Research article

Automated explanation of machine learning models of footballing actions in words

Journal of Sports Analytics
Vol. 11(0): 1–14

DOI: 10.1177/22150218251353089
journals.sagepub.com/home/san

Pegah Rahimian[1], Jernej Flisar[2] and David Sumpter[1,2]

Abstract

While football analytics has changed the way teams and analysts assess performance, there remains a communication gap between machine learning practice and how coaching staff talk about football. Coaches and practitioners require actionable insights, which are not always provided by models. To bridge this gap, we show how to build wordalisations (a novel approach that leverages large language models) for shots in football. Specifically, we first build an expected goals model using logistic regression. We then use the coefficients of this regression model to write sentences describing how factors (such as distance, angle and defensive pressure) contribute to the model's prediction. Finally, we use large language models to give an engaging description of the shot. We describe our approach in a model card and provide an interactive open-source application describing shots in recent tournaments. We discuss how shot wordalisations might aid communication in coaching and football commentary, and give a further example of how the same approach can be applied to other actions in football.

Keywords

Soccer analytics, explainable AI, expected Goal, language models

Received: 7 April 2025; accepted: 31 May 2025

Introduction

The field of soccer analytics has witnessed a rapid evolution, with machine learning models playing a crucial role in evaluating player and team performance (Decroos et al., 2019; Dick and Brefeld, 2023; Fernandez et al., 2019; Gyarmati and Stanojevic, 2016; Peralta Alguacil et al., 2020; Rahimian et al., 2022, 2023). One of the most widely used models is the Expected Goals (xG) model, which is used to evaluate the quality of scoring opportunities by assigning probabilities to shots based on factors such as location, angle, and defensive pressure (Pollard and Reep, 1997; Sumpter, 2016). Several studies train machine learning models using predictors such as shot type, distance to goal, and angle to goal to estimate xG (e.g., Bransen and Davis, 2021; Eggels et al., 2016; Herbinet, 2018; Pardo, 2020; Rathke, 2017; Sarkar and Kamath, 2021; Tippana, 2020; Wheatcroft and Sienkiewicz, 2021). Recent work has also explored integrating tracking and performance data to improve predictive performance and model explainability in xG modelling, highlighting the importance of both accurate classification and interpretable feature contributions (Cefis and Carpita, 2025a, 2025b). Additionally, detailed studies have investigated the role of data sources on model performance (Davis and Robberechts, 2020); how defensive positioning and goalkeeper placement enhances the estimation of goal probabilities (Lucey et al., 2015); and the use of neural networks to estimate scoring probabilities (Ruiz et al., 2015).

An important consideration when building xG models is that we should be able to explain their implications to coaching staff. Many xG models are black boxes, producing numerical probabilities without offering clear explanations of how the different features of a shot determine the probability that it will result in a goal (Davis et al., 2024). To

[1]Dept. of Information Technology, Uppsala University, Uppsala, Sweden
[2]Twelve Football, Stockholm, Sweden

Corresponding author:
Pegah Rahimian, Uppsala University Department of Information Technology, SE-751 05 Uppsala, Sweden.
Email: pegah.rahimian@it.uu.se

address this challenge, one approach is to use SHAP (SHapley Additive Explanations) (Lundberg and Lee, 2017) to explain the contribution of each feature to a model's prediction. Anzer and Bauer (2021) applied SHAP to xG models, demonstrating that shot distance is the most influential factor in goal probability. Another approach is to build models that are interpretable by design. For example, building on work by Morales (2016), Sumpter (2016) proposes a logistic regression model that incorporates how much of the goal the shooter can see and distance from goal as variables. This expected goals model can thus be explained in terms of the shooter sight on goal, a simple to communicate coaching concept.

Even when adopting these approaches, there remains a gap between what a machine learner practitioner and coaching staff might consider as an explanation. Indeed, while SHAP values give a numerical representation of feature contributions, these do not automatically translate into actionable insights for football practitioners. This is part of a larger issue within sports analytics where very few studies explain how adopting recommendations from a model impact performance (Goes et al., 2021). To bridge this gap, we adapt an approach introduced by Caut et al. (2025) known as *Wordalisation*. The key idea is to use large language models (LLMs) to convert numbers into natural language narratives. One example in Caut et al. (2025) is a football scout, which uses rankings of players in key metrics to describe their skills. Wordalisations are thus concise, easily digestible narratives that summarize data-driven observations without directly reporting numerical values. Prompt engineering, the practice of crafting effective input instructions for LLMs, is a key to using these systems (Brown et al., 2020; Reynolds and McDonell, 2021; Wei et al., 2022). By careful framing of prompts, users of LLMs can significantly improve the relevance, accuracy, and creativity of generated outputs. By engineering prompts from data, we can transform abstract metrics into accessible explanations.

Our contribution extends this concept by shifting from the wordalisation of raw data values to the wordalisation of machine learning model outputs–specifically, xG predictions and their feature-level contributions. This enables the communication of how each input factor influenced the predicted xG value of a shot, providing coaches and analysts with meaningful, contextualised interpretations rather than just descriptive statistics or abstract numbers. Unlike prior work that primarily focuses on developing more accurate xG models, the goal of this paper is to enhance the interpretability of model outputs for non-technical users. We do not aim to introduce a novel predictive model for xG estimation–instead, instead we build on the models that already exist in the literature and in practice to make them interpretable by coaches. Our emphasis is on generating understandable, actionable textual explanations from model predictions. While we use logistic regression in this study due to its interpretability and established effectiveness in xG modelling, the system is model-agnostic and can be adapted to other machine learning models with decomposable or explainable outputs (e.g., SHAP-compatible models).

Figure 1 illustrates the overall workflow of our proposed approach, which is divided into two main components: the model pipeline and the wordalisation process. The data pipeline extracts data from databases and APIs, generates relevant features, trains corresponding machine learning models, and calculates the contribution of each feature to the output. The second component, wordalisation, uses LLMs to generate intuitive, text-based narratives that explain xG values based on feature contributions. To document our approach, we provide a structured model card (Mitchell et al., 2019) detailing design, capabilities, and limitations for transparency and reproducibility. We provide an open-source Streamlit application that enables users to import their own shot dataset and explore xG explanations interactively. The tool is available at https://shotsgpt.streamlit.app/. We also provide the code online: https://github.com/Peggy4444/shotsGPT/tree/main.

Materials and methods

We start by describing the dataset and features used to train the xG model. We then explain the model justification and interpretability. Next, we outline the steps for constructing prompts in our wordalisation process. Finally, we introduce metrics to evaluate the wordalisation by analyzing the trade-off between engagement and accuracy.

Data description and feature generation

The dataset used in this study was obtained from the Hudl-StatsBomb events and StatsBomb360 datasets for the following available competitions: EURO Men 2024 and 2022, National Women's Soccer League (NWSL) 2018, FIFA 2022, Women's Super League (FAWSL) 2017, and Africa Cup of Nations (AFCON) 2023.[1] These datasets were accessed using the `statsbombpy` API.[2] The StatsBomb events dataset comprises 110 columns detailing various aspects of each event, while the StatsBomb360 dataset includes 7 columns describing the positions of players visible in the frame of the action. These datasets were merged to provide a comprehensive view of the events for all matches played by all teams participating in the respective competitions and seasons. After filtering for open-play, non-header shots with goalkeeper tracking data available, the number of shots used in our analysis per competition is as follows: EURO 2024 (1029 shots), EURO 2022 (972 shots), FAWSL 2017–18 (2270 shots), NWSL 2018 (811 shots), FIFA 2022 (1130 shots), and AFCON 2023 (869 shots). These shot-level data formed the input to the expected goals (xG) and wordalisation pipeline described in this study.

