**AI 工具使用详情**

# 工具总览

表 1 所使用AI的相关信息

| **序号** | **工具名称** | **型号/版本** | **开发机构/公司** | **使用日期** |
| --- | --- | --- | --- | --- |
| AI-1 | ChatGPT | GPT-5 Thinking | OpenAI | 2025-09-4 |
| AI-2 | PyCharm AI Assistant（JetBrains） | Claude 4 Sonnet | Anthropic | 2025-09-5 |
| AI-3 | GitHub Copilot（Microsoft） | Claude 4 Sonnet | Anthropic | 2025-09-6 |

# 使用环节与合规判定

总体声明

本参赛队在竞赛期间**核心建模与分析全部由队员独立完成**；为提高文字表达与模型比较的规范性，我们**有限度地使用了大语言模型**进行**模型优劣对比的文字整理、数模假设与流程的语言润色、以及公式的规范化书写**。所有AI输出均**由队员复核与修改**，不得用于自动生成核心结果、调参决策或替代数据分析。

 **选题与文献检索:**AI协助确定选题并搜索相关文献线索，由团队谈论思考并作出决策。

 **建模方法筛选与每问流程规划:**AI提供备选数学模型与每问任务清单，由人工确定最终技术方案。

 **数据处理与可视化:**AI生成数据清洗、统计及可视化的脚本初稿，由人工改写与校验。

 **代码调试与错误定位:**AI辅助定位依赖、路径或语法问题，由人工完成最终修复与测试。

 **结果分析与图表说明:**AI生成初步的分析摘要和图注草稿，由人工进行事实核对与改写。

 **论文润色与排版:**AI用于语言表达优化和图表格式微调，不修改任何核心数据与结论。

# 关键交互记录

## 选题

**参赛队输入（提炼后的提示词）**：“我们列出历年训练题（切割/干扰/灌溉/疾病预测等），请按难度与团队适配度评估今年 A/B/C 三题，并给出各题可行方法栈与风险点，用于核验我们对 C 题的预判是否存在偏差。”

**AI核心回复（≤6行要点）**：

1. **结论**：C**最稳**、A**可打**、B**谨慎**。
2. **A（烟幕）**＝3D相对运动+几何遮挡+多机多弹**时间窗调度**（可MILP/启发式），目标最大化遮蔽时长；先做遮挡评估器再排程。
3. **B（外延层）**＝**色散+多光束（Airy）+双角度联合拟合**，厚度与折射率强耦合、对初值极敏感，容错低。
4. **C（NIPT）**＝**区间删失生存**（T4% 达标时刻）+**GAMM/Beta**估计+**最优分箱/DP**给BMI组别与时点；Q4女胎用**Logistic/L1**基线+**XGBoost+校准**+**SHAP**。
5. **偏差校准**：A不是“算距离”这么简单；B不是“代条纹公式”；C不是“回归分组”即可，需处理删失与重复测量。
6. **MVP 路线**：A先遮挡闭式评估→贪心/小规模MILP；B先FFT初值→双角度两束→必要时多束；C先生存面→组内DP与置信界→女胎判别与校准。

## 建模方法筛选与文献检索:

**队伍提问（选题后方案校验）**：Q1用GAMM做相关与显著性；Q2用logistic-GAMM+（MLE/贝叶斯）+DP分组+-保障+期望风险；Q3在Q2基础上做多因素+误差；Q4用XGBoost。此路线能否很好解决？

**AI要点答复**：总体可行；建议把Q2的“logistic回归”统一改成**离散时间生存GAMM**（person-period）直接得到，并在DP中以-保障+最小样本+惩罚防过切。

## 数据处理与可视化:

* **队伍提问（摘要）**：为保证可复现，需独立重做数据清洗与基础可视化；请提供一键脚本与产出清单，我们本地跑全流程。
* **AI 应答（要点）**：给出q1\_clean.py一键脚本：
  1. 读取原始 Excel（男女胎两个表），统一列名与类型；
  2. 解析“孕周（周+天）→连续周数”；
  3. 计算体质指数列与孕妇ID标准化；
  4. **男胎**生成 Y浓度\_Beta调整（SV 边界微调，供 Beta 回归/GAMM）；
  5. 质控指标做 **QC-PCA** 得检测质量主成分1/2；
  6. 标注边缘样本；
  7. 导出**清洗结果\_Q1.xlsx**（男女胎清洗表）、**清洗\_描述统计\_Q1.xlsx**、以及同名 CSV/Parquet。并说明后续 Q1 GAMM 的输入/交叉验证分组方式与可视化接口。

