

大语言模型结构化领域评估的一致性边界

多模型企业流程分类学研究

Tim Ou

2026 年 3 月

报告提纲

- ① 研究动机与问题
- ② 研究设计
- ③ 核心发现
- ④ SDAF 理论框架
- ⑤ 假设检验汇总
- ⑥ 验证与贡献
- ⑦ 实践建议
- ⑧ 局限与未来
- ⑨ 结论

现实背景：LLM 作为自动评估者

- LLM 正被广泛用于结构化评估任务
 - 风险分类、影响评级、合规评分
 - 企业流程的 AI 影响评估
- 关键假设：不同 LLM 对同一内容**应给出一致判断**
- 但这个假设从未被系统验证过

核心问题

当多个前沿 LLM 独立评估同一结构化领域时，
它们的一致性有多高？

四个研究问题

- RQ1 **一致性量级**: 前沿 LLM 间的总体一致性水平如何? 是否在不同评估维度间系统性变化?
- RQ2 **结构性决定因素**: 哪些因素解释了模型间一致性的差异? (提示词版本、分类层级、领域新颖性、评估置信度)
- RQ3 **偏差特征**: 各模型是否存在可刻画为稳定“指纹”的系统性偏差?
- RQ4 **共识效用**: 尽管两两一致性很低, 多模型共识机制能否恢复有用信号?

评估领域：O'Process 框架

来源	节点数	覆盖领域
APQC PCF 7.4	1,921	跨行业流程
ITIL 4	141	IT 服务管理
SCOR 12.0	164	供应链
AI-era 扩展	99	AI 治理/MLOps
合计	2,325	

- 四级层级结构（部分五级）
- 中英双语节点描述
- **9 个评估维度**
 - 5 个分类维度（变化状态、渗透级别等）
 - 18 个数值维度（0–100 量表）

四模型实验设计

模型	来源	架构	提示词
Gemini 2.5 Flash	Google (US)	MoE	v2.2
DeepSeek V3.2	DeepSeek (CN)	MoE	v2.2
Qwen3 235B	Alibaba (CN)	Dense	v2.2
GPT-5 mini	OpenAI (US)	未公开	v2.2

- **统一提示词 v2.2**（消除提示词混淆）
- 温度 = 0.0（确定性输出）
- JSON 结构化返回
- 文化平衡：2 美 + 2 中
- 178 个自动化断言验证

数据规模

4 模型 × 2,325 节点 × 23 维度
= **213,900** 个独立判断

设计优势

所有模型使用相同提示词，
确保一致性差异反映模型本身
而非提示词工程的伪影

一致性度量 (7 族)

- Cohen's κ (两两一致性)
- Fleiss' κ (多评估者)
- Krippendorff's α
- ICC (组内相关)
- Gwet's AC1/AC2
- 加权 κ (有序数据)
- NMI (信息论)

分析框架

- 信息熵 & 互信息
- PCA 降维
- Bootstrap 置信区间
- LOO 稳定性检验
- ANOVA & 效应量
- 逻辑回归预测模型
- 混合效应方差分解

发现 1：低一致性符合预期——但业界假设相反

四模型研究结果

- 均值 Cohen's $\kappa = 0.078$
- Fleiss' $\kappa = 0.032$ (“轻微”)
- Mean Gwet's AC1 = 0.587
- Mean ICC(2,1) = 0.174
- 多数共识 ($\geq 3/4$) : **80.3%**

为什么低一致性是预期的

- 无金标准答案（主观评估）
- 无约束输出空间（非选择题）
- 模型训练数据/架构各异

但业界的隐含假设是……

“换一个 LLM 评估，
结果应该差不多”

这个假设是错误的。
模型选择本身就是一个
影响结论的关键决策。

本研究的贡献不是发现低一致性，而是**量化并解构**了不一致的内部结构。

发现 2：一致性的结构性依赖

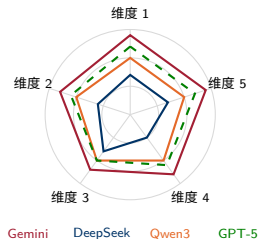
因素	发现	效应量	显著性
提示词对齐	相同提示词 κ 高 2×	$\Delta\kappa = +0.065$	$d = 0.58$
领域新颖性	AI 时代 56.6% vs 传统 42.2%	$h = 0.83$	$p = 0.005$
层级深度	越深一致性越低	$\rho = -0.046$	趋势
可观测性	可观测维度 ICC 8× 于主观维度	ICC: 0.25 vs 0.03	$p < 0.001$
框架来源	APQC vs SCOR 差距 11×	κ : 0.011–0.131	显著

