

# GAMMA Fusion: A Generalized Athelete Model for Medal Analysis based on Gamma Function Summary

Predicting the medal counts table at the 2028 Olympic Games is no doubt an intriguing topic. This paper introduces a GAMMA Fusion, which is comprised of a **Core Medal Prediction Model** patched by other impact factors, in order to provide better predictions.

First, we designed our model as an **integrated macro-micro model**, based on the **Gamma Model** we built for each (athlete, event) to show the strength change over time. We used **Least Square Method** to fit them into the model, with a final average loss of **0.7557**. We also built a **Country / Sport Gamma Model** concluded from individual ones, dynamically changing over time.

Next, we constructed **Core Medal Prediction Model** based on the models above, separating countries into two layers. In the first layer, elite countries' athletes were divided into **recruits** (new athlete) and **veterans** (old athletes). For recruits, we utilized **EWMA** recruits number prediction and Country/Sport Gamma Models patching. For veterans, we adopted **Logistic Regression** for dropout rate and **Bayesian Inference** for revision. We used these results to predict preliminary 2028 model counts. We got **Prediction Intervals** by adopting **Monte-Carlo Simulation**. In the second layer, we grouped unexceptional countries and considered both random effect and stable improvement, using **ARIMA** and the Gamma Model respectively.

Then, we used **Principal Component Analysis (PCA)** model to take sports advantage into further account. First, the data-set is down-scaled through **Pearson Correlation Coefficient** and **ReLU** process, and the General Advantage is analyzed with PCA, gathering similar sports and unveiling countries' important sports. Second, Home Country Effect (HCE) is analysed with simple **regression equation fitting**, and **features (Hierarchy, the Trailing Effect)** is summarized, offering a general 16.7% boost in the number of medals.

Besides, Great Coach Effect (GCE) is brought out to complete our model. We first employed **time-series differencing method**, sorting out the possible evidence and comparing to historical fact, where coaches were **divided into world class and national greatness**. This method achieved a remarkable **65%** accuracy. Then we estimated the impact and added other factors like **outcomes and costs** to the model, refining its performance. Finally, suggestions were given to **Canada, New Zealand, Australia** to take advantage of GCE after employing our model.

Moreover, this paper extends its analysis to other factors found during the model refining and outliers analysing process. Based on data facts, we explore and explained effects including **Monopoly of Big Countries, Sudden Injuries, Geographical Differences, and Further HCE**, and finally offered some suggestions to the Olympic Committee, in order to further the **JUSTICE criterion**.

Finally, we used the patched model to predict 2028 medal distribution, giving USA, CHN, GBR, JPN, KOR to be top 5. Our model achieved 86.67% stability in maintaining classification, and have 82% similarity with historical American host Olympics data.

**Keywords:** Olympics medal prediction, Gamma distribution, Bayesian inference, EWMA, Time Series Analysis, PCA, Monte Carlo Simulation

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Assumptions and Notations</b>	<b>4</b>
2.1	Assumptions . . . . .	4
2.2	Notations . . . . .	4
<b>3</b>	<b>Data Preprocessing</b>	<b>4</b>
3.1	Data Initialization . . . . .	4
3.2	Data Overview . . . . .	5
<b>4</b>	<b>Model Basis</b>	<b>5</b>
4.1	The Athlete Gamma Model . . . . .	5
4.2	The Country / Sport Gamma Model . . . . .	6
<b>5</b>	<b>Core Medal Prediction Modal</b>	<b>7</b>
5.1	EWMA Patching Unreliable Recent Data and Predicting Recruits . . . . .	7
5.2	Bayesian For Veteran Medal Prediction . . . . .	8
5.3	Veterans Quitting Possibilities & Preliminary Prediction . . . . .	9
5.4	Predictions for unexceptional Countries . . . . .	11
<b>6</b>	<b>Events effect</b>	<b>12</b>
6.1	Data Processing . . . . .	12
6.2	Data Analysis . . . . .	13
6.3	Home Country Effect Presentation . . . . .	14
<b>7</b>	<b>Great Coach Effect</b>	<b>15</b>
7.1	Evidence Search . . . . .	15
7.2	Impact Estimation . . . . .	17
7.3	Result Application . . . . .	17
<b>8</b>	<b>Other Discovery</b>	<b>18</b>
8.1	Monopoly of the Big Countries . . . . .	19
8.2	Sudden Injuries . . . . .	19
8.3	Geographical Differences . . . . .	19
8.4	Home Country Effect for Other Countries . . . . .	20
<b>9</b>	<b>Final Prediction</b>	<b>20</b>
<b>10</b>	<b>Sensitivity Analysis</b>	<b>21</b>
<b>11</b>	<b>Conclusions</b>	<b>22</b>
11.1	Strengths . . . . .	22
11.2	Weaknesses . . . . .	22
<b>Reference</b>		<b>23</b>
<b>Appendices</b>		<b>24</b>
<b>Appendix A</b>	<b>Modeling Details</b>	<b>24</b>
<b>Appendix B</b>	<b>The derivation of Bayes' theorem.</b>	<b>25</b>

# 1 Introduction

The 2024 Paris Olympic Games saw the United States top the medal count with 126 total medals, tying with China for the most gold medals (40 each). France, the host country, ranked 5<sup>th</sup> in gold medals and 4<sup>th</sup> in total medals, while Great Britain ranked 7<sup>th</sup> in gold medals and 3<sup>rd</sup> in total medals. In addition to the medal results that have captured much attention, there are also some equally noteworthy outcomes. Countries like Albania, Cabo Verde, Dominica, and Saint Lucia won their first Olympic medals. We are interested in the probability of the occurrence of both these two cases.

We need to develop a model to predict medal counts for the 2028 Los Angeles Olympics, including prediction intervals and comparisons of medal situations before and after. In addition, we are tasked with predicting first-time medal winning countries and analysing the relationship between events and medal counts, while exploring the influence of Home Country preferences on event selection. Next, we'll assess the impact of 'great coaches' on performance and provide recommendations for investment in coaching for three countries. Finally, additional findings based on the model will be presented.

Our whole project is carried out in four steps: **Data processing**, **Model construction**, **Final prediction**, **Sensitivity analysis**. The Model construction step is the core of our project. We first created a rough and general model in Section 4 and 5. Then in Section 6, 7, and 8, we made use of the outliers from our model, explored the factors affecting the outliers, and build patches according to the different factors like Sports advantage, Great Coach Effect, etc. to refine our

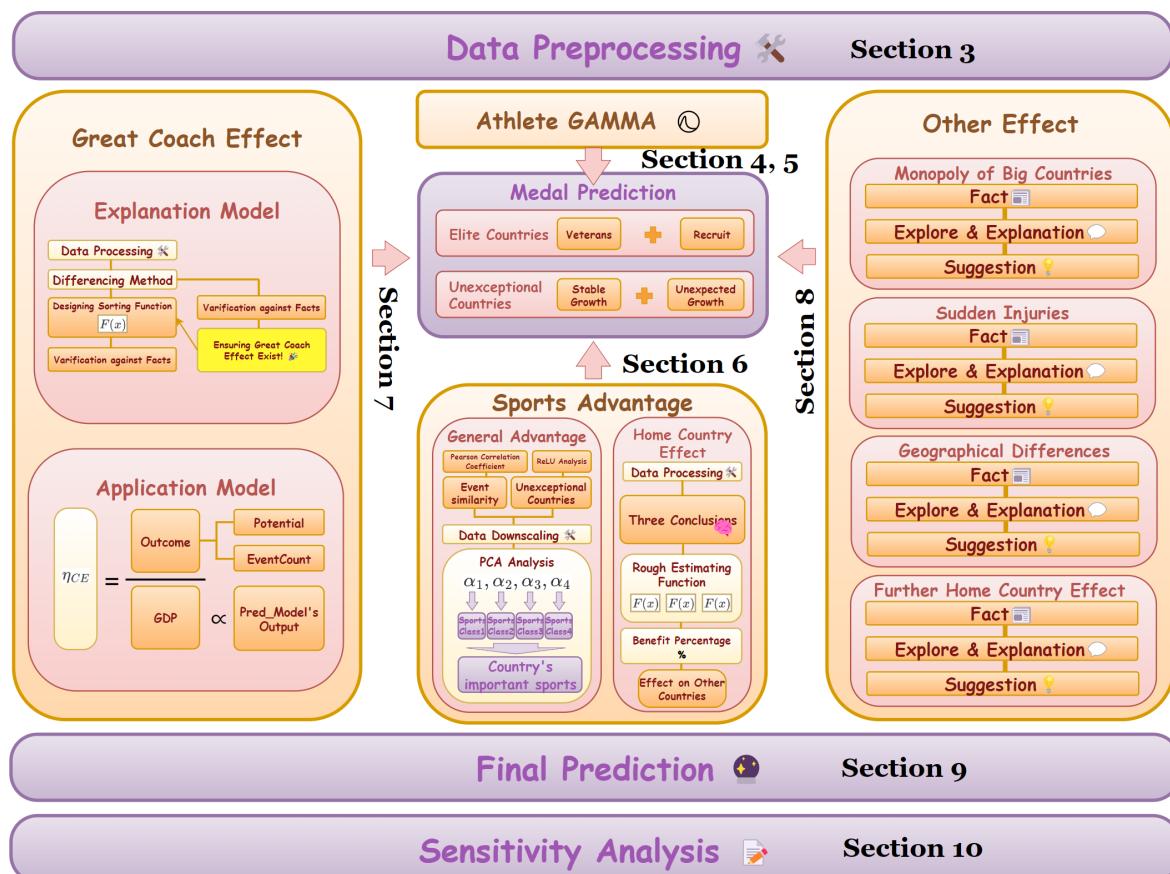


Figure 1: Overall Workflow

models. It was after we had built the model that we gave out our final prediction, examining our project's power.