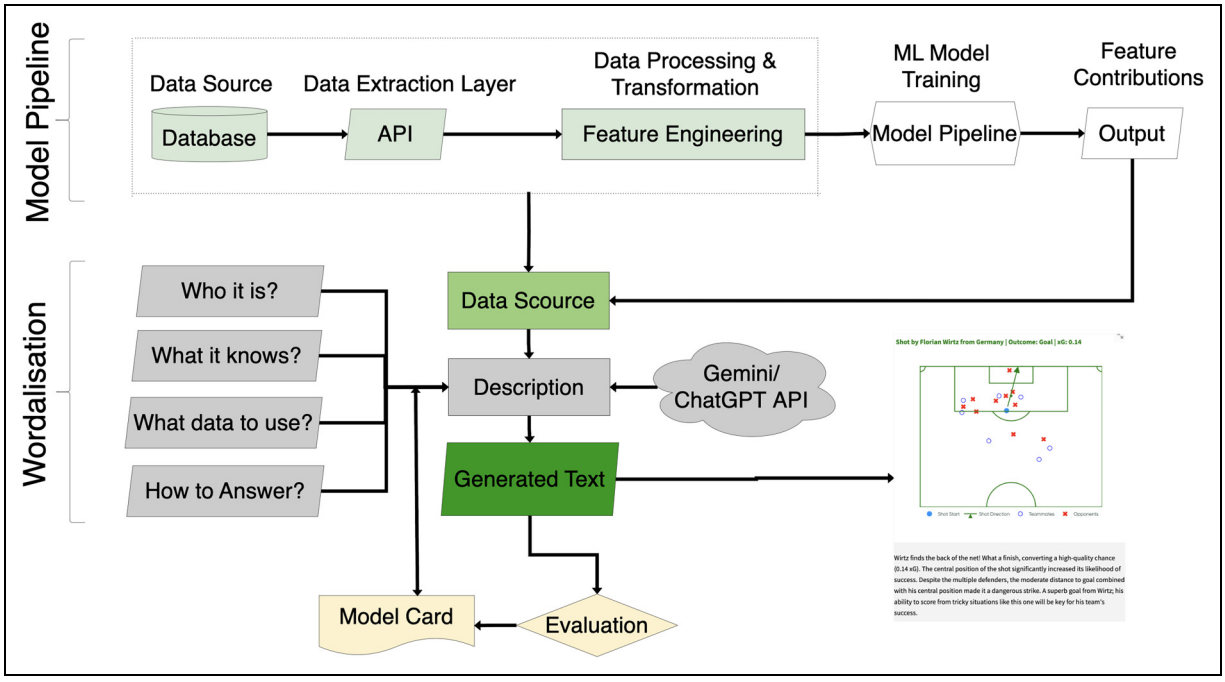

Figure 1. Overview of the proposed workflow, comprising the data pipeline for feature extraction, model training, and feature contribution analysis, along with the wordalisation process that integrates data source, description, and LLM chat modules. The output is an engaging and accurate LLM generated text.

From the shot data, we generated a set of features categorized into body part-related, play pattern-related, goalkeeper-related, shot location-related, and opponent-related features. These features collectively provide a detailed understanding of the shot and its surrounding context. We assume a fixed pitch size of 105 meters in length and 68 meters in width for all games.

The body part-related features include a binary indicator for whether the shot was taken with the left foot, called `shot with left foot`. Play pattern-related features consist of binary indicators for `shots after throw-ins`, `corners`, and `free kicks`. Goalkeeper-related features include the Euclidean distance between the shot location and the goalkeeper's position (`distance to goalkeeper`), the goalkeeper's distance to the centre of the goal (105, 34) (`goalkeeper distance to goal`), and the angle between the shot location and the goalkeeper's position (`angle to goalkeeper`), measured relative to the goal line.

Shot location-related features encompass the vertical distance from the shot to the centerline of the pitch (`vertical distance to centre`), the angle between the shot location and the goalposts (`angle to goal`), and the Euclidean distance from the shot location to the goal line (`distance to goal`). Opponent-related features include the count of opposition players within 3 meters of the ball at the time of the shot (`nearby opponents in 3 meters`), the number of opponents within a triangular area formed by the shot location and the goalposts (`opponents in triangle`), the minimum Euclidean distance to the nearest opponent (`distance to nearest opponent`), and the angle to the nearest opponent from the shot location (`angle to nearest opponent`). These features, which are illustrated in Figure 2, cover shot positions and position of opposing team, all explaining the success factors of shots.

Feature selection and model training

Our data pipeline is designed to fit a model to that data and build wordalisations which compare the importance of features of a particular shot compared to other shots in that dataset. This means we fit different logistic regression xG models independently to each of the six competitions (to see the individual models go into https://shotsgpt.streamlit.app/shots and select the competition of interest from drop-down in the left sidebar and look at the summary table of the trained model for the selected competition). Our aim is not to build the "best" expected goals model but to be able to explain the probability of scoring a particular shot relative to other shots in the same competition. Depending upon the competition we expect different variables to have different weights in the final model.

In designing this pipeline, we used one of the competitions (Euros 2024) to identify if any of the variables could be dropped. We found that `angle to goalkeeper` and `angle to goal` were highly correlated (Pearson correlation R=0.88), so we dropped the second of these. Similarly,

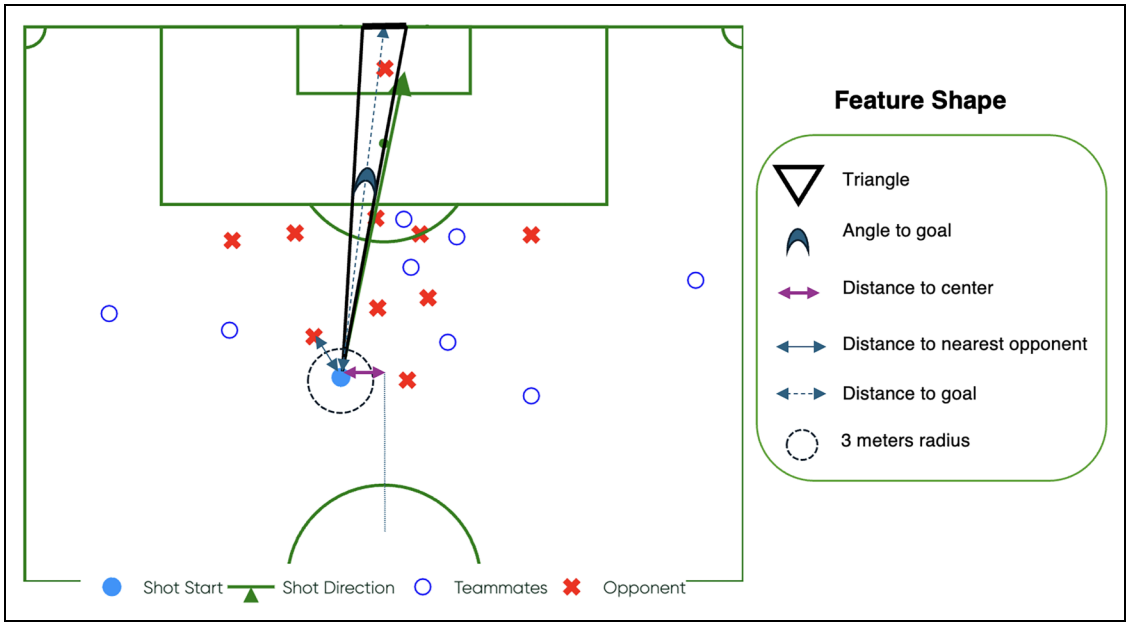

Figure 2. Illustration of various football features including shot location, goalkeeper position, opponent pressure, and teammates' positions.

