Models for Olympic Medal Tables

2025 MCM Problem C

Sofia Bonyadi, Aryan Dalal, Jiayi Cora Guo
Julie Quan, Shani Shaham

Spring 2025

A project submitted in partial fulfillment of the requirement for the class of MATH 142: Mathematical Modeling

in the

Department of Mathematics

of the

University of California, Los Angeles (UCLA)

Committee In Charge:
Mason A. Porter, *Professor of Mathematics*
Haoyang Lyu, *Graduate Teaching Assistant*

Contents

1 Summary

With the 2028 Los Angeles Olympic Games approaching, global viewers and countries are increasingly interested in the potential medal outcomes of their national teams. Such forecasts enable countries to allocate funding in training resources, especially to events with highest medal prospects. In this problem, we focus on predicting the total medal counts for each country, countries who will win their first medal, and how will 2028 results differ from Paris 2024.

Building on this, the improved model introduces historical medal performance into medal assignment. We used an exponential time-decay function to weight each country's medal record by recency, so that more recent Olympic success carries greater influence. Each country is limited to entering a maximum of three athletes per event, in line with actual Olympic regulations. We also introduced a "host-effect" to the United States. As the host for LA 2028, the United States receives a 5% boost in both participation and medal-winning probabilities, reflecting increased qualification rates and home-field advantage. Furthermore, We assigned a small nonzero baseline probability (= 10) to every entry in the medal matrix to allow for the possibility of first-time medalist countries. After running 1,000 simulations, we rounded the fractional medal counts using a Hamilton-style apportionment strategy to ensure integer medal counts that match the total number of events. We scaled expectations to guarantee one medal to countries with a nonzero expectation, and used a rounding pass that adjusted final tallies by no more than ± 1 medal.

As a result, our model predicts that the United States will mark first place in the 2028 medal table with approximately 188 medals, followed by China, Great Britain, and Australia. Compared to Paris 2024, countries like Italy and Hungary are expected to improve, while France, Korea, and Japan decline due to the loss of prior host advantages. Liberia emerges as the most likely candidate to win its first medal, with a 7.1% chance.

While the model provides reasonable results into LA 2028, there are few limitations and weaknesses that may reduce realism. The event list was based on 2024 data due to the lack of a finalized 2028 schedule. Also, all events were modeled as individual competitions, omitting proper treatment of team-based events. Furthermore, our model treats all events as independent, not accounting for correlations across similar disciplines (e.g., multiple swimming events). With more time, we would expand the model to handle these areas to build a more accurate projection.

Through these modeling choices and simplifications, we created a framework that balances complexity with interpretability, offering insight into likely outcomes and structural biases in Olympic medal distributions.

2 Introduction

2.1 Problem

The Olympic Games attract billions of viewers from all over the world as viewers cheer on athletes. Fans connect with one another as they witness exciting competition and support athletes representing their nation. Fans not only tune in to watch individual events, but also follow medal counts to see what countries are leading in medals.

Our goal is to build a model that forecasts the medal counts for each country for the 2028 summer Olympics based on the given data. We aim to predict countries that earn their first medals and analyze what countries improve. We hope to explore the correlation between participating countries and the events they chose to compete in.

For many countries, earning a medal brings national recognition, pride, and global attention. Countries such as Albania, Dominica, and Cabo Verde won their first-ever medals at the 2024 Paris Olympics, while many nations have yet to win their first medals. For this reason, fans track medal count to see what countries are ranking in medals.

The data provided was recorded by the International Olympic Committee during each respective Olympics. The first data set includes all athletes with their representing country, year, respective sport, event, and result (medal type or none) from 1896 to 2024 Summer Olympics. Countries that have won at least one medal are listed in medal count tables from 1896 to 2024 Summer Olympics. Host countries for all summer Olympics from 1896 to 2032 are also provided to analyze what effect the host country has on medal results. The last set features all sports with their sports governing body, discipline, and the number of events for that discipline for all summer Olympics from 1896 to 2024.

Although Olympic prediction is frequently conducted, researchers use different strategies to do so. Schlembach *et al.* constructs a two-stage Random forest, a machine learning model to predict medal count for the Olympics from 2008 to 2020. The parameters include socioeconomic data such as GDP, population size, human development index (HDI) and using past medal counts to predict total medals in the first stage to then forecast gold medal count in stage 2. Bernard and Busse present a regression model to analyze how population and GDP affect medal count. They specifically use a log-linear regression model to predict the Sydney 2000 Olympics using historical data and taking the logarithm of population and GDP per capita.

We analyzed the given data and made general assumptions to implement a probabilistic model based on athlete participation. However, it does not consider countries that are historically strong. Resulting in an improved model that accounts for historical medal data. We also consider the impact of "host effect" for the United States on medal outcomes. The model was simulated over many iterations and found United States to be in first place for overall medal count, then China in second, Great Britain in third, and Australia in fourth. Countries such as Italy and Hungary are expected to improve, however, France and Japan are expected to decline. We explore how initial assumptions have determines the results of our model and what improvements could be made to assumptions to strengthen our findings.

3 Mathematical Preliminaries

In order to give a sound introduction to this paper, we want to introduce some basic definitions which will be referenced throughout the text. The reader should note that these definitions are by no means their rigorous versions as we don't assume measure theory as a pre-requisite. However, the reader should note that measure-theoretic definitions to the below are their rigorous versions. As such, let's begin with what exactly is a *random variable*.

3.1 Basic Probability Theory

Suppose there exists Ω, named *sample space*. This sample space takes in values $\omega \in \Omega$ which represent all possible outcomes of the space. Now, we are ready to define a random variable.

Definition 3.1 (Random Variable)**.** A *random variable* is a function that sends outcomes from Ω into $\mathbb{R}$, given as

$$X(\omega) : \Omega \longrightarrow \mathbb{R}$$

Remark. What makes a random variable extremely powerful is that it's a function that's considered *measurable*. What constitutes a measurable function? In general, a measurable function is a special mapping that preserves structural properties between two *measurable spaces*. Now, what exactly is a measurable space? A measurable space requires us to understand the formalism of σ-algebra. We don't concern ourselves with such levels of abstraction.