## 2 Assumptions and Notations

### 2.1 Assumptions

To clarify the problem and facilitate model construction, the following assumptions are made

- **Each athlete-event pair and its intrinsic data can be treated as an independent unit.** We assume that the performance of athletes in different events does not influence each other as the external factors occurring between competitions tend to nullify the influences.
- **Medal Weight.** We have a projection from Medal to weights, showing the importance and difficulty of different medals.
- **The strength of the athletes directly affect the medals they get.** Since there is little data, it is reasonable to make such an assumption, which discards other factors causing big fluctuations.
- **Issues about dropping-one-year athlete** To simplify the model, we see athletes who attended 2020's Game, missed 2024's Game, but participated in 2028's Game as newcomers.

### 2.2 Notations

#### Symbols and Abbr. Description

$HC$	Home Country
$age$	Representing the number of times athletes participated in the Olympics.
$score$	$\omega_{Gold} \cdot x_{Gold} + \omega_{Silver} \cdot x_{Silver} + \omega_{Bronze} \cdot x_{Bronze} + \omega_{NoMedal} \cdot x_{NoMedal}$
$\alpha_i$	The $i$ _th eigenvector in descending order of eigenvalue
$n_x$	The medal number. $x$ is the name of the medal.
$\omega_x$	The medal weight. $x$ is the name of the medal.
$\lambda_x$	The coefficient before the function of $x$ . $x$ is any independent variable.

## 3 Data Preprocessing

### 3.1 Data Initialization

We based our study on the data from International Olympic Committee (IOC) 's website[1]. Firstly, we analyzed and amended the four datasets. The amendments we made are listed below.

- `summerOly_programs.csv`: We replaced all nonsense cells and empty cells with 0. We also renamed the sports in accordance with the `summerOly_athletes.csv`, incorporate similar disciplines such as Rugby Union and Rugby Sevens.

- `summerOly_athletes.csv`: We removed entirely duplicate data, replaced all invalid entries, and weighted medals based on individual and team events to facilitate subsequent calculations of the actual medal count. We removed data before 1972, considering that due to complicated reasons (such as WWII, collapse of the Soviet Union), the early data hardly have any rule. We also integrated the data from West Germany with that from East Germany.

## 3.2 Data Overview

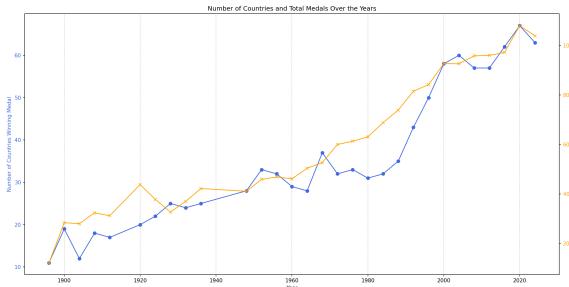


Figure 2: The line chart of the total number of medals and the number of medal-winning countries for each year.

We did a primary visualisation of the data, as is shown in the line chart. the overall number of the medals and countries winning the medal is growing on a general basis. This provided us some ideas of performing the data preprocessing stage.

## 4 Model Basis

After reviewing the data, we decided to construct an **integrated macro-micro model with a focus on the micro perspective**.

Prior to developing this model, we attempted to predict national medal counts directly using time-series models, only to find the results unreliable (see Section 4.3 for details). We concluded that relying solely on macro-level time-series models does not explicitly represent the latent variables for analysis, which led us to shift our focus to the micro-level athlete perspective. However, using this micro-level approach, athletes who have only competed once in recent Olympics are difficult to fit due to insufficient data, and predicting the performance of potential newcomers is particularly challenging. In such cases, we must return to the macro-level national perspective.

We therefore decided to divide athletes into two categories: **veterans** (those who have competed before) and **recruits** (potential newcomers). The performance of veterans is predicted mainly on the basis of their individual past performance while receiving a little revision from the the country's strengths, while the number and performance of recruits is predicted solely on the basis of the country's past performance. The detailed process of model construction is as follows.

### 4.1 The Athlete Gamma Model

We fit each athlete into a Gamma model [2] for **each event**  $\Gamma(P_{\max}, a_{\max}, \alpha, \beta)$ :

$$\Gamma(\text{age}) = P_{\max} \cdot \left( \frac{\text{age}}{a_{\max}} \right)^\alpha \cdot e^{-\beta(\text{age} - a_{\max})}$$

where  $P_{\max}$ ,  $a_{\max}$ ,  $\alpha$  and  $\beta$  are the particular parameters for each individual athlete.  $P_{\max}$  represents their peak strengths.  $a_{\max}$  implies whether an athlete is at their peak or in their dark hours when they enter the first Olympics.  $\alpha$  controls the increase rate of one's strength while  $\beta$  controls the decay rate. It's worth mentioning that the Gamma model of each athlete of a certain country can reflect a country's strength in a certain sport field and have some latent trends as the time grows which can be predicted by time series analysis. We'll dive into this later in the following section.

We obtained  $(age, \omega_{medal})$  from the dataset and utilized **Least Square Method** to fit every athlete into the model. Since the datasets for each (Athlete, Event) are extremely small, we introduced **regularization terms** to prevent overfitting, punishing model parameters (especially  $\alpha$  and  $\beta$ ) when extremely large when calculating the loss. The loss function is revised as follows:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^n (\Gamma(age) - \omega_{medal})^2 + \lambda_{\alpha} \cdot e^{\alpha} + \lambda_{\beta} \cdot e^{\beta} + \lambda_{P_{\max}} \cdot P_{\max}^2$$

For athletes with only a single data point, we assigned them an average distribution based on the medal they earned in that only year. In practice, we get the mathematical expectation for simplicity. Also, special treatment was applied to the team event data (see Appendix A for details). Fig ?? visualized the distribution of the variance, and the average loss is 0.7557, which is pretty good.

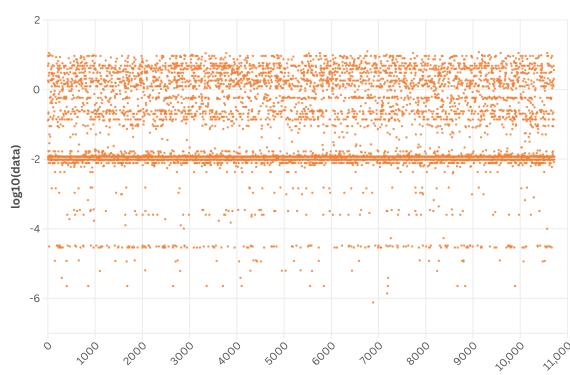


Figure 3: Gamma Loss

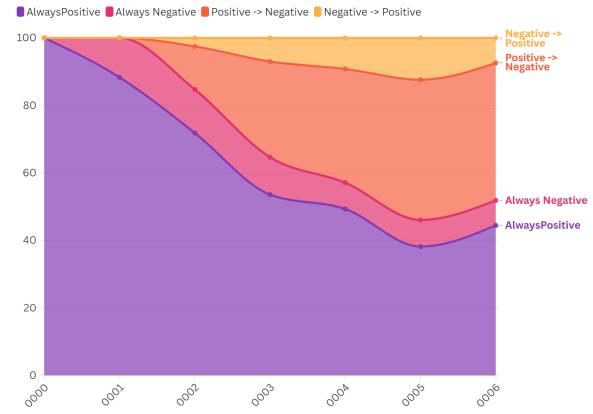


Figure 4: Change Curve

According to the data presented in Fig. 4, the quantity of athletes with a performance curve that decreases first and then increases is extremely minimal, therefore it's reasonable for us to stimulate most of the athletes with Gamma function. For example, an athlete with decreasing performance has a small  $a_{\max}$ .

## 4.2 The Country / Sport Gamma Model

We define the Gamma Model of a country for a certain sport as the **distribution** of the Gamma model (a four dimensional vector  $[P_{\max}, a_{\max}, \alpha, \beta]$ ) of the athlete that came out from this country. The distribution **changes over time**. To simplify our model, we arguably supposed that the distribution  $P(NOC, Sport, Year) \sim \mathcal{N}(\mu, \Sigma)$ , where the  $\mu$  is a four-dimensional vector and the  $\Sigma$  is the covariance matrix.

From this definition, we can easily conclude the fact that the closer the year of the Country / Sport Gamma Model is to the current year, the more **unreliable** it is. It is because we aren't sure

whether the athletes debuting in this year are actually good or bad, given that we don't know their actual performance afterwards. In section 5.2, we'll solve this trouble by adopting both Time Series Analysis and Bayesian approach.

Note that time series of the Country / Sport Gamma model can be well fit into a EWMA model[5][6]

$$S_t = \lambda_{\text{EWMA}} \cdot X_t + (1 - \lambda_{\text{EWMA}}) \cdot S_{t-1}$$

This can be validated by the fact that the probability of an acceptable prediction (average deviation of the four parameters  $< 0.5$ ) is 90.2% on the average. (We predicted 2008, 2012, 2016) It's worth mentioning that when predicting data for a certain year, we masked the data of that year and those after. The reason behind the outliers will be explored in later sections.

## 5 Core Medal Prediction Model

To predict next year's medal counts, we built extra models and made detailed analysis on the basis of our model. In general, the prediction of a country is a four-dimensional vector, containing  $\mathbb{E}[\text{'Gold'}]$ ,  $\mathbb{E}[\text{'Silver'}]$ ,  $\mathbb{E}[\text{'Bronze'}]$  and  $\mathbb{E}[\text{'No Medal'}]$ .

$$\begin{aligned} \text{Prediction} &= \mathbb{E}[\text{Veterans}] + \mathbb{E}[\text{Recruits}] \\ &= \sum_{i=1}^n x_i(p_i \cdot \text{Medal}) + \text{Medal}_c \cdot \mathbb{E}[\Delta n] + \epsilon_c + \sum \epsilon_i \end{aligned}$$

where  $\mathbb{E}[\cdot]$  is the mathematical expectation,  $\Delta n$  is the number of new athletes of this event of this country,  $\epsilon_c$  and  $\epsilon_i$  are random effects term of countries and individuals respectively,  $\text{Medal}_c$ ,  $\text{Medal}$  are **four-dimensional one-hot vectors** and the former is sampled from the Country / Sport Gamma model distribution.