`distance to goalkeeper` was also highly correlated with `distance to goal` (Pearson correlation R=0.81), so we dropped the first of these. We then fitted the logistic regression model and looked at the P-values for each of the remaining variables. Among those, `goalkeeper distance to goal` and `angle to the nearest opponent` had P-values higher than 0.05, so we dropped them from the model. However, we found that squaring the `vertical distance to centre` made it a significantly explanatory variable, so included it in the model, calling it `squared distance to centre`.

While accuracy is a concern when building predictive models, for our study interpretability is emphasised because we want to build wordalisations to explain the underlying factors that contribute to a shot's outcome. For this reason, we maintain certain features in the model even if their p-value is greater than 0.05, as they provide valuable insights into the shot context. For instance, features like `shot after throw in` or `nearby opponents` are kept in the model, as they help explain the circumstances around the shot.

After this feature selection and transformation process, we arrive at the final set of features that are either (or both) statistically significant and interpretable. The features retained for the final xG models are as follows: `squared distance to centre`, `euclidean distance to goal`, `nearby opponents in 3 meters`, `opponents in triangle`, `goalkeeper distance to goal`, `distance to nearest opponent`, `angle to goalkeeper`, `shot with left foot`, `shot after throw in`, `shot after corner`, `shot after free-kick`. The same features are included for every competition, although the coefficients vary since they are estimated per competition. As a future work, we plan to adjust and distinguish these feature list according to the domain knowledge, playing style, climate situations, and gender adaptability of each competition. Although our primary focus is on interpretability rather than predictive optimization, the xG models achieve solid performance across all six competitions, with ROC scores ranging from 76% to 84% and low Brier scores (0.05–0.09), indicating good calibration. See model card for details of evaluation at both the main app and the GitHub repository at: https://github.com/Peggy4444/shotsGPT/blob/main/model%20cards/model-card-shot-xG-analysis.md.

Explainable components and feature contribution weights

In logistic regression, the predicted probability of an event is modelled as a function of the input features using log-odds. The log-odds can be expressed as a linear combination of the input features, where each feature contributes to the final prediction based on its coefficient. The log-odds for a given shot (feature vector) are defined as:

$$\log - \text{odds}(\mathbf{x}) = \beta_0 + \sum_{\mathrm{j}=1}^{\mathrm{M}} \beta_{\mathrm{j}} \mathrm{x}_{\mathrm{j}}, \tag{1}$$

where β_0 is the intercept (baseline), β_j is the coefficient, and x_j is the value of feature j.

We calculate the contribution of each feature to the log-odds by first mean-centreing the feature values. This

step adjusts each feature value x_j by subtracting the mean of that feature across the dataset, ensuring that each feature's contribution is measured relative to its baseline value. The mean-centred feature value for a given shot is denoted as $\tilde{x}_j$, calculated as:

$$\tilde{x}_j = x_j - \mu_j,$$

where μ_j is the mean of feature x_j across all feature vectors (ie. shots) in the dataset.

The contribution of each feature x_j to the shot's predicted xG is then calculated as

$$\text{Contribution of x}_\text{j} = \beta_\text{j} \cdot \tilde{\text{x}}_\text{j}. \quad (2)$$

Here, β_j is the coefficient associated with feature x_j, and $\tilde{x}_j$ is the mean-centred value of that feature for the specific shot. This approach isolates the unique effect of each feature relative to other observations in the data set. In the context of expected goals, the contribution tells us what was unusual (or not) about this particular shot relative to the other shots in the dataset.

Note that the log-odds are converted to a probability via the logistic function, yielding the final predicted xG:

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-\text{log-odds}(\mathbf{x})}}. \quad (3)$$

giving the overall probability of a shot being a goal, i.e the expected goals value. By calculating the contributions of each feature for every shot, we can understand the specific factors driving the model's prediction. For example, an (unsuccessful) shot in the 56th minute from a match between Germany and Scotland in EURO 2024 is shown in Figures 3(a) and 3(b)), with an xG of 0.03. In this case, there are 4 opponents in the triangle blocking the shot path, leading to a large negative contribution from the `opponents in triangle` feature. This can be seen in the distribution plot, where each point is a single shot contribution (i.e. $\beta_j \cdot \tilde{x}_j$) for each of the model variables. The fact that the shooter is closely marked by opponents, which significantly reduces the chances of scoring, is thus reflected in the plot, where the value for `opponents in triangle` is far to the left, indicating a strong negative influence on the xG.

In contrast, the second successful shot, in Figures 3(c) and 3(d), features a slightly higher xG than the first shot. In this instance, there is only one opponent (the goalkeeper) in the triangle blocking the path, and this results in a positive contribution to the xG. The distribution plot shows the `opponents in triangle` feature far to the right, indicating a positive impact on the xG.

Wordalisation: Step by step prompt

While the approach above explains shot success in terms of the variables, such as defensive pressure (measured by the number of opponents in the shooting triangle), these do not automatically allow communication with practitioners. Visualizations like distribution plots can fall short in conveying actionable or intuitive understanding to coaches, players, or non-technical stakeholders. To address this gap, we adapt the wordalisation approach of Caut et al. (2025), described in the introduction.

There is a structured, four-step approach for creating prompts underlying wordalisations. Each step is designed to provide clarity and context, ensuring the generated descriptions are coherent, aligned with practitioner needs and accurate. These steps are as follows: 1) Tell it who it is, 2) Tell it what it knows, 3) Tell it what data to use, 4) Tell it how to answer. In our case, the aim is to describe these steps tailored for interpreting the contributions of different variables to estimated xG values. An overview of the approach is given in Figure 4. We now outline the four Wordalisation steps, for, what we call, a shot commentator.

Tell it who it is: a large language model's **system prompt** establishes the context by specifying the role the assistant should fulfill when generating responses. In the case of our shot commentator, we use the system prompt shown in the top right box of 4.

Tell it what it knows: The next step involves defining the assistant's knowledge base through example question-and-answer pairs provided by human expert. These examples help the language model understand both the domain-specific knowledge it should convey and the style in which it should respond. An example is shown in Figure 4. In total we provide 43 question/answer pairs.[3]

Tell it what data to use: The next step is to convert the numerical values of the overall expected goal value (i.e. equation (3)) and the individual contributions (i.e. equation (2)) into words. This is a a very sensitive stage in creating the wordalisation, since it requires us to explain to a coach without a mathematical background, what these equations tell us about football. At this stage, it does not, for example, suffice to simply print out the variable values or the contributions. We need to carefully explain what those values imply about football.

