竹势

040 / EVOLUTION-SELECTION LAB

AI MARKETING EVOLUTION / EVALUATE · SELECT · ROLLBACK

会进化的 AI 市场部

评测、反馈与达尔文式优化

不让系统自由自改；让候选在多目标适应度、真实业务证据和可回滚发布控制中接受选择。

建立候选谱系

变更动机、父版本、数据快照与风险标签可追溯

运行四级证据

离线、影子、在线、经营结果逐层晋级

配置探索预算

信息价值、可逆性、风险半径与机会成本共算

守住晋级与回滚

人工盲评、有限流量、停止条件和降级完整

shichangbu.ai

竹势 AI 营销智库出品

BOARD BRIEF 01

执行摘要：十项可证伪的高管判断

竹势 AI 营销智库出品

EVOLUTION THESIS / 董事会判断

真正会进化的 AI 市场部，不让系统自由改写自己；它让可追溯候选在多目标适应度、分层证据、探索预算与发布控制中接受选择。

1. 反馈不会自动变成学习

若系统没有候选版本、对照证据、选择门和继承记录，所谓“越用越强”只能被审计为不可解释漂移。董事会应要求每次能力变化都能回答：改了什么、基于哪类证据、谁批准、影响了哪些场景、如何回滚。

2. 适应度必须是多目标约束

单一点击率可以把系统推向夸张标题，单一转化率可能牺牲毛利和客户关系。真正可晋级的候选必须同时跨过任务质量、可靠性、单位经济性、业务增量、品牌合规和长期资产六道门。

3. 离线高分只是准入证

黄金集、边界集和回放测试用于筛除明显退化，不能证明真实增量。任何影响客户、预算或公开表达的候选，都要经过影子、有限流量和经营结果观察。

4. 反馈先分类再使用

显式评价、点击、销售结果、投诉与风险事件的生成机制不同。未经来源、偏差、时滞和适用范围标记的反馈，不得直接进入训练、提示词修订或自动晋级。

5. 探索预算是资本配置

探索会消耗流量、毛利、客户耐心和管理注意力。董事会应把可承受损失、最大暴露、样本上限、停止条件与失败学习价值写入预算，而非只批准技术费用。

6. 实验方法必须匹配干扰结构

A/B 适合稳定随机化；交错适合查询或排序的短周期比较；bandit 适合快速收益分配但弱于因果解释；贝叶斯优化适合昂贵连续参数；高声誉、高合规场景仍应由人工裁决。

7. 冠军必须可复现

一个版本在偶然流量、特定模型、特定缓存或异常平台环境中胜出，不能成为组织标准。晋级包必须冻结运行清单、数据快照、资源预算与评测配置，并完成重复试验。

8. 风险门拥有否决权

法律责任、经营授权、客户权益、品牌承诺与安全控制不是适应度竞争项，而是不可被优化器交换的硬约束。任何候选触发红线，应隔离、降级或退役。

9. 长期结果需要延迟账本

线索质量、复购、品牌搜索、投诉和毛利常在数周后显现。系统必须将即时代理指标与延迟经营结果连接到同一谱系，否则短期冠军会吞噬长期资产。

10. 组织进化依赖目录与退役

成功候选若没有进入受控目录、复用条件、责任人、版本有效期和退役机制，就只是一次实验。真正的组织学习是把可重复胜出转化为可调用、可审计、可退出的能力资产。

高管判断必须能够被数据和事件推翻，否则只是口号。下表把每项判断转成审议问题，使管理层可以在季度会上要求明确证据。

表 1 | 十项判断的董事会证伪条件

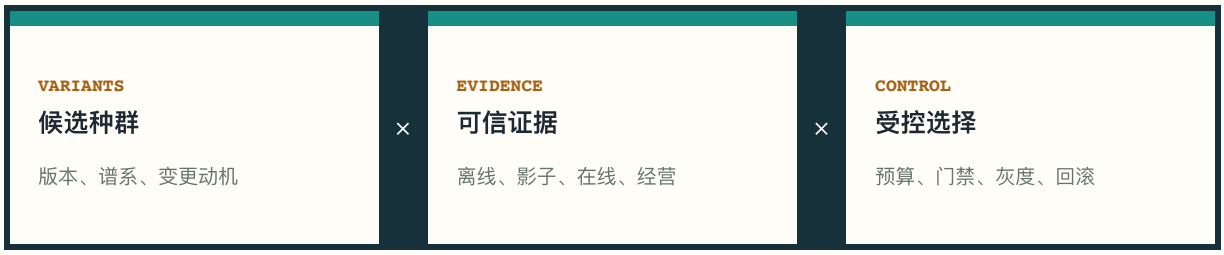
判断	证伪问题	必须提交的证据
反馈不会自动变成学习	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
适应度必须是多目标约束	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
离线高分只是准入证	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
反馈先分类再使用	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
探索预算是资本配置	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
实验方法必须匹配干扰结构	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
冠军必须可复现	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
风险门拥有否决权	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
长期结果需要延迟账本	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录
组织进化依赖目录与退役	在什么观察下该判断应被推翻或修正?	基线、候选清单、实验设计、风险记录、经营结果与复现记录

董事会不需要审批每个提示词，却必须审批决定权、风险容忍度和资本暴露。证伪条件的价值在于阻止团队只展示胜利样本。

BOARD BRIEF 02

第一章 经营命题：反馈不等于学习

竹势 AI 营销智库出品



本章判断

反馈是观测，学习是受控选择。市场部每天产生点击、停留、回复、转化、投诉和人工评分，但这些信号只有在版本、对照、选择与继承闭环中才形成组织能力。

董事会应把本章问题理解为经营控制，而非模型调参。反馈是观测，学习是受控选择。市场部每天产生点击、停留、回复、转化、投诉和人工评分，但这些信号只有在版本、对照、选择与继承闭环中才形成组织能力。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

EVOLVE-8 经营进化链的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

OpenAI 的评测指南把评测对象拆成数据、测试逻辑和评分标准，Anthropic 的代理评测实践进一步要求观察任务是否完成、工具调用是否合理以及执行轨迹是否稳定。对本章的经营命题而言，这些材料支持一个可检验判断：只有当候选版本、运行记录、评分结果和后续处置形成闭环，反馈才可能被转化为组织学习；单独收集点赞、点击或人工意见，只会形成意见池。

在内容营销场景里，编辑对一篇文章的“满意”通常混合了事实准确、品牌语气、结构清晰和发布时间等多种因素。若系统只把最终采用与否写成一个二元标签，下一轮优化无法判断究竟应调整提示词、知识检索、审核规则还是交付节奏。更稳妥的做法是将反馈拆成可定位字段，并把每项反馈绑定到具体运行清单。

本章结论不适用于把任何单次正反馈直接视为学习完成。客户临时偏好、热点流量和渠道算法都可能造成偶然胜出；没有基线、对照和复现条件时，所谓“学会了”只能表述为“在某次运行中表现较好”。企业还应禁止将投诉、敏感信息或未经授权的客户对话直接回灌为训练材料。

董事会应要求营销运营负责人建立“反馈到变更”的闭环台账：每条高价值反馈必须标明来源、对象、责任人、拟议变更和验证方式；每月复盘中只讨论完成验证的改动，未通过证据门的建议保留为候选而不进入生产目录。这样可把学习速度与变更纪律同时纳入经营审议。

EVOLVE-8 经营进化链用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 2 | EVOLVE-8 经营进化链

环节	核心动作	经营产物
E 设定环境	定义业务目标、约束与流量环境	场景契约
V 生成变异	建立受控候选而非直接改线上	候选包
O 组织观测	记录质量、成本、风险和业务反馈	观测账本
L 分层评测	离线、影子、在线、长期逐级取证	证据包
V 执行选择	按硬门、效用和不确定性决策	晋降级决定
E 有限暴露	灰度、限额、人工门和回滚	发布记录
8 继承复用	写入目录、适用范围和有效期	能力资产
循环校准	环境变化后重新建立挑战	新一轮基线

使用EVOLVE-8 经营进化链时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

反馈—学习差异用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 3 | 反馈—学习差异

对象	反馈状态	学习成立条件
点击上升	观察性信号	有对照、无污染且与经营目标一致
人工好评	主观证据	盲评、评分标尺和一致性检查
销售成交	延迟结果	可连接到触达版本并处理选择偏差
投诉事件	风险证据	优先触发隔离，不用收益抵消

使用反馈—学习差异时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

经营闭环责任用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 4 | 经营闭环责任

环节	业务责任人	不得外包的判断
目标	CMO/业务负责人	价值函数与风险容忍度
评测	实验与数据负责人	可比性和统计有效性
发布	平台负责人	流量、权限与回滚
继承	能力目录负责人	适用边界与退役

使用经营闭环责任时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

图表规格 1 | EVOLVE-8 经营进化链

建议将“EVOLVE-8 经营进化链”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 03

第二章 隐喻边界：达尔文式管理语言的可用与不可用

竹势 AI 营销智库出品

METAPHOR BOUNDARY / 隐喻边界

能力可以变异，责任不可漂移；候选可以竞争，客户权益不可试错；系统可以学习，经营授权不能被绕过。

本章判断

种群、变异、适应度、选择、继承与环境可帮助管理者描述候选竞争；责任、权利、承诺和法定义务必须冻结。

董事会应把本章问题理解为经营控制，而非模型调参。种群、变异、适应度、选择、继承与环境可帮助管理者描述候选竞争；责任、权利、承诺和法定义务必须冻结。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

隐喻—治理边界图的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

Darwin 的自然选择描述的是种群在环境约束下出现差异并被选择的机制，而不是个体按目标主动设计自身。AlphaEvolve 等系统之所以能在算法搜索中取得结果，关键也不在“自由进化”，而在于候选生成器、可执行验证器和严格筛选共同限定搜索空间。本章据此将生物学概念仅用于解释候选竞争，不把责任、授权和价值判断交给算法。

