Universität Konstanz

Forecasting the UEFA EURO 2024: Comparing FIFA23 and LLM Simulations.

Social Media Data Analysis

Jannik Wirtheim

21.08.2024

GitHub Repository:

https://github.com/wirthy21/SMDA-Project.git

Motivation

The UEFA European Championship 2024 (EURO24) has set new benchmarks and captivated fans like never before. Hosted in Germany, the tournament achieved an unprecedented attendance milestone, drawing approximately 2.67 million fans to 51 matches across ten stadiums, resulting in an estimated ticket revenue of 300 million euro (Sim, 2024). This highlights the immense dedication of fans and the prominent status of football in Europe. But the excitement isn't limited to the pitch, about a quarter of Germans participate in at least one betting game (Statista, 2024). To enhance their predictions, bettors utilize a variety of data sources. While many bettors depend on their intuition, others trust the odds provided by major betting banks, which are derived from robust statistical models. Additionally, some utilize data from popular video games like FIFA Manager, which European football clubs have also used to scout new talents (The Guardian, 2015). Since 2022, many bettors have also started incorporating predictions from large language models (LLMs) (Bamber, 2024). Unlike the balanced statistical models of betting odds, the outputs from LLMs are a black-box, making it difficult to understand how predictions are generated. But are these results reliable? This study investigates this question by simulating the EURO24 using data from the video game FIFA23 and predictions from state-of-the-art LLMs, then comparing the models' results with the actual tournament data. The following research questions were derived from this:

How good could the data from the video game FIFA23 forecast the UEFA European Championship 2024 results?

How good does' predictions of LLMs perform in the same task? Does using newer or more up-to-date models lead to better forecasts compared to older models?

Data Retrieval & Processing

To obtain current statistics about the EURO24, various values were scraped from the football website transfermarkt.de (Transfermarkt, 2024). Transfermarkt is a portal that provides up-to-date statistics on professional and amateur football leagues, including current market values or players' transfer history, among other things. This allows current data on the participating nations, such as the total market value or the average age, to be combined with the data from FIFA23.

Kaggle offers a data set consisting of the data of the video game series FIFA15 to FIFA23 (Leone, 2022). In addition to player statistics, the data set also consists of a variety of team-based metrics divided into women's and men's football. Only the men's team data set is relevant for the simulation. As each team is included in the data several times due to the different versions, only the latest team statistics were used. Since not all participating teams of the EURO24 are included in the data set or are updated every year, the data had to be pre-processed. For example, the data for Slovenia and Switzerland from the FIFA22 game was the most recent and was used accordingly. The nations of Turkey, Serbia, Georgia, Slovakia and Albania are not included in the data set. Performance scores were determined separately for these nations. To obtain a single Performance Score that reflects the strength of the team, various numerical metrics from the dataset were considered. In addition to a general overall value of the team, there are key figures for attack, defence, midfield, free kicks, corners, and penalties. These metrics were Min-Max normalized for comparability and adjusted relative to the number of metrics. The resulting "Performance Score" thus reflects the football strength of a team with values between 0 and 1. Additionally, to obtain the scores for countries not included in the FIFA23 dataset, the average performance score of all nations was adjusted relative to the total market value of the other countries, and a suitable score was assigned based on the country-specific market values.

To simulate the tournament using predicted values from various LLMs, prompt engineering was applied with three of the most popular models. OpenAI's ChatGPT-4 model (released on March 14, 2023) was one of the models used. With a context window of up to 128.000 tokens, 96 transformer layers, and approximately 175 billion parameters, it was the most powerful model used at the time (OpenAI, 2024). ChatGPT was trained on diverse datasets, including books, websites, articles, and other forms of written content across various topics. As a competitor, Google's Gemini 1.0 (launched on December 6, 2023) was employed (Google, 2023). It was trained on a dataset that includes web documents, books, code, images, audio, and video data (Gemini Team, 2024). Unfortunately, detailed model statistics about its characteristics were not documented in Google's research paper. Other reported parameters vary significantly and often pertain to the newest models. Finally, Llama38B (released on April 18, 2023) was used, featuring a total of 8.03 billion parameters, a context length of 8.000 tokens, and trained on over 15 trillion tokens of data from publicly available sources (Hugging Face, 2024). Obtaining the performance scores with prompt engineering presented a challenge since LLMs typically have limited predictive capabilities in gambling contexts. The models only generated the necessary scores when the task was framed as a scientific experiment rather than a game of chance. The methods by which these LLMs determine their results are mostly unclear. With ChatGPT, it was observed that the model accessed the official FIFA website during prompt execution. ChatGPT and Gemini utilized their online applications, while for Llama3, the Hugging Face API was used. Llama3 was presented as a Soccer Analyst Bot to generate values within the required range. This approach encountered issues, as the model never outputted all team values correctly in a single prompt. Consequently, data was compiled from multiple prompts. With both the LLMs and FIFA23 performance scores determined, the next step was to simulate the tournament.

