题目匹配准确率评估

好的！让我详细列出完善评估系统需要做的具体事情，不涉及代码，纯粹是操作清单：

## 📋 完善评估系统操作清单

### ****阶段一：建立评估标准 (人工工作)****

#### 1. ****创建金标准数据集****

* **任务**: 人工标注每道题目与每个知识点的相关性
* **具体操作**:
  + 拿出您的11道测试题目
  + 对每道题，人工判断它与"概率的统计定义"的相关程度 (0-5分)
  + 对每道题，人工判断它与"实验和结果"的相关程度 (0-5分)
  + 记录判断理由
* **时间**: 约30-45分钟
* **输出**: 一个标注表格

#### 2. ****定义相关性评分标准****

* **5分**: 高度相关，直接考察该知识点
* **4分**: 较相关，需要该知识点作为基础
* **3分**: 中等相关，涉及该知识点的某些概念
* **2分**: 弱相关，有一定联系但不是重点
* **1分**: 几乎无关，仅有表面联系
* **0分**: 完全无关

### ****阶段二：多维度评估指标 (分析工作)****

#### 3. ****精确率与召回率分析****

* **精确率**: 推荐的题目中，有多少是真正相关的？
* **召回率**: 所有相关的题目中，推荐了多少？
* **F1分数**: 精确率和召回率的综合指标
* **分析维度**: 整体 + 按知识点分别计算

#### 4. ****推荐排序质量评估****

* **NDCG@5**: 前5个推荐的排序是否合理？
* **MAP**: 每个位置的推荐质量如何？
* **位置偏差**: 相关题目是否排在前面？

#### 5. ****覆盖度分析****

* **题目覆盖**: 是否遗漏了重要的相关题目？
* **知识点覆盖**: 每个知识点的推荐是否充足？
* **类型覆盖**: 不同difficulty、不同topic的题目覆盖如何？

### ****阶段三：细粒度质量分析 (诊断工作)****

#### 6. ****推荐质量分级****

* **EXCELLENT**: 算法高分 + 人工高分 (>=4分)
* **GOOD**: 算法中分 + 人工中高分 (3-4分)
* **ACCEPTABLE**: 算法低分 + 人工中分 (2-3分)
* **POOR**: 算法推荐但人工低分 (1-2分)
* **WRONG**: 算法推荐但人工0分

#### 7. ****错误案例分析****

* **误推荐**: 哪些不相关题目被错误推荐了？为什么？
* **漏推荐**: 哪些相关题目被遗漏了？原因是什么？
* **排序错误**: 哪些题目的排序不合理？

#### 8. ****算法表现诊断****

* **关键词匹配**: 在什么情况下有效/失效？
* **Topic标签**: 哪些标签组合容易误判？
* **题目类型**: 哪类题目(纯概率/混合/几何等)匹配效果好/差？

### ****阶段四：阈值优化分析 (调优工作)****

#### 9. ****最优阈值寻找****

* 测试不同匹配分数阈值 (1分、2分、3分...)
* 分析每个阈值下的精确率、召回率
* 找到F1分数最高的阈值
* 考虑实际应用场景 (宁可多推荐 vs 宁可少推荐)

#### 10. ****参数敏感性分析****

* 关键词匹配权重 vs Topic匹配权重
* 不同知识点是否需要不同的匹配策略？
* 题目长度、难度是否影响匹配效果？

### ****阶段五：对比基准测试 (验证工作)****

#### 11. ****基准对比****

* **随机推荐**: 随机推荐题目的效果如何？
* **纯关键词**: 只用关键词匹配的效果？
* **纯Topic**: 只用Topic标签的效果？
* **当前算法**: 关键词+Topic的综合效果？

#### 12. ****不同数据集测试****

* 在11道测试题上的表现
* 在真实AMC题库上的表现 (如果可用)
* 跨领域迁移：数学题目 vs 其他学科题目

### ****阶段六：改进方向指导 (决策工作)****

#### 13. ****瓶颈识别****

* 当前最大的问题是什么？(精确率低/召回率低/排序差)
* 哪类错误最需要优先解决？
* 哪个环节的改进能带来最大提升？

#### 14. ****改进优先级****

* **高优先级**: 影响大、实现容易的改进
* **中优先级**: 影响大但实现复杂的改进
* **低优先级**: 边际改进或特殊场景优化

#### 15. ****下一步行动计划****

* 具体需要改进哪个算法模块？
* 需要收集什么额外数据？
* 应该先实现语义匹配还是LLM验证？

## 🎯 实施建议

### ****最小可行版本 (今天就能做)****

1. **人工标注11道题目** (30分钟)
2. **计算基础精确率/召回率** (分析现有结果)
3. **识别3-5个最明显的问题**

### ****完整版本 (本周完成)****

1. 实施上述15个步骤的全部内容
2. 生成详细的评估报告
3. 制定具体的改进路线图

### ****持续优化版本 (长期)****

1. 建立评估流水线，每次算法更新后自动评估
2. 收集用户反馈，不断完善金标准
3. A/B测试不同算法版本的实际效果

## 📝 您需要准备的材料

1. **11道测试题目清单** (您已有)
2. **2个知识点描述** (您已有)
3. **30-45分钟时间** 进行人工标注
4. **Excel表格或笔记本** 记录标注结果

您想从哪个阶段开始？我建议先做最小可行版本，快速获得改进方向！