The xG value (from equation (3)) quantifies the likelihood of scoring. Instead of giving a numerical value, we use percentiles. Specifically, we translate xG values into qualitative descriptions of scoring chances. We categorize the xG values based on predefined percentiles into five categories: "slim chance" for the 25th percentile (<0.028xG), "low chance" for the 50th percentile (<0.056xG), "decent chance" for the 75th percentile (<0.096xG), "high-quality chance" for the 90th percentile (<0.3xG), and "excellent chance" for values above the 90th percentile (>0.3xG). An example of the outputted text is shown in blue in Figure 5(a).

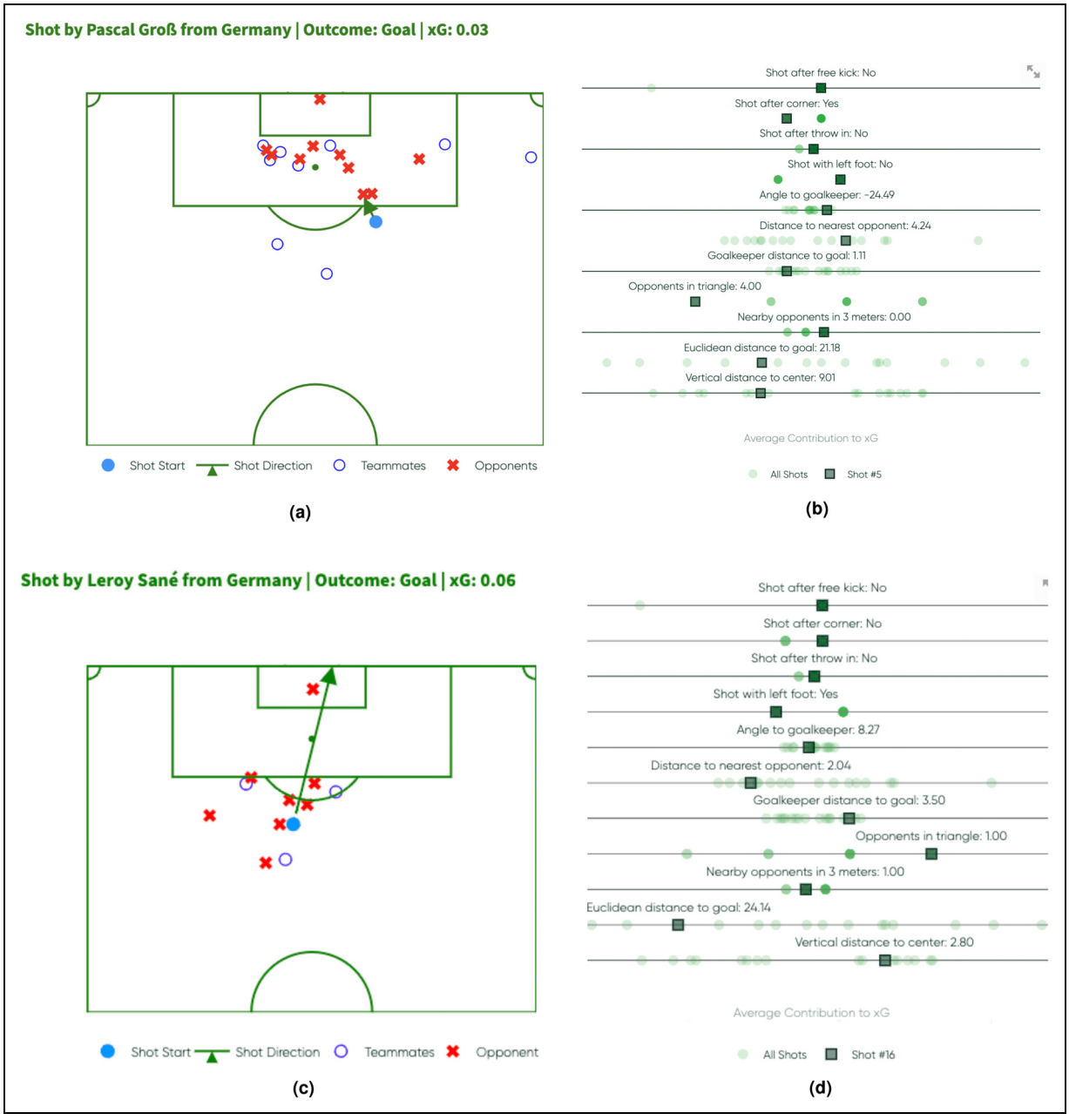

Figure 3. Analysis of two shots from Germany vs. Scotland in EURO 2024. The top row shows the 56th-minute shot, with the pitch visual on the left and the contribution plot on the right. The bottom row shows the 85th-minute shot, with the pitch visual on the left and the contribution plot on the right. (a) 56th minute shot - Pitch Visual, (b) 56th minute shot - Contribution Plot, (c) 85th minute shot - Pitch Visual and (d) 85th minute shot - Contribution Plot.

We also use percentile ranges to describe continuous features such as `euclidean distance to goal` and `angle to goalkeeper`, grouping them into categories like ”close-range” or ”tight angle.” For binary features, we employ mappings, such as:

> ”The shot was taken with the left foot.” if the value of `shot with left foot` feature is True.

An example of these contributions is shown in red in Figure 5(a).

To then explain how different factors contributed to the xG value, we use the contributions as shown in Figure 3. By ranking these contributions, positive factors (e.g., ”close proximity to the goal”) and negative factors (e.g., ”poor shooting angle”) are highlighted in the text. For example, if a shot shows large contribution values for the following

