评估

RAG 模型抽取结果评估分析报告

生成时间: 2025-10-17 15:12:07

1. 模型对比汇总

模型	评估文件数	实体总数	实体正确	实体错误	实体准确率(%)	关系总数	关系正确	关系错误	关系准确率(%)
DeepSeek	10	289	282	7	97.58	234	217	15	92.74
Gemini	10	304	297	7	97.7	225	221	4	98.22
Kimi	10	163	158	5	96.93	143	122	21	85.31

2. 关键发现

- **实体抽取最优**: Gemini (97.7%)
- **关系抽取最优**: Gemini (98.22%)

3. 各模型详细分析

DeepSeek

- 实体准确率: 97.58%
- 关系准确率: 92.74%
- 实体总数: 289
- 关系总数: 234
- 详细数据: `deepseek/paper_details.csv`

Gemini

- 实体准确率: 97.7%
- 关系准确率: 98.22%
- 实体总数: 304
- 关系总数: 225
- 详细数据: `gemini/paper_details.csv`

Kimi

- 实体准确率: 96.93%
- 关系准确率: 85.31%
- 实体总数: 163
- 关系总数: 143
- 详细数据: `kimi/paper_details.csv`

4. 文件说明

```
evaluation_results/
├── evaluation_summary.md     # 本文件（汇总报告）
├── deepseek/
│   └── paper_details.csv     # DeepSeek 每篇论文的详细结果
├── gemini/
│   └── paper_details.csv     # Gemini 每篇论文的详细结果
└── kimi/
    └── paper_details.csv     # Kimi 每篇论文的详细结果
```

5. 改进建议

1. 分析错误实体和关系的具体类型，针对性优化 Prompt
2. 对准确率较低的实体/关系类型增加更多 few-shot 示例
3. 人工复核错误案例，调整 schema 定义
4. 考虑使用集成方法结合多个模型的优势