen and Havard Loras. 2019. Monitoring training and match physical load in junior soccer players: Starters versus substitutes. *Sports*, 7(3):70.

Martina Galletti and Valeria Cesaroni. 2025. From end-users to co-designers: Lessons from teachers. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 505–516, Vienna, Austria. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation. *Journal of Artificial Intelligence Research*, 61:65–170.

Marcus P. Hannon, Nicholas M. Coleman, Lloyd J. F. Parker, John McKeown, Viswanath B. Unnithan, Graeme L. Close, Barry Drust, and James P. Morton. 2021. Seasonal training and match load and micro-cycle periodization in male premier league academy soccer players. *Journal of Sports Sciences*, 39(16):1838–1849.

J. A. Harley, C. A. Barnes, M. Portas, R. Lovell, S. Barrett, D. Paul, and M. Weston. 2010. Motion analysis of match-play in elite U12 to U16 age-group soccer players. *Journal of Sports Sciences*, 28(13):1391–1397.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, K. V. Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Clifford Konold, Trevor L. Higgins, Stephen J. Russell, and Khalil Khalil. 2015. Data seen through different lenses. *Educational Studies in Mathematics*, 88(3):305–325.

Ellen B. Mandinach and Edith S. Gummer. 2016. What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education*, 60:366–376.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Konstantinos Michos, Maria-Luisa Schmitz, and Dominik Petko. 2023. Teachers' data literacy for learning analytics: A central predictor for digital data use in upper secondary schools. *Education and Information Technologies*, 28:14453–14471.

Samraj Moorjani, Adit Krishnan, Hari Sundaram, Ewa Maslowska, and Aravind Sankar. 2022. Audience-centric natural language generation via style infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1919–1932, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sankalan Pal Chowdhury, Nico Daheim, Ekaterina Kochmar, Jakub Macina, Donya Rooein, Mrinmaya Sachan, and Shashank Sonkar. 2025a. Large language models for education: Understanding the needs of stakeholders, current capabilities, and the path forward. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1–10, Vienna, Austria. Association for Computational Linguistics.

Sankalan Pal Chowdhury, Terry Jingchen Zhang, Donya Rooein, Dirk Hovy, Tanja Käser, and Mrinmaya Sachan. 2025b. Educators' perceptions of large language models as tutors: Comparing human and AI tutors in a blind text-only setting. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 356–374, Vienna, Austria. Association for Computational Linguistics.

Luiz Henrique Palucci Vieira, Christopher Carling, Fabio Augusto Barbieri, Rodrigo Aquino, and Paulo Roberto Pereira Santiago. 2019. Match running performance in young soccer players: A systematic review. *Sports Medicine*, 49(6):917–931.

Kseniia Petukhova and Ekaterina Kochmar. 2025. Intent matters: Enhancing AI tutoring with fine-grained pedagogical intent annotation. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 860–872, Vienna, Austria. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6908–6915.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

Hieu Tran, Zonghai Yao, Lingxi Li, and Hong Yu. 2025. Readctrl: Personalizing text generation with readability-controlled instruction learning. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 19–36, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Natercia Valle, Pavlo Antonenko, Kara Dawson, and Anne Corinne Huggins-Manley. 2021. Staying on target: A systematic literature review on learner-facing learning analytics dashboards. *British Journal of Educational Technology*, 52(4):1724–1748.

Sam Wiseman, Stuart Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.