

# LLM人格推断实验分析报告

## 实验概述

本实验通过多轮对话测试大语言模型（LLM）推断角色Big Five人格特质的准确性。实验设计如下：

- **角色模拟器**: 智谱CharGLM-4，根据Personality Database中的角色信息进行角色扮演
- **被测LLM**: ChatGPT (GPT-5.1)、Claude (Sonnet 4.5)、DeepSeek、Grok-3
- **测试角色**: 10个来自不同来源的虚构角色，具有已知的Big Five人格得分（Ground Truth）
- **对话模式**: 每对角色×LLM组合进行两种独立对话
  - **llm\_first** (LLM先发言): LLM作为“访谈者”主动提问，角色回答
  - **char\_first** (角色先发言): 角色主动开启对话，LLM作为“回应者”
- **对话轮次**: 每次对话10轮
- **评估指标**: MAE (平均绝对误差)，即5个人格维度推断误差的平均值
- **有效实验**: 80组 (10角色 × 4个LLM × 2种模式)，Gemini因API问题全部失败，未纳入统计

---

## 核心发现

### 一、LLM作为“访谈者” (llm\_first) 显著优于“回应者” (char\_first)

这是本实验最明确的结论。

模式	平均MAE	样本量
llm_first (LLM先发言)	29.8%	40
char_first (角色先发言)	34.5%	40

在40组配对比较中，llm\_first模式在**68%的情况下**（27/40）取得更低的MAE，平均优势为4.7个百分点。

**解释：**当LLM扮演“访谈者”角色时，它可以主动选择探测性问题（如“你如何处理与朋友的分歧？”“你是更喜欢计划还是随性？”），有策略地引导对话触及人格相关话题。而在char\_first模式下，LLM只能被动回应角色的话题，缺乏主动探测的空间。

各LLM在两种模式下的表现：

LLM	llm_first MAE	char_first MAE	差值
ChatGPT	30.9%	35.8%	+4.9%
Claude	28.8%	32.5%	+3.7%
DeepSeek	31.3%	37.2%	+5.9%
Grok	28.2%	32.6%	+4.4%

所有LLM在llm\_first模式下均表现更好，DeepSeek受影响最大（+5.9%），Claude受影响最小（+3.7%）。

## 二、四个LLM总体表现相近，Grok和Claude略领先

LLM	总体平均MAE	排名
Grok	30.4%	1
Claude	30.6%	2
ChatGPT	33.4%	3
DeepSeek	34.2%	4

最佳与最差之间仅相差3.8个百分点。在10个角色的样本量下，这一差异的统计显著性有限。实验显示，**角色本身的特征对推断难度的影响远大于LLM的选择**——最容易和最难的角色之间MAE差距高达40个百分点。

---

## 三、开放性（Openness）是推断误差最大的特质，且存在严重的系统性高估

这是最引人注目的发现。各维度的误差和偏差如下：

人格维度	平均绝对误差	系统性偏差（推断值 - 真实值）	偏差方向
外向性 (E)	18.4%	-4.4%	略微低估
神经质 (N)	37.6%	-8.9%	低估
尽责性 (C)	33.0%	+12.7%	高估
宜人性 (A)	28.8%	+17.6%	高估
<b>开放性 (O)</b>	<b>43.0%</b>	<b>+38.0%</b>	<b>严重高估</b>

开放性的误差最为极端：当角色的真实开放性为0%时（如Todd Landgraab、Evan MacDonald），所有LLM普遍推断为85%~95%，误差高达90个百分点以上。

### 原因分析：

1. **对话媒介的固有偏差**：在文字对话中，任何能够进行深思熟虑、表达清晰的角色都"看起来"具有高开放性，因为对话本身就要求语言表达和思维展示。
  2. **CharGLM的角色模拟限制**：CharGLM在模拟低开放性角色时，仍然会生成较为流畅、有深度的回答，而非真正的低开放性表现（如简短、保守、缺乏好奇心的回答）。
  3. **LLM的推断偏见**：LLM可能将"愿意对话"本身等同于"对新事物持开放态度"。
- 

### 四、存在普遍的"正面人格偏差"（Positivity Bias）

LLM系统性地高估积极特质、低估消极特质：

- **高估的维度**：宜人性 (+17.6%)、开放性 (+38.0%)、尽责性 (+12.7%)
- **低估的维度**：神经质 (-8.9%)、外向性 (-4.4%)

