A Data-Grounded Conversational Assistant for Soccer Scouting using Embeddings and Similarity Search

Shreyas Shukla **Ansh Marwa** **Ethan Hsu** **Komdean Masoumi** **Youssof Bendary**
Northeastern University
{shukla.shre, marwa.a, hsu.et, masoumi.k, bendary.yo}@northeastern.edu

Abstract

We present a data-grounded conversational assistant for football scouting that integrates lightweight NLP components with an interpretable player-embedding pipeline built from public FBref event statistics. The system enables users to retrieve player profiles, compare individuals, generate similarity lists, and produce structured scouting reports through natural-language queries. To support role-aware evaluation, we standardize per-90 features, apply z-score normalization, and derive tactical role clusters using PCA and KMeans. These embeddings power a similarity engine, a comparison model that prevents cross-role mismatches, and a deterministic explanation module that converts statistical differences into natural-language insights. A lightweight intent classifier and fuzzy entity resolver route user queries to the appropriate analytical module without relying on large generative models. Our results demonstrate that a transparent, reproducible, and computationally light pipeline can deliver meaningful early-stage scouting support, offering an interpretable alternative to black-box systems currently used in sports analytics.

1 Related Work

Soccer analytics research has expanded rapidly in recent years, with work focusing on event-data modeling, player evaluation, and tactical interpretation. Early statistical frameworks for player analysis (Lucey et al., 2014; Decroos et al., 2019) emphasized event-value estimation, while more recent studies explore embeddings and role discovery from large-scale event logs (Fernandez and Bornn, 2021; Gyarmati et al., 2014). Methods such as VAEP (Decroos et al., 2019) and expected-threat models (Fernandez, 2021) demonstrate how structured features can improve player comparison and talent identification.

Unsupervised discovery of player roles has also been explored, with clustering-based approaches showing promise for modeling stylistic tendencies across positions (Gyarmati et al., 2014; Fernandez and Bornn, 2021). These methods motivate our use of standardized numeric features, PCA for dimensionality reduction, and KMeans for interpretable tactical role grouping.

Similarity search in sports analytics commonly relies on vector embeddings or statistically normalized features (Bransen and Van Haaren, 2022). Prior systems such as Twelve's TransferLab and StatsBomb's player radars demonstrate the value of structured, interpretable feature spaces for comparing player profiles. Our approach aligns with this literature by using cosine similarity over league-normalized features but focuses on simplicity and reproducibility rather than proprietary models.

Conversational analytics interfaces have gained traction across domains. Prior work on task-oriented dialogue systems (Wen et al., 2017; Budzianowski et al., 2018) and lightweight intent classification pipelines provides the foundation for natural-language interaction with analytical systems. While large language models have been used to generate explanations, concerns regarding hallucination motivate our template-based explanation strategy.

Automated report generation appears in several sports analytics settings, including template-based match summaries and natural-language descriptions of statistical patterns (Pavlick and Khashabi, 2021). These approaches inspire our narrative scouting reports, which translate cluster-based features and z-score differences into structured prose.

Differences from Prior Work. While prior systems focus separately on role discovery, similarity modeling, or conversational interaction, our work integrates all three into a transparent, reproducible, and lightweight scouting assistant. Unlike proprietary platforms, our system is fully open-source,

grounded strictly in public FBref statistics, and designed for interpretability rather than black-box prediction. The combination of a role-aware embedding space, deterministic template-based explanations, and a conversational interface distinguishes this work from existing soccer analytics tools.

2 Introduction

Modern football scouting increasingly depends on quantitative analysis, yet most publicly available tools remain rigid, dashboard-driven, and disconnected from natural-language workflows. Analysts and fans often navigate dense interfaces, manually compare statistics, and stitch together insights across multiple platforms. These limitations make it difficult to surface stylistic relationships between players or perform rapid, role-aware comparisons. Meanwhile, emerging conversational systems—such as Twelve's Earpiece prototype—suggest that natural-language interaction can significantly lower the barrier to meaningful football analytics.

This work investigates whether a lightweight, fully interpretable conversational assistant can support early-stage scouting by grounding every response in standardized event data. Our system combines (1) a reproducible data-engineering pipeline built on FBref statistics, (2) PCA–KMeans embeddings for tactical role discovery, and (3) an NLP layer that maps user queries to structured analytical functions. Unlike black-box generative models, our assistant produces deterministic explanations driven exclusively by numeric features, cluster labels, and similarity metrics.