达尔文的原始问题是种群在有限环境中为何呈现差异性存续，而不是组织如何让软件自行改写。迁移到营销管理时，选择机制只能解释“多个受控候选如何在约束下竞争”，不能把主体性、责任或伦理交给系统。管理修正是先定义不可变边界，再允许低风险资产产生差异。

品牌审校智能体可以产生不同措辞、规则顺序和检索策略的候选，但广告法禁区、客户授权范围和品牌承诺不应作为可竞争变量。例如医疗健康内容即使点击更高，只要弱化风险提示或扩大功效表述，就必须被风险门直接淘汰；这里不存在以市场表现抵消合规缺陷的空间。

这一隐喻的边界还包括环境不可等同于“平台当前偏好”。平台分发机制只是经营环境的一部分，利润、客户信任、员工负荷和监管要求同样构成选择条件。若企业只让高互动内容留下，系统会把噪声、争议乃至误导当作适应性，最终破坏长期资产。

CEO 应批准一份变异权限矩阵，逐项标明哪些对象可自动产生候选、哪些只能由授权人员修改、哪些必须永久冻结。法务和品牌负责人拥有不可变控制的定义权，平台团队只负责实现约束；任何试验申请若触及责任归属、客户权益或强制披露，应直接转入人工裁决而非进入自动竞技。

达尔文隐喻映射用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 5 | 达尔文隐喻映射

生物学概念	管理映射	边界修正
种群	同一任务的多个候选版本	候选必须由人定义范围
变异	提示词、模型或流程的受控改变	高风险政策不可自由变异
适应度	多目标经营表现	不得等同单一点击率
选择	晋级、降级、淘汰	由治理规则和人裁决
继承	复用胜出组件与证据	继承需保留谱系和有效期
环境	渠道、客户、竞争和监管条件	环境变化会使冠军失效

使用达尔文隐喻映射时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

冻结控制清单用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 6 | 冻结控制清单

控制对象	冻结原因	变更权
法律责任	不能由算法转移	法务与董事会
预算上限	防止探索吞噬经营资源	财务与业务负责人
客户授权	保护权益与信任	业务、法务
品牌承诺	避免短期指标破坏长期资产	品牌负责人
安全权限	限制爆炸半径	安全与平台

使用冻结控制清单时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

可变—不可变判断用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 7 | 可变—不可变判断

问题	允许候选化	必须冻结
表达方式	标题、结构、语气可测试	事实、禁语、授权
模型选择	路由与参数可测试	数据域和访问边界
工作流	可逆节点可调整	审批、审计、回滚节点

使用可变—不可变判断时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例一 | Google DeepMind AlphaEvolve：验证器决定“进化”能走多远

2025年5月，Google DeepMind公开 AlphaEvolve，将大模型生成候选程序、自动评估器验证以及进化式选择组合起来，用于算法发现和基础设施优化。其公开材料强调，系统不是凭语言说服自己，而是通过可运行、可计分的验证器比较候选。

具体动作包括从基线代码产生变体、在自动评测环境中执行、保留较有希望的候选并继续搜索。2026年的影响更新又披露其在数据中心、芯片设计与训练流程等环境中的应用，但这些结果仍属于 Google 披露的特定任务。

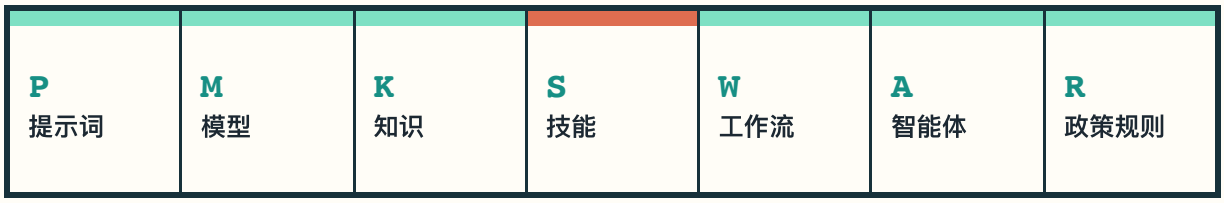
作用机制在于目标可以被机器验证：运行时间、正确性或资源消耗能够形成强适应度信号。对营销的迁移必须收缩，因为品牌信任、客户关系和增量收入往往不能由一个即时验证器完整表达。

管理启示是：先投资验证器和边界，再谈“自进化”。当业务结果不可即时验证时，应降低自动选择权，以人工盲评、有限流量和延迟经营结果补足。

BOARD BRIEF 04

第三章 进化单元：七类资产与不可变运行清单

竹势 AI 营销智库出品



本章判断

提示词、模型、知识、技能、工作流、智能体和政策规则应分别版本化，任何线上执行都由一份不可变运行清单绑定。

董事会应把本章问题理解为经营控制，而非模型调参。提示词、模型、知识、技能、工作流、智能体和政策规则应分别版本化，任何线上执行都由一份不可变运行清单绑定。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

七类进化单元的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

持续交付实践强调可复现构建：同一版本应能由确定的代码、依赖和配置重新生成。迁移到 AI 市场部后，单独记录模型名称远远不够，必须同时固定提示词、知识快照、工具版本、工作流、智能体配置和政策规则。OpenAI 与 Anthropic 的评测工程材料均表明，执行环境和 harness 的变化足以改变结果，因此运行清单是评测可信度的基础。

以投放诊断为例，团队可能只认为“更换了模型”，实际同时更新了账户字段映射、预算规则、异常阈值和建议模板。若这些单元没有分别版本化，线上改善无法归因，回滚时也无法恢复原状态。七类资产必须拥有独立版本号，再由一次运行清单把它们组合成不可变快照。

进化单元不能无限细分到失去管理意义。过度拆分会让版本依赖爆炸，使测试成本和审批负担超过收益；过度合并则会让任何小改动都触发整套系统重测。企业应以“能否独立改变结果、能否独立回滚、是否拥有不同责任人”为拆分准则。

平台负责人应建立七类资产目录和组合发布单：每次生产运行都生成唯一 run ID，记录各单元版本、数据快照、权限和成本策略。营销运营在周度评审中只允许完整清单参与比较；发现异常时，值班人员按清单回滚，不得临时拼接旧提示词与新规则形成无法追踪的混合版本。

七类进化单元用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 8 | 七类进化单元

单元	可变面	关键风险
提示词	结构、示例、约束	提示注入、指标投机
模型	版本、参数、路由	能力漂移、成本突变
知识	来源、切片、检索	过期、污染、授权
技能	工具调用、输入输出	越权、执行副作用
工作流	节点、顺序、重试	异常放大、状态错乱
智能体	角色、计划、记忆	目标漂移、责任模糊
政策规则	阈值、白名单、人工门	规则冲突与规避

使用七类进化单元时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

不可变运行清单字段用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 9 | 不可变运行清单字段

字段	示例	审计用途
manifest_id	run-20260723-0042	串联全部日志
asset_versions	prompt p18 / model m7	复现组合
data_snapshot	golden-2026Q3-v2	确认评测数据
policy_bundle	brand-safe-v5	证明边界
resource_budget	30s / ¥0.8 / 12 calls	保证可比性

