数据驱动精准营销：从流量到品牌的 AI 全链路

作者：**SVSAC 硅谷战略顾问委员会**

核心判断

精准营销应对经营结果负责，企业需要连接流量、客户价值与品牌积累，而不是被局部转化指标牵着走。对企业而言，值得优先推进的是：统一客户与订单口径，区分新增需求和渠道转移，分阶段验证投放、内容与复购运营的增量贡献。本篇结合公开资料与分析，讨论企业可以如何识别机会、组织能力并安排下一步投入。

关键词： 广告增量；随机实验；曝光选择；campaign 级预测；利润贡献；品牌测量

SVSAC - sv-sac.org

SVSAC | sv-sac.org

精准营销应对经营结果负责，企业需要连接流量、客户价值与品牌积累，而不是被局部转化指标牵着走。

我们的建议是：统一客户与订单口径，区分新增需求和渠道转移，分阶段验证投放、内容与复购运营的增量贡献。企业可以先据此梳理自己的产品、客户、能力与投入边界，把讨论推进到可以分工、可以验证的具体任务。

对管理层来说，这个议题值得放进经营议程。围绕“先做什么、凭什么做、做到哪一步再扩大投入”形成共识，往往比同时铺开多个方向更有推进价值。下文保留相关资料、比较和推演，供企业判断自身条件与行动顺序。

1. 从“更会找人”到“增加了什么”：核心问题与结果层级

1.1 精准不等于因果

数字平台可以估计某类用户点击、访问或购买的概率，也可以把内容送到预估响应较高的人群。这样的精准能力回答的是“在当前投放和选择规则下，谁更可能出现一个可观察事件”。营销决策更关心另一个问题：如果企业改变广告、预算或受众分配，与维持原有政策相比，客户行为、品牌和企业收益会发生什么变化？前一个问题是结果预测，后一个问题是政策反事实。两者相关，但并不等价。

若一位客户在未看广告时本来就会购买，模型把广告投给他可能提高被归因的转化，却没有增加购买。反之，广告也可能让客户记住品牌、稍后通过搜索或经销商下单，此时末次点击归因会漏掉影响。更高的点击率可能由创意吸引力提升带来，但若吸引的是低毛利客群，或者挤占了自然成交，企业净收益仍可能下降。故“精准营销”的核心问题不能以精准命中作为结论，而须追问：目标变量是什么、相对于哪个反事实、由谁在何时被处理、企业承担哪些成本。

本篇的主问题是：AI 辅助的广告与客户分析，在何种数据、投放和组织条件下，能识别真实增量并把预算配置到有贡献的触达；又在何种条件下，流量指标或模型归因会高估短期商业效果，或不能支持长期品牌结论？这里“AI 辅助”涵盖内容/受众优化和广告效果预测的决策流程，不预设任何平台当前具备某项实验能力，也不把同一工具包的全部结果都归给模型本身。

1.2 五类结果不能靠一条转化漏斗相乘

至少要分开五类结果。第一，投放交付，如可触达合格用户数、曝光频次、展示成本和有效触达。第二，行为反应，如点击、访问、加购、注册或购买。这些事件各有定义，可能由同一人重复发生。第三，广告增量，即相对于未投放或

SVSAC | sv-sac.org

替代投放策略新增的行为或购买。第四，经济贡献，包括新增订单毛利、折扣、退货、渠道费用、广告支出、履约与服务成本，以及被蚕食的其他产品销售。第五，品牌或长期客户结果，包括记忆、认知、信任、偏好、价格接受度、后续选择和复购。前一层可以是后一层的预测变量或中介，但不能仅凭相关性把漏斗各阶段的比率相乘并称为净 ROI。

每一项还需固定分析单位。曝光事件、用户、账户、campaign、市场簇和品牌系列可能是不同单位。如果一个用户收到 20 次展示，曝光次数不等于 20 名独立客户；如果一个 campaign 统计多个转化事件，事件数也不等于独立实验数。重复展示还会改变潜在剂量和平台竞价，因而“每次曝光效应”“每个合格人群政策的 ITT”“每美元增量贡献”是不同估计量。

1.3 商业问题应包含预算反事实

若企业仅在两个创意中选择，其目标是相对创意效果，不能据此知道广告比不投放产生了多少增量。若在固定总预算下把钱从渠道 A 转到渠道 B，估计对象是预算再分配政策；两渠道独立各自测出的效果不能直接相加，因为拍卖价格、受众重叠和频次可能变化。若想估计新增预算的边际收益，则需要观察额外预算对照或可信的剂量变化，不能把现有平均 ROI 外推到新增预算。

业务终点也要与决策时点一致。预算决策通常在 campaign 开始前或运行中作出；如果模型预测依赖 campaign 结束后才知道的点击、曝光率或实验组结果，它可能对事后活动效果预测有用，却不能被称为事前投放预测。若要用结果决定下期预算，要说明历史 campaign、目标 campaign 的市场和投放机制是否相似、未来信息何时可用，以及模型更新后的分配是否反过来改变标签分布。

2. 从目标选择到被看见：广告因果链中的选择机制

2.1 被投放者与未被投放者不是天然可比

广告并非随机送达。竞价和优化系统会依据用户历史、设备、上下文、预计响应和广告主规则选择机会；广告主也会定向高意向客群。若直接比较“看见广告的人”与“没看见的人”的购买率，前者可能原本就有较高的购买倾向。该差距同时包含用户选择和广告作用，不能单靠模型控制变量就保证拆开。

Lewis 和 Rao 对 25 项大型数字广告现场随机实验的分析揭示了测量难度和选择偏差风险。研究由美国主要零售商和金融服务企业的广告实验构成，总支出约 280 万美元；各实验通过随机保留用户不接收广告来形成对照。作者报告零售活动的个体销售标准差相对均值很高，常见变异系数约 10，ROI 置信区间中位宽度

SVSAC | sv-sac.org

超过 100 个百分点；其计算显示，要区分细小的经济差异，部分活动需远超百万量级的 person-week 观察。[1]这些数字描述的是 2007–2011 年前后、特定在线展示广告及合作企业数据中的统计经济性，不是所有今天广告渠道的固定样本量要求，也不是广告无效的证据。研究明确指出，对基线销量波动低、广告几乎带来全部客户或品牌认知低的企业，适用性可能不同；更高人均支出可能提高统计功效，但也可能因边际收益递减降低经济回报。[1]

2.2 分配、可投放资格、实际展示和点击是不同状态

一次线上广告实验可以从“是否被随机分到可投放组”开始，但被分到处理组不代表一定产生实际曝光。合格用户可能没有浏览广告位；拍卖可能输给其他广告主；平台建议的广告可能被发布商拒绝；用户可能使用屏蔽工具或快速离开。随机分派可识别的是按原始分配计算的政策效应；实际曝光者效应则需进一步识别哪些人若处于处理组会被曝光，并需说明不依从、外溢和对照反事实。

点击又在曝光后发生，而且通常由广告内容和用户特质共同决定。只在点击者中比较购买或只将点击加入因果模型，不会自动得到广告增量。若广告影响点击，点击处于处理之后；若点击也反映购买意愿，条件化点击还会改变样本构成。企业可把点击作为可观察结果或预算策略的预测特征，但需要讲清楚它用于预测、分层还是因果估计，不能不加说明地充当“广告被看见后才产生的购买”的证据。

2.3 自然成交、替代渠道和同一市场的拍卖竞争

广告效果的反事实不是抽象的“什么都没发生”。在用户未看到焦点广告时，平台仍可能展示其他广告；竞争者可能获得曝光；用户仍可能通过品牌搜索、门店、经销商、邮件或自然流量购买。测试策略如果抑制焦点广告，可能改变拍卖分配、竞争者占用或企业的其他触点。因此净增量取决于对照条件及市场环境。广告既可能“拉新”，也可能防止已考虑品牌的客户转向竞争者；两种机制对利润、市场份额和总品类需求的含义不同。

一个比较创意的 A/B 测试可以识别创意 A 相对创意 B 的分派差异，却不识别两者相对于不投放的增量；若对照组被竞品广告填充，处理效果也可能包含竞争替代关系。媒体覆盖重叠和多个 campaign 同时投放，会让一个 campaign 的控制人群接受另一项活动，破坏独立处理。故评估要保存账户/用户资格、所有并行广告计划、投放频次、其他媒体支出和时间窗，而不是只导出焦点 campaign 的归因面板。

SVSAC | sv-sac.org

2.4 观察窗改变结果含义

短观察窗容易把结果限定在即时点击或当周交易；长观察窗可捕捉延迟回访、重复购买和品牌变化，但也会累积价格、库存、季节、促销、竞品和其他渠道影响。时间窗不是技术设置，而是估计对象的一部分。将 7 日末次点击与 30 日实验结果对比，不仅是“窗口长短”不同，也改变了哪些用户事件会被归给广告。

品牌结果特别容易被短窗错配。广告可能先改变记忆或考虑集，再由稍后的搜索、门店体验和同伴推荐影响选择；一次态度问卷的差异也可能因熟悉度或回答迎合而出现，并不等于持续品牌资产。若要声称品牌提升，需要预先选择构念、测量方式、比较基线和观察时间；若要声称利润提升，还需从记忆或偏好连接到实际购买、价格、毛利和渠道成本。二者不能因“品牌漏斗”图上相邻就直接视作同一因果链。

3. 主要广告测量研究能说明什么

3.1 随机化可以消除一类选择，但仍可能没有足够信息

Lewis 与 Rao 的正式 QJE 研究将广告投放的随机保留设计与常见观察性“看过/没看过”比较放在同一测量问题中。25 项活动中大多数覆盖数十万乃至百万用户，但购买发生较少、每笔销售金额波动较大，使 ROI 效应相对于个体销售噪声很小。作者指出，投放设计本来就会选择人群，观察方法必须正确控制与广告投放和购买共同变化的因素；若遗漏变量或函数形式误设，估计偏差可能远大于期望效应。[1]

这并不意味着实验自动解决企业全部广告决策。作者模拟在活动之外增加独立 person-week 观察以评估不同 ROI 假设的检验力；“扩大多少倍”依赖该样本、成本、结果变异和目标差异。真实企业需要预先定义最低经济意义差异，按自身转化基率、毛利和集群/用户结构计算所需精度，并承认有些 campaign 可能经济上难以精确测量。若活动不能承受随机抑制曝光，可探索分地区或分时段的方案，但处理单元、外溢和时趋势又会成为识别条件；不能把困难的随机设计直接改称“模型归因”。

3.2 Ghost Ads 的关键是对照人群如何构造

Johnson、Lewis 和 Nubbemeyer 的 2016 年 Ghost Ads 会议稿仅作曝光反事实研究的待核线索。[2]未取得原段，不确认模拟拍卖、预测曝光标记的实现或局部

SVSAC | sv-sac.org

效果结果。本篇分析的问题是，广告资格、随机分派和实际展示是否对应同一人群；这些关系须由对称记录和明确假设判断，不能从一个方法名称推定已解决。

作为本篇的设计要求，分派日志、投放机会与实际展示应分开记录，并注明缺失和版本。若展示与预设资格不一致，必须判断差异是否受处理影响，不能只保留最终曝光者比较。这里不把发布商拒绝、技术兼容或并行活动的特定实现归属于未核文献，也不据此声称某平台方法已经验证。

本篇未核 Ghost Ads 会议稿或 JMR 正式版的具体方法原段，不用它证明当前平台可部署、生成式创意收益或跨渠道 ROI。可独立提出的问题是：若只有最终曝光日志，没有分派与机会记录，未曝光反事实如何识别？下文有关资格和局部效果的分析是作者条件推导，不是该文实现的复述。

3.3 PIE 把因果测量与效果预测分成两阶段

Gordon、Moakler 和 Zettelmeyer 的 PIE 方法进一步区分“随机实验识别活动增量”与“用实验标签预测未做实验的其他 campaign”。作者的 2026 年 10 月正式论文有期刊元数据；本次方法核读使用 2026 年 4 月 1 日 arXiv v2 作者稿，不能把它称为出版社排版全文。模型以 campaign 级 RCT 估计为标签，学习 campaign 特征与因果增量之间的关系，再对没有独立 RCT 的活动预测效果。作者特别说明，可将 campaign 结束后才形成的处理组均值、曝光率、点击或末次点击转化作为预测特征；这些变量若用于因果回归控制可能有问题，但用于预测一个由 RCT 识别的 campaign 级效果是不同的统计任务。[3]

“不同任务”不等于“新活动的效果已被实验识别”。该方法的标签由实验得到，未实验 campaign 的预测依赖训练样本覆盖与跨 campaign 映射稳定。其公开稿描述的 Meta 数据中，839 个原始多触点/多臂实验经拆分后形成 998 个处理–对照组合；这些组合测量 2,226 个转化事件，作者把“实验×事件”作为 RCT 观察。因此 2,226 不是 2,226 个独立广告主或 campaign。[3]研究按美国广告主 2019 年 11 月至 2020 年 3 月的大型 Meta 测试抽样，设定至少百万测试组用户和至少 5,000 次测试组转化；在该样本和评价权重下，PIE 的 out-of-sample R^2 为 0.88，原始 7 日末次点击单指标为 0.19。作者进一步做新广告主和新活动细分的留出验证；广告主外推与细分外推准确度会下降，细分下降幅度不同。[3]

这些数值的结果变量是“每美元增量转化”的预测，不是收入或利润预测，也不是品牌效果。文章解释广告平台需有代表性的 RCT 训练样本、定期更新以监测数据陈旧，预测模型不自动消除活动间差异。PIE 也允许后定 campaign 特征，故它适合在 campaign 结束后预测效果以辅助下轮预算，而不能因为其使用点击或曝

SVSAC | sv-sac.org

光数据便称为 campaign 启动前的因果预测。若用于新广告主、其他平台、不同国家或长期品牌目标，还须重新验证范围与标签定义。

3.4 有限证据与作者分析中的断点

可读广告实验与 PIE 资料分别支持其范围内的精度、选择问题及实验标签预测说明。本篇据此区分四个命题：处理后较高转化、实验局部增量、新活动预测准确，以及净利润增加。它们所需证据不同；最后一项还须核毛利、替代预算、服务成本和成熟品牌结果。未核的曝光标记文献不加入共同已证事实链。

3.5 广告指标的变化可能是好事，也可能只是换了分母

优化目标若从点击改为转化，或从总转化改为增量转化，模型学习的策略也会不同。点击率是点击/展示，访问率可能是访问用户/合格用户，转化率可能是购买用户/被追踪用户；当过滤、设备识别或归因窗口变化时，即使数值方向相同，样本母体也可能不同。前后指标若不在相同用户资格、成本口径、观察窗和平台版本下计算，就不能说“转化率提高了 X%”。

处理后指标如点击、归因购买会同时受到广告影响及用户倾向影响。它们可以帮助刻画广告活动结果，却不能单独识别增量。在 PIE 中，后定特征是以 RCT 活动为训练样本的 campaign 级预测器输入，预测目标是实验标签；这不同于在同一 campaign 中把点击者与未点击者直接比较。本篇后续会把“模型对效果标签的预测准确度”与“模型驱动预算决策的真实因果效果”分开检验。

4. 研究方法：建立从分派到经营结果的可检验链

4.1 证据边界和审查方式

本篇为定向范围梳理，不是所有广告技术或渠道的系统综述。Lewis 与 Rao 公开论文用于其已读实验、变异、ROI 精度及选择原段；PIE 方法按 arXiv v2 已读范围使用，2026 年 10 月期刊记录仅作发表身份核对。[1],[3]Ghost Ads 会议稿原文未取得，仅留待核线索，不声称选读或读取 JMR 终稿。各来源分别记录环境、随机单位、结果和外推限制。

本篇不会从摘要级或有限选段来源外推其未读方法细节；也不会把研究中的 Meta、Google 或单个零售商样本写成所有平台的普遍规律。被引用数据只在对应来源边界内说明。作者构造的账本、识别假设和企业协议另列为分析，不将其包装成实测或文献结果。

SVSAC | sv-sac.org

4.2 估计对象的层级

至少有四个估计对象需要分开。第一是政策分派效应：把合格单位随机分派到某广告策略或控制策略，按原始分派比较结果，反映现实政策包括未曝光和不依从。第二是实际曝光者的局部因果效应：要通过随机工具变量、对称预测曝光等设计识别，目标人群受曝光规则和顺从性限定。第三是 campaign 级效果预测：从一组 RCT 标签学习其他活动的效果映射，评价单位是 campaign 或实验事件，预测对目标活动的可迁移性需验证。第四是经营政策结果：固定预算和可执行范围下，采取一组投放/停投/换渠道规则后对增量贡献、品牌和客户长期价值的总影响。前三者可作为第四者的组成信息，却不是第四者的替代品。

每一估计量必须锁定目标总体、分派单位、处理对照、窗口和结果。政策分派 ITT 往往包括未被实际展示广告的人；这是该策略对所有合格用户的平均部署结果，不是实际曝光效应。本篇所讨论的局部效果取决于预先定义的相应人群；它不能自动外推到低倾向、不合格或被平台拒绝的人群。campaign 级预测的 out-of-sample 分割需按目标场景选择：同一广告主不同 campaign 的插值，不等于未见广告主的冷启动；同一行业中的随机留出，也不等于留出整个新行业。

4.3 用共同账本连接广告增量与净贡献

可以将一段预算政策的净贡献写成：两政策在同一完整订单范围和窗口内的净收入减成本之差，再减尚未计入订单账的增量广告、数据及实验成本。折扣、渠道、履约、退货与服务项目各计一次；若已包含在每单净毛利中，不得再次扣除。仅在单位贡献可比且成本分项不重叠时，才可简化成“增量订单数×扣除下列增量项之前的单位贡献基数−这些增量成本”。若广告替代了自然成交或其他产品销售，需将相应蚕食纳入上述总账；若订单增量已覆盖付费、自然及其他产品的净变动，不得再重复扣同一损失；若长期客户价值包含未来毛利，则折现率、复购周期、留存与观察成熟规则也须明确。曝光、点击和末次点击转化可帮助诊断机制，但不能直接替代订单增量或毛利。

这里的账本是作者框架而非某企业已实现的 ROI 估计。每个组件都应有来源或假设标记：随机实验中观察到的新增购买、企业真实毛利、财务确认广告支出、归因窗口内外结果、品牌测量值及未成熟订单分别记录，不在同一单元混合。若实际毛利、自然销量或退货记录缺失，净贡献就部分不可识别；应报告可观察的中间结果或范围，而不是由模型补成确定性利润。

SVSAC | sv-sac.org

4.4 品牌与流量结果的时间对齐

短期 campaign 测量可将指定期间内的购买作为主要终点，但长期品牌结果应另设计测量。记忆、品牌考虑和态度可能先于行动，且会随重复暴露、其他品牌活动、产品体验、口碑和市场价格改变。若同一调查同时测多个品牌维度，需记录量表、基线、抽样、地区和时间点；若计划比较多个市场，应验证构念和量尺的可比性。品牌状态的中间测量可以解释潜在机制，不自动证明销售或利润影响。

为了识别品牌活动的增量，可按市场簇、账户簇、用户资格或投放时段分派不同预算/创意政策，保留有意义的未处理或替代政策对照，并观察多个预定时间点。若处理人群会互相传播、竞价共用预算或广告挤占同一用户的其他触点，干扰需在分派单位和分析中处理。短窗态度结果和长窗购买结果需要分别陈述；若试验周期不足，应将品牌后续结果登记为未成熟，而非视作无效或成功。