A random variable can be *discrete* or *continuous*. A discrete random variable takes in a countable number of distinct values whereas a continuous random variable takes in an uncountable infinite number of values on an interval.

Given our newly learned definition of a random variable, let's define expectation.

Definition 3.2 (Expected Value)**.** Let X be a discrete random variable for outcomes x_i. The expect value of X, denoted $\mathbb{E}[X]$ is defined by

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i P(X = x_i)$$

Intuitively, this is imply a weighted average of the possible outcomes of a discrete random variable. The fundamental idea of employing a Monte-Carlo method is to solve for some type of expectation that we intend to derive.

Now, what if X was a continuous random variable? Surely we can't use sums, right? Indeed, for the expectation of a continuous random variable, we use an integral while the integrand (or the summand a priori) remains the same. Now, we permit X to take continuous values.

Definition 3.3 (Probability Distribution)**.** Let X be the set of countable points. THe *discrete probability distribution* on X is defined by the function $p(x)$ on X such that $0 \leq p(a_i) \leq 1$ and $\sum_{i=1}^{\infty} p(a_i) = 1$.

Once again, what if we wanted a continuous probability distribution? Replace the sum with an integral and the integrand (or the summand) remains the same.

3.2 Stochastic Processes

Definition 3.4 (Monte-Carlo Simulations)**.** A *Monte-Carlo method* is a branch of experimental mathematics wherein stochastic dynamics are inherent to the design of the experiment. A *Monte-Carlo Simulation* is simply a simulation of a stochastically defined system.

In general, we can use Monte-Carlo simulations for anything that we expect to contain some level of randomness. In particular, we want to understand the behavior of the system without necessarily observing the system itself and so, we simulate it via randomness.

4 Methodology

Our goal was to create a model to predict the 2028 Olympic medal tables. Our central question became: at the elite level of Olympic competition, does skill actually make a significant difference, or does winning a medal come down to chance? This question influenced our decision to build a probabilistic, simulation-based model.

4.1 Baseline Model

We began our modeling work with a baseline model, utilizing the Monte Carlo simulation framework to estimate future medal performance based solely on historical participation data. The central assumption of this model is that all athletes in each event have an equal chance of winning a medal, regardless of nationality or historical performance. This allows the model to reflect the intuitive idea that higher participation translates to greater medal opportunity, even under neutral skill assumptions. We chose to utilize historical Olympic data from 1992-2024 to avoid biases introduced by widespread boycotts in earlier Games.

4.1.1 Assumptions

General:
(1) The 2028 Los Angeles Olympics will have the same events as the 2024 Paris Olympics.
Justification: Since we did not have access to the 2028 Olympic event list, we assumed it will mirror the 2024 Games. This is a reasonable forecasting assumption in the absence of a confirmed 2028 event schedule, especially considering major event changes are uncommon within a single quadrennial cycle.
(2) For each event, 16 athlete slots are sampled and filled.
Justification: This is a realistic approximation of typical Olympic event finals (track and field or swimming often have 8–16 finalists). It allows for variability while still reflecting limited competition per event.
(3) Countries sampled to participate based on historical frequency
Justification: This assumption is rooted in realism. Countries tend to consistently enter events they've historically competed in.
(4) All events are simulated as individual as opposed to team events Justification: This is a simplifying measure for our model, based on the fact that most Olympic events are individual.
Baseline Model Assumptions:
(1) There is an equal chance of winning for all participants in a single event.
Justification: This simplification removes bias from historical medal counts and emphasizes randomness in elite competitions, where margins are often razor thin. It also enables us to isolate the effect of participation volume on medal outcomes.

4.1.2 Estimating Participation Rates

To realistically simulate the structure of Olympic events, we first calculated the frequency with which each country has historically participated in each event. This allowed us to capture not just whether a country has competed in an event, but how consistently it sends athletes to that event over time. Using athlete-level data from Olympic Games held between 1992 and 2024, we counted the number of appearances by each country in each event and normalized this by the number of Olympic Games in the period. This produced an average participation rate for each country-event pair.

$$\text{AvgParticipation}_{e,c} = \frac{\text{No. of athletes from country } c \text{ in event } e}{\text{No. of Olympic Games since 1992}} \tag{1}$$

We stored these values in a participation matrix $P \in \mathbb{R}^{E \times C}$, where E is the set of events and C the set of countries. For each event e, we treat the row P_e as a probability vector over countries c, and use it to sample athlete slots in the simulation. This matrix formed the backbone of our simulation. For each event, we treated the row P_e as a probability distribution over countries and used it to sample which countries would send athletes to the event. This structure reflects the assumption that countries that have historically sent more athletes to an event are more likely to do so again, thus increasing their chances of winning a medal in a model where all entrants are treated as equally skilled.

4.1.3 Uniform Medal Assignment

Using the P matrix, we were able to sample athlete slots in each event in the 2024 Olympics. For each event, we removed countries with weak participation (defined as average participation < 0.3). We simulated n=16 athlete slots with replacement, meaning a country could appear multiple times in a single event. This reflects the notion that countries with deeper delegations have more chances to win. Events lacking sufficient medal data ("Mixed Relay" and "Relay Only Athlete") were excluded from the simulation to ensure consistency in medal assignment. From the sampled athlete slots, we required at least 3 unique countries to ensure medals could be awarded properly. Then, we selected 3 unique countries at random to receive Gold, Silver, and Bronze medals for that event. Selected winners were recorded by type in our medals matrix M.

4.1.4 Monte Carlo Averaging

This process described above simulates one cycle of Olympic Games. We repeated the simulation $N = 1000$ times. The expected number of medals of type $m \in \{\text{Gold}, \text{Silver}, \text{Bronze}\}$ for country c was calculated by:

$$\text{ExpectedMedals}_{c,m} = \frac{\#\text{ of medals of type } m \text{ won by } c}{N} \tag{2}$$

The final result was an expected medal table sorted by total medals per country. Randomness is introduced in this model at two key points: first, in sampling countries for athlete slots in each event using probabilities derived from participation rates, and second when randomly assigning medals among the eligible athletes in each simulation.

4.2 Improved Model

This baseline model served as a neutral starting point, allowing us to isolate the impact of participation volume. However, it lacked predictive power for historically strong nations. This motivated the development of our improved model, which incorporates historical medal outcomes into performance based weighting, along with additional structural enhancements.