## 代码调试与错误定位

**队伍提问：**  
我们运行 q1\_clean.py 的时候，脚本在写 Parquet 的地方报错中断，清洗结果\_Q1.xlsx 能生成，但 **描述统计 Excel 没有生成**。  
报错核心信息：

ArrowInvalid: Could not convert '≥3' ... Conversion failed for column 怀孕次数

我们怀疑是不是缺少某些库导致的？

**AI回答：**

* **根因定位**：不是缺库问题，你们的 pandas/numpy/scikit-learn/openpyxl 都齐全。真正的原因是 **列「怀孕次数」里混有字符串（如 '≥3'）和数值**，pyarrow 在写 Parquet 时强制要整列是 int64，结果无法序列化，导致脚本在写 Parquet 这一步崩溃，后续导出中断。
* **现场修复（稳定版补丁）**：
  1. **输出顺序调整**：先写 CSV & Excel（主产物），最后再尝试写 Parquet；
  2. **类型护栏**：新增 safe\_to\_parquet()，在写入前将所有 object 列统一转为 string dtype，失败则捕获并跳过，不影响主流程；
  3. **可选依赖**：若仍希望产出 Parquet，可安装 pyarrow 或 fastparquet；
  4. **可观测性**：控制台打印 QC-PCA 解释率、缺失率，并明确提示 Parquet 成功/跳过。
* **复测结果**：修复后主产物（清洗结果\_Q1.xlsx、清洗\_描述统计\_Q1.xlsx、CSV）均可生成；Parquet 若再遇异值，只会提示并跳过，不再影响主流程。
* **遗留风险提醒**：如果后续要把**「怀孕次数」**作为数值变量用在建模中，建议在清洗时显式映射（例如 '≥3'→3 或直接分箱为有序分类），避免模型阶段出现隐式转换问题。

## 结果分析与图表说明

**队伍提问：**

这是改了吗？我怎么感觉图还是老样子？260 多个孕妇，不可能一点堆都不扎吧？还是说目前这样就是最合理的？孕妇的 BMI 会呈现正态分布，所以出现黄色的几个其实集中度很高？

**AI回答：**

* **原因解释**：不是模型坏掉，而是因为观测周太离散 + 我们当时没有补齐中间周。男胎样本的检测周集中在 12、13、16、20、26 周这些点，所以 person-period 风险集只会在这些整数周有人，掩膜阈值一筛，就只剩几条“黄杠”。
* **调整建议**：
  1. **补齐中间周**：在首次达标之前，把中间缺的周也加上，BMI 用 LOCF（前推）或线性插值。这样风险集人数会在更多周数上出现，不会只剩稀疏的几条带。
  2. **放宽掩膜阈值**：把 BMI 分箱调大一点（1.0→2.0）、邻域窗口放宽（±2 周、±2 档）、最低支持人数从 8 降到 4–6，就能看到更连贯的“块状”有效区。
  3. **解释合理性**：BMI 本来接近正态分布，中间区间样本多，所以黄色高支撑带集中在那儿是合理的；只是周数稀疏让整体看着“不成片”。
* **小结**：图看着“怪”，根本原因是 **采样周稀疏 + 掩膜阈值苛刻**；解决方法是 **补周 + 放宽窗口**，结果会更平滑、更合理。