4.5 研究限制

本篇不拥有企业级随机广告数据、跨渠道支出和真实成本记录，也未执行品牌研究或生成式 AI 投放实验。已读来源集中在数字广告实验、平台方法和 campaign 效果预测，不能代表电视、线下零售、B2B 采购或所有当代广告平台。Lewis–Rao 数据跨行业但仍是特定年份/合作企业；Ghost Ads 原文未核，不计为已读方法；PIE 验证使用大规模 Meta campaign 实验及特定标签与留出方法。因而我们提出的统一估计链、预算账本和后续协议是可供实证检验的框架，不是已有的企业效应结论。

5. 从随机分派到实际曝光：设计决定能回答什么

5.1 资格池与政策分派先于平台优化

评估投放政策时，应先定义决策时刻的合格总体，而非事后可见的曝光用户。资格可以包括市场、地区、产品可售状态、客户关系或同意状态，但必须在随机分派前锁定，并说明哪些条件会随模型分数或实时竞价变化。若营销团队先看响应排名、再只让高分账户进入实验，研究对象就是该高分资格池，不代表全部客户。若平台在分派后还会根据实时行为调整竞价，处理就是“获准进入这套优化流程”，不是固定展示一条广告。

基础设计是在合格总体内按用户、账户或市场随机分派到投放政策和对照政策，保留初始分派字段。主结果按原始分派计算，即处理组平均结果减控制组平

SVSAC | sv-sac.org

均结果，称为意向处理效应（ITT）。若处理政策同时规定渠道、预算、创意和定向规则，ITT 估计的是整个组合政策对该总体的平均影响。它会包含未赢得竞价、没有可用广告位、发布端拒绝、用户未实际曝光等实施结果。按实际曝光重新筛选处理组，再与全部控制组比较，会丢掉随机分派所建立的可比性。

例如，一家公司只在高购买倾向用户中投放。分数和合格名单应在分派前冻结；处理组获准投放，控制组抑制焦点广告。这样的试验回答在高倾向名单中启用该策略的平均政策效果。即使模型选人准确，结果也不能说“AI 使全体受众购买增加”，除非全体受众被纳入或外推经过独立验证。名单选择、名单内分派、实际曝光和最终交易是四个不同阶段。

5.2 构造算例：曝光者差异不等于曝光效果

以下是作者构造的期望概率，不是平台实测。假设合格池有 1,000 个账户，高倾向和低倾向各 500 个。不看广告时，两组转化概率分别为 20%和 2%；广告使高倾向组提升至 22%、低倾向组提升至 3%。试验将“只对高倾向组投放”这一政策随机分配给一半合格池；控制组不投焦点广告。假设处理分派不改变其他渠道、没有干扰且结果完整观察。在这个各臂高、低倾向账户各半的期望构造中，处理政策组内高倾向账户全部收到广告，低倾向账户不曝光，因此处理臂总曝光率为 50%；控制组不投焦点广告。计算的是总体期望值，不是整数交易记录，也没有构造置信区间。

控制总体平均转化率是(20%+2%)/2=11%；处理政策总体平均转化率是(22%+2%)/2=12%，因此政策 ITT 为 1 个百分点。若只把实际曝光者 22%与全部控制者 11%相减，差异是 11 个百分点，其中大部分来自高倾向用户与总体用户的基线差异，并非广告增量。

在额外假设下可算局部曝光效应：假定随机进入投放策略只通过焦点广告曝光影响结果，不改变竞品曝光、其他渠道、竞价或价格；分派独立于潜在结果；曝光测量准确；没有控制组交叉曝光；策略只增加而不抑制曝光。此时用 ITT 除以策略导致的曝光率差，可得到依从者的 Wald 比率。本例曝光率差为 50 个百分点，1/0.5=2 个百分点，代表高倾向组中可曝光者的局部效应。它不是所有合格客户都曝光的平均效应，也不是点击者效果。

如果构造表再直接给出低倾向组曝光效应为 1 个百分点，则全体曝光相对不曝光的平均效应是(2%+1%)/2=1.5 个百分点。但上述策略并未随机向低分客户投放，1.5 个百分点不能从该实验单独识别；它只是构造题面已经提供了两组潜在结果。企业若要决定是否扩展到低分人群，需在那里增加随机探索，或明确外推

SVSAC | sv-sac.org

假设并作敏感性分析。这个例子把总体政策 ITT、可曝光子群局部效果和全体曝光效应区分开来。任何一个转化差都不是利润，除非还计入毛利、广告支出、退货、服务成本及自然成交替代。

5.3 预测曝光标签与事后曝光的区别

对优化投放平台，我们建议区分原始资格、分派与实际展示。若不同创意触发不同受众选择，某种控制广告的接收人群未必代表焦点广告的潜在受众。这是需以平台日志检验的条件风险，不是 Ghost Ads 具体实现或已验证结果的转述；没有对称资格记录时，不应仅凭最终曝光者估计部署效果。

但这个标签不是自然真值。模拟拍卖、发布商接受规则、实时底价、库存和素材资格要与真实投放足够一致。若控制组模拟显示“会推荐”，但发布商本会拒绝，或处理组更新了竞价而控制组模拟不更新，对称性就会受损。算法版本和日志因此属于处理定义；不记录版本、资格变化和竞价过程，技术变更可能被误认为受众反应。

实际曝光者也不一定是分派前固定群体。曝光受实时竞价、用户访问和库存影响。若平台因用户点击而增加展示，最终频次既是剂量，也可能是处理后的反应；事后按频次筛选或控制可能改变估计对象。研究至少应保留资格、竞价、赢标、推荐、接受、展示、可视性和互动的时间戳，而不是只导出曝光总数。

5.4 估计对象应随问题命名

协议可先写清结果语句，例如：“在已合格且可购买的客户中，获准进入固定版本投放政策的 8 周 ITT，相对现行政策，是否提高 90 日确认的净订单贡献？”随后需指定个人、账户还是地区随机；预算是否相同；对照接受何种替代投放；重复订单怎样计数；退款、自然搜索和未成熟订单怎样处理。若目标是“预测可曝光者中的广告效果”，应另写局部效应及适用群体。

固定总支出下把预算从渠道 A 转到 B，比较的是组合政策，不是绝对广告回报；A 和 B 单独试验的效果不能简单相加，因为竞价价格、受众重叠和频次会变化。新增预算试验则需说明边际预算位置和边际收益假设。销售增加与广告同时发生，不足以识别因果，因为季节、上市、促销和销售跟进也可能变化。

SVSAC | sv-sac.org

6. 用 RCT 标签预测其他 campaign：可迁移性的条件

6.1 预测不是把随机标签扩展成普遍因果

PIE 先从随机试验取得 campaign 级增量标签，再学习 campaign 特征与标签的关系，预测没有独立实验的活动。[3]这解决的是如何将稀缺实验信息延伸到未测活动，不是每个未实验活动都拥有随机识别。实验标签在原试验范围内依赖分派和结果测量；对新活动的预测还依赖训练样本覆盖、标签噪声、特征可得性和关系稳定。

若训练实验集中在某一广告目标、市场、品类、促销周期或成熟客户群，模型在其他活动上的精确分数也可能只是支持域外外推。随机拆分 campaign 行，如果同一广告主、产品、市场、季节或实验批次同时进入训练和测试，测到的可能是已知环境内插值，不是冷启动。PIE 报告的验证结果随新广告主和新活动细分而异，作者也强调代表性 RCT 样本和关系更新的重要性。[3]

6.2 划分单位应匹配上线对象

如果模型每周为既有广告主的新活动排序，可按时间切分，只用早期活动训练并预测较晚活动；若会接触新广告主，应按广告主成组留出；若将跨国使用，则需将国家或市场作为独立外推维度。总体测试集上的 R^2 不能替代这些具体上线场景。

还需防止标签泄漏。同一 RCT 拆成多个处理–对照组合或多个转化事件时，如果一个组合进入测试、同一实验的另一组合留在训练，测试标签可能间接泄露。PIE 源稿区分 839 个原始实验、998 个处理–对照单元和 2,226 个“实验×转化事件”观察；不能把最后一项说成 2,226 个独立 campaign。[3]评估权重也要说明按观察行、试验、广告主还是拟投预算加权。

campaign 级 RCT 标签可能很噪。把点估计当作无误差真值，会令模型拟合抽样误差，尤其当样本量和精度与活动特征相关时。加权、收缩或多层方法可以改变不同精度标签的影响，却不能修复试验执行错误、未覆盖活动或缺失结果。报告应把实验估计不确定性与效果预测不确定性分开。

6.3 R^2 与“能否正确投预算”不是同一评价

R^2 描述给定样本和权重下预测值解释结果变异的程度；预算决策关心模型选活动后是否优于现行规则。如果仅少数 campaign 处在预算阈值附近，整体 R^2 很高仍可能在临界排序中错；若模型改善的都是极小预算活动，统计误差下降也可能没有经营意义。反过来，整体 R^2 一般，但高支出候选排序有改善，可能更有

SVSAC | sv-sac.org

用。应报告阈值附近误判、预算加权损失、排序稳定性和相对现行规则的策略增量。

PIE 报告的 R^2 目标是增量转化/美元，并与 7 日末次点击指标比较。[3]若企业目标为利润，还要用经验证的每次增量转化净价值、产品毛利、退货和服务成本，而不是将统一转化标签等同于利润最大化。模型在单个活动上预测得准，也不能直接回答多个活动同时竞争共享预算时的组合效果。

6.4 后定特征可预测，不等于投放前可用

PIE 允许使用活动结束后才形成的组均值、曝光率、点击率和末次点击转化作为预测特征，因为它预测的是由 RCT 估出的 campaign 效果。[3]一个变量不宜作为处理后的因果控制，不代表它不能用于另一项预测任务；但可用时间决定它支持何种决策。活动结束后的特征可用于下期预算或复盘，不能支持本次活动开始前的定向。

如果系统运行中滚动决策，每个快照只能用当时已形成的信息，不能回填未来点击、最终转换或最终支出。还要区分预测分数和实际策略执行：是否真正改变竞价，平台是否受库存限制，多个活动是否共享预算。独立活动标签训练出的响应预测，不必然等于同时优化整个组合后的结果。

6.5 决策前设置拒绝外推的规则

预测器应允许“不知道”。若产品、市场、预算、创意形式或受众策略超出训练数据范围，可拒绝输出精确增量，改为要求实验或报告宽区间。拒绝规则本身也要验证：若只检查被保留的高分活动，忽略被拒绝活动，评估集会产生选择性。高预算或高风险活动可保留随机试验，逐渐积累新标签后扩大预测范围。

一个可检验的上线方案是按广告主和时间锁定训练、验证、测试窗口；在验证集决定拒绝域与排序规则；未来活动仍保留随机审计；最后比较模型建议政策与原规则的预算加权 ITT，而不止比较预测分数。若模型胜出，还应核查收益是否来自少数大客户、某一季节、成本变化或标签泄漏。若效果只在同一广告主内部成立，结论应限于此范围。

7. 竞价与共用预算下的干扰：活动不是独立插槽

7.1 个体随机不一定意味着市场环境固定

在线拍卖将广告主和活动置于共同稀缺的注意力和广告位中。对一名用户隐藏焦点广告后，平台可能向其展示竞品；提高处理组出价可能改变其他活动价

SVSAC | sv-sac.org

格；同一家庭或账户成员也可能跨设备接触广告。因此，用户随机无误仍不保证结果只取决于本人是否处理。若个体潜在结果依赖他人处理状态，独立个体假设就不合适。

考虑一个条件场景：焦点广告退出后，其他广告仍可竞争该位置。此时“投放公益广告”与“焦点广告缺席而其他广告照常竞价”不是同一政策。是否发生竞争替代、方向及幅度须用数据判断。该场景是本篇的反事实分析，不作为未核 Ghost Ads 论文发现。

7.2 外溢路径与可观测诊断

外溢可以来自同一用户跨组接触、家庭或企业联系人共享优惠，也可来自地区预算占用、竞价价格、库存和频次变化。多个 campaign 共用受众或预算时，平台可能把资源从一个 campaign 转给另一个，单项试验改变的可能是整个活动组合供给。

可以检查重复标识、跨组广告曝光、展示价格和频次、预算与赢标率、并行 campaign、转化路径和设备匹配。诊断不能证明无干扰：平台没有竞价日志时，低频接触可能不可见；曝光率平衡也不能证明拍卖外部性为零。报告应列明可见范围，并估算未记录触点须多强才会改变结论。

7.3 选择分配单位与推广目标

如果主要风险是共享预算和拍卖价格，可考虑按地区、市场或时间窗随机，减少簇间溢出。代价是独立簇更少、精度降低。若风险仅在同账户联系人之间，以账户为分派单位更合理。随机单位应由实施和干扰机制决定，而非为了获得更多样本行。

地域簇试验估计的是给定处理饱和度和市场配置下的政策差异。若只让 20% 地区处理，不能直接推算全市场采用后的效果，因为价格、库存和对照可用性会变化。若条件允许，可在多个处理饱和度下分组随机，估计覆盖比例变化下的直接效应和外溢；做不到时应限制外推。商业问题多半是固定账户和预算约束下，模型政策相对原规则的总贡献，而非逐人效果的简单求和。

7.4 版本变化改变处理定义

文案、图片、受众、时段、竞价和频控同时变化，测得的是策略包，而非单一模型贡献。若要拆分，可用多臂或因子设计：固定平台、预算和受众，仅改变内容环节；或固定创意，比较定向模型；或在明确交互的设计中比较预算规则。

SVSAC | sv-sac.org

若模型改变竞价和曝光剂量，曝光就是机制之一，不宜在主 ITT 后按实际曝光条件化以“剥离模型”。

持续学习系统还会版本漂移。处理组每天更新而控制组固定时，版本变更本身就是处理差异。应记录模型、训练数据、特征、素材与竞价规则的时间戳。能锁定时可锁定一轮；必须在线更新时，应把更新策略作为处理，按阶段保存日志。某一静态版本的结论不自动适用于后续版本。

8. 从活动评估到预算政策：竞争解释与实施边界

8.1 策略包效应不自动拆成模型贡献

若团队同时启用 AI 创意生成、受众扩展、自动竞价和不同频控，随机实验识别的是新策略包相对旧策略包的总效应。不能据此分配“多少来自生成模型、多少来自排序或竞价”。要识别组件贡献，需要随机化单项或使用足够支持交互判断的因子设计；若模型必须与新运营流程配合，单独模型的边际效果可能并无可识别对象。

替代解释包括促销、库存、销售跟进、价格变化、平台资源倾斜、创意更新和受众组成变化。可操纵因素应尽量正交化或分层随机，辅助行动和版本需预先记录，并将结论限定在策略包层级。回归控制或事后归因分摊不能替代识别，尤其当辅助行动由处理状态触发时。

8.2 固定预算和推广后的价格反应

固定预算下把资金移至高预测收益活动，可能提升该活动边际竞价价格、降低新增受众质量或造成频次饱和。用旧活动的平均效果外推加预算后的增量，隐含边际响应不变。可预设数个预算梯度或随机预算分配，测量有限范围的剂量响应，并记录花费、独立触达、频次、竞价成本和边际增量。机会成本来自被挤出的替代活动，而非泛用外部报价。

比较两个渠道时，应固定支出或明确政策金额。若处理组花费更多，结果可能是预算增量；等预算比较则回答同样的钱投哪里更好。这两类问题不可混成一个 ROI。每美元比率还可能因小分母极端波动，应同时报告支出、增量、总贡献及不确定性。

8.3 决策规则还受不确定性与执行能力约束

把预测增量转成行动时，系统需要处理区间、最低预算、广告库存、频次上限、品牌安全和地区排除。预估略为正但区间很宽的活动，可分配探索预算而非

SVSAC | sv-sac.org

立即全量；负向预测也需排查追踪缺失、时间错位或自然销售缺失。模型输出应呈现支持样本、范围和拒绝原因，而非把分数显示成确定事实。

执行容量也是策略一部分。品牌团队审核素材，法务核查声明，渠道团队核预算和落地页，数据团队验事件，销售和客服承接结果。若这些环节延迟，活动窗口可能错过季节或缺货期。应记录提交、批准、上线、暂停时间和原因，把无法执行作为政策结果，不能删掉未上线活动后只报告成功案例。

8.4 分阶段验证及失败条件

可按四个问题推进。第一，在按时间和广告主留出的验证中，模型是否比现行规则更准，能否对超出支持域的活动拒绝预测。第二，在新 campaign 中，原始分派比较模型建议与现行规则是否提高预算加权增量贡献。第三，在固定预算下，差异计入素材、实验、服务、退货和自然成交替代后是否仍成立。第四，若声称长期品牌价值，是否预先定义构念、市场可比性、重复测量和延迟窗口，并将态度与购买、毛利分开报告。

每步都可能失败。更准的预测未必改善预算，因为活动无法执行或转化价值不同；点击变多可能是低意向流量；销售增加可能由促销、季节或供应变化造成；态度提升也可能不迁移到价格或购买。将预测验证、政策试验、经营核算和长期品牌测量分开，才能定位断点，而不把误差压成一个“AI 提升比例”。

9. 从增量转化到经营贡献：分子、分母和预算替代

9.1 先确认“每美元”里的美元是谁的

广告指标常以每次点击成本、每千次展示成本或每美元转化表示，但单位成本不等于增量回报。每美元末次点击转化把被归因的转化除以广告费用；每美元增量转化则把实验估计的净增转化除以可比支出。前者只需归因窗口和支出，后者还依赖对照反事实、随机分派、成本范围与活动窗口。两者数值可能接近，也可能方向相反。报表应保留支出数、增量转化估计和对应的不确定性，而不只显示效率比率。

若焦点策略改变了费用结构，还需确定比较的是固定总预算、固定人均/账户预算，还是预算可变的完整政策。固定预算政策比较回答同等资源在候选活动之间如何分配；总预算增加比较回答增加投入的边际产出；以每美元为目标排序则隐含可分割资源、边际回报稳定且活动间无相互影响。共享竞价和频次饱和会使这一假设失效。不同设计的政策结论不能在一个“ROI”字段里混用。

SVSAC | sv-sac.org

9.2 构造账本：增量交易不自动等于净收益

一个透明的作者构造账本可设对照组 1,000 个合格账户中出现 100 笔已确认订单，处理政策组同样 1,000 个账户中出现 112 笔；毛增量是 12 笔。若每笔订单扣除下列额外折扣、退货与服务项目之前的贡献基数相同、均为 40 元，增量毛利为 480 元。假定两组政策预算差为 300 元，处理组额外折扣、退货净损失和服务支出合计各项后净增成本为 140 元，则该例净贡献为 480−300−140=40 元。数字仅用于展示会计结构，不代表显著性或某企业 ROI。若实验效应区间是−4 至 28 笔，毛利贡献范围为−160 至 1,120 元；扣除 440 元成本后净贡献区间为−600 至 680 元，点估计为正并不足以证明政策稳赚。

账本中一笔订单只能进一个明确状态。退款前的下单数、已付款数、完成履约数和退货后的净订单是不同终点。如果处理政策引来更多低价促销订单，订单数量可能上升而贡献毛利下降；如果退款或取消晚于观察窗，结果只是未成熟。成本也要与窗口一致：广告消耗可能即时确认，素材审核和实验工程是固定或一次性成本，客服或履约负担可能随着客户数变化。将固定成本在不同 campaign 间摊销会改变单位回报，但不改变总贡献，因此应同时保留总额与摊销口径。

9.3 自然购买与替代成交如何进入反事实

被广告触达者之后通过自然搜索、门店、经销商或销售代表下单，不代表广告没有作用，也不代表该单都由广告造成。对照组在同样时间内的相关购买构成基线；主效果比较应按随机分派保留所有合格账户并按统一窗口计数。多渠道路径可用于描述活动如何参与，而不能把平台末次点击独占为因果功劳。若业务目标是某一渠道销售额，需说明其他渠道被忽略还是作为竞争结果纳入。

