

# 大语言模型在结构化领域评估中的一致性极限

基于业务流程分类的四模型实证分析

---

OUJIANFENG

2026 年 3 月 · 213,900 条判断

# 研究背景

- LLM 被广泛用作自动评估者
  - 风险评估、合规打分、AI 影响判断
- 多模型评估可靠性缺乏实证研究
- 本研究规模
  - 4 款前沿 LLM
  - 2,325 个流程节点
  - 23 个评估字段
  - = 213,900 条独立判断

## 实验设计

四盲设计

模型间无法访问彼此响应

统一提示词 v2.2

温度 = 0.3

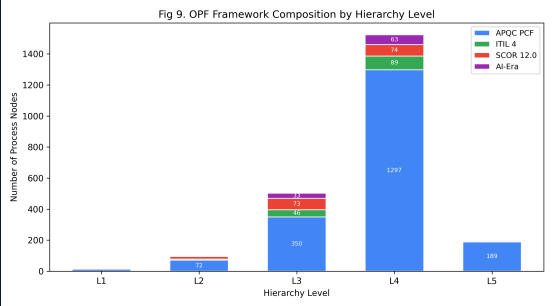
# 三个核心研究问题

**RQ1** 前沿 LLM 在结构化评估中  
能达到什么水平的一致性？

**RQ2** 一致性如何随**维度、领域和层级**变化？

**RQ3** 系统性模型偏差能否从  
任务固有歧义中**分离**？

# OPF 评估框架构成



框架	节点数
APQC PCF 7.4	1,921
ITIL 4	141
SCOR 12.0	164
AI 时代扩展	99
合计	2,325

4 级层级结构  
中英双语描述

# 四款前沿 LLM 评估模型

Fig 15. LLM Model Specifications

Model	Provider	Release	Architecture	Context	API
Gemini 2.5 Flash	Google	May 2025	MoE	1M	Gemini API
DeepSeek V3.2	DeepSeek	Jun 2025	671B MoE	128K	DashScope
Qwen3 235B	Alibaba	Jul 2025	235B MoE	128K	DashScope
GPT-5 mini	OpenAI	Aug 2025	Undisclosed	128K	OpenAI API

Gemini 2.5 Flash  
Google

DeepSeek V3.2  
DeepSeek

Qwen3 235B  
阿里巴巴

GPT-5 mini  
OpenAI

# 统一提示词 v2.2：9 维度 AI 影响评估

v2.2 指令每个 LLM 对流程节点进行结构化的 AI 影响评估，输出 JSON 格式，覆盖 9 个维度：

维度		子项
D1	AI 渗透力	3 项 × 1-5 分
D2	变化状态	已变/将变/稳定
D3	变革性质	4 类型 × 0-5
D4	人机边界	类型 1-4
D5	不确定性	高/中/低
D6	信号质量	5 源分布
D7	流程结构刚性	3 项 × 1-5
D8	数据生态位	4 项 × 1-5
D9	AI 改造就绪度	4 项 × 1-5

输出：5 分类字段 + 18 数值维度 = 23 字段/节点

## v2.2 关键改进 (vs v2.1)

- **锚定量化：** D7/D8/D9 引入明确阈值  
如 D9: 1=<5%, 3=15-30%, 5=>50% 成本节约
- **置信度校准** (规则 10):  
类型 B 证据 + 无行业统计 → 强制“低”
- **状态约束** (规则 9):  
“已变”需 ≥30% 企业采用 + A 类证据
- **全量表使用** (规则 8):  
禁止聚集 3-4 分，强制区分度
- **格式严控：**  
边界类型限定“类型 N”正则输出

# 实验设计与统计方法

## 实验控制

**四盲设计**：模型间无法访问彼此响应

**统一提示词** v2.2：消除提示词混淆

**温度 = 0.3**：近确定性输出

**JSON 结构化返回**：178 条断言验证

**文化平衡**：2 美 + 2 中

## v2.2 的 10 条强制规则

规则	约束
1-2	循证评估，区分领先/主流
3-4	计划 $\neq$ 变化，诚实不确定性
5-6	标注证据类型，依据 $\leq 30$ 字
7-8	D3 允许 0 分，全量表区分
9-10	状态 + 证据约束，置信度校准

## 统计方法

Cohen's/Fleiss'  $\kappa$  • Gwet's AC1 • ICC • Bootstrap CI • 混合效应 SD-ANCOVA 分解 • ANOVA  $\eta^2$   
• 5-fold CV 逻辑回归 • 蒙特卡洛模拟

# 流程节点来源与可靠性

框架	节点	版本	权威性与可靠性
APQC PCF	1,921	7.4	全球最广泛的跨行业流程分类标准；600+ 会员企业；ISO 对标
ITIL	141	4.0	ISO/IEC 20000 对齐的 IT 服务管理国际标准
SCOR	164	12.0	ASCM 供应链运营参考模型 (Plan/Source/Make/Deliver/Return)
AI-era	99	1.0	本研究自建：AI 治理、MLOps、数据伦理，填补传统框架空白

