DOI: https://doi.org/10.36910/6775-2524-0560-2026-64-01
UDC 004.94:796.332
**Hotovych Volodymyr**, Ph.D., Associate Professor
https://orcid.org/0000-0003-2143-6818
**Sorokivskyi Oleksandr**, PhD student
https://orcid.org/0009-0006-6477-5878
Ternopil Ivan Puluj National Technical University, Ternopil, Ukraine

# AUTOMATED TACTICAL ANALYSIS OF FOOTBALL MATCHES USING LARGE LANGUAGE MODELS

**Hotovych V., Sorokivskyi O. Automated tactical analysis of football matches using Large Language Models.** The paper proposes a system for automated tactical analysis of football match data using large language models. The system uses two LLMs: a main model, which analyzes the input data and draws conclusions, and a second model (a judge model) that checks the conclusions made by the main model. Before the analysis, the input data are pre-processed into an enriched match document – a data structure that holds values aggregated from the input dataset. The system follows an evidence-based principle: every conclusion of the main model must be backed by references to facts, that is, fragments of the input information. The system can also search for patterns – recurring tactical actions of the same team across several matches. Its work combines four stages: computing metrics that describe the actions of football teams from the input data; using an LLM to generate conclusions about a team's tactics (insights); automatically assessing the quality of these insights with the «LLM-as-a-judge» scheme; and dropping insights that fail this check. The system was tested on data from the open source StatsBomb. The results produced by the system are shown to be valid.
**Keywords:** Large Language Models, tactical analysis of football matches, LLM-as-a-judge, insight validation, pattern search, sports analytics, StatsBomb.

**Готович В. А., Сороківський О. В. Автоматизований тактичний аналіз футбольних матчів з використанням великих мовних моделей.** В роботі пропонується система для автоматизованого тактичного аналізу даних про футбольні матчі на основі застосування великих мовних моделей. Дана система складається із двох LLM: основної (здійснює аналіз вхідних даних та генерує висновки) та допоміжної (т. зв. модель-суддя, що валідує висновки, отримані основною моделлю). При цьому відбувається попередня обробка вхідних даних (отримання т. зв. збагаченого документа матчу – структури даних з агрегованими на основі вхідного набору даними). Система функціонує із застосуванням принципу доказовості (кожен висновок основної моделі повинен підкріплюватися посиланнями на факти – фрагменти отриманої на вхід інформації). Також пропонована система дозволяє здійснювати пошук патернів (повторюваних тактичних дій однієї і тієї ж команди впродовж кількох матчів). Робота даної системи поєднує чотири етапи: обчислення на основі вхідних даних метрик, які характеризують дії футбольних команд, генерація за допомогою LLM висновків про застосовувану командою тактику (інсайтів), автоматизована оцінка якості цих інсайтів за схемою «LLM-як-суддя», відкидання інсайтів, які не пройшли валідацію. Оцінку роботи системи здійснено на основі даних, отриманих з відкритого джерела StatsBomb. Продемонстровано адекватність отриманих системою результатів.
**Ключові слова:** великі мовні моделі, тактичний аналіз футбольних матчів, LLM-як-суддя, валідація висновків, пошук патернів, спортивна аналітика, StatsBomb.

**Problem statement.** Tactical analysis is an important part of the coaching staff's work in professional football. Understanding how a team presses, builds attacks, moves between phases of play, involves individual players, and so on lies at the core of pre-match preparation, in-match adjustments, and post-match review. However, this analysis is still often done by hand: coaches and analysts watch video, study statistics, and compile reports [24, 25]. This approach is labor-intensive and hard to scale, especially when many matches have to be analyzed. Developing tools, methods, and algorithms for the automated analysis of a team's tactics during a football match is therefore a relevant task.

Modern tools make it possible to simplify and automate this task. In particular, open match datasets are now available for analysis, for example: StatsBomb Open Data [26] – a public repository with detailed event logs collected for over 3,000 matches in more than 40 competitions; Wyscout [27] – a dataset covering seven competitions of the 2017/18 season; and SkillCorner [28] – data recorded from video broadcasts.

This work is based on the StatsBomb Open Data dataset, downloaded from the official GitHub repository. In it, each event is recorded as a separate entry linked to a player, a position on the field (coordinates), and a timestamp. The data are structured in JSON format. Access to them is open and provided on a non-commercial basis.

One important result of a match's tactical analysis is a set of numerical indicators – metrics. They describe both the actions of the team as a whole and the actions of individual players. Interpreting metric values has usually required an expert (an experienced analyst). Large language models (LLMs) can handle this part of the tactical-analysis task, because they are able to analyze structured numerical data and generate coherent text [8–10] – that is, to move from abstract numbers to concrete meaning. The main risk of this

approach, common to LLMs in general, is hallucination [21]: the model may produce plausible tactical conclusions that are not supported (or are poorly supported) by the data provided for analysis, or that directly contradict it. Tools and methods are therefore needed to reduce this risk.

**Analysis of Research and Publications.**

**The use of LLMs in football-analytics tasks.** The use of large language models for sports-analytics tasks is a new and promising research direction. Work in this area covers match-outcome prediction, tactics interpretation, and the generation of analytical texts in natural, human-readable language.

The closest work to the one proposed here is [8]. In it, the authors fine-tuned a GPT-3-scale model on English Premier League match data, teaching it to predict a number of tactical actions. Although [8] shows that LLMs can detect patterns in how match events unfold, it works in a next-token prediction mode and does not require the model to reference numerical evidence; that is, hallucinations in the LLM output are not detected automatically here. This is a fundamental shortcoming that the system proposed by the authors of this paper avoids.

The review [9] systematizes work on applying LLMs to football-match analytics tasks and identifies three types of the most popular tasks today, together with their corresponding systems: retrieving information from the processed match data, predicting how events will unfold, and generating descriptive insights. The authors of [9] note that the category of descriptive insights remains the least studied to date. FootGPT, a model developed by the authors of [10], has about 1 billion parameters. It was combined with a powerful commercial LLM that generated a training dataset for it. Such an approach requires considerable effort to prepare the training data, and the result depends heavily on the format of the input data itself. The system proposed by the authors of this paper is free of this shortcoming: instead of a fine-tuning stage, pre-structured data are passed directly into the context of the request to the main LLM. Although some computational cost is required for this data pre-processing, the approach is still more flexible, more reproducible, and less resource-intensive.