We spotted that  $\epsilon_i$  can be easily neutralized (Section 5.2) if the athletes of a country are enough, but countries with lower ranks suffered a lot from this uncertainty. As a result, in order to also predict how many will earn their first medal in the 2028 Olympics, we get a **two-layer** prediction model. In the first layer, we predicted for elite countries where we treat each country separately. In the second layer, however, we put unexceptional countries together to predict '**sudden medals**' on a group level. Fig 5 presents our prediction model.

The first three subsections will dive into the first layer, while the final one will address the second layer.

### 5.1 EWMA Patching Unreliable Recent Data and Predicting Recruits

In Section 4.2, we mentioned the 'unreliable' problem. However, the Country / Sport Gamma models of recent year, especially 2024, are significantly important for predicting the quality of the recruits. As a result, we utilized **EWMA** to conduct time series analysis to **patch** the recent Country / Sport Gamma models. To be specific, we introduced an exponential function:

$$f(T) = \frac{2}{1 + e^{\lambda_{\text{patch}}(T - T_p)}} - 1$$

where  $T_p$  is the year to predict, which equals 2028 in this case. This function represents the percentage that we trusted the initial model. Then, we used **EWMA** to predict the Country / Sport Gamma Model for each year sequentially over time, and then patch the model for each year using the following formula:  $P_{\text{revised}} = f(T) \cdot P_{\text{original}} + (1 - f(T)) \cdot P_{\text{predicted}}$ . This  $\lambda_{\text{patch}}$

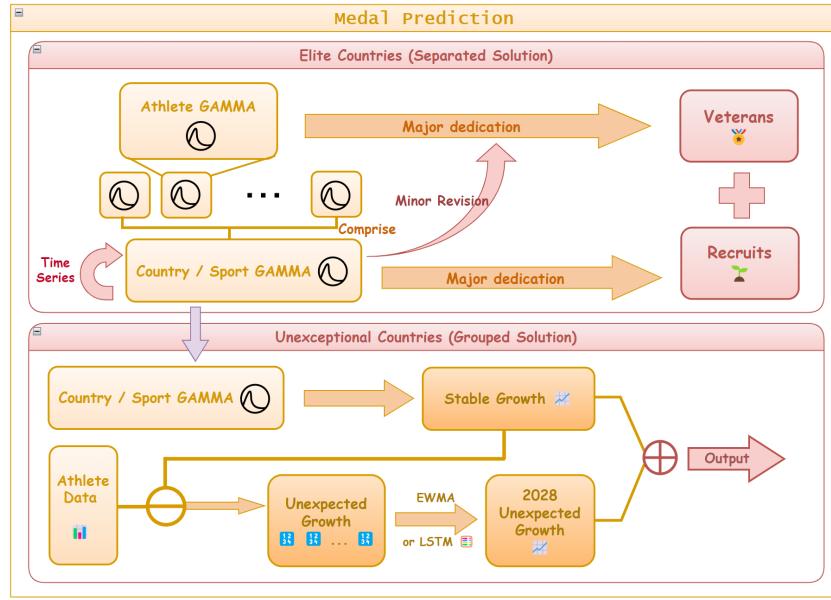


Figure 5: Prediction Model

can be learned through Linear Regression. However, to save time, we manually set four lambdas in place of  $f(T)$  for 2024, 2020, 2016 and 2012.

The patching process is illustrated in Fig 6. We'll soon make further use of the patched models in the following section. Here, we used the predicted model to get  $Medal_c$ . As for  $\Delta n$ , we introduced another EWMA model to predict. The precision of  $\Delta n$  is quite accurate. (Fig 8) Take 2016 as an example. The average of the absolute value of deviation is 0.3874, and the max one is 3.83339963. Those max ones nearly all comes from the Gymnastics team of Switzerland. Fig 7 shows this result. The detail of how we deal with the data are introduced in Appendix A. (Including treatment of the HC effect) Finally, we used **Monte-Carlo Simulation** to predict the expected medal count of recruits (we predicted Gold, Silver, Bronze separately) as well as the **Prediction Interval (PI)**. For details, please refer to Appendix A.

## 5.2 Bayesian For Veteran Medal Prediction

As for the task of predicting the medal of a veteran will win next year, we adopted a **Bayesian approach** to derive our predictions from both the past performance of a country and the athletes themselves. The logic of our approach is that the veterans each have a distribution of Gamma

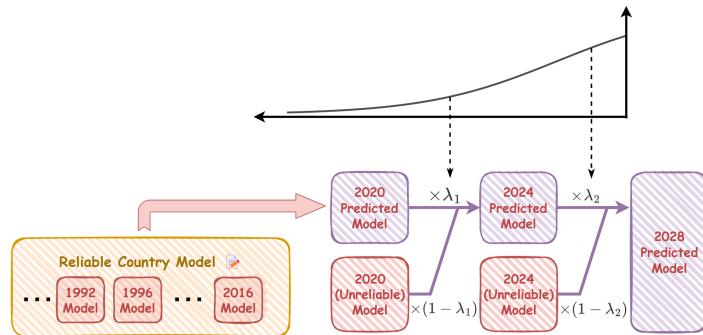


Figure 6: The Patching Process

NOC	Sport	Event	Predicted	Deviation
SUI	Gymnastics	Gymnastics Men's Individual	0.16660037	3.83339963
KOR	Gymnastics	Gymnastics Women's Horse Vault	0.3648548	3.6351452
SUI	Gymnastics	Gymnastics Men's Horse Vault	0.3648548	3.6351452
KOR	Gymnastics	Gymnastics Women's Individual	0.57800127	3.42199873
SUI	Gymnastics	Gymnastics Men's Floor Exercise	0.60285533	3.39714467
SUI	Gymnastics	Gymnastics Men's Rings	0.60285533	3.39714467
SUI	Gymnastics	Gymnastics Men's Horizontal Bar	0.76945569	3.23054431
SUI	Gymnastics	Gymnastics Men's Individual All-Around	0.76945569	3.23054431
SUI	Gymnastics	Gymnastics Men's Parallel Bars	0.76945569	3.23054431
SUI	Gymnastics	Gymnastics Men's Pommelled Horse	0.76945569	3.23054431
SUI	Gymnastics	Gymnastics Women's Balance Beam	0.82623582	3.17376418
SUI	Gymnastics	Gymnastics Women's Floor Exercise	0.82623582	3.17376418

Figure 7: Enter Caption

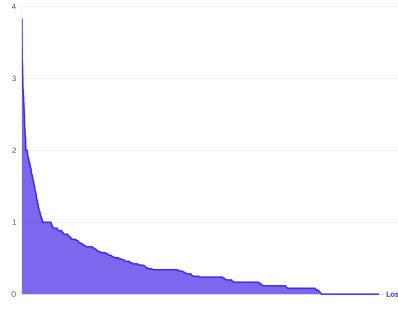


Figure 8: Added Number Loss

model (In the Veteran task, we simplify it as mathematical expectation as mentioned above), which is comprised of **prior probabilities**. We want to revise those probabilities by some known information, getting the **posterior probabilities**, which are more reliable. We let  $s$  denote the Gamma model of a certain athlete should be,  $pre$  indicate the predicted model derived from his or her past scores and  $c$  is the Country / Sport Gamma model at this time. Since  $p(c)$  is a macroscopic distribution, averaging over many athletes and being patched by previous data, we can arguably assume two things:  $pre$  and  $c$  are conditionally independent given  $s$ ,  $pre$  and  $c$  are independent in the absence of conditions. So, we can safely get the following equation: (The concrete derivation is unfolded in Appendix B.)

$$p(s | pre, c) = \frac{p(s | c) \cdot p(s | pre)}{p(s)}.$$

As a result, in order to get  $p(s | pre, c)$ , we only cared about  $p(s | c)$ ,  $p(s | pre)$  and  $p(s)$ .  $p(s | c)$  equals to  $p(c)$  according to our assumption,  $p(s | pre)$  can be easily fit using past data, and  $p(s)$  can be interpreted as the average performance of an Olympics player which can be directly derived from the whole datasets. It's also worth mentioning that this process is actually another patching process, and if the athletes come from an early age and already have a great  $p(s | pre)$ , we needn't revise it much. So we introduced a  $\lambda_{rev}$  which is determined by his or her *age*, and also follows a pattern similar to Section 5.1:  $P_{revised} = \lambda_{rev} P_{original} + (1 - \lambda_{rev}) P_{predicted}$

After the model was built, we validated its effectiveness. We selected the most suitable medal data from 2012 and 2016 to test the model, and used the differences to assess the model's accuracy. The total sum of all differences was calculated to be  $-4.510305625746792$ . The probability that the predicted medals exactly matched the actual results was 71.37%.

It's worth noting that during the model development process, we found that the HC effect mainly manifests as an increase in the number of athletes from that country and a rise in the number of dominant events. However, each individual's performance changes little. Therefore, the Country / Sport Gamma Model is barely affected by this effect.

### 5.3 Veterans Quitting Possibilities & Preliminary Prediction

Since we build a model for each athlete, it's particularly important whether a person will participate in the next competition. So, we define a probabilistic model  $p$  referring to the possibility that *a person* will attend next competition.

$$p(m_{last}, t) = \sigma(\alpha \cdot \omega_{m_{last}}) - \beta e^t)$$

where  $\sigma(\cdot)$  denotes the Log-Likelihood Function used in **Logistic Regression**,  $m_{last}$ ,  $t$  refer to the medal the person won in the previous year and the times he or she has attended the event

respectively.  $\alpha$  and  $\beta$  are the parameters that need to be learned.

We used **Gradient Descent** to solve it and the average training score is  $Score_t = 0.8369 > 0.8$ , which demonstrated a satisfactory training performance (the distribution of  $Score_t$  is shown in Fig 9).