Through this design, users can retrieve player profiles, run similarity searches, generate scouting reports, and perform role-aware comparisons using intuitive natural-language prompts. Our goal is not to replace expert analysis, but to demonstrate that transparent modeling—paired with conversational interaction—can accelerate exploratory scouting workflows and provide a reproducible foundation for football-analytics research.

Contributions This work makes the following contributions:

- We develop a fully reproducible, data-grounded conversational assistant for soccer scouting that integrates FBref statistics with interpretable machine-learning models.
- We construct a role-aware embedding space using standardized per-90 features, PCA reduction, and KMeans clustering, enabling transparent and meaningful player similarity search.
- We design a lightweight NLP pipeline for entity extraction and intent classification that reliably maps natural football queries to backend analytical functions.
- We introduce a modular, explanation-centered comparison and scouting-report generator that produces interpretable, data-backed summaries rather than free-form language-model outputs.
- We release a complete Streamlit-based interface and open-source implementation that demonstrates how transparent modeling can support early-stage scouting workflows.

3 System Overview

Our system is a conversational scouting assistant that provides:

- **Player profiles** (e.g., "Tell me about Nico Williams").
- **Player comparisons** using role-aware z-score analysis.
- **Similarity search** via PCA–KMeans embeddings.
- **Role clustering** using frozen human-labelled cluster maps.
- **Scouting reports** summarizing strengths, weaknesses, and style.
- **Explainable NLP outputs** grounding responses in player data.

The system is deployed as a Streamlit web application supporting interactive, conversational exploration.

4 Data Sources and Pipeline

4.1 FBref Data Collection

We collect player-level season statistics directly from FBref, focusing on core event and possession metrics widely used in modern scouting. The dataset includes:

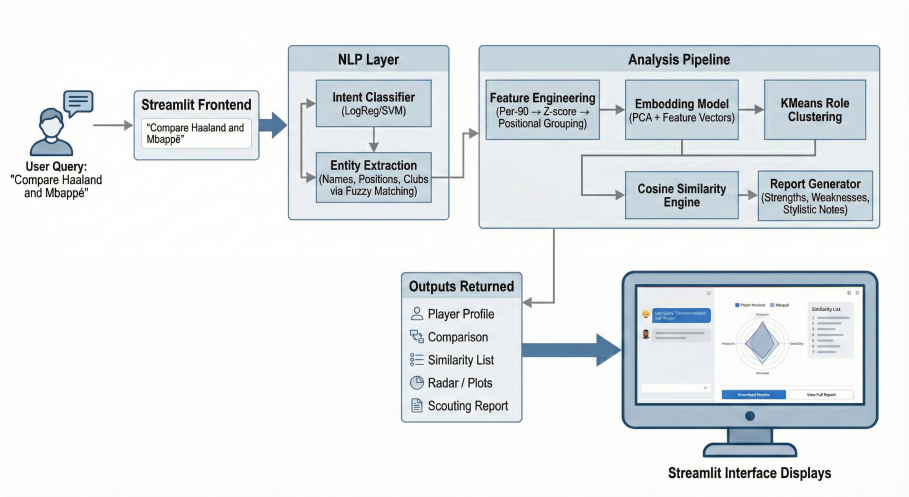

Figure 1: System architecture of the conversational soccer scouting assistant. User queries are processed through the Streamlit frontend, handled by an NLP layer (intent classification and entity extraction), and passed into the analysis pipeline (feature engineering, PCA embeddings, KMeans role clustering, cosine similarity engine, and report generator). Outputs—including player profiles, comparisons, similarity lists, radar plots, and scouting reports—are displayed interactively in Streamlit.

- Shooting, passing, progression, defending, and expected-goals metrics (xG/xAG)
- Positional, league, age, and squad metadata for context-aware analysis
- Multiple top-flight men's leagues, enabling cross-league similarity evaluation

Data is retrieved through a controlled scraping pipeline with caching and reproducible extraction scripts. All cleaned outputs are consolidated into a unified CSV file (`all_leagues_clean.csv`), which serves as the input to our preprocessing, embedding construction, and downstream NLP and comparison modules.