所谓蚕食也需定义对象。广告可能将本来就会买的客户从自然渠道移到付费渠道，增加广告费而未增加总订单；也可能将客户从高毛利产品带到折扣产品；亦可能阻止客户选择竞品。用焦点 campaign 归因订单而不观察品牌总交易、替代产品和竞品缺失，会无法区分增量与重分配。一个可行分析是以账户为单位记录所有可追踪成交类别，并按预先定义的主终点估计总贡献；次级渠道归因用于解释路径，不能反向替换主终点。

9.4 比率指标的不稳定性

每美元增量转化是一个比率估计，分母花费若小或受到随机波动，比例可能极不稳定。假设两组都只有 1 笔增量，花费分别 100 元和 10 元，比率会显示 0.01 和 0.1 次/元，但小样本不支持宣称后一政策效率高十倍。若花费本身由处理后的投放策略决定，简单把增量估计除以处理组实际花费也不一定对应固定预算政策

SVSAC | sv-sac.org

效果。可报告按预先指定预算的净贡献和支出差，并用联合不确定性或随机化推断分析，而不是只对两个点估计做比值。

零或负的花费差、不可比较的曝光强度、极端小样本和成本记账缺失会令比率失去直观解释。对每一决策，应定义是否比较等预算方案；若不是，应展示投入变化与响应曲线。Lewis 和 Rao 关于测量精度的研究特别提示，效应相对销售噪声极低时点估计 ROI 可能仍有极宽区间。[1]模型排序若按预测均值而不计预测不确定性，也可能把高方差活动推到预算前列。

10. 归因窗口、数据完整性和客户路径

10.1 归因规则是描述口径，不是反事实

末次点击可用于运营对账：在某个回看窗口内，找到转化前最后一次点击广告并登记为归因来源。它不回答如果用户没有看到该广告是否仍会购买。窗口从 7 天改为 30 天，或从点击归因改为曝光归因，会改变被记入活动的交易集合；数字升高可能只是规则覆盖更多路径。归因逻辑需要随数据版本记录，前后报表应在一致的用户、终点和窗口下重算。

PIE 所比较的 7 日末次点击是 campaign 级预测基准，与 RCT 标签的 incremental conversions per dollar 不同。[3]其 R^2 差异应在作者数据和目标定义下解释，不意味着每家企业的末次点击必然高估或低估同一幅度。一个渠道的点击常与购买意图相关，也可能是广告引起的中间行为；仅凭 R^2 无法判断业务实际增量、毛利或品牌影响。

10.2 用户识别与跨设备拼接改变观测总体

路径分析依赖是否能把广告曝光、点击、访问和交易连到同一人或账户。匿名设备、登录状态、cookie 限制、共享终端和线下成交会产生漏接、重复和错误拼接。若处理政策本身提高登录或打开应用概率，能够被追踪到的用户也会由处理改变；只分析可匹配路径可能产生处理后选择。若对照更难识别自然交易，表面增量可能是数据完整性差异。

因此应将“事件发生”“可观测事件”和“成功归因事件”分别记账。平台聚合报告可用于投放交付，但不能与企业 CRM 订单按身份连接时，应标注数据缺口和不可识别部分。企业可对随机抽取且具有明确支付记录的子样本进行独立结果核验，但核验样本如何选取、未匹配账户如何纳入、误匹配率怎样估计都要写明。无法得到全路径数据时，可把主要结论限制为可观测订单 ITT，避免推断全市场路径归属。

SVSAC | sv-sac.org

10.3 处理后的活动指标与选择性留存

广告效果模型可能使用展示、点击、加购、频次和渠道回访等活动后特征。若它预测的目标是实验中已估的 campaign 效果，这些变量可以作为预测信息；若用它们去“调整”同一实验的处理效应，则可能因处理后变量或共同原因引入偏差。数据仓库中还会出现只保留活动完成者、预算未耗尽活动或至少有一次转化的 campaign 等选择。若成功活动较易形成完整记录，训练集可能系统性偏向可测且表现较好的活动。

一个可靠的版本表应保留资格规则、随机分派、预算和创意版本、实际曝光日志、点击事件、转化回传、身份匹配状态、撤销与迟到更正、观察窗和数据抽取时间。这样才能识别结果变化是活动本身、埋点覆盖、追踪技术或数据修订造成。PIE 作者将活动结束后的聚合信息用作跨 campaign 预测器输入，但这并不说明这些指标没有选择或版本问题；它说明的是在给定实验标签和预测任务下可学习关联。[3]

10.4 多触点分摊不等于增量分摊

多触点归因常将同一订单价值拆分到若干接触点。分摊规则可以是线性、首末触点或模型权重，适合内部描述渠道参与，但分摊总和等于订单价值只是账面约束，不代表每个渠道都产生该比例的因果增量。如果删去某个触点后用户通过其他渠道仍购买，归因权重不能直接等于该触点的可避免损失。增量归因需要相对于替代政策的对照，而非仅凭路径顺序。

渠道互补还会使单渠道实验和全组合政策不同。品牌搜索可能承接社媒产生的记忆，零售商促销可能改变付费广告转化；一个渠道单独被抑制时，另一个渠道或自然路径可能补位。可通过正交因子或市场簇设计估计互补/替代交互，但高阶组合很快增加实验臂和样本需求。资源有限时，应根据预算决策优先选少量关键交互，其他路径作为假设而非在归因图中伪装成已估贡献。

11. 品牌结果的延迟与测量边界

11.1 品牌不是短期点击的别名

品牌结果可能包括识别、回忆、关联、信任、考虑、偏好、支付意愿和实际选择。它们彼此相关却非同一构念。曝光后的问卷记忆提高，可能来自近期熟悉而非稳定偏好；搜索量增加可能代表兴趣，也可能是疑问或投诉；交易短期上升

SVSAC | sv-sac.org

可能来自折扣，不一定形成品牌资产。研究需明确想测哪个构念、哪个人群、相对于何种政策和基线。

若要声称品牌长期价值，需安排超出即时点击窗口的观测，考虑重复曝光、产品体验、客服、口碑、竞品活动和价格环境。同期品牌活动会干扰个体随机试验；若市场级传播强，按用户随机可能低估或污染组间差异。以地区或市场簇分派可减少部分传播，却需要足够独立市场并校验市场间趋势，且估计对象变为特定饱和度下的部署策略。

11.2 重复测量与问卷缺失

同一客户在多个时间点参与问卷时，记录是重复测量而非独立样本。处理组和控制组的回访率不同也可能使观测到的态度变化偏离目标总体。若广告内容吸引更愿意回答的人，或者处理组出现退订/屏蔽，完整案例分析会只留下特定响应者。应预先指定样本框、邀请策略、主量表、失访处理和分析对象，并同时报告邀请率、应答率和组间差异。

问卷量表还可能有响应迎合、理解差异及跨地区语义差别。若多市场比较品牌态度，应验证题项在市场间的含义、量尺和基线是否可比；否则所谓“品牌提升两分”可能是作答尺度或翻译差异。辅助行为指标可验证问卷方向，但不能因为问卷与点击相关就把构念等同。对长期品牌结论，应同时报告测量边界和市场覆盖。

11.3 态度、购买、毛利与客户留存各自归因

假设随机广告提高了品牌回忆，但未观察到购买差异。可能原因是观察期短、产品库存不足、价格不合适、样本不足，或记忆变化不能迁移到购买；“不显著”无法区分这些机制。若观察到购买增加却毛利不增，折扣、品类替代和退货可能抵销；若复购提升，需等足够客户成熟并控制同期促销。品牌态度可作为中间结果，不是未来毛利的直接观测值。

对于客户留存，需定义首次接触、活跃、流失或复购事件及风险时间。即使广告先于后续复购，也不能把留存差全归因于广告，因为高倾向客户会更可能受到广告并更会留下。随机政策 ITT 可以回答在指定期间对合格总体的后续结果差；实际曝光者或购买者限定分析需要额外识别条件。品牌或留存模型可帮助预测客户价值，却不自动证实广告创造了该价值。

SVSAC | sv-sac.org

12. 跨渠道政策与归因的统一验证方案

12.1 先画因果图，再决定是否能连成一条链

可以把流程表述为：广告政策分派影响竞价与曝光，曝光影响注意和互动，互动可能影响访问与购买，产品体验与服务影响重复购买/态度，最后才形成收入、毛利和潜在品牌变化。自然需求、价格、库存、销售跟进、竞争者投放和用户意图可能同时影响多个节点。观测路径记录提供顺序信息，但并不自动识别每个箭头。特别是只在点击、访问或实际购买的人群中估计下游结果，往往改变了受众构成。

若目标是整个广告政策的贡献，主要结果应按最初分派总体估计，不要把处理后中间节点作为筛选条件。机制分析可以报告广告对点击、访问或记忆的影响，另用明确的中介假设解释路径；不应把这些中间效应连乘成利润。若确需估计中介效应，需要说明时变混杂、处理后混杂和额外识别假设，并将其与随机化的总效应分开。

12.2 推荐的最小可检验方案

一项可复核的企业试验至少要保存以下信息：预先定义的目标总体和排除条件；随机单位与分层变量；策略版本、预算和对照状态；从获准投放到实际呈现的阶段事件；重复曝光频控；跨 campaign 和跨渠道干扰；主结果及独立订单/财务来源；退款、取消、未匹配和延迟交易；结果观察窗和数据冻结时点。分析先报告原分派 ITT 及区间，再分别描述曝光、点击和渠道路径；若估计局部曝光效应，则陈列工具变量条件和目标人群；若训练 campaign 效果预测器，则另做与部署相匹配的时间/广告主留出。

最小方案还需要一套可决策规则：根据每次增量订单价值、投放与执行成本、预算替代和不确定性，确定是否扩大、继续试验或停止。决定规则应在看结果前写下，避免事后挑选窗口、受众、转化事件和利润分母。长期品牌结果可作为分期随访，而非强行塞进短期主终点；如果随访未成熟，就如实报告未决。

12.3 决策建议必须允许被推翻

该框架有明确可反驳预测。如果在固定随机分派、对照条件和预算下，点击变化在预注册的跨广告主/市场留出中稳定预测由试验独立估计的订单、贡献和成熟品牌效果，那么点击可以在该覆盖范围内作为经验证的预测代理；这仍不反驳“单凭点击相关性不能识别因果效果”，也不保证换政策后代理继续有效。反之，如果模型只改善归因分数或短期预测，而策略随机试验的贡献没有增加，机

SVSAC | sv-sac.org

制是更好的预测或记账，不是创造了利润。若效应主要来自小部分高倾向客户，全面扩展后的竞价价格和频次变化可能削弱平均收益；这要求在更高覆盖饱和度下再检验。

研究还应将模型增量与流程改进分开。有可能只是团队开始统一目标、清理订单事件、固定预算和加强复盘，就改善了决策；这种改善可以有商业价值，却不证明某个模型算法单独有效。通过对照流程、版本日志、相同预算和充分执行的策略实验，才能判断是数据质量、运营纪律、模型排序还是内容/竞价处理带来差异。所有尚未执行的设计都只是待检验方案，本篇没有真实企业数据或投放结果。

13. 把内容、受众与预算拆成可识别的处理

13.1 内容实验回答的是内容差异，不是模型存在与否

实践中常把“AI 创意”作为一个处理标签，实际上其中可能包括生成文字、生成图片、素材筛选、地区改写、审核方式、迭代速度、竞价规则和目标人群等多项变化。若新素材同时由模型生成、团队改过、增加了预算并换了定向，实验只识别这些变化合成的策略包。要归因于某个内容组件，至少要让其他条件尽量相同，并预先规定实验比较对象。

一个可操作的两臂设计可以先固定人群资格、渠道、预算、频控和落地页，只随机呈现人工版本或 AI 协助版本。两组都应经同一品牌和事实审核，以免把审核质量差异误当生成模型效果。若生成版和人工版同时在语气、信息量、优惠强度、图像品质上不同，估计仍是成套素材策略效应。若业务关心的不是创作来源，而是“更快交付、同预算下能测试更多版本”的流程收益，就应把产能/迭代机制定义为处理，不宜用单个素材的点击率替代。

多臂或因子设计可分别随机内容来源、内容个性化、受众策略和预算，但臂数增加会稀释样本。实验单位也应与内容传播单位一致：用户可以评价个体广告，市场或账户可能更适合投放政策，素材审核可在素材簇层级。相同素材被展示给数万用户并不意味着创意生成有数万个独立样本；真正独立的创意簇可能很少，组间推断应考虑素材聚类和重复调用。

13.2 品牌主张、匹配内容与令人不适的精准感

更个性化不必然更有说服力。内容和用户兴趣越匹配，广告可能更容易被注意；但若文案暴露了过多隐私推断、对用户处境作出未经证实的判断，信任或品牌好感可能下降。相关效果可能因数据来源的可接受程度而异：用户明确提供的

SVSAC | sv-sac.org

信息与系统推断的敏感特征，对同一商品的呈现方式并不等价。团队不能仅从转化模型最优分数推出“用户愿意接受这种个性化”。

测试可以区分内容个性化的相关性与其披露方式。比如固定事实、优惠和素材质量，随机比较一般品牌信息、依据站内明确行为生成的内容、依据跨站推断生成的内容；测量即时注意、对品牌信任、投诉/屏蔽和实际购买。若涉及受限或敏感信息，则先由产品和法务按适用规则界定资格，不把不适宜处理的内容放进实验臂。结果可能是点击上升而信任下降，或态度更好但购买不变；这些不是相互矛盾，而是不同终点。

对生成内容的“准确性”也要分成可验证事实、品牌立场、产品适配和创意风格。一个模型能模仿历史语气，不代表新的主张符合当前产品条件；同一优惠在不同地区的限制、库存、价格和交付条件可能不同。若同一文案被复制到多市场，处理差异可能不仅是语言而是承诺本身。实验或评估需将主张版本与目标市场绑定，避免事后以“AI 整体表现”抹平事实错误和市场适配。

13.3 内容适配与受众定向的交互

若模型对高分用户生成个性化素材、对低分用户给普通素材，内容和受众两个处理被缠在一起。高转化结果可能来自选对人，也可能来自选对内容，或者两者互补。可在支持范围内做 2×2 设计：通用素材/个性化素材×旧受众/模型受众，保持每臂预算、渠道、频控和投放时间可比。主效应与交互需分别解释；若样本不足以估计交互，结论只能为策略组合，不能归因某一模块。 这里应在共同合格总体内随机分派“受众选择政策”与素材政策，而非把天然不同的旧受众和模型受众直接当成随机组；按所有原始分派单位统计结果，不能只比较最终触达者。

交互不一定为正。个性化素材可能只在用户已表达兴趣时有效，陌生受众反而觉得被监视；更广泛的受众可能需要清晰品牌介绍，而非细分商品细节。平均上两个模块单独都有效，也不能假设叠加一定更有效；若同时切换后观察到提升，仍要排除预算、素材审查和上线速度的同步变化。多臂试验若只有少数创意模板，要以模板簇作为推断层级或增加模板数量，不可把每名受众当作独立生成版本。

13.4 生成速度带来的候选数量与赢家偏差

生成工具使候选广告数量增加，意味着团队可以更广泛探索，也意味着从大量候选中挑出最高点击素材时会产生赢家选择偏差。假设 100 个真实效果完全相同的候选各自有随机测量误差，最高观测值仍倾向高于真实值。若将胜出素材当

SVSAC | sv-sac.org

成真实高效应再与历史平均比较，部分“提升”只是筛选极值。素材数量越多、试投样本越小、选择步骤越多，这种偏差越值得关注。

较稳妥的流程是先在探索集筛选创意，再用独立确认样本比较最终候选与既有基准；确认集不能再用于反复改文案和挑赢家。另一选择是把多轮筛选机制整体随机化，比较“模型生成+筛选政策”和“现行创作+筛选政策”，此时估计的是整个开发流程的价值。若上线后继续迭代，则每轮新选择需要保留新确认集或随机探索流量，不能一直用同一人群作筛选和证实。

生成候选较多的价值也不只体现在胜出素材的平均点击。团队可能通过覆盖更多需求、减少低质量创意的制作时间或更快发现品牌禁区而受益；反过来，大量候选增加审核排队、法务检查和渠道配置成本。完整经济比较应计入可用候选数、被淘汰比例、审核返工、最终上线比例、创意寿命和增量结果。候选生产量增加而可上线质量不变，可能只是把瓶颈从创作移到审查。

14. 个体异质性：谁被预测会买，与谁因广告多买

14.1 响应概率与处理效应的差异

一个模型若准确预测“谁会购买”，仍可能把广告投给本来就会购买的人。假设顾客甲无广告购买概率 50%，广告后 55%，顾客乙无广告概率 5%，广告后 20%。甲的响应概率较高，但其广告增量为 5 个百分点；乙的基础响应低，增量却是 15 个百分点。若目标是新增交易，按无条件购买概率排序会选甲；若目标是增量转化且两者成本相同，理想排序会选乙。若乙购买毛利更低或触达成本更高，利润排序又可能改变。

因此需要区分基础风险、处理效应、经济价值。响应模型预测 Y(0)或当前政策下结果概率；增量模型试图预测 Y(1)−Y(0)；净价值还需将购买差乘以边际毛利并扣除触达成本和替代影响。对用户的个人因果效应无法在同一人身上同时观察处理和对照结果。即便模型给每人一个精细 uplift 分数，那也是在可识别假设与训练资料上估出的条件平均量，并不是该个体潜在结果的真实标签。

14.2 分层实验比个体分数更容易审计

可以按分派前的倾向分层或业务价值层构建随机实验，估计各层 ITT，并检验模型排序是否把预算放到增量更大的层。分层必须在分派前定义，层内有处理和控制支持。若某细分几乎全被投放、没有未投放对照，不能从该细分的现有日志单独得出其增量。模型在目标域的支持概率接近零时，重加权会极端不稳定；这一点应当以重叠图、样本量和权重分布呈现，不能只报一个总体平均效果。

SVSAC | sv-sac.org

简单切出一组“高 uplift 人群”并报告处理组比控制组多买，仍可能是筛选噪声或多重尝试结果。更好的做法是把学习样本与检验样本按客户簇或时间拆开：第一部分开发分层规则，第二部分锁定规则并随机比较策略，第三部分保留未来时间验证迁移。确认分组和阈值需在看到确认结果前冻结；否则同一组结果既定义“高增量人群”又证明其高增量，论证循环。

14.3 细分稳定性与预算权重

平均增量可能掩盖异质性，但碎片化细分也会造成噪声。按语言、设备、地域、历史购买或品类拆分时，每个层级都有自己的样本量与结果方差；同时做数十种切分并只报告最好的几组，会夸大效果。分析应预先挑选有业务理由的异质性变量，报告全部预定层、交互估计及区间，并在独立数据上确认重要差异。

统计平均层效果也不一定等于决策价值。若每个账户贡献毛利不同，按账户平均的增量购买可能低估高价值产品影响；若某层体量很小、成本很高，其单位增量虽大，总政策贡献可能有限。应明确目标权重：当前合格客户人口权重、计划触达预算权重，还是未来增长目标权重。若未来政策会改变各层预算，则使用固定的历史权重可能不能代表新政策总体。

14.4 异质性模型的拒绝与探索

当历史数据在某些人群或市场缺少随机支持时，模型的细分分数可能靠代理变量外推。适宜的动作不是把不确定分数变成精准定向，而是将其标为未知，安排有界探索，或维持原策略直到有足够证据。探索流量会带来短期机会成本，但也产生后续训练标签；在评价政策时，需把探索成本与信息价值分开报告，且不要把探索期间受众组成和成熟窗口当作普通投放结果。