Another research direction is the automatic description of in-match events in natural language. In [11], an open LLM was fine-tuned to generate real-time match commentary. In [12], an LLM is used to explain the rating of a particular shot: the model computes the probability of a goal and describes in words why this probability is high or low, taking into account the distance from the point of contact with the ball to the goal, the position of the attacking player, and the pressure on that player from the opponents' defenders (pressing). Both approaches [11, 12] are close in idea to the one proposed by the authors of this paper – to turn numerical match data into human-readable text. However, the authors of [11, 12] focus on explaining individual events rather than forming tactical conclusions about the match as a whole, and they do not automatically check the validity of each statement produced by the LLM.

A notable approach is presented in [13]: to retrieve facts from an array of match data, the SoccerRAG model combines vector search with an LLM-based answering interface and includes a verification chain to reduce hallucinations. Most directly related to ours is work [14], in which a RAG-based approach [23] achieved an accuracy of 63.3% correct answers to descriptive user queries about matches. Compared with it, our approach makes it possible to assess the quality of the insights obtained from the LLM. Unlike the systems proposed in the literature, which answer specific user queries, our system proactively generates tactical insights from match data, without the user needing to formulate a query.

**Analytical systems without LLMs and their limitations.** Before LLMs appeared, football analytics was dominated by machine-learning and clustering methods. Analysis systems based on them produce numerical results but leave the interpretation of these results to a human analyst. For example, [15] applies a player-clustering method that made it possible to identify 11 functional player types based on 27 features. In [16], each team attack is represented as a graph of passes between players, and on this basis a method is proposed for automatically identifying the players most important to a team's attacking actions. In [17], a method is proposed for detecting the typical tactical schemes used by teams by clustering match data that have first been transformed by an artificial neural network into compact numerical vectors.

These approaches share a common limitation: none of them solves the task of automatically formulating tactical conclusions about a match in natural language while accounting for the risk of hallucination. This is where the system proposed by the authors of this paper qualitatively differs from them.

**The aim of the study.**

The aim of this paper is to develop a system for the automated analysis of previously collected information about football matches that generates tactical conclusions (insights) in natural language and

supports each such conclusion with evidence – references to facts, that is, fragments of the input information. The system must also automatically validate (check) the correctness of the conclusions it produces.

In this work, a tactical insight is understood as a short natural-language statement that describes a specific behavior of a team or an individual player from a tactical point of view, supported by references to numerical match metrics and by an explanation of how these numbers confirm the conclusion made.

The central principle of the developed system is evidence: every generated insight must contain references to specific fields (elements) of the input dataset that hold metric values. These references are checked by the system automatically. If a field does not exist in the dataset, the insight is sent back for correction or discarded. The quality of accepted insights is assessed by a separate LLM instance – the judge – against two criteria: reliability (whether the insight matches the stated numerical values) and relevance (whether these numerical values are appropriate evidence for the given insight).

The system proposed in this paper combines four stages that have not previously been considered together within a single study:

- computing metric values from a match dataset;
- generating insights about teams' tactical actions using an LLM;
- automated quality assessment of the resulting insights with the «LLM-as-a-judge» scheme;
- reducing the number of hallucinations.

Compared with TacticalGPT [8], we move from predicting the next event to a tactical narrative with verified evidence and add a full evaluation cycle. Compared with the RAG-based approach [14], we move from the accuracy of query answers to the quality of insights (reliability + relevance) and add a software-based reference validator. Compared with the clustering approaches [15, 16], we apply a further step in the analytical chain: the automatic generation, checking, and evaluation of tactical conclusions in natural language.

**Description of the proposed system.**

**General algorithm.** The algorithm of the proposed system consists of five main steps (Fig. 1).

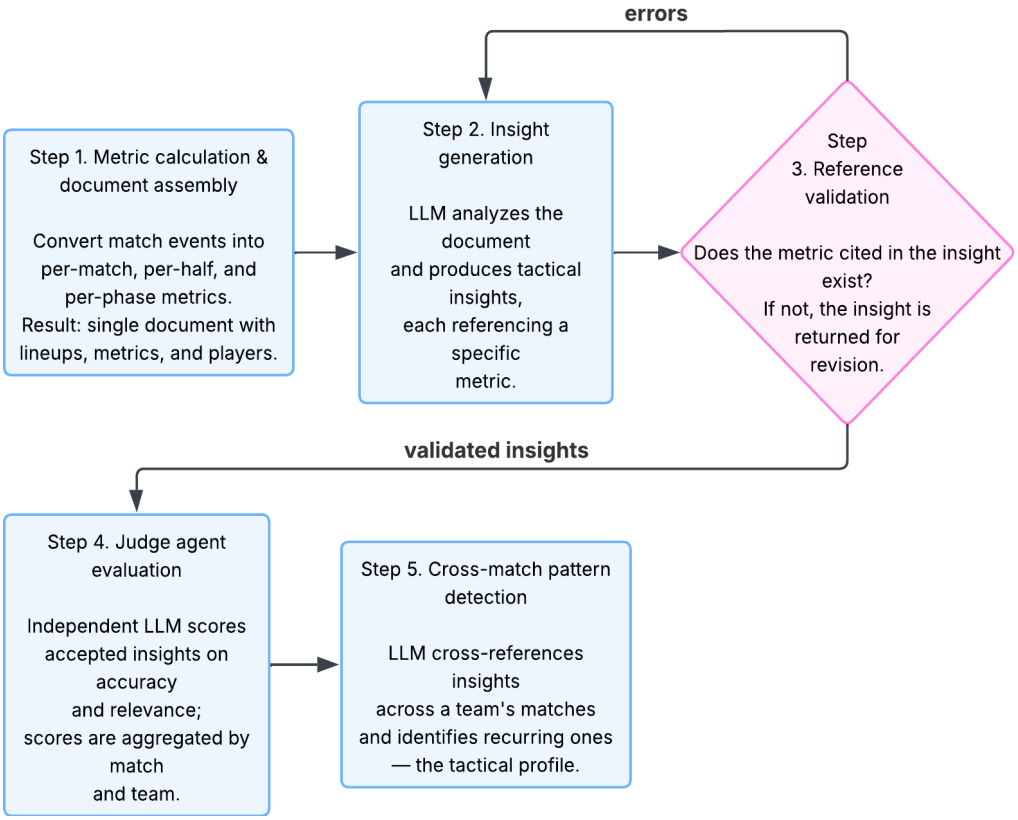


Fig. 1. General algorithm of the system's operation

From the dataset of a particular match, the numerical values of the metrics are computed and a file with a clear structure – the enriched match document – is formed. This document, together with the