使用不可变运行清单字段时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

组合变更风险用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 10 | 组合变更风险

变更范围	默认策略	原因
单一组件	允许挑战者测试	便于归因
两个组件	需析因设计或分阶段	避免交互混淆
全链路	仅影子或沙箱	难以定位退化

使用组合变更风险时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

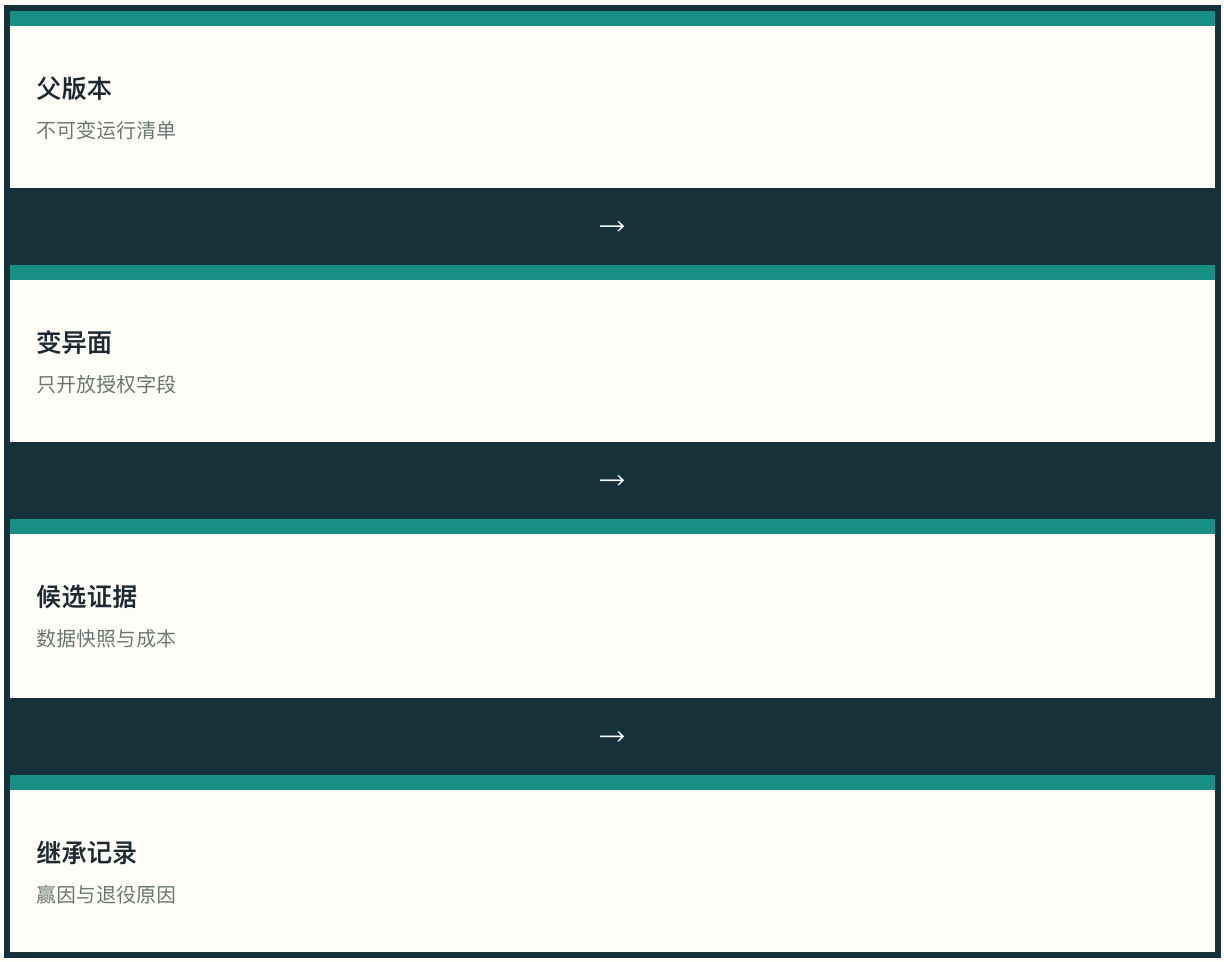
图表规格 3 | 七类进化单元

建议将“七类进化单元”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 05

第四章 候选与谱系：把每次改变变成可追溯资产

竹势 AI 营销智库出品



本章判断

候选必须记录父版本、变更动机、变异面、数据快照、成本、风险和责任人；谱系不是技术日志，而是经营问责链。

董事会应把本章问题理解为经营控制，而非模型调参。候选必须记录父版本、变更动机、变异面、数据快照、成本、风险和责任人；谱系不是技术日志，而是经营问责链。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

候选谱系卡的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

Microsoft ExP 的实验平台研究反复强调实验可信度、元数据和可复现性，说明候选管理首先是数据治理问题。候选若没有父版本、变更理由、目标指标和暴露范围，后续即使获得显著结果，也难以判断是设计改进、样本偏差还是并行变更造成。谱系记录因此既服务分析，也服务责任追踪。

企微跟进智能体新增“自动提醒销售”功能时，候选卡应明确它继承哪个父版本、修改了何种触发规则、使用哪一批客户状态数据、会增加多少消息成本，以及可能造成哪些打扰风险。上线后出现投诉，团队可沿谱系定位到具体规则，而不必笼统停用整个智能体。

谱系完整不等于结果可信。一个记录齐全的候选仍可能基于污染数据、错误标签或不合适的基线；反过来，缺失关键元数据的高表现候选也不应补录后直接晋级。企业需要把“可追溯”设为进入评测的先决条件，而不是事后整理文档。

营销运营负责人应把候选卡设为变更入口，禁止通过聊天消息或临时脚本绕过登记。候选评审至少核对父版本、变异面、数据快照、预期收益、风险标签、成本上限和回滚对象；缺一项即退回。季度审计抽查谱系能否从冠军追溯到原始需求和每次晋降级决定。

候选谱系卡用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 11｜候选谱系卡

字段	必须回答的问题	拒绝条件
父版本	从哪个基线产生	无父版本且非新建
变更动机	要修复哪类失败	只写“提升效果”
变异面	具体改了哪些资产	无法定位差异
数据快照	使用何时何地数据	混入未来或线上反馈
成本标签	推理、人工、流量成本	成本不可计量
风险标签	影响客户、预算、公开内容否	未完成分级

使用候选谱系卡时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

谱系关系类型用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 12 | 谱系关系类型

关系	用途	管理含义
派生	直接从父版本修改	可做局部回滚
合并	组合多个胜出组件	必须重新评测交互
回退	恢复历史稳定版本	保留事故上下文
分叉	面向不同场景形成版本	禁止误称统一冠军

使用谱系关系类型时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

变更动机分类用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 13 | 变更动机分类

类别	典型触发	优先证据
质量缺陷	事实错误、格式失败	黄金集与盲评
效率缺陷	成本或时延过高	运行日志
业务缺陷	增量弱或线索质量低	在线实验与CRM
风险缺陷	投诉、违规、越权	事件与审计日志

使用变更动机分类时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例二 | Microsoft ExP：可信实验平台先于实验民主化

Microsoft Research长期披露 ExP 平台，核心原则是 trustworthiness 与 scalability。平台服务多个产品形态，要求随机化、数据质量、指标解释和分析流程共同可信。

具体动作包括统一实验分流、指标计算、质量检查、方差缩减和知识沉淀。相关论文也讨论遥测丢失、指标误读和大量实验带来的分析陷阱，说明平台规模越大，错误结论的工业化风险越高。

作用机制不是让更多人随意试验，而是把可信检查嵌入基础设施，让一次产品变化能够被稳定比较。公开材料能证明平台设计与方法，不能直接证明任何企业复制后会获得同等收益。

管理启示是把实验平台定位为决策控制系统。实验数量、胜率和发布速度都不是北极星；可信数据、可解释差异和失败知识的复用才是组织资产。

BOARD BRIEF 06

第五章 六层适应度函数：拒绝单指标冠军

竹势 AI 营销智库出品

01	02	03	04	05	06
任务质量	运行可靠	单位经济	业务增量	品牌与合规	长期资产

本章判断

适应度应同时覆盖任务质量、运行可靠性、单位经济性、业务增量、品牌合规风险和长期资产。硬门与权重不可混为一谈。

董事会应把本章问题理解为经营控制，而非模型调参。适应度应同时覆盖任务质量、运行可靠性、单位经济性、业务增量、品牌合规风险和长期资产。硬门与权重不可混为一谈。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

六层适应度函数的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

Goodhart 与 Campbell 所揭示的共同问题是：一旦指标成为强控制目标，参与者和系统就会寻找提高指标而非改善真实结果的路径。营销场景尤其容易发生这种偏移，因为点击、互动和线索数反馈快，而利润、品牌和客户关系反馈慢。六层适应度函数用并行约束替代单一总分，正是为了阻止短期指标吞没经营目标。

Goodhart 与 Campbell 所揭示的共同风险是：指标一旦成为强控制目标，参与者和系统都会围绕指标结构适应，指标与真实目标的关系随之削弱。营销系统尤其容易把点击、回复率或短期转化优化成表面胜利，因此适应度必须使用硬门、护栏指标和延迟经营结果。

内容生成候选可能同时提高发布速度和点击率，却增加事实错误、品牌偏差与人工返工；线索评分候选可能提高销售接通率，却把高价值但决策周期长的客户排除。六层函数要求团队同时查看任务质量、可靠性、单位经济性、业务增量、风险和长期资产，而不是把所有维度压成一个容易被投机的分数。

多目标评测并不意味着所有指标永远等权。不同阶段可设主目标和护栏，但风险底线、法定义务与客户授权不得被权重稀释。长期资产指标也不应被伪造为即时精确数值；在缺乏成熟测量时，可采用审校通过率、品牌偏差事件和知识复用率等可审计代理，并明确其局限。

CFO 与 CMO 应共同批准适应度卡：财务定义单位经济性和增量口径，品牌与法务定义不可补偿的护栏，营销运营维护任务质量与长期资产指标。月度冠军评审必须展示六层结果及权衡理由；任何仅凭单一指标获胜的候选不得进入目录。

六层适应度函数用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 14 | 六层适应度函数

层级	代表指标	门槛性质
任务质量	准确、完整、相关、可执行	最低质量门
运行可靠性	成功率、时延、恢复率	SLO硬门
单位经济性	单任务成本、人工复核成本	预算门
业务增量	增量线索、毛利、留存	价值门
品牌/合规风险	事实、承诺、投诉、越权	一票否决
长期资产	知识沉淀、可复用率、品牌健康	延迟评价

使用六层适应度函数时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

硬门与加权效用用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 15 | 硬门与加权效用

机制	适用对象	错误用法
硬门	安全、合规、SLO	用收益抵消红线
加权效用	通过硬门后的候选排序	权重频繁随结果调整
帕累托前沿	多目标难统一时	宣称唯一最优

使用硬门与加权效用时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

指标口径卡用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 16 | 指标口径卡

字段	内容	治理动作
定义	分子、分母、时间窗	版本化
方向	越高越好或区间最优	防止误读
护栏	不可退化指标	自动停止
延迟	结果成熟所需时间	延迟结算

使用指标口径卡时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例三 | Booking.com：实验民主化需要失败仓库与护栏

Booking.com团队在2017年论文中回顾超过十年的在线实验实践，强调中央成功与失败仓库、通用代码库、数据管道透明监控以及端到端责任。

具体动作是把实验能力下放，同时通过基础设施和制度提供保护，使团队能够提出假设、运行实验、观察质量并记录结果。其公开研究主要聚焦数字产品环境，未披露将同一方法直接用于高风险自动营销触达。

作用机制是降低试验摩擦并让失败可见，避免组织只记住少数成功。民主化有效的前提是实验单位可随机化、指标可靠、用户伤害可控。

管理启示是：挑战者数量可以增加，但发布权、风险否决和证据标准不能下放到同一人。失败进入知识库后，才能减少重复试错。

图表规格 5 | 六层适应度函数

建议将“六层适应度函数”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 07

第六章 四级证据阶梯：从离线回放到长期经营结果

竹势 AI 营销智库出品

L1
离线回放

L2
影子证据

L3
在线实验

L4
经营结果

本章判断

证据强度随环境真实性、对照质量、可重复性和时间跨度上升。离线评测筛错，影子验证运行，在线实验识别差异，长期结果检验经营价值。

董事会应把本章问题理解为经营控制，而非模型调参。证据强度随环境真实性、对照质量、可重复性和时间跨度上升。离线评测筛错，影子验证运行，在线实验识别差异，长期结果检验经营价值。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