一个可检验的比较是锁定原策略和模型策略，在同一目标总体与同一预算下按账户簇随机；主要结果为固定窗口确认的净贡献 ITT。次要分析评估预测分层中的效果差是否在独立时间段重现。若总体策略无改善或关键分层排序反转，应检验是模型校准、成本价值、随机波动、竞价变化还是目标权重错配；不能只重新训练直到历史数据出现期望排序。

15. 从分数到资金：固定预算下的排序与边际收益

15.1 排序目标需要统一单位

预算系统常把模型输出的预期转化、预测 ROI、用户价值或排序分数转成竞价。若不同活动的标签单位不同，例如转化、销售额、毛利或留存价值，分数不

SVSAC | sv-sac.org

能直接比较。即便全部转换成预测利润，也要检验利润定义、模型校准、成本覆盖和不确定度是否一致。广告平台可能以点击、应用安装或销售事件优化；广告主自己的目标则可能是贡献利润、客户留存或品牌搜索。平台优化目标与业务目标相同，需要实际核实，而非根据界面名称推断。

固定预算下的理想决策不是将每个 campaign 按平均 ROI 机械排序，而是选择一组可执行行动，在约束下最大化总预期贡献。行动可能不可分：某些活动需要最小投放、特定市场要同步开播、素材制作有固定成本。一个活动的下一元预算和当前平均回报不同。若忽略这些约束，连续分配模型的结果可能无法执行；应把预算可行集合、最低/最高消耗和渠道承接写入方案。

15.2 边际回报与频次饱和

受众中的最初触达可能带来新注意，后续重复展示收益递减或导致厌烦。竞价价格也可能随预算增加而上升。历史 campaign 平均效果并不等于新增预算的边际效果。一个简单构造：某渠道前 1000 元增加 20 笔增量订单，下 1000 元增加 8 笔；每笔净毛利 30 元时，前段毛利 600 元、后段 240 元，尚未扣广告费前者比后者更有效，但两段分别净亏 400 元和 760 元：前段相对更有效也不代表相对不投广告值得支出，后段更不能据此扩量。平均 ROI 或模型按初始段排序不能直接授权扩大支出。

预算剂量试验可以随机给同一策略分配低、中、高预算，尽量固定受众资格、时间和创意，并记录频次、独立覆盖、边际拍卖价格和结果。若高预算组实际上只把更多钱花在同一批已经曝光用户上，估计的是频次政策而非拓宽触达的效果。若不同臂触达不同市场，需处理市场趋势和跨臂拍卖干扰。试验必须在业务可承受区间内，并设定预算上限与停止条件。

15.3 阈值决策与不确定性

只按预测期望值排序，在估计区间宽时可能将资金集中于噪声较大的活动。设活动 A 预期增量利润为 100，标准差为 120；活动 B 预期为 80，标准差为 20。若按均值，A 排第一；但若企业不愿承担较大下行风险，B 可能更合适。风险厌恶不是模型客观性问题，而是决策者对短期现金流、品牌风险或履约容量的取舍，应预先说明。此为构造说明，不代表通用风险分值。

可以采用下置信界、概率达标或探索/利用规则，但这些规则各有代价。下界排序可能排除真正高潜力但样本不足的新活动；探索规则可能短期牺牲回报；单纯按概率“正效应”也忽视贡献幅度。对于目标毛利大但不确定度高的市场，可以

SVSAC | sv-sac.org

设置有限实验份额并更新证据；对重大预算行动，先用边界范围计算最坏可接受损失，而不把一个模型输出当成保证。

15.4 从投放排序变为组合政策试验

要验证排序策略本身，可以在若干可比账户或市场簇中随机使用“模型排序”或“现行排序”，固定总预算、可用行动、时间和审批规则。主要结果是每个原始分派簇的总净贡献，采用分派前计划的客户或预算权重。需要保留被模型拒绝的对象和被替代的原行动，否则评估只看模型选中的赢者。若实验组总预算、候选池或执行支持高于对照，比较的就不只是模型排序。

单一活动分组实验无法告诉企业排序政策面对竞争性预算时会怎样重配全部支出。政策试验应记录预算转移矩阵、活动被挤出的支出、各项实际花费和可执行失败。若预算不足以覆盖所有活动，未被选择的对象并非“零成本未投放”：其中一些可能会被原规则选中，差异也是策略效果的一部分。更可靠的识别对象是整套资源分配规则，而不是把事后表现最好活动单独拿出来。

16. 策略更新、反馈和跨期学习

16.1 更新后的数据不再来自旧策略分布

投放算法一旦上线，分数会影响谁被触达、谁留下点击和购买标签。历史训练集若主要来自旧策略，更新后的模型看到的是被新策略筛选过的结果。高分用户可能因为被多次触达而增加疲劳，低分用户因少被投而缺少标签；新的点击、归因和转化率因此是政策产生的数据，而非中性抽样。模型若将“被选中且有标签”当作真实偏好，可能强化最初的选择而忽略未触达的潜在客户。

若想继续学习增量效果，应保留随机 holdout 或一定探索概率，并记录每个单位被各政策选中的概率、触达资格与未触达原因。若某一类永远不被选择，其政策反事实可能没有正概率支持，单靠加权不能恢复。此时应指出缺乏 positivity，重新设计受控探索，或将结论限于当前已覆盖的人群。

16.2 预测准确改善与政策因果效果改善要分别监测

模型上线后，预测误差可能下降，因为新标签更贴合当前投放；与此同时，策略净贡献未必上升，因为内容疲劳、拍卖价格、库存、竞品和团队执行发生变化。每次更新应分别看校准/预测指标、被推荐活动覆盖范围、预算决策变化、分派试验结果及利润账本。只看点击模型预测准确，不足以说预算政策变好；只看一季实验正向，也不能证明下次更新稳定有效。

SVSAC | sv-sac.org

若频繁查看同一在线实验结果并据此停止、改组或增加预算，名义置信区间和预先设定的比较不再简单成立。更新规则可以采用明确的阶段门和随时监测方案；每次改动都记录时间、版本、所见数据和触发理由。上线前后时序比较可用于发现异常，但若季节和投放策略同时改变，不能把变化全归因于模型。

16.3 可证伪的反馈设计

可在合格总体中保留小比例随机实验：模型策略、现行政策和有限探索臂并行，按账户簇或其他适当单位分派；模型更新在预先指定周期执行，实验组在该周期内维持版本稳定。记录随机概率、实际曝光、成本、订单、渠道替代和拒绝/投诉等保护结果。政策 ITT 按原始分派报告，并监测模型策略的覆盖变化和不同批次校准。如果模型策略只是把预算推向已有高响应客户，新增增量应在控制基线倾向后仍表现为政策总贡献提升；若只改善付费渠道归因或表面 ROAS 而总贡献不变，说明可能是渠道迁移而非净创造。

这些条件仍不自动解决所有跨期问题。品牌学习和口碑可能跨越实验窗，市场竞争者可能跟随复制，长期频次或隐私反馈可能改变可持续性；短期保留组只识别对应政策期的估计对象。若市场扩展引起竞价均衡变化，需逐步提高处理饱和度或按市场簇测试。若无法继续随机化，则应提出清楚的准实验假设和敏感性，而不能把在线模型更新日志称为自然实验。

17. 搜索广告的增量、自然位移与交易总量不是同一个问题

17.1 付费触达可能重分配注意力

品牌方看到一条付费搜索广告带来的点击，不能据此判断平台因此多创造了一笔交易。搜索结果页的付费位和自然位往往争夺同一屏幕上的注意力；消费者可能只是从自然链接转向付费链接。此时广告主的付费归因增加、自然流量减少，但消费者购买总量未必变化。反过来，若广告促使用户发现原本不会找到的产品，总交易可能增加；若付费位挤压了竞争品牌的自然曝光，平台交易也可能在品牌间重新分配。估计对象须先说明是广告主的品牌交易、平台总交易、全品类购买，还是消费者搜索成本。

自然位移还可能发生在广告主自己的品牌词上。用户已主动搜索品牌名称时，广告可能只是截取原本会通过自然结果到达的点击。若把被广告触达者与未触达者按最后点击归因比较，投放选择和搜索意图会使付费路径看起来格外有效。可比反事实需要在合格的搜索/账户总体中对是否展示品牌词广告作分派，按

SVSAC | sv-sac.org

原始分派统计付费与自然来源合计的订单、贡献和新客，并保留跨设备和延迟回访的观察规则。广告位的“点击份额”不是新增需求。

17.2 平台层反例及其边界

一项电商平台的现场研究摘要报告，研究者对 3%的访客阻断全部赞助搜索广告；摘要据此报告的是赞助广告的存在减少平台经手的总交易，而不是阻断广告减少交易；摘要同时称，赞助广告从自然搜索列表夺取了一部分业务，但广告收入足以抵消佣金损失，使该平台在财务上受益。[5] 这个结果揭示了几个估计对象可能朝不同方向变化：自然成交下降、平台总交易下降、平台收入上升。它不表示广告主获利，也不能推出消费者福利改善或品牌资产增加。

该研究只能按已实际读取的官方摘要范围引用。摘要提到针对 3%访客阻断广告，但没有公开足够信息让本篇复核分派方法、具体窗口或效果量；因此此处不推断随机分派细节，也不将 3%写成可复现的处理率。平台名称未在摘要中披露，不猜测其身份。更不能把单个平台的搜索广告案例外推成所有电商、所有品牌词或 AI 排序的规律。其价值在于说明平台的佣金和广告收入目标可能冲突，而不是提供广告主 ROI 系数。

若决策者只看平台广告后台的归因交易，可能把平台收入、商家新增毛利和自然成交位移混在一起。对广告主而言，需要另外核对买家是否本来会购买、是否从自营/经销渠道迁移、促销是否同时变化、毛利是否为增量以及退款是否成熟。平台层实验可以提示机制问题，却不能替代广告主的固定总体政策试验。

17.3 搜索意图的分层与品牌词试验

品牌词、品类词和问题解决型查询的先验意图不同。将它们合并后，平均点击率或转化率可能主要反映流量构成，而非广告增量。品牌词用户可能已经接近交易；品类词用户仍在比较；问题词用户可能尚未确定需求。统一的“转化提升”会掩盖这些阶段差异。即使结果按查询类型分层，也需注意查询词本身可能由广告之前的搜索经历或其他营销影响，不能在处理后挑选有利词类。

更可审计的设计是先冻结可纳入的查询/市场清单，按预处理搜索意图、品牌关系与用户阶段分层，再在层内随机广告政策。主要结果应包括付费加自然的总订单，以及品牌方贡献；付费渠道点击、自然点击和页面访问作为机制结果，不替代总量端点。还要记录竞争竞价、实际展示、广告花费、排名与拒绝/未展示原因。若没有足够的品牌词对照，结论应限定为有支持的查询层，而不是对未投放品牌词作外推。

SVSAC | sv-sac.org

18. 从展示到购买：漏斗与延迟效应的证据

18.1 分母会在漏斗每一步变化

展示、可见、访问、独立访客、转化用户、交易笔数和销售额是不同的计量对象。广告可以提高访问量，却吸引购买意图较低的增量用户；这样总购买可能仍增加，但每次访问的转化率下降。若团队只报转化率，分母扩大可使比例变低；若只报总订单，则没有显示每笔订单的成本。报告应同时给出分子和分母，明确重复访问、重复订单、取消与退款规则。

一项覆盖 432 个 Google 展示广告现场实验的作者研究稿报告了显著异质的漏斗表现：其摘要给出的访问提升中位数为 17%，转化提升中位数为 8%，但跨实验分布宽。表 2 中，184 项有转化结果的实验之相对转化提升估计，第 10 至第 90 百分位范围为−8.9%至 83.4%（中央 80%）；这是跨实验点估计的分布范围，不是单项或汇总处理效应的置信区间。[4] 该材料是作者公开稿，不是本篇逐页核读的期刊最终版；研究对象是特定历史时期平台上自选择参加的广告主实验。它支持“访问和转化效应不等、实验异质性大”的限定性经验观察，不能作为今天所有平台或生成式 AI 创意的预期提升率。

这一漏斗差异至少有几种竞争解释：增量访问者的购买意图较低；广告带来的是原有访问提前发生；测量遗漏或 cookie 流失改变了后续访问记录；展示广告增加了无法归因的品牌搜索；或者转化目标与真正商业价值不一致。单一漏斗比例不能判别这些机制。应保留原随机分派单位，并分别追踪事件与时间，避免只在实际访问者或完成归因者中条件化而破坏分派可比性。

18.2 Carryover 不是固定的“品牌记忆系数”

同一作者稿还报告活动结束后的后续访问/访客结果，并讨论跨期延续；摘要所述平均后续访客和访问提升与中位数、实验间分散程度并不相同，且 cookie 流失会使追踪衰减。[4] 这说明短期广告的影响可能在活动结束后仍被观察到，却不足以给所有行业套用一个统一的延续系数。后续曝光、再营销、季节变化和购买周期也可能与广告效应共同变化。

把结束后的结果称为品牌资产需要额外测量。活动后访问可能来自记忆、促销提醒、购买研究、广告造成的即时好奇、竞争对手反应，或用户在多个设备间转移。品牌认知、偏好、支付意愿、实际复购和长期贡献毛利是不同结果。一个短期访问效应既不是长期品牌存量，也不保证后续购买有利润。

可将观察窗口预先分成即时、活动后短期和更长期阶段，保持同一随机分派总体，分别报告各端点及成熟比例。若在后续窗口中发生新活动，需要记录并区

SVSAC | sv-sac.org

分：估计“初始活动加上正常后续营销”的政策效果，还是“阻止其他营销后的纯延续效应”。后一问题可能需要控制后续处理，但不能通过删掉后来被触达的人来实现，因为后续曝光本身可能受初始处理影响。若身份链接随时间失效，则报告失访比例和边界；不能把未观测当作无访问或无购买。

18.3 促销效果、品牌效果和记忆效果的时间结构

促销广告可能将订单从未来拉到当前，而非创造长期新增需求。只观察活动期会把提前购买当成净增；只观察活动后又可能漏掉即时收益。若消费者在通常购买周期内补货，短期增量可能随时间收敛。分析时应把“购买发生时间”与“购买者是否新增”区分，并结合购买周期、价格折扣和退货成熟期设定终点。

品牌广告的假设机制则可能是认知、联想或偏好变化，随后才影响搜索和交易。检验这一机制需要在随机分派后固定问卷构念、样本抽取和测量时间，并考虑重复问卷造成的提示效应。问卷态度改善不必然转化为选择；实际搜索或购买变化也未必是品牌认知机制，可能来自价格或便利性。若品牌量表存在跨市场测量差异，跨语言均值的差异不能直接当作品牌效应差异。

一个可证伪的结果模式是：若认知或态度提升，但在购买窗口成熟后，品牌搜索、复购或贡献利润没有相应变化，则“态度改变已转化为商业价值”的链条未获支持；如果交易增加却态度不变，则需检查价格、库存和渠道便利等替代机制。负结果不表示广告无价值，而是说明指定机制和终点之间的推断不能越级。

19. 从单次活动到多触点：归因、迁移与预算组合

19.1 末次触点把路径当成了增量

客户旅程常经过自然搜索、社媒内容、邮件、经销商页面和线下销售。末次触点会把订单记给最后一次可追踪的点击，但“最后发生”不等于“没有它就不会买”。首触点也可能只是很早的接触记录，未必是决定性原因。多触点归因模型分配的是既定路径上的信用，不自动识别各触点的因果贡献。

若每个渠道都各自运行实验，且客户跨渠道受到处理，简单相加各渠道增量可能重复计算；若一个渠道让用户迁移至另一个渠道，某渠道结果变化也可能是转移。企业可在可控制的客户簇或市场上比较完整渠道政策包，报告总交易和渠道构成变化；只有在各臂保持其他渠道规则稳定、或明确采用因子设计时，才可把总效果拆分为渠道组件。拆分估计需要足够的独立簇与无严重干扰条件。

SVSAC | sv-sac.org

19.2 展示广告现场研究对漏斗的反证作用

上述 432 项现场研究稿还提醒，广告归因端点和业务端点之间存在阶段性损失：增量访问者的后续转化不一定与原访客相同，跨活动的转化效应分布可以跨过零。[4] 因此，AI 创意使互动增加的主张需要经过独立访问、成熟交易、单位经济与品牌结果的层层检验。即使某模型改善创意点击预测，若实际投放只将既有需求分配到收费入口，广告主净贡献仍可能没有提高。

外推到新广告主和新平台还会遇到样本选择。参与平台实验的广告主可能具备不同预算、数据质量、品牌成熟度和投放能力；无法把实验数直接视作代表全市场的广告主样本。若只保留完成报告或提供购买标签的活动，缺失可能依赖结果。可报告实验来源、参与门槛、活动数与广告主数的不同层级，并用目标广告主分布重新加权，但加权不能补出没有支持的行业/规模区域。

19.3 平台收益不等于广告主收益

平台可能同时从广告展示收费和交易抽佣获益。即使广告挤占自然结果、使总交易下降，只要新增广告收入超过损失的佣金，平台总收入或利润仍可能上升；Moshary 的摘要级案例描述了这种方向组合。[5] 这与广告主是否获得新增利润是两件事：平台广告收入是广告主支出；商家要比较新增毛利、渠道迁移、折扣、退货和广告成本。平台的最优广告库存与广告主的最优获客政策不必一致。

同一广告行动可能让多方结果不同。广告主可能多卖而降利；平台可能提高广告收入但减少平台总交易；消费者可能更快找到产品，也可能被更相关的付费位引导到较贵选择。现有摘要证据只涉及匿名电商平台的交易和平台收入方向，不能判断消费者福利或广告主利润。跨主体评价应明确主体和转移支付：把广告费视作广告主成本、平台收入，不能在合并社会总量里重复当作新价值。

19.4 预算政策的多方账本

一个可用于试验方案的账本至少把四层分开：平台收取的广告费、广告主渠道总订单、广告主增量贡献利润、消费者搜索/购买结果。前两者能在某个平台增长，不意味着后两者改善。若进行市场级或平台级试验，还应追踪被挤出的自然位、其他渠道、竞品曝光及经销商订单，说明哪些是净增、哪些可能只是内部转移。

账本中的对照必须保持总量口径一致。处理组广告费增加而自然订单减少时，应计算处理组所有渠道的总贡献差；不可将处理组付费归因销售减去对照组总销售，也不能把不同归因窗口下的订单直接对比。退款未成熟、跨期复购缺失和经销商库存变化应显示为未决组件或区间，而不是以模型分数填充成确定收

SVSAC | sv-sac.org

益。此建议是作者提出的验证结构，不是 Moshary 摘要已经执行的广告主利润分析。

20. 将增长信号转化为可检验的商业主张

20.1 一组分层终点而非单一总分

可把决策终点组织为顺序而非加权混合分：首先看合格总体内的分派与触达是否按方案实施；其次看广告成本、访问与交易的增量；然后看扣除折扣、退货和渠道迁移后的贡献；最后看预先指定的品牌记忆、复购或长期客户价值。每层都回答不同问题。若把点击、订单、品牌态度和毛利压成一个综合分，权重变化会掩盖具体失败位置，且不同业务的权重没有天然客观值。

层级也不是必然的因果漏斗。比如品牌搜索可能是中介也可能受竞品活动影响；未点击用户也可能受广告影响。把“点击者中的转化”当作下一层因果效果，会对处理后的中介条件化。应以全体原始分派单位估计主要 ITT，再把点击/访问等作为机制描述或在明确因果中介设计下分析。对于多阶段决策，可随机化策略而非只比较实际点击者。

