详细方案设计

一、数据预处理与特征工程

1. 数据整合

- 将四个数据集通过"年份"和"国家"字段关联，构建统一分析单元（国家-年份级面板数据）。
- 补充关键衍生特征：
  - **东道主标识**（来自_hosts.csv，当前届次主办国=1）
  - **项目多样性指数**（基于_programs.csv计算各届项目种类的Shannon熵）
  - **运动员参与度**（从_athletes.csv统计每国每届参赛人数）
  - **项目突变标记**（若某届新增/取消项目数超过历史平均3σ，标记为1）

2. 零值国家筛选

- 定义"零值国家"：历史上从未获得任何奖牌
- 构建二元分类标签：下一届是否可能突破零奖牌（根据历史突变事件定义正样本）

二、第一阶段：零膨胀模型（Zero-Inflated Model）

1. 模型选择依据

- 使用Augmented Dickey-Fuller检验验证奖牌数时间序列的平稳性
- 通过Vuong检验比较零膨胀泊松（ZIP）与零膨胀负二项（ZINB）的拟合优度
- 最终选择ZINB模型（通常更适合过离散数据）

2. 特征设计

- 核心预测因子：
  - 近期参赛人数增长率（3届滑动窗口）
  - 同区域邻国的奖牌突破事件（空间滞后项）
  - 新增项目与本国优势项目的匹配度（需结合_athletes.csv中的运动项目历史表现）

3. 输出处理

- 输出为概率值p，设定阈值θ（通过Youden指数确定），当$p>θ$时判定可能突破零奖牌

三、第二阶段：混合预测模型

1. 时间序列基模型构建

- 使用时间序列回归模型预测奖牌数量，考虑什么滑动窗口那些
- 关键外部变量：
  - 滞后3届奖牌数的指数衰减加权平均
  - 东道主效应的时变强度系数（通过历史数据估计主办国平均增益）
  - 项目数量变化与本国优势项目的交互项
  - ……

2. 残差学习机制

- 构造残差数据集：$R_t = Y_t - \hat{Y}_t^{TS}$，其中$Y_t$为实际奖牌数
- XGBoost输入特征：
  - 基模型预测值及其置信区间宽度
  - 非线性特征：
    - 优势项目突变强度（优势项目数变化 × 历史占比）
    - 运动员新人比率（首次参赛者占比）
    - ……
- 特征筛选：通过SHAP值分析保留|SHAP|>0.01的特征

3. **预测合成技术**

- 点预测：$\hat{Y} = \hat{Y}^{TS} + \hat{R}^{XGB}$
- 区间合成：
  $$U_{total} = \sqrt{(U_{TS}^2 + \sigma_{residual}^2)}$$
  其中U_{TS}为时间序列预测区间的半宽，$\sigma_{residual}$为XGBoost在验证集上的残差标准差

四、不确定性传播机制

1. **蒙特卡罗模拟**
   - 对时间序列模型进行1000次轨迹采样，每次采样结果作为XGBoost的输入特征
   - 在XGB预测阶段注入高斯噪声$N(0,\sigma_{residual})$
   - 最终预测区间取2.5%和97.5%分位数

五、验证策略

1. **时域交叉验证**

- 采用滚动窗口验证：以4届奥运会为窗口，每次滚动1届
- 评估指标：
  - 点预测：对称MAPE（处理零值问题）
  - 区间预测：区间覆盖概率（ICP）与平均间隔宽度（MSIS）

2. **对抗性测试**

- 构造虚拟国家的极端场景：
  - 突然成为东道主
  - 优势项目被移出奥运
  - 运动员数量激增300%
- 检验模型在这些场景下的响应是否符合领域知识