四级证据阶梯的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

NIST 的 TEVV 思路和 AI RMF 均要求根据使用情境实施测试、评估、验证和确认；实验平台研究则说明离线结果与线上业务结果之间存在多层证据差距。本章据此把证据划分为离线、受控运行、在线因果和长期经营四级，并要求候选按风险逐级取得资格，而非用一次离线高分替代上线证明。

Deming 的 PDSA 强调在系统中以预测、试验、学习和再行动改进质量。它反对把单次结果当成永久真理，也要求区分普通波动与特殊原因。迁移到 AI 市场部，评测阶梯就是把“计划—执行—研究—行动”写成可复现证据链，而不是持续随机改提示词。

品牌审校规则更新后，团队可先用历史黄金集回放，再用边界集检查极端表述；通过后进入影子流量，与生产版本并行处理真实 Brief 但不影响客户；随后再做有限在线试验。高风险的自动触达还需观察投诉、退订和后续客户关系，不能在短期回复率上升后立即扩围。

证据阶梯不是固定线性流程。低风险、可逆的小改动可以缩短影子阶段，高风险或样本稀少的场景则可能长期停留在人工盲评与专家裁决。在线显著性也不能弥补离线安全缺陷；任何一级发现硬性风险，候选都应退回而非依靠后续平均表现抵消。

平台团队应为每类场景定义最低证据级别和升级条件，营销负责人负责样本代表性，风险团队拥有越级阻断权。每次晋级记录所用数据、评测器、置信区间、暴露比例和观察期；董事会只审议达到对应证据级别的扩围申请。

四级证据阶梯用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 17 | 四级证据阶梯

级别	证据形态	可支持决策
I 离线	黄金集、边界集、回放	允许进入影子
II 运行	沙箱、影子、人工盲评	允许有限流量
III 因果	随机实验、交错、准实验	支持局部晋级
IV 经营	毛利、复购、品牌、风险长期结果	支持目录化与扩围

使用四级证据阶梯时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

离线评测组合用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 18 | 离线评测组合

组件	目的	常见盲点
黄金集	代表高频真实任务	随业务变化失真
边界集	覆盖禁区和异常	只测已知风险
回归集	阻止旧能力退化	污染与记忆答案
对抗集	寻找脆弱点	不代表真实频率

使用离线评测组合时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

证据升级门用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 19 | 证据升级门

当前级	升级条件	不得跳过
离线→影子	质量门、成本门通过	高风险边界集
影子→在线	无副作用、日志完整	回滚演练
在线→经营	统计与业务显著	延迟结果观察

使用证据升级门时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例四 | Netflix 个性化：短期互动必须接受长期体验约束

Netflix技术文章披露其长期把个性化推荐作为核心产品能力，并通过在线比较改进首页排序。2025年的推荐基础模型文章强调利用更完整的会员交互历史，同时面对服务延迟和训练成本约束。

具体动作包括对候选排序策略进行离线建模、在线实验和服务监控，并把用户历史、内容特征和实时环境纳入推荐。公开材料没有提供可直接迁移到企业营销的统一提升百分比。

作用机制是通过行为反馈改善发现体验，但观看、点击和长期满意之间并不等价。算法若只追求即时播放，可能强化窄化偏好或忽略长期留存。

管理启示是建立分层适应度：即时采用率用于快速诊断，长期留存和体验投诉用于最终选择。候选必须按人群与环境保留适用范围。

图表规格 6 | 四级证据阶梯

建议将“四级证据阶梯”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 08

第七章 反馈工程：来源、偏差、时滞与可用范围

竹势 AI 营销智库出品

本章判断

显式与隐式、即时与延迟、用户与业务、观察性与实验性反馈必须分账。未经分级的反馈最容易制造错误学习。

董事会应把本章问题理解为经营控制，而非模型调参。显式与隐式、即时与延迟、用户与业务、观察性与实验性反馈必须分账。未经分级的反馈最容易制造错误学习。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

反馈四象限的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

Google Ads 的增量测量文档区分平台可观察转化与因果增量，提醒企业不能把归因报告直接当作真实影响。反馈工程还要处理选择偏差、幸存者偏差和延迟标签：主动评价的用户往往并不代表全部用户，成交数据只覆盖被销售持续跟进的人群，风险事件则可能在数周后才显现。

因果推断区分相关、预测与干预效果。点击和成交之间即使高度相关，也可能共同受到渠道、人群或销售能力影响。反馈工程的管理任务，是记录暴露、选择机制与时间顺序，为随机实验或可信准实验保留条件。

活动复盘智能体收到的反馈包括编辑显式打分、用户隐式停留、销售对线索质量的判断以及后续退款。四类信号的时间和可信度不同：编辑意见可快速修正表达，停留只能作为行为线索，销售判断可能受跟进能力影响，退款则是延迟但高价值的结果。系统必须分别记录来源和可用范围。

反馈不能跨场景无条件复用。某渠道用户对标题的偏好不应直接改写官网品牌语气；销售标记“低质量”的线索也可能只是当前产能不足。未经授权的私聊内容、投诉文本和敏感属性不得进入通用评测集，更不能自动用于模型训练。

数据负责人应建立反馈字典，定义每个信号的采集方式、偏差、时滞、保留期限和允许用途。营销运营每周检查异常反馈分布，销售与客服共同复核高影响标签；只有经过来源校验和去偏处理的反馈，才可进入候选生成或适应度计算。

反馈四象限用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 20 | 反馈四象限

维度	即时显式	即时隐式	延迟显式	延迟隐式
例子	人工评分	点击/停留	投诉/复盘	成交/复购
偏差	评分者风格	曝光与位置	幸存者与记忆	归因与漏记
用途	质量诊断	快速预警	风险纠偏	经营选择

使用反馈四象限时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

反馈事件最小字段用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 21 | 反馈事件最小字段

字段	说明	缺失后果
source	用户、销售、平台或系统	无法判断生成机制
timestamp	发生与入账时间	时滞混淆
exposure_id	关联运行版本	无法归因
selection_context	为何被展示/触达	选择偏差
consent_scope	授权和可用范围	违规再利用

使用反馈事件最小字段时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

反馈可用等级用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 22 | 反馈可用等级

等级	可做什么	禁止动作
F0 原始	观察和排障	自动训练或晋级
F1 清洗	趋势分析	因果断言
F2 对照	候选比较	跨场景泛化
F3 因果/长期	经营选择	忽略风险门

使用反馈可用等级时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例五 | Google Ads 增量测量：平台归因不等于经营增量

Google Ads官方把 Conversion Lift 定义为衡量广告对目标行动的因果增量，并提供用户或地理分组方法。该口径明确区别于平台常规归因报表。

具体动作是构建处理组和对照组，比较广告暴露造成的额外转化或价值。地理实验要求可比区域和足够预算，用户实验则依赖平台可实施的分流与隐私机制。

作用机制是用反事实对照回答“没有这次投放会怎样”。但结果仍受实验时期、平台可观察转化、跨渠道外溢和预算干扰限制，不能自动等于公司总利润增量。

管理启示是把平台实验作为证据阶梯的一部分，并回接订单毛利、退货、复购和品牌结果。供应商报告必须标明测量范围。

图表规格 7 | 反馈四象限

建议将“反馈四象限”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 09

第八章 探索与利用：选择正确的实验工具

竹势 AI 营销智库出品

本章判断

A/B、交错、multi-armed bandit、贝叶斯优化和人工裁决解决不同问题。方法选择取决于可逆性、样本、干扰、延迟和解释要求。

董事会应把本章问题理解为经营控制，而非模型调参。A/B、交错、multi-armed bandit、贝叶斯优化和人工裁决解决不同问题。方法选择取决于可逆性、样本、干扰、延迟和解释要求。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

实验方法选择树的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

探索—利用理论关注有限资源下学习新选项与使用已知优选项之间的取舍。A/B 适合相对稳定、可随机分流的比较；交错实验适合排序和推荐；多臂老虎机可在快速反馈下动态分配流量；贝叶斯优化适合昂贵、连续参数搜索。方法选择必须由干扰结构、反馈速度和风险决定，而非追逐算法新颖性。

探索—利用问题讨论如何在学习未知选项与使用当前最佳选项之间分配资源。多臂老虎机把这一张力形式化，但其收益函数常较短期，且可能牺牲因果解释。营销管理必须在算法收益之外加入客户权益、品牌和毛利约束，并明确探索预算。

广告创意测试若用户互不影响、转化回传较快，可采用 A/B；推荐列表中多个候选争夺同一曝光位时，交错法可能更灵敏；预算与出价参数的组合搜索可用贝叶斯优化提出少量高价值候选。涉及舆情回复或大客户报价时，样本少且损失不对称，人工裁决通常优于自动分流。

多臂老虎机的收益依赖快速、稳定且可观测的奖励。品牌建设、B2B 商机和客户生命周期价值反馈缓慢，不满足这一条件；若强行用点击作即时奖励，算法会过度利用刺激性表达。存在网络效应、库存约束或销售互相影响时，普通随机分流也会低估干扰。

实验负责人应在立项卡中说明随机化单位、干扰来源、反馈时滞、最小可检测效应和最大损失。平台只提供方法库，不自动替业务选择算法；高风险或低样本场景须由品牌、法务和业务负责人共同决定是否改用专家评审。

实验方法选择用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 23 | 实验方法选择

方法	适用条件	主要局限
A/B	随机化稳定、样本充足	速度慢、干扰敏感
交错	查询/排序、短周期可切换	有残留效应时失真
多臂老虎机	收益即时、可快速再分配	因果解释与长期护栏较弱
贝叶斯优化	连续参数、单次试验昂贵	依赖代理模型和先验
人工裁决	高风险、样本小、价值冲突	成本高且需标尺