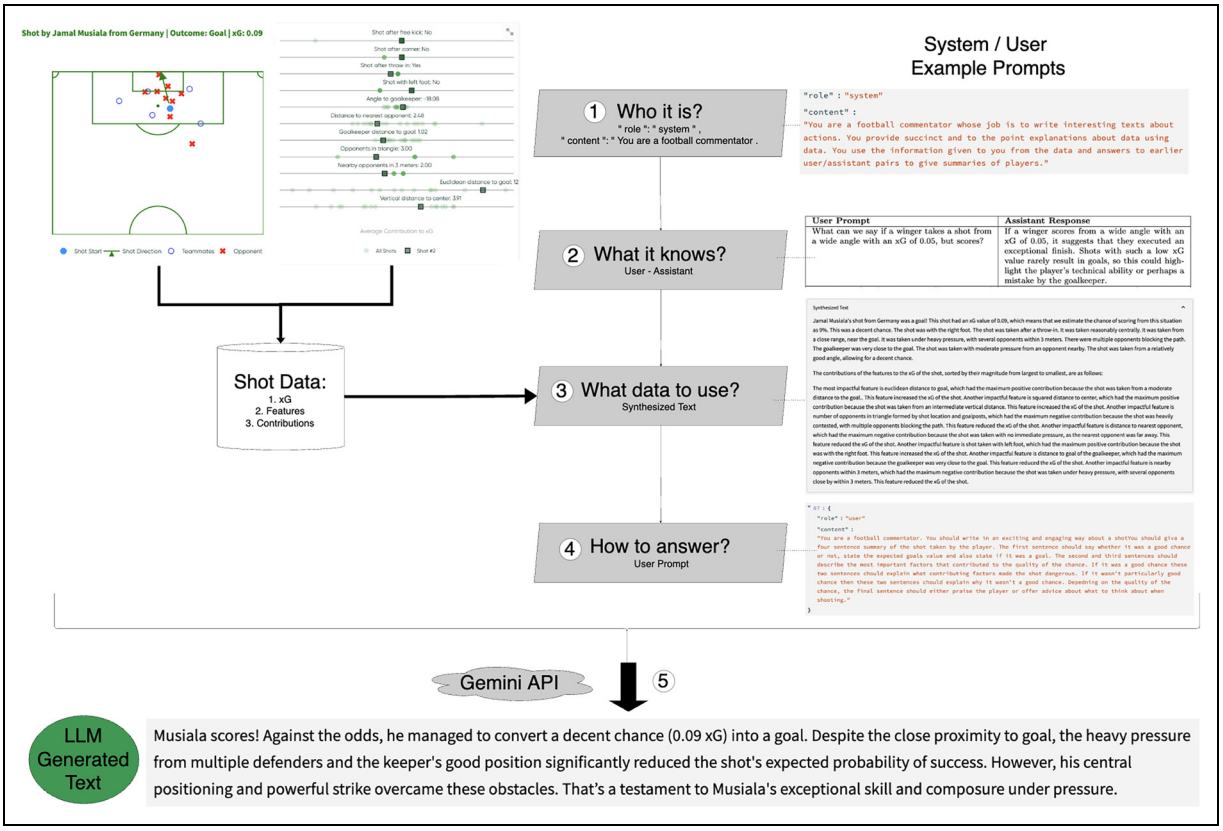

Figure 4. Wordalisation workflow for shots.

features: `euclidean distance to goal`, `nearby opponents in 3 meters`, `opponents in triangle`, the following text is automatically generated:

> "The high chance of scoring was influenced by the player's close position to the goal and minimal defensive pressure."

Only features with contributions greater than 0.1 or less than -0.1 in the log-odds are included in these descriptions, as they are more likely to influence the outcome of the shot prediction. While this is a somewhat heuristic threshold, it is guided by two considerations. First, from a statistical perspective, a change of ± 0.1 in the log-odds corresponds to only a small shift in predicted probability (less than 2.5 percentage points for probabilities around 0.5), which is typically negligible in practical football analysis. Second, we conducted empirical sensitivity checks and observed that feature contributions within this range rarely aligned with meaningful changes in xG and were often noise-like or difficult to interpret consistently across datasets. Moreover, we aimed to strike a balance between *informativeness* and *brevity* in the generated explanations: including too many features with weak effects can reduce clarity and overwhelm the user with marginal details. This threshold thus improves both the interpretability and relevance of the textual output. The value is adjustable and can be revised based on analyst preferences or use cases.

It is noteworthy that this synthesis step is fully automated: we define a function once that dynamically generates a per-shot textual summary by extracting and transforming the relevant shot data. The function can be applied to thousands of shots without additional manual intervention. A full list of such functions used for assigning contributions can be found in the description class of our code https://github.com/Peggy4444/shotsGPT/blob/main/classes/description.py. The gray text in Figure 5(a) explains the impact of individual characteristics on the xG value, ranked by their contribution magnitude.

Tell it how to answer: This stage is focused on crafting the specific instructions and examples that guide the LLM in generating more engaging responses from synthesized text. It gives very specific instructions about the type of text we would like to generate, specifying how many sentences should be written and what each sentence should contain. In addition, we include explicit human-generated examples, a technique known as few-shot prompting (Schulhoff et al., 2024). Figure 5(b) presents a human-written example of a few-shot prompt for the synthesized text shown in Figure 5(a). We provide three training examples of this type for the wordalisation.

```
"Florian Wirtz's shot from Germany was a goal! This shot had an xG value of 0.14,
which means that we estimate the chance of scoring from this situation as 14%.
This was a high-quality chance, with a good probability of scoring."
"The shot was with the right foot. The shot was taken after a throw-in. It was
taken from very close to the center of the pitch. It was taken from a moderate
distance from the goal. It was taken with moderate pressure, with one opponent
within 3 meters. There were multiple opponents blocking the path. The goalkeeper
was very close to the goal. The shot was taken with no immediate pressure from any
close opponent, with the nearest opponent far away. The shot was taken from a
relatively good angle, allowing for a decent chance."
"The contributions of the features to the xG of the shot, sorted by their
magnitude from largest to smallest, are as follows: The most impactful feature is
vertical distance to center, which had the maximum positive contribution because
the shot was taken closer to the center of the pitch (less vertical distance).
This feature increased the xG of the shot. Another impactful feature is euclidean
distance to goal, which had the maximum positive contribution because the shot was
taken from a moderate distance to the goal. This feature increased the xG of the
shot. Another impactful feature is number of opponents in triangle formed by shot
location and goalposts, which had the maximum positive contribution because the
shot was heavily contested, with multiple opponents blocking the path. This
feature increased the xG of the shot. Another impactful feature is shot taken with
left foot, which had the maximum positive contribution because the shot was with
the right foot. This feature increased the xG of the shot."
```

(a)

```
What a strike from Wirtz! Turning a high quality (0.15 xG) chance into a stunning
goal. He opened up a position for himself in a central position on the pitch,
where he had the time and space to carefully place his shot. Despite multiple
opponents between him and the goal, he executed expertly. Wirtz's clinical
finishing turned a throw-in into a valuable goal for his team.
```

(b)

Figure 5. Example synthesized text and few-shot example prompt. (a) Synthesized text. The text highlighted in blue is the initial description of the expected goals. The red text explains the features in footballing terms, and the grey text describes the feature contributions and (b) Few-shot example.

Engagement and accuracy

We do automated evaluation of our wordalisations based on two key criteria: engagement and accuracy. We compare five distinct cases. For case 1, the text provided to the evaluation (denoted as [Case text] in the evaluation prompts below) consists only of shot quality and features (as shown as coloured texts in Figure 5(a)). The idea is to test whether the LLM (Gemini in the examples used here) already has the ability to assess shot value just from a description of the shot, but without additional data. Case 2 extends the text provided in the evaluation prompt to include contributions (as well as shot quality and features). This provides a comprehensive explanation of the shot and the factors influencing its quality. Case 2 tests the engagement and accuracy of a purely descriptive text.

Cases 3 and 4 test the wordalisations. Case 4 produces a text following the complete wordalisation approach by following all the steps described in section "Wordalisation: Step by Step Prompt". Case 3 omits the 'tell it what it knows' and 'tell it how to answer' stages, to help assess how important these parts of the prompts are in shaping an accurate answer. Finally, case 5 serves as a baseline, providing only numerical feature values without any textual explanation or narrative.

The aim of our **engagement** evaluation is to measure how interesting the generated descriptions are to readers. To calculate engagement using an LLM, we first provide the text we want to evaluate, then ask *"Rank this text on a scale from 0 to 5 for how interesting and engaging it is."* This process is repeated for all shots, and the engagement score for each description is then averaged. To ensure robustness, the system includes error handling mechanisms in case of failed responses, retrying the request multiple times.