4.2.1 Additional Assumptions

In addition to the general assumptions previously listed, the improved model assumes:
(1) Historical medal performance informs current medal probabilities.

Justification: This is rooted in the fact that countries with strong infrastructure, coaching, and athletic programs tend to remain strong over time. Empirical Olympic data also shows high medal concentration consistently remains with few top performers.
(2) Countries with no historical medals still have a small nonzero chance of winning.

Justification: In order to incorporate randomness into our model and preserve realism of rare breakthrough medal wins.

(3) Countries cannot send more than 3 athletes per event in the simulation.

Justification: Following the 2024 Paris Olympics, there is a maximum of three athletes per eligibility of the National Olympic Committee.

(4) The host nation receives a small boost in participation and performance probabilities.

Justification: This reflects the "host effect" phenomenon, where host countries tend to perform better than expected, often credited to a "home-field advantage."

4.2.2 Key Model Enhancements

To improve our model results and better utilize data provided, we moved beyond solely examining historical participation rates to incorporate historical medal performance into probability distributions. We assign a decaying weight to medals earned in each Olympic year:

$$w_y = e^{-\alpha(2028-y)} \quad \text{with } \alpha = 0.15 \tag{3}$$

This ensures recent results carry more weight. For example, medals from 2024 are worth more than those from 2004. Then, for each country c, event e, and medal type m, we compute:

$$\text{WeightedMedals}_{e,c,m} = \sum_{y=1992}^{2024} \text{MedalsWon}_{e,c,m,y} \cdot w_y \tag{4}$$

These form three medal probability matrices: `gold_matrix`, `silver_matrix`, and `bronze_matrix`. Each row is an event, and each column is a country.

Athletes are sampled utilizing the participation matrix P, following the method from the baseline model. However, in this improved model, we enforced a cap of 3 athletes per country per event. This was incorporated to better adhere our model to Olympic rulings that restrict each NOC to three athletes per individual event.

For each event, we assign medals using weighted probabilities. For country c in event e, the probability of winning a medal is:

$$P(c \text{ wins medal}) \propto \left(\text{WeightedMedals}_{e,c,m} \cdot \text{Slots}_c\right)^2 \tag{5}$$

This formula ensures that a country's chance of winning is proportional to both its past success and current event presence. $\text{WeightedMedals}_{e,c}$ reflects historical strength in the event as previously displayed in Equation 4. Slots_c represents the number of athlete slots sampled in the event. An exponent of 2 is utilized to sharpen the competitive edge and make strong countries more dominant.

In order to take the "host effect" phenomenon into account, we introduced a two part host country advantage for the host of the 2028 Olympic Games. Firstly, the host country receives a 5% boost in participation probability when sampling athletes:

$$\text{P}_{e,c} \leftarrow \text{P}_{e,c} \cdot 1.05 \quad \text{if } c = \text{USA} \tag{6a}$$

where $\text{P}_{e,c}$ represents each country c's probability of being sampled for a given event e. This increases the likelihood that the USA will draw more athlete slots per event, reflecting the fact that host nations typically qualify for more events due to automatic entries. Secondly, if the host country is eligible to win a medal (i.e., has at least one sampled athlete), its weight in the medal matrix is also scaled up by 5%, giving the host nation a slight advantage in medal assignments as well:

$$\text{WeightedMedals}_{\text{USA}} \leftarrow \text{WeightedMedals}_{\text{USA}} \cdot 1.05 \tag{6b}$$

reflecting the assumption that athletes perform better at home, due to factors like familiarity and larger crowd support. To ensure that countries with no prior medal history were not completely excluded from winning, we added a small baseline value $\varepsilon = 10^{-6}$ to each entry in the medal probability matrices. Providing each country with small nonzero weights allows us to retain a degree of randomness in our simulation, avoiding overly deterministic results that are dominated by historical data.

4.2.3 Summary of Improved Model Pipeline

1. Compute participation probabilities and weighted historical medal matrices.
2. For each event, simulate 16 athlete slots using participation rates (capped at 3 per country).
3. Assign medals using weighted probabilities reflecting strength $\times$ slots.
4. Repeat over 1000 simulations and average results to estimate medal table.

4.3 Rounding and Medal Allocation Strategy

1. **Scale to event totals**
   For each medal type, multiply every country's fractional expectation by a constant so that the column sums exactly to the Olympic targets:

   $$356 \text{ gold}, \quad 356 \text{ silver}, \quad 356 \text{ bronze}.$$

2. **Guarantee one medal where deserved**
   If a country's scaled expectation is positive but still < 1, bump it up to a single medal. This prevents small but meaningful probabilities from being rounded to zero.

3. **Single rounding pass with minimal adjustment**

   - Round every scaled value to the nearest integer.
   - If the rounded column is short (or long) by a few medals, adjust using a largest-remainder tweak: give the remaining medals to—or take them from—the countries whose fractional parts are *closest* to the 0.5 threshold.
   - No nation's tally changes by more than ± 1 medal, and the final totals return to $351 - 351 - 351$.

Outcome. The resulting integer medal table (i) matches the fixed number of Olympic events, (ii) keeps each nation's share as close as mathematically possible to its simulated expectation, and (iii) still rewards emerging NOCs with at least one medal whenever their modeled probability is non-zero.

4.4 The "Great Coach" Effect

To model the dynamics of a "Great Coach" effect on the Olympic medal tables, we must first understand how do we define what constitutes a "Great Coach". A simple question like this can often result in various answers and so, to preemptively remove the possibility of ambiguity, we define the "Great Coach" effect as a metric that can result in a positive sporting impact on both an individual as well as the team. For the purpose of the greater model of this paper, we are assuming the Olympics as an individual event and so, we can forgo teams. As such, suppose there exists two countries X and Y. Country X and country Y have athletes that have an equal chance of winning a medal; however, country Y has been performing better as a result of a highly skilled coach who once was an Olympian themselves. Country X, in an attempt to invest in its infrastructure decides to poach country Y's coach. We now ask ourselves two questions:

1. Does country Y's performance increase in this event?
2. Does country X's performance decrease as a result of losing a "great coach"?