Simulation

The schedule for international football competitions follows a strict plan that can be divided into the group stage and knockout stage (UEFA, 2022). During the group stages, the participating teams are divided into six groups of four teams. In this stage, each team in a group plays against each other. Wins (3 points), draws (1 point), and losses (0 points) are accordingly scored and form a table for each group. Additionally, head-to-head results and goal differences determine who advances to the knockout stage. The top two teams from each group automatically advance to the next round, with the four best third-placed teams also progressing. In the knockout stage, matches are formed based on the rankings of the teams. For example, in the round of 16, first-placed teams do not play against each other but face second and third-placed teams. The winner of a match directly advances to the next round. If a match is tied after regular time, an additional 30 minutes of extra time is played. If the score remains tied after extra time, a penalty shootout takes place. This knockout style remains consistent through all final stages. These rules have been implemented to ensure a consistent progression of the tournament. A function has been defined to simulate the group matches as well as the knockout rounds. The performance scores of the teams meeting each other form the basis for the strength of each team. These scores are used as weights in the distribution of goals, which are randomly assigned to the teams between 0 and 5. In the background, a table is generated for each group, listing points, goals, goals against, and the goal difference. Based on the group results, the knockout stage matches are determined. The simulation of the knockout stage also includes extra time and possible penalty shootouts. The winner directly advances to the next round. Both functions are combined into one to simulate the entire tournament multiple times. Each time a nation advances to the next round, a counter increases. The results of the final tournament simulations (n = 10.000) are subsequently presented for the performance scores from FIFA23. The decision for the number of simulations was made based on a trade-off between available computer resources and runtime. The results of the LLMs are shown in Appendix A. The values represent the probabilities of a team reaching the respective stage. The results are sorted descending by the team with the highest probability of winning the tournament.

Country	Round of 16	Quarter Final	Semi Final	Final	Winner	Label
Spain	92.89	59.26	36.20	22.91	13.51	Winner
England	90.54	57.82	34.07	21.12	12.58	Final
Italy	86.41	50.54	28.21	16.00	8.65	QF
Germany	83.51	48.74	26.59	15.36	8.57	QF
France	78.59	47.34	26.95	14.32	7.59	SF
Czech Rep.	87.66	49.46	27.08	13.25	6.79	SF
Netherlands	75.65	43.62	24.37	12.40	6.43	QF
Belgium	81.54	42.71	22.65	10.72	5.08	QF
Ukraine	80.25	41.37	21.51	10.04	4.80	Ro16
Hungary	69.01	33.45	15.76	7.62	3.42	Ro16
Denmark	69.10	32.35	13.85	6.31	2.88	Ro16
Romania	72.96	34.06	16.32	6.87	2.81	Ro16
Croatia	68.94	31.72	14.02	6.26	2.69	Ro16
Poland	59.15	28.59	13.13	5.35	2.40	Group
Scotland	61.34	27.27	12.00	5.50	2.32	Ro16
Türkiye	68.10	28.54	12.00	4.93	1.94	Ro16
Austria	57.44	26.94	11.72	4.73	1.87	Group
Switzerland	60.06	26.37	10.69	4.82	1.82	Group
Portugal	64.55	25.26	10.66	3.97	1.43	Ro16
Slovenia	54.38	21.08	7.95	3.09	1.07	Group
Serbia	52.13	19.51	7.33	2.78	0.93	Group
Georgia	39.17	11.51	3.56	0.77	0.19	Group
Slovakia	29.27	8.39	2.27	0.60	0.14	Group
Albania	17.36	4.10	1.11	0.28	0.09	Group

Table 1: Predicted Tournament Outcomes for Teams based on FIFA23 data (SF: Semi Final, QF: Quarter Final, Ro16: Round of 16)

Analysis

To compare the results obtained from the models with the actual results of the EURO24, the outcomes need to be labelled. Nations will be classified using a function into the classes: Group, Round of 16, Quarter Final, Semi Final, Final, and Winner. This classification is based on the probabilities of reaching the next stage. For example, Albania in Group B has the lowest probability within the group of reaching the Round of 16. Therefore, the function automatically assigns the label "Group" to each team that finishes last in its group. Similar to determining the best third-placed teams, this ranking is also performed within the function, where the two lowest-ranked third-placed teams are also labelled as "Group." The remaining teams are assigned to the Round of 16 based on their ranking. For each match-up, a comparison is made to determine which of the two teams has a higher probability of reaching the Quarter Finals. The team with the lower value receives the label "Round of 16." This system is maintained throughout the rest of the tournament. The winner of the final is labelled "Winner." Based on the assigned labels, various classification metrics can be calculated by comparing these predicted labels with the true labels. For this purpose, Spearman's correlation, permutation tests, as well as precision, recall, accuracy, and the confusion matrix will be used. Spearman's correlation is particularly useful for non-parametric data and measures the strength and direction of the monotonic relationship between two ranked variables (Wiśniewski, 2022). It assesses how well the relationship between two variables can be described using a monotonic function, by converting the data to ranks and calculating the Pearson correlation coefficient between these ranks. A permutation test assesses the significance of an observed effect by comparing it to the distribution of effects generated by randomly shuffling the data (Welch, 1990). It involves repeatedly rearranging the data labels and calculating the test statistic for each permutation to create a distribution of the statistic under the null hypothesis. The p-value is then determined by the proportion of permuted test statistics that are as extreme as or more extreme than the observed test statistic, providing a measure of how likely the observed effect is due to chance. The weighted average was deliberately used as the basis for the calculation, as classes are highly unbalanced and were therefore adjusted to the group size. Precision measures the percentage of true positive predictions out of all positive predictions, reflecting the accuracy of the model's positive predictions (Buckland & Gey, 1994). Recall indicates the percentage of true positive cases correctly identified by the model, demonstrating its ability to detect all actual positive instances. Accuracy represents the proportion of all correct predictions, including both true positives and true negatives, out of the total predictions, serving as a comprehensive performance metric. A confusion matrix (Appendix B) is a table that evaluates a classification model's performance by displaying true positives, true negatives, false positives, and false negatives (Scikit-learn developers, n.d.).

