

大语言模型在结构化领域 评估中的一致性极限

引理 一一

基于业务流程分类的四模型实证分析

张垒 | 2026年3月 · 213,900条 引理 设计

研究背景

1. LLM被广泛用作自动评估者（风格评估、合报打分、AI影响判断）
2. 多模型评估可靠性缺乏实证研究
3. 本研究：4款前沿LLM \times 2,325流程节点 \times 23评估等级 = 213,900判断
回盲设计，模型问题无法窥见答案，避免偏见

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、维度和层级变化?

RQ3: 系统性模型偏差从维度和层级变化?

RQ3: 系统性偏差是否随维度和层级变化? 是否一致?

OPF评估框架构成

APQC PCF 7.4: 1,921节点

ITIL 4: 141节点

SCOR 12.0: 164节点

AIB时代扩展：供应链，4级层级结构

AIB时代扩展：供应链系统
总计2,325个供应链节点

创建一张幻灯片

四款前沿LLM评估模型



模型使用2.5种不同，精度=0.0

核心发现：一致性惊人地低

均值 Cohen's $\kappa = 0.078$

Bootstrap 95% (仅比随机略高)

Bootstrap 95% CI: [0.047, 0.110]

Landis-Koch量表：轻微一致用此警告警示这个发现的可靠性

Kappa悖论论：同一数据， 截然不同的答案

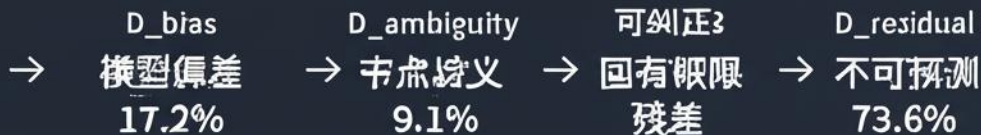
最新摘要案例高亮：boundary_type字码： Fleiss

$\kappa = 0.041 \rightarrow$ “轻微一致” Gwet

Gwet AC1 = 0.889 \rightarrow “凡平完美”

荒唐结论：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？



关键洞察：残差主导意味着看大模型分歧是特异性的

局部结论：LLM是否一致？答案取决于你选择哪个指标。

模型偏差并非均匀分布

关键发现：PCF流程偏差最高（3.7倍，需推理AI相关性）
AI时代流程最低（AI关系自明）

penetration等级 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k ≠ 无信息

核心数据：完全共识(4/4)：仅42节点(1.8%)

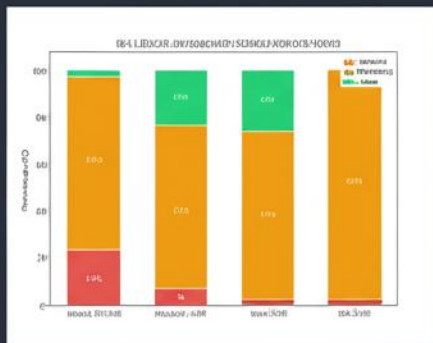
精英节点：多数共识($\geq 3/4$)：80.3%的判断可恢复

非精英节点(≥ 3 节点无多数)：215节点(9.2%)

底层节点：多模型集成可从不可靠数据差异达成共识

性能提升：PCF 不可靠评分分者中提取有用信号

创建一张幻指纹片



Gemini: 最激进, 23.7%标 '已变'

DeepSeek: 最保守, 分布最分散

Qwen3: 高度集中, 0.7%'已变'

GPT-5 mini: 极端单峰, 97.5%'将变'

后发制人月28战第2.1, 信月升快难

因个市等持且持久欠能盛响隆签名

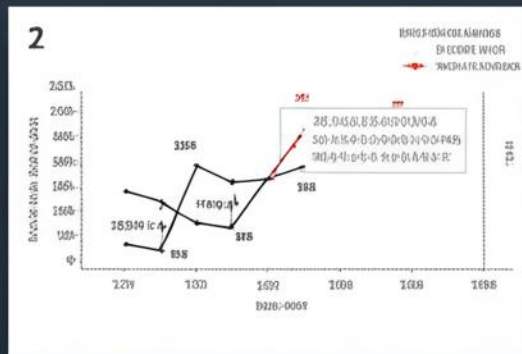
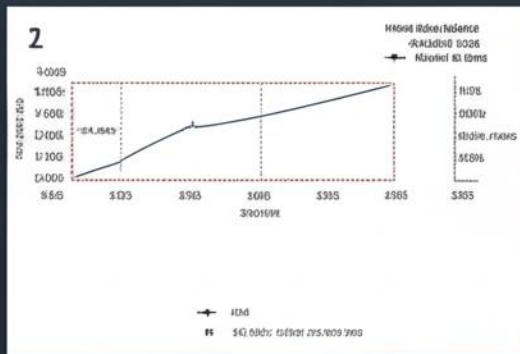
四模型 张幻偏差名雷达图

关键发现：GPT-5 mini高估乱垃圾标注率92.9% (极端校准偏差)

Gemini在所有维度上皆为激进

DeepSeek皆为保守字均高偏差 幻觉与幻觉持续存在

创笔性走廊与天花板



46 268 237 204 :

[illegible]

三大学术贡献

① 一致性不确定性

(同一数据同时既满足 0 假设 ($\lambda=0.076$), 中($AC1=0.587$), 可恢复(80.3%)

② 结构化分歧分析框架

将分歧结构化图析为“一致”“误差”三个可操作组分
(SDAF)

③ 反偏差可信度

(Counter-biasing and differential bias) 具具诊断价值

实践启示

单模型部署 →  风险高

(验证集性能波动大) 模型泛化能力不可靠 (AUC: 0.907), 低泛化能力模型

多模型集成 →  多数投票恢复60%精度

高偏差导致 →  优先校准 (D_bias可纠正)

高残差导致 →  接受风险 (不可预测)

新领域评估 →  偏差低但残差高需谨慎

置信度分层 →  用AUC=0.877区分可信值

结论与展望

LLM一致性的答案取决于你如何定义一致性

三行总结：低 κ 开不套筛筛无信息——分
分枝本身是可分析的信号

SDAF框架在低 κ 以黑箱变为白盒模型而提升
正型儒屋是可纠正工程
熊盖掩操作和，但73.6%残差代表根本极限
底部

詹皓宇 · 欧剑锋 · jianfang.ou@example.com