We also have some other follow-up questions such as, how fast does country Y's performance increase with a "great coach" and does country X remain dominant in the event or do they lose their performance. For simplicity, we assume that for now, we don't concern ourselves with problems of convergence, that is, given a "great coach", we don't particularly care about how *long* it takes to reach optimal performance, rather, we only care *if* they reach optimal performance.

As such, we are now ready to lay out some ground assumptions.

1. At every iteration of the Olympic games, coaches are either free to continue with their current team or they may switch teams.
2. There exists a natural order of more skillful and less skillful coaches. Not every coach is assumed to have the same coaching ability. We ignore the factors that may result in how a coach becomes more skillful.
3. Countries that historically performed well in the Olympics are generally able to attract skillful coaches but not always.

4.4.1 A Model Proposition for Great-Coach Effect

Given this set of assumptions, we observe that first, coach movement is stochastic in nature so we expect a submodel that is a stochastic process. So, given

$$\mathcal{C} = \{C_1, C_2, \ldots, C_n\}$$

and

$$\mathcal{K} = \{K_1, K_2, \ldots, K_m\}$$

for countries and coaches respectively for n not necessarily equal to m. We can say that every coach has some skill level, denoted $s_l \sim N(\mu_s, \sigma_s)$, that is, skill level is normally distributed.

We intend to utilize the Markov-Chain Monte-Carlo framework to model this phenomenon. Why? We utilize MCMC because we want to draw samples from a distribution such that each draw only depends on the state of the previous draw, that is, these samples form a Markov-Chain. Indeed, we can determine medal outcome for each country by

$$M_{\text{gold, C}} \sim \text{Poisson}\left(E_C^G\right)$$
$$M_{\text{Total,C}} \sim \text{Poisson}\left(E_C^T\right)$$

where E_C^G and E_C^T are expectation values for Gold type medals and Total medals respectively given $\mathcal{C}$. We also assume that the number of medals that a country wins is in-line with a Poisson distribution. We are defining

$$E_C^G := x_G\left(a_c + \gamma_G s_{l_{f(c)}}\right)$$

$$E_C^T := x_T \left(a_c + \gamma_T s_{l_{f(c)}}\right)$$

where $s_{f(c)}$ represents the skill level of the coach for country C and the parameters γ_G represent how much coach skill influences the ability to win gold medals while x_G represents a simple constant. The same carry for E_C^T apart from notational differences.

4.4.2 The Case of Switching Coaches

To have a proper dynamic of how coaches can leave or join teams, we want to propose that in our simulation, we randomly select a non-zero amount of coach country pairs and propose a swap that's random in nature. Notice that this suggests that one of our assumptions is not active in that all teams now have an equally random chance of being able to poach a coach from another team and that countries with greater historical performance are also subject to the same randomness. This now allows us to define what's called the Metropolis-Hastings Rule.

Suppose we want to find a way to understand how we can simulate switching of coaches amongst teams, we can do this by

$$\text{CoachAccept} = \min\left(1, \frac{P\left(\text{new coach}\right)}{P\left(\text{current coach}\right)}\right)$$

Suppose we choose a coach such that $P\left(\text{new coach}\right) > P\left(\text{current coach}\right)$, by the Metropolis-Hastings acceptance rule, we accept the coach switch because it results in a better coach. However, if its the converse, we still accept the switch but do so proportional to the ratio $P\left(\text{new coach}\right)/P\left(\text{current coach}\right)$. We do so because if we chose to not accept, then we will simply have an order of teams that will each result in a better switch, contrary to our modeling purpose where we want to see what can happen if a switch occurs but it was actually not a good coach switch.

We can now run the MCMC with a Metropolis-Hastings forcing Acceptance Rule for $N = 1000$ runs which should allow us to understand if the Great Coach effect plays a large enough influence over multiple team switchings. We note that since we are limited by data available, the Great Coach effect's model is independent from the rest of the paper given its over-reliance on placeholder data rather than historical data.

5 Results and Analysis

5.1 Overall Medal Projections for LA 2028

Using 1 000 Monte-Carlo runs, our model simulates every Tokyo-2024 event that is also scheduled for Los Angeles. A mild host multiplier (1.05) boosts U.S. performance, while an exponential time-decay down-weights older Games. Table 1 reports the mean medal counts and their 95 % prediction intervals (2.5-th and 97.5-th percentiles across the replicates).

Table 1: Projected LA 2028 medal counts with 95 % prediction intervals (hybrid format: integer point estimate, decimal bounds).

Country	Gold	Silver	Bronze	Total
USA	36 (36.0–36.4)	39 (39.1–39.5)	30 (29.9–30.3)	106 (105.2–105.9)
CHN	20 (20.3–20.6)	15 (15.2–15.5)	14 (13.8–14.1)	50 (49.4–50.1)
GBR	12 (11.8–12.1)	11 (11.3–11.5)	17 (17.3–17.6)	41 (40.5–41.0)
AUS	13 (12.4–12.7)	14 (13.5–13.8)	14 (13.6–13.9)	40 (39.7–40.2)
ITA	7 (7.2– 7.4)	10 (10.1–10.4)	13 (12.6–12.9)	30 (30.1–30.5)
GER	9 (8.4– 8.7)	7 (7.4– 7.6)	7 (7.3– 7.5)	23 (23.2–23.6)
JPN	8 (8.3– 8.5)	7 (7.2– 7.5)	8 (7.4– 7.7)	23 (23.1–23.5)
FRA	5 (5.2– 5.5)	9 (8.6– 8.9)	7 (7.0– 7.2)	21 (21.0–21.5)
NED	7 (7.3– 7.6)	7 (6.5– 6.7)	7 (6.4– 6.6)	21 (20.4–20.8)
CAN	7 (6.9– 7.1)	5 (4.8– 5.0)	6 (5.7– 5.9)	18 (17.5–17.9)

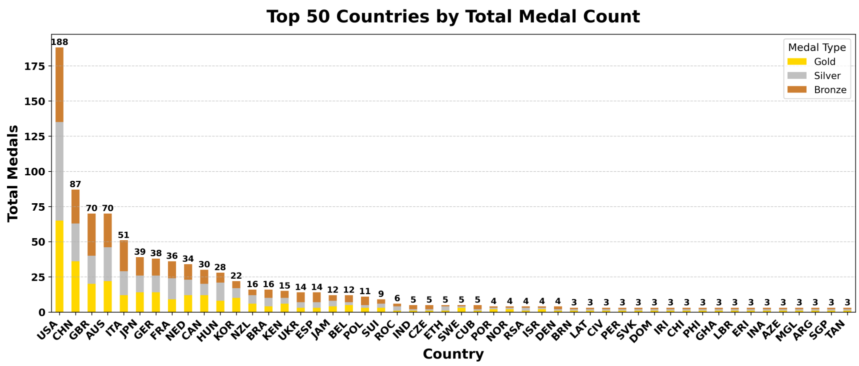

Figure 1: Stacked bar chart corresponding to Table 1.