Results

The FIFA23 data shows a weak positive correlation in predicting EURO24 outcomes, with a Spearman's correlation coefficient of 0.282. However, this is not statistically significant (p-value: 0.204). The permutation test results reveal statistically significant precision (p-value: 0.0039) and recall (p-value: 0.0103), while accuracy (p-value: 0.0592) is marginally significant. These findings suggest that the model provides better predictive insights compared to random labels. The precision, recall, and F1-scores vary across different stages, with perfect scores for the later stages (Semi-Final, Final, Winner) but poor performance in earlier stages (Group, Round of 16, Quarter-Final), resulting in an overall accuracy of 41%. This indicates that while FIFA23 data offers valuable predictive insights, especially in terms of precision and recall for later stages, its overall accuracy in predicting EURO24 outcomes remains somewhat limited.

Label	**Precision**	**Recall**	**F1**
Group	0.38	0.43	0.40
Round of 16	0.29	0.25	0.27
Quarter Final	0.33	0.33	0.33
Semi Final	1.00	1.00	1.00
Final	0.00	0.00	0.00
Winner	0.00	0.00	0.00
accuracy			0.36
macro avg.	0.33	0.34	0.33
weighted avg.	0.36	0.36	0.36

Table 2: Classification Report of FIFA23 results

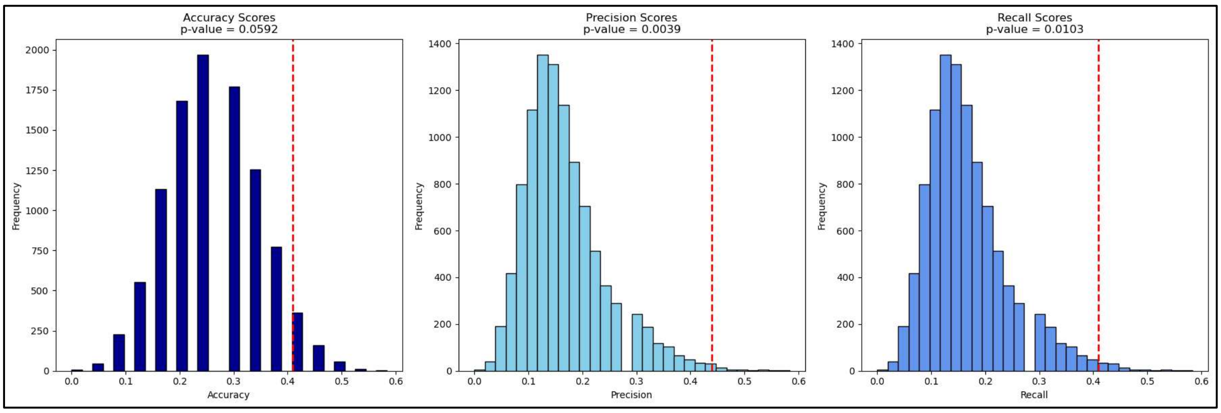


Illustration 1: FIFA23 results of permutation tests

The ChatGPT predictions for EURO24 show a negative Spearman's correlation coefficient of -0.079, indicating no meaningful correlation with actual outcomes (p-value: 0.728). The permutation test results suggest that while the precision and recall are statistically significant (both p-value: 0.0465), the accuracy is not (p-value: 0.2642). Performance metrics across different stages highlight a notable drop in accuracy (32%) compared to FIFA23 data, with especially low scores for the later stages (Final and Winner), suggesting limited predictive power overall.

Label	Precision	Recall	F1
Group	0.38	0.43	0.40
Round of 16	0.33	0.25	0.29
Quarter Final	0.25	0.33	0.29
Semi Final	0.50	0.50	0.50
Final	0.00	0.00	0.00
Winner	0.00	0.00	0.00
accuracy			0.32
macro avg.	0.24	0.25	0.25
weighted avg.	0.32	0.32	0.32

Table 2: Classification Report of ChatGPT results

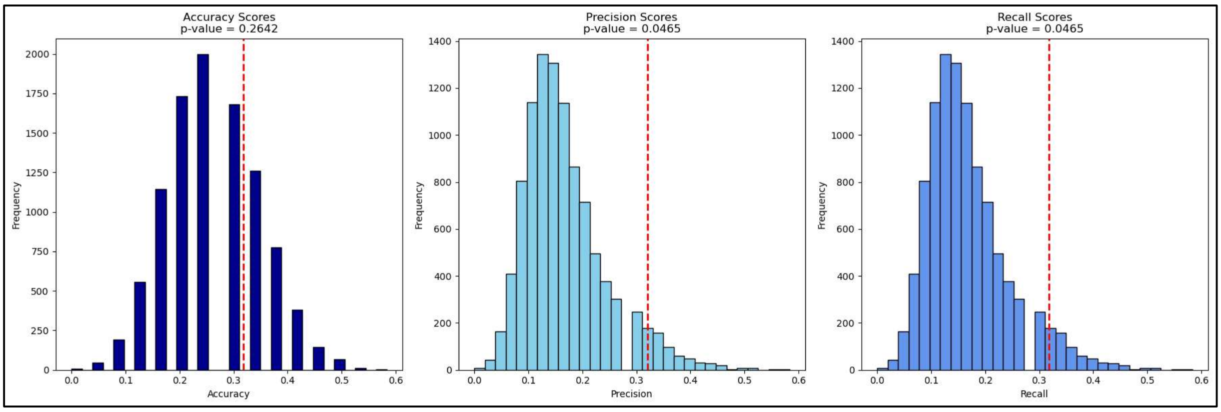


Illustration 2: ChatGPT results of permutation tests

The Gemini predictions for EURO24 demonstrate a weak positive correlation with actual outcomes, indicated by a Spearman's correlation coefficient of 0.110 (p-value: 0.625), suggesting no significant relationship. The permutation test results show significant precision and recall (both p-value: 0.0447), but the accuracy is not statistically significant (p-value: 0.2616). Like the ChatGPT results, Gemini's accuracy (32%) and predictive performance, particularly for later stages (Final and Winner), are limited, indicating low overall predictive power for EURO24 outcomes.