核心洞察

近随机的总体一致性之下，隐藏着 0.46 个单位的 κ 变化幅度——不一致不是噪声，而是**可预测、可干预的**。

发现 3：模型偏差指纹

模型	偏差特征
Gemini	乐观主义 ：最高“已变”率，上调偏差
DeepSeek	保守主义 ：最高“稳定”率，高区分度
Qwen3	集中化 ：95%+ 置信度为“中”
GPT-5 mini	超集中 ：97.5% “将变”，分类空间坍塌



- 53/54 个模型-维度对显著偏差 ($p < 0.001$)
- 偏差跨所有 2,325 个流程持续存在
- 属于模型的**稳定属性**，非采样伪影
- GPT-5 mini 提供**新偏差类型**：
非乐观/保守轴，而是空间坍塌

发现 4: Kappa 悖论 \rightarrow 定理 1

同一数据，矛盾结论

- Fleiss' $\kappa = -0.012 \Rightarrow$ “低于随机水平”
- Gwet's AC1 = **0.762** \Rightarrow “显著一致”

定理 1 (一致性不确定区间):

若 Herfindahl 指数 $H > \frac{1+c}{q+1-c}$, 则存在

$$W = \frac{(q+1-c)H - (1+c)}{q-1}$$

使 $\kappa < 0$ 与 $AC_1 > c$ **同时成立**。

实例: $q=4, H=0.85 \rightarrow W = \mathbf{0.621}$

学术贡献

扩展 Warrens (2010) 和 Feinstein & Cicchetti (1990):

- 首个**精确边界条件**
- 首个**封闭宽度公式**
- 最大实证实例 ($W=0.621$)

实践含义:

- “LLM 是否一致?” 无度量无关的答案
- 应报告“可靠性走廊”而非单一数字

发现 5：共识恢复有用信号

- 多数投票 ($\geq 3/4$) 解决 **80.3%** 节点
- 全票一致节点聚集在语义连贯族群
 - 供应链、IT 运维、网络安全
 - 高面效度
- LOO 稳定：移除任何单模型，共识**不翻转**
- 215 个“硬节点”（无共识）→ 标记人工审核

有序一致性优势

加权 κ (change_status) = **0.299**

名义 κ = 0.171

提升 **+74.9%**

预测模型

逻辑回归预测多方一致：

AUC = **0.923**

主导特征：响应多样性

发现 6：反偏差信号放大（命题 3）

命题 3： 偏差模型做出**违背自身偏差**的判断时，

$$P(T_j = c \mid c \neq A_i) \geq \frac{1 - \varepsilon}{1 - \delta\varepsilon} > 1 - \varepsilon$$

即反偏差判断**严格优于**基线准确率。

模型	偏差方向	反偏差	跨模型同意率
Gemini	已变 (乐观)	稳定	86.0%
DeepSeek	稳定 (保守)	将变	96.7%
Qwen3	中 (集中)	低/高	78.3%
GPT-5m	将变 (超集中)	稳定	66.7% [†]

[†] N=6, 样本不足

放大倍数

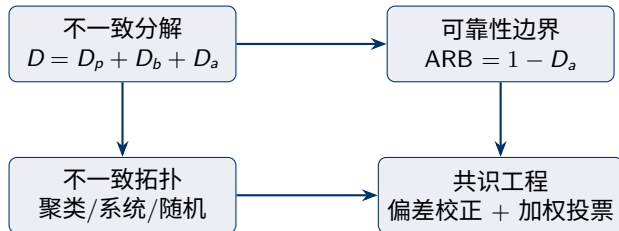
- 基线同意率：~45%
- DeepSeek 反偏差：96.7%
- 放大倍数 $A \approx 2.1\times$

实践意义

多模型共识系统中，反偏差判断应获得**更高权重**。

DeepSeek 说“将变” > Gemini 说“将变”
(前者克服保守偏差，后者顺从乐观偏差)

结构化不一致分析框架 (SDAF)



核心思想：不一致不是需要消除的噪声，而是揭示评估领域**内在模糊性**的信息信号。

三个数学结果相互强化：定理 1（度量不确定）→ 定理 2（校正有天花板）→ 命题 3（最优信号定位）。

SDAF 不一致分解（四模型混合效应）

定义 1（不一致分解）：

$$D_{total} = D_{bias} + D_{ambiguity} + D_{residual}$$

混合效应模型（统一提示词 v2.2）：

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

- α_i ：节点随机效应（任务模糊性）
- β_j ：模型固定效应（系统偏差）
- ϵ_{ij} ：残差（模型 × 节点交互）

成分	方差比
D_{bias} （模型偏差）	17.2%
$D_{ambiguity}$ （节点模糊性）	9.1%
$D_{residual}$ （交互残差）	73.6%