Figure 1 confirms that the United States, aided by the host-nation effect, is projected to top the table by a wide margin. China remains a solid second, while Great Britain and Australia are in a close race for third place. Italy, Germany, and Japan occupy the middle of the top ten, with the Netherlands and Canada completing the list.

5.2 Change Relative to Paris 2024

Instead of raw counts, Figure 2 compares each country's *share* of the overall medal pool in Paris 2024 (blue) with its simulated share for Los Angeles 2028 (orange). Percentages are shown atop the bars so small differences are visible.

Most likely to improve.

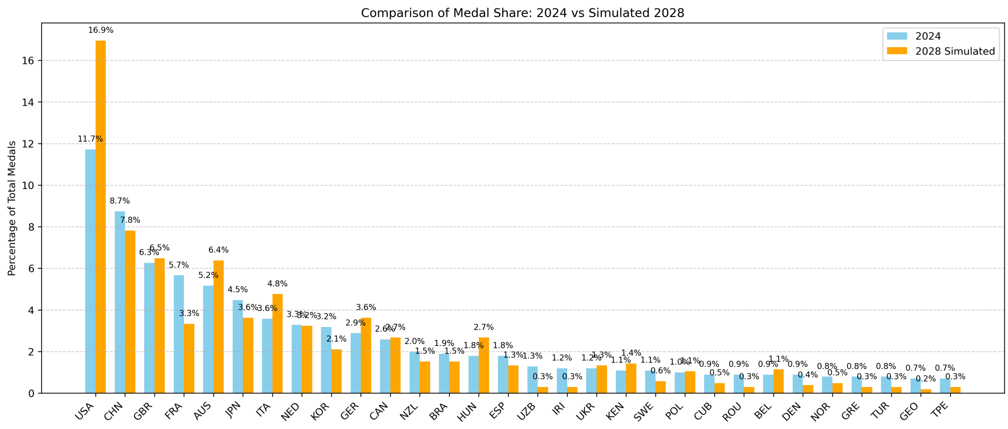

Figure 2: Share of total medals: Paris 2024 results (blue) versus our LA 2028 simulation (orange). Labels give the exact percentage of all medals won.

- **United States** grows from 11.7 % to 16.9 % of total medals (+5.2 pp), the largest jump in the field.
- **Australia** rises from 5.2 % to 6.4 % (+1.2 pp), overtaking France and closing on Great Britain.
- **Italy** increases from 3.6 % to 4.8 % (+1.2 pp)
- **Hungary** bumps its share from 1.8 % to 2.7 % (+0.9 pp),

Most likely to decline.

- **France** falls from 5.7 % to 3.3 % (–2.4 pp) once the host-nation boost of Paris is removed.
- **Korea** drops from 3.2 % to 2.1 % (–1.1 pp),
- **China** decreases slightly from 8.7 % to 7.8 % (–0.9 pp),
- **Japan** slips from 4.5 % to 3.6 % (–0.9 pp)

Overall, only a handful of countries are forecast to change their medal *share* by more than one percentage point; most remain within the statistical noise band implied by the 95 % prediction intervals.

5.3 First-Time Medal Probabilities

Only a handful of countries with no prior Olympic podium finish are forecast to break through in Los Angeles. Figure 3 ranks the twenty most likely debutants and shows their 95 % prediction intervals.
Liberia leads with a 7.1 % chance of securing at least one medal (CI ±1.6 pp). Guinea, the Maldives, and Mali each sit near 0.2 %. Every other candidate lies below 0.2 %.

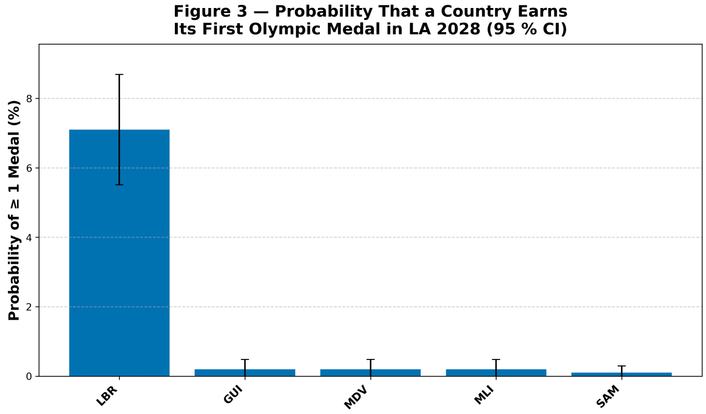

Figure 3: Probability that a nation earns its *first* Olympic medal in LA 2028. Error bars denote 95 % prediction intervals from 1 000 Monte-Carlo runs.

5.4 Home–Country Bias

To gauge the "host advantage," we reran the entire Monte-Carlo pipeline with the `host_multiplier` set to 1 (no boost) and contrasted the outcome with our baseline run that applies a modest 5 % bonus to every U.S. entry. Figure 4 shows the two scenarios side-by-side for the top-20 nations; the hatched bar is the no-host run, while the solid bar includes the host boost.