20.2 一个假设账本：相同归因订单，不同净价值

以下数值仅用于复算口径。假设两个投放政策各面对 1,000 个合格账户，90 日内均记录 100 笔平台归因订单。政策 A 广告费为 12,000，折扣与可变履约成本共 8,000，退款成熟后留下的净订单为 80；政策 B 广告费为 8,000，相关成本为 9,000，留下净订单 90。假设每笔净订单扣除下列折扣和可变履约支出之前的贡献基数相同、均为 250，且退款损失未再次计入这些支出，则 A 的已观测账面贡献为 80×250−12,000−8,000=0；B 为 90×250−8,000−9,000=5,500。虽然归因订单相同，净订单与费用结构不同，策略价值也不同。

但这仍不是增量利润：账本没有给出无广告对照。如果 A 的 80 笔订单中有 70 笔在不投放时也会发生，B 的 90 笔中只有 65 笔原本会发生，则增量订单假设分别为 10 和 25；但这两个反事实数量不可从上述归因数直接观察。需在固定合格账户总体内随机分配政策，成熟后估计每个原分派账户的净贡献 ITT，并把退款未决账户单独报告。假设的反事实仅说明归因数和毛利账本都无法单独识别广告新增。

SVSAC | sv-sac.org

20.3 结果模式及其解释边界

若广告策略提高付费点击和归因销售，但总渠道贡献不变，可能是自然/其他渠道迁移、边际广告成本抵消毛利、或归因重复；需要分别检查总量和成本。若总交易增加而利润不增，可能是折扣或低毛利产品占比上升。若品牌态度上升而交易未变，可能是窗口过短、态度未转成行为、或品牌端点测量不稳定。若模型预测的活动增量很好而政策组合结果无改善，预测目标与预算目标可能错位，也可能存在执行干扰或均衡效应。

这些模式不能靠事后挑选一个有利指标来证明主张。应预先指定一个主要商业终点、少数机制终点、观察窗口和退款成熟规则，并记录所有预设结果。次要和探索性指标可用于生成新假设，但需在独立样本或后续周期确认。若结果因精度不足而无法排除有意义的正负差异，应称不确定，而非“无效”或“成功”。

综上，点击与交易之间需要证据桥梁，交易与利润之间还需要全成本与反事实，利润与品牌长期价值之间则需要更长观察与可复核的品牌测量。每一级的证据增加了约束，也缩小了能直接声称的范围；这种分层让失败点可被定位，而不是把整条价值链统称为“精准”。

21. 把市场间干扰纳入试验设计

21.1 随机单位不是默认的“用户”

投放策略常在用户层随机，但竞价、预算、创意、经销覆盖和组织执行可能跨用户共享。一个用户被分到控制组，不代表其所在市场没有广告影响：竞价竞争会改变该地其他竞品的展示价格；销售团队可能因看到处理市场的表现而提前调整控制市场；多触点活动也可能令同一客户在不同设备、账户或渠道间被重复分派。因此实验单位需要与政策实际改变的资源边界匹配，而不是因为日志每行都是用户就选择用户随机。

在完全隔离的市场、账户或独立预算组中随机，能减少某些跨单位干扰，但不能自动保证无溢出。需要先画出处理如何传播：相同账户的多名联系人是否共用邮件或销售人员；一个市场投放是否改变另一个市场的竞价；一个创意池是否把审查和预算从对照市场抽走。若跨簇路径无法完全消除，可以把它作为设计变量，而不是事后仅在回归中“控制掉”。

SVSAC | sv-sac.org

21.2 两阶段饱和度设计与估计对象

一种可检验的构造是先将市场簇随机到不同投放饱和度，再在部分饱和度簇内随机选择符合资格的账户接受模型策略。比如设零投放、半数账户投放、全量投放三种市场策略。每簇内部的账户名单和预算规则在分派前冻结。零投放与全量投放之差针对的是市场政策包的总效果；在半投放簇中比较被分派与未分派账户，可估计指定饱和度下的直接分派差；半投放簇的未分派账户与零投放簇账户之差，则可检验潜在的簇内溢出。三种差异不能互换。

为说明估计对象差别，设无投放市场的每户总交易贡献为 100 单位；在半饱和度市场，分派投放户为 106、未分派户为 98。该构造中，半饱和度下直接差为 8 单位，而市场平均为(106+98)/2=102，比无投放基线多 2 单位。若全量投放市场的户均结果另为 103，则全量政策与不投放的差为 3 单位。三个数均是假设潜在结果，不是观测数据；它们说明直接处理差不能代替总政策收益，半投放饱和度也不能自动揭示全量推广结果。

干扰方法研究提出的设计逻辑包括明确随机分派机制、把分派映射为实际暴露状态、再定义对应因果对比；其权重和方差估计依赖已知分派概率与可陈述的暴露映射。[6] 该研究讨论的是一般干扰框架，并以学校社会网络试验为例，不是广告平台的现场证据。将其用于营销时，簇边界、暴露映射和可观测概率都需在本企业环境另行成立。

21.3 设计无法覆盖全部竞争反应时

若广告库存通过同一拍卖分配，处理与控制市场即便按地理位置区分，仍可能共享竞价者、预算上限或受众池。此时实验控制市场可能因处理侧多花预算而获得更便宜展示，或者更贵展示；处理差异同时包括直接触达、拍卖价格和竞争者反应。账户随机只能识别当前混合饱和度下的分派政策，不能无条件外推到全量投放后的均衡。

应记录处理前后可见的投放份额、竞价价格、频次、跨市场账户重叠及同步活动。若没有随机化饱和度，竞争反应只能作为未识别机制或敏感性分析，不能用稳定的竞价日志证明“没有干扰”。实务中也可以把问题收窄为固定预算下的相对分配政策：在共同预算池里随机使用模型排序或现行规则，比较原始预算簇的总贡献。该估计回答两种规则如何分配同一资源，不是广告全开与全关的市场均衡效应。

SVSAC | sv-sac.org

22. 从设计目标到现场协议

22.1 冻结资格总体与可操作的策略包

预注册首先冻结合格总体。名单需带有账户/客户唯一键、市场、入组时点、历史购买和当前渠道资格；重复身份的合并规则必须在随机分派前确定。若一个消费者有多个账号或同一企业有多个品牌账户，需说明分派层级并将可识别的关联账户聚在同一簇。未通过资格、无法合法触达、库存不足或不在目标市场者，应从预先定义的分母排除并记录原因，不可在看到分派结果后替换名单。

接着定义可执行处理包：候选创意版本、目标受众、预算、频次限制、竞价/出价规则、落地页和促销条件。若策略臂使用 AI 生成内容，对照臂使用现行创作流程，必须明确制作时间、审查人员和修订轮数是否属于处理包。若要分离模型与流程，需另设能支持比较的臂；若业务只能部署组合策略，主要结论应描述组合效果而非算法独立效果。

分派前要确定平台/内部日志所需字段：原始随机概率与随机数、政策版本、创意 ID/哈希、预算批准值与实际消耗、合格/未投放原因、展示与点击事件时间、账户映射版本、订单和退款链接、跨渠道去重键。关键日志丢失应按预定缺失处理报告，不能只保留可归因的成功曝光记录。

22.2 主要终点与审计顺序

为避免某个部门把有利指标选成主结论，主要端点应在试验前选定并定义经济主体、观察单位和窗口。示范性主要端点可以是每个原始分派账户在 90 日内经财务口径核验的净贡献；若品牌周期明显超过 90 日，则 90 日结果只能作为短期端点，另设更长期随访。计算式需说明销售扣除折扣、退款、可变履约成本和广告费用的时点。采购订单、支付、发货、退货、回款不能当作同一事件。

次要端点可以包括总订单、自然/付费渠道来源、独立访问、品牌搜索、复购与品牌测量；每项写明分母和测量版本。若品牌调查由实验引发，邀请与答复会产生选择，应保留邀请名单、回应率，并在可能时对不回应作边界分析。审核员可以不知道广告策略臂，但负责财务归类者通常会接触源记录；应将其工作规则锁定，抽样双评争议并保存原始裁决轨迹。

结果表需同时呈现 ITT、实际投放率、平均广告费、未投放原因、缺失/失访比例、退款成熟度和成本组件。只有在原随机分派对比下，才把效果称为策略 ITT；根据实际曝光或点击重新分组只能作为补充的不同估计对象，不能替代主分析。

SVSAC | sv-sac.org

22.3 预先制定停止与错误成本规则

现场投放可能遇到品牌风险、价格错误、库存不足、投诉或预算失控。停止规则应明确由谁触发、哪些事实足以暂停、暂停后账户如何仍纳入原始分派分析、已发广告如何处理。安全/合规审核不是随机取消的处理变量；不适宜素材应在进入实验前排除。若需要比较审核流程，比较的只能是都满足底线要求的合格方案。

分析方面，应指定主要对比、有限的预设异质性、成熟窗口和中期查看规则。若团队多次查看区间后选择停止或变更预算，普通固定终点置信区间可能不再对应原始计划。可采用预先写明的序贯规则，或限制中期查看并标记所有探索性结果。对未达到业务精度的结果，应保留“不确定”状态，不能用预测模型填补为确定商业效果。

23. 成熟结果、退订与缺失不是同一种失败

23.1 固定窗口与不成熟交易

广告政策可能改变回访和购买时间，实验组与对照组在固定截止日的成熟比例可能不同。固定 90 日窗口统一的是观察截点，不意味着所有交易达到同样的成熟状态。若退款通常晚于订单，按订单日结束实验会把尚未退货的金额误算成最终贡献。简单剔除尚未成熟的账户也会根据处理后结果选择样本。

可在研究开始前定义“财务成熟”为订单发货并经过预设退货/退款窗口，另报告固定日历截点下的暂估总额、成熟账户比例和未决金额。主要 ITT 仍按全部原分派单位计算；对未决价值给出明确区间，或等待预先规定的财务成熟日后重算。若经营决策必须在更早时间发生，则应报告该早期可得估计，并将其与最终成熟账本区分，而非把早期预测称作已实现利润。

23.2 失访、未触达、取消与不合格

账号失联、广告未成功投放、用户退订、订单取消、竞品中标以及客户暂缓购买不能合并成一个“未转化”标签。未触达常是处理实施结果；退订是可能的保护端点；订单取消或退款是交易路径状态；竞品中标对固定项目机会可能是竞争终止，但账户未来仍可能购买；延期则可能仍处于风险集。每种事件应按主体、机会 ID、截止时间和原端点定义标注。

退出分析前先保留原始分派总体及每种状态数量。对有共同截止日的二元结果，可同时报告最坏/最好边界：把处理组未决账户都视为未达成、对照组未决都

SVSAC | sv-sac.org

视为达成，得到保守下界；反向赋值为上界。此界可能很宽，但比只比较“最终知道结果者”透明。若事件时间方法要求独立删失或其他结构，必须明确假设并另做敏感性；延期和取消并非天然可以删失或视作同一竞争事件。

23.3 跨期品牌结果和政策更新

品牌搜索、复购或态度随时间测量时，同一账户的重复观测相关，不能将每一次访问当独立样本。调查样本损耗可能按市场、兴趣或处理而异。建议同时报告原始分派组别的响应率、各波次留存与被观察的可比性；模型调整可以提高精度，但不能证明失访随机。若营销策略在随访中继续更新，估计对象是“初始策略后按约定继续运营”的动态政策，而不是停止所有营销的纯品牌记忆效应。

若需单独识别持续品牌影响，可固定后续活动、或者把是否继续营销也纳入分派设计；这会增加样本与运营成本。许多现实企业无法长时间保持市场隔离，因此必须限定外推到实际测试的持续活动制度。试验结束后的品牌搜索增量也可由公关、产品上新或竞品变化造成，需记录重大同期事件，不能把时间相关性当成广告记忆机制。

24. 结果解释与反证条件

24.1 事前写下不同主张的可推翻模式

“生成内容更快”是生产率主张；“更多人看到”是交付主张；“更多点击”是互动主张；“增加交易”是需求主张；“提高贡献利润”是经济主张；“积累品牌资产”是长期品牌主张。这些主张需要不同的数据和反事实。生产时间缩短且可上线率降低，尚不能单独推翻“端到端内容产能提高”：须按同一期间或资源约束比较候选数×按时合格上线比例，只有该可用产出未提高，才不支持对应产能主张；广告费归因订单上升而付费+自然总订单不变，不能支持新增需求；总订单增加但净贡献下降，不能支持获利；态度短期提升但长期搜索/复购没有方向一致的变化，则长期品牌机制仍未确立。

不应因其中一个端点未达预期就改换为另一个结果宣称成功。事前将主张、估计对象、责任主体、观察窗口和可推翻条件写在一张预注册表内，分别报告估计值和区间。对多终点或多个市场的结果，明确哪个是主要、哪些是探索，不根据方向或显著性临时重新排序。

SVSAC | sv-sac.org

24.2 替代解释的检验优先于模型增复杂度

若结果变化与模型上线同步，首先查执行、媒体价格、预算、活动日历、库存、竞争者和测量版本是否也变了。若广告策略效果在不同市场反向，分析目标权重和处理覆盖；如果复购统计降低，查客户身份关联或观察窗口；如果付费 ROI 改善而总利润未变，查渠道重分配和费用遗漏；如果点击变多但成熟购买未变，查流量质量、归因窗口和退款成熟。

模型更复杂不自动解决这些偏差。把更多行为特征加入预测器可能改进排序，却不能恢复未投放人群的无广告结果或自动识别跨市场溢出。若分析指出关键对照缺失，应缩小结论、补充随机支持，或报告边界；不要把学习算法的拟合输出误称为观测反事实。

24.3 试验结果怎样进入扩展决策

即使受控试验显示正向 ITT，预算扩展仍须考虑推广饱和度、拍卖价格变化、创意审查产能、销售承接和可服务库存。一个区域内有限预算的效果不等于全部区域同时扩张后的均衡收益。可先按事前选定的市场簇逐步提高覆盖，在每个阶段保留同期对照并更新投放价格、总渠道贡献和保护结果；每个阶段回答一个预先声明的问题，避免“滚动上线”导致无对照的长期趋势叙述。

阶段门应区分可行性、因果效果和扩展价值。可行性判断问日志、审核和资金是否能够完成；因果判断问固定政策是否改变了总体端点；扩展判断问新增预算在更高饱和度时是否还有正边际贡献。第一阶段顺利只说明可执行，并不证明后两层。若竞争反应或容量约束在扩张后变化，应重新识别，而非把早期局部效果线性放大。

24.4 本章方法的边界

上述协议是可检验的研究方案，不是本次分析已经执行的企业实验。本篇没有任何企业级广告投放日志、平台分派机制、客户关系数据或财务账本。公开研究中的广告主样本、历史平台和结果定义不能代替特定企业的技术与商业验证。干扰方法文献提供的是在给定设计、暴露映射和概率支持条件下如何定义与估计因果量的框架，并不自动给出适用于市场竞价的正确簇边界。

当企业无法随机预算、无法看到自然渠道结果或无法长期保留客户身份时，可做的结论会受到结构性限制。预测模型可能在已有数据上帮助选活动，但“工具分数更准”不等于广告有增量、“局部实验有效”不等于全市场扩张获利。对无法识别的项目应标明缺失证据，优先设计能区分最重要竞争机制的下一步试验，而不是把不确定性隐藏在综合 ROI 中。

SVSAC | sv-sac.org

25. 品牌资产不是一个曝光后的指标

25.1 先定义品牌构念，再选择测量

“品牌提升”容易成为未经定义的汇总词。消费者基础品牌资产的经典框架把它定义为品牌知识对消费者营销反应产生的差别影响，品牌知识由品牌知名度与品牌形象/联想构成。[7] 这一定义强调反事实比较：相同产品或营销元素贴上真实品牌名称，与贴上虚构或无品牌名称时，消费者反应是否不同。它是概念与测量框架，不是 AI 广告或某企业利润变化的实证结果。

品牌知名度还需区分提示后的识别与无提示回忆；形象不是一个单项好感分，而包括联想的类型、强度、好感、独特性和相互一致。即时广告回忆可以表明受访者记得素材，但不说明品牌联想更清晰；广告喜好也不能代替品牌偏好。若实验只问“你喜欢这个广告吗”，其估计对象是广告评价，不是消费者基础品牌资产。

因此品牌指标在试验前应与待验证的经营主张对应。若主张是提高认知，预先定义无提示回忆/提示识别、受访总体与测量时点；若主张是改变联想，说明相关性、可信度、独特性如何操作测量；若主张是形成品牌差异反应，就比较真实品牌营销元素和产品属性尽可能相同的无品牌/虚构品牌版本。若主张是财务品牌价值，则还需识别品牌知识如何改变选择、价格反应、复购和成本，不可由一项态度量表跳到估值。

25.2 间接测量与直接测量互补

Keller 区分两类测量思路：间接法测品牌知识本身，用多种指标理解知名度与联想结构；直接法测试同一营销元素在真实品牌和虚构/无品牌标识条件下，消费者反应是否存在差异。[7] 单一指标不能覆盖多维构念，间接与直接测量回答不同问题。间接问卷发现某联想增强，说明可能的知识结构变化；直接对照发现营销反应变化，说明品牌知识可能改变了消费者对营销组合的反应。两者结合可定位“知道了什么”与“这是否改变选择”。

数字广告实验可采用分阶段设计。第一阶段随机政策可估计曝光安排对总体认知/联想指标的影响；第二阶段在独立样本或预先指定测试中比较品牌标识与中性标识下对相同卖点、价格、渠道条件的反应。把两阶段结果相乘来计算品牌收益并不成立，除非中介机制、样本关系、时间和尺度都满足额外假设。品牌知识的变化也可能是活动后购买、客服服务或经销商体验的结果，而非广告独立作用。

SVSAC | sv-sac.org

直接品牌对照在情绪型或高度视觉化创意中尤其难做到“完全相同”。真实标识本身会改变配色、包装、人物或熟悉度；拿掉 logo 后，素材可能变得不自然。此时更诚实的比较是产品/创意政策包的总体差异，再将识别到的总结果称为品牌化营销的效果，而不是品牌资产单独贡献。概念框架提供了需要测量什么的方向，并不保证实验能构造完美无品牌反事实。

25.3 生成内容可能强化熟悉语气，也可能稀释独特联想

生成模型可能复用历史高点击元素，使内容看起来更一致、更熟悉；也可能根据通用训练模式生成竞品都能使用的模板，使品牌独特性下降。二者对短期注意力未必同向：熟悉表达可能更容易理解，独特表达可能更难被即时识别，却有助于形成区分。单看点击量无法判断品牌联想是否更强、更独特。

一个可检验的创意比较需要先定义品牌核心联想，再盲评或消费者测量素材是否传达这些联想，并同时测量联想的强度与独特性。素材可以随机给同一市场的受众看，但同一模板的多个渲染不等于独立品牌策略。按模板簇分派或抽样，记录生成提示、人工修改和版本，避免“模型写了更多不同句子”被误报为“品牌更加独特”。若内容质量评审由知道来源的内部团队完成，评价可能受到对 AI 或设计团队的预期影响，应与盲评消费者结果分开呈现。

26. 识别“广告有效”与“品牌带来差异反应”

26.1 处理定义必须避免把两个问题合并

广告是否促成结果与品牌名称是否改变消费者反应，是两个因果问题。一个两臂广告/无广告试验可估计在当前品牌下投放政策的整体效果，却不能单独识别品牌名称的贡献。另一方面，同一广告在真实品牌和虚构品牌之间的比较，可测量品牌知识造成的差异反应，但不回答投放本身比不投放更好。