Label	Precision	Recall	F1
Group	0.38	0.43	0.40
Round of 16	0.29	0.25	0.27
Quarter Final	0.33	0.33	0.33
Semi Final	0.50	0.50	0.50
Final	0.00	0.00	0.00
Winner	0.00	0.00	0.00
accuracy			0.32
macro avg.	0.25	0.25	0.25
weighted avg.	0.31	0.32	0.32

Table 3: Classification Report of Gemini results

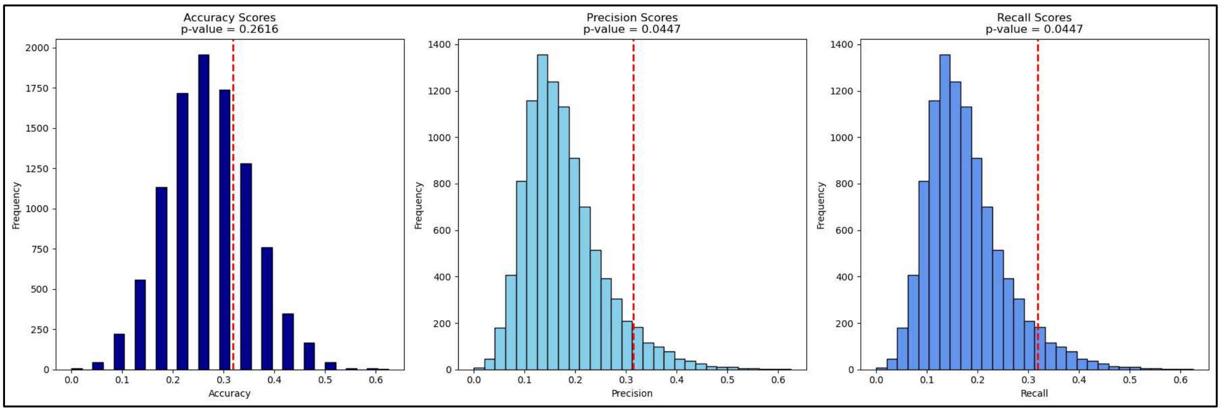


Illustration 3: Gemini results of permutation tests

Label	Precision	Recall	F1
Group	0.38	0.43	0.40
Round of 16	0.29	0.25	0.27
Quarter Final	0.33	0.33	0.33
Semi Final	1.00	1.00	1.00
Final	0.00	0.00	0.00
Winner	0.00	0.00	0.00
accuracy			0.36
macro avg.	0.33	0.34	0.33
weighted avg.	0.36	0.36	0.36

Table 4: Classification Report of Llama3 results

The Llama3 predictions for EURO24 reveal a weak negative correlation with actual outcomes, with a Spearman's correlation coefficient of 0.147 (p-value: 0.514), indicating no significant relationship. The permutation test results highlight significant precision and recall (both p-value: 0.0229), but the accuracy is not statistically significant (p-value: 0.1412). Llama3's overall accuracy (36%) is slightly better than ChatGPT and Gemini, with more reliable precision and recall, especially for early stages, though it struggles significantly with the later stages (Final and Winner), reflecting limited predictive power.

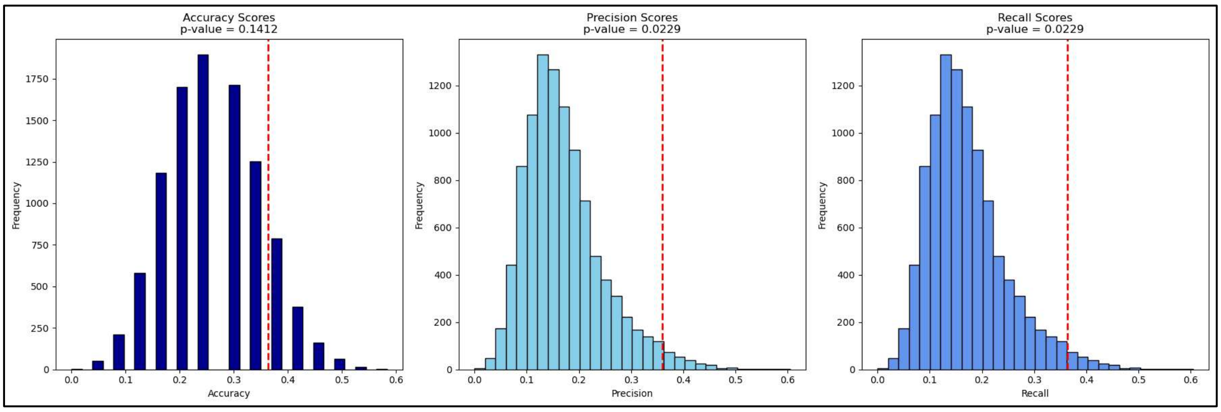


Illustration 4: Llama3 results of permutation tests

Among the four models, FIFA23 shows the highest accuracy (41%) and significant permutation test results for precision and recall, while ChatGPT, Gemini, and Llama3 demonstrate lower accuracy and limited predictive power, particularly struggling with the later stages of EURO24. Despite the appropriate results of the classification report, the Spearman correlation shows negligible values for all four models. The coefficients do not indicate a correlative relationship.

