实验加速器：基于内容感知排序的 A/B 测试可解释洞察与创意建议

Zhengmian Hu*Adobe Research 圣何塞，加利福尼亚州，美国
zhengmianh@adobe.com

Lei Shi*Adobe Research 圣何塞，加利福尼亚州，美国
leis@adobe.com

Ritwik Sinha
Adobe Research 圣何塞，加利福尼亚州，美国
risinha@adobe.com

Justin Grover Adobe Digital Experience 圣何塞，加利福尼亚州，美国
jgrover@adobe.com

David Arbour
Adobe Research 圣何塞，加利福尼亚州，美国
arbour@adobe.com

arXiv:2602.13852v1 [cs.AI] 14 Feb 2026

摘要

现代在线实验面临两大瓶颈：流量稀缺，迫使团队艰难抉择要测试哪些变体；而事后提取洞察的过程依赖人工、缺乏一致性，且往往不考虑内容。与此同时，各组织未能充分利用历史 A/B 测试结果和丰富的内容嵌入，而这些信息本可用于指导优先级排序和创意迭代。我们提出了一个统一框架，旨在：(i) 确定优先测试哪些变体；(ii) 解释胜出方案为何胜出；(iii) 发掘有针对性的新机会，以开发潜力更高的变体。借助处理方案嵌入和历史结果，我们训练了一个点击率排序模型，并针对上下文变化纳入固定效应，在对候选方案进行评分的同时兼顾价值与内容多样性。为提高可解释性和理解度，我们将处理方案投影到经过整理的语义营销属性上，并通过符号一致的稀疏约束 Lasso，在这一空间中重新表达排序器，从而得出各属性系数和带符号贡献，用于生成可视化解释、前 k 个驱动因素及自然语言洞察。随后，我们计算机会指数，将属性重要性（由排序器得出）与当前实验中属性表达不足的程度相结合，以标示缺失的高影响力属性。最后，大语言模型将排序后的机会转化为具体的创意建议，并估算学习潜力和转化潜力，从而加快测试周期，使其信息量更高、效率更好。这些组件已集成到 Adobe 的一款真实产品中，即**实验加速器**，为客户提供基于人工智能的洞察和扩展实验规模的机会。我们还评估了该框架在 Adobe 企业客户开展的一些真实世界实验中的表现，验证了生成流程的高质量。

CCS 概念

•计算方法 → 自然语言生成;**主成分分析**; **• 以人为中心的计算 → 用户研究**; **• 计算理论 → 结构化预测**; **• 信息系统 → 排序学习**; **情感分析**.

关键词

迁移学习；生成式人工智能驱动的实验；学习排序；可解释性

*两位作者对本研究的贡献相同。

1 引言

在线随机实验是现代数字营销的基石：它能够评估产品变更、大规模个性化用户体验，并将客户互动转化为可衡量的参与度和收入提升。它还能对变更效果进行无偏估计和不确定性量化，从而帮助理解因果关系、获得可靠的业务洞见，并做出更明智的决策。

然而如今，随着实验规模不断扩大，使实验保持可靠的机制开始遇到实际限制。在本文中，我们研究了两种在规模扩大时变得更为严峻的持续性约束。首先，待测试的变体数量正在增加，尤其受到生成式人工智能成功的推动。每增加一个变体都会稀释统计功效、提高最小可检测效应并延长测试时长。这迫使人们在固定的用户预算和营销活动时间线下，对上线内容做出艰难选择，使简单的“全部测试”策略变得不可行。其次，两项关键的下游任务通常是临时进行的：**洞察提取** 和**机会生成**。所谓**洞察提取**，是指将原始A/B结果和变体内容转化为业务洞察——例如，识别哪些语义属性（语气、激励、紧迫性）能够解释观察到的表现差距。所谓 **机会生成**，是指系统地提出对潜在变体的改进，并将洞察转化为可执行的创意变更。若人工完成，这两项任务在众多并行测试中成本高昂，且容易受到分析师个人偏见的影响。

借助生成式人工智能和表征学习的最新进展，我们提出了一个统一框架，利用两项高价值资源来应对上述挑战：历史 A/B 测试结果档案（包含效应量和不确定性），以及能够捕捉创意语义结构的丰富内容嵌入。具体而言，我们的方法能够（i）优先确定要测试的变体，（ii）解释某些变体胜出的原因，以及（iii）揭示可用于开发潜力更高的新变体的定向机会。该框架分为三个相互衔接的阶段。**排序**。 我们基于历史处理方案及其结果的嵌入训练点击率（CTR）排序模型。在推理时，该模型能够预测新实验中候选变体的相对排名。**洞察**。 为提高可解释性并加深理解，我们将处理方案投影到一组经过筛选的语义营销属性上，以获得属性得分，然后在该属性空间中重新表达排序器。由此得到的各属性系数能够量化每个属性对预测点击率的影响