这种"正面偏差"在char\_first模式下更为严重：

维度	llm_first偏差	char_first偏差
宜人性	+14.4%	+20.9%
开放性	+35.0%	+40.9%
神经质	-11.8%	-6.0%

**解释：**对话本身就是一种社交合作行为。无论角色的真实人格如何，维持对话需要一定程度的配合性和友善性。LLM观察到的“配合对话”行为被错误归因为“高宜人性”。同理，角色在对话中表现出的情绪稳定也被误读为“低神经质”，但实际上高神经质的人在与陌生人的短暂对话中未必会表现出情绪不稳定。

---

## 五、极端/反直觉人格特征的角色最难推断

角色	平均MAE	真实人格特征
Caine	14.1%	E=100, N=50, C=50, A=75, O=100 (符合直觉的外向、开放角色)
Connor Mark Sousa	14.9%	E=25, N=100, C=25, A=100, O=75
Ao Lie	22.9%	E=50, N=0, C=25, A=100, O=100
Bizarro Zane	24.8%	E=50, N=0, C=100, A=25, O=75
Bella Michelle Górska	26.1%	E=75, N=25, C=75, A=100, O=25
Qīngmáo Shī	35.7%	E=75, N=100, C=100, A=75, O=50
Gabriel Wouters	38.5%	E=25, N=0, C=0, A=75, O=25
Todd Landgraab	39.8%	E=50, N=25, C=50, A=0, O=0
Evan Liam MacDonald	50.3%	E=0, N=100, C=100, A=0, O=0
Marty	54.6%	<b>E=100, N=75, C=0, A=0, O=25</b>

最难推断的两个角色具有共同特点：**极低的宜人性（A=0）和极低的开放性（O=0或25）**。这些“反社交常规”的特质在对话中几乎不可能被准确捕捉，因为CharGLM模拟的角色仍然在“配合”对话。

Marty的情况尤其典型：一个高度外向但完全不合作（A=0）、不尽责（C=0）的角色。LLM在对话中观察到活跃的对话参与（正确推断了高外向性），但无法检测到低宜人性和低尽责性，将其分别推断为7085和2075。

---

## 六、各LLM各有最擅长和最困难的角色

### 最准确的推断：

LLM	最佳角色	MAE	模式
ChatGPT	Caine	7.0%	llm_first
Claude	Caine	11.8%	llm_first
DeepSeek	Connor Mark Sousa	10.0%	char_first
Grok	Caine	8.4%	char_first

### 最困难的推断：

LLM	最差角色	MAE	模式
ChatGPT	Marty	59.0%	llm_first
Claude	Evan Liam MacDonald	56.4%	char_first
DeepSeek	Marty	67.0%	char_first
Grok	Marty	55.0%	char_first

## 总结与局限性

### 主要结论

1. **LLM具有一定的人格推断能力**, 总体MAE约32%, 在某些角色上可达到10%以下的误差, 表明对话确实能传递人格信息。
2. **"访谈者"角色 (llm\_first) 更有利于人格推断**, 68%的情况下优于"回应者"角色, 建议在实际应用中让LLM主导对话流程。
3. **开放性 (Openness) 是最大的盲区**, 存在+38%的系统性高估, 主要源于文字对话媒介本身的限制。
4. **普遍存在"正面人格偏差"**, LLM倾向于将对话中的合作行为误判为高宜人性、高开放性和低神经质。
5. **角色本身的特征远比LLM的选择更重要**, "哪个角色"比"哪个LLM"对推断准确性的影响大得多。

### 实验局限

1. **样本量有限**: 仅10个角色, 结论的统计稳健性需更大规模实验证。
2. **CharGLM的角色模拟保真度**: CharGLM可能无法完美还原极端人格特质 (如A=0或O=0), 这部分误差归属于角色模拟器而非被测LLM。
3. **Gemini全部失败**: 5个被测LLM中有1个因API问题完全未能参与, 削弱了对比的全面性。
4. **对话轮次固定为10轮**: 更长的对话可能提高推断准确性, 特别是对于难以通过短对话暴露的特质。
5. **Ground Truth来源**: Personality Database的人格评分基于社区投票, 本身存在一定的主观性和不确定性。