corresponding instruction (prompt), is passed to the LLM, which generates conclusions about the teams' tactical actions (insights) with explicit references to specific fields of the document. Each such reference is checked programmatically (a deterministic step): if a field does not exist in the document, the insight is returned to the LLM for review and possible correction, or it is discarded. The quality of accepted insights is assessed by a separate LLM (the judge) on the basis of the corresponding prompt, against the criteria of reliability and relevance. Finally, when several matches of the same team are analyzed, the LLM compares insights across matches and identifies recurring insights.

**Data source and match sample.** The system analyzes the open StatsBomb dataset. Unlike aggregated statistics, here every game event is recorded as a separate entry (fact) linked to a player, to the coordinates on the football pitch where it occurred, and to a timestamp relative to the start of the game. The pitch is described in a coordinate system: the x-axis $x \in [0,120]$ is the length of the pitch in the direction of attack in yards, and the y-axis $y \in [0,80]$ is its width in yards.

The match data include the following event types: passes, shots, pressing, a player carrying the ball forward, duels, player substitutions, and changes in the tactical positioning of players on the pitch. For every shot on goal, StatsBomb also provides a pre-computed value of the xG metric. The data for each match contain facts about the lineup (player identifiers, positions, numbers) and player-repositioning events that explicitly record changes between the tactical schemes used by a team – for example, from a 4-3-3 to a 3-5-2 scheme.

A single «raw» StatsBomb event – for example, one pass – is provided in the supplementary repository [29].

Each such event is an atomic and independent fragment of data. On their own, such records contain no aggregated indicators and say nothing about the tactics used by a team. To draw certain conclusions about tactics, one has to collect and analyze hundreds of such events. This is exactly what the document-enrichment step performs (see the subsection «The enriched match document»).

**Match sample.** To evaluate the system's operation, we selected data on matches played in seven competitions across four leagues: the Bundesliga, the FA Women's Super League, the Indian Super League, La Liga, Ligue 1, the Premier League, and Serie A. For each of the 23 teams, the season with the largest amount of available data on played matches was chosen automatically, after which up to ten matches were analyzed in chronological order. Processing the match data while preserving the chronological order makes it possible to detect a specific team's recurring tactical actions (patterns) across matches (the last step of the system's algorithm). In total, data on 240 matches involving 23 teams were processed.

**The enriched match document.** The «raw» match dataset is too detailed and too large to be passed directly to an LLM: it is unstructured and contains no aggregated indicators. It is therefore not suitable for processing by an LLM. For this reason, the source dataset of each match is transformed algorithmically – by deterministic computation of metrics using predefined formulas (without any LLM) – into an enriched document: a structured JSON object that combines information about the team lineups, the computed metric values, and the tactical segmentation of the match (the division of the match into time fragments during which a given team maintained a particular tactical scheme of actions). This document serves as the single context for all subsequent LLM steps.

An abridged example of such a document – for the match AC Milan – Palermo (Serie A, 2015/2016) – is provided in the supplementary repository [29].

Each field of such a document can be accessed programmatically, in dot notation. For example, field phase_metrics.first_half.pressing_intensity.home.avg_recovery_time corresponds to the value 13.24, and field player_metrics[Romagnoli].progressive_carries to the value 6.

The complete set of fields, with their types and descriptions, is documented in the enriched match document provided in the supplementary repository [29].

**Tactical phases.** The match is divided into tactical phases according to events that change the arrangement of a team's players on the pitch [7]. Each phase covers a fragment of the total playing time during which the tactical scheme and the team's lineup remained unchanged, and the enriched document always contains a separate set of metric values for it. This allows the system to track how coaching decisions – for example, a switch to a more defensive scheme after conceding a goal – are reflected in the metric values.

The boundary in time between adjacent phases is defined by the moment when the StatsBomb log records a change in the tactical scheme or lineup of one of the teams. All events in the interval from the

start of the match (or from the previous phase change) up to this moment belong to a single phase. The corresponding time values are stored in the fields start_minute and change_minute, while the change itself is recorded by the new formation and lineup of the team (see the subsection «Building the document»). If both teams change their on-pitch player arrangement at the same time – for example, at the start of the second half – both changes are recorded as the boundary of a single phase, not two separate ones. The metrics for each phase are computed using the same formulas as the per-half or whole-match metrics (see the subsection «Computed metrics»), but only on the events that fall within that phase.

**Computed metrics.** All metrics are computed separately for each team and aggregated at three levels: the full match, the first and second halves, and each individual tactical phase of the match (Table 1). Below, each of them is described briefly.

Table 1. Summary of the metrics

| Metric | Computation | Interpretation |
|---|---|---|
| xG | Sum of xG over all of the team's shots | Higher value – better-quality attack |
| PPDA | Ratio of the opponent's passes in the attacking third to the team's defensive actions in the same part of the pitch | Lower PPDA – more aggressive pressing |
| ΔPPDA | Difference between second-half and first-half PPDA | A negative value – pressing intensified in the second half |
| Possession percent | Ratio of the team's ball-possession time to the total match duration | Higher value – more time on the ball |
| xT of an action | Difference in xT between the ball's final and initial positions | A positive value – the ball was moved into a more dangerous zone for the opponent's goal |
| Recovery time | Average time between a pressing event and the next ball reception by the same team | Shorter time – more effective pressing |
| xT/90 | Ratio of a player's xT to the minutes they played, divided by 90 | The player's normalized contribution to the team's actions, independent of time spent on the pitch |

Expected goals (xG) reflects the quality of a team's attacking actions [1, 2]. For every shot on the opponent's goal, StatsBomb provides a pre-computed xG value – the probability of scoring a goal given the position on the pitch and the game context (the arrangement of the teams' players). Summing over all shots gives the team's total xG during the match.