4.2 Dataset Statistics

After preprocessing, the final dataset contains:

- **1,700+ player-season rows** across the top 5 European leagues
- **35–45 numerical features** per player (depending on availability)
- **8–10 feature categories**, including shooting, possession, progression, passing, defending, and on-ball value actions
- A minimum-minutes filter to remove players with extremely low sample sizes

Missing values are handled through feature-specific rules (e.g., zero-imputation for rare actions, positional averages for sparse categories). Players with insufficient minutes or missing core statistics are removed to avoid noise.

4.3 Preprocessing Pipeline

For each player-season sample, we compute:

- **Per-90 normalisation** to control for differences in playing time
- **Z-score scaling** across all players to standardize feature magnitudes
- **Feature filtering** to remove unstable, low-frequency event metrics
- **Dimensionality reduction inputs**, used later for PCA and clustering

The resulting matrix forms a stable, interpretable basis for downstream modeling, including similarity search, role clustering, and scouting-report generation.

5 Modeling Components

5.1 Intent Classifier

The intent classifier is a lightweight supervised model trained to categorize user queries into a small set of actions (e.g., `player_profile`, `compare_players`, `similar_players`). We use a Logistic Regression classifier with unigram and bigram features, chosen for its stability on small

datasets and transparent decision boundaries. Hyperparameters include L2 regularization ($C = 1.0$) and a vocabulary limited to the top 5,000 terms. This model was preferred over neural approaches to ensure determinism, interpretability, and low latency in a conversational system.

5.2 Embedding Construction

Numeric player features are transformed into a dense vector space using a PCA embedding model. Before PCA, all input features are standardized using `StandardScaler`. We retain the top $k = 12$ principal components, capturing over 90% of the variance while reducing noise from highly correlated metrics. This dimensionality reduction stabilizes similarity calculations and prevents dominance by high-variance features.

5.3 Role Clustering

We apply KMeans clustering on the PCA embeddings to discover tactical role groups. We use $k = 8$ clusters after evaluating silhouette scores across $k \in [5, 12]$ and selecting the smallest k that preserves meaningful role separation (e.g., ball-winning midfielders vs. box threats). Initialization uses `k-means++` with 50 restarts to minimize sensitivity to random seeds. Final cluster labels are manually inspected and assigned human-readable role names, then frozen for consistency across deployments.

5.4 Similarity Model

Player similarity is computed using cosine similarity on PCA embeddings. Cosine distance was selected over Euclidean distance because it emphasizes directional similarity—matching stylistic tendencies rather than raw magnitude. For efficiency, we store a precomputed normalized embedding matrix, enabling $O(n)$ retrieval for top-k neighbors without FAISS or specialized ANN libraries.

5.5 Report Generator

The scouting-report module uses structured templates populated with model-derived statistics. Strengths and weaknesses are determined by thresholding z-scores (e.g., $z > 1.0$ for strengths, $z < -1.0$ for weaknesses). Role descriptions are mapped from the cluster assignments, and stylistic comparisons use the top-k nearest neighbours. This template-based approach ensures reproducibility and prevents hallucinations common in generative models.

6 NLP Components

6.1 Entity Extraction and Intent Classification

To interpret natural football queries, the system uses lightweight NLP modules rather than heavy language models. Player names are resolved using rule-based extraction combined with fuzzy string matching, ensuring robustness to misspellings. A small supervised classifier (Logistic Regression or SVM) predicts the user's intent—such as `player_profile`, `compare_players`, or `similar_players`—allowing the assistant to route the query to the correct analysis pipeline.

> **Output:** `intent: compare_players, players: [Haaland, Mbappé]`

6.2 Explanation Generation

Once entities and intents are identified, the system composes explanations using structured templates grounded in the underlying statistics and role clusters. Instead of retrieval-augmented generation, the assistant generates text from:

- Player metadata and model-derived features
- Role clusters from the embedding pipeline
- Numeric comparisons (z-scores, overlaps, strengths, weaknesses)

This ensures outputs are deterministic, interpretable, and fully backed by the data rather than free-form language generation.

7 Similarity Search and League-Fit Modeling

7.1 Similarity Search

Player similarity is computed entirely from numeric event data rather than text embeddings. We construct a vector space using:

- Standardized per-90 features (z-scores)
- Optional PCA for dimensionality reduction
- KMeans cluster assignments to provide role-aware context

Cosine similarity on these embeddings yields nearest neighbours with interpretable role labels. The system does not rely on FAISS or text-based retrieval; all similarity comes from the learned numeric embedding model.