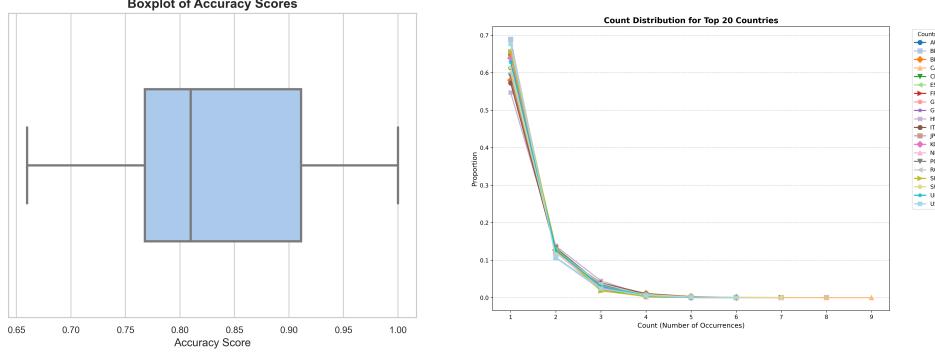


Figure 9: Distribution  $Score_t$  obtained through training.

Figure 10: Count distribution of the top 20 countries.

For further verification of this model, we've proved that the probability is not concerned with *the country* of the athletes by adopting the **Kruskal-Wallis Test**. We calculated p-value:  $p = 0.99675 > 0.05$  and it's acceptable.

For veterans, however, since we replaced distributions with mathematical expectations for simplicity, it's hard to get a PI for every individual. But the  $p_i$  discussed in this section is a probabilistic model, offering possibilities for simulations. So we applied similar **Monte-Carlo Simulation** to get the PI for veterans when predicting 2028 medal counts. So far, the **elite countries** part comes to an end, we directly added up the PI of veterans and the PI of recruits and get the following results. (In this section, we simply assumed that the sport of 2028 is the same as 2024, and we consider the change of the event amount in existing sports)

	Loss	Variance			
		Gold	Silver	Bronze	Total
Veteran	2012	12.1834	7.0192	7.6786	15.0648
	2016	9.1383	8.5047	11.5399	22.7080
All	2012	134.4594	293.8373	401.1429	602.2451
	2016	106.5606	144.4103	185.4984	361.1033

Table 1: Variance Table with Empty Top Left Cell

We particularly focused on the top two countries, United States and China. To our satisfaction, the model can accurately predict a phenomenon that United States usually have similar gold, silver, bronze counts and are all very high and China have high gold counts but relatively low silver, bronze counts. Table 2 is an example, showing the real counts and predicted counts in 2016.

Table 2: Medal Predictions and True Values

	Gold	Silver	Bronze
USA All True	46	37	38
USA All Predicted	43.45084855	39.56983619	44.90724692
China All True	26	18	26
China All Predicted	34.9829136	19.74928595	21.64751901

It's worth mentioning that we didn't choose to predict 2020 due to the fact that literally **all** no athletes from the 2016 dataset participated in the 2020 Games, and we included rough consideration of new events (based on Country / Sport Gamma model) when predicting results for years before 2028.

From the form, we delightedly spotted a good prediction result of veteran performance, but when it comes to the recruits, the losses are unacceptable. Starting from Section 6, we'll continue to patch this model to correct the accuracy.

For the 2028 medal count prediction, we listed top 10 countries' medal counts and whether a country improves or becomes worse is illustrated by **rank change** and **medals differences** in the final prediction section (i.e., Section 9). Improve Countries: USA, GBR, JPN, KOR, IRI, ESP, PRK, JAM. Worse Countries: CHN, AUS, FRA, ITA, NED.

It's worth noting that this is just the **preliminary prediction result**, because we assumed that the events of the 2028 didn't change compared with 2024 in this model. We'll conduct more detailed analysis relevant to event change in the later section and modify this table a little.

Table 3: 2028 Predicted Medal Table

	Gold	Silver	Bronze	Total
<b>USA</b>	$49.00 \pm 4.27$	$42.06 \pm 3.89$	$34.59 \pm 4.56$	$125.7 \pm 6.12$
<b>CHN</b>	$39.88 \pm 3.45$	$20.44 \pm 2.78$	$23.90 \pm 3.12$	$84.21 \pm 4.56$
<b>GBR</b>	$19.32 \pm 1.23$	$18.58 \pm 1.45$	$17.31 \pm 1.67$	$55.21 \pm 2.12$
<b>JPN</b>	$18.99 \pm 1.34$	$10.45 \pm 1.12$	$19.57 \pm 1.45$	$49.01 \pm 1.98$
<b>KOR</b>	$14.98 \pm 1.23$	$4.769 \pm 1.01$	$10.56 \pm 1.34$	$30.31 \pm 1.56$
<b>GER</b>	$12.51 \pm 1.12$	$14.26 \pm 1.23$	$16.77 \pm 1.45$	$43.54 \pm 1.89$
<b>AUS</b>	$10.21 \pm 1.01$	$6.707 \pm 1.12$	$12.94 \pm 1.23$	$29.86 \pm 1.67$
<b>FRA</b>	$8.663 \pm 1.12$	$14.05 \pm 1.23$	$14.78 \pm 1.34$	$37.49 \pm 1.78$
<b>IRI</b>	$7.983 \pm 1.01$	$0.655 \pm 0.12$	$5.685 \pm 0.23$	$14.32 \pm 0.67$
<b>ITA</b>	$7.879 \pm 1.12$	$7.740 \pm 1.01$	$12.76 \pm 1.23$	$28.38 \pm 1.56$

## 5.4 Predictions for unexceptional Countries

In this layer, we mainly focused on predicting unprecedented medals, i.e. how many countries win their first medal in a certain sport, which will be denoted by event  $A$ , and if this medal is exactly the first Olympic medal for this country (not only for a certain sport), represented by event  $B$ . Obviously,  $\mathbb{E}[B]$  is the answer for "how many will earn their first medal in the next Olympics" (which we denote as  $res$  later).

We counts the occurrence of event  $A$  and event  $B$  respectively and the results are visualized below. They already seem to have a strong relation but we need more detailed analysis. (It's worth noting that the outlier 2020 is the consequence of COVID-19: lots of events are changed.)

According to our model, event  $A$  attributes to two possible factors: one is the **stable improvement** of a country in one sport (event  $A_1$ ), which can be predicted by our previous Country / Sport Gamma Model, and the other is the **random effect** (event  $A_2$ ). Obviously,  $\mathbb{E}[A_2] = \sum(\epsilon_c + \sum \epsilon_i)$ . Intuitively, this expression is a steady time series and can be predicted. Since we can't judge the specific traits of this time series only by intuition, here we tried EWMA, ARIMA and LSTM [7]. We found that ARIMA fits the best. ( $N(A_2) = N(A) - N(A_1)$ , where  $N(A_1)$  is derived from our previous model)

Once we separate these two events, we found that the  $P(B|A_2)$  fits a time series model more possibly than  $P(B|A)$  intuitively, and this is also proven by the results, despite the fact that the improvement is minor because the dataset is too small.

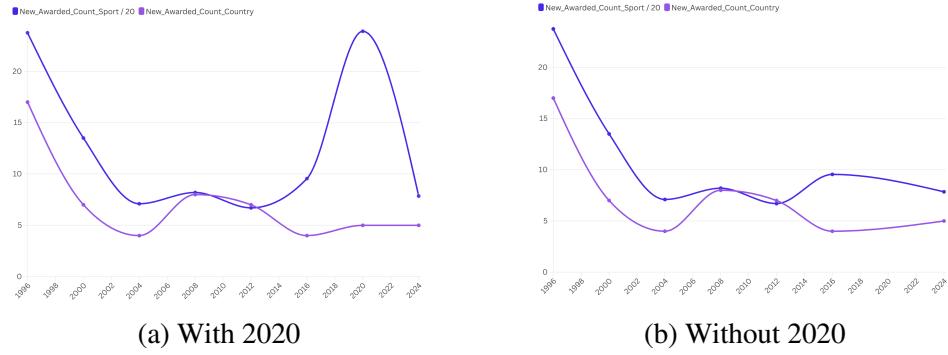


Figure 11: The occurrence of A and B.

Finally, we predicted the  $E[A_2]$  and the  $P(B|A_2)$  of 2028. The final result  $res = [\mathbb{E}[A_2] * P(B|A_2)] = [5.8452] = 6$  We are quite certain of this result as we only gets a *variance* = 1.627 when we applied the same logic to previous years.

## 6 Events effect

The type and number of specific events significantly influence a country's medal count. Indeed, as China excels in table tennis and the United States dominates in basketball, many countries have their own advantageous events, which often serve as major sources of medals. **Principal Component Analysis (PCA)** [8] is particularly suitable for calculating the correlation between events and the number of medals won by countries.

### 6.1 Data Processing

As a prerequisite for implementing PCA, dimensionality reduction is applied to the data.

**STEP 1** To examine pairwise relationships between a large number of events, all events are arranged as both rows and columns in a square matrix, and **Pearson correlation coefficients**

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

are used to assess whether there is a linear relationship between any two events.

Based on this matrix, we find that events belonging to the same sport generally have higher correlation coefficients, which provides a basis for grouping events into sport categories for analysis.

**STEP 2** We generated a table where the horizontal coordinate is the sport, the vertical coordinate is the country, and the data are the scores that each country has earned in that sport. The scores were obtained by weighting and summing the gold, silver and bronze medal awards. To determine the results of advantageous sports, we reprocessed this table by standardizing it according to sports categories using the **The z-score standardization method** as follows:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (1)$$

where  $\mu_j$  is the average score of event  $j$ , and  $\sigma_j$  is its standard deviation. For each country, the top ten events were selected and ranked based on their significance, identifying them as the country's advantageous events.

However, considering cases where the data values are too small (below -0.15), we argue that even if this event is relatively advantageous for the country compared to its other events, it still lacks absolute competitiveness and is unlikely to result in medal wins.

Therefore, we generated a heatmap Using ReLU function,(Fig. 12) in which data marked in blue indicate insufficient strength and are thus considered invalid, while only the data highlighted in orange are deemed effective. The heatmap suggests that most countries show limited actual performance and relatively low medal potential in their "top ten advantageous events." These countries have a relatively minor influence on our subsequent PCA analysis.

Here, we focus our analysis only on countries with a significant number of effective values. Certainly, our analysis remains applicable to others; however, prioritizing the development of their most proficient sports is likely the most effective strategy for those countries.