Conclusion

The analysis of FIFA23 data and various LLMs in predicting EURO24 outcomes underscores the distinct strengths and limitations of each approach. To calculate a parametric test in addition to the non-parametric permutation test, the Spearman correlation was used. The coefficients are showing that none of the models offer a strong or statistically significant correlation between their predictions and the actual outcomes. The highest correlation is seen with FIFA23, but even this is weak and not statistically significant. The results of the permutation test highlight the effectiveness of FIFA23 data in comparison to the other models. The FIFA23 model demonstrates good performance, achieving an overall accuracy of 41%, as indicated by the permutation test, which allows for comparison with random outcomes, and comparisons with LLM models. While the accuracy shows marginal significance when compared to the permutation test results, the precision and recall metrics are highly significant. Notably, the FIFA23 model outperforms the LLM models, highlighting its reliability, particularly in predicting outcomes during the knockout stages where other models tend to struggle. Given the complexity, vast data, and unpredictability of football, the FIFA23 model's performance is despite the non-significant correlation still commendable.

As with the FIFA23 models, the basic predictive strength of the LLM models was examined. In addition, the simulation was used to investigate whether newer LLMs make better predictions than older ones. As already described, all three LLM models perform worse than the FIFA23 model. When comparing the performance of the LLMs based on their recency, there is no clear indication that newer models perform better in forecasting EURO24 outcomes. Despite differences in release dates ChatGPT (March 2023), Gemini (December 2023), and Llama3 (April 2023) all models exhibited similar overall accuracy and faced challenges in predicting the later stages of the tournament. The slight variations in accuracy and performance metrics across different stages suggest that while advancements in model development may enhance certain aspects of prediction, they do not necessarily translate into superior forecasting accuracy for complex and dynamic events like a football tournament.

Overall, the FIFA23 data, with its 41% accuracy, appears to offer the most reliable predictive insights for EURO24 outcomes, particularly for the later stages of the tournament. In contrast, LLMs, regardless of their recency, demonstrate limited predictive power with lower accuracy rates. These findings imply that while LLMs are valuable tools for various applications, their effectiveness in specific forecasting tasks like predicting sports tournament outcomes remains limited. Consequently, the integration of domain-specific data (such as FIFA23 data) with advanced language models could be a more effective strategy for improving predictive accuracy in such contexts.

Critique

The FIFA23 game data was utilized for this study due to its open access availability, as data from the subsequent FC24 game was not yet accessible for the EURO24 simulation. Had the latest data been available, it could have potentially enhanced the accuracy of the results and avoided the generation of performance scores based on market values. Although the video game series offers a robust source of football-specific data, the model did not include some crucial metrics. For instance, current team statistics from qualifying rounds and friendly matches, as well as information on injuries or tactical changes, were omitted. Additionally, psychographic factors such as the number of fans in the stadium or the home-field advantage were not considered, further limiting the model's predictive power.

The use of LLMs represents a black-box approach, as it is unclear how these models generate their predictions. Originally designed for natural language processing, LLMs are not specifically tailored for predicting sporting events. Additionally, during data collection, it was observed that the models exhibit biases based on the language in which the prompt is given. For instance, Spain received the highest performance score when the prompt was in Spanish, while France scored highest with a French prompt. For this experiment, the results were presented in English. Moreover, comparing the models based solely on their "age" does not provide a comprehensive understanding of their overall architecture, as they differ significantly in parameters such as training data, number of layers and weights.

Although the statistical tests used to assess the reliability of the models are appropriate, they have several limitations. For instance, precision, recall, and accuracy are sensitive to class imbalance, which is influenced by the varying stage sizes in the study. To address this, a weighted average adjusted for class size was used to calculate the permutation tests, though the macro average yielded significantly higher results. Permutation tests also have limitations, particularly in their assumption of exchangeability. They presume that observations are exchangeable under the null hypothesis, which might not hold true in cases with complex dependencies among data points. Additionally, they are often criticized for having lower power compared to parametric tests. Due to this, the spearman correlation was calculated as an additional parametric test. The Spearman’s correlation focusses on rank but not prediction quality. The test evaluates how well the predicted ranks match the actual ranks, but it doesn't differentiate between different types of prediction errors. For example, misclassifying a "Winner" as "Group" (a large error) and misclassifying a "Semi-Finalist" as "Finalist" (a minor error) are both considered in terms of rank difference, but the impact on real-world outcomes can be very different. This is especially problematic in imbalanced datasets where minority class errors are often more severe. In imbalanced data, many instances may belong to the same majority class, leading to ties in ranks. Spearman's correlation can be sensitive to the way these ties are handled, which might lead to misleading conclusions about the model's performance. Comparing the results of the permutation tests and the Spearman’s correlation we identify a discrepancy in the reliability of the results. High precision, recall, and accuracy metrics can create a false sense of confidence in the model's overall performance. These metrics may suggest that the model is performing well, but they might not fully capture issues with ranking, especially when imbalanced classes lead to a disproportionate focus on correctly predicting majority classes.