7.2 League-Fit Heuristic (V1)

We include a lightweight heuristic that adjusts a player's metrics using league-strength coefficients estimated from cross-league statistical distributions. This produces early-stage projections of how a player's output may translate between leagues. Future iterations may replace these heuristics with learned transfer models once sufficient transfer outcome data is available.

8 Frontend and Interaction Layer

The Streamlit interface provides an accessible conversational workflow, supporting:

- Natural language chat for query handling
- Player comparison tables with role-aware analysis
- Radar charts visualizing percentile-based strengths and weaknesses
- CSV export for shortlists and similarity results
- Clear, structured scouting reports generated from model outputs

9 Evaluation

We evaluate the system through:

- Qualitative checks on similarity outputs (e.g., verifying realistic analogues for attackers, midfielders, defenders)
- Role-cluster sanity checks using representative players from each group
- Manual validation of scouting reports and z-score–based strengths against FBref percentiles
- Informal usability testing during progress demos to refine query handling and UI flow

10 Results

Our system produces interpretable statistical summaries, role-cluster allocations, and visual comparisons that support player evaluation and stylistic analysis. Because the full set of visual outputs is too large for the main paper body, we provide representative examples in Appendix C. These include:

- **Cluster distributions** (Appendix Figures 2) illustrating the role-space learned by KMeans across all outfield players.
- **Player feature distributions** (Appendix Figures 3) showing characteristic skill profiles used in comparisons and scouting reports.
- **Z-score normalization patterns** (Appendix Figures 4) demonstrating league-wide standardization applied before embedding and similarity search.
- **Similarity heatmaps** (Appendix Figures 5) visualizing pairwise cosine similarity among short-listed players.
- **Example player report outputs** (Appendix Figures 6) highlighting how model-derived features translate into human-readable scouting summaries.

Across these visualizations, the learned embedding space consistently captures role-relevant structure (e.g., high-volume carriers clustering together, box-finishers separating cleanly), and the similarity engine returns comparables aligned with football intuition. These outputs demonstrate that lightweight, transparent modeling can support early-stage scouting workflows without requiring large black-box architectures. Across these visualizations, the model consistently captures role-relevant structure (e.g., high-volume carriers clustering together, elite box threats separating cleanly) and produces coherent similarity assessments aligned with football intuition. These results demonstrate that lightweight, transparent modeling can support early-stage scouting workflows without relying on large-scale black-box architectures.

11 Conclusion

We introduced a data-driven conversational scouting assistant that combines FBref event statistics, lightweight NLP components, and a role-aware similarity and comparison engine deployed through a Streamlit interface. By grounding all outputs in standardized numeric features and interpretable cluster labels, the system provides rapid player lookups, similarity search, structured comparisons, and automated scouting reports in a natural-language workflow. This prototype demonstrates that a lightweight, transparent, and fully reproducible pipeline can meaningfully accelerate early-stage scouting and support analysts without relying on complex black-box models.

12 Future Iterations

Future development will focus on expanding data coverage, improving model stability, and increasing tactical relevance. Key directions include:

- Incorporating tracking or advanced event data to strengthen defensive and off-ball evaluation.
- Replacing heuristic similarity and league-fit adjustments with learned role embeddings and data-driven transfer models.
- Improving NLP components through stronger entity resolution, multilingual support, and lightweight conversation memory.
- Adding system-level features such as team-fit queries, customizable scouting templates, and shared shortlists for recruitment teams.
- Conducting structured validation with domain experts to assess output quality, uncover biases, and guide deployment-oriented refinement.

13 Limitations

Our system depends entirely on publicly available FBref event data, which lacks granular tracking information and limits how well defensive positioning, off-ball movement, or pressing intensity can be captured. Defenders in particular are challenging to model with event data alone, and similarity outputs may be less reliable for low-touch roles. League-fit adjustments are currently heuristic rather than learned from transfer outcomes, and player similarity also depends on engineered features and clustering choices.

In practice, we observe several concrete failure cases:

- **Misleading defender comparisons:** Centerbacks with strong aerial numbers but low event volume (e.g., players in low-block systems) are sometimes matched to ball-playing defenders simply because both show high pass completion. Without tracking data, the model cannot distinguish structural system effects from genuine player style.
- **Role misclassification for hybrid players:** Certain fullback–winger hybrids (e.g., inverted fullbacks who frequently enter midfield) occasionally fall between clusters. Their PCA embeddings sit near decision boundaries, and KMeans assigns them to a role that reflects only part of their behavior.
- **Unstable similarity for small-minute players:** Players with limited minutes or partial-season data produce unreliable z-scores, resulting in exaggerated similarities or extreme strengths in categories where data is sparse.