数据质量保障

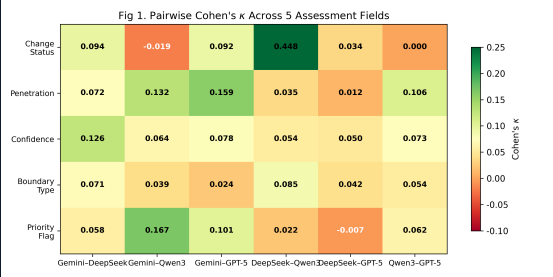
中英双语节点描述  
4 级层级结构（部分 5 级）  
ID 唯一性 + 父子引用完整性校验  
9 项自动化质量门禁

为什么选择流程分类？

结构化、层级化、有明确分类边界  
覆盖传统到 AI 新兴领域  
为 LLM 评估一致性提供理想测试基底



# 核心发现：一致性惊人地低



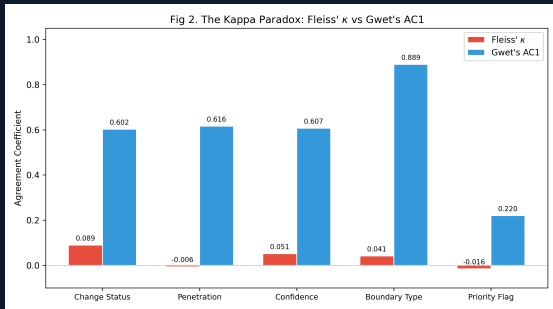
关键数据

均值 Cohen's  $\kappa$  = 0.078  
仅比随机略高

指标	数值
Bootstrap 95% CI	[0.047, 0.110]
Fleiss' $\kappa$	0.032
Gwet's AC1	0.587
多数共识	80.3%

Landis-Koch 量表：轻微一致

# Kappa 悖论：同一数据，截然不同的答案



## 最极端案例

boundary\_type 字段：

Fleiss  $\kappa = 0.041$

→ “轻微一致”

Gwet AC1 = 0.889

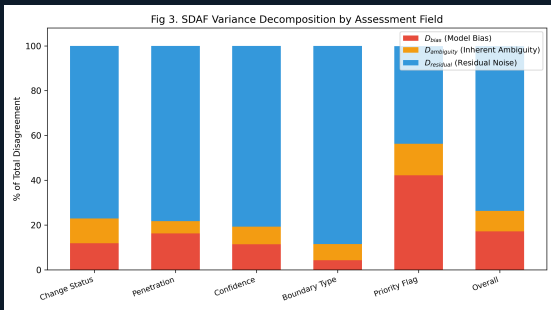
→ “几乎完美”

## 结论

LLM 是否一致？

答案取决于你选择哪个指标。

# SDAF 方差分解：分歧从何而来？



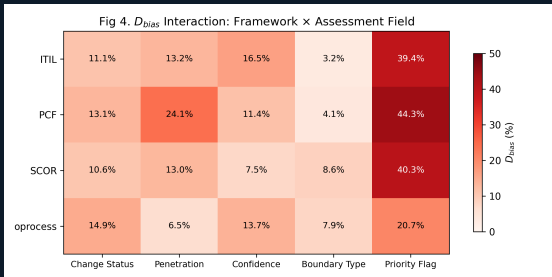
结构化分歧分析框架 (SDAF):

组分	占比	
$D_{bias}$	17.2%	可纠正
$D_{ambiguity}$	9.1%	固有极限
$D_{residual}$	73.6%	不可预测

## 关键洞察

残差主导 → 大多数分歧是特异性的，无法预测

# 模型偏差并非均匀分布



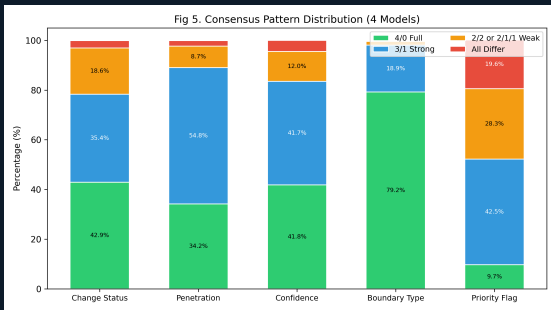
## 关键发现：

- **PCF 流程偏差最高**  
传统业务需推理 AI 相关性
- **AI 时代流程偏差最低**  
AI 关系自明
- penetration 字段  
PCF vs AI-era 差异达 **3.7 倍**