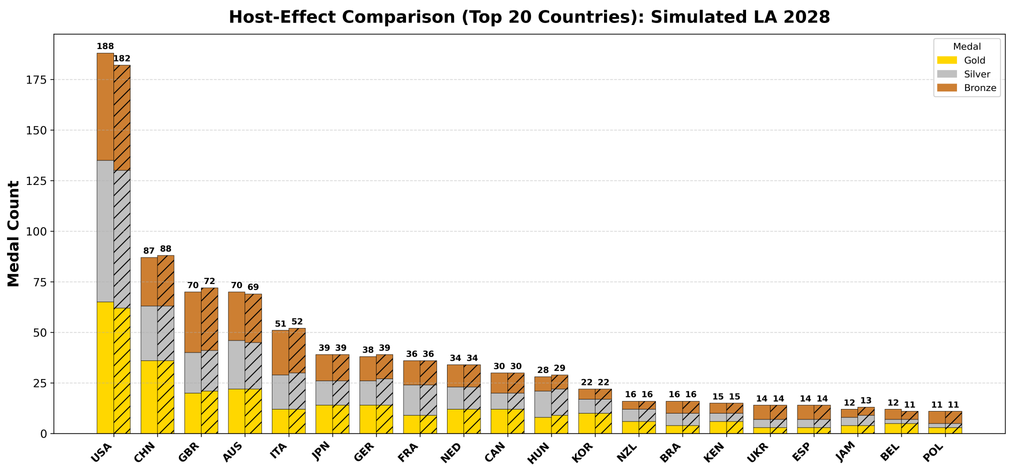

Figure 4: Effect of the host bonus: solid bars = baseline run (`host_multiplier`=1.05); hatched bars = rerun with no host boost.

Key observations.

- The United States gains **+6 medals** (188 vs 182), a 3 % lift that is spread fairly evenly across gold, silver, and bronze. The bump is modest compared with the large 20–30 % uplifts seen for previous hosts such as China (2008) and Japan (2021).
- Most other nations shift by at most ±1 medal. China drops slightly (–1), while Great Britain picks up +2 as events redistribute.

- The overall ranking of the top-20 remains unchanged, confirming that a 5 % bonus affects totals but does not overturn the medal hierarchy.

Hence, with a conservative host multiplier, the "home-field" effect is present but not dominant: it enlarges the U.S. haul by only a single-digit margin and leaves other countries' projections essentially intact.

6 Conclusions and Discussion

6.1 Conclusion

In this project, we successfully developed a probabilistic model using Monte Carlo simulations to forecast the 2028 Olympic medal table. Our approach incorporated both historical participation data and past medal performance to account for structural advantages among nations, while preserving fairness through randomized simulations. While our model aligns well with intuitive and historical trends, it remains limited by simplifying assumptions, such as treating all events as individual and excluding non-sporting variables like GDP or geopolitical factors. Future improvements may include sport-specific modeling, enhanced regression frameworks, or even graph-based learning techniques to better capture the complexities of international athletic competition.

6.2 Rounding Error

To convert fractional medal expectations into whole-number allocations, we applied a Hamilton-style rounding procedure that ensures each medal type (gold, silver, bronze) exactly matches the number of Olympic events. This method maintains fairness by keeping each country's total as close as possible to its scaled expectation, while also satisfying the hard constraint of integer values.

Crucially, the rounding error is tightly controlled: no country deviates by more than ± 1 medal from its ideal count in any category, and the overall error across all nations sums to zero for each medal type. To promote inclusivity, countries with non-zero expectations are guaranteed at least one medal. While this may slightly inflate totals for low-expectation countries, it aligns with Olympic principles of broad global representation.

Though minimal, this process introduces small edge effects—particularly around the rounding threshold of 0.5—that can marginally benefit or penalize countries on the cusp. Nonetheless, the method strikes a deliberate balance between statistical precision and symbolic fairness, ensuring the final medal table is both mathematically sound and globally representative.

6.3 Strengths and Weaknesses

While our model provides generally acceptable results and insights into the 2028 Olympic medal distribution for each nation, there are a few limitations and variables that we could incorporate to improve predictive accuracy. First, our model was built using 2024 event lists. It is heavily based on historical data since we did not have a complete list of events for the 2028 Games, Thus, we scaled the percentage of 2024 wins by the total number of 2028 events. This approach may not fully account for the addition or removal of specific events in 2028.

Another limitation we faced was limited access to the data. Not only do historical patterns in medal wins influence projected medal counts, other factors are also significant in giving insight into medal projection such as GDP, economic growth, and recent athletic performance. These variables can significantly impact the medal results of each nation. However, data access constraints may have limited the model's accuracy in forecast.

Our model depends on historical data, and this makes it difficult to predict medal outcomes for nations with no past data or prior medal wins. Reliance on historical data inherently favors countries that have won medals in the past. Thus, we arbitrarily added a small probability baseline to countries with no previous medals. However, this attempt did not seem to make significant change in new or emerging countries. This again underscores the need for variables other than past records.

Another limitation in our models predictive power is the fact that we treated all events as individual, potentially distorting medal count outcomes for team based events.

6.4 Future improvements

We initially planned to decompose results by sport (e.g., athletics, swimming, cycling) in order to pinpoint each NOC's core strengths. In practice, this would require mapping each of the 287 simulated events back to a broader sport category and recomputing shares. Due to time constraints, we leave that sport-level aggregation as a clear avenue for future work.

One improvement would be a more sophisticated treatment of team vs. individual events. To better reflect how team events operate in the Olympics, we would separate the events list into individual and team categories. Team events should be sampled without replacement to reflect real constraints on country participation. This would also prevent the possibility of a country receiving two medals in one team based event.

Additionally, we would like to introduce correlations between related events within the same sport. For example, a country that excels in the 100m freestyle is also likely to perform well in the 200m freestyle. Currently, our model treats each event independently. Introducing intra sport correlations would increase realism by capturing shared performance trends across similar disciplines.

Finally, access to more recent (in-between) performance data, such as world rankings or continental qualifiers, could further improve our ability to predict first time medal wins and enhance overall medal table forecasts.

References

Bernard, Andrew B and Meghan R Busse (2004). “Who wins the Olympic Games: Economic resources and medal totals”. In: *Review of economics and statistics* 86.1, pp. 413–417.

Csurilla, Gergely and Imre Fertő (Feb. 2, 2023). “The less obvious effect of hosting the Olympics on sporting performance”. In: *Scientific Reports* 13.1. Publisher: Nature Publishing Group, p. 819. ISSN: 2045-2322. DOI: 10.1038/s41598-022-27259-8. URL: https://www.nature.com/articles/s41598-022-27259-8 (visited on 06/04/2025).