PPDA (Passes Per Defensive Action) measures pressing intensity [3] as the ratio of the number of the opponent's passes to the number of the team's own defensive actions, counted only in the final 40 metres of the pitch before the opponent's goal – exactly where active pressing takes place:

$$PPDA = \frac{N_{opp.passes\,(final\ 40\,m.)}}{N_{def.actions\,(final\ 40\,m.)}},$$

where $N_{opp.passes\,(final\ 40\,m.)}$ is the number of the opponent's passes made in the final 40 metres of the pitch in front of the opponent's goal, and $N_{def.actions\,(final\ 40\,m.)}$ is the number of the team's own defensive actions performed in the same zone. Both quantities are counted over the aggregation interval under consideration: the full match, a half, or an individual tactical phase.

Defensive actions include pressing, interceptions and tackles, and fouls by the opposing team's players. A lower PPDA value means more aggressive pressing. In addition, $\Delta PPDA$ is computed – the difference between the values of this metric during the first and second halves. This is needed to detect changes in a team's pressing intensity over the course of a match [4].

Ball possession percentage is the share of playing time during which the ball was with each team's players:

$$Possession = \frac{S_{team}}{S_{total}} * 100\%,$$

where $S_{team}$ is the total duration of the ball-possession sequences of the team under consideration, and $S_{total}$ is the total duration of the possession sequences of both teams, that is, the total playing time during which the ball was in play. Both values are measured in seconds and are computed from the timestamps of the match events.

Pressing intensity over time. To capture how pressing changes during a match, the number of pressing events is counted within 10-minute intervals (0–10 min, 10–20 min, …, 80–90 min). In addition, the average time between the moment a particular pressing event occurs and the moment of the next successful ball recovery by the same team is computed, as an indirect indicator of pressing effectiveness.

Expected threat ( $xT$ ) reflects how much a specific on-the-ball action moves the team closer to the opponent's goal [5]. Singh's model is used: the pitch is conventionally divided into a 12×8 grid of cells, each of which stores the probability of scoring from that zone. For a pass or a carry:

$$xT_{action} = xT(c_{destination}) - xT(c_{start}).$$

Here the starting cell and the destination cell are the grid cells corresponding to the ball's position before and after the action, and $xT(c)$ is the threat value for the cell c pre-computed by Singh's model. In other words, the $xT$ of an action is the difference in threat between the cell the ball arrived at and the cell it started from. Only positive values are summed – that is, actions that move the ball into a zone with higher threat [6].

**Player metrics.** For each player, the following are computed: the number of minutes played, total $xT$ , $xT$ per 90 minutes (to compare players regardless of time on the pitch), the number of progressive carries (moving the ball $\geq 10$ m. toward the opponent's goal within a single action, for example a pass), the number of pressing-type actions, and passing accuracy under pressure from the opponent's players.

**Insight generation and reference checking. Model parameters.** Based on the enriched document, the system queries the main LLM to obtain insights. In doing so, the model analyzes all indicator values present in the data. The key requirement for the LLM's analysis result, stated in the prompt text, is that every insight must contain explicit references to specific metrics (fields of the enriched document) in dot notation, together with an explanation of how these values support the overall conclusion.

Both calls to the main model (insight generation and correction of references to the enriched-document fields) use the same LLM configuration (Table 2). The judge (see the subsection on the «LLM-as-a-judge» evaluation) is not an agent with access to actions or a dialogue, but a second, fully independent call to the same LLM with its own system prompt and no access to the state of the insight generator. It uses the same configuration; the only difference is the Pydantic output schema (the PatternEvaluation parameter instead of TacticalInsights).

Table 2. LLM parameters

| Parameter | Insight generator | Judge (LLM) |
|---|---|---|
| Model | gpt-4o | gpt-4o |
| Temperature | 1,0 (default) | 1,0 (default) |
| Output mode | Structured output – TacticalInsights schema | Structured output – PatternEvaluation schema |
| Maximum number of tokens | unlimited (default) | unlimited (default) |
| Calls per dataset for a specific match | 1–2 calls | 1 call |

The Temperature value of 1,0 is a deliberate choice: since the output is forced into a fixed Pydantic schema through OpenAI's structured-output mechanism, a higher temperature does not affect the structure of the response, but it allows the model to produce more diverse tactical interpretations of the same numerical data. Separate model instances with no shared state are used for the «judge» and the «generator».

The full text of this prompt is provided in the supplementary repository [29]. It instructs the model to analyze the enriched match document from one team's perspective across pressing and out-of-possession play, in-possession build-up, transitions, set pieces, individual impact, and metric interpretation, and – critically – to support every claim with a citable metric_key, making no assertion that cannot be tied to a specific metric.

**The schema used for the main model's analysis result.** The model's response is a fixed data structure (TacticalInsights) that includes: a short summarizing narrative of the match (match_summary); a list of tactical insights (each with a pattern label, a description, a list of metric values metric_keys, and an evidence_description explanation); and lists of the team's strengths and weaknesses and its key players.

**The reference-checking and correction loop.** After each response, every reference to a field of the enriched document is checked programmatically – a dedicated path-checking module verifies each key against the structure of the enriched document. If all keys are found, the insights are passed on. If a key is not found, it is marked as invalid and returned to the model in a correction prompt, which asks the model to repair the unresolvable metric_keys – pointing them to real paths or removing them, without hallucinating paths – and to return the complete corrected TacticalInsights object (provided in the supplementary repository [29]).