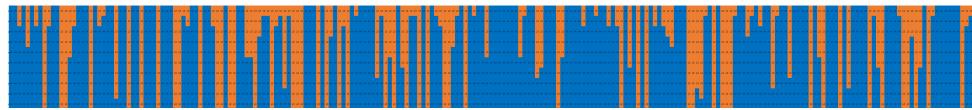


Figure 12: Heatmap of Relative Competitive Strength of Countries in Different Olympic Sports

Through the aforementioned steps, we successfully reduced the data dimensionality to a  $46 \times 74$  matrix.

## 6.2 Data Analysis

Using the z-score standardization method, the filtered ( $z_{ij}$ ) values can be used to construct the covariance matrix of the standardized country-event score matrix  $Z = (z_{ij})$ , ensuring the validity of  $\text{Cov}(X) = \frac{1}{n-1} Z^T Z$ , where  $Z$  is the standard matrix.

Next, its eigenvalues and eigenvectors are computed. The eigenvalues represent the amount of variance explained by each principal component, while the eigenvectors indicate the direction of the principal components (i.e., the linear combination coefficients). Thus, we can further decompose Equation (??): $\text{Cov}(X) = Q\Lambda Q^T$ , where  $\Lambda$  is the diagonal matrix of eigenvalues, and  $Q$  is the corresponding matrix of eigenvectors.

We sorted the eigenvalues in descending order and selected the principal components that explain the largest proportion of variance. Our results are presented in the Table 13 and further visualized in Fig. 14.

Principal Component No.	Proportion of Variance Explained (%)
1	39.78
2	10.23
3	7.27
4	6.54
5	4.88
Cumulative 1-4	63.82

Figure 13: Proportion of Variance Explained by Principal Components

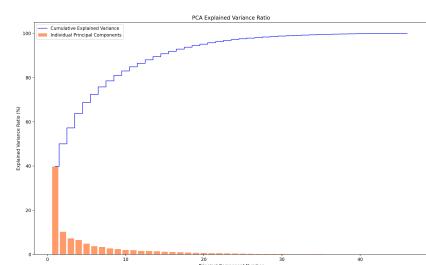


Figure 14: PCA Explained Variance Ratio

From Fig. 14 and Table 13, it can be concluded that the eigenvalues ranked fifth and beyond each account for less than 5% of the total variance, while the top four eigenvalues collectively explain 63.82% of the variance. Therefore, we focus only on the first four eigenvectors. We selected the four most influential factors in each eigenvector and queried GPT to summarize the similarities within each category (For details, please refer to the *Report on Use of AI*). Our findings are as follows:

- **Recurrence** : There are some sports like Badminton and Artistic Swimming that appears important in different eigenvectors. They even have similar importance in different eigenvectors.
- **Similarities** : For each eigenvector, all of these sports have relatively centralized similarities.
- **Differences** : At the same time the generalization of similarities exhibited by different eigenvectors varies

In Fig. 15, we list internationally significant sports, compute the dot product of their eigenvectors with the corresponding country rows, and derive the performance of different countries across the four eigenvector categories. This allows us to identify the advantageous sports for each country. The heatmap Fig. 16 indirectly illustrates the key sports for different countries through the four eigenvector-based categories, highlighting the sports through which they achieve high scores, or favorable medal distributions. Meanwhile, the relationship between the events and how many medals countries earn.

Eigenvector No.	Sports (arranged from most important to less important)				Summary (by ChatGPT)
1	Swimming	Athletics	Basketball	Equestrian	Traditional and classic competitive sports
2	Weightlifting	Badminton	Artistic Swimming	Wrestling	Sports that heavily rely on technique and skill"
3	Badminton	Table Tennis	Artistic Swimming	Cycling Track	Sports that emphasize quick reflexes and precise control
4	Skateboarding	Softball	Baseball	Beach Volleyball	Modern sports with strong entertainment value

Figure 15: PCA Explained Variance Ratio

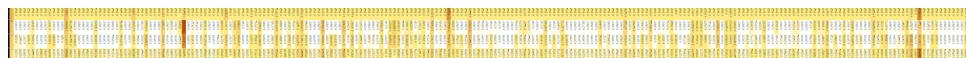


Figure 16: Heat Map of Key Sports for Different Countries

### 6.3 Home Country Effect Presentation

When examining medal trends by event, we also observe that host countries can gain a competitive advantage by applying to add more events and categories in which they excel, thereby increasing their medal potential.

Year	Country	Previous Rank	HC Rank	After Rank	ev1	ev2	ev3	ev4
2004	Greece	15	15	58	0.318747	0.698429	0.165367	0.042413
2008	China	2	1	2	0.749359	7.026537	11.21709	1.044544
2012	Britain	4	3	2	2.417473	1.008809	2.762119	0.23125
2016	Brazil	22	13	12	0.86239	-0.10289	0.211689	1.95809
2020	Japan	6	3	3	1.039236	2.557074	2.693423	7.24568
2024	France	8	5	Null	1.843487	0.734572	1.372699	0.761891

Figure 17: Home Country Effect

By analyzing the data in Fig. 17, we can draw the following conclusions:

- **Home Country Effect DO Exist:** For all host countries, there was a notable improvement in their rankings compared to the previous year, alongside a more pronounced increase in the number of medals across each category.
- **Hierarchy:** The extent of this rise is hierarchical. For countries that were originally ranked higher, the rise in national rankings was smaller. For countries with lower rankings, the magnitude is larger. By statistically processing the actual number of awards over the years, we find that for higher rankings, the difference in strength between them is generally greater, which is a reflection of the ranking density effect. Thus the extent to which they reach a rise in rankings through the home country effect is weaker.
- **The Trailing Effect:** Each country is able to maintain its Home Country Effect relatively well in the second term as home country, continuing to perform at a high level, but returning to its original level in the second term and beyond.

Based on the relevant data, we roughly derived the following computational formula:

$$[Increase] = \sum 0.4731 \times x_i$$

This implies that the value corresponding to this country in the  $i$  eigenvector equals the increase in the number of gold medals. Moreover, since the number of silver and bronze medals is dynamically conserved, under the assumption "Medal Weight" that 1 silver medal is equivalent to 2 bronze medals, the weighted medal count can be calculated as follows:

$$[Counts] = 9.2137 - 2.5402 \times x_i$$

Based on calculations, the Home Country effect overall **leads to a 16.7% increase** in the number of medals.

The analysis above indicates that the Home Country effect primarily relates to the number of advantageous events, the performance gap between athletes, and factors such as local support and adaptation to the natural environment. Given the limitations of the dataset, the current model remains relatively crude and cannot capture detailed patterns with high accuracy. We will explore these aspects further in the concluding section of the article.

More, if some of the countries have similar advantageous items as the host country (which can be simulated by our PCA method), then we can say that these countries will also be favored by the home country effect. As analyzed in the previous data, no two countries have exactly the same advantageous items, so this favor is weaker than the host itself, similar to a differential mode.

## 7 Great Coach Effect

During our earlier analysis of the model, we realized that it is not entirely complete, and certain additional factors could affect the prediction results. Next, we will delve into the influence of the "Great Coach" factor.

### 7.1 Evidence Search

To identify evidence of changes that might be attributed to the "great coach" effect, we start with the **definition of the great coach effect**: a great coach enables the same team or athletes to

achieve significant success. Given the four-year interval between Olympic Games, we assume that a coach deemed great can lead the team to substantial breakthroughs within this period, reflected in a notable improvement in medal outcomes.

Therefore, we employ a time-series **differencing method**. We first calculate a weighted sum of medal counts to derive a score, then compute the first-order difference of this score over time, and divide it by the number of participants. This process yields the following statistical data in Fig. 18, with two points worth noting:

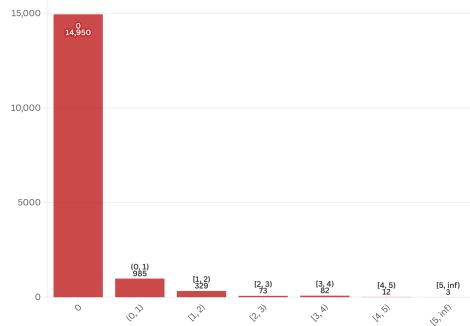


Figure 18: Absolute value of first-order differences in scores for the same country and sport

- 1) Here we used the weights of [16,4,2,1] instead of [8,4,2,1] weights in the assumptions to more prominently highlight the significance of winning gold medals, and thereby making potential evidence of great coach effects more apparent.
- 2) Based on data analysis, we find that great coaches like Lang Ping[3] and Béla Károly[4] have a similar impact on all events within a sport category. This provides a rationale for dimensionality reduction, allowing us to analyze each country's performance at the sport level.

Based on our real-life experiences, we recognize that great coaches are rare, and their presence or absence usually leads to a sustained impact. Therefore, we set a threshold of 4 for the score differencing values to filter the data. When analyzing the filtered result, we found that some countries exhibit a two-year alternating high score differences in certain sports, with the magnitude of these differences being similar or identical. Through an analysis of overall trends, we conclude that these fluctuations are likely due to transient anomalous factors in the country's performance in that sport, rather than the influence of a great coach.

After a second round of data cleaning, we compare our findings with real-world data and arrive at the following analysis map for the success rate of identifying great-coach effects (Fig. 19):

We categorise great coaches into two levels: world level and national level. World-level coaches are those who have achieved widespread success in coaching and are recognised globally, while national-level coaches are those who have significantly advanced the sport especially in their country and gained national recognition. Our method achieves a 65% accuracy, while all the world-class great coaches (actually limited in number) is well reflected in our results. This shows that the data for evidence of changes that might be due to a “great coach” effect selected by our method is sufficiently robust.

