指数案例说明

案例设定

- 评测对象：5 个“AI 办公助手”品牌（同赛道可比）
- 案例品牌：**灵犀办公**（演示全链路计算）
- 指数口径
  - **AIP = 0.3A + 0.3B + 0.2C + 0.2D**
  - **AIC = (E1+E2+E3)/3**（v1=v2=v3=1/3）
  - **AIR = 0.3R1 + 0.3R2 + 0.2R3 + 0.2R4**
  - **UAF / MCI** 映射规则照文档（UAF∈[0.7,1]；MCI∈[0.8,1]）
  - 综合指数（可选）：

    AIBP2 = MCI × [λ·AIP + (1-λ)·(α·AIC + β·AIR)] × UAF，取 λ=0.6，α=0.6，β=0.4

Step 1：数据怎么来（保证可复现）

按文档流程：题集 → 多模型批量问答 → 存档输出 → 结构化抽取与标注 → 计算指标 → QA/QC → 形成 AIP/AIC/AIR + UAF/MCI。

并且题集中要有一定比例“公开题 + 暗题/随机题”用于防操纵。

Step 2：先把“灵犀办公”的 AIP 算出来（从原始计数到分数）

> 这里我用文档给的指标定义做一组可复现的原始数据。

2.1 维度 A（认知准确度）

- **A1 事实正确描述率** = 86/100 = 86
- **A2 核心要素覆盖度** = 8/10 = 80
- **A4 品牌区分度**：混淆 4/120 → ConfuseRate=0.0333 → A4=96.67
- 本例维度 A 取简单平均：A = (86+80+96.67)/3 = **87.56**

2.2 维度 B（可见性与召回）

- **B1 指名召回率** = 19/20 = 95
- **B2 类目/需求召回率** = 30/40 = 75
- **B3 召回排序指数**：先按位置权重加权计数，再归一到 0–100（文档给了形式，归一方式由项目定义）
  - 本例归一后：B3 = 86.3
- **B4 长尾覆盖度** = 16/20 = 80

- 维度 B 简单平均：B = (95+75+86.3+80)/4 = **84.1**

2.3 维度 C（场景适配）

C1/C2/C3 的定义与评分方式按文档：C1 1–5 分折算 0–100；C2 覆盖率；C3 合理性评分。

- C1=82，C2=70，C3=78 → C = (82+70+78)/3 = **76.7**

2.4 维度 D（一致性与调性）

D1 跨模型一致性；D2 跨时间稳定；D3 调性匹配。

- D1=78，D2=65，D3=72 → D = (78+65+72)/3 = **71.7**

2.5 计算 AIP

AIP = 0.3A + 0.3B + 0.2C + 0.2D（权重如文档示例）

- AIP = 0.3×87.56 + 0.3×84.1 + 0.2×76.7 + 0.2×71.7 = **81.18**

Step 3：再算 AIC、AIR、UAF、MCI（并给出综合分）

3.1 AIC（建设指数）

AIC = (E1+E2+E3)/3

- E1=80（知识库/语料建设）
- E2=60（落地应用成熟）
- E3=70（战略与实践透明）

  → **AIC=70**

3.2 AIR（风险与稳定）

- R1（严重错误控制度）：Err_severe=1%，Err_max=5% → R1=80
- R2（负面偏差控制度）：NegBias=3% → R2=97
- R3（稳定性得分）：CV=0.12，CV_max=0.30 → R3=60
- R4（操纵疑似）：Susp=0.10 → R4=90
- AIR = 0.3R1+0.3R2+0.2R3+0.2R4 = **83.1**

3.3 UAF & MCI（校准因子）

- UAF_raw = 0.4Usage + 0.3Trust + 0.3BrandUsage；UAF=0.7+0.3·UAF_raw
  - 假设：Usage=0.70，Trust=0.60，BrandUsage=0.65 → UAF_raw=0.655 → **UAF=0.8965**
- MCI_raw = 0.25S+0.25Q+0.25T+0.25M；MCI=0.8+0.2·(MCI_raw/100)
  - 假设：MCI_raw=70 → **MCI=0.94**

3.4 综合指数 AIBP2（用于“排名”）

按文档综合公式与参数（λ=0.6，α=0.6，β=0.4）

- **灵犀办公 AIBP2 = 66.41**

> 传播与报告层面，文档也建议不要只给一个综合分，至少同时展示 AIP/AIC/AIR/UAF/MCI。

Step 4：做“排名”演示（同赛道 5 品牌）

统一采用同一批次的 UAF=0.8965、MCI=0.94（行业与方法层面一致），各品牌 AIP/AIC/AIR 如下（示例）：

品牌	AIP	AIC	AIR	AIBP2(综合分)	排名
灵犀办公（案例）	81.18	70	83.1	**66.41**	1
云帆AI	73	85	76	64.35	2
极光写作	84	65	60	63.71	3
北斗助手	79	55	90	63.20	4
星河智答	60	45	70	48.88	5

排序规则（可执行）

1. 先按 AIBP2 降序；2) 若分差 <0.5，优先比 AIP；3) 仍相近再比 AIR（风险更低者优先）；4) 最后看 MCI（低置信度不给高排名）。

三种“落地实现路径”（你可以按资源选）

方案 1：“演示版”

- 做法：20–50 道题 × 2–3 个模型 × 人工标注 + Excel 公式
- KPI：能输出 AIP/AIC/AIR/UAF/MCI + 排名；复现误差 <±2 分
- 实操路径：题集（公开/暗题）→ 批量问答 → 表格标注（A1/B1/...）→ 公式出分 → 排名看板

方案 2：“标准版”（可用于对外报告）

- 做法：200–500 道题 × 4–8 模型 × 双人复核标注 + QA/QC
- KPI：MCI≥0.92；跨批次波动（关键指标 CV）进入可控区间
- 实操路径：题库规范化 → 多批次测量（跨时间）→ 异常检测（公开题 vs 暗题差异）→ 出具等级与说明

方案 3：“自动化生产版”（持续监测 + GEO 闭环）

- 做法：数据管道（抓取/调用/存档）+ LLM 辅助结构化抽取 + 人类抽检
- KPI：每周更新；召回/负面风险预警；可定位到“哪类问题导致丢分”
- 实操路径：建立“品牌知识库+FAQ语料”→ 定期测评 → 对低分指标定向补语料/改表达 → 再测验证提升
