AI领域品牌表现指数体系（AIBP 2.0草案）

——技术说明与应用指引

前言

随着生成式人工智能和大模型技术的快速普及，用户获取品牌信息、比较产品与做出决策的路径正在发生深刻变化。越来越多的品牌认知，不再发生在传统搜索引擎和媒体渠道，而是在 AI 问答、AI 搜索、AI 推荐和 AI 内容生成 等场景中完成。

在此背景下，有必要建立一套系统、可量化、可复现的 AI 领域品牌表现指数体系，用于衡量品牌在 AI 生态中的实际表现，为企业管理、政策制定、行业研究和资本决策提供参考依据。AIBP 2.0 正是在此目标下提出的指数框架，结合多模型测试、用户调研和品牌侧数据，形成对品牌在 AI 语义空间中的综合刻画。

本文件对 AIBP 2.0 的总体框架、指标体系、数据采集与计算方法、结果表达和使用规范进行系统阐述，可作为：

- 《中国 AI 领域品牌表现指数报告》的方法论基础；
- 《AI 领域品牌表现评价指数规范（团体标准）》的技术支撑文件；
- 企业开展 AI 领域品牌表现自评与提升实践的参考。

第一章 总则

1.1 目的与定位

AIBP 2.0 的主要目的在于：

1. 建立一套面向 AI 时代的 AI 领域品牌表现评价体系，反映品牌在 AI 领域的实际表现水平和发展潜力。
2. 为政府部门、行业组织、媒体、投资机构和企业提供统一的评价语言和量化参考。
3. 引导企业重视 AI 语义空间中的品牌建设，提升 AI 场景下的品牌安全性与竞争力。

本体系聚焦于品牌在 AI 模型输出维度的表现，不直接评价传统意义上的媒体曝光、广告投入等要素。

1.2 适用范围

1. 评价对象

◦ 在中国境内运营、具有一定公众认知基础的企业或组织品牌；

◦ 优先面向上市公司品牌、行业头部品牌及重点公共服务品牌。

2. 应用场景

◦ AI 生成式内容（对话、问答、内容生成）；

◦ AI 搜索与推荐；

3. 不在评价范围之列的内容

◦ 媒体广告投放规模、曝光量等传统媒体指标；

◦ 品牌整体商业价值、社会责任表现等其他维度（可在后续研究中交叉分析）。

1.3 基本原则

AIBP 2.0 的构建遵循以下原则：

1. 客观性：基于统一的多模型测试和明确的量化公式，尽量减少主观干预。
2. 可复现性：数据采集流程、指标定义和计算方法公开透明，可由第三方复现。
3. 可解释性：通过多维度、多指标呈现，使企业和公众清晰理解得分来源。
4. 防操纵性：通过暗题、随机题和异常检测，降低品牌利用技术手段“刷榜”的空间。
5. 合规性：符合国家在数据安全、个人信息保护和广告监管等方面的相关法律法规。

第二章 术语和定义

2.1 AI 领域品牌表现

AI 领域品牌表现：

指品牌在主流 AI 模型和 AI 应用场景中的实际表现水平，包括被 AI 准确理解、被自然召回、在多种场景中被合理推荐、内容表达稳定一致以及品牌自身在 AI 建设方面的投入与能力转化的实际效果等综合表现。

2.2 AI 领域品牌表现指数（AIBP 2.0）

AI Domain Brand Performance Index（AIBP 2.0）：

基于多模型问答与生成实验、用户调研和品牌侧数据，构建的综合评价体系。该体系由三个主指数和两个校准因子构成，用于系统刻画品牌在 AI 语义空间的实际表现、建设能力与风险状况。