Ma, Jiayi, Yunyi Chen, and Xiang Lin (Mar. 19, 2025). “Study on Medal Prediction Based on XGBoost and ARIMA Models”. In: *Scientific Journal of Technology* 7, pp. 215–222. DOI: 10.54691/stbqbk94.

Monte Carlo Simulation - an overview — ScienceDirect Topics (2025). URL: https://www.sciencedirect.com/topics/economics-econometrics-and-finance/monte-carlo-simulation (visited on 06/04/2025).

“Olympic medal prediction based on Seq2Seq model and TPE optimization” (2025). In: *World Journal of Sport Research* 3.1. ISSN: 29599946. DOI: 10.61784/wjsr3008. URL: http://www.upubscience.com/News11Detail.aspx?id=1229&proid=53 (visited on 06/04/2025).

Raychaudhuri, Samik (n.d.). “INTRODUCTION TO MONTE CARLO SIMULATION”. In: ().

Schlembach, Christoph et al. (2022). “Forecasting the Olympic medal distribution–a socioeconomic machine learning model”. In: *Technological Forecasting and Social Change* 175, p. 121314.

STGCN-LSTM for Olympic Medal Prediction: Dynamic Power Modeling and Causal Policy Optimization (2025). URL: https://arxiv.org/html/2501.17711v1#S4 (visited on 06/04/2025).

Zhu, Zhaoyu et al. (Feb. 28, 2025). “Application of Data-Driven Prediction and Strategic Optimization in Olympic Medal Distribution”. In: *International Journal on Advanced Science Engineering and Information Technology* 1, pp. 10–26. DOI: 10.70731/c3q02875.

7 Author Contributions

Below, we summarize every author's contributions to the paper. Every author made significant technical contributions during the modeling process; some authors primarily worked on the modeling.

Sofia Bonyadi
Sofia worked on the non-mathematical writeup, introduction and background sections of the report, along with guidance on model building and preliminary data analysis.

Aryan Dalal
Aryan led the effort to analyze the "great coach effect", along with providing mathematical preliminaries for our model. Aryan also helped to consolidate and format our findings into our presentation through beamer and LaTeX.

Jiayi Cora Guo
Cora led to effort to evaluate our model results and organized figures for this report, along with implementing the rounding strategy.

Julie Quan
Julie compiled our summary sheet, along with organizing our key findings in our conclusion section. Julie also aided in model building and data analysis.

Shani Shaham
Shani spearhead the model building process with help from all group members. Shani also drafted the methods section in this report.

Acknowledgment

The authors of this paper would like to thank Professor of Mathematics, Mason A. Porter[1] and Graduate Teaching Assistant, Lyu Haoyang[2] for their inputs during the project timeline. They were an invaluable resource during our modeling process.

We made use of external resources up to and including past attempts at modeling medal prediction for the Olympics games. These results helped us in gaining insight into what has been attempted so far and what could be done more comprehensively to reach a feasible conclusion. While there were no explicit results used from these references, they were helpful in providing information about our modeling strategy.

We also utilized Generative Language Models to answer simple questions such as, “What possible model domains can be used to predict olympic medal count”–throughout the modeling process, we refrained from using Language Models such as ChatGPT to attempt the modeling, rather, our use cases involved simplifying our assumptions and cleaning data.

Authors
Sofia Bonyadi, sbonyadi04@ucla.edu
Aryan Dalal, aryandalal@ucla.edu
Jiayi Cora Guo, jguo33@ucla.edu
Julie Quan, julieq03@ucla.edu
Shani Shaham, sshaham512@ucla.edu

Names listed in alphabetical order of surname.

Committee In Charge
Mason A. Porter, *Professor of Mathematics*, mason@math.ucla.edu
Haoyang Lyu, *Graduate Teaching Assistant*, lyuhaoyang@math.ucla.edu

UCLA Department of Mathematics
Mathematical Sciences Building
520 Portola Plaza, MS 6363
Los Angeles, CA 90095

[1]Department of Mathematics, University of California, Los Angeles
[2]Department of Mathematics, University of California, Los Angeles

8 AI Use Report

1 OpenAI ChatGPT (May 16 2025, ChatGPT 4o)
Query: What possible model domains can be used to predict Olympic medal count?
Output: "There are several model domains commonly used to predict Olympic medal counts, including: Regression models (e.g., linear, Poisson), Classification models (e.g., logistic regression), Simulation-based models (e.g., Monte Carlo methods), Time-series analysis, Machine learning approaches (e.g., random forests, XGBoost), Bayesian hierarchical models. Each domain offers different trade-offs in terms of interpretability, scalability, and data requirements."

2 OpenAI ChatGPT (May 21 2025, ChatGPT 4o)
Query: How should we filter the events list to exclude events with insufficient participation data?
Output: "To filter events with sparse data: (1) Set a threshold for minimum number of countries or athletes per event (e.g., greater than 10). (2) Exclude events missing data in multiple years or with inconsistent reporting. (3) Use participation ratios (e.g., countries participating / total possible). (4) Consider excluding events that are newly added or discontinued frequently. These filters help maintain model stability and reduce noise from under reported events."

3 OpenAI ChatGPT (May 24 2025, ChatGPT 4o)
Query: How can we incorporate historical bias from host countries in data preprocessing?
Output: "To address host country bias: (1) Add a binary feature indicating whether a country is the host in a given year.(2) Normalize medal counts using z-scores or percent changes from non-host years.(3) Exclude host years in baseline trend models to assess overperformance.(4) Compare simulations with and without host effects to quantify the differential impact."

4 OpenAI ChatGPT (May 30 2025, ChatGPT 4o)
Query: What is the best way to present expected value results from Monte Carlo simulations?
Output: "Effective ways to present expected value results include: Tables listing mean, median, and standard deviation of medals per country. Bar plots or heatmaps for visual comparison. Box plots showing distribution spread from simulation runs. Cumulative distribution functions (CDFs) to show medal outcome probabilities. It's also helpful to annotate plots with historical outcomes for context."