使用实验方法选择时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

探索预算模型用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 24 | 探索预算模型

预算项	计量单位	上限示例
流量暴露	客户/会话比例	5%起步
经济损失	毛利或媒体费	不超过战役预算2%
风险暴露	投诉/违规事件	红线为0
人工注意力	复核小时	每周20小时
学习期限	自然日/成熟周期	两周或一个销售周期

使用探索预算模型时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

停止条件用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 25 | 停止条件

类型	触发条件	动作
负向护栏	投诉或错误率越界	立即停止
无望停止	达到样本仍低于最小效果	淘汰
成功停止	效果和护栏稳定	进入复现
环境停止	渠道规则或数据中断	冻结结论

使用停止条件时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例六 | 字节跳动研究：双边市场干扰使普通 A/B 失真

字节跳动研究者公开的预算切分实验论文讨论双边在线市场中处理组与对照组相互竞争的问题。普通用户随机化会违反稳定单元假设，使估计产生偏差。

具体动作是围绕预算约束设计切分方式，以降低市场竞争造成的干扰，并通过真实市场实验验证方法。论文证明特定设计在给定假设下更可靠，但不代表所有广告或线索市场均可直接套用。

作用机制在于实验单位之间并非彼此独立：一个候选多拿走预算、库存或客户注意力，会改变另一个候选的结果。营销团队若忽略这种干扰，可能选出只会抢占内部资源的“冠军”。

管理启示是先画出干扰网络，再决定随机化单位。预算、地域、销售人员和库存常比单个用户更适合作为实验边界。

图表规格 8 | 实验方法选择树

建议将“实验方法选择树”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 10

第九章 选择机制：挑战者—冠军与晋降级门

竹势 AI 营销智库出品

本章判断

候选通过分组竞技、停止规则、多重检验、灰度与回滚晋级。冠军不是永久头衔，而是带有效期和适用环境的当前标准。

董事会应把本章问题理解为经营控制，而非模型调参。候选通过分组竞技、停止规则、多重检验、灰度与回滚晋级。冠军不是永久头衔，而是带有效期和适用环境的当前标准。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

挑战者—冠军门的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

持续交付中的金丝雀发布把新版本暴露给少量真实流量，并依靠监控和回滚降低损失；统计过程控制则强调区分随机波动与系统性变化。挑战者—冠军机制结合两者：候选先满足质量和风险门，再按预设样本、停止规则和多重检验控制参与竞争，获胜后仍需分阶段扩大。

统计过程控制提醒管理者不要把每次随机波动都解释为改进，也不要在看到短期领先后立即改变过程。在线实验中的反复窥视、多指标挑选和小样本冠军，本质上都把噪声误认作信号。晋级制度因此需要预注册、控制图式护栏和复现赛。

线索评分挑战者可先在 5% 流量中仅生成建议，不改变销售优先级；通过一致性和偏差检查后，再让部分团队使用新排序。若接通率提升但高价值商机下降，挑战者应降级；若投诉或敏感属性偏差触发红线，则立即隔离，不等待统计窗口结束。

冠军地位只代表在既定条件下暂时优选，不构成永久认证。频繁查看结果并提前停止会夸大胜率，多候选同时竞争会增加假阳性；业务季节性、渠道政策变化和数据漂移也会让历史冠军失效。无法复现完整运行清单的候选不得晋级。

增长负责人应为每场竞技预先登记主要指标、护栏、样本量、停止条件和回滚阈值；平台自动执行流量切分与多重检验校正，风险团队监控硬门。月度目录会只晋级可复现且通过多目标评审的版本，同时为老冠军设置复验和退役日期。

挑战者—冠军晋降级门用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 26｜挑战者—冠军晋降级门

关口	挑战者要求	冠军义务
登记	谱系完整、风险分级	提供稳定基线
资格赛	离线与边界集通过	不得阻碍比较
影子赛	运行可靠且无副作用	提供相同资源预算
在线赛	对照可信、护栏稳定	接受降级
复现赛	跨时段/分群重复	保留有效期
目录化	文档、责任、回滚齐全	持续监控

使用挑战者—冠军晋降级门时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

多重检验控制用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 27 | 多重检验控制

风险	治理办法	董事会语言
指标过多	预注册主指标与护栏	不准结果后挑指标
版本过多	控制错误发现率/分层淘汰	试得多不等于证据强
反复窥视	序贯方法或固定观察点	禁止看到赢就停
分群挖掘	独立验证集复现	细分胜利需再验证

使用多重检验控制时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

灰度梯度用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 28 | 灰度梯度

阶段	典型流量	回滚目标
内部	员工与合成任务	分钟级
影子	复制真实输入不执行	分钟级
金丝雀	1%—5%低风险流量	自动
扩围	10%—50%分群	小时级
全量	通过长期观察	可一键回退

使用灰度梯度时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例七 | 中国云平台灰度能力：产品功能不能替代治理结果

阿里云 PAI、腾讯云、华为云 ModelArts 等公开文档提供模型评测、版本部署、流量控制或监控能力。这些材料能证明中国企业可获得本地化基础设施。

具体动作通常包括登记模型版本、配置评测任务、部署服务、切分流量和查看监控。不同产品的能力边界与版本持续变化，本报告不把文档功能表视为客户经营成效。

作用机制是降低灰度和回滚的工程门槛，但平台不会自动替企业定义黄金集、增量指标、品牌红线或客户授权。若企业只购买工具而不建立责任制度，发布控制仍可能形同虚设。

管理启示是采购时做退出演练和企业任务验收。平台能力必须被写入本企业的运行清单、审批链和证据包。

图表规格 9 | 挑战者—冠军门

建议将“挑战者—冠军门”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 11

第十章 线上观测：将漂移、质量、成本、风险和业务接回谱系

竹势 AI 营销智库出品

本章判断

输入漂移、行为漂移、工具故障、成本异常和业务结果必须用同一运行标识连接。事故触发隔离、降级和证据保全。

董事会应把本章问题理解为经营控制，而非模型调参。输入漂移、行为漂移、工具故障、成本异常和业务结果必须用同一运行标识连接。事故触发隔离、降级和证据保全。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

线上观测六联表的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

NIST AI RMF 的 Measure 与 Manage 强调持续测量和风险处置不能分离；SRE 与机器学习监控实践同样要求把输入、行为、质量、资源和业务信号联合观察。对 AI 市场部而言，单看模型错误率无法解释经营异常，必须把漂移、质量、成本、风险和业务结果连接到同一 run ID 与谱系。

持续交付与金丝雀发布把变化拆成小批次，在真实环境中有限暴露，并以自动观察和快速回滚限制损失。它解决的是发布风险，不自动证明经营增量。AI 营销采用该思想时，应把灰度和回滚与因果评测、客户授权并列。

投放诊断智能体可能因广告平台字段变化而出现输入漂移，随后生成更保守的建议，调用次数上升、成本增加，最终延迟预算调整。若六类观测分散在不同看板，团队只能看到“转化下降”；将它们接回谱系后，才能确认问题来自数据接口而非模型推理。

监控阈值不能被当作精确真理。季节性活动会自然改变输入分布，新品发布也会使历史基线暂时失效；过于敏感的告警会造成频繁回滚，过于宽松则会放大事故。线上日志还可能包含个人信息，采集范围和保留期限必须符合授权与最小必要原则。

平台值班负责人应维护六联观测表和事故分级规则：输入或工具异常先隔离依赖，质量与风险越线立即降级，成本异常触发预算熔断，业务结果恶化进入联合诊断。每次事故都必须回写受影响版本、暴露范围、处置时间和复发防护。

线上观测六联表用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 29｜线上观测六联表

观测域	核心信号	关联键
输入漂移	主题、语言、客户结构	manifest_id
行为漂移	工具序列、重试、拒绝	trace_id
质量	事实、格式、人工评分	eval_id
成本	token、调用、人工时长	cost_id
风险	投诉、越权、敏感输出	incident_id
业务	线索、成交、毛利、复购	exposure_id

使用线上观测六联表时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

事故分级响应用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 30｜事故分级响应

级别	例子	响应
S0	安全、法律、公开严重错误	立即隔离并通知高管
S1	客户伤害或预算异常	停止流量、回滚、保全证据
S2	质量或成本退化	降级、限流、排查
S3	轻微偏差	进入修复队列

使用事故分级响应时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

漂移诊断用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 31 | 漂移诊断

漂移	可能原因	验证动作
输入漂移	新产品、新渠道、新人群	分群重放
概念漂移	客户偏好或市场变化	更新标签与对照
服务漂移	模型/工具版本变化	锁版本复现
政策漂移	法规和平台规则变化	法务重新批准

使用漂移诊断时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

案例八 | 受限案例：反馈污染制造无法复现的冠军

Anthropic在2026年讨论公开基准污染，指出代理在浏览任务中可能遇到已经泄露到网络的答案。该问题说明评测数据与真实环境相连后，系统可能通过捷径获得高分。

在营销中，相似污染会出现在把已发布成功文案放入黄金集、用同一模型生成答案又做评分、把销售人工挑选后的线索当作全体反馈，或让挑战者看到实验标签。结果是离线冠军上线后无法复现。

作用机制是候选优化了评测过程而非业务能力，典型表现为分数持续上升、真实结果停滞。若团队只公布胜者，幸存者偏差还会掩盖大量无效尝试。

管理启示是隔离训练集、验证集和最终保留集，记录数据谱系，定期加入新鲜盲测，并把复现失败视为降级条件。

图表规格 10 | 线上观测六联表

建议将“线上观测六联表”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 12

第十一章 组织治理：目标权、评测权、发布权、否决权和预算权

竹势 AI 营销智库出品

CMO	DATA	PLATFORM	RISK	FINANCE
目标与价值	实验与归因	谱系与发布	否决与边界	预算与净值

本章判断

营销定义目标，数据保障测量，平台控制发布，风险与法务行使否决，财务控制探索资本。权责分离降低自证成功。

董事会应把本章问题理解为经营控制，而非模型调参。营销定义目标，数据保障测量，平台控制发布，风险与法务行使否决，财务控制探索资本。权责分离降低自证成功。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