After a correction round, the paths are checked again (Fig. 2): valid insights are passed on, while those whose references remain invalid are discarded. This approach differs from free-text self-correction [22]: the validator makes a binary judgment (the key either exists or it does not). The checking module is fully deterministic. The LLM is called at most twice per insight.

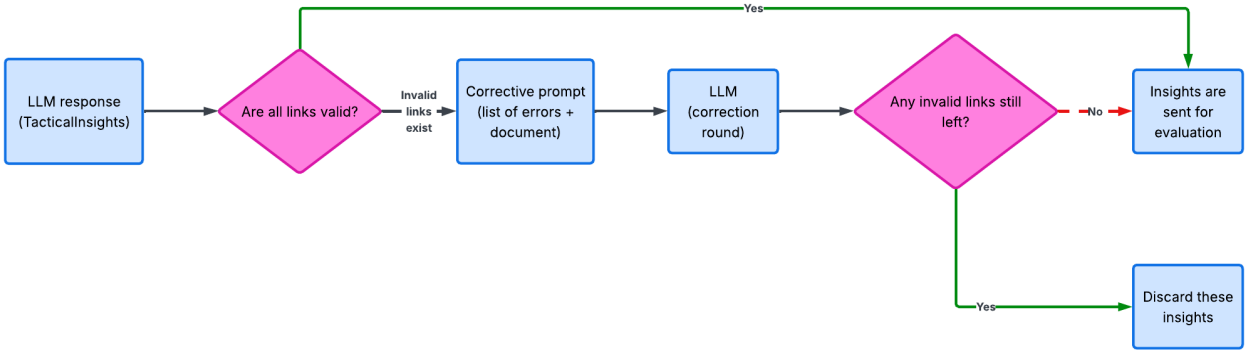


Fig. 2. The reference-checking and correction loop for generated insights

**Detecting recurring patterns.** Analyzing a single match provides information about the tactics a team used within that match only. However, it is also possible to detect a team's typical (recurring) tactical techniques by analyzing several consecutive matches played by that team [14, 15, 16]. This paper proposes such a method. For each team whose data are analyzed, the insights from all analyzed matches are collected and passed to a separate LLM call with the task of identifying insights that recur in two or more matches. The result is a list of the team's stable tactical characteristics, ordered by frequency of occurrence and stored with references to the specific matches in which they were recorded.

At the current stage of the research, cross-match aggregation is performed by a separate LLM call (gpt-4o) – another independent instance, distinct from the insight generator and the judge. Its input is the validated insights of all the team's analyzed matches; its output is a list of patterns that recur in at least two matches, each labelled and traced to the specific match_ids in which it appeared, ordered by frequency (Fig. 3; the full prompt is provided in the supplementary repository [29]).

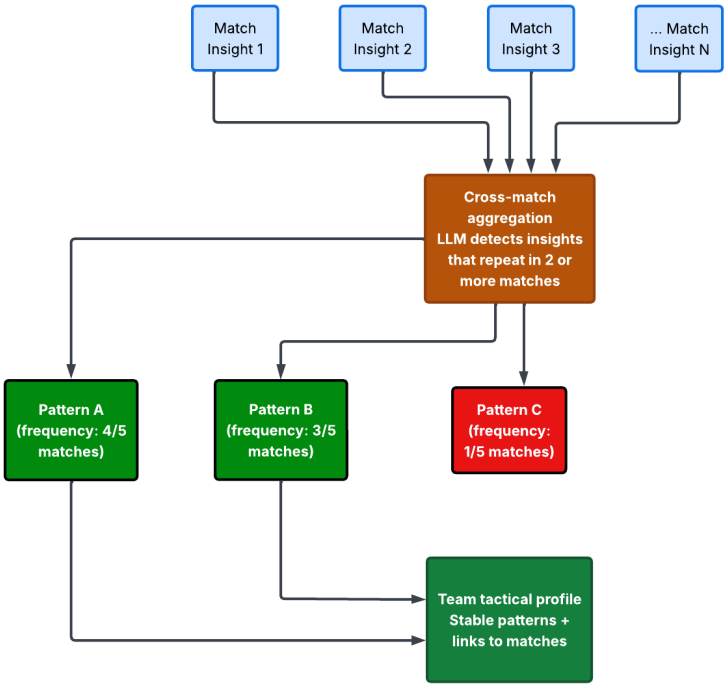


Fig. 3. Scheme for detecting a team's recurring tactical patterns across matches

The «at least two matches» threshold is a deliberately minimal requirement: a single occurrence is considered situational, whereas a repetition in two or more matches is already a tactical tendency. Across the whole corpus, 83 recurring patterns were found for 23 teams (median 3 per team, range 2–8; see the subsection «Analysis of recurring patterns»).

Future work will develop a more formalized method for determining the stability of patterns: for each tactical insight, the system should automatically determine whether it is a systematic feature of the team's style or merely a situational deviation [17].

**The «LLM-as-a-judge» evaluation.** Because manual annotation of tactical insights requires time and domain expertise, a separate LLM – the «judge» – is used for automated evaluation within the proposed system. This approach has found application in the literature [18, 19, 20]. It scales well to data of any volume.

The judge runs as an instance of the LLM that is fully independent of the insight generator (the parameters of both are given in Table 2). The judge has a separate system prompt and no access to the state of the generation LLM. The programmatic reference-checking step filters out invalid citations before they reach the judge, so the judge focuses solely on the semantic quality of already-verified insights.

The full judge prompt, including the scoring rubric for each level, is provided in the supplementary repository [29].

The judge receives, for each insight, its verbal description and the pre-validated metric values, and assigns two scores on a 1–5 scale:

- Faithfulness – how accurately the verbal description of the insight reflects the stated numerical metric values. A score of 1 means a direct contradiction with the data; a score of 5 means every statement is fully supported by the metrics.
- Relevance – how appropriate the metric values are as evidence for the verbal description. A score of 1 means the metrics are not related to the description at all; a score of 5 means they are precise and exhaustive confirmations.

The overall score of an insight is the average of these two dimensions. The overall scores are averaged over each match and over the whole dataset to obtain aggregated quality indicators for the system.