For **accuracy** we evaluate how well the generated descriptions align with the true contribution of individual features to the expected goals (xG) value. To do this, an LLM is provided with a prompt that asks it to assess whether a particular feature (such as Euclidean distance to goal or vertical distance to centre) is a positive, negative, or neutral contributor to the xG value. Specifically, we

Figure 6. Feature contribution analysis and LLM generated text of two shots from Washington Spirit - North Carolina Courage in National Women's Soccer League (NWSL) 2018. (a) 2th minute shot by McCall Zerboni and (b) 10th minute shot by Jessica McDonald.

write the prompt *"In the following text [Case text] was [Feature] a positive, negative, or not contributing factor? Respond with one of ['positive', 'negative', 'not contributing']"* The output labels are then compared to the ground truth, where features are considered positive if their contribution exceeds 0.1, negative if it is below -0.1, and neutral (not contributing) if it lies between -0.1 and 0.1. The accuracy score is calculated as the percentage of correct assessments made by the LLM across all shot descriptions.

Results

Shot description application

In order to demonstrate our approach we built a shot description application in Streamlit https://shotsgpt.streamlit.app/. The application allows the user to select a match from one of the available tournaments, then a shot from that match and it compares the selected shot to the other shots in the match in a distribution plot, shows the location of players and the ball in that shot and writes a short commentary about the shot. The application also allows the user to see the steps used in building the wordalisation: the model summary of the fitted logistic regression; the synthesized text at the "tell it what data to use" stage; and the full sequence of messages sent to the language model. We provide the full code for this application on Github: https://github.com/Peggy4444/shotsGPT/tree/main.

Feature contributions

The contribution plots visualize feature importance (Figure 6). Each horizontal band in the plot represents a feature, with its width indicating the magnitude of its contribution. Shots with values to the right of the vertical axis have more xG, while those to the left have less xG. In these plots, `euclidean distance to goal` and `vertical distance to centre` generally emerge as dominant factors. For these

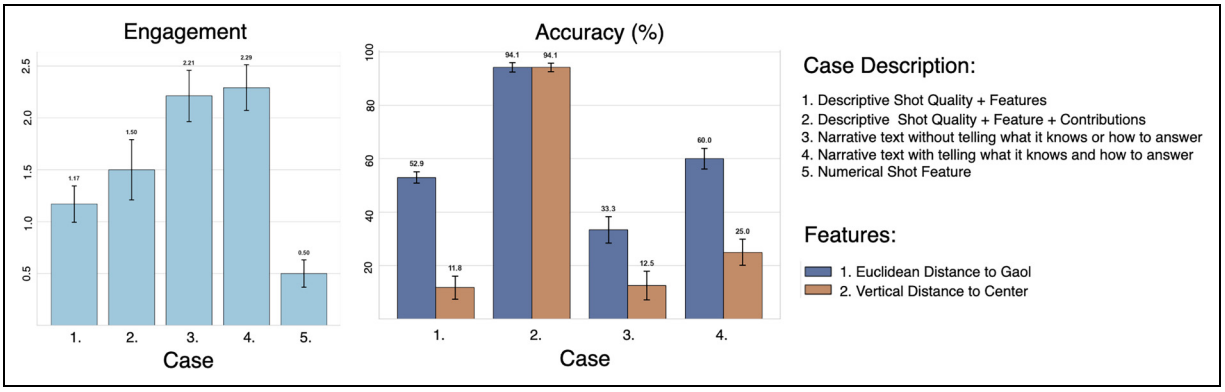

Figure 7. Engagement and Accuracy scores. The results are averaged over 10 runs and standard deviation is shown on top of the bars.

variables, the shots are spread out further on the scale: shots from large distances (far on the left), for example, have a reduced probability of goal, while those at short distance (far to the right) have an increased probability. Figure 6 highlights two contrasting cases. In Figure 6(a), the shot (which did result in a goal) had a relatively large `distance to nearest opponent`, increasing the chance of scoring. Conversely, Figure 6(b) illustrates an unsuccessful shot where a small `distance to nearest opponent` significantly reduced the xG. Notably, in both instances, this opponent proximity feature outweighs the typical dominant variables, underscoring the context-dependent nature of feature contributions. The accompanying LLM-generated analysis aligns with these observations, accurately capturing how situational factors alter the relative impact of features on xG predictions.

Model card

The model card provides a comprehensive overview of the system's design, capabilities, and limitations. It documents both components of the pipeline: the xG prediction model and the language model used for wordalisation. For the xG model, the card specifies the use of logistic regression, the rationale for its selection (interpretability), and details of the training data used per competition. To ensure transparency about predictive performance, the model card includes evaluation metrics for each competition, namely ROC-AUC (discrimination), Brier score (calibration), and log-loss (probabilistic confidence). These metrics are reported separately for datasets such as EURO 2024, FIFA 2022, FAWSL, and AFCON 2023, allowing the assessment to the robustness and reliability of the xG estimates across contexts. The card also includes structured prompt templates used for generating natural language explanations from feature contributions, as well as limitations of the language model outputs, including sensitivity to input phrasing and lack of domain-specific grounding. Under ethical considerations, the model card highlights issues related to dataset bias (e.g., league-specific play styles), reproducibility (open-sourced code and reproducible data pipeline), and responsible AI usage (exclusion of goalkeepers, and transparency of feature attribution logic). It follows the model card framework (Mitchell et al., 2019) and aims to support informed deployment in decision-making environments. The full model card is publicly accessible via: https://github.com/Peggy4444/shotsGPT/blob/main/model%20cards/model-card-shot-xG-analysis.md

Evaluation

Figure 7 shows that there is a trade-off between engagement and accuracy in the generated descriptions. Case 2 (i.e., descriptive shot quality + features + contributions) achieves the highest accuracy, for key features such as `Euclidean distance to goal` and `vertical distance to centre`, as it explicitly lists and explains feature contributions. This makes it easier for the LLM to identify correct contribution labels. These two features were chosen for accuracy evaluation because they exhibit the strongest influence on xG values, as evidenced by their wider distribution in the contribution plot (Figure 3). However, Case 2's engagement score is low. In contrast, case 4 (i.e., full wordalisation), which leverages contextual examples and narrative elements, strikes an optimal balance between accuracy and engagement. It achieves the second-highest accuracy while maintaining the highest engagement score, making it the most suitable for practical use. This balance ensures that the generated descriptions are not only reliable and explainable but also accessible and engaging for football practitioners. The results demonstrate that Case 4 effectively addresses the needs of analysts and coaches by providing insightful, interpretable, and actionable insights into xG values and feature contributions.