## 论文润色与排版环节

| **任务** | **主要提示词（摘要）** | **工具要点输出** | **采纳与人工修改** |
| --- | --- | --- | --- |
| 行文润色 | “压缩‘变量选择策略’段落，保持技术细节不删减” | 给出更简洁的表述，并统一了专业术语 | 我们逐句核对，保留简化语句，删除与实验结论不符的措辞 |
| 符号规范 | “将Beta-logit GAMM的模型写为统一符号，并给出连接函数与方差项” | 输出了 μ、ϕ、logit 的统一记号，公式排版规范化 | 我们对照核查，确认符号一致，少量参数符号手工微调 |
| 叙述润色 | “调整结果分析部分语气，避免‘绝对化结论’，改为统计置信口径” | 建议将‘显著影响’改为‘在显著性水平下有统计学证据’ | 我们采纳语气修饰，保证结论表述合规且更符合审稿标准 |

**提问：**帮我润色一下，尽量不要大幅度改动我的论文，要求风格与前文一致

**GPT回答**（截取部分内容）：

目标是在不增加采样成本的前提下，构建**女胎异常判定器**并给出**可复核、可解释、可部署**的流程。我们使用的特征仅包含与标注无直接因果联动的观测量：染色体 Z 值（13/18/21/X）、测序质控（GC 含量、总读段 reads、有效读段比例 reads\_ratio）、孕妇 BMI 与年龄等；以**孕妇ID分组**做K折交叉验证，阻断同一受试者的“数据泄漏”。异常样本占比偏低，训练时采用**代价感知/带权对数损失**以凸显少数类。为保障可解释性，构建可解释的GAM作为**基线模型**，并以**SHAP（TreeSHAP）**解释主模型的特征贡献。

**提问**（除此之外还询问了GPT一些其他的建议）：模型评价这一部分的表格后面就没话了，有一种戛然而止的感觉，是不是在表格后面添加一句话更合适？

**GPT回答：**

* 是的，加一句“收尾 + 承上启下”的话最顺。位置：在“模型评价 → 第一问模型评价”里，表 3（分层交叉验证指标）紧跟其后，新起一段，粘贴下面这一句：
* 综上，按孕妇 ID 分层的 5 折交叉验证表明模型在不同子集上泛化稳定（测试集 MAE=0.02790±0.00182、Spearman ρ=0.148±0.028；各折样本量均衡：训练 866±8、测试 216±8），无显著折间波动，因而可据此在后续章节开展达标概率与时点决策。
* 结论：5 折 CV 稳定（MAE=0.02790±0.00182，ρ=0.148±0.028），样本量均衡，无折间异常——足以支撑后续基于该模型的决策分析。

**提问**（此外GPT也帮助我理解了其他的文件材料）：这是所有的数据，这里男胎与女胎清洗版就是进行缺失值处理以及标准化吗？

**GPT回答：**

没问题，给你一份“文件→用途”一屏看懂。先回答核心疑问：“男胎\_清洗版\_Q1.csv / 女胎\_清洗版\_Q1.csv 是不是只做缺失值处理和标准化？”  
不止。它们一般同时完成：①缺失值处理（BMI用体重/身高²回填等）②单位统一（孕周=周+天/7；kg、m）③SV映射把V映到(0,1)④质控筛样（GC、比对率、读段等）⑤（可选）对J,K做中心化/标准化。男胎版用来做Q1建模；女胎版多用于对照/附录，Q1本体不参与建模。

# 采纳/不采纳清单

## 采纳说明（选题阶段）

我们团队在充分讨论后，采纳了 AI 工具关于选题的分析意见，最终坚定选择**C题（NIPT）**。原因主要有三点：

### 对比分析

* B 题虽然表面只有三问，看似“物理光学+干涉公式”很直观，但我们敏锐地注意到第 3 问多光束干涉+色散建模极其复杂，容易让队伍“做到一半崩盘”。
* A 题的结构，我们联想到 2024 国赛 A 题“板凳龙”：同样是几何判定+时序调度的大工程，虽然可做，但代码量大、验证成本高，注定人力消耗重。
* C 题虽然带有医学背景，数据量大、信息显得“唬人”，但本质是我们熟悉的 生存分析+分组优化+判别模型，路径成熟，稳健可控。

### 博弈考量

我们预测：很多队伍会被 B 题“短小易懂”的表象吸引而扎堆，但最后大概率掉坑；部分队伍会因 A 题与“板凳龙”相似而投入，但实现难度大，竞争激烈；而 C 题会让不少队伍望而却步（因生物学背景与大数据量吓退），从而减少对手数量。因此，选择 C 题不仅稳，还能形成 **差异化竞争优势**。