**Examples of the system's output.** Below is an example of an insight generated for an AC Milan match. The model formulated a tactical observation, cited specific metrics as evidence, and described how the metric values support the conclusion.

*Insight:* Compact defensive shape

*Description:* AC Milan maintained a narrow and compact defensive shape, effectively covering the central zones, especially in the first half.

*Metric references:*

field phase_metrics.first_half.pressing_intensity.home.avg_recovery_time → 13.24 s
field phase_metrics.first_half.ppda.home → 10.0

*Rationale:* The average ball-recovery time in the first half and a PPDA of 10,0 correspond to an organized mid-block rather than active pressing – which is consistent with the described compact shape.

The judge receives the same insight together with the metric values and assigns scores (Table 3).

Table 3. The judge's evaluation of the example insight

| Dimension | Score | Judge's reasoning |
|---|---|---|
| Faithfulness | 5 of 5 | The cited metrics fully support the description of a compact defensive shape with effective coverage of the central zones |
| Relevance | 5 of 5 | The metrics accurately illustrate AC Milan's disciplined defensive organization in the first half |
| Average | 5,0 | |

For comparison, another insight from the same match claimed «intense pressing from the start», whereas the first-half PPDA was 10,0 (low intensity) and the second-half PPDA was 2,35. The judge detected the contradiction and assigned a faithfulness score of 3 of 5.

**Evaluation dataset.** The system was evaluated on a dataset covering seven competitions: the Bundesliga, the FA Women's Super League, the Indian Super League, La Liga, Ligue 1, the Premier League, and Serie A. For each team, the season with the largest amount of available match data in the StatsBomb open data was chosen automatically, and up to ten matches were analyzed in chronological order (Table 4).

Table 4. Evaluation dataset summary

| Indicator | Value |
|---|---|
| Teams | 23 |
| Matches processed | 240 |
| Insights evaluated | 691 |
| Matches with no evaluated insights | 33 (13.8%) |

Of the 240 matches, 33 (13,8%) produced no evaluated insights. The reasons for this are described in the subsection «Error rate». The remaining 207 matches produced 691 evaluated insights, an average of 3,34 insights per match.

**Aggregated quality metrics.** Scores are computed at the level of individual insights. Table 5 gives the mean and standard deviation across all 691 evaluated insights.

Table 5. Aggregated quality metrics across the 691 evaluated insights (23 teams, 207 matches)

| Dimension | Mean ± standard deviation |
|---|---|
| Faithfulness | 4,15 ± 0,84 |
| Relevance | 4,16 ± 0,91 |

The system achieves an overall score of 4,16 out of 5. The mean faithfulness and relevance values are almost equal. The standard deviation indicates moderate dispersion – most insights are concentrated in the 4–5 range, with a long tail of lower scores for cases where the model cited too few metrics or the metric values had a weak connection to the insight's claim.

Table 6. Distribution of scores across all evaluated insights

| Score range | Patterns | Amount |
|---|---|---|
| 1,0–2,0 | 4 | 0,6% |
| 2,0–3,0 | 35 | 5,1% |
| 3,0–4,0 | 150 | 21,7% |
| 4,0–5,0 | 502 | 72,6% |

72,6% of insights score in the 4–5 range, and fewer than 6% score below 3. The four insights scoring below 2,0 share a common cause: the model formulated claims about tactical phenomena (for example, set-piece variations or attack-development routes) for which the enriched document has no

corresponding metric. This confirms that the programmatic path-checking step removes invalid references but cannot detect references where the metric value is found correctly yet does not support the claimed insight.

**Results by team.** All 23 evaluated teams – drawn from the five major European leagues, the FA Women's Super League, and the Indian Super League – achieve an average overall score above 3,8 on the 1–5 scale, which confirms the stable output quality of the system across different competitions and tactical contexts. Team-level means are tightly clustered, ranging from 4,40 (AS Roma) down to 3,83 (Bastia), with faithfulness and relevance tracking closely for almost every team and no competition showing a systematic disadvantage. The full per-team breakdown is provided in the supplementary repository [29].

The teams with the highest scores (AS Roma – 4,40, Athletic Club – 4,29) benefit from clear tactical patterns that are well reflected by the available metrics. The teams with lower scores are mostly clubs from Ligue 1 and the Indian Super League – matches in these competitions may contain less distinct numerical indicators, owing to lower tactical variability or fewer transitions between phases.

**Error rate.** The results of processing 33 of the 240 matches (13,8%) contain no evaluated insights. In these cases, the judge returned pattern names worded differently from the **original** ones, so the matching algorithm could not associate any score with a pattern. Such records are considered potentially incorrect and are excluded from all aggregated scores.

**Qualitative examples.** To illustrate the range of quality, the single highest-scored insight (overall score 5,0) and the single lowest-scored insight (overall score 1,0) from the processed dataset are presented below.

Example 1. Medium-intensity pressing and counter-pressing – Borussia Mönchengladbach (overall score: 5.0).

Mönchengladbach maintained medium-intensity pressing throughout the game, often pressing the opponent in their own half after goal kicks and turnovers. The team consistently began pressing immediately after losing the ball, with noticeable intensity.

Metrics: ppda.away = 3.51, first_half.ppda.away = 2.44, second_half.ppda.away = 5.43, Dahoud.pressures_applied = 27, Stindl.pressures_applied = 15

Faithfulness 5 of 5: The description accurately reflects the metrics – a low PPDA across the match indicates high pressing intensity, confirmed by the individual players' pressure counts.

Relevance 5 of 5: PPDA and player-level pressure counts are direct evidence for the pressing pattern; the split by halves shows the phase-level detail the description refers to.

This insight illustrates the ideal case: each claim is backed by metrics that unambiguously correspond to the described behaviour, with no gap between the number and the conclusion.

Example 2. Set-piece effectiveness – Bayer Leverkusen (overall score: 1.0)

Leverkusen tactically used a variety of set-piece deliveries, alternating short corners and in-swinging balls, which contributed to the number of goals.

Metrics: (none)

Faithfulness 1 of 5: No metrics are provided; the description is entirely speculative.