Further applications

So far we have applied this method to evaluating shots, but the same concept can be used to evaluate other actions in

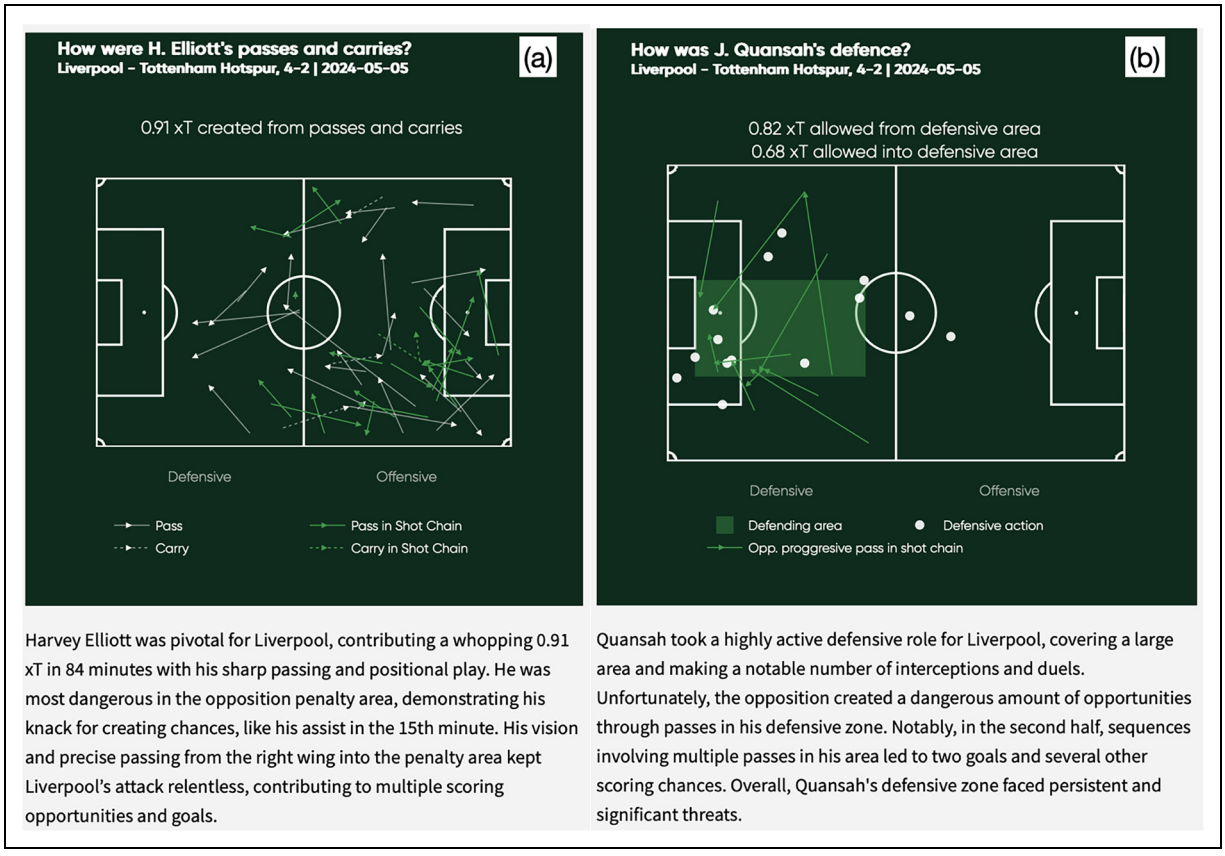

Figure 8. Wordalisation applications in passes, carries and defensive actions.

football. An example is shown in Figure 8 for both attacking (8a) and defensive (8b) actions. For attack, we assign an "expected threat" value to every pass and carry made by a player and use this to describe the most common type of pass the player makes. The first step in this process was to create an action-based expected threat model (Sumpter, 2023), using three seasons of event data across the French, English, German, Spanish and Italian leagues. This model is a logistic regression predicting the probability that a pass, will eventually be part of a chain of actions leading to a goal. The model is then interpreted in the context of individual players by creating synthetic descriptions which summarises the best passes and carries by a player, and also including details of where those passes occurred on the pitch. These synthesized texts are then passed to GPT4o along with both "Tell It What It Knows" question-answer pairs and "Tell It How To Answer" examples. The output is an engaging text, explaining how the player's passes contribute to the team.

Discussion

There are three steps to the process we have outlined for generating natural language narratives describing football shots. The first is to create a mathematical model of the probability of a shot being a goal, in our case a logistic regression. By focusing on variables which are interpretable, we ensure that, at the second step, we can convert the outcome of this model into words. Neither of these steps uses language models and instead we use "old-fashioned" statistical models to fit an expected goal model to data. The linear nature of the logistic regression ensures that we make a correct interpretation of the variables. The third and final step involves combining the "tell it what data to use" text with a series prompts to produce an engaging text about the shots. The resulting text is both engaging and factually correct.

In terms of explaining what makes a chance good (or poor), we see our approach as an improvement on the SHAP-based feature importance approach Anzer and Bauer (2021). The wordalisations not only retain the model's accuracy but also make its outputs more accessible and actionable for end-users. Our automated evaluation methods show that there is a trade-off between an engaging description and an accurate description of all aspects of the shot. This is to be expected, if a coach were to describe the quality of a shot to a player, we would not expect them to give all details in every description.

Our work contributes to the theoretical foundation of wordalisations (Caut et al., 2025), by extending its

application to logistic regression models and LLM-generated explanations. Unlike previous implementations that focused on raw numerical rankings, we ask LLMs to interpret the output of machine learning models. Our approach can be extended to other models, such as the expected threat model in Figure 8. Similarly, any model — such as pitch control (Spearman et al., 2017) and off-ball runs (Peralta Alguacil et al., 2020)— which describes positioning and actions of players can, by following the three steps outlined here, be converted into an informative wordalisation.

Broadly speaking, models of football can be divided into two approaches: those which give an explicit description of a mechanisms (as discussed in the previous paragraph) and those which use machine learning to make predictions. Although we don't do so here, our approach can potentially be adapted to a more general machine learning setup through the SHAP values of a model (Anzer and Bauer, 2021; Lundberg and Lee, 2017). When introducing SHAP, Lundberg and Lee (2017) let g denote the explanation model approximating the original predictive function f. The explanation for a prediction $f(\mathbf{x})$, where $\mathbf{x}$ is the input, is modelled using a linear equation:

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i',$$

where: - $\mathbf{z}'$ are the simplified input features, - ϕ_0 is a baseline value, typically the mean prediction of the model, - ϕ_i represents the contribution of each feature z_i' to the model's prediction. For our logistic regression model, equation 1 provides a ready-made $g(\mathbf{z}')$ in the form of log-odds, because it is linear in the features. This is not the case for most machine learning models. The challenge then to building wordalisations for general machine learning models is to select first the simplified input features and then to automate the production of tests around those features. Anzer and Bauer (2021) make the first of these steps for an expected goals model based on xGBoost, the further step of wordalising this approach remains an interesting and open research challenge. An important question, however, is whether such an approach is really needed when a more mechanistic approach (based on logistic regression) is so effective.

Coaches often require insights that are not only accurate, but also easily digestible and actionable Forcher et al. (2024); Goes et al. (2021). Our system provides insight by converting complex numerical outputs into intuitive, text-based narratives that highlight key factors influencing xG values, such as shot distance, angle, and defensive pressure. This could allow coaches or players to quickly grasp why a shot has a high or low xG value, enabling more informed decision-making during training and matches. Our system is designed to generalise across competitions by training separate logistic regression models for each dataset using the same set of input features. This ensures that while the explanatory framework remains consistent, the model coefficients–and hence, the weighting of contextual factors–are tailored to the specific style and dynamics of each competition. The modular design also allows for domain-specific adaptation: practitioners can revise the feature set based on league-specific knowledge, and the resulting wordalisations will automatically reflect these changes. We see this as an opportunity for future work, particularly in supporting analysts working across different footballing environments.