若研究同时要看两个问题，理论上可设置多个随机条件：真实品牌广告、同一卖点的虚构/匿名品牌广告、无广告对照；必要时再比较品牌文案与模型生成文案。主要对比应分别是品牌广告政策相对无广告的效果、真实品牌呈现相对匿名呈现的差异、以及创意来源策略间的差异。每个对比都需明确是否包含品牌熟悉度、素材制作、分发和媒介成本。品牌广告对无广告若高于零，但真实品牌与匿名版本无差异，说明广告可能有效，却没有证据表明增量来自已有品牌知识。

现实中无广告臂常难以完全实现，因为其他营销活动仍在发生。可以把“当前其他营销照常”的政策设为共同背景，对一个目标渠道随机投放与保持不投放。此

SVSAC | sv-sac.org

估计是目标渠道在既有营销环境中的边际贡献。若该渠道与其他活动有互动，应保留并报告这种策略环境，不宣称孤立广告的效应。

26.2 目标人群、品牌熟悉度与构成效应

知名品牌的触达样本可能本来就更熟悉；新市场或冷启动受众则可能不了解名称。若模型把已有互动者优先排序，广告反应较好可能是样本选择，而非创意提高品牌价值。实验可在分派前测量或用既有数据预设熟悉度层，再报告各层的意向处理效果。层内样本需要同时有投放与对照支持；若低熟悉度市场几乎不被触达，就无法从原日志估计广告对该市场的品牌提升。

品牌熟悉度问卷本身可能提示受访者注意品牌，影响后续记忆测试。因此事前熟悉度可用独立样本或已有的、事先收集指标，也可在设计中采用不被试验处理污染的基线测量子样本。不可将处理后回忆题用来定义“原本熟悉的人”。市场之间的品牌翻译、同名品牌冲突、经销商覆盖和渠道服务差异也可能改变测量含义；跨市场比较需先验证量表和品牌识别规则的可比性。

26.3 评分机制的复核

品牌调查的响应不是随机发生。广告可能提高受访者兴趣或增加烦扰，进而改变答卷率；只分析完成问卷者会按处理后响应选择样本。记录邀约、打开、完成和被排除各步数量，按原分派报告回应率。若回应率差异明显，应给出非回应敏感性，而非假设未回答者与已回答者相同。

开放式联想编码也会引入观察者偏差。编码前固定字典和新增词规则，抽取一部分由两名不知道处理臂的评审分别编码，报告一致性、争议类别和裁决过程。编码器不能看到生成模型版本或投放成绩。若新联想超出旧字典，应由盲态评审依据预定准则扩充，保存字典版本和原始回答；只把有利于品牌主张的新词加入指标，会引入结果导向的测量。

27. 从记忆变化到长期行为的时间与数据结构

27.1 回忆、联想、选择和复购是不同层次

一次曝光可短时提高广告回忆，但记忆会衰退；反复接触可能巩固品牌联想，也可能使素材疲劳。真实购买还受价格、产品质量、库存和竞品影响。研究需区分广告识别、品牌回忆、联想检索、偏好、实际选择与重复购买，不能把不同时间点和构念拼成一个“品牌 lift”。

SVSAC | sv-sac.org

可以预先设置曝光后即时、数周后和购买周期后的测量。若无提示回忆在即时波次提升、后续波次收敛，支持短期记忆而非持久资产；若联想结构持久变化但购买未变，表明认知影响尚未转化为行为或可能受其他约束；若购买提升但测量指标不变，需要检查价格促销、便利性、产品供给等替代机制。所有这些解释都要与实验精度相配，不能以“不显著”证明不存在。

27.2 观察窗与重复问卷的反应性

同一受访者多次回答品牌问题，可能因为重复提问而更容易回忆品牌；追踪样本也可能比初始总体更忠诚或更愿意答问。重复测量可帮助观察个人变化，但需要处理波次相关、失访和问卷反应性。另抽独立横截面可避免同一人反复受提示，却可能混入人群构成变化。两种方案估计对象不同，应说明是同一人的变化还是目标总体均值变化。

若同时追踪账户行为和调查，可预先将问卷抽样与主广告试验分开，或只随机邀请一个测量子样本，以降低问卷本身对全体行为的影响。调查结果与交易数据联结必须遵循适用的授权和数据边界；没有可联结的观察单位时，不要声称态度变化解释了订单变化。对失访者进行加权需要可观测的响应概率和足够的重叠；当未回应依赖未测品牌态度时，加权不能自动校正。

27.3 广告记忆的比较和因果中介

如果研究者希望说明“品牌回忆上升导致购买增加”，仅观察到回忆与购买相关不够。购买意图、品牌熟悉度、既有忠诚和广告分配都可能同时影响两者。随机广告处理可识别广告政策的总效果，却不随机化中介回忆；对处理后回忆条件化可能引入偏差。要识别中介机制需要更强的设计或假设，例如额外随机化能影响记忆的干预且不会直接改变购买，但这种排除条件常难在广告中成立。

更稳健的表述是分别报告随机分派对回忆、联想和行为的总处理效应，再讨论这些端点是否形成与机制相容的模式。相容不等于中介已证实。若总效果存在但回忆指标无变化，仍可能有其他品牌机制；若回忆提高而行为未变化，经营价值尚需后续验证。模型能够预测哪些人更可能记住广告，仍然不说明提高他们的曝光会创造净价值。

SVSAC | sv-sac.org

28. 品牌指标到财务价值的桥接

28.1 品牌态度不是财务估值

品牌资产有消费者视角、客户行为视角和财务估值视角。消费者基础品牌资产框架关注品牌知识如何改变对营销要素的反应；它讨论潜在的选择、忠诚、价格和渠道结果，但不会直接给出某品牌的货币估值。[7] 一项广告试验即使证明品牌知识有差异影响，也需要进一步估计销售量、价格、毛利、获客和留存成本，再计算可归因的现金流。用调查好感分直接乘以客户终身价值，缺少识别桥梁。

可将估值拆成若干可检验组件：品牌呈现改变选择概率多少；在可比价格下这是否影响转化；是否改变可承受价格或折扣；购买者是否更可能留存；维护认知的未来营销成本如何；现有产品质量和分销能否兑现承诺。每个组件需要自身数据和反事实。若产品质量升级与品牌活动同时发生，品牌溢价可能反映产品而非品牌知识；若经销商服务改善，长期复购也不一定来自广告。

28.2 一个假设的品牌差异反应账本

设一个构造实验把受访者随机看到同一款产品、相同价格和同一卖点，但产品标记为真实品牌或匿名品牌。假定真实品牌组 1000 人中有 440 人选择，匿名组 1000 人中 400 人选择，表面差为 4 个百分点。这个差值可作为该测量情境中的品牌标识反应差，但若结果来自态度问卷或假设选择，它不等于实际购买；它也不说明数字广告造成了这份品牌知识。

若下一步在真实购物环境复现同一差异，还需保证价格、库存、产品属性、页面位置、促销、物流和竞品供给可比，并以随机化分派为单位估计购买差。即便真实交易增加，也要验证每笔增量的毛利、退货、渠道迁移和广告费用。如果每新增选择需要折扣或媒体支出，净价值可能为负。以上人数和选择比例全是构造值，用来说明测量层级，不能当作任何品牌的实证。

28.3 价格溢价与需求弹性的边界

观察到品牌产品价格高而销量仍高，并不能单独识别消费者愿为品牌支付溢价。品牌通常与产品质量、渠道覆盖、服务、包装和营销同时变化；可比价格对照不充分时，价格差可能来自这些属性。即使消费者实验显示愿付价格不同，调查中的支付意愿也可能高估实际成交，且价格变动会影响购买人群和竞品反应。

价格与品牌机制可通过预设价格阶梯、随机化品牌信息或随机价格呈现来研究，但高价策略可能带来经销商行为、竞争者反应和不同客户构成。短期小样本的价格弹性不能直接推到全年组合利润。应分别报告实验场景下的品牌差异反

SVSAC | sv-sac.org

应、真实市场的价格/销量变化和总贡献，并对价格、促销和渠道共变给出敏感性。若无法随机价格或品牌呈现，结论应停留在相关关系或明确假设的模型估计。

28.4 价值兑现与反证

品牌记忆只有在后续产品与服务持续兑现品牌承诺时，才可能转成偏好和复购。若广告承诺超出库存、售后或经销网络能力，短期关注可能转化为投诉和负面联想；若当地分销不足，品牌认知增长也可能没有购买机会。AI 创意量增加无法补足产品质量或交付能力缺口。测试应记录客户投诉、退货原因和服务失败，不只看广告端点。

品牌价值主张可以被若干结果推翻或限制：同一品牌条件下模型内容提升点击但未提升品牌联想，削弱“创意建立品牌记忆”；联想提高但品牌/匿名对照反应没有差异，未显示品牌知识改变营销反应；短期品牌选择差异在价格与渠道控制后消失，说明原差异可能由市场供给解释；品牌量表提升但长期贡献不变，商业估值链仍未闭合。相反结果也需在独立样本、明确目标总体和可重复测量中确认，不能从一次多指标搜索中挑出一项做长期价值保证。

29. 已有生成式广告试验支持什么，不能支持什么

29.1 个性化视频的现场结果并非“AI 整体增益”

一项已发表于 *Marketing Science* 的现场实验在 WhatsApp 上与一家销售环保产品的直营电商品牌合作，随机把用户分配到三种广告条件：基于生成式 AI 的个性化视频、个性化图片广告和通用非个性化视频。摘要报告，AI 个性化视频的互动表现比两个基线高 6 至 9 个百分点。[8] 这是比概念演示更接近真实投放的证据，说明在该品牌、渠道与受众范围内，生成式广告个性化可能改善被测互动结果。

但三组处理同时改变了视频/图片格式、是否个性化，以及 AI 内容生成流程；按已读摘要范围，不能将差异单独归因于模型本身。研究结果是互动改善，不等于下单、毛利、回购、消费者福利或品牌资产。摘要没有提供足以复核各项结果细节的完整方法表格，因此本篇不推断处理效应的方差、利润收益、长期品牌变化或可迁移的用户分层。作者也明确指出生成式 AI 营销应用的效力和可推广性可能因场景而异。[8]

这个例子应改变核心问题的精度：不再问“AI 广告是否更好”，而是问“什么处理包，在什么受众、平台和结果口径下产生何种差异”。若管理者想知道生成模型

SVSAC | sv-sac.org

独立作用，必须在相同视频格式、信息、受众和预算下改变创意生产机制；若只比较商业上可部署的整体方案，结论就应是整体方案效应。外部研究不能代替本企业的分解试验。

29.2 从互动到品牌结果还缺哪些桥接

在该现场实验的设计中，用户购买历史被用于个性化素材，而通用视频对所有用户呈现一致品牌信息。[8] 如果互动上升，可能来自购买历史与内容相关，也可能来自视频与图片的刺激差、表达形式或创意长度。要判断品牌机制，还需观察准确的产品理解、品牌联想和后续购买；只用互动率不够。

本篇不能据其摘要回答新客与老客的不同增益，也不推断信息个性化对用户信任的方向。可检验的后续设计应把个性化依据（用户明确提供的信息、已发生购买、模型推断兴趣）、创意格式和生成方式分别控制或随机化，并预先定义投诉、退出和购买等端点。若互动改善只出现在已有购买者，增长可能是对既有偏好的更有效回应，未必带来新客；若新客中同样改善，仍需检查增量交易是否覆盖媒体和履约成本。

29.3 外部效度需要目标总体和支持域

某项直营电商品牌的 WhatsApp 结果不自动适用于以搜索广告、零售媒体或品牌社媒为主要渠道的公司。渠道用户构成、广告形式、关系基础和成交路径均不同。目标公司应在比较前界定自己的合格用户总体、媒体渠道、产品品类和购买周期，再判断已有实验在哪些变量范围内有重叠。若没有重叠，来源结果可提出假设，但不能被重新加权包装成目标企业效果。

更一般地，模型的内容生成能力、数据获取方式和实际投放环境可能随着时间变。2025 年实验说明的是当时该系统与具体组合的结果，不能作为 2026 之后所有模型或平台的默认参数。复现时需记模型/产品版本、提示模板、人工修改、素材审核、受众配置、平台政策和最终曝光方式。即使同一模型输出相似文本，广告竞价和受众注意也可能改变结果。

30. 预测赢家与在线实验：模型不等于评审者

30.1 纯语言模型对内容表现的判断有边界

LOLA 研究利用 Upworthy 新闻出版商的历史标题 A/B 测试资料，检验提示、嵌入和微调语言模型能否预测哪条标题更吸引点击。作者工作稿报告，纯 LLM 方法对最佳标题的识别能力有限，不能替代内容实验；随后提出把 LLM 预测作

SVSAC | sv-sac.org

为在线实验先验并与 UCB 流量分配结合的 LOLA 方法。[9] 研究的目标是新闻标题点击，后半部分基于历史资料的数值实验，而不是 AI 广告现场试验或品牌/利润效果。

这项结果重要之处不是泛称“模型预测很差”，而是区分了两个任务：模型可在有限证据下为候选排序提供先验；最终要选择的内容仍需受众响应数据。实验数据则承担校准和校验角色。把“模型说这条文案最好”直接变成投放决定，绕开了消费者实际反应；反之，将模型评分用于候选筛选或分配初始探索资源，仍保留了实验所提供的证据。

论文中的 Upworthy 资料包含大量历史标题和点击，但新闻标题的生命周期短、目标是获得阅读点击。来源不支持把点击选优外推为商业广告的增量购买、毛利或长期品牌效应。它对营销的启发是方法结构：候选生成、先验评分和实际流量试验可以组合，但奖励端点必须与公司目标一致。

30.2 平衡测试与自适应分配的取舍

传统 A/B 测试在预定阶段把流量分到不同臂，结束后再选择表现更好的版本。优势是设计和推断容易解释；代价是流量会继续流向已表现较弱的创意。在线学习可随证据动态调整流量，减少探索期内不佳版本造成的即时机会损失；但小样本时，早期随机波动可能令算法过早集中到次优臂。LOLA 把模型预测作为辅助样本或先验后，再以自适应规则分配流量，作者在历史出版数据上评估其点击回报。[9]

营销团队若使用自适应分配，仍需保存每次选择的概率和当时可用候选集，否则难以重建不同策略下的因果比较。算法随着观测更新，处理分配不再是固定比例；若分析忽略动态概率，常规简单均值差可能不再回答预期估计对象。可事前采用适配设计和对应的推断程序，但该方法在新闻点击目标上的结果不能证明其在企业贡献利润上的表现。

探索流量应被视为真实成本。低效素材占用展示机会，可能影响销售；而几乎没有探索，后续就缺少新创意、市场或时段的支持。对新品或小市场，短时点击量很少，算法的“赢者”可能只是噪声；对季节促销，等待足够证据也可能错过有效窗口。设计时需为不确定性、上新周期、预算和用户风险选择探索比例，并报告策略本身的机会成本，而不是声称自适应必然优于 A/B。

30.3 目标错配会使优化算法更有效地做错事

若系统按点击分配曝光，而最终目标是贡献利润，算法可能更迅速地发现“最会吸引点击”的版本，却不一定发现最有利润的版本。构造一个每臂各展示 1,000

SVSAC | sv-sac.org

次、费用均为 50 单位的例子：创意 A 点击率 5%，点击后购买率 2%，每单毛利 40，预期 1 笔订单、毛利 40，净贡献−10；创意 B 点击率 3%，点击后购买率 10%，每单毛利 70，预期 3 笔订单、毛利 210，净贡献 160。按点击数选择 A，按构造的净贡献则选 B。所有输入均为假设，说明奖励函数定义决定了被优化的结果，不是实际活动数据。 此算例另假设所有购买均经观察到的点击发生、无点击外成交，且每单毛利已包含除列示广告费外的相关成本；现实不满足时不能据该漏斗乘积推总成交或因果增量。

把优化目标改为购买，也仍可能遗漏退货、折扣、渠道转移和长期客户价值。若不同产品每单毛利、退款概率和履约成本不同，订单数也不是统一经济单位。系统可以优化任何编码后的奖励，但不会自动补充未进入目标函数的利益相关者结果。上线前要由财务和业务负责人选择主奖励、保护指标和停止条件，并明确哪些端点只用于观察而不作承诺。

31. 创意策略的贡献分解与实验结构

31.1 三个问题决定三种设计

若问题是“生成系统是否比人工更能产出可用创意”，就应比较产出成本、事实/品牌审核通过、可投放率和创意效果，同时尽可能保持渠道、受众和频次相同。若问题是“个性化内容是否增加说服力”，就需固定内容生产方式和媒介格式，只改变个性化规则。若问题是“整体 AI 工作流能否改善预算决策”，处理包可能同时包含生成、筛选、编辑和投放，需要以整套政策对比现行流程。

这三个问题对应的随机单位不相同。个体曝光可以按用户/账户随机；创意生产能力需要多独立素材簇；完整营销流程可能要按品牌、市场或周期簇随机，以免同一审稿者和创意库污染各臂。把许多用户看到同一支视频误当作许多独立 AI 生成策略，会低估对创意簇的不确定性。测试需要同时报告参与的独立素材数、内容模板数、分派用户数及重复曝光结构。

若设计只能支持一个组合策略比较，应明确哪些组件无法拆分，保留可再检验的组件假设，不从总效应中倒推出“AI 模型贡献了全部提升”。若一个处理臂在生成后经历更多人工编辑和事实检查，结果包含这些劳动输入；这可能是值得部署的工作流，但不能称为模型的纯自动效果。

31.2 版本与候选选择的独立确认

每个候选创意应保留模型版本、系统提示、输入资料快照、人工编辑、品牌与法务审核时间以及最终素材哈希。上线候选最好在筛选阶段与确认阶段分离：

SVSAC | sv-sac.org

探索流量用于生成或筛选；确认流量比较冻结的最终候选与基准。若确认失败后修改素材再使用同一确认人群反复挑选，结果已参与下一轮决策，不再是独立证据。

创意模板的共性会让不同广告相关。若同一提示和布局只替换产品名称生成几十个版本，这些版本共享生产机制，不应被当作相互独立的创意策略。统计不确定性至少要反映素材/模板簇；若最终政策要扩展到新的内容类别，也要有足够的独立模板覆盖。纯用户级标准误不能解决创意层样本不足。

31.3 把学习速度与最终质量分开

自适应或 AI 先验可能更快选出一条表现不错的创意，但“更快找到赢家”与“赢家长期表现更好”不是一回事。内容生命周期短时，节约的探索展示具有经济价值；若品牌广告持续数月，短期点击胜出不代表后续疲劳更少。报告可分两项：在达到预定证据门槛前累计的投放机会成本，以及冻结版本在独立确认期的点击、交易和品牌端点。若只报自适应期总点击，策略会把学习阶段的所有数据和候选更替混合，难以判断终选质量。

32. 从候选效率到全周期经济评价

32.1 生成成本不是工作流成本

生成一个可投放素材涉及模型使用、提示准备、内部数据整理、设计编辑、翻译或市场校对、产品事实核验、品牌与法务审查、平台适配、投放和后续监控。即便单次模型推理费用很低，新增素材会扩大人工审核和实验成本。成本单位需分辨按素材、按模板、按市场、按账户和按活动发生；一次性模型接入或素材系统建设与每轮重复成本也不同。

评估一个完整政策时，可列成本组件：一次性接入/培训、固定人员工时、按产量变动的审查工时、媒体支出、实验流量机会成本、失败素材的返工、未通过审核的处置和退款/投诉风险。应把现金支出和机会成本分开，并防止同一人工时间既计入编辑工时又重复计入“生产损失”。若人员工时只是被释放而非减薪或裁员，它是容量收益，不是自动现金节省。

32.2 质量门与成本门同时报告

假设生成工作流每周产出创意增多，但最终审核通过比例下降，净上架数可能仍增加，也可能审查排队导致活动错过窗口。需同时报告原始候选数、事实核验通过率、品牌适配通过率、发布数、返工次数和从需求到上线时间。发布率是