Relevance 1 of 5: Without numerical evidence, the insight is unsupported.

This is the most typical kind of error: the system generated a seemingly plausible conclusion but could not support it with any metrics – because the delivery type and set-piece triggers are not represented in the available metric schema.

**Analysis of recurring patterns.** The cross-match pattern analysis identified 83 recurring tactical patterns across all 23 teams (median: 3 per team, range: 2–8).

A recurring pattern is a tactical action that the system independently detects in two or more matches of the same team. A pattern is considered recurring if it appears – possibly in different wording but with the same tactical essence – in at least two matches. Recurring patterns reflect choices that the coaching staff repeats against different opponents, rather than a reaction to a specific game situation (Table 7).

Table 7. Number of recurring patterns by team and the average number of repetitions per pattern

| Team | Patterns | Avg. rep. |
|---|---|---|
| Borussia Dortmund | 2 | 9,00 |
| Bayern Munich | 4 | 7,50 |
| Barcelona | 3 | 6,67 |
| Birmingham City WFC | 3 | 6,67 |
| Atalanta | 2 | 6,50 |

| Team | Patterns | Avg. rep. |
|---|---|---|
| Bologna | 4 | 6,00 |
| Bayer Leverkusen | 4 | 5,75 |
| Arsenal | 3 | 5,67 |
| Athletic Club | 4 | 5,50 |
| Bordeaux | 3 | 5,33 |
| Augsburg | 4 | 5,25 |
| AS Monaco | 4 | 5,00 |

The table 7 is presented in abbreviated form. The full version of the table can be found in the supplementary repository [29].

Three examples illustrate the concept:

*Example 1. Bayer Leverkusen – intense pressing (8 matches, Bundesliga, 2023/24 season).*

The system recorded intense pressing in every match of Leverkusen in the sample:

*vs RB Leipzig (19.08.2023):* «Bayer Leverkusen applied intense pressing early in the match to disrupt the opponent's build-up.»

*vs FC Heidenheim (24.09.2023):* «Leverkusen consistently pressed high up the pitch, using goal kicks as triggers for quick ball recovery, confirmed by a low PPDA.»

*vs Freiburg (29.10.2023):* «Bayer Leverkusen showed a tendency toward intense pressing immediately after losing the ball, especially in the first half.»

The pattern persisted across eight matches, capturing the team's characteristic pressing style in an unbeaten season.

*Example2. Barcelona – fullbacks in attacking actions (8 matches, La Liga, 2020/21 season):*

*vs Real Betis (07.11.2020):* «Using wide areas, Barcelona's fullbacks played a key role in attack build-up.»

*vs Osasuna (29.11.2020):* «Barcelona organized play from the goalkeeper and centre-backs. The in-possession shape transformed, often using overlapping runs of the fullbacks.»

*vs Sevilla (04.10.2020):* «The use of overlapping fullback runs was a key mechanism for Barcelona to widen the attacking range.»

*Example 3. AC Milan – attack build-up through the flanks (4 matches, Serie A, 2015/16 season):*

*vs Palermo (19.09.2015):* «AC Milan's attack used width through overlapping fullback runs, while Bonaventura made progressive carries.»

*vs Sassuolo (25.10.2015):* «AC Milan's build-up started from the defenders with progressive carries, such as Romagnoli with 6 carries, and then shifted to the flanks.»

**Stability analysis.** If the system detected arbitrary regularities rather than genuine tactical tendencies, recurring patterns would not be expected to score higher than one-off ones – since the judge evaluates each match separately and does not know which patterns are later deemed recurring.

Table 8. Quality of pattern-related insights versus the rest

| Group | N | Faithfulness | Relevance | Overall |
|---|---|---|---|---|
| Insights that are manifestations of patterns | 348 | 4,36 | 4,39 | **4,38** |
| Other insights | 354 | 3,96 | 3,96 | **3,96** |

Insights from recurring patterns scored higher – 4,38 versus 3,96 (+0,42). Since the judge evaluates each match separately and does not know which insights will later be recognized as part of a pattern, the higher score is not the result of bias – it indicates that insights recurring across matches are also better supported by numerical evidence.

**Conclusions.**

In this work, a system has been developed and proposed that, using an LLM, automatically generates tactical insights from football match data obtained from the StatsBomb database, supporting each insight with references to metrics pre-computed from the input data. The system can automatically check the correctness of these insights and discard incorrect ones. The evaluation was conducted on 691 insights from 23 teams. The average score was 4,16 ± 0,85 on a 1–5 scale, with 72,6% of insights falling in the 4–5 range – a result that is stable across all teams and competitions.

The system uses the principle of evidence: in its conclusions it must reference specific metrics, which is a key factor for ensuring the quality of the resulting conclusions. Insights with numerical support consistently receive higher scores: the more metrics the model cites and the more precisely they match the tactical claim it formulates, the higher the judge's score. Detail at the level of tactical phases further improves quality compared with the overall match indicators.

The cross-match analysis identified 83 recurring patterns for 23 teams. Insights that are part of such patterns scored 0,42 higher than the rest (4,38 vs 3,96) – even though the judge evaluates each insight independently, without knowing whether it is part of a pattern. This indicates that the system consistently detects a team's real tactical tendencies rather than random observations within a single match.

Three types of errors that can be present in the insights obtained from the LLM were identified:

1. Absence of metrics: the insight describes a tactical phenomenon – for example, a set-piece variation or player positioning during an attack – for which the enriched document has no corresponding supporting metric.
2. Semantic inversion: the path to a metric value in the enriched document is found correctly, but the model misinterprets the arithmetic sign of the metric (for example, treating a low PPDA value as a sign of weak pressing rather than the opposite).
3. Matching errors: in 33 matches (13,8%), the judge did not return a correct insight name, so these insights were excluded from the further quality evaluation of the developed system as a whole.