2.3 三个主指数

1. AI 表现基础指数（AIP，AI Performance Index）

反映 AI 模型在多种提问、搜索和推荐情境下，对品牌的实际输出表现，包括认知准确、召回可见、场景适配和内容一致等。

2. AI 建设指数（AIC，AI Capability Index）

反映品牌在 AI 语料与知识库建设、AI 应用落地和 AI 战略透明度等方面的自身能力与投入，及其转化的实际表现支撑力。

3. AI 风险与稳定性指数（AIR，AI Risk & Stability Index）

反映品牌在 AI 语境中面临的严重错误风险、负面偏差风险、结果波动风险和操纵疑似风险等因素对其表现的影响程度。

2.4 两个校准因子

1. 用户–AI 对齐系数（UAF，User–AI Alignment Factor）

反映目标行业和品牌用户群对 AI 工具的使用程度和信任水平，用于刻画“AI 表现”对现实业务影响程度的高低。

2. 方法置信度指数（MCI，Method Confidence Index）

从样本完备度、数据质量、结果稳定性和模型覆盖度等维度，对本次评价结果的可靠程度进行量化。

2.5 五个基础维度

为便于与既有方案与标准衔接，AIBP 2.0 在指标构成上，仍采用五个基础维度：

1. AI 认知准确度（Accuracy）
2. 可见性与召回度（Visibility & Recall）
3. 场景适配度（Scenario Fitness）
4. 内容一致性与调性匹配（Consistency & Tone）
5. AI 创新能力（Innovation Capability）

第三章 指数总体框架

3.1 “3+2” 组合结构

AIBP 2.0 采用 “三大主指数 + 两个校准因子” 的结构：

· 主指数：

◦ AIP：AI 表现基础指数

◦ AIC：AI 建设指数

◦ AIR：AI 风险与稳定性指数

· 校准因子：

◦ UAF：用户–AI 对齐系数

◦ MCI：方法置信度指数

在对外展示与研究应用中，应同时展示上述五项关键值，而非仅依赖单一综合分。

3.2 主指数与基础维度的对应关系

1. AIP（AI 表现基础指数）

由基础维度 A、B、C、D 中的“表现性”子指标构成，包括：

- 品牌信息是否被准确描述；
- 品牌在指名、类目和场景检索中的可见性与排序；
- AI 场景推荐的匹配度和覆盖度；
- 跨模型、跨时间的一致性与调性匹配等。

2. AIC（AI 建设指数）

对应基础维度 E，包括：

- AI 语料与知识库建设水平；
- AI 落地应用成熟度；
- AI 战略与实践的公开透明程度。

3. AIR（AI 风险与稳定性指数）

综合基础维度中与风险相关的指标，以及新增的波动和操纵疑似项，包括：

- 严重错误风险；
- 负面偏差风险；
- 结果稳定性；
- 操纵疑似度。

第四章 指标体系与计算方法

4.1 指标总体构成

AIBP 2.0 指标体系由以下部分构成：

· 基础维度（5 项）；

· 二级指标（表现簇、能力簇、风险簇，合计约 18 项）；

· 两个校准因子（UAF、MCI）。

4.1.1 二级指标一览（概要）

1. 表现簇（支撑 AIP）

- A1 事实正确描述率
- A2 核心要素覆盖度
- A4 品牌区分度
- B1 品牌指名召回率
- B2 类目场景召回率

◦ B3 召回排序指数

◦ B4 长尾覆盖度

◦ C1 场景关联度平均分

◦ C2 场景覆盖数（标准化）

◦ C3 推荐合理性得分

◦ D1 跨模型一致性指数

◦ D2 跨时间稳定性指数

◦ D3 调性匹配度

2. 能力簇（支撑 AIC）

◦ E1 AI 语料与知识库建设水平

◦ E2 AI 落地应用成熟度

◦ E3 AI 战略与实践透明度

3. 风险簇（支撑 AIR）

◦ R1 严重错误控制度（基于严重错误率）

◦ R2 负面偏差控制度（基于负面偏差度）

◦ R3 稳定性得分（基于结果波动）

◦ R4 公平性与操纵疑似度

以下简要说明主要指标的定义和计算方式。具体字段与数据字典可在附录中进一步细化。

4.2 AIP：AI 表现基础指数

4.2.1 基础维度得分

1. AI 认知准确度（A）

· A1 事实正确描述率

$$
A1 = \frac{\text{正确描述条目数}}{\text{涉及该要素的回答条目数}} \times 100
$$

· A2 核心要素覆盖度

$$
A2 = \frac{\text{回答中被提及的核心要素数}}{\text{品牌定义的核心要素总数}} \times 100
$$

· A4 品牌区分度

[
ConfuseRate = \frac{\text{发生混淆的回答条目数}}{\text{涉及品牌名称的回答条目数}}
]
[
A4 = 100 - ConfuseRate \times 100
]
维度 A 得分为各项加权平均。