SVSAC | sv-sac.org

过程量；客户投诉、事实错误、广告屏蔽和订单退款则是结果或保护端点。任何一项都不能替代销售增量。

成本效果可按“每个最终审核通过并上线的创意成本”和“每单位增量贡献成本”两层报告。前者帮助诊断产能，后者回答经济价值。若成本按每个候选摊销，失败候选和返工的成本也要计入分子，否则工作流看起来会因大量低成本尝试而变廉价。若上线创意数不变但编辑时间下降，应报告释放的时间及其后续用途；没有将时间转为新增活动、缩短上市周期或现金节省的证据，不能直接称为利润提升。

32.3 做敏感性而不是伪精确点估计

企业初期可能不知道审核返工率、素材寿命、增量毛利和身份链接比例。可将这些量分别设为可观察区间，计算净贡献或单位收益在不同组合下的范围，不把单个模型预测值当成基线真值。若结果只有在非常高的长期品牌溢价、极低审核成本或没有退货的假设下才转正，决策者应看到这一依赖。

敏感性分析不能替代随机对照。它回答在一组假设下决策结果如何变化，而不是这些假设是否真实。重要缺失参数应通过下一阶段实验、成本记录或成熟财务数据补充。若不同目标之间不可直接换算，例如品牌联想和即时现金利润，可先报告多目标结果与资源约束，让管理者明确选择，而不是任意给每个指标加权成一个看似精确的 AI 总分。

33. 品牌词广告的流量替代：两个现场设计究竟比较了什么

33.1 eBay 实验的对照并不是“广告开与关”的简单前后差

Blake、Nosko 与 Tadelis 的 eBay 研究提供了一个有用的警示：显眼的付费点击不等于同等数量的新增访问。作者报告的品牌词实验中，eBay 在 Yahoo 和 MSN 上暂停含“eBay”的关键词广告，同时继续在 Google 购买品牌词广告；后者被用作季节变化的对照。按作者 2014 年 8 月 12 日公开的 forthcoming 稿，控制 Google 趋势后，MSN 来源的总点击量损失约 0.529%，作者据此写作约 99.5%的总点击保留。脚注又给出一个常被省略的换算：总点击损失的 0.5%约相当于全部付费点击的 1.5%。这两个比率回答不同问题。前者以该搜索平台送至网站的总点击为基准，后者以付费点击为基准；把“99.5%替代”直接与另一研究“约一半付费流量未被补回”并列，会造成分母错配。[10]

该研究并非所有平台、所有日期都进行了随机用户分派。品牌 MSN 实验主要依靠平台间的前后比较，以 Google 的品牌搜索流量校正季节变化；随后另有

SVSAC | sv-sac.org

Google 平台暂停并恢复的试验，但该段缺少可比的同期控制。把它概括成一个覆盖用户、平台和品牌政策的统一 RCT，会夸大识别范围。作者还报告品牌搜索的归因销售基本被自然搜索吸收；但短期总点击和特定销售口径仍不能自动回答长期品牌记忆、竞争者获客或多渠道贡献。[10]

33.2 地理分派与竞争条件的待检验场景

Coviello、Gneezy 与 Goette 的 Edmunds 工作论文仅作品牌搜索研究待核线索。[11]原文未取得，本篇不确认其 DMA 数量、分派、观察窗、竞价条件或流量结果，也不将这些信息作为 eBay 案例的实证比较组。后文改用明确条件场景讨论停投政策。

比较流量变化必须先写清分母：相对基线总流量的百分点、相对付费点击的百分比及总访问变化是不同量。本篇不再引用未核表 2 数值或假定其实际停投程度。实际分析应分别报告分派政策、执行强度和端点；有未执行者时，意向停投效果不等于完全停投的政策效果。

可读 eBay 作者稿中 99.5%对应其 MSN 受控比较的总点击留存。[10]该数字不能与未核 Edmunds 结果合成平均替代率。跨研究比较须先核对象、分母、对照与停投范围；仅有相似标题或比例外观不足以比较。

33.3 对照设计的经济含义

eBay 在部分搜索引擎停投、Google 继续投放的设计用于其特定对照。[10]另可设想按地理市场分派、竞争者仍在场的政策，但这里不称其为已核 Edmunds 实验。若竞争者随停投调整出价，结果可能包含竞争反应；它不同于清空所有广告后的自然流量反事实。

因此，“停止自己的品牌广告”的商业问题必须与真正拟部署的政策一致：按搜索平台停投、跨平台全停、保留竞品位、限制竞品位、只降低出价，都是不同处理。平台和竞价市场可能对预算变化作出响应，长期均衡也可能不同于数周内的短期流量移动。若目标是判断某个品牌是否削减预算，局部停投的实验结果可提供其观察窗内的增量证据，却不会自动给出全行业同时停投后的均衡结果。

34. 不同现场结果不是“规模效应”的随机检验

34.1 观察到异质性，不能直接识别其来源

“大品牌靠自然搜索、小品牌必须守品牌词”仍是待检验机制，不由一个可读案例和另一个未核研究构成验证。企业规模与品牌熟悉度、竞品强度、平台覆

SVSAC | sv-sac.org

盖、自然排序、查询组合、设备、地域及时间可共同变化，跨公司差异不能自行分解这些因素的因果贡献。

若拟检验基线付费流量占比是否调节停投效果，应预先定义交互与分层，并区分随机政策和非随机市场属性。基线占比还可能反映竞争、品牌需求或自然排序；即使观察到交互，也不能直接解释为增加竞争者使广告更有效。这里是作者的检验建议，不报告未核 Edmunds 异质性结果。

34.2 可检验机制需要在一个设计内操纵

至少有三种机制可产生相似的观察结果。第一，品牌导航习惯强时，用户更可能点自然结果或直接输入网址。第二，竞争者在广告位展示时，停投会把部分点击让给竞争品牌。第三，品牌词流量在某些地域更常来自购买意图，而在其他地域更常来自信息查询。仅凭总流量变化无法区分这三种解释；自然点击、直接访问、竞品展示、下游订单及新客标签需要同时定义并跟踪。

能区分机制的现场研究可以在账户或市场层随机安排至少两个处理维度：自有广告开/关与竞品广告位开放/受限，或自有广告强度与竞品可见性。在用户之间共享拍卖、排名或预算时，随机化单位必须覆盖干扰范围，或者设计不同处理饱和度并预设直接效应与市场溢出效应。品牌词查询量本身也可能被早期处理影响；若只把处理后出现的查询纳入分析，会按处理后的行为选择样本。更稳妥的主分析在实验前锁定合格市场/受众框，并按原始分派计算总结果，同时把查询发生、广告展示、自然点击作为机制结果。

34.3 样本内外推需要明确加权对象

把某一高付费依赖市场的估计推广到其他市场，需要目标分层权重和层内迁移假设；缺共同支持时不能用一个总平均补空白。可读 eBay 结果也不自动代表经销商、区域服务商或新品牌。[10]这是迁移要求，不是声称未核 Edmunds 已提供可用分层估计。

一个报告应分别给出：已随机化市场的意向处理效应；按预先确定的目标市场权重标准化的效应；没有重叠支持的目标层及其边界。若市场层有效样本少，不能用大量用户点击数伪装独立样本量；不确定性应尊重 DMA/市场分派层级。比较研究间点估计时还要并列说明观察期、平台、停投强度、竞品是否继续展示、自然与直接流量口径，以及是否观察订单或财务结果。

SVSAC | sv-sac.org

35. 流量恢复与商业价值可能方向相反

35.1 一个显式构造的流量–贡献账本

下面的算例仅用于说明终点转换，不代表任何企业。假设某品牌在两个可比市场各有 10,000 次基线总访问，其中处理市场的品牌词广告停投。对照市场在观察窗内保持 10,000 次访问；处理市场付费访问减少 1,000 次，自然访问增加 400 次，直接访问增加 100 次。因此，处理市场净少 500 次总访问，广告位来源的 1,000 次付费访问中，500 次通过自然或直接路径补回。这个“50%补回”是对构造流量账本的描述，不能替代随机试验估计，也不能套用 eBay 或 Edmunds 数值。

接下来假设这些被保留的访问中，不同来源的转化率和订单贡献不同。处理组相对对照少了 500 次访问；若未恢复的访问中有 2%会形成订单，平均每单贡献为 80 货币单位，点估计的未实现贡献为 500×0.02×80=800。假设停投节省广告支出 1,200 单位，则该构造中的净贡献变化为+400。若未恢复访问转化率只有 1%，则损失贡献为 400，净变化为+800；若转化率为 4%，损失贡献为 1,600，净变化为−400。相同流量替代比例可以对应不同净价值，差异由未恢复访客的反事实购买率、每单边际贡献和广告节省共同决定。

这里将流量变化乘以转化率，只有在“被处理改变到达的访问”具有相应转化概率且没有其他订单路径变化时，才能被解释为因果贡献。观察性来源转化率不是未访问者的反事实转化率；访问被广告截获与访问本来会自然发生的用户并非可互换。实际评估应优先估计原分派对总订单和贡献的影响，而不是用已点击者的平均转化率补算未发生访问。

35.2 多终点报告防止单一流量比率越权

流量分析仍有价值：它能说明自然结果、直接导航和竞品展示如何替换。但对于预算决策，至少需要将终点分层：平台流量、网站访问、合格线索或订单、退货后的净订单、扣除媒体与履约增量成本后的贡献，以及更长期的复购/品牌指标。前一层的效果可以解释机制，却不能代替后一层的估计。尤其是一个访问可能来自多个触点，若把每个触点归因订单相加，会重复计算同一交易；应同时维护去重后的账户/订单总账。

有些成本也不随点击线性变化。竞价费可能因流量变化而省下，但固定创意、代理或平台最低支出不会自动消失；竞品展示增加可能影响后续自然搜索，短期窗口也可能遗漏回访。预算重新分配到非品牌词或其他渠道之后，整体政策价值还取决于替代活动是否挤占已有预算。把“停投节省”当成全部机会收益，忽略替代项目净贡献，会把局部账本误当全渠道净效应。

SVSAC | sv-sac.org

36. 面向 AI 营销的可证伪扩展

36.1 先分开模型能力与投放政策

品牌词停投研究检验的是广告可用性的政策反事实，不检验 AI 是否提升创意、受众选择或竞价。AI 可以改变候选素材、查询扩展、出价时机和受众划分；每一种都可能改变被处理对象及预算分布。若模型只提高点击预测准确率，却将更多支出导向本来会访问的品牌导航者，预算回报可能下降；反过来，即使平均点击概率预测无变化，若模型改善了对新增访客的筛选，也可能提升单位预算贡献。预测准确度与增量政策价值不能互相替代。

36.2 一个可执行的比较设计

对有足够品牌词流量的企业，可先冻结市场、搜索平台、品牌词词典、自然排序和竞品状态的基线记录，再按市场或不互相干扰的账户簇，随机分配“现行规则”与“AI 建议政策”。两组均保留法律、品牌安全和最低服务约束；AI 组的建议须记录模型版本、输入特征时间戳、推荐出价、实际执行和人工覆盖。若实际出价因平台拍卖随其他市场预算变化，应将分派解释为现行竞争环境中的政策效应，而不是固定环境下的模型效应。

主要结果可设为固定观察窗内的去重订单净贡献，按原始分派计算；品牌搜索流量拆分、自然点击、竞品点击和活动支出用作中间结果。预先定义订单成熟窗、退货/取消处理、贡献口径和缺失规则。一个预定的市场分层可检验效应是否随基线自然排名或品牌认知变化，但不应在多种切分中挑选显著结果。若要区分创意生成与投放策略，可增加一个“相同投放政策、不同创意流程”的独立随机臂；样本不足时，宁可报告组合政策效果，也不声称识别了单一 AI 模块。

36.3 预先写明什么结果会推翻建议

若 AI 政策组点击增加、但去重净贡献不升且支出增加，不能称为精准营销成功；若品牌词流量提升仅来自自然访问重分配而新客或利润无变化，预算向品牌词扩张的解释也不成立。若估计区间仍覆盖实质性正负贡献，结论应为精度不足，而不是按点估计扩大投放。若模型带来的收益只存在于训练/选型市场，独立确认市场无效，则更可能是选择或过拟合，而非稳定的政策改进。

这类研究仍不能凭一个季度的市场随机试验回答长期品牌资产问题，也不能把在平台允许范围内的出价试验推广到其他国家的拍卖规则。它能回答的范围更窄也更有用：在明确的受众框、竞争状态、预算约束、模型版本与观察窗内，改

SVSAC | sv-sac.org

变 AI 辅助政策是否使净商业结果优于当时的基准做法；如果答案不确定，哪些数据或额外随机化才足以缩小不确定性。

37. 稀少商业事件：零转化结果可以排除什么

37.1 观测到零次不等于真实概率为零

对于广告试验，品牌切换、退订、投诉、取消订单、极高价值新客等事件可能远少于点击。若试验期内没有观察到某类结果，不能据此说其概率为零。一个透明的有限样本推导可以说明证据上限：假定有固定的 n 个相互独立、结果均已成熟且具有同一个事件概率 p 的合格机会，观察到零事件的概率是(1−p)^n。令这个概率等于单侧显著性水平 α，得到零事件情形下的单侧上界 p_U=1−α^(1/n)。在 α=0.05 时，n=50 的上界约 5.82%，n=300 约 0.994%，n=3,000 约 0.0998%。这些值只是该二项模型下的计算，不是通用样本量、绩效门槛或安全保证。

这个界回答的是“在上述模型下，尚未被数据排除的单一事件概率有多大”，不是“效果为零的后验概率”，更不是处理组减控制组的因果风险差区间。若处理和对照各有 n 次且都未出现事件，需要分别形成组内概率界，再对组间差作相应推断；不能把单组的 p_U 当作差值上界。若问题是严重投诉是否因 AI 广告增加，关键量是两组预设合格机会的投诉率差以及随机化不确定性，而不是全样本合并的零事件计数。

37.2 独立机会比原始曝光数更窄

上述公式的 n 不能随意取平台展示次数。一次创意模板可能被重复投放数百万次，但同一文案、受众算法、市场和投放日期共享生成过程，结果会聚类。若聚类内事件正相关，按独立曝光计算的区间过窄。实际设计应先界定随机单位和可重复决策单位，例如市场簇、账户簇、创意模板簇或预先定义的合格触达机会；推断应在对应层级聚合，或用与分派机制相容的设计型随机化推断。

同样，机会必须在观察窗内有机会成熟。若购买周期长，在仅观察到活动结束而尚未走完整个预定随访窗口时，未购买不能当作已成熟负例；否则 n 混入右删失结果。反之，完整固定窗口内确认没有购买，应就“窗内购买”端点记 0，不因可能未来成交而改为缺失；它不等于永久不购买。把所有展示都当成已成熟负例，会人为压低事件率；只保留已经转化或已经完全成熟的账户，又可能受处理影响的回访速度与样本选择影响。协议应预先定义统一的入组时间、观察窗、成熟日期、失联处理以及哪些结果属于延期而非失败。

SVSAC | sv-sac.org

当团队先逐日查看结果，出现“显著”就停、零事件就延长样本时，固定 n 公式不再适用于未调整的可选停止过程。一个可执行方案是预设最大样本和唯一正式读数时间；若业务需提前安全监测，应另外定义序贯界限和升级规则，不能在多次查看中沿用一次性 95%界限。模型、市场或内容版本变化后，旧版本的观察不能不加区分地充作新版本的相同机会。

37.3 概率上界不约束金额尾部

概率小不代表经济损失小。假设一次严重投诉事件在 300 个成熟独立机会中为零，二项界只说明在模型前提下事件率的单侧 95%上界约 0.994%。若每次事件的损失没有上界，则这个概率上界不能给出期望损失的有限上界。即使事件率被限制在 1%以内，只要单次损失可能极大，最坏情形的期望金额仍可能很高。反过来，一次高价订单也可能显著改变一小样本的平均收益。风险率与损失严重度必须分开记录，必要时报告各自的界限或情景，而不是把小概率直接换成“可接受”。

可复算的构造例：假设某 AI 广告组的 300 个独立成熟机会中没有发生退款欺诈，但某事件一旦发生的额外成本暂按 10,000 单位封顶。在二项独立假设下，概率上界约 0.009936；概率乘以封顶损失得到每机会约 99.36 单位的粗略上界情景。它只有在损失上限真实成立、成本口径完备、机会独立同分布且报告的是相应概率边际时才有意义。若企业没有合同或历史依据支持 10,000 上限，这个乘积不是统计推论，而是人为设定的情景。潜在的品牌损害、诉讼或跨账户传播也不一定被单次损失封顶覆盖。

37.4 处理风险与基准风险应分开估计

对于 AI 与现行规则的风险比较，分别报告处理组和控制组的绝对事件率、风险差、相对风险及其不确定性，前提是样本量允许。若两组都观察到零事件，数据通常无法支持“风险相同”；它只能对两边各自的概率给出较宽上界。若处理组零次、对照组出现事件，也不能把小样本的点估计差直接解释为 AI 降低风险。对于零事件与少数事件，需使用与随机分派和小样本相容的区间或随机化推断，并明确估计对象、单侧/双侧水平以及是否对多种事件作过选择。

更重要的是，安全决策往往针对严重性分层而不是总投诉数。一般误点、退订和涉及错误承诺的投诉不应相互抵消。对于严重内容错误，可以预设事件分类、盲态判定和升级处置；未知类别应作为未分类结果保留，而不由生成系统自行将其归入“无问题”。若事件稀少到无法排除有意义的风险差，正确结论是风险证据不足，采用更窄的试验范围或增加独立成熟机会，而不是声称零风险。

SVSAC | sv-sac.org

38. 从实验效果到全渠道部署：识别规模化后的变化

38.1 一个试验臂的效果不是市场全面部署效果

随机化市场或账户簇试验估计的是给定当前竞价、曝光和其他参与者行为下的分派效果。逐步把同一策略扩至更多预算、市场或消费者，可能改变拍卖价格、自然排名、竞品反应、渠道内外流量组合和素材疲劳。因而“试验中每增加一单位预算的平均贡献”不是任意规模下恒定的边际回报。此处的关键区分是：试验识别现行竞争环境下的局部政策效果；全面部署还需要对处理饱和度、竞争均衡和预算替代作出假设或另行识别。

一个构造例可展示这种非线性。假设固定媒体预算为 1,000 货币单位，小规模试验在该预算下产生 1,300 单位扣除履约、但尚未扣广告支出的增量贡献；这不是已观察企业数据。把同一预算扩大到更多市场时，若竞价使每单位触达成本上升 20%，固定预算能买到的机会数量因子是 1/1.20，而不是 0.80。再假设新增市场的单位贡献比原样本低 15%，则扩张后的贡献为 1,300/1.20×0.85≈920.83，扣除 1,000 媒体支出后净值约−79.17；基准方案的媒体前贡献减媒体费为+300。成本上涨 20%与机会数量下降 20%不是同一比例。两个折减仍完全是假设，不是广告市场定律；例子说明从局部试验到全量建议需要测量边际竞价成本和可服务池，而不能线性放大点估计。

38.2 逐级饱和试验能识别哪些部署问题

若预算扩张会影响同一市场中多个用户，单纯用户级随机分派可能把用户间拍卖干扰留在处理与控制之间。更可操作的方案是在市场簇随机分派低、中、高广告预算饱和度，若还要区分直接与间接效果，须如 21.2 节那样在各饱和层内再随机分派个体处理，保留可比的未分派单位，并预定义暴露映射。由此才可比较相同饱和度下的直接分派效果和未分派单位跨饱和层的间接效果；若簇内全用同一政策，跨簇只能识别预算政策包的总效果，不能自动拆出间接效应。两种设计都可另评总预算变化对应的增量净贡献。其有效性仍依赖市场间干扰可忽略、簇数足够以及每一饱和层有足够重叠。