**Limitations.** The evaluation carried out with the proposed system relies solely on the judge; no study involving human analysts was conducted. The judge and the insight generator use the same model (GPT-4o), which may introduce a bias in favour of a similar response style. A different model would probably give somewhat different results. The list of metrics available in the source data numbers a little over 20, but it is not exhaustive and does not represent all tactically important events in a match – in particular, dynamic tactical schemes and set pieces (corners, free kicks, etc.) are not studied. These data are available in StatsBomb, but using them would require more complex aggregations and would further increase the already large enriched document, which raises the risk of hallucinations and a decline in insight quality.

**Prospects for further research.**

The list of metrics contained in the fields of the enriched document, on which the LLM must rely, can be expanded with data on the positioning of players on the pitch, which would allow better coverage of patterns of player positioning and their movement off the ball. Such data partly already exist in the open StatsBomb 360 dataset – the coordinates of visible players at the moment of individual events, without being tied to a specific player's name. However, this is not continuous tracking but individual frames, and coverage is extremely limited: among the competitions considered in this work, 360 data are available only for some Bundesliga matches.

An interesting stage is developing a clear methodology for finding recurring patterns for a specific team on a set of data about the matches it has played.

Currently, the search for recurring patterns across matches is performed as a separate step after all available data have been processed. Automatically updating the list of found patterns for a specific team after each newly processed match would make it possible to gradually build up the team's tactical profile in real time.

**References**

1. Caley M. Shot matrix I: Shot location and expected goals // Cartilage Free Captain. 2014.
2. Brechot M., Flepp R. Dealing with randomness in match outcomes: How to rethink the measurement of a team's offensive and defensive performance in association football // Journal of Sports Economics. 2020. Vol. 21, no. 4. P. 335–362. DOI: 10.1177/1527002519897962.
3. Trainor C. PPDA: A pressing stats glossary // StatsBomb. 2014.
4. Bekkers J., Sahasrabudhe S. Pressing intensity: An intuitive measure for pressing in soccer. 2025. DOI: 10.48550/arXiv.2501.04712.
5. Singh K. Introducing expected threat (xT) // karun.in. 2018.
6. Poropudas J. Extended model for expected threat // Proceedings of the New England Symposium on Statistics in Sports (NESSIS). 2021.
7. Oruç A. et al. Phase-level tactical analysis in professional football // Proceedings of the StatsBomb Conference. 2023.
8. Caron M., Müller O. TacticalGPT: Uncovering the potential of LLMs for predicting tactical decisions in professional football // Proceedings of the StatsBomb Conference. 2023.
9. Hernandez et al. Querying football matches: Towards using LLMs for sports analytics // Proceedings of the International Advanced Computing and Communication Systems Symposium (IACSS). Springer, 2024.
10. Kucukoglu Y. et al. FootGPT: A large language model for football analytics. 2023. DOI: 10.48550/arXiv.2308.08610.

11. Cook A., Karakuş O. LLM-Commentator: Novel fine-tuning strategies of large language models for automatic commentary generation using football event data // Knowledge-Based Systems. 2024. Vol. 300. Art. 112219. DOI: 10.1016/j.knosys.2024.112219.
12. Caut R. et al. Wordalisations: Converting numerical sports model outputs to natural language. 2025. DOI: 10.48550/arXiv.2504.00767.
13. Midoglu C. et al. SoccerRAG: Multimodal soccer information retrieval via natural language queries. 2024. DOI: 10.48550/arXiv.2406.01273.
14. Badr A. et al. Dynamic descriptive analytics in football using retrieval-augmented generation // Procedia Computer Science. 2025.
15. Trower M., Graham N., Cottrell N., Hengster Y. Clustering women's football players: Identifying functional patterns for performance optimisation // Proceedings of the StatsBomb Conference. 2023.
16. Upadhyay P., Backhaus J. Identifying key players and playing styles of 10 EPL teams during offensive sequences in 2021/22 // Proceedings of the StatsBomb Conference. 2023.
17. Demir E., Sahin Y., Ure N. How do football teams play? Deep embedded clustering for playing style analysis // Proceedings of the International Advanced Computing and Communication Systems Symposium (IACSS). Springer, 2025.
18. Zheng L. et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena // Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). 2023.
19. Liu Y. et al. G-Eval: NLG evaluation using GPT-4 with better human alignment. 2023. DOI: 10.48550/arXiv.2303.16634.
20. Gu J. et al. A survey on LLM-as-a-Judge. 2024. DOI: 10.48550/arXiv.2411.15594.
21. Huang Y. et al. Hallucination-to-Truth: A systematic review of hallucination mitigation in large language models. 2025. DOI: 10.48550/arXiv.2508.03860.
22. Bhargava et al. Multi-stage self-verification for LLM factual accuracy. 2025. DOI: 10.48550/arXiv.2509.05741.
23. Lewis P. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Proceedings of the Conference on Neural Information Processing Systems (NeurIPS). 2020.
24. Wright C., Atkins S., Jones B., Todd J. The role of performance analysts within the coaching process: Performance Analysts Survey "The role of performance analysts in elite football club settings" // International Journal of Performance Analysis in Sport. 2013. Vol. 13, no. 1. P. 240–261.
25. Wright C., Carling C., Collins D. The wider context of performance analysis and its application in the football coaching process // International Journal of Performance Analysis in Sport. 2014. Vol. 14, no. 3. P. 709–733.
26. StatsBomb. StatsBomb Open Data : GitHub repository. 2018–. URL: https://github.com/statsbomb/open-data.
27. Pappalardo L. et al. A public data set of spatio-temporal match events in soccer competitions // Scientific Data. 2019. Vol. 6, no. 1. Art. 236. DOI: 10.1038/s41597-019-0247-7.
28. SkillCorner. SkillCorner Open Data : GitHub repository. 2023–. URL: https://github.com/SkillCorner/opendata.
29. Sorokivskyi O., Hotovych V. Automated Tactical Analysis of Football Matches Using Large Language Models – supplementary materials (data examples and prompts) : GitHub repository. 2026. URL: https://github.com/PajariPy/Automated-Tactical-Analysis-of-Football-Matches-Using-Large-Language-Models.

Історія статті:
Отримано: 29.05.2026 Доопрацьовано: 07.08.2026 Прийнято до друку: 23.09.2026 Опубліковано: 29.09.2026