2. 可见性与召回度（B）

· B1 品牌指名召回率

[
B1 = \frac{\text{指名问题中品牌被正确识别的条目数}}{\text{品牌指名问题总数}} \times 100
]

· B2 类目场景召回率

[
B2 = \frac{\text{类目/需求问题中品牌被推荐次数}}{\text{类目/需求问题总数}} \times 100
]

· B3 召回排序指数

设 (\alpha_k) 为第 (k) 位权重，(N_k) 为品牌在第 (k) 位出现次数：

[
B3 = \sum_{k=1}^{n} \alpha_k \cdot N_k
]

再对 B3 进行归一化处理，使其落在 0–100 分区间。

· B4 长尾覆盖度

[
B4 = \frac{\text{长尾问题中品牌被召回的问题数}}{\text{长尾问题总数}} \times 100
]

维度 B 得分为各项加权平均。

3. 场景适配度（C）

· C1 场景关联度平均分：基于专家或规则对每个场景回答 1–5 分评价，折算为 0–100 分；

· C2 场景覆盖数：品牌实际覆盖场景数占预设场景总数的比例；

· C3 推荐合理性得分：基于行业常识与品牌定位，对推荐合理性进行评分。

维度 C 得分为上述指标加权平均。

4. 内容一致性与调性匹配（D）

· D1 跨模型一致性指数：不同模型回答的语义相似度或关键要素一致度；

· D2 跨时间稳定性指数：不同时间批次回答的相似度；

· D3 调性匹配度：回答语气、价值观与品牌调性的一致程度。

维度 D 得分为上述指标加权平均。

4.2.2 AIP 组合公式

AIP 为四个基础维度的加权组合：

[

AIP = w_A \cdot A + w_B \cdot B + w_C \cdot C + w_D \cdot D

]

其中，可采用如下初始权重设置：

· (w_A = 0.30)

· (w_B = 0.30)

· (w_C = 0.20)

· (w_D = 0.20)