分配单位必须与媒介交易方式相符。若同一广告主在多个市场使用共享预算，平台会把某个市场的消耗变化重新分配到别处，市场间就不再独立。可选择账户级预算池作为簇，或在设计中冻结共享预算并记录平台实际迁移；若无法控制或测量迁移，结果只能解释为这一共享池内的政策包效果。不能将素材曝光用户数当作独立的预算分派单位。

SVSAC | sv-sac.org

38.3 多阶段账本把“流量到品牌”拆成可观察转换

一个完整评估可以把过程分为：合格机会、实际竞价、赢得展示、有效观看/点击、首次到站、合格线索、订单、净订单、边际贡献、复购和品牌记忆。每一步都要有分母和去重规则。例如点击率的分母可以是展示，也可以是可见展示；从线索到订单的分母可以是线索队列，也可以是最终成熟账户。分母变化会改变比率，因此模型预测提高某阶段通过率，并不自动提高每个原始入组机会的最终贡献。

若某处理影响了中间选择，例如更多点击但来自低意向访问，按“点击者”比较后续转化会条件化于处理后变量，可能制造偏差。主分析应保留预先定义的合格总体和原始分派；对中间层级的条件比率用于过程诊断，不能未经额外识别就当成因果机制。若渠道归属可变、客户跨设备，订单按账户/交易 ID 去重，并对归属规则变化做敏感性分析。

38.4 当预算限制成为目标的一部分

当团队只有固定预算时，策略比较不应只问“AI 臂总收入是否更高”，还要问其是否挤掉了哪些原有项目。若一个广告活动从其他活动挪来 5 万预算，实验结果应计入被替代活动的反事实贡献；否则只是将收入重新归类。需要分别报告固定总预算下的组合政策效果，与预算额外增加时的增量效果。前者回答资源重新配置，后者回答扩张是否值得，两者需要不同的随机政策或预算档位设计。

38.5 流量“替代率”是比率估计量，不是流量差本身

可把处理对付费、自然和直接访问的原始分派效应分别记作 τP、τO、τD；如果停投令付费访问减少，描述性的补回比例可以定义为 $S=(\tau O+\tau D)/(-\tau P)$。它与总流量效应 $\tau T=\tau P+\tau O+\tau D$ 相关，却不是同一个量。若 $S=1$，观察到的自然和直接访问增加恰好抵消付费访问减少；$S<1$ 时有未补回流量；$S>1$ 可能表示额外自然需求、渠道互补或其他共同变化，不能仅据此说停投创造了访问。S 接近分母为零时会非常不稳定，即使总流量效应估计相对精确，补回比例的区间也可能很宽。

一个构造的随机市场对照差：与控制市场相比，处理政策使付费访问减少 1,000 次、自然访问增加 600 次、直接访问增加 200 次，则 $S=(600+200)/1{,}000=80\%$，总访问变化为−200 次。这里每项都需在同一观察窗和同一合格市场框内按原始分派估计。若只使用平台归因报表，品牌词曝光后自然点击的变化可能被重新归类或遗漏直接访问；若处理前后还发生搜索算法更新，单纯前后差也不能归因给广告停投。该比率适合描述渠道流量如何重新分布，但商业决策仍需订单与净贡献端点。

SVSAC | sv-sac.org

不确定性也不能忽略：S 是两个估计效应的比值，τP 随机接近零时正态近似的常规区间可能失真。报告可单列三个流量分量和 τT，再使用与随机化相容的比率推断或敏感性区间；若付费流量下降不够明确，应直接说明替代率不可稳定估计。相比一个看似精准的百分比，这种处理更诚实地显示数据实际识别到的是流量方向还是流量比例。

39. 证据综合、替代解释与外推边界

39.1 不同层次证据组成的是链条，不是一个大实验

本次分析引用的证据包括：点击和归因的测量问题、展示广告与搜索广告的现场实验、广告效果预测、品牌资产概念框架、生成式创意摘要级现场证据、历史标题选择实验，以及可读 eBay 品牌词实验；未核 Edmunds 不计已证案例。它们分别观察不同处理、单位、时期和结果。把这些研究串在一起，可以提出“预测–分派–曝光–行为–净贡献–长期品牌”链条中的待检验节点；但跨论文并置不能证明整条因果链已经验证。特别是一个研究的点击增量不能填补另一个研究的利润结果缺口。本篇对设计和端点的区分说明，流量证据不能直接升级成利润主张。[10]

39.2 替代解释应与 AI 解释竞争

即使采用 AI 后某些渠道指标改善，也可能来自内容模板更标准、管理者更频繁查看表现、投放预算改变、受众选择更加集中或活动期需求变强。若 AI 组同时获得新的工作流程、培训或额外媒体预算，随机化识别的是整个处理包，无法单独归因于模型。分臂设计要使“模型建议”“工作流程标准化”“预算变化”尽可能可区分；如果实践中无法拆分，报告需把效果归于可复制的处理包，而不称为算法本身的效果。

反证条件也应预先具体化：独立确认市场不复现、点击提高而贡献不变、品牌记忆提高但偏好或购买不变、利润依赖少数未成熟高额订单、或控制组自然流量补偿后全渠道差异接近零，均会收窄结论。反证不意味着方法无用，它可能指出当前的效用只在特定人群、渠道或预算范围内。

39.3 时间和市场的转移限制

历史研究中的搜索排序、竞价制度、平台覆盖、消费者隐私设置和归因能力都可能变化。跨国家部署还面临语言、文化、广告监管、货币、支付和物流差异。即使随机化内部有效，也不能不加条件地从美国 DMA 迁移到其他国家，或

SVSAC | sv-sac.org

从 2012 年的搜索广告迁移到当代个性化推荐。[10]迁移计划至少应明确目标总体、基线变量、处理版本、竞争状态和数据覆盖，并在目标环境中保留前瞻性验证。

衡量品牌资产也有时间和文化含义：回忆、联想、信任、购买偏好与销量并不等价。跨语种量表的词义和响应方式可能不同；若量表不具备基本测量可比性，潜在均值的差异可能是测量结构变化，而不是品牌效果差异。短期调查不能代替持续观察复购、自然搜索、价格溢价或留存，但这些长期结果也会被新品、渠道促销和竞争动作共同影响。

39.4 信息价值与投放决策的可逆性

研究结果未必只有“扩大”或“停止”两个选项。某些 AI 政策可以按市场逐步开启，另一些会要求一次性整合数据、素材和代理商合同。若结果区间跨过零，决策应同时考虑可逆性、追加试验成本和扩大后被锁定的支出。统计不确定性本身不能替决策者选择风险偏好，但可帮助说明哪些政策在合理的不确定范围内仍有价值。

一个纯构造的两状态例子：假设企业拟扩大到 10 万个合格机会，固定部署成本为 8,000 单位。管理者基于先验判断认为，效果良好状态的概率为 0.4，每个机会净增量贡献 0.40；效果不佳状态概率为 0.6，每个机会损失 0.20。立即扩张的期望净值为 0.4×(100,000×0.40−8,000)+0.6×(−100,000×0.20−8,000)=−4,000。立即停止的增量净值定义为零。若假设能够免费且完美识别状态，先识别再决定的期望净值为 0.4×32,000+0.6×0=12,800，所以“完美信息价值上限”是 12,800；它不是追加样本真实能达到的价值。真实复验既有成本，也可能无法完全区分状态，因此要估计其信息价值，需要说明新试验结果与状态的关系，而不能只把当前区间宽度说成潜在利润。

这一区分也显示了一个常见错误：如果立即扩张是不可逆、且固定成本在试验后无法收回，企业对负收益尾部的厌恶会改变可接受门槛；若政策可以逐市场试行并及时撤销，分阶段部署提供了期权价值。撤销成本、预算承诺、客户体验和长期品牌损害都应进入情景边界。研究报告可以给出不同损失权重下的行动阈值，让管理层明确价值判断来自哪里，而不是暗中将单一统计显著性等同于商业许可。

39.5 构造的组成效应分解：同一分层结果如何产生不同总量

为说明企业间总量差异不能自动归因为“品牌规模”，考虑一个明确构造的两类查询模型。导航意图强的查询在停投后有 98%的访问被自然或直接路径保留；较易被竞品争夺的查询保留 50%。假设这两种类内保留率在两个市场完全相同。

SVSAC | sv-sac.org

市场甲的导航类占 90%，其总体保留率是 0.90×0.98+0.10×0.50=93.2%；市场乙导航类占 20%，总体保留率为 0.20×0.98+0.80×0.50=59.6%。仅仅由于查询构成不同，两市场总量相差 33.6 个百分点，即使没有任何类内政策效应差异。

以上构造并未解释真实公司之间的差异。本篇没有经统一核验的查询分类、分层率与目标权重，不能完成跨研究标准化。[10]它说明要识别公司层面差异，至少须有共同的市场/查询分层变量和重叠范围，并按事先选定的目标权重重新计算。反之，若同一层内仍存在不同效应，剩余差异也可能来自自然排序、竞品广告或执行强度。没有共同支持的层不能靠加权公式“补出来”，而要单列为外推部分或给出边界。

在新研究中，可先估计各预定义意图层的随机化政策效应，再用目标总体的层权重汇总；同时报告原样本权重与统一权重下的结果。若汇总方向对合理权重变化很敏感，决策结论就依赖目标市场组成，而不应写成无条件的品牌词规律。分层权重必须取自处理前可定义的目标总体，不能根据实验后表现挑出更有利的市场权重。

39.6 不完美诊断试验的信息价值：从“可能有效”到是否部署

前文 39.4 节的完美信息价值是上限，实际复验通常会有假阳性和假阴性。继续使用 39.4 节明确构造的两种真实状态：良好状态概率 0.4 时规模化净值 +32,000，较差状态概率 0.6 时净值−28,000；立即扩张的期望净值为−4,000，当前最佳选择是停止，增量净值为 0。现在假设一项额外诊断试验的灵敏度和特异度都为 0.8。这两个数是决策者为演示而设定的，不是观测数据，也不是本次分析的试验效果。

良好状态且出现阳性信号的联合概率为 0.4×0.8=0.32，假阳性概率为 0.6×0.2=0.12，故 P(+)=0.44；由贝叶斯公式，P(良好状态|+)=0.32/0.44≈0.727。部署的状态概率阈值来自 p×32,000+(1−p)×(−28,000)>0，即 p>28,000/60,000≈0.467。因此阳性后的后验 0.727 高于部署阈值；阳性后的条件期望部署净值约 15,636。阴性时 P(良好|−)=0.08/0.56≈0.143，低于部署概率阈值 0.467；条件期望部署净值约−19,429，小于零。因此，按风险中性准则只在阳性后部署。

若诊断免费且不改变未来机会、市场状态、固定成本或预算，按信号决策的期望增量价值为 0.32×32,000+0.12×(−28,000)=6,880；其中后项是阳性误报后仍部署的损失。若试验现金成本为 3,000 且没有其他延迟或机会成本，净信息价值为 3,880，高于当前停止的 0，也低于完美信息上限 12,800。若信号与状态无信息，

SVSAC | sv-sac.org

例如阳性和阴性各半且均维持原先 0.4 良好概率，则两个信号下部署的期望仍是 −4,000，理性决策都停止，信息价值为 0；支付试验费就会降低期望净值。

这个计算说明复验价值不是由当前区间宽度单独决定，而取决于它是否有机会改变行动、信号区分状态的能力及其成本。它也依赖强假设：试验不消耗或改变未来可部署机会，真实状态在试验期间不漂移，固定部署成本不变，且每个状态内的贡献可以按该方案规模线性汇总。若等待导致市场窗口消失、账户被竞争者触达、试验组挤占广告预算，或市场效果存在未建模异质性，应把这些延迟、干扰和机会成本纳入新的决策树。灵敏度/特异度只是构造概率，不是置信区间或 p 值；不能从一个估计区间的宽度反推诊断试验的 EVSI。

40. 行动优先级与决策建议

40.1 最有价值的下一步研究

首先，复制不同企业与平台上的品牌词干预，但保持定义一致：记录竞品是否仍能竞价、自然排序、总点击与付费点击、直接访问、订单和贡献，并公开分母。其次，在 AI 策略评估中同步随机化模型/流程与广告政策，避免只看模型预测或点击率。第三，扩展到预算饱和度和跨市场干扰，测量投入扩大后的拍卖与客户池变化。第四，将品牌记忆和商业结果按同一预先注册样本持续跟踪，并独立确认长期效果。每一步都应预设停止规则、成熟期和对多重结果的控制。

40.2 对题目的回答

AI 可以帮助营销团队更快生成和筛选候选、识别可能响应的人群或管理预算，但这些能力只有在改变可执行的触达政策，并改善按原始合格总体计算的净结果时，才构成营销增量。观测到的点击、归因订单、预测分数、品牌态度或流量恢复，都各自回答有限问题；它们不能单独证明 AI 带来了新客户、利润或长期品牌价值。

因此，从流量到品牌不是一个可以由单一模型分数跨越的漏斗。可靠判断需要把具体 AI 处理、可识别的对照、竞价和渠道干扰、结果成熟、全成本、品牌测量以及外推目标连在同一个研究设计中。若证据支持的仅是某平台、某受众和某窗口内的流量或参与度变化，结论就应停留在该层；只有当订单、贡献和品牌结果在独立且足够成熟的样本中得到相应证据，才可提出更强的商业主张。

SVSAC | sv-sac.org

40.3 有限的产业启示

对企业而言，优先级不是先选“最先进”的营销模型，而是先决定哪一条可改变的政策值得评估：停投品牌词、调整非品牌预算、选择素材、选择受众，还是改进频次。把基准政策、结果窗口、增量成本和可接受的不确定性写明，再以小范围可逆试验获得与决策相关的证据。无法形成对照时，可以做描述性诊断，但不应把相关性结果包装成因果 ROI。

最终结论保持条件性：AI 营销价值取决于任务与总体定义、模型是否改变真实触达、市场反应和成本结构、结果是否成熟，以及证据能否迁移。我们提出的是一套可被反驳的评估逻辑，不是对任何平台、企业或 AI 系统效果的保证。

SVSAC - sv-sac.org

SVSAC | sv-sac.org

给企业的下一步建议

精准营销应对经营结果负责，企业需要连接流量、客户价值与品牌积累，而不是被局部转化指标牵着走。因此，我们建议把下一步工作收敛为一项有明确交付物的专项任务。

具体起点是：统一客户与订单口径，区分新增需求和渠道转移，分阶段验证投放、内容与复购运营的增量贡献。在此基础上，管理层可以按证据充分程度、资源投入与业务影响确定优先级，先验证关键假设，再安排后续合作与投入。

要把上述判断用于实际经营，可以围绕具体产品或项目、目标市场、现有能力和决策难点，进一步形成营销链路诊断、客户价值分层、投入评估与增长试点方案。这类工作需要把行业研究、企业诊断与合作条件核验衔接起来，让分析结果成为管理层可以使用的决策依据。

资料与分析范围

下列说明列明资料范围、分析依据与适用条件，便于读者辨别文中的事实、推演和建议。市场进入、交易安排及适用规则仍需结合具体项目核实。

企业常把曝光、点击、访问、转化、客户价值和品牌变化放进同一张营销看板，并把能预测这些指标的模型称作“精准营销”。但预测谁会购买不等于知道广告造成了多少购买；广告带来增量购买也不等于产生了正的净贡献；短期转化更不能自动代表品牌资产或长期客户价值。本篇研究 AI 辅助的投放与广告测量在何种条件下能改善预算决策，重点区分按分配政策的意向处理效果、实际曝光者局部效果、campaign 层级因果效果预测、利润贡献和延迟品牌结果。本篇采用定向证据综述与方法分析，核读广告回报实验测量，以及以随机实验为标签预测其他 campaign 增量效果的公开研究版本。证据显示，定向曝光下的观察性对比会受到购买倾向选择影响，零售广告实验的个体销售噪声可能远大于待识别效应；改进实验估计效率有价值，但仍须保留合适的目标总体和反事实。PIE 一类方法展示了如何用实验标签训练 campaign 级效果预测，同时也要求区分训练标签的因果识别与新 campaign 预测的可迁移性。本篇据此提出一个可检验的分析框架，将流量、增量购买、毛利、投放和服务成本、品牌延迟结果分开核算；所有尚未执行的试验设计与数字例均标作作者方法或构造。本篇没有企业投放数据或真实 ROI 观测，不作 AI 营销效果保证。

SVSAC | sv-sac.org

参考文献

[1] Randall A. Lewis; Justin M. Rao. The Unfavorable Economics of Measuring the Returns to Advertising. The Quarterly Journal of Economics, 130(4), 1941–1973 (2015). https://gwern.net/doc/economics/advertising/2015-lewis.pdf

[2] Garrett A. Johnson; Randall A. Lewis; Elmar I. Nubbemeyer. Ghost Ads: Improving the Economics of Measuring Ad Effectiveness. NBER Economics of Digitization conference author manuscript, February 18, 2016 (2016). https://conference.nber.org/confer/2016/EoDs16/Johnson_Lewis_Nubbemeyer.pdf

[3] Brett R. Gordon; Robert Moakler; Florian Zettelmeyer. Predicted Incrementality by Experimentation for Ad Measurement. Management Science; published online October 8, 2026 (2026). https://arxiv.org/html/2304.06828

[4] Garrett A. Johnson; Randall A. Lewis; Elmar I. Nubbemeyer. The Online Display Ad Effectiveness Funnel & Carryover: Lessons from 432 Field Experiments. Working paper; SSRN abstract 2701578 (not a journal-final article) (2017). https://marketing.wharton.upenn.edu/wp-content/uploads/2017/08/Johnson-Garrett-PAPER-VERSION-2.pdf

[5] Sarah Moshary. Does Sponsored Search Advertising Augment Organic Search? Evidence from an E-Commerce Platform. Management Science 71(11), 9687–9709; published online 2025-03-28 (2025). https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2024.04792?journalCode=mnsc

[6] Peter M. Aronow; Cyrus Samii. Estimating Average Causal Effects under General Interference, with Application to a Social Network Experiment. The Annals of Applied Statistics, 11(4), 1912–1947 (2017). https://arxiv.org/pdf/1305.6156

[7] Kevin Lane Keller. Conceptualizing, Measuring, and Managing Customer-Based Brand Equity. Journal of Marketing, 57(1), 1–22 (1993). https://people.duke.edu/~moorman/Marketing-Strategy-Seminar-2015/Session%203/Keller.pdf

[8] Anuj Kapoor; Madhav Kumar. Frontiers: Generative AI and Personalized Video Advertisements. Marketing Science, 44(4), 733–747; published online 2025-05-08 (2025). https://pubsonline.informs.org/doi/abs/10.1287/mksc.2023.0494

[9] Zikun Ye; Hema Yoganarasimhan; Yufeng Zheng. LOLA: LLM-Assisted Online Learning Algorithm for Content Experiments. Marketing Science, 44(5), 995–1016; published online 2025-03-24 (2025). https://faculty.washington.edu/hemay/LOLA.pdf

[10] Thomas Blake; Chris Nosko; Steven Tadelis. Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. Econometrica 83(1), 155–174 (2015). https://faculty.haas.berkeley.edu/stadelis/BNT_ECMA_rev.pdf

[11] Lorenzo Coviello; Uri Gneezy; Lorenz Goette. A Large-Scale Field Experiment to Evaluate the Effectiveness of Paid Search Advertising. CESifo Working Paper No. 6684, September 2017 (2017). https://www.ifo.de/DocDL/cesifo1_wp6684.pdf

SVSAC | sv-sac.org