## 结论

模型可靠性是领域依赖的

# 共识分析：低 $\kappa \neq$ 无信息



## 指标

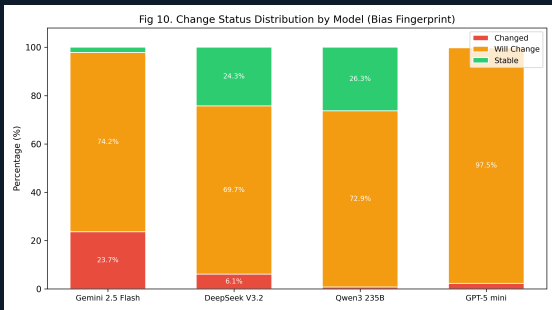
## 数值

完全共识 (4/4)	42 (1.8%)
多数共识 ( $\geq 3/4$ )	80.3%
难节点 ( $\geq 3$ 字段无多数)	215 (9.2%)

## 结论

多模型集成可从  
不可靠评分者中  
提取有用信号

# 模型偏差指纹



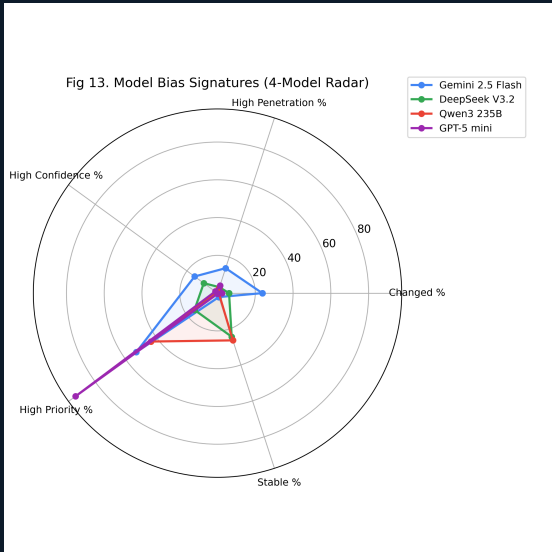
模型	已变%	特征
Gemini	23.7%	最激进
DeepSeek	6.1%	最保守
Qwen3	0.7%	高度集中
GPT-5m	2.3%	极端单峰

GPT-5 mini: 97.5% 标记为“将变”

## 结论

每个模型有独特且持久的  
偏差签名

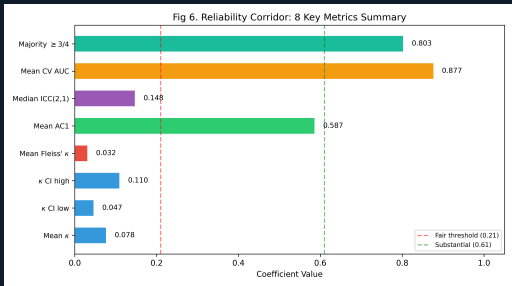
# 四模型偏差签名雷达图



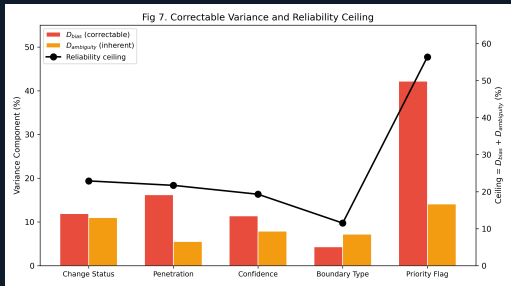
## 关键发现：

- GPT-5 mini 高优先级标注率  
92.9%（极端校准偏差）
- Gemini 在所有维度最激进
- DeepSeek 最保守均衡
- 偏差签名在  
所有领域持续存在

# 可靠性走廊与天花板



8 指标可靠性走廊



可纠正方差与天花板

## 核心结论

即使完美消除所有偏差，一致性天花板仅 **26.3%**。

73.6% 残差方差 = 不可系统性纠正的极限。高偏差字段反而**更可预测**。



# 三大学术贡献

## 一致性不确定性 (Agreement Indeterminacy)

同一数据同时是**低的** ( $\kappa=0.078$ )、**中等的** ( $AC1=0.587$ )、**可恢复的** (80.3%)。  
指标选择本身就是一个研究决策。

## 结构化分歧分析框架 (SDAF)

将分歧分解为可操作组分：

$D_{bias}$  (可校准纠正) +  $D_{ambiguity}$  (固有极限) +  $D_{residual}$  (特异性噪声)

## 反偏差可信度 (Counter-Bias Credibility)

模型偏离自身基线时的异见**最具诊断价值**。

DeepSeek 说“已变” → 高信号 (它通常很保守)

场景	建议
单模型部署	风险高 — 模型选择比数据更影响结果
多模型集成	推荐 — 多数投票恢复 80% 信号
高偏差字段	优先校准— $D_{bias}$ 可纠正
高残差字段	接受极限—不可预测
新领域评估	偏差更低但残差更高
置信度分层	用 $AUC=0.877$ 区分可信/不可信共识

### LLM 一致性的答案取决于你如何定义一致性

- 低  $\kappa$  并不意味着无信息——分歧本身是**可分析的信号**
- SDAF 框架使分歧从黑箱变为**可操作的诊断工具**
- 模型偏差是可纠正的，但 73.6% 残差代表**根本极限**

感谢聆听 ▪ OUJIANFENG