Appendix

Appendix A: Predicted Tournament Outcomes for Teams

ChatGPT:

Country	Round of 16	Quarter Final	Semi Final	Final	Winner	Label
France	85.82	56.93	36.61	21.94	13.18	Winner
Spain	85.31	53.15	31.34	19.00	10.78	Final
England	86.79	53.38	30.53	17.76	10.30	SF
Portugal	84.90	48.85	28.98	15.98	8.69	Ro16
Germany	84.59	48.26	27.23	15.16	8.21	QF
Netherlands	74.59	43.79	24.46	12.74	6.70	SF
Italy	78.83	44.80	24.14	12.94	6.57	QF
Belgium	82.39	43.89	22.27	10.77	5.08	QF
Croatia	69.47	35.76	18.03	8.83	4.11	Ro16
Switzerland	74.06	36.59	18.40	8.82	3.99	Group
Türkiye	69.41	32.33	15.14	6.81	2.99	Ro16
Denmark	67.40	32.31	14.00	6.37	2.60	Ro16
Serbia	61.92	27.80	11.66	5.47	2.22	Group
Austria	53.41	25.61	11.62	4.70	1.97	Group
Ukraine	65.88	27.41	11.56	4.67	1.90	Ro16
Poland	53.42	25.55	11.38	4.68	1.77	Group
Scotland	58.96	24.51	10.10	4.07	1.70	QF
Czech Rep.	59.63	23.54	9.46	3.68	1.59	Ro16
Romania	61.04	24.35	9.18	3.35	1.33	Ro16
Hungary	52.57	21.14	8.08	3.11	1.16	Ro16
Slovenia	50.61	20.19	7.39	2.84	1.03	Group
Slovakia	54.45	19.32	7.47	2.56	0.91	Group
Georgia	48.88	17.62	6.76	2.36	0.80	Group
Albania	35.67	12.92	4.21	1.39	0.42	Group

Appendix A1: Predicted Tournament Outcomes for Teams based on ChatGPT data (SF: Semi Final, QF: Quarter Final, Ro16: Round of 16)

Gemini:

Country	Round of 16	Quarter Final	Semi Final	Final	Winner	Label
France	84.03	53.53	32.22	18.90	11.30	Winner
Germany	87.30	50.66	28.91	16.95	9.63	QF
Belgium	87.78	52.06	29.78	17.06	9.48	QF
Portugal	87.46	50.95	29.56	16.56	9.33	Final
Spain	81.51	49.12	28.60	15.72	8.61	Ro16
England	82.19	46.23	25.17	13.70	7.20	SF
Italy	73.97	40.14	21.28	11.06	5.49	QF
Netherlands	72.44	40.02	21.32	11.13	5.46	SF
Croatia	74.50	41.00	21.65	11.01	5.41	Ro16
Serbia	69.09	33.13	15.57	7.66	3.41	Group
Switzerland	72.79	34.84	16.67	7.55	3.36	Group
Denmark	70.04	34.00	15.62	7.18	3.23	Ro16
Czech Rep.	72.60	34.88	16.63	7.28	3.18	QF
Poland	61.12	30.63	15.23	6.76	3.02	Group
Ukraine	62.33	26.67	10.74	4.48	1.85	Ro16
Slovakia	62.37	26.76	11.32	4.59	1.84	Group
Austria	51.39	23.18	9.90	4.16	1.66	Group
Türkiye	63.06	25.99	10.80	4.43	1.65	Ro16
Hungary	56.54	22.60	8.71	3.17	1.17	Ro16
Romania	50.80	19.30	6.98	2.61	0.97	Ro16
Scotland	50.65	19.46	6.93	2.37	0.89	Ro16
Slovenia	47.75	17.83	6.74	2.46	0.79	Group
Albania	41.05	15.53	5.87	2.03	0.74	Group
Georgia	37.24	11.49	3.80	1.18	0.33	Group

Appendix A2: Predicted Tournament Outcomes for Teams based on Gemini data (SF: Semi Final, QF: Quarter Final, Ro16: Round of 16)

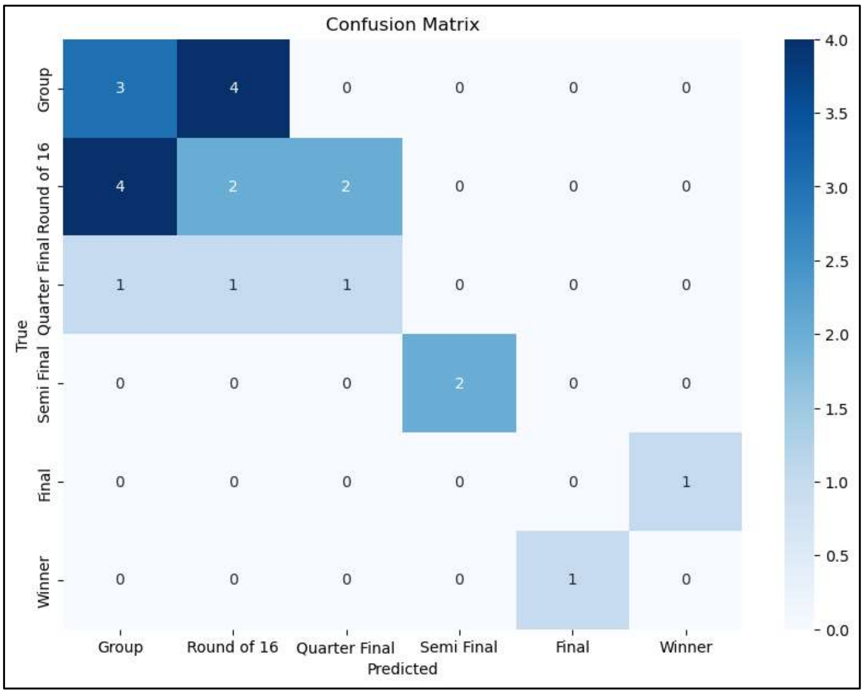
Llama3:

Country	Round of 16	Quarter Final	Semi Final	Final	Winner	Label
England	89.15	54.30	31.56	18.41	10.24	Winner
France	85.99	52.16	30.98	17.17	9.98	SF
Spain	85.93	52.95	29.74	17.08	9.73	Final
Germany	87.58	49.84	29.17	16.58	9.15	QF
Italy	83.25	49.20	27.73	15.48	8.59	QF
Belgium	84.09	49.35	28.01	15.37	8.12	QF
Portugal	84.29	48.84	27.75	14.70	8.09	Ro16
Denmark	79.98	40.97	20.79	10.68	5.26	Ro16
Netherlands	78.16	43.26	22.71	11.07	5.14	SF
Serbia	74.05	36.02	16.47	7.83	3.71	Group
Ukraine	68.55	33.10	15.49	7.26	3.24	Ro16
Croatia	66.05	32.55	14.83	6.83	2.94	Ro16
Czech Rep.	66.61	30.87	14.35	6.28	2.68	QF
Switzerland	65.20	27.31	11.66	4.94	2.03	Group
Romania	64.19	28.12	12.41	5.22	1.99	Ro16
Georgia	62.21	27.13	11.69	4.74	1.96	Group
Scotland	63.67	26.93	10.88	4.52	1.79	Ro16
Poland	56.95	24.70	10.42	4.15	1.58	Group
Slovakia	49.15	19.07	7.17	2.53	0.97	Group
Hungary	52.32	19.93	7.64	2.81	0.91	Ro16
Türkiye	50.54	18.40	6.94	2.36	0.77	Ro16
Austria	44.75	16.67	6.35	2.29	0.70	Group
Albania	33.67	11.46	3.46	1.13	0.29	Group
Slovenia	23.67	6.87	1.80	0.57	0.14	Group

Appendix A3: Predicted Tournament Outcomes for Teams based on Llama3 data (SF: Semi Final, QF: Quarter Final, Ro16: Round of 16)

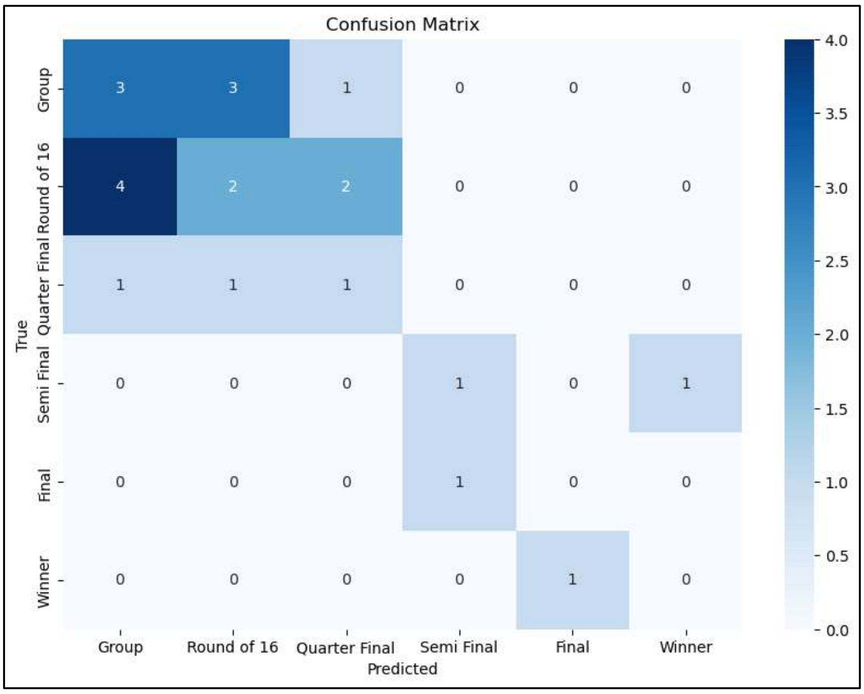
Appendix B: Predicted Tournament Outcomes for Teams

FIFA23:



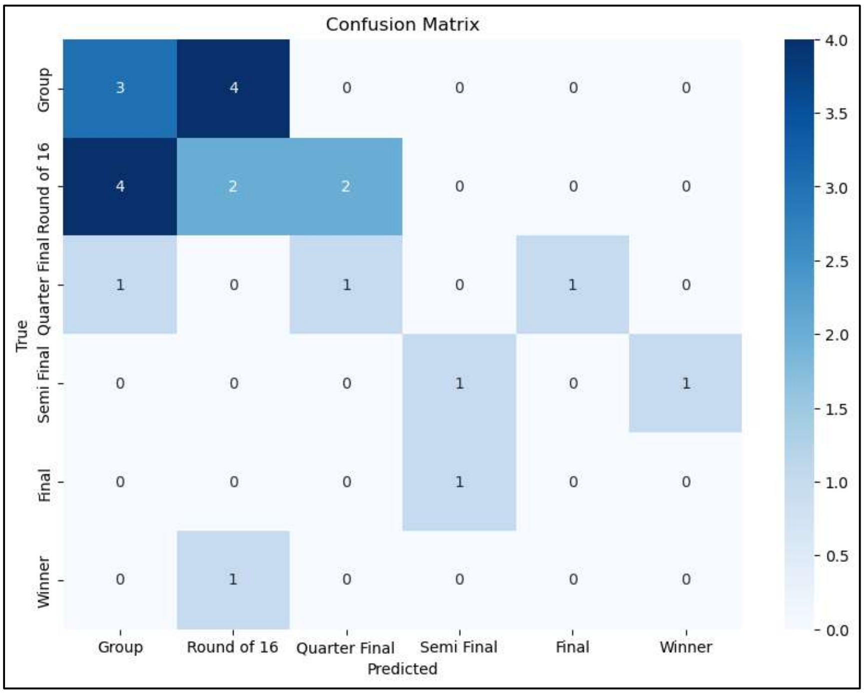
Appendix B1: Confusion Matrix of FIFA23 data