### 团队共识

我们团队一旦确定方向就不会再更改，所以没有“备选”概念。综合自身特长（疾病预测、风险建模经验）与 AI 的辅助判断，我们一致认为 C 题是 **既稳妥又能发挥优势的唯一正确道路**。

## 采纳说明（C题方法路线）

* **Q1**：采纳AI提议，维持Beta-GAMM框架，结构选用，并执行按孕妇分组的交叉验证与模型诊断（k-index、concurvity），保证显著性检验与稳健性。
* **Q2**：采纳将Logistic-GAMM 统一为离散时间生存 GAMM 的建议，结合-保障与期望风险最小化；在 BMI 维度采用 DP 自适应分组，并加入惩罚项与最小样本约束，以避免过切。
* **Q3**：采纳在 DP 计算中引入 Monte-Carlo 抽样处理组内多因素与检测误差的建议，以不确定性惩罚（置信下界/后验分位）替代点估计，增强鲁棒性。
* **Q4**：采纳“基线 Logistic(L1)+XGBoost 强化”的双模型策略，执行按孕妇/批次分组的交叉验证，处理类不平衡（scale\_pos\_weight），并进行概率校准与可靠性曲线绘制，同时利用 SHAP 实现可解释性；报告指标统一为“固定 FPR 与代价最小阈值”。
* **风险清单**：采纳 AI 强调的关键约束——防止样本泄漏、比例数据避免误设为高斯、DP防止过切（+最小）、阈值必须在校准后概率上选择——并全部纳入实现与审核流程。

## 采纳说明（数据处理与可视化）

* **统一清洗脚本**：采纳“一键式 Python 脚本”方案，固定输入/输出路径，确保复现实验可追踪（文件指纹可核）。
* **孕周解析**：采纳“多格式正则解析→连续周数”策略，消除源数据书写差异对建模的影响。
* **比例边界处理**：采纳**Smithson–Verkuilen**微调，生成 Y浓度\_Beta调整，与后续 **Beta(link=logit)** 家族对齐。
* **质控降维**：采纳 **QC-PCA(PC1/PC2)** 合成质量主成分，替代多重共线 QC 原始列，便于可视化与稳健建模。
* **缺失与边缘标注**：采纳“缺失率报告 + 边缘样本标记（不删，只标）”，为后续敏感性分析与掩膜可视化做准备。
* **导出规范**：采纳“双工作簿”产出（清洗总表 + 描述统计）并冗余 CSV/Parquet，以便 Python/R/Matlab 跨环境复用。
* **可视化衔接**：采纳“清洗→网格/透视表接口”的组织方式，直接支撑后续等高线图、热图与风险集图的绘制（不在清洗脚本内硬编码图形，保持职责单一）。
* **验证与日志**：采纳“控制台摘要 + 关键比率打印”的做法，便于队内同学快速核对清洗是否按预期执行。

注：以上采纳严格限定在“数据处理与可视化（基础）”环节；后续 Q1–Q4 建模与图形产出调用此清洗结果作为唯一数据源，避免多版本数据造成审计难题。

## 采纳说明（代码调试与错误定位）

* **采纳-1｜输出顺序保护**：固定“**先 CSV/Excel、后 Parquet**”流程，确保主产物不再因 Parquet 失败而丢失。
* **采纳-2｜类型统一**：在写 Parquet 前对 **全部 object 列统一转为 string**，规避混型崩溃；建模前再按需**显式数值化**。
* **采纳-3｜容错与日志**：对 Parquet **try/except** 兜底，并在控制台输出**明确的成败与原因**，便于审计与复现。
* **采纳-4｜最小依赖**：不再误判为“缺库问题”；仅在需要 Parquet 时建议安装 pyarrow/fastparquet，其余依赖保持当前版本即可。
* **采纳-5｜可复现承诺**：将“稳定版脚本”作为**唯一清洗入口**，锁定输入/输出文件名，支撑后续 Q1–Q4 全部图表与模型复现。