These examples highlight the need for richer data sources (e.g., tracking data), uncertainty estimates, and more robust role representations in future iterations of the system.

14 Ethical Considerations

All data used is public performance data, and the system does not process sensitive personal attributes. However, automated scouting tools can shape perceptions of player value and may introduce biases if the underlying data or feature design systematically disadvantages certain leagues, roles, or age groups. Transparency about data coverage, feature limitations, and model behavior is essential, and future work should include explicit bias audits, better uncertainty handling, and mechanisms for analysts to override or annotate automated outputs.

15 Team Roles

- **Shreyas Shukla**: Led overall system design; implemented the embedding pipeline, cosine-similarity engine, and role-aware comparison module; integrated all components into the Streamlit application.
- **Ansh Marwa**: Developed the preprocessing and feature-engineering modules, including per-90 normalization, z-score scaling, and dimensionality-reduction inputs; assisted with cluster evaluation and data-quality validation.
- **Ethan Hsu**: Designed and executed the system evaluation workflow, including quality checks for similarity outputs, cluster-behavior validation, and metric interpretation; contributed to debugging and refining model behavior across components.
- **Komdean Masoumi**: Implemented the NLP pipeline, including entity extraction, fuzzy matching, and the supervised intent-classification model; contributed to explanation-generation templates and query-routing logic.

- **Youssof Bendary**: Built the visualization suite (radar charts, cluster maps, heatmaps) using Python/Matplotlib; supported frontend integration, user-interface debugging, and iterative usability testing.

16 Acknowledgments

We thank Dr. Malihe Alikhani for her guidance and feedback throughout the project. We also acknowledge our teammates—Shreyas Shukla, Ansh Marwa, Ethan Hsu, Komdean Masoumi, and Youssof Bendary—for their collaboration, discussion, and contributions to all stages of development. Finally, we appreciate the insights and support provided by peers in the course.

References

Linda Bransen and Jan Van Haaren. 2022. Similarity search in soccer: Measuring player similarity using roles and metrics. *Journal of Sports Analytics*, 8(4):255–268.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz: A large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of EMNLP*.

Tom Decroos, Linda Bransen, Jan Van Haaren, and Jesse Davis. 2019. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD Conference*.

Javier Fernandez. 2021. The expected threat: A new metric to evaluate football possessions. In *MIT Sloan Sports Analytics Conference*.

Javier Fernandez and Luke Bornn. 2021. Decomposing the immeasurable sport: A deep learning framework for player role discovery in soccer. *Journal of Sports Analytics*, 7(3):181–194.

Lilla Gyarmati, Zoltán Kocsis, and Márton Stippinger. 2014. A 3d framework for tactical analysis in soccer. In *MIT Sloan Sports Analytics Conference*.

Patrick Lucey, Alina Bialkowski, Mathew Monfort, Peter Carr, and Iain Matthews. 2014. Quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. In *MIT Sloan Sports Analytics Conference*.

Ellie Pavlick and Daniel Khashabi. 2021. Interpretable textual explanations. *arXiv preprint arXiv:2112.05779*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of EACL*.

A Pseudocode

Algorithm 1 Entity Extraction

$t \leftarrow \text{normalize}(text)$
$tokens \leftarrow \text{tokenize}(t)$
$players \leftarrow \text{exactOrFuzzyMatch}(tokens, playerNames)$
$leagues \leftarrow \text{phraseMatch}(t, leagueList)$
$stats \leftarrow \text{keywordMatch}(tokens, statKeywords)$
return $\{players, leagues, stats\}$

Algorithm 2 Intent Classification

$t \leftarrow \text{normalize}(text)$
if patternRuleMatches(t) **then**
  return patternIntent
end if
$scores \leftarrow \text{keywordScoring}(t)$
if $\max(scores) > \tau$ **then**
  return $\arg\max(scores)$
end if
if MLModel exists **then**
  return MLModel.predict(t)
end if
return "unknown"

Algorithm 3 Embedding Pipeline

$X \leftarrow \text{clean}(X_{\text{raw}})$
$X_{\text{scaled}} \leftarrow \text{StandardScaler}(X)$
$E \leftarrow \text{PCA}(X_{\text{scaled}})$
$C \leftarrow \text{KMeans}(E)$
return $\{E, C\}$ {embeddings and cluster labels}