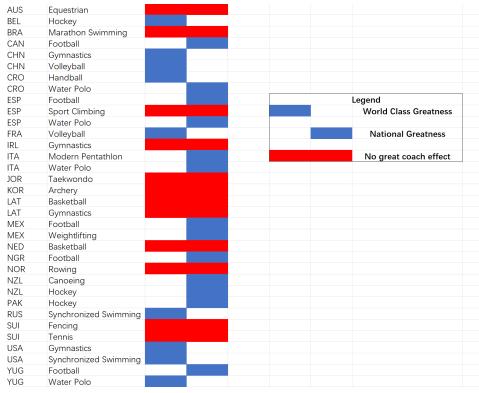


Figure 19: Great Coach Effect Judgment Success Analysis Chart

## 7.2 Impact Estimation

We calculate the extent to which such an effect contributes to medal counts and **great coach effect (GCE)** impact with this formula:

$$[Effect] = [GreatCoachIndex] \times [EventCounts] \times [ShiftSpace]$$

We present our result in Fig. 20, which indicates that:

- The impact of a great coach at the world-level great coach is highly significant.
- The impact of a great coach at the national-level is relatively less significant than that at the world-level. However, for specific countries, they have shown outstanding effects, even comparable to those of world-class great coaches.

By incorporating this GCE analysis, we improve our model, making it more precise and improving the accuracy of future medal distribution predictions.



Figure 20: Great Coach Effect Impact Analysis Chart

## 7.3 Result Application

In order to select three countries and estimate the impact of investing in a great coach, we evaluate the cost-effectiveness of this investment and quantitatively analyze its components using the model established above. We will analyze the issue from both the outcome and cost perspectives.

- **Outcomes.** According to the model above, the outcome can be divided into two components: the change in the country's strength in the sport category and the number of events in that sport. While the number of events may vary by year and host country's choices, it generally follows a stable distribution and has a clear interpretation. In contrast, quantifying "changes in strength" is more complex, and we define it as follows:

$$Outcome = potential \times EventCounts$$

The potential can be estimated using the model we previously established through the Country / Sport Gamma function, the size of the ratio between the number of awards and the number of participants will also be one of the considerations for potentials. When we observe that the parameters of the Gamma curves for most events in two countries are similar, we can reasonably conclude that the overall levels of the two countries are comparable. Consequently, these few events may hold significant potential. On the other hand, the  $\alpha$  and  $a_{max}$  of a certain Country / Sport Gamma function can directly reflect a country's potential. For example, if a Country / Sport Gamma function has small  $\alpha$  and large  $a_{max}$ , it means that the athletes of this country of this sport have large potential and are yet to be tapped.

- **Costs.** It represents the financial cost required for the country to recruit the great coach, which we assume to have a near-linear relationship with the ratio of the GDP of the coach's home country to that of the hiring country. Additionally, based on researches (For example, Xun's [9]), GDP is strongly correlated with overall medal counts, allowing us to derive an expression for cost using the given dataset.

$$GDP \propto f(MedalCounts)$$

In summary of the two perspectives, the cost-effectiveness (denoted as Merit here) can be calculated using the following formula:

$$[Merit] = \frac{Outcome}{GDP} = \alpha \times \frac{Potential \times EventCounts}{f(MedalCounts)}$$

where  $\alpha$  is a pending coefficient,  $f(MedalCount)$  is a function related with medal counts. As for the selection of three countries to estimate, we choose:

- **Canada (CAN) investing in a great coach for fencing.** Based on the model simulation, investing in this program (e.g., hiring a coach from Italy) could increase the potential by 23.5%, indicating highly favorable cost-effectiveness.
- **New Zealand (NZL) investing in a great coach for swimming.** Based on the model simulation, investing in swimming (e.g., hiring a coach from the United States) could increase the potential by 5.1%, indicating moderately cost-effectiveness.
- **Australia (AUS) investing in a great coach for gymnastics.** Based on the model simulation, investing in this program (e.g., hiring a coach from Denmark) could increase the potential by 9.8%, demonstrating relatively high cost-effectiveness.

## 8 Other Discovery

In Section 4.1, when we are calculating Gamma Loss, we spotted some athletes with abnormally high loss which is up to 18.655.

In Section 4.2, we compared some year's predicted country's Gamma model and true Gamma model and looked into the reason why a few country received a considerable unexpected boost in certain sport.

## 8.1 Monopoly of the Big Countries

During the first period of data analysis, we noticed the distribution of medals is imbalanced. The top 15 countries accounted for 65.45 % of the medals and 72.56 % of the gold medals, as we can see in the figure. It is clear that the big countries often have a monopoly on winning medals and remain stable in their performance.

Therefore, the Olympic committees can focus on obtaining the number of athletes from large countries as well as largely increasing the number of athletes from small countries. This may add possibilities for the small countries to win medals from the big ones, reducing the monopoly of the big countries.

## 8.2 Sudden Injuries

One abnormal value is "Sara Kolak F, Croatia, Athletics Athletics Women's Javelin Throw Gold" [10]. We find out that she's actually an excellent javelin thrower, but her strength has been greatly affected by injuries. Therefore, sudden injuries is a large error that needs attention and correction.

Unfortunately, because of the way our model is designed and the training data we've got, we can't yet accurately predict this factor. So, at the moment, we can only guess that athletes are developing steadily. To improve this, we're going to keep a closer eye on athletes' daily training and performance in the next competition, so we can reduce the impact of sudden changes in data due to injuries and illnesses.

The Olympic Committee should also endeavour to minimise permanent injuries caused by accidents during the Olympics. In addition, the Committee should present the public with more exciting competitions by enhancing attention and protection of athletes' safety

## 8.3 Geographical Differences

When dealing with important sports in different countries in PCA, we found that for different classes of sports exhibited by the eigenvectors, the distribution of countries that specialise in them has certain geographical characteristics. For example, Europe and America are better at team ball sports (football...), and Asia is better at fast reaction and precise control sports (badminton, table tennis...).

In particular, the inclusion of a predictive model as a patch can be done by this Eq:  $\Delta n_{medal\_geo\_effect} = p_0 \times \Gamma_{geo}$  where  $p_0$  is the worldwide average performance, and  $\Gamma_{geo}$  is the geographical difference coefficient.

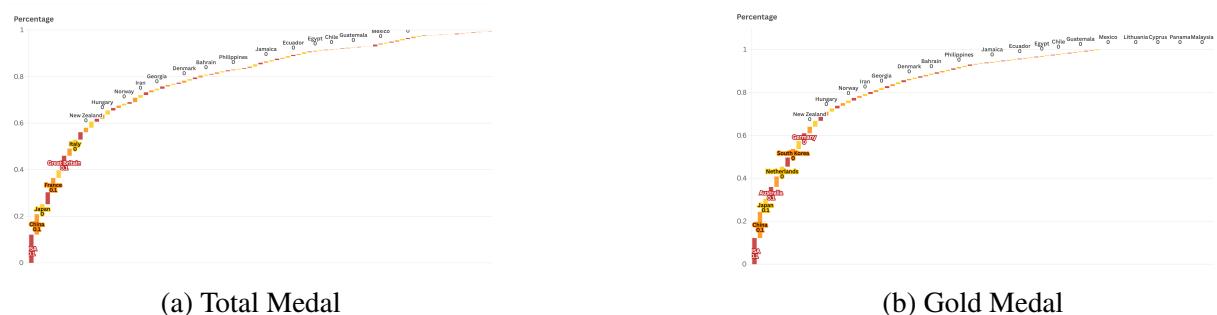


Figure 21: Country Occupation Percentage of Total Medal and Gold Medal

The total number of Olympic medals is shown in the Figure 22. We also found that most of the high number of awards are distributed in the northern hemisphere, while there is less distribution near the equator and more distribution in temperate and subarctic countries, among other geographical characteristics. It can be seen that the geographical characteristics of a country also affect the ceiling of different sports in a country and the development of native athletes.

The Olympic Committee should also make sure that there is a balance in the number of different types of events, and that the overall competition is similar for different countries. This will make sure that being excellent in some events does not have a big effect on the final result. This will make sure that the competition is fair, reasonable and exciting.

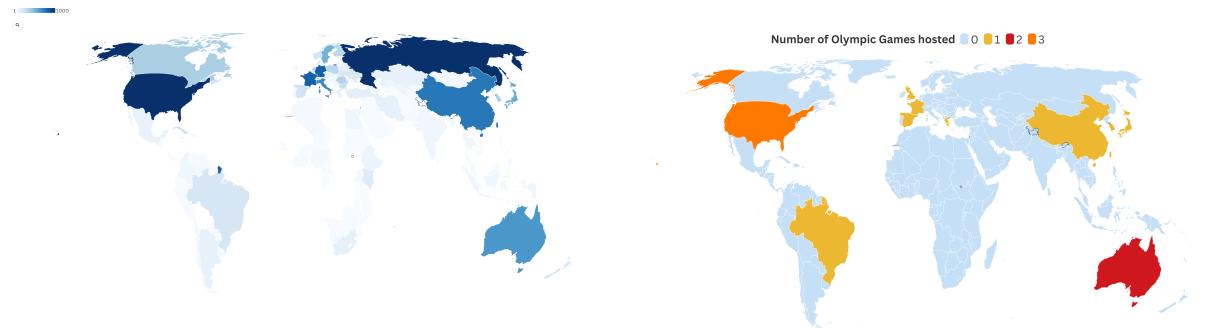


Figure 22: Heatmap of Relative Competitive Strength of Countries in Different Olympic Sports

Figure 23: Number of Olympic Games hosted

## 8.4 Home Country Effect for Other Countries

Our data shows a significant variation in the number of medals won by different countries from one session to the next. This indicates that it is largely due to the HC Effect, which impacts the host country as well as other ones. We have analysed the impact on countries with similar strong sports to the host country. The HC's unique weather, diet, sports pitch... can be extremely similar or vastly different to those in the athletes' native countries. This can lead to not adapting to the environment. Figure 23 shows the number of Olympics hosted after 1984.

The Olympic Committee should ensure that the HC are diverse and balanced in terms of factors like time zones, size, climate, so that the overall performance of each country is stable.

Furthermore, there are also factors like 2020 Japan (influenced by Covid-19), 2024 France (Russia cannot participate due to the political issue) that cannot be predicted. They are also important points that remain for the Olympic Committee to consider about. In our opinion, the first and only criteria is to remain JUSTICE and only consider the welfare of mankind.