This document was typesetted with LaTeX

Appendix A: Simulation Code for LA2028 Medal Predictions

```
# --- PARAMETERS ---
n_simulations = 1000
n_athletes = 16
min_participation_threshold = 0.3
host_country = 'USA'
host_multiplier = 1.05  # 5% host advantage

# --- DECAY WEIGHTS ---
relevant_years = list(range(1992, 2025, 4))
year_decay = {year: np.exp(-0.15 * (2028 - year)) for year in relevant_years}

# --- WEIGHTED MEDALS ---
medal_data = athletes[
    (athletes['Year'].isin(year_decay)) &
    (athletes['Medal'].isin(['Gold', 'Silver', 'Bronze']))
].copy()
medal_data['Weight'] = medal_data['Year'].map(year_decay)

weighted_medals = (
    medal_data.groupby(['Event', 'NOC', 'Medal'])['Weight']
    .sum()
    .reset_index(name='WeightedCount')
)

gold_matrix = weighted_medals[weighted_medals['Medal'] == 'Gold'].pivot(index='Event', column
silver_matrix = weighted_medals[weighted_medals['Medal'] == 'Silver'].pivot(index='Event', co
bronze_matrix = weighted_medals[weighted_medals['Medal'] == 'Bronze'].pivot(index='Event', co

# --- PARTICIPATION MATRIX ---
recent_athletes = athletes[athletes['Year'] >= 1992]
participation = recent_athletes.groupby(['Event', 'NOC']).size().reset_index(name='AthleteCou
games_count = recent_athletes['Year'].nunique()
participation['AvgPerOlympics'] = participation['AthleteCount'] / games_count
country_participation = participation.pivot(index='Event', columns='NOC', values='AvgPerOlymp

# --- STEP 1 FIX ---
all_countries = country_participation.columns
for matrix in [gold_matrix, silver_matrix, bronze_matrix]:
    missing_cols = all_countries.difference(matrix.columns)
    for col in missing_cols:
        matrix[col] = 0
    matrix += 0.01
    matrix = matrix[sorted(matrix.columns)]

# --- EVENT LIST CLEANING ---
events_2024 = set(athletes[athletes['Year'] == 2024]['Event'].unique())
missing_gold = [e for e in events_2024 if e not in gold_matrix.index]
missing_silver = [e for e in events_2024 if e not in silver_matrix.index]
missing_bronze = [e for e in events_2024 if e not in bronze_matrix.index]
```

```
excluded_events = list(set(missing_gold + missing_silver + missing_bronze))
events_cleaned = [e for e in events_2024 if e not in excluded_events and e in country_partici

# --- MEDAL ASSIGNMENT FUNCTION ---
def get_weighted_winner(matrix, sampled_slots, event):
    if event not in matrix.index:
        return np.random.choice(sampled_slots)
    counts = pd.Series(sampled_slots).value_counts()
    eligible = [c for c in counts.index if c in matrix.columns]
    if not eligible:
        return np.random.choice(sampled_slots)
    base_weights = matrix.loc[event, eligible] * counts[eligible].values
    if host_country in base_weights.index:
        base_weights[host_country] *= host_multiplier
    weights = base_weights ** 2
    if weights.sum() == 0:
        return np.random.choice(eligible)
    probs = weights / weights.sum()
    return np.random.choice(eligible, p=probs)

# --- SIMULATION LOOP ---
all_medals_per_sim = []
for sim in range(n_simulations):
    medals_this_sim = []
    for event in events_cleaned:
        p = country_participation.loc[event]
        p = p[p > min_participation_threshold].clip(upper=3.0)
        if len(p) < 3:
            continue
        if host_country in p.index:
            p[host_country] *= host_multiplier
        probs = p.values / p.sum()
        countries = p.index.tolist()
        raw_slots = np.random.choice(countries, size=n_athletes, replace=True, p=probs)
        slot_counts = pd.Series(raw_slots).value_counts().clip(upper=3)
        slots = []
        for country, count in slot_counts.items():
            slots.extend([country] * count)
        if len(set(slots)) < 3:
            continue
        gold = get_weighted_winner(gold_matrix, slots, event)
        silver = get_weighted_winner(silver_matrix, slots, event)
        bronze = get_weighted_winner(bronze_matrix, slots, event)
        medals_this_sim.extend([
            {'Country': gold, 'Medal': 'Gold'},
            {'Country': silver, 'Medal': 'Silver'},
            {'Country': bronze, 'Medal': 'Bronze'}
        ])
    medal_df_sim = pd.DataFrame(medals_this_sim)
    medal_counts_sim = medal_df_sim.groupby(['Country', 'Medal']).size().unstack(fill_value=0
    medal_counts_sim['Total'] = medal_counts_sim.sum(axis=1)
    all_medals_per_sim.append(medal_counts_sim)
```

```
# --- BUILD ARRAY ---
all_countries_set = set()
for df in all_medals_per_sim:
    all_countries_set.update(df.index)
country_list = sorted(all_countries_set)
medals = ['Gold', 'Silver', 'Bronze', 'Total']
medal_array = np.zeros((n_simulations, len(country_list), len(medals)))
for i, df in enumerate(all_medals_per_sim):
    for j, country in enumerate(country_list):
        if country in df.index:
            for k, medal in enumerate(medals):
                medal_array[i, j, k] = df.loc[country].get(medal, 0)

# --- CALCULATE MEAN AND CONFIDENCE INTERVAL ---
mean_medals = medal_array.mean(axis=0)
std_medals = medal_array.std(axis=0)
ci95 = 1.96 * std_medals / np.sqrt(n_simulations)

# --- FORMAT OUTPUT ---
mean_df = pd.DataFrame(mean_medals, index=country_list, columns=medals)
ci_df = pd.DataFrame(ci95, index=country_list, columns=[f"{m}_CI95" for m in medals])
mean_df['Total'] = mean_df[['Gold', 'Silver', 'Bronze']].sum(axis=1)
medal_with_ci_sorted = pd.concat([mean_df, ci_df], axis=1).sort_values('Total', ascending=Fal

# --- CHECK FIRST-TIME MEDALISTS ---
past_medalists = athletes[
    (athletes['Year'] < 2028) &
    (athletes['Medal'].isin(['Gold', 'Silver', 'Bronze']))
]['NOC'].unique().tolist()
simulated_2028_medalists = medal_with_ci_sorted[medal_with_ci_sorted['Total'] > 0].index.toli
first_time_medalists = sorted(set(simulated_2028_medalists) - set(past_medalists))
```