Algorithm 4 Similarity Search

$v \leftarrow \text{embeddings}[id]$
$sims \leftarrow \text{cosineSimilarity}(v, \text{embeddings})$
$sims[id] \leftarrow -\infty$ {exclude self}
$top \leftarrow \text{topK}(sims, k)$
return top

Algorithm 5 Role-Aware Player Comparison

$\Delta \leftarrow z[A] - z[B]$ {per-feature deltas}
strengths $\leftarrow \{f \mid \Delta_f > 0.5\}$
weaknesses $\leftarrow \{f \mid \Delta_f < -0.5\}$
scores $\leftarrow$ aggregateByGroup(Δ)
summary $\leftarrow$ generateNarrative($A, B, scores$)

return $\{strengths, weaknesses, summary\}$

Algorithm 6 Scouting Report Generation

strengths $\leftarrow$ topPositiveZ(id)
weaknesses $\leftarrow$ topNegativeZ(id)
comps $\leftarrow$ FIND_SIMILAR(id, embeddings)
role $\leftarrow$ clusterLabel(id)
return formatReport(id, role, strengths, weaknesses, comps)

B Dataset

All player statistics were collected from the public FBref database:
https://fbref.com/en/

C Project Resources

Live Application:
https://shuklashreyas-soccer-scouting-bot-srcappapp-jewwpm.streamlit.app/
Source Code Repository:
https://github.com/shuklashreyas/Soccer-Scouting-Bot

D Result Visualizations

This appendix provides extended visual examples referenced in the main paper, including cluster maps, radar charts, comparison profiles, and similarity heatmaps.

Cluster 6: Creative Wide Playmakers / High-Volume Carriers

Cluster size: 81

	Player
47	Jarrod Bowen
94	Amad Diallo
102	Jeremy Doku
131	Cody Gakpo
134	Alejandro Garnacho
150	Jack Grealish
181	Callum Hudson-Odoi
191	Alex Iwobi
230	Mohammed Kudus
273	Bryan Mbeumo
285	Yankuba Minteh
300	Iliman Ndiaye

Figure 2: KMeans role clusters derived from standardized feature vectors.

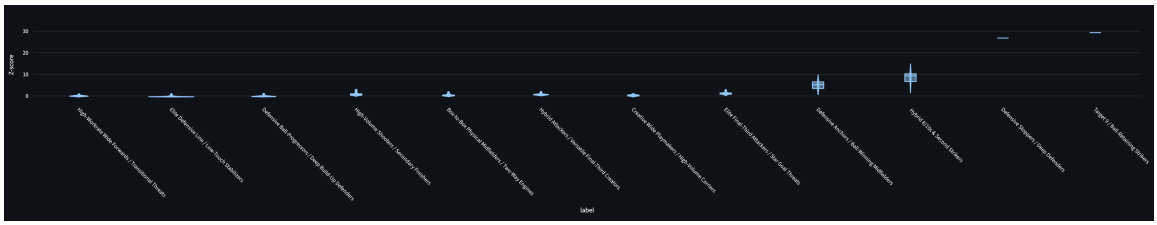

Figure 3: Skill distribution used for player feature comparisons.

Figure 4: League-wide z-score normalization across player statistics.

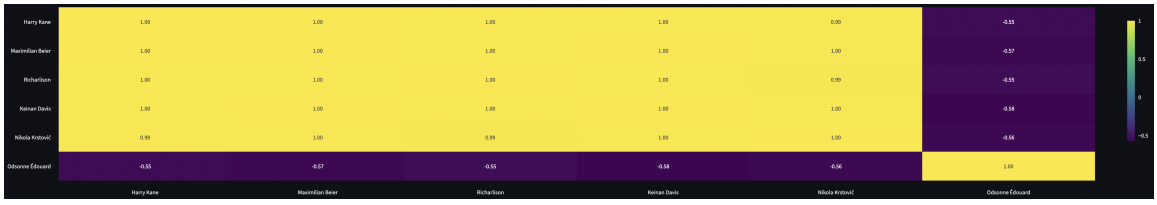

Figure 5: Similarity heatmap illustrating pairwise cosine similarity across sample players.

Figure 6: Example scouting profile generated for Erling Haaland.