The current system's output has been evaluated using large language models (LLMs), which rated the descriptions based on engagement and alignment with feature contributions. While this provides a scalable proxy for quality of the texts, future work will involve designing structured evaluation studies. These will assess how textual explanations affect the perception and understanding of xG by users with varying levels of football and statistical expertise, evaluating both the practical impact of the narratives on decision-making, as well as the accuracy of the automated evaluation. A natural next step is to look at human evaluation by coaching staff. Do the coaches find these descriptions accurate? And, even more importantly, are they useful in coaching situations?

In summary, we have taken an approach which emphasizes model explainability, not just in a statistical sense, but also in the sense that our models explain the value of a shot in plain language. We believe that machine learning practitioners should endeavour to take this approach, which will further help analysts and coaching staff better utilize data without requiring deep expertise in machine learning.

ORCID iD

Pegah Rahimian https://orcid.org/0000-0002-0293-2739

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Notes

1. https://statsbomb.com/news/statsbomb-release-free-euro-2024-data/
2. https://github.com/statsbomb/statsbombpy
3. https://github.com/soccermatics/twelve-gpt-educational/blob/shots/data/describe/action/shots.xlsx

References

Anzer G and Bauer P (2021) A goal scoring probability model for shots based on synchronized positional and event data in football (soccer). *Frontiers in Sports and Active Living* 3. https://api.semanticscholar.org/CorpusID:232387328

Bransen L and Davis J (2021) Women's football analyzed: Interpretable expected goals models for women. In: *AI for Sports Analytics (AISA) Workshop at IJCAI 2021*. Montreal, Canada.

Brown TB, Mann B, Ryder N, et al. (2020) Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877–1901.

Caut AM, Rouillard A, Zenebe B, et al. (2025) Representing data in words. https://arxiv.org/abs/2503.15509.

Cefis M and Carpita M (2025a) Accuracy and explainability of statistical and machine learning xg models in football. *Statistics* 59(2): 426–445. DOI: 10.1080/02331888.2024.2445305

Cefis M and Carpita M (2025b) A new xg model for football analytics. *Journal of the Operational Research Society* 76(1): 1–13. DOI: 10.1080/01605682.2024.2323669

Davis J, Bransen L, Devos L, et al. (2024) Challenges in sports analytics: Methodological and evaluation considerations. *Machine Learning* 113: 6977–7010.

Davis J and Robberechts P (2020) How data availability affects the ability to learn good xg models. In: *Proceedings of the 7th International workshop of machine learning and data mining for sports analytics*.

Decroos T, Bransen L, Van Haaren J, et al. (2019) Actions speak louder than goals: Valuing player actions in soccer. In: *In ACM KDD*.

Dick U and Brefeld U (2023) Action rate models for predicting actions in soccer. *AStA Advances in Statistical Analysis* 107: 29–49. DOI: 10.1007/s10182-022-00435-x

Eggels H, Van Elk R and Pechenizkiy M (2016) Explaining soccer match outcomes with goal scoring opportunities predictive analytics. In: *Proceedings of the workshop on machine learning and data mining for sports analytics 2016 co-located with the 2016 European conference on machine learning and principles and practice of knowledge discovery in databases*. Garda, Italy.

Fernandez J, Bornn L and Cervone D (2019) Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. In: *13th MIT Sloan sports analytics conference*.

Forcher L, Forcher L and Altmann S (2024) How soccer coaches can use data to better develop their players and be more successful. In: *Individualizing training procedures with wearable technology*, pp.99–123. Springer.

Goes F, Meerhoff L, Bueno M, et al. (2021) Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European Journal of Sport Science* 21(4): 481–496.

Gyarmati L and Stanojevic R (2016) A merit-based evaluation of soccer passes. In: *ACM KDD Workshop on Large-Scale Sports Analytics*.

Herbinet C (2018) *Predicting Football Results Using Machine Learning Techniques*. Imperial College London: Meng thesis.

Lucey P, Bialkowski A, Monfort M, et al. (2015) quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In: *MIT Sloan sports analytics conference*.

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. *NeurIPS* 30.

Mitchell M, Maziarka P, Kamar E, et al. (2019) Model cards for model reporting. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp.1–13.

Morales CA (2016) A mathematics-based new penalty area in football: Tackling diving. *Journal of Sports Sciences* 34(24): 2233–2237.

Pardo M (2020) *Creating a model for expected goals in football using qualitative player information*. Master's thesis, Universitat Politecnica de Catalunya.

Peralta Alguacil F, Fernandez J, Piñones Arce P, et al. (2020) Seeing in to the future: using self-propelled particle models to aid player decision-making in soccer. In: *In Proceedings of the 14th MIT sloan sports analytics conference*.

Pollard R and Reep C (1997) Measuring the effectiveness of playing strategies at soccer. *Journal of the Royal Statistical Society Series D: The Statistician* 46(4): 541–550.

Rahimian P, Van Haaren J, Abzhanova T, et al. (2022) Beyond action valuation: A deep reinforcement learning framework for optimizing player decisions in soccer. In: *16th MIT sloan sports analytics conference*.

Rahimian P, Van Haaren J and Toka L (2023) Towards maximizing expected possession outcome in soccer. *International Journal of Sports Science and Coaching* 19: 230–244.

Rathke AAT (2017) An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise* 12: 514–529. https://api.semanticscholar.org/CorpusID:148713007

Reynolds L and McDonell K (2021) Prompt programming for large language models: Beyond the few-shot paradigm. *arXiv preprint arXiv:2102.07350*.

Ruiz H, Lisboa PJG, Neilson P, et al. (2015) Measuring scoring efficiency through goal expectancy estimation. In: *The European symposium on artificial neural networks*.

Sarkar S and Kamath S (2021) Does luck play a role in the determination of the rank positions in football leagues? a study of europe's big five. *Annals of Operations Research*: 245–260. DOI: 10.2202/1559-0410.1014.

Schulhoff S, Ilie M, Balepur N, et al. (2024) The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Spearman WR, Basye AT, Dick GJ, et al. (2017) Physics-based modeling of pass probabilities in soccer. In: *MIT Sloan sports analytics conference*.

Sumpter D (2016) *Soccermatics: Mathematical Adventures in the Beautiful Game*. Bloomsbury Publishing.

Sumpter DJT (2023) Expected threat - action-based. https://soccermatics.readthedocs.io/en/latest/lesson4/xTAction.html.

Tippana T (2020) How accurately does the expected goals model reflect goalscoring and success in football? Bachelor's thesis, Aalto University, School of Business. https://aaltodoc.aalto.fi/server/api/core/bitstreams/43d75b97-553d-40eb-8aaf-79dadbe54514/content.

Wei J, Wang X, Schuurmans D, et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Wheatcroft E and Sienkiewicz E (2021) A probabilistic model for predicting shot success in football. *arXiv preprint arXiv:2101.02104*.