🔍 三种评价方法对比报告

生成时间: 2025-08-08 21:14:31 **对比团队数**: 15 个 **评价方法**: 人工评价 vs 机器评价 vs LLM评价

🎯 关键发现

相关性分析

- **人工 vs LLM**: 0.888 (高度相关)
- **人工 vs 机器**: 0.167 (低相关)
- **机器 vs LLM**: -0.099 (负相关)

主要结论

- **LLM评价与人工评价高度一致**，相关系数达到0.888
- **机器评价与人工评价差异较大**，相关系数仅为0.167
- **机器评价与LLM评价呈负相关**，说明评价标准存在根本差异

📊 详细排名对比

团队名称	人工排名	人工分数	机器排名	机器分数	LLM排名	LLM分数	人工-机器差异	人工-LLM差异
Team cake	1	0.91	19	0.468	1	0.828	+18	0
mkrisnai	2	0.89	16	0.475	3	0.802	+14	+1
Team MarSan_AI	3	0.89	12	0.504	10	0.680	+9	+7
erehulka	4	0.88	4	0.543	7	0.713	0	+3
SomethingAwful	5	0.86	7	0.522	4	0.790	+2	-1
Team_SINAI	6	0.85	26	0.413	2	0.825	+20	-4
Team NLPunks	7	0.84	15	0.489	9	0.680	+8	+2
gleb.shnshn	8	0.74	20	0.462	8	0.698	+12	0
Team Iron Autobots	9	0.74	29	0.345	5	0.772	+20	-4
ZhongyuLuo	10	0.73	11	0.506	13	0.595	+1	+3
backtranslation_baseline	11	0.73	10	0.506	12	0.605	-1	+1
nikita.sushko	12	0.70	3	0.553	14	0.525	-9	+2
VitalyProtasov	13	0.69	6	0.531	11	0.623	-7	-2
mt5_baseline	14	0.68	24	0.418	15	0.497	+10	+1
delete_baseline	16	0.47	22	0.447	17	0.259	+6	+1

🏆 前5名对比

人工评价前5名

1. **Team cake** - 分数: 0.91
2. **mkrisnai** - 分数: 0.89
3. **Team MarSan_AI** - 分数: 0.89
4. **erehulka** - 分数: 0.88
5. **SomethingAwful** - 分数: 0.86

机器评价前5名

1. **nikita.sushko** - 分数: 0.553
2. **erehulka** - 分数: 0.543
3. **VitalyProtasov** - 分数: 0.531
4. **SomethingAwful** - 分数: 0.522
5. **backtranslation_baseline** - 分数: 0.506

LLM评价前5名

1. **Team cake** - 分数: 0.828
2. **Team_SINAI** - 分数: 0.825
3. **mkrisnai** - 分数: 0.802
4. **SomethingAwful** - 分数: 0.790
5. **Team Iron Autobots** - 分数: 0.772

🎯 一致性分析

三种评价都在前50%的团队 (2 个)

- **SomethingAwful**: 人工第5名, 机器第7名, LLM第4名
- **erehulka**: 人工第4名, 机器第4名, LLM第7名

⚠ 显著差异分析

人工评价 vs 机器评价显著差异 (差异≥10位)

- **Team cake**: 人工第1名 vs 机器第19名 (机器评价更低18位)
- **mkrisnai**: 人工第2名 vs 机器第16名 (机器评价更低14位)
- **Team_SINAI**: 人工第6名 vs 机器第26名 (机器评价更低20位)
- **gleb.shnshn**: 人工第8名 vs 机器第20名 (机器评价更低12位)
- **Team Iron Autobots**: 人工第9名 vs 机器第29名 (机器评价更低20位)
- **mt5_baseline**: 人工第14名 vs 机器第24名 (机器评价更低10位)

人工评价 vs LLM评价显著差异 (差异≥5位)

- **Team MarSan_AI**: 人工第3名 vs LLM第10名 (LLM评价更低7位)

📈 各评价方法特点分析

人工评价

- **最高分**: Team cake (0.91)
- **最低分**: delete_baseline (0.47)
- **平均分**: 0.77
- **标准差**: 0.12

机器评价

- **最高分**: nikita.sushko (0.553)
- **最低分**: Team Iron Autobots (0.345)
- **平均分**: 0.479
- **标准差**: 0.056

LLM评价

- **最高分**: Team cake (0.828)
- **最低分**: delete_baseline (0.259)
- **平均分**: 0.659
- **标准差**: 0.152

💡 结论和建议

主要结论

1. **LLM评价是人工评价的良好替代**: 两者相关系数高达0.888，说明LLM能够较好地模拟人工评价标准
2. **机器评价与人工评价存在根本差异**: 相关系数仅为0.167，说明机器评价可能采用了不同的评价维度
3. **评价方法的互补性**: 不同评价方法关注的重点不同，可以提供多角度的评价视角

建议

1. **优先使用LLM评价**: 在需要大规模自动化评价时，LLM评价可以作为人工评价的有效替代
2. **结合多种评价方法**: 综合考虑人工、机器、LLM三种评价结果，获得更全面的评价
3. **深入分析差异原因**: 对于评价结果差异较大的团队，需要进一步分析其原因