智能路由系统数据报告

生成时间：2025-04-17 20:15:00

1. 系统概述

1.1 系统状态

- 初始化时间：2025-04-17 20:01:25
- 运行状态：正常
- 主服务端口：3100
- 数据库位置：/Users/xinghailong/Documents/solana_MEVbot/intelligent_routing/data/routing_data.db

1.2 硬件配置

- 处理器：Apple M4
- GPU内存：16383 MiB free
- 系统类型：macOS

2. 模型配置

2.1 可用模型

- OpenAI: gpt-4o, gpt-4.1 (已配置)
- DeepSeek: deepseek-v3.1, deepseek-r1 (已配置)
- Gemini: gemini-pro, gemini-pro-vision, gemini-1.5-flash (已配置)
- Claude: 通过Cursor企业版访问
- Llama3: 本地部署（8B参数版本）

2.2 Llama3模型详情

- 参数规模：8B
- 上下文长度：8192 tokens
- 加载时间：2.01秒
- 平均响应时间：15-16秒
- 量化方式：Q4_0
- 文件大小：4.33 GiB

3. 数据收集

3.1 数据库结构

- queries：查询记录表
- routing_decisions：路由决策表
- user_feedback：用户反馈表
- model_performance：模型性能表

3.2 收集数据类型

1. 查询记录：
    - 查询内容（MD5哈希）
    - 语言类型
    - 领域分类
    - 复杂度评分
2. 路由决策：
    - 选择的模型
    - 置信度
    - 选择理由
    - 响应时间
3. 用户反馈：
    - 评分（1-5分）
    - 评论内容
4. 模型性能：
    - 成功率
    - 平均响应时间
    - 用户评分

4. 隐私保护

4.1 数据安全措施

- 查询内容使用MD5哈希存储
- 不存储用户身份信息
- 只记录必要的性能数据
- 数据本地存储，不发送到外部服务器

4.2 访问控制

- API密钥集中管理
- 端口访问控制
- 系统资源保护
- 异常处理机制

5. 系统性能

5.1 响应时间

- Llama3加载时间：2.01秒
- 平均响应时间：15-16秒
- 最大上下文长度：8192 tokens

5.2 资源使用

- GPU内存：16383 MiB free
- 模型缓存：1024.00 MiB

- 计算缓冲区：560.00 MiB

6. 建议与优化

6.1 短期优化

- 增加请求缓存
- 优化日志记录
- 完善错误处理

6.2 中期改进

- 实现负载均衡
- 添加监控告警
- 优化资源使用

6.3 长期规划

- 分布式部署
- 自动扩缩容
- 智能调度系统

7. 系统限制

7.1 当前限制

- 依赖Cursor企业版访问Claude
- 本地资源有限（M4芯片）
- 单点部署风险
- 缺乏自动扩缩容

7.2 注意事项

- 定期检查系统状态
- 监控资源使用情况
- 及时更新模型配置
- 定期备份数据