ARB = 1 - $D_{residual}$ ：

跨字段范围 **11.5%–56.4%**

核心结论

17% 可校正 | 9% 任务难度 | 74% 不可约

评估可靠性边界 (ARB)

定义 2: $ARB^f = 1 - D_{ambiguity}^f$

数学基础:

$$Y_m^{(i)} = g(\theta^{(i)}, \phi_m, \epsilon_m^{(i)})$$

- $\theta^{(i)}$: 节点潜在难度
- ϕ_m : 模型偏差参数
- $\epsilon_m^{(i)}$: 特异性噪声

偏差校正后最大一致性上界:

$$\kappa^* \leq 1 - \frac{E[\text{Var}_m[P(Y_m = k|\theta)]]}{E[\text{Var}[P(Y = k)]]}$$

实践意义

- 低一致性**不主要是**模型质量问题
- **也不主要是**提示词工程问题
- 73.6% 的不一致是**不可约**的模型 × 节点交互
- 17.2% 可通过偏差校正改善

11 项假设检验结果

H	零假设	结果	关键统计量	效应
H1	$\kappa_F = 0$	边际	$\kappa_F=0.032$	“轻微”
H2	$\kappa = 0$ (两两)	拒绝	多数对 $p < 0.05$	$\bar{\kappa}=0.078$
H3	同/异提示词无差异	拒绝	$\Delta\kappa=+0.065$	$d=0.58$
H4	领域新颖性无效应	拒绝	+14.4 pp	$h=0.83$
H5	深度与一致性无关	趋势	$\rho=-0.046$	可忽略
H6	模型均值相等	拒绝 (18/18)	全 $p < 0.001$	$\bar{\eta}^2=0.39$
H7	观测 = 随机	拒绝	超额 3.1%	有限
H8	多数投票不稳定	未拒绝	0 翻转	LOO 稳定
H9	κ 依赖样本量	未拒绝	$N \approx 100$ 稳定	—
H10	可观测 = 主观 ICC	拒绝	8× 差距	大
H11	$\kappa_F \equiv \text{AC1}$	拒绝	AC1 0.50–0.80	悖论

验证设计

- 分层抽样 123 个节点
- 高/中/低共识各 31/51/41 节点
- 领域专家独立标注

关键发现

专家-模型 $\kappa = 0.042$
低于模型间 $\kappa = 0.078$
专家-共识准确率 = 39.7%

含义:

- 人类与模型的不一致甚至**更大**
- 评估任务对大多数节点**缺乏明确的答案**
- 支持 SDAF 的核心论断：
不一致反映**任务本身的模糊性**
- 专家低置信节点确实显示更多模型分歧

3 个数学结果（附证明）

- 定理 1：一致性不确定区间（封闭公式）
- 定理 2：可靠性天花板（校正上界）
- 命题 3：反偏差信号放大 ($A=2.1\times$)

5 项实证贡献

- A. 374,325 判断 + 178 断言
- B. SDAF 不一致分解框架
- C. 4 模型偏差指纹刻画
- D. 可观测性原理 (ICC $8\times$)
- E. 有序一致性层 ($\kappa_w + 75\%$)

6 个可检验假设

- H1. 可观测性缩放
- H2. 偏差持久性
- H3. 递减收益
- H4. 反偏差推广
- H5. 富化阈值
- H6. 边界集中

6 项实践建议

- 多模型评估 · 偏差档案 · 可靠性走廊
- 不一致 = 信号 · 可观测性分级 · PCA 简化

合计：3 数学 + 5 实证 + 6 假设 + 6 建议 = 20 项贡献

三个数学结果（附证明）

定理 1：一致性不确定区间 (Section 6.8)

若 $H > (1 + c)/(q + 1 - c)$ ，则 $\kappa < 0$ 和 $AC_1 > c$ 可同时成立。

不确定宽度 $W = [(q + 1 - c)H - (1 + c)]/(q - 1)$ 。

首个封闭公式，扩展 Warrens (2010)。实测极端实例 $W = 0.621$ 。

定理 2：可靠性天花板 (Section 6.9)

偏差校正最大改善上界： $\Delta\kappa_{max} \leq 1 - D_{residual} = D_{bias} + D_{ambiguity}$

实测字段天花板范围 11.5%–56.4%，为校准投资提供优先级排序。

命题 3：反偏差信号放大 (Section 7.4)