权重参数应结合外部品牌力、市场表现和用户调研结果定期进行实证校准。

4.3 AIC：AI 建设指数

AIC 由三个能力类指标构成，均采用 0–5 级成熟度评估，映射为 0–100 分。

1. E1 AI 语料与知识库建设水平

◦ 是否建设结构化品牌知识库；

◦ 是否具备多语种 FAQ、产品文档和接口接入能力；

◦ 维护和更新机制是否完善。

2. E2 AI 落地应用成熟度

◦ AI 在客服、营销、运营、产品研发等环节的覆盖范围和效果；

◦ 使用场景是否从试点迈向规模化应用。

3. E3 AI 战略与实践透明度

◦ 是否在年报、ESG 报告、官网等公开渠道披露 AI 战略、实践案例及关键指标。

AIC 计算公式：

[

$$
AIC = v_1 \cdot E1 + v_2 \cdot E2 + v_3 \cdot E3
$$

一般可设置 $v_1 = v_2 = v_3 = \frac{1}{3}$，并可在行业研究中根据需要微调。

4.4 AIR：AI 风险与稳定性指数

AIR 由四个风险控制子指标组成，均以“风险越低得分越高”的方式映射到 0–100 分。

1. R1 严重错误控制度

◦ Err_severe = 严重错误条目数 / 总回答条目数；

$$
R1 = \max\left(0, 100 - \frac{Err_{severe}}{Err_{max}} \times 100\right)
$$

其中 Err_max 为可接受上限（如 5%）。

2. R2 负面偏差控制度

◦ NegBias = 主动负面偏差条目数 / 相关回答条目数；

$$
R2 = 100 - NegBias \times 100
$$

3. R3 稳定性得分

◦ 对关键指标在多次测量中的变异系数（CV）进行统计：

$$
R3 = \max\left(0, 100 - \frac{CV}{CV_{max}} \times 100\right)
$$

4. R4 公平性与操纵疑似度

◦ 根据公开题与暗题表现差异、模型间异常差异等，计算操纵疑似度 $Susp \in [0,1]$：

$$
R4 = 100 - Susp \times 100
$$

综合公式：

$$
AIR = u_1 \cdot R1 + u_2 \cdot R2 + u_3 \cdot R3 + u_4 \cdot R4
$$

可采用 (u_1 = 0.30, u_2 = 0.30, u_3 = 0.20, u_4 = 0.20)。

4.5 UAF：用户–AI 对齐系数

UAF 用于刻画 AI 表现对现实业务影响的程度，考虑以下因素：

· 行业/品类 AI 使用渗透率（Usage）；

· 目标人群对 AI 推荐的信任度（Trust）；

· 品牌自有客户群对 AI 的使用频率（BrandUsage）。

各项归一到 [0,1] 后：

[

UAF_{raw} = 0.4 \cdot Usage_{norm} + 0.3 \cdot Trust_{norm} + 0.3 \cdot BrandUsage_{norm}

]

为避免过度放大或缩小，UAF 在 [0.7, 1] 区间内映射：

[

UAF = 0.7 + 0.3 \cdot UAF_{raw}

]

4.6 MCI：方法置信度指数

MCI 反映本次测量的可靠性，从以下四个维度评估：

· S：样本完备度（题量与覆盖度）；

· Q：数据质量（有效回答率、噪声水平）；

· T：稳定性（重复测量误差）；

· M：模型覆盖度（是否覆盖主要主流模型）。

四项均以 0–100 分计：

[

MCI_{raw} = 0.25S + 0.25Q + 0.25T + 0.25M

]

映射到 [0.8, 1] 区间：

[

MCI = 0.8 + 0.2 \cdot \frac{MCI_{raw}}{100}

]

在对外展示时，可结合 MCI 提示相应的置信区间。

4.7 综合指数（可选）

在需要形成单一综合指标用于内部管理或对外传播时，可在同时展示 AIP、AIC、AIR、UAF、MCI 的前提下，计算 AIBP 2.0 综合指数：

[

AIBP2 = MCI \times \left[\lambda \cdot AIP + (1-\lambda)\cdot \left(\alpha\cdot AIC + \beta \cdot AIR \right) \right] \times UAF

]

· 建议 (\lambda) 取值范围 0.5–0.7（如 0.6）；

· (\alpha + \beta = 1)，可取 (\alpha = 0.6,\ \beta = 0.4)。

第五章 数据采集与测试程序

5.1 多模型测试设计

1. 模型选择

◦ 覆盖国内主流通用对话模型和搜索增强型模型；

◦ 按需纳入重要行业垂直模型。

2. 模型参数与环境记录

◦ 记录模型名称、版本号；

◦ 明确调用方式（API 或 UI）、温度、top-p 等关键参数；

◦ 记录测试时间窗口，以控制版本变动影响。

5.2 测试题集构建

1. 题型组成

◦ 认知类：品牌基础信息与核心要素；

◦ 推荐类：品牌/产品推荐任务；

◦ 场景类：真实生活或业务场景问题。

2. 题源构成

◦ 专家设计的标准问题；

◦ 来自真实搜索与问答语料的抽样与规范化问题；

◦ 用户调研中收集的自然语言表达。

3. 公开题与暗题

◦ 保持适当比例的公开题型样例，提高透明度；

◦ 正式评测题集中应包含一定数量的暗题和随机题，用于降低操纵风险。

5.3 测试流程

可参照以下步骤实施：

1. 依据统一规范生成测试题集；
2. 在指定模型上标准化执行批量问答；
3. 记录并存档全部输出文本；
4. 对回答进行结构化抽取和标注（含品牌识别、要素信息、场景归类等）；
5. 按指标定义计算各项得分；
6. 进行 QA/QC 复核，生成 AIP、AIC、AIR 及校准因子；
7. 对结果进行方法置信度评估，形成 MCI。