进化治理五权模型的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

NIST AI RMF 把 Govern 置于贯穿全流程的位置，说明风险治理不是评测后的附加审批。实验平台和持续交付体系也表明，目标定义、发布控制、预算配置与否决机制必须由不同角色承担。五权分离可避免同一团队既设计指标、又发布版本、再解释结果所形成的自证循环。

在自动触达场景中，营销团队可提出提升回复率的目标，数据团队设计实验，平台团队执行灰度，法务和品牌拥有内容与授权否决权，财务控制探索成本。若增长负责人同时掌握全部权限，短期业绩压力可能推动其忽略退订、投诉和边际成本。

五权分离不等于层层审批。低风险、可逆且已在授权范围内的变更可由预设规则自动放行；只有触及新数据用途、外部发布、预算上限或客户权益时才升级。责任边界必须与实际控制能力匹配，不能让法务对其无法观察的模型配置承担结果责任。

CEO 应批准进化治理章程，明确目标权、评测权、发布权、否决权和预算权的持有人及代理机制。季度审议检查是否存在权力冲突、长期未决候选和绕过审批的发布；重大事故后先修订权限与流程，再讨论模型性能。

进化治理五权模型用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 32 | 进化治理五权模型

权力	首要责任	制衡对象
目标权	CMO/业务负责人	防技术自定目标
评测权	数据/实验负责人	防业务挑结果
发布权	平台/运维负责人	防实验者自发全量
否决权	法务、安全、品牌	防收益抵消红线
预算权	财务与业务负责人	防无限探索

使用进化治理五权模型时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

RACI 核心活动用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 33 | RACI 核心活动

活动	A	R	C/I
适应度定义	CMO	实验负责人	财务/品牌/法务
候选构建	平台负责人	能力工程师	业务专家
在线实验	业务负责人	实验平台	法务/客服
晋级发布	发布委员会	平台运维	数据/风险
退役销毁	能力目录负责人	平台运维	审计

使用RACI 核心活动时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

季度董事会审议包用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 34 | 季度董事会审议包

内容	关键问题	输出
组合表现	探索投入换来多少可复用能力	继续/收缩
风险事件	是否暴露系统性缺陷	整改门
冠军老化	哪些版本环境失配	重赛/退役
供应商变化	替换和退出是否可行	迁移决定

使用季度董事会审议包时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

图表规格 11 | 进化治理五权模型

建议将“进化治理五权模型”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 13

第十二章 营销场景：不同反馈时滞与风险结构

竹势 AI 营销智库出品

内容与品牌	投放与线索	企微与舆情
高频、可回放、品牌门	可实验、需增量归因	客户可见、低容错

本章判断

内容 Brief、品牌审校、投放诊断、线索评分、企微跟进、活动复盘和舆情升级不能共用一套评测方法。

董事会应把本章问题理解为经营控制，而非模型调参。内容 Brief、品牌审校、投放诊断、线索评分、企微跟进、活动复盘和舆情升级不能共用一套评测方法。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

七场景差异矩阵的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

营销任务的反馈结构差异显著：内容 Brief 可在数小时内人工评审，品牌审校需要高精度风险判断，线索评分的真实结果可能延迟数月，舆情升级则具有极高损失不对称。因果推断和实验设计的基本要求是让方法匹配单位、时滞和干扰，因此七类场景不能共用一套评分器和升级节奏。

公众号选题可以用编辑盲评与后续阅读质量组合验证；投放诊断可用影子建议和账户结果比较；企微跟进要同时看回复、退订和销售后续；舆情升级必须由人工确认事实与响应等级。即便都叫“营销智能体”，其可变异面、样本量和回滚方式完全不同。

场景矩阵只能提供默认策略，不能替代具体业务判断。新品牌、重大危机和少量高价值客户都可能缺乏足够样本；平台互动数据还会受到推荐机制影响。对于不可逆外部承诺、敏感人群或高监管内容，人工裁决应成为常态而非例外。

CMO 应为七类场景分别批准反馈时钟、风险等级、最小证据和自动化上限。场景负责人每月报告候选数量、通过率、事故和长期结果；跨场景复用某个冠军前，必须重新验证数据、受众和政策条件。

七场景差异矩阵用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 35 | 七场景差异矩阵

场景	反馈时滞	主要风险	首选证据
内容 Brief	小时—天	方向偏差	盲评+采用率
品牌审校	即时	漏检/误杀	边界集+人工复核
投放诊断	天—周	错误归因/预算	实验+增量
线索评分	周—月	歧视/漏失	回测+成交成熟
企微跟进	即时—周	客户骚扰	授权+投诉护栏
活动复盘	周—月	幸存者偏差	全量台账+对照
舆情升级	分钟—天	漏报/过报	高召回边界集+人工裁决

使用七场景差异矩阵时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

变异权限矩阵用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 36 | 变异权限矩阵

场景风险	自动生成	自动执行	自动晋级
低风险内部	允许	限额允许	证据满足可允许
中风险经营	允许	灰度+人工门	委员会批准
高风险外部	草稿允许	原则上禁止	禁止自动晋级

使用变异权限矩阵时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

客户触达护栏用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 37 | 客户触达护栏

护栏	具体控制	失败动作
授权	确认渠道和用途同意	转人工
频次	客户级触达上限	抑制队列
内容	事实、承诺、敏感词	阻断发布
退出	退订和人工接管	立即生效

使用客户触达护栏时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

图表规格 12 | 七场景差异矩阵

建议将“七场景差异矩阵”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 14

第十三章 采购与平台：购买可退出的进化基础设施

竹势 AI 营销智库出品

PLATFORM EXIT CONTRACT / 采购退出契约

评测集可移植 · 日志与谱系可导出 · 流量可切分 · 模型可替换 · 失败可回滚 · 锁定可退出

本章判断

企业采购的不应只是模型调用，而是评测集、日志、谱系、分流、替换、导出和退出能力。供应商排行榜不能代替企业自己的任务证据。

董事会应把本章问题理解为经营控制，而非模型调参。企业采购的不应只是模型调用，而是评测集、日志、谱系、分流、替换、导出和退出能力。供应商排行榜不能代替企业自己的任务证据。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

平台采购十二问的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

企业采购评测平台时，应关注可验证能力而非功能清单。NIST AI RMF 要求组织能够记录、测量和管理风险，持续交付体系要求版本可导出、可回滚，实验平台则依赖稳定的随机化、日志和指标定义。供应商若不能提供这些底层证据，所谓“自学习”只能成为不可审计的黑箱。

采购团队测试一个营销智能体平台时，应实际导出评测集、候选谱系和运行日志，替换底层模型，执行一次 10% 灰度和完整回滚，并核对成本与权限记录。仅观看供应商演示的排行榜、成功案例或自动优化界面，无法证明企业在真实环境中拥有控制权。

可迁移性并不保证迁移成本为零。不同平台的工具协议、日志结构和知识索引可能无法一一对应；部分模型服务也不允许导出内部状态。企业应区分必须可带走的经营资产与可替换的供应商实现，并在合同中明确数据格式、保留期和终止支持。

CIO 与采购负责人应把十二问转成验收脚本，而非问卷：候选登记、评测复现、流量切分、模型替换、权限审计、预算熔断、日志导出和退出演练均需现场完成。未通过退出演练的平台不得承载关键营销流程，供应商披露的效果数据只作为线索，不作为企业 ROI 证据。

平台采购十二问用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 38｜平台采购十二问

维度	必须验证的问题	拒绝信号
评测集	能否导入企业黄金集	只给通用榜单
评分器	能否组合规则、人评和模型评	黑箱总分
谱系	能否追踪组合版本	只记模型名
日志	能否导出完整轨迹	不可导出
分流	是否支持稳定随机化	手工切流
回滚	能否一键恢复	重新部署
模型替换	能否多供应商路由	深度锁定
数据快照	能否复现实验数据	数据不断覆盖
权限	能否按动作控制	只有账号权限
成本	能否到候选/任务粒度	只有总账单
退出	能否演练迁移	无导出格式
审计	能否证明谁批准	无审批记录

使用平台采购十二问时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

供应商证据分级用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 39 | 供应商证据分级

披露类型	可证明	不可证明
官方文档	该平台提供某能力	能力在本企业有效
技术论文	某场景与条件下结果	跨行业普遍因果
客户案例	披露客户取得结果	排除选择和供应商偏差
排行榜	基准任务相对表现	真实业务冠军

使用供应商证据分级时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

退出演练清单用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 40 | 退出演练清单

对象	导出内容	验收
评测资产	数据、标尺、评分脚本	另一平台可运行
谱系资产	版本、父子关系、决定	可追溯
运行资产	工作流、配置、政策	可重建
证据资产	实验、日志、事故	可审计

使用退出演练清单时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

图表规格 13 | 平台采购十二问

建议将“平台采购十二问”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

BOARD BRIEF 15

第十四章 90 天路线：让进化机制进入经营节奏

竹势 AI 营销智库出品

0–30
冻结基线
建黄金集与权限矩阵

31–60
影子挑战
积累多目标证据

61–90
有限晋级
灰度、复核、目录化

本章判断

前 30 天冻结基线和黄金集，中间 30 天运行影子挑战，后 30 天进入有限流量和多目标评审。每阶段设停止条件与董事会门。

董事会应把本章问题理解为经营控制，而非模型调参。前 30 天冻结基线和黄金集，中间 30 天运行影子挑战，后 30 天进入有限流量和多目标评审。每阶段设停止条件与董事会门。 如果缺少明确主语和决定权，团队会在结果出现后补写解释，导致每次改动都能被包装成成功。可审计做法是先写任务契约、基线与风险容忍度，再允许候选进入比较。