## 9 Final Prediction

Our prediction is produced not only from our Core Medal Prediction Model, but also from the patches we designed to boost the accuracy and robustness of the model. After combining the results of both, along with the corresponding weighting and function compositing, the data shown in Figure 23 was obtained.

The first figure shows the overall predicted ranking of the countries below. It is summarized from the early years' results that three classes can be separated (ranking 1-2, 3-10, 11-15), and our prediction tend to follow this pattern.

The second figure shows the predicted Gold, Silver, Bronze, Total medal counts of the top 15th countries. The Gold and Total count is ranked in the third and fourth figure, comparing with the situations of 2024.

We asked ChatGPT what are the possible events of the new sports in 2028, and searched the countries that are capable of these sports. Moreover, we take Home Country Effect's 0.16% boost into account, and also calculated the effect for other countries. Finally, as a reference, we compared the data of 1996 and 1984, which also has the Home Country of America. We noticed that in those datasets, America gained a huge lead in the gold and total medal count, which is identical to our prediction, providing a 82% similarity.

Gold Ranking: USA, CHN, GBR, JPN, KOR, GER, AUS, FRA, IRI, ITA, CUB, NED, ESP, PRK, JAM. Total Ranking: USA, CHN, GBR, JPN, GER, FRA, AUS, KOR, ITA, NED, ESP, IRI, CUB, PRK, JAM.

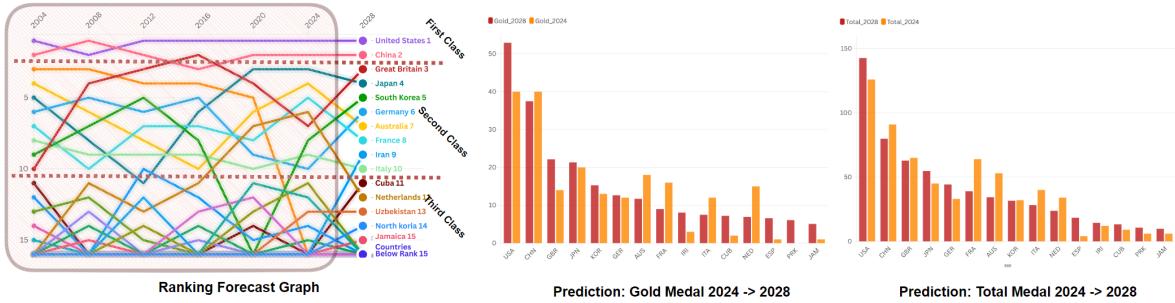


Figure 24: Our Final Prediction Graph Set

Table 4: 2028 Predicted Medal Table

NOC	Gold	Silver	Bronze	Total
<b>USA</b>	52.92	49.63	40.13	142.7
<b>CHN</b>	37.48	19.42	22.94	79.84
<b>GBR</b>	22.13	21.07	19.64	62.83
<b>JPN</b>	21.34	11.41	21.98	54.73
<b>KOR</b>	15.28	4.960	11.41	31.64
<b>GER</b>	12.63	14.40	17.28	44.31
<b>AUS</b>	11.70	7.832	14.83	34.36

<b>FRA</b>	9.009	14.61	15.37	38.99
<b>IRI</b>	8.063	0.6879	5.685	14.44
<b>ITA</b>	7.485	7.353	13.40	28.23
<b>CUB</b>	7.227	2.270	3.758	13.25
<b>NED</b>	6.938	5.510	11.32	23.77
<b>ESP</b>	6.593	4.805	7.006	18.40
<b>PRK</b>	6.071	2.072	2.572	10.72
<b>JAM</b>	5.081	3.090	1.640	9.811

## 10 Sensitivity Analysis

This sensitivity analysis aims to evaluate the impact of **hyperparameter variations** on the model's performance and identify the most influential parameters. Since our model is very complicated and is comprised of several parts. We carried out sensitivity analysis for each part of the prediction model.

We found that these following parts are robust: Sport Gamma,  $\Delta n$  prediction, veteran performance prediction. However, the recruit performance prediction is quite sensitive to parameter changes. The sensitivity of Sport Gamma to two important hyperparameters  $\lambda_\alpha$  and  $\lambda_\beta$  is illustrated by Fig 25 and the sensitivity of a small perturbation of the Country / Sport GAMMA model parameters to the 2028 gold medal prediction for USA is illustrated by Fig 26.

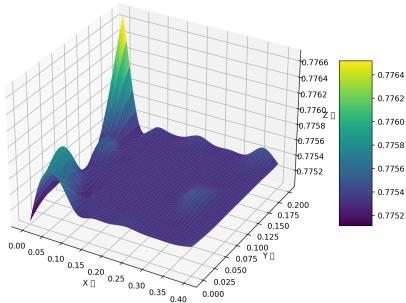


Figure 25: Sensitivity of recruit performance prediction

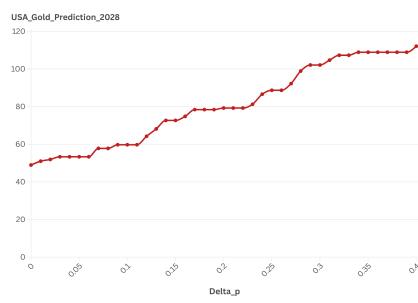


Figure 26: Sensitivity of recruit performance prediction

## 11 Conclusions

### 11.1 Strengths

Our core model is not only a macroscopic model which has low interpretability and low accuracy, it also looks into each individual, which directly influences the medal counts. So our model is more accurate (especially for the veteran part) and has strong interpretabilities, so that we can gain insights by looking into the outliers.

Besides, our model is not a single core model, but one with numerous of patches that explain and predict the large amount of seemingly 'outliers'. As is shown above, our core model itself only provides overall and rough prediction, while with the help of our various patch development, in cases like HC effect, Great Coach Effect, our comprehensive model can identify these potential effects and take them into account, refining and modifying our rough prediction, achieving a better performance.

### 11.2 Weaknesses

The weakness of our model is mostly decided by these recruits, the regularity of whose appearance is not yet perfectly understood due to the limited number of factors considered. Additionally, as the data is so limited, the model parameters we simulated were, to a great extent, determined by random data and general laws. Furthermore, the simulation result is discrete, leading to a large deviation.

According to our model, the annual number of veterans per country is largely determined by the number of veterans in previous years, which is relatively rough as the improvement in the performance of athletes from 4 years ago could motivate the increase in veterans.

In our patched models, some of the models require further data support in order to get a more precise result and prediction. For some other models, like the Great Coach Effect model, the overall system may be too simple that it cannot achieve an accuracy up to 90%+, because some other effects may also act similarly to Great Coach Effect. This needs further research to find more factors. Moreover, the factors themselves may be unpredictable, which will always bring us some outliers. But that's the meaning of prediction - everything is possible.

## References

- [1] COMAP 2025. Problem C data set.2025
- [2] Griffin, J., Hinoveanu, L. , Hopker J. (2022). Bayesian modelling of elite sporting performance with large databases. *Journal of Quantitative Analysis in Sports*, 18(4), 253-268.
- [3] Olympics.com Biography, Lang Ping, <https://olympics.com/en/athletes/ping-lang>
- [4] USA Gymnastics Hall of Fame, <https://usagym.org/halloffame/inductee/coaching-team-bela-martha-karolyi/>
- [5] Cox,D.R., Prediction by Exponentially Weighted Moving Averages and Related Methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23: 414-422 (1961).
- [6] James M. Lucas , Michael S. Saccucci , Exponentially Weighted Moving Average Control Schemes: Properties and Enhancements, *Technometrics*, 32:1, 1-12 (1990)
- [7] Box, G.E., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (1978). Time Series Analysis: Forecasting and Control. *The Statistician*, 27, 265-265.
- [8] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24, 498-520.
- [9] Bian, Xun. (2005). Predicting Olympic Medal Counts: the Effects of Economic Development on Olympic Performance. *Honors Projects*. 13.
- [10] Wikipedia.org, Sarah Kolak,[https://en.wikipedia.org/wiki/Sara\\_Kolak](https://en.wikipedia.org/wiki/Sara_Kolak)

# Appendices

## Appendix A Modeling Details

Sec.	Theme	Activity
4.1	Team event	We queried GPT to filter team events, cleaned the dataset, and added a column in the <code>summerOly_athletes.csv</code> for the actual number of participants per team event. For participant number prediction, no changes are needed, but calculating the loss requires dividing by the number of participants, which also applies to medal count predictions.
5.1	The strategy of filling missing value for Country / Sport Gamma distribution time series prediction.	When performing time series prediction for the country's Gamma distribution, for default data, in order to ensure data completeness, we didn't fill in the default value. Instead, we designed the following algorithm. (The pseudocode used for data processing is in algorithm 1.) This approach is used because the strength of the athletes produced by the country has not changed, and the missing data is due to various reasons such as specific events not being included in the competition that year. Comparing with simply setting all parameters to 0, this approach is more intuitive.
5.3	The possibilities of veterans participating in the next game.	If the $p$ for a sport is greater than the $p$ from training without distinguishing the sport, use the sport's own $p$ ; if it's smaller, use the $p$ from the training without distinguishing the sport (however, we avoid sports with limited data, such as breaking, where the $p$ is NAN).

```

default_values = {p_max ↪ 0.5, a_max ↪ 5, alpha ↪ 0.2, beta ↪ 0.05};
for i from 0 to length(merged_df) - 1 do
    if merged_df.loc[i, 'p_max'] is NaN then
        if i > 0 and merged_df.loc[i - 1, 'p_max'] ≠ default_values['p_max'] then
            merged_df.loc[i, 'p_max'] = merged_df.loc[i - 1, 'p_max'] × 0.9;
            merged_df.loc[i, 'a_max'] = merged_df.loc[i - 1, 'a_max'];
            merged_df.loc[i, 'alpha'] = merged_df.loc[i - 1, 'alpha'];
            merged_df.loc[i, 'beta'] = merged_df.loc[i - 1, 'beta'];
        end
        else
            merged_df.loc[i, 'p_max'] = default_values['p_max'];
            merged_df.loc[i, 'a_max'] = default_values['a_max'];
            merged_df.loc[i, 'alpha'] = default_values['alpha'];
            merged_df.loc[i, 'beta'] = default_values['beta'];
        end
    end
return merged_df;

