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 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.

2 System Overview

Our system is a conversational scouting assistant that provides:

- **Player profiles** (e.g., "Tell me about Nico Williams"), including key statistics, role labels, and radar visualizations.
- **Player comparisons** (e.g., "Compare Haaland and Mbappé"), using role-aware z-score analysis and structured narrative summaries.
- **Similarity search** (e.g., "Find me a left-back like Alphonso Davies"), powered by a PCA–KMeans embedding model and cosine similarity.
- **Role clustering** through a frozen, human-labelled cluster map that ensures interpretable and stable player roles.

- **Scouting reports** combining strengths, weaknesses, comparables, and style explanations based on model outputs.
- **Explainable NLP outputs** that translate statistical insights into natural football language.

The system is deployed as a Streamlit web application, enabling rapid prototyping, interactive exploration, and an accessible user experience across all functions.

3 Data Sources and Pipeline

3.1 FBref Data Collection

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

- Shooting, passing, progression, defending, and expected-goals metrics (xG/xAG)
- Positional and squad metadata needed for clustering and comparison
- Multiple top-flight men's leagues to support cross-league similarity analysis

All data is obtained through a controlled scraping pipeline with caching and reproducible extraction scripts. Processed records are consolidated into a unified CSV (`all_leagues_clean.csv`), which serves as the input to feature engineering, embedding construction, and downstream modeling.

3.2 Feature Engineering

For each player-season entry, we derive a structured feature vector suitable for clustering, comparison, and similarity search. The pipeline applies:

- Per-90 normalisation to standardise contributions across different playing times
- Z-score scaling across all players to ensure comparable feature magnitudes
- Position-aware feature grouping so wingers, midfielders, fullbacks, and strikers are evaluated on role-relevant metrics
- Final numerical vectors used for the embedding model, KMeans role clustering, and downstream similarity and scouting analyses

4 NLP Components

4.1 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]`

4.2 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.

5 Similarity Search and League-Fit Modeling

5.1 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.

5.2 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.

6 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

7 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

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

9 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.

10 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.

11 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. Player similarity and comparisons also depend on engineered features and clustering choices, and results may shift if statistical distributions change or if certain players have limited minutes.

12 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.

13 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

A Dataset

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

B Project Resources

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

C 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 1: KMeans role clusters derived from standardized feature vectors.

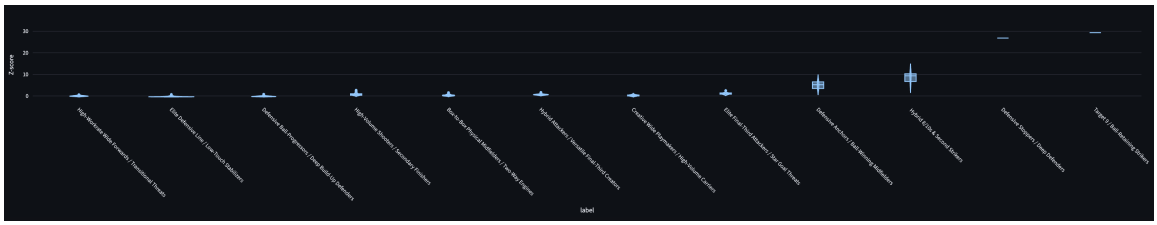

Figure 2: Skill distribution used for player feature comparisons.

Figure 3: League-wide z-score normalization across player statistics.

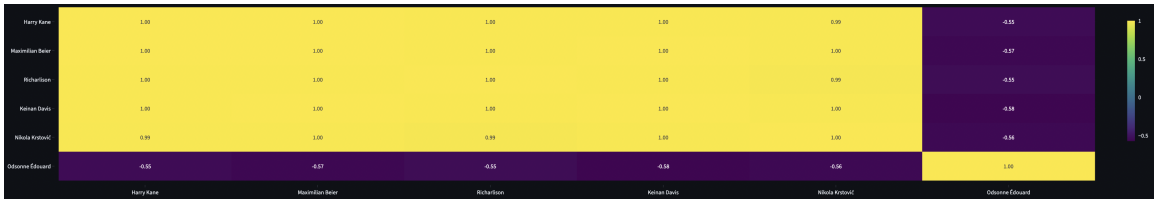

Figure 4: Similarity heatmap illustrating pairwise cosine similarity across sample players.

Figure 5: Example scouting profile generated for Erling Haaland.