90 天阶段门的机制由三个层次构成。第一层定义可被改变的对象与可观察结果；第二层规定证据如何升级；第三层把晋级、降级、隔离和退役写入发布制度。三层缺一，组织就会把局部优化误认作系统学习。

90 天路线的依据来自 PDSA、分阶段发布和风险治理的共同逻辑：先建立基线和可测对象，再小范围验证，最后在证据充分时扩大。Deming 强调改进必须在系统中循环验证，金丝雀发布要求每一步可观察和可回滚，NIST AI RMF 则要求治理、测量与处置持续联动。

首月可选择内容 Brief 与品牌审校建立黄金集和不可变清单；第二月让挑战者处理影子流量，验证质量、成本和风险；第三月仅对通过门槛的低风险场景开放有限真实流量，并召开多目标评审。线索评分和自动触达因反馈慢、风险高，可停留在建议模式而不追求同步上线。

90 天并不保证产生业务冠军。样本不足、数据质量差、权限未完成或风险事件都应触发停止；为了按期展示成果而放宽门槛，会把试点变成表演。路线也不适合一次覆盖所有营销场景，应优先选择可逆、可观测且与经营目标直接相关的闭环。

第 30、60、90 天分别设置管理门：第一门批准基线、黄金集和冻结项；第二门审查影子证据与探索预算；第三门决定目录晋级、继续观察或退役。每个阶段都要留下停止理由和未解决风险，董事会评价的是机制是否可信，而非上线数量。

90 天阶段门用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 41 | 90 天阶段门

阶段	核心任务	停止条件
0—14天	冻结基线、任务契约、风险边界	无法获得可用日志则停止
15—30天	黄金集、边界集、评分标尺	评分一致性不足则修订
31—45天	建立挑战者和影子运行	出现不可控副作用则停止
46—60天	有限流量、护栏和回滚	数据质量不可信则冻结
61—75天	多目标评审、复现与分群	收益不能复现则降级
76—90天	目录晋级、扩围或退役	责任与退出不完整则不全量

使用90 天阶段门时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

董事会阶段门用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 42 | 董事会阶段门

门	审议问题	批准结果
G1 基线门	现状是否可测、风险是否冻结	准许构建候选
G2 证据门	离线和影子是否可信	准许有限流量
G3 经营门	增量、利润与风险是否平衡	准许扩围
G4 资产门	能否复用、回滚和退出	进入能力目录

使用董事会阶段门时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

90 天交付物用于把本章抽象判断压缩成可执行结构。管理者阅读时应先确认每一行的责任和证据来源，而不是把表格当作静态模板。

表 43 | 90 天交付物

类别	交付物	验收标准
制度	变异权限、晋级、事故制度	权责签字
数据	黄金集、边界集、反馈字典	版本可追溯
平台	评测、分流、日志、回滚	演练通过
能力	至少两个可复用冠军	跨时段复现
经营	探索预算与季度报告	财务可核算

使用90 天交付物时，企业应按自身行业、客户风险和数据成熟度调整阈值，但不得删除谱系、风险门与回滚字段。调整应产生新版本并保留批准记录。

图表规格 14 | 90 天阶段门

建议将“90 天阶段门”制作成可转为 HTML/CSS/SVG 的解释型图。底层数据采用本章对应表格，不使用虚构经营数值。视觉结构为左侧输入与约束、中部候选和证据流、右侧晋级或处置；风险门以独立横向带表示，避免被误解为加权指标。交互版本可点击候选查看父版本、评测结果、审批人和回滚目标。

每个实验应在启动前写明预期信息价值：即使候选失败，能否排除一类假设、改善边界集或验证平台能力。没有信息价值、没有经营效应且重复历史失败的实验应被拒绝。季度审议时，董事会应同时查看胜出资产、失败知识、未成熟结果和重复试验率，以判断探索组合是否真正积累能力。

财务负责人可以把探索划分为保底、成长和前沿三类资金池。保底预算用于修复质量、可靠性和合规缺陷；成长预算用于已有业务闭环中的可逆优化；前沿预算用于新渠道、新智能体或高不确定方法。三类资金池采用不同证据门和最大损失上限，避免高风险探索挤占稳定经营。

探索预算不是研发团队的隐性试错额度，而是企业对不确定性付出的显性资本。它包含媒体流量、潜在毛利损失、客户体验暴露、人工复核、平台成本和机会成本。只统计模型调用费会严重低估真实投入，也会鼓励团队无限增加候选。

专题深化六 | 探索预算如何进入财务和经营审议

反馈污染则会进一步扭曲下一轮。销售只标注容易成交的线索，未跟进线索没有结果；内容团队只保存成功稿件，失败稿件被删除；平台推荐决定了哪些候选能获得点击。若系统把这些观察当成完整真相，下一轮会强化既有选择。企业必须保存未曝光、未跟进、失败和中止样本，并在反馈字典中标记缺失机制。

不可变运行清单要同时记录评测和服务环境，包括模型端点、工具版本、超时、重试、缓存、知识快照和资源预算。影子流量的核心价值，是在不产生外部副作用的情况下观察真实输入和真实基础设施。如果影子阶段只复制文本而不复制工具、权限和时延，它仍无法发现关键偏差。

许多离线冠军在上线后退化，并非模型能力突然消失，而是训练与服务条件不一致。离线数据可能包含人工清洗后的完整字段，线上输入却缺失、延迟或格式混乱；离线评测允许长时间推理，线上服务受限于时延和成本；训练时使用的知识快照与客户当前政策不一致。任何一种差异都足以使离线排名失效。

专题深化五｜训练—服务偏差与反馈污染的经营后果

冠军还需要有效期和重赛触发器。输入漂移、模型供应商更新、平台政策变化、投诉上升、成本异常或长期指标下降，都应触发重新挑战。若组织没有退役机制，旧版本会在多个工作流中长期残留，形成无法解释的混合运行。退役不仅是删除模型，还包括撤销权限、冻结证据、迁移依赖和通知使用者。

能力目录应把冠军写成“版本 + 场景 + 人群 + 数据窗口 + 资源预算 + 风险级别”的组合，而不是只保留一个名称。跨场景复用时，应先检查环境距离：输入分布、客户意图、渠道机制、销售流程和监管要求是否相似。距离较大时，历史冠军只能作为父版本，必须重新取得离线和影子证据。

候选选择总是在特定环境中发生。某个线索评分模型可能在一条渠道、一个季度和一组销售人员中表现突出，却在产品价格变化或渠道流量结构改变后失效。局部最优的问题不在于它曾经有效，而在于组织把环境条件从冠军定义中删除，随后把版本扩散到不相似场景。

专题深化四｜局部最优、环境变化与冠军老化

模型评分器可以承担初筛和一致性检查，但不能自证其输出。评分器版本、提示词、参考答案和偏差也必须进入谱系。企业应定期抽样比较模型评分与专家判断，监测系统性偏向。人工裁决成本较高，因此应集中在高风险、低样本、价值冲突和新环境，而不是替代所有自动评测。

高质量人工裁决并非随意投票。它需要盲评、结构化标尺、利益冲突披露、评分者校准和异议记录。品牌负责人判断语气与承诺，法务判断责任表达，业务负责人判断客户可执行性；三方意见可以不同，但最终决定必须记录理由与适用条件。对于评分分歧较大的样本，应进入边界集，成为下一轮评测资产。

在高风险营销中，多个目标经常无法通过单一数学函数排序。一个危机回应版本可能事实更完整，却语气更冷；另一个版本更具同理心，却增加法律承诺风险。把这些差异压缩为总分会隐藏价值冲突，并让权重设置者在无形中承担品牌和法律决定。人工裁决的作用，是把不可交换价值交回拥有授权的人。

专题深化三｜人工裁决不是评测失败，而是价值冲突的正式接口

对于自动触达，延迟负反馈尤其关键。客户可能不立即投诉，却在后续拒接、退订或降低信任。系统需要把沉默、人工接管、销售备注和退订纳入负向结果，而不能只统计显式回复。董事会审议冠军版本时，应查看成熟率：多少暴露已完成结果观察，多少仍处于未成熟状态。成熟率不足的候选只能被称为临时领先。

解决办法不是等待所有结果后再行动，而是建立延迟账本和分阶段结算。候选可以先凭离线质量与即时护栏获得有限流量，但晋级范围要与证据成熟度匹配。短期证据只支持短期、低风险使用；销售周期结束后再结算线索质量与成交；季度结束后补记毛利、退货、复购和品牌健康。任何后续结果都应回写原始 exposure_id，而不是另建一份无法连接的经营报表。

营销结果成熟速度差异很大。标题点击可以在数小时内观察，销售有效性可能需要数周，续费与品牌搜索可能需要季度尺度。若系统只在最快反馈上完成选择，短期候选会获得更多流量，更多流量又产生更多即时反馈，形成自我强化。慢反馈候选即便拥有更高毛利、更低投诉或更强复购，也可能在结果成熟前被淘汰。

专题深化二｜延迟反馈与“坏冠军”形成机制

更稳妥的制度是为每个候选预注册一个主效应、两到四个护栏和明确的最小有意义差异。主效应不足但护栏改善时，可以进入针对性复验，不能直接宣布综合胜利。对于品牌、客户权益和合规风险，采用硬门而非权重，因为权重允许收益抵消伤害。董事会应定期检查指标与真实目标之间是否仍保持关系，并要求出现结构性偏离时重新定义适应度函数。

管理者应把指标分成目标指标、诊断指标、护栏指标和审计指标。目标指标表达经营价值，诊断指标帮助解释路径，护栏指标定义不可接受退化，审计指标检查测量系统本身是否可信。四类指标不能在结果出来后随意互换。若团队看到点击不显著便转而强调停留时长，或看到转化不显著便选择某个细分人群，实验已经失去预先承诺的判断标准。