ChatGPT:



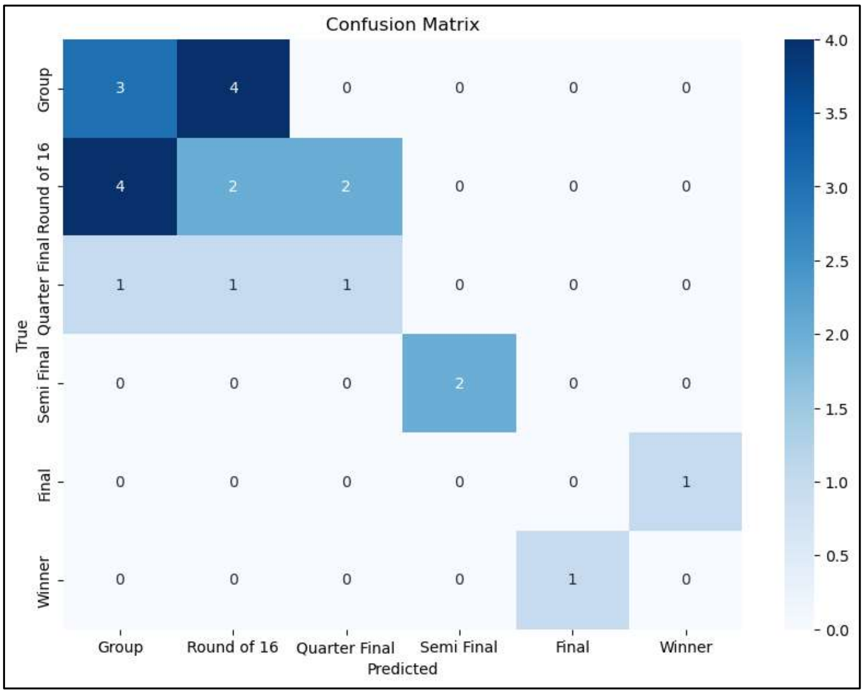
Appendix B2: Confusion Matrix of ChatGPT data

Gemini:



Appendix B3: Confusion Matrix of Gemini data

Llama3:



Appendix B4: Confusion Matrix of Llama3 data

References

Bamber, F. (2024). Gambling advice from AI: Can we trust LLMs? *The AI Journal*. https://aijourn.com/gambling-advice-from-ai-can-we-trust-llms/

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the Association for Information Science and Technology.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, et al. (2023). *Gemini: A Family of Highly Capable Multimodal Models*. Retrieved from https://doi.org/10.48550/arXiv.2312.11805.

Google. (2023, December 6). *Google Gemini AI: Availability*. Retrieved from https://blog.google/technology/ai/google-gemini-ai/#availability

Hugging Face (2024). *Meta Llama 3 8B*. Hugging Face. Retrieved from https://huggingface.co/meta-llama/Meta-Llama-3-8B

Leone, S. (2022). *FIFA 23 Complete Player Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/stefanoleone992/fifa-23-complete-player-dataset?resource=download&select=male_teams.csv

OpenAI. (2024, August 8). *GPT-4*. OpenAI. Retrieved from https://openai.com/index/gpt-4/

Scikit-learn developers. (n.d.). *Model evaluation: Classification report*. Scikit-learn. Retrieved from https://scikit-learn.org/stable/modules/model_evaluation.html#classification-report

Sim, J. (2024, July 18). *Euro 2024 in numbers: Attendance records, viewership peaks and a boost in alcohol-free beer sales*. SportsPro Media. Retrieved from https://www.sportspromedia.com/insights/analysis/euro-2024-stats-tv-ratings-viewership-attendance-social-media-sales/

Statista. (2024). *Fußball EM: Teilnahme an Tippspielen*. Statista. Retrieved from https://de.statista.com/statistik/daten/studie/1475312/umfrage/fussball-em-teilnahme-an-tippspielen/

The Guardian. (2015). *Why clubs use Football Manager as a scouting tool*. The Guardian. Retrieved from https://www.theguardian.com/technology/2014/aug/12/why-clubs-football-manager-scouting-tool

Transfermarkt. (2024). *Europameisterschaft 2024 Teilnehmer*. Transfermarkt. Retrieved from https://www.transfermarkt.de/europameisterschaft-2024/teilnehmer/pokalwettbewerb/EM24

Union of European Football Associations (UEFA). (2022). *Regulations of the UEFA European Football Championship 2022-24*. https://documents.uefa.com/r/Regulations-of-the-UEFA-European-Football-Championship-2022-24/Article-1-Scope-of-application-Online

Welch, W. J. (1990). Construction of Permutation Tests. Journal of the American Statistical Association, 85(411), 693–698. https://doi.org/10.1080/01621459.1990.10474929

Wiśniewski, J. (2022). The possibilities on the use of the Spearman correlation coefficient. *ER*, V(1), 151-162. Retrieved from https://www.researchgate.net/publication/362218857_THE_POSSIBILITIES_ON_THE_USE_OF_THE_SPEARMAN_CORRELATION_COEFFICIENT