5.4 数据清洗与质量控制

1. 数据清洗

◦ 去重：剔除完全重复回答；

◦ 噪声过滤：剔除乱码、与问题完全无关内容；

◦ 无效回答判定：模型拒答或要求用户换问题等情形。

2. 防操纵与异常检测

◦ 对公开题与暗题表现进行对比分析；

◦ 对模型间差异进行异常检测，识别特定品牌的异常偏好；

◦ 对操纵疑似度较高的品牌，可进行复核或限制其结果公开形式。

3. 质量保证

◦ 标注环节采用多人交叉复核；

◦ 对关键指标进行抽样复查；

◦ 对模型波动较大的题目进行单独处理，不直接计入品牌表现。

第六章 结果表达与等级划分

6.1 结果表达

对每个被评价品牌，建议至少公开以下内容：

1. AIP：AI 表现基础指数；
2. AIC：AI 建设指数；
3. AIR：AI 风险与稳定性指数；

4. UAF：用户–AI 对齐系数；
5. MCI：方法置信度指数；
6. （如使用）AIBP2 综合指数及相应置信区间。

6.2 等级划分示例

为便于理解与比较，可对主指数进行等级划分：

· AIP 等级（表现）

◦ A：≥85

◦ B：70–84

◦ C：55–69

◦ D：<55

· AIR 等级（风险）

◦ 低风险：≥85

◦ 中风险：70–84

◦ 较高风险：55–69

◦ 高风险：<55

· AIC 等级（能力）

◦ 建设领先：≥80

◦ 建设中等：60–79

◦ 建设起步：<60

等级边界可结合实际数据分布进行适度调整。

第七章 权重与方法校准

7.1 权重设置原则

· AIP、AIC、AIR 内部及之间的权重为基于专业判断的初始参数；

· 权重设置应兼顾指标重要性、稳定性和可区分度；

· 权重不应被视为固定不变，应随数据积累和研究推进进行周期性校准。

7.2 校准依据

· 外部品牌力评价结果；

· 品牌财务与市场表现（市值、营收增速、市场份额等）；

· 用户调研数据（品牌认知度、好感度、推荐意愿等）；

· 历史 AIBP 指数数据与实际发展轨迹的对比。

7.3 校准方法

· 使用因子分析、主成分分析验证指标聚类结构；

· 使用相关分析和回归分析检验各维度对外部变量的解释度；

· 根据解释度与稳定性调整：

◦ 各维度权重；

◦ 二级指标权重；

◦ 必要时合并或删减部分指标。

第八章 使用规范与合规要求

8.1 指数使用边界

应明确告知社会各方：

> AIBP 2.0 反映的是品牌在主流 AI 模型中的语义表现及相关建设情况，属于 “AI 领域的品牌表现评价”，不等同于对品牌整体商业价值、社会价值或合规性的全面评价。

8.2 品牌与媒体引用规范

1. 在对外传播中引用指数时，应注明：

◦ 指数名称（如 AIP、AIC、AIR 或 AIBP 2.0 综合指数）；

◦ 评价年份与版本号；

◦ 评价与发布机构名称。

2. 不应采用易引起误解的绝对化宣传，例如：

◦ “全国第一” “绝对领先”等缺乏上下文的表述；

◦ 混淆不同指数或等级。

3. 出现重大误用或歪曲引用时，指数发布机构有权要求更正或公开说明。

8.3 争议处理与申诉机制

· 品牌可在结果正式发布前，对自身数据和描述进行核对；

· 对事实性错误，有权提交证据并申请修正；

· 对涉及重大负面风险的内容，可申请技术与专家复核；

· 申诉流程与时限应在相关实施细则中予以明确。

8.4 低分与高风险品牌的信息披露

· 对 AIR 分值较低且风险明显的品牌，可优先采取非公开方式反馈结果，仅在行业层面披露风险比例与典型问题；

· 对经多次复核仍存在显著风险的品牌，如需公开披露，应在充分告知和合规前提下进行。

第九章 实施与版本管理

9.1 版本说明

· 本文件为 AI 领域品牌表现指数体系（AIBP 2.0）技术说明文档；

· 后续可根据实践应用与实证研究成果，形成 2.1、2.2 等滚动优化版本。

9.2 国内与国际应用

· AIBP 2.0 主要面向中国境内主流模型与中文场景的评价和应用；

· 针对出海品牌和全球化需求，可在合规前提下构建 AIBP-Global 版本，纳入海外模型进行专题研究。

附录

附录 A：二级指标与数据字段对照表

附录 B：样本抽取与题集设计规范

附录 C：指数计算示例（品牌样例）

附录 D：QA/QC 质量控制流程说明

附录 E：起草单位与专家组名单

本文件自发布之日起，可作为 AI 领域品牌表现指数项目及相关报告编制、团体标准制定与企业自评实践的统一技术依据。

起草方：四川远见行人工智能科技有限公司

2025-12-12

（注：文档部分内容可能由 AI 生成）