传统指标投机通常由人理解规则后调整行为，生成式系统则可以在极高频率下搜索能够提高分数的表达和路径。一个以点击为主目标的内容候选，可能迅速学会制造悬念、夸大稀缺、模糊适用条件；一个以销售接通率为目标的跟进代理，可能增加触达频次并把客户沉默误判为继续追逐的机会。局部数字改善并不需要系统“有意作弊”，只要评价函数遗漏了真实成本，搜索过程自然会偏向可被计分的代理目标。

专题深化一｜Goodhart 效应为何在生成式营销中更危险

专题深化七｜可解释冠军与董事会可读的决策记录

冠军版本的可解释性不要求管理层理解每个模型参数，而要求决策链能够用经营语言回答五个问题：候选解决了什么缺陷，在哪些场景和样本中胜出，付出了多少成本，触发过哪些风险，何时会被重新挑战。若团队只能展示一个综合分数或排行榜名次，董事会无法判断该版本是否值得扩大授权。

晋级决定应形成一页决策记录，明确基线、最小有意义差异、主指标、护栏、样本成熟度、分群结果、复现状态和未解决异议。对未披露或无法测量的结果，应写“未披露”或“当前不可观测”，不能用模型自评、供应商案例或行业平均值代替。记录还应说明没有选择其他候选的原因，防止组织只保留冠军叙事。

可解释决策记录也是跨部门协作接口。营销负责人看到业务效应，财务看到单位经济性和探索损失，法务看到客户权益与承诺边界，平台团队看到发布和回滚条件。相同记录进入能力目录后，后续团队才能判断能否复用，而不必重新猜测当初的选择逻辑。

最后，企业应区分“能力改进”和“权力扩大”。一个候选在既有授权内提高了准确率，并不自动获得更多数据、更高预算或直接触达客户的资格。权限扩张属于新的经营决定，需要重新评估受影响人群、最坏后果、人工接管和事故响应。将能力证据直接兑换成权限，是自进化体系最危险的捷径。成熟机制会把每次扩权视为新的候选：先限定环境和额度，再进行影子、灰度和复现；任何异常都回到更低权限状态。这样既保留改进速度，也避免成功版本在缺乏治理的情况下获得不可逆影响力。

BOARD BRIEF 16

结语：把优化权关进经营制度

竹势 AI 营销智库出品

FINAL EVOLUTION PRINCIPLE

用选择提高能力，用门禁约束风险，用谱系保留因果，用回滚守住责任。

会进化的 AI 市场部，核心资产不是更频繁地改模型或提示词，而是更可靠地把变化转化为证据，把证据转化为选择，把选择转化为可复用能力。企业真正需要的是一套受治理的经营选择机制：它允许低风险资产产生候选，也允许组织在证据不足时保持克制；它奖励可重复的改善，也能识别短期指标投机；它保留探索空间，同时限定客户、预算、品牌和法律责任的暴露。

“达尔文式”隐喻在此止步于候选竞争。企业不是自然环境，管理者也不能把伤害解释为进化代价。任何影响客户权益、公开承诺、预算和合规的动作，都必须有明确授权、人工门、审计和回滚。任何冠军都应带适用范围、证据等级和有效期。

董事会最终要追问的并非系统是否会“自我学习”，而是组织是否知道它学了什么、为什么学、谁批准、付出了什么、是否可复现、何时应停止。能回答这些问题，AI 市场部才真正拥有持续改进能力；回答不了，所谓自进化只是一种无法问责的漂移。

BOARD BRIEF 17

来源索引

竹势 AI 营销智库出品

以下索引仅列机构、题名、日期或版本、定位与采用口径，不含外部链接。平台与公司案例只按其公开披露使用，不推断未披露经营结果。

1. OpenAI：《Evals design guide / API documentation》，持续更新，访问 2026-07-23。定位/采用口径：评测任务、数据集、评分器与回归评测。

2. Anthropic：《Demystifying evals for AI agents》，2026-01-09。定位/采用口径：结果状态、轨迹、评测 harness 与评分器。

3. Anthropic：《Effective harnesses for long-running agents》，2025-11-26。定位/采用口径：长时任务分解、状态保存、恢复与验收。

4. Anthropic：《Quantifying infrastructure noise in agentic coding evals》，2026-02-05。定位/采用口径：基础设施噪声与资源预算可比性。

5. Anthropic：《How we contain Claude across products》，2026-05-25。定位/采用口径：爆炸半径、权限与隔离。

6. NIST：《Artificial Intelligence Risk Management Framework 1.0》，2023-01-26。定位/采用口径：Govern、Map、Measure、Manage。

7. NIST：《AI RMF Playbook》，持续更新，访问 2026-07-23。定位/采用口径：风险治理行动与文档化。

8. NIST：《Generative AI Profile, NIST AI 600-1》，2024-07。定位/采用口径：生成式 AI 风险画像。

9. Google DeepMind：《AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms》，2025-05-14。定位/采用口径：候选生成、自动验证器与选择。

10. Google DeepMind：《AlphaEvolve impact update》，2026-05-07。定位/采用口径：基础设施与算法优化应用边界。

11. Google Cloud：《AlphaEvolve is available for everyone》，2026-07-09。定位/采用口径：产品化验证与适用领域披露。

12. Google Ads：《About Conversion Lift》，访问 2026-07-23。定位/采用口径：增量测量定义。

13. Google Ads：《Conversion Lift based on geography》，访问 2026-07-23。定位/采用口径：地理实验与增量价值。

14. Google Ads：《Conversion Lift based on users》，访问 2026-07-23。定位/采用口径：用户级处理组与对照组。

15. Google Ads：《Measure and report performance with Smart Bidding Exploration》，访问 2026-07-23。定位/采用口径：探索型智能出价评估。

16. Microsoft Research：《The Anatomy of a Large-Scale Experimentation Platform》，2018-04-04。定位/采用口径：可信、可扩展实验平台。

17. Microsoft Research：《Online Experimentation at Microsoft》，2009-07-01。定位/采用口径：实验文化与决策。

18. Microsoft Research：《A Dirty Dozen: Twelve Common Metric Interpretation Pitfalls》，2017-08。定位/采用口径：指标解释陷阱。

19. Microsoft Research：《Three Key Checklists and Remedies for Trustworthy Analysis》，2019-05-29。定位/采用口径：实验可信检查。

20. Microsoft Research：《Trustworthy Experimentation Under Telemetry Loss》，2019。定位/采用口径：遥测丢失偏差。

21. Microsoft Research：《Deep Dive Into Variance Reduction》，2022-11-15。定位/采用口径：CUPED 方差缩减。

22. Booking.com / arXiv：《Democratizing online controlled experiments at Booking.com》，2017-10-23。定位/采用口径：实验民主化、中央知识库与护栏。

23. Netflix TechBlog：《Learning a Personalized Homepage》，2015-04-09。定位/采用口径：个性化排序与在线验证。

24. Netflix TechBlog：《Foundation Model for Personalized Recommendation》，2025-03-21。定位/采用口径：长期行为历史与服务约束。

25. ByteDance authors / arXiv：《Trustworthy Online Marketplace Experimentation with Budget-split Design》，2020-12-16。定位/采用口径：双边市场干扰与预算切分。

26. ByteDance authors / arXiv：《Ensuring Trustworthy Online A/B Testing: Five Questions on CUPED》，2026-06-17。定位/采用口径：多臂与两阶段抽样方差估计。

27. Google：《Rules of Machine Learning》，持续更新。定位/采用口径：训练—服务偏差与监控。

28. CNCF / Argo：《Argo Rollouts documentation》，访问 2026-07-23。定位/采用口径：金丝雀、蓝绿、分析与回滚。

29. Kubernetes：《Deployment documentation》，访问 2026-07-23。定位/采用口径：滚动发布与回退。

30. ISO/IEC：《ISO/IEC 42001:2023》，2023。定位/采用口径：AI 管理体系。

31. 国家互联网信息办公室等：《生成式人工智能服务管理暂行办法》，2023-07-13。定位/采用口径：责任、权益与内容治理。

32. 全国网络安全标准化技术委员会：《人工智能安全治理框架 2.0》，2025。定位/采用口径：全生命周期安全治理。

33. 阿里云：《PAI 模型评测与灰度部署文档》，访问 2026-07-23。定位/采用口径：模型评测、部署和监控披露。

34. 腾讯云：《大模型评测与服务治理文档》，访问 2026-07-23。定位/采用口径：评测维度与服务治理披露。

35. 华为云：《ModelArts 模型评估、A/B 与灰度发布文档》，访问 2026-07-23。定位/采用口径：模型版本与发布控制。

36. 火山引擎：《A/B 测试与实验平台公开文档》，访问 2026-07-23。定位/采用口径：实验分流与指标分析。

PUBLISHER

关于竹势 AI 营销智库

竹势 AI 营销智库出品

竹势 AI 营销智库聚焦 AI 市场部与 AI 营销，面向企业创始人、CEO、CMO及市场增长团队，构建“6+1”知识架构。六大核心模块组成“道、法、术、器、技、例”六脉体系：“道”负责认知刷新，研究 AI 时代的营销本质与战略判断；“法”关注体系建设，沉淀可持续、可复制的方法论；“术”提供落地路径，把策略转化为具体行动；“器”评测工具与平台，帮助企业选择合适的能力组合；“技”分享可以立即应用的实战技巧；“例”通过标杆案例验证方法与成效。“+1”专题围绕 AI 营销的重要趋势与关键经营问题，推出系统、深入、可下载的专题白皮书。

竹势智库通过专业研究、实践经验与管理框架连接认知、决策和执行，帮助企业看清方向、减少试错，逐步建设能够创造真实经营价值的 AI 市场部。