```

**Algorithm 1:** Fill Missing Values

## Appendix B The derivation of Bayes' theorem.

The meanings of the notations are already explained above. Here are the concrete derivation.

$$\begin{aligned}
 p(s | pre, c) &= \frac{p(pre, c | s) \cdot p(s)}{p(pre, c)} \\
 p(s | pre) &= \frac{p(pre | s) \cdot p(s)}{p(pre)} \\
 \frac{p(s | pre, c)}{p(s | pre)} &= \frac{\frac{p(pre, c | s) \cdot p(s)}{p(pre, c)}}{\frac{p(a | s) \cdot p(s)}{p(pre)}} \\
 &= \frac{p(pre, c | s) \cdot p(pre)}{p(pre, c) \cdot p(pre | s)}
 \end{aligned}$$

According to above analysis, we have:

- **pre** and **c** are conditionally independent given **s**:

$$p(pre, c | s) = p(pre | s) \cdot p(c | s).$$

- **pre** and **c** are independent in the absence of conditions:

$$p(pre, c) = p(pre) \cdot p(c).$$

Under these conditions:

$$\frac{p(s | pre, c)}{p(s | pre)} = \frac{p(pre | s) \cdot p(c | s) \cdot p(pre)}{p(pre) \cdot p(c) \cdot p(pre | s)} = \frac{p(c | s)}{p(c)}.$$

According to Bayes' theorem:

$$\frac{p(c | s)}{p(c)} = \frac{p(s | c)}{p(s)}.$$

Thus, under the conditions of independence and conditional independence between **pre** and **c**, the equation holds:

$$p(s | pre, c) = \frac{p(s | c) \cdot p(s | pre)}{p(s)}.$$

# Report on Use of AI

1. OpenAI ChatGPT (Nov 5, 2023 version, ChatGPT-4)

- Query1:** (1) Swimming Athletics Basketball Equestrian  
(2) Weightlifting Badminton Artistic swimming Wrestling  
(3) Badminton Table Tennis Artistic swimming Cycling Track  
(4) Skateboarding Softball Baseball Beach Volleyball

Please summarize the common characteristics of each row, which should be distinctive and highly representative.

**Output:** I understand that you want to identify a distinctive and representative common characteristic for each row. We can find the most representative feature based on the nature of the sports in each row. Here is the summary for each row:

Row 1: Swimming, Athletics, Basketball, Equestrian

Common Feature: "Traditional and Classic Competitive Sports"

These sports are classic competitive events with a long history and widespread participation, often featured in major international competitions like the Olympics. They are highly competitive and attract a diverse range of countries and regions, drawing a broad global audience.

Row 2: Weightlifting, Badminton, Artistic Swimming, Wrestling

Common Feature: "Sports Highly Dependent on Technique and Skill"

The sports in this row emphasize the athlete's technique, skill, and precision. Whether it's the technical movements in weightlifting, the skillful strokes in badminton, the coordination in artistic swimming, or the tactical skills in wrestling, all these sports rely not only on strength but more on skill and precise movement control.

Row 3: Badminton, Table Tennis, Artistic Swimming, Cycling Track

Common Feature: "Sports Emphasizing Quick Reactions and Precise Control"

The common feature of these sports is the high demand for reaction speed and precise control. Badminton, table tennis, and artistic swimming all require extremely fast reactions and fine motor coordination, while track cycling demands precise control over the trajectory of movement, emphasizing quick reactions and precise control abilities.

Row 4: Skateboarding, Softball, Baseball, Beach Volleyball

Common Feature: "Modern and Highly Visually Appealing Sports"

These sports have a strong modern feel and high visual appeal. Skateboarding, beach volleyball, and baseball (especially with baseball's recent return to the Olympics) are popular among young people and have a strong visual impact. Particularly skateboarding has become a focus of street culture and Olympic competition, offering strong entertainment and visual appeal.

Summary:

Row 1: "Traditional and Classic Competitive Sports"

Row 2: "Sports Highly Dependent on Technique and Skill"

Row 3: "Sports Emphasizing Quick Reactions and Precise Control"

Row 4: "Modern and Highly Visually Appealing Sports"

Each common feature highlights the core attributes of the sports in that row and is distinctive and unique, providing a good summary of the type of sports in each row. If you have any other specific requirements, feel free to let me know!

**Query2:** Help me create a dictionary, mapping all the countries to NOC.

**Output:** Here is a Python dictionary that maps countries to their respective NOCs (National Olympic Committees): country\_to\_NOC = {  
    "United States": "USA",  
    "Great Britain": "GBR",  
    "China": "CHN",  
    "Russia": "RUS",  
    "Germany": "GER",  
    "Japan": "JPN",  
    "France": "FRA",  
    "South Korea": "KOR",  
    "Italy": "ITA",  
    "Australia": "AUS",  
    "Netherlands": "NED",  
    "Hungary": "HUN",  
    "Brazil": "BRA",  
    "Spain": "ESP",  
    "Kenya": "KEN",  
    "Jamaica": "JAM",  
    "Croatia": "CRO",  
    "Cuba": "CUB",  
    >New Zealand": "NZL",  
    "Canada": "CAN",  
    "Uzbekistan": "UZB",  
    "Colombia": "COL",  
    "Switzerland": "SUI",  
    "Iran": "IRI",  
    "Greece": "GRE",  
    "Argentina": "ARG",  
    "Denmark": "DEN",  
    "Sweden": "SWE",  
    "South Africa": "RSA",  
    "Kazakhstan": "KAZ",  
    "Ukraine": "UKR",  
    "Serbia": "SRB",  
    "Poland": "POL",  
    "North Korea": "PRK",  
    "Belgium": "BEL",  
    "Thailand": "THA",  
    "Slovakia": "SVK",  
    "Georgia": "GEO",  
    "Azerbaijan": "AZE",  
    "Belarus": "BLR",  
    "Turkey": "TUR",  
    "Armenia": "ARM",  
    "Czech Republic": "CZE",  
    "Ethiopia": "ETH",  
    "Slovenia": "SLO",  
    "Indonesia": "INA",

```
    "Romania": "ROU",
    "Bahrain": "BRN",
    "Vietnam": "VIE",
    "Chinese Taipei": "TPE",
    "Bahamas": "BAH",
    "Independent Olympic Athletes": "IOA",
    "Ivory Coast": "CIV",
    "Fiji": "FIJ",
    "Jordan": "JOR",
    "Kosovo": "KOS",
    "Puerto Rico": "PUR",
    "Singapore": "SGP",
    "Tajikistan": "TJK",
    "Malaysia": "MAS",
    "Mexico": "MEX",
    "Venezuela": "VEN",
    "Algeria": "ALG",
    "Ireland": "IRL"
}
```

**Query3:** What are the possible events that may occur in 2028 Summer Olympics, according to the added sports list: Baseball/softball, cricket, flag football, lacrosse, squash?

**Output:** Projected breakdown of programs:

1. Baseball/Softball

Baseball: Major League Baseball (MLB): the standard professional baseball league, primarily distinguished from other levels of the game in terms of player strength, pace of play, and athletic demands.

Fast Pitch Baseball (Fast Pitch) and Slow Pitch Baseball (Slow Pitch): These are the two main types of baseball and softball pitching, with Slow Pitch being more commonly seen in recreational and women's games, and Fast Pitch being more commonly used in professional games.

Softball: Fast Pitch Softball: The game is faster and the rules are closer to those of baseball.

Slow Pitch Softball: A slow pitch with a curved path that requires a high degree of skill and strategy.

Wall Softball: Using multi-angle wall bounce to increase the difficulty, more competitive.

2. Cricket

International Cricket (International Cricket): such as One Day International (ODI), Twenty20 (T20), etc., is currently the most popular form of cricket around the world.

Test Cricket: A traditional cricket match that lasts five days and emphasizes endurance and strategy.

Ten-a-Side Cricket: Reduced to 10 players per team, the game is faster paced.

Street Cricket: Informal form of cricket, commonly played on city streets, backyards, etc.

3. Flag Football

Men's/Women's Flag Football: divided into men's and women's games, the rules are less different, mainly in the physical ability of the participants, speed.

7-a-side Flag Football: Compared to traditional 11-a-side football, the field and personnel are smaller, and the game is more tactically oriented.

Mixed Sex Flag Rugby: A format in which both men and women play together, with rules adjusted for physical demands and an emphasis on tactical cooperation.

#### 4. Lacrosse

Men's Lacrosse: In contrast to Women's Lacrosse, Men's Lacrosse is more physically demanding and utilizes hard protective equipment.

Women's Lacrosse: The rules are relatively mild and emphasize skill and speed.

Box Lacrosse: Played on closed courts, the game is faster paced and more physical, and is commonly played in Canada.

Field Lacrosse: Unlike indoor lacrosse, it is usually played on a large grass field and the pace of the game is relatively slow.

#### 5. Squash

Singles Squash: Two players play against each other using singles rules.

Doubles Squash (Doubles Squash): four players participate in the game, the rules of the game is similar to the singles, but there is more movement and cooperation on the court.

Soft Squash: A softer ball, suitable for beginners or older players.

Hardball Squash: A more challenging game using hard balls, commonly played in the United States.

**Query4:** Help me identify all the team sports listed below:

B-Girls

Women's Sabre Individual

Men's 67kg

Jumping Individual

Women's 200m Freestyle

Basketball Men's Basketball

Judo Men's Extra-Lightweight

Women's +87kg

Football Men's Football

Athletics Men's Long Jump

...

**Output:** Here is the identification of team sports from the list provided:

Team Sports: 1. Basketball Men's Basketball – Team sport

2. Football Men's Football – Team sport

3. Hockey Men's Hockey – Team sport

...

2. DeepL (Version 25.1.4.15077+495e4eb7b7cc47c7c43de87ac335c76ea3e6a019)

We used it as a translator for some special nouns and words we do not know.