$P(T_j = c \mid c \neq A_i) \geq (1 - \varepsilon)/(1 - \delta\varepsilon) > 1 - \varepsilon$

反偏差判断**严格优于**基线准确率。DeepSeek 反保守同意率 = **96.7%**（基线 45%）。

相互强化：定理 1 证明单一度量不可信 → 定理 2 限定偏差校正上界 → 命题 3 指出最高信号判断。

六个可检验假设 (H1-H6)

ID	假设	核心预测	证伪标准
H1	可观测性缩放	$ICC \propto \text{可观测性指数}$	主观维度 $ICC > 0.20$
H2	偏差持久性	方向稳定 ≤ 12 个月	偏差方向在版本间反转
H3	递减收益	$\Delta \kappa_k \propto 1/k^\alpha$	$k > 7$ 时 κ 线性增长
H4	反偏差推广	$A > 1.3\times$ (人类面板)	多场景 $A \leq 1.0$
H5	富化阈值	存在 C^* , 残差 $= D_{ambi}$	κ 随上下文线性增长
H6	边界集中	$>80\%$ 分歧在相邻类别	分歧均匀分布

从实证到假设生成

6 个假设将单一大规模研究转化为**未来研究纲领**——每个假设都有明确的实验设计和证伪路径。

六项实践建议

- ① 永远不要信任单模型评估——使用 ≥ 3 个独立模型并报告一致性
- ② 量化偏差指纹——每个部署的 LLM 应配有偏差档案
- ③ 区分可观测与主观维度——对主观维度降低置信度阈值
- ④ 报告一致性走廊 (κ 和 AC1 双报告) ——而非单一数字
- ⑤ 不一致 = 信号——高不一致节点标记人工审核而非强制共识
- ⑥ 评估工具可简化——18 维 \rightarrow 3 个主成分 (64.6% 方差)

核心启示

模型不一致经过恰当分析后，揭示的是评估领域本身的**内在模糊性结构**——这是特征，不是缺陷。

主要局限

- 单一评估领域（企业流程）
- 时间窗口固定（2026 Q1 模型版本）
- 无人类金标准（仅 123 节点专家验证）
- 提示词自身的构念效度未经专家面板评审
- 温度 = 0 消除了随机变异

未来方向（由 H1–H6 驱动）

- 跨领域推广：H1 可观测性缩放验证
- 纵向追踪：H2 偏差持久性检验
- 最优集成规模：H3 递减收益曲线
- 人类面板推广：H4 反偏差可推广性
- 上下文富化实验：H5 富化阈值
- 细粒度量表：H6 边界集中效应

结论：核心发现与数学结果

- ① LLM 是**接近随机**的评估者： $\kappa = 0.078$ ，超出随机仅 3.1%
- ② 不一致具有**可利用的结构**： κ 范围跨 0.46 个单位
- ③ 每个模型拥有**稳定的偏差指纹**：持续跨 2,325 个流程
- ④ SDAF 将不一致**转化为信号**： $D_{bias}=17.2\%$ （可校正）， $D_{residual}=73.6\%$
- ⑤ 多模型共识**恢复有用信号**：80.3% 解决率，LOO 稳定
- ⑥ **定理 1**：“LLM 是否一致？”无度量无关的答案（ $W=0.621$ ）
- ⑦ **定理 2**：偏差校正上界 $\leq 26.3\%$ （73.6% 不可约）
- ⑧ **命题 3**：反偏差判断放大 $2.1\times$ （DeepSeek 96.7%）

中心论点

不一致经恰当分解后是**信息性的**——其边界可**数学刻画**（定理 1-2），其信号可**定向放大**（命题 3）。

指标	数值
评估流程节点	2,325
总判断数（主要 + 复制）	374,325
统计表格	42
图表	29 (PNG + PDF)
假设检验	11 (H1-H11)
可检验假设生成	6 (H1-H6, Section 7.10)
自动化断言	178（全部通过）
定理/命题（附证明）	2 定理 + 1 命题
分析脚本	36 (25 主要 + 11 复制)
引用文献	45+
贡献	3 数学 + 5 实证 + 6 假设 + 6 建议

RQ	问题	答案
RQ1	一致性量级	接近随机: $\kappa=0.078$, $\kappa_F=0.032$
RQ2	结构性因素	提示词 ($\Delta\kappa=+0.065$)、领域新颖性 ($p=0.005$)、可观测性 ($8\times$)
RQ3	系统性偏差	是: 普遍且稳定。17.2% 方差 (可校正)
RQ4	共识效用	是: 80.3% 多数共识, LOO 稳定

谢谢! 欢迎提问。