差距。该方法还能生成数据驱动的自然语言洞察。**机会**。 我们构建了一个机会指数，将属性重要性（来自排序器）与当前实验中的属性表达不足相结合，以检测缺失且影响力较大的属性；大语言模型（LLM）会将这些机会按优先级转化为具体的创意建议，并评估其学习潜力和转化潜力。

我们的贡献主要有三点：

(1) 一种排序模型，通过嵌入迁移历史实验中的知识，从而预测候选变体的相对排名。这一步利用高质量 A/B 测试数据集中的信息，并为后续任务奠定了坚实且统计严谨的基础。(2) 一种洞察提取模块，通过约束 Lasso 将排序模型的模型权重投影到语义属性空间，生成带符号的逐属性贡献，以及标准化的洞察产物（可视化内容、前-k 驱动因素、自然语言理由）。(3) 一种机会生成模块，基于有原则的机会指数，找出缺失的高影响力属性，并将其转化为由大语言模型生成的创意改进建议。

这些组件已在生产系统实验加速器中实现，目前已面向 Adobe 客户上线[1]。我们在真实世界实验中评估了该框架，结果表明，这条流水线能够提供令人满意且忠实于关键营销属性的解释，以及可执行的机会，从而加快测试周期，使其更具信息量且更高效。

2 相关工作

内容感知排序。 相关工作包括以下内容 [2,3], ，这些工作探讨了利用处理嵌入学习预测模型，以预测多个变体的结果或排名。这些工作的一项关键假设是，存在一个大型实验数据集，可直接用于预测处理方案的相对排名，而评估通常基于随机留出样本。在现实中，许多品牌都处于数据稀疏环境，需要从不同领域已完成的实验中学习，这引出了一个问题：这些学习成果能否迁移应用。

点击率预测与反应建模。 广告的点击率预测主要关注预测准确性 [8, 11, 14, 15]。社交媒体帖子的反应建模也旨在从过往帖子中学习社群反应，并在发布前提供预测模型 [12]。相比之下，我们的方法不仅预测处理方案的排名，还通过营销属性和排序模型系数提供可解释的洞察与机会。这也有助于为未来的实验确定测试优先级并推动创意改进。

***LLM* 作为评判者。** 提示大语言模型或使用 LLM-as-a-judge 进行文本生成评估也已有相关探索，例如 [5]。这种方法不会将处理方案与真实世界 A/B 测试数据集中的结果数据相关联。相比之下，我们的方法从针对特定实验任务训练的定量模型中提取洞见。

内容优化与创意生成。 相关研究还包括内容优化与创意生成 [1, 6, 7]。 [6] 通过语言模型微调进行内容优化。 [1]开发了一个排序模型，以及一个基于深度学习的创意生成模型。 [7] 应用贝叶斯优化来最大化处理组的奖励。然而，这些研究通常直接在嵌入空间中运作，而不是在营销属性层面进行，因此难以解释哪些属性对历史实验最为重要，以及哪些属性需要改进。我们的方法以更具可解释性的方式实现了这一目标。

3 语义排序模型

本节将介绍实验加速器骨干排序模型核心流程的技术细节。我们要探讨的核心问题是：给定某项实验中的一组变体，我们能否根据它们的预期结果预测其最终排名？这需要基于一些历史实验数据构建模型，以探索处理组背后的语义信息，并将这些知识迁移到新实验中。在我们的方法中，我们专注于根据点击率对处理组进行排序：

$$\text{CTR} = \frac{\text{Number of Clicks}}{\text{Number of Impressions}} \times 100\%.$$

3.1 训练数据和测试数据的选择

首先要考虑的问题是，构建此类排序模型应使用哪种训练数据，以及使用哪些数据来评估模型。需要考虑多个方面：

首先，我们需要的是数字营销场景中的数据集，而不是用于其他目的的通用文本数据集。确保数据集与我们的业务目标高度契合至关重要。不同的应用场景可能会收集截然不同的文本材料，即使是相同的文本材料，在不同情境下也可能有不同的解读。因此，许多用于通用文本分析应用（如分类或语义聚类）的开放基准数据集并不适用。

其次，我们需要使用与实验形式高度一致、且目标结果相似的 A/B 测试数据集。A/B 测试数据的重要性，体现在其严谨的随机化上，这能确保结果预测中不存在未测量的混杂因素。使用这类数据集，可以避免模型受到与文本语义信息无关的因素所引致的隐藏偏差影响。此外，针对相似的结果进行建模，也能确保所学知识可以合理地迁移和解释。

第三，我们需要利用公共数据集进行训练，并使用直接获取的客户数据进行测试。我们的目标是构建一个平台，为其他公司提供企业对企业（B2B）服务，帮助它们更好地开展 A/B 测试。尽管我们可以收集许多客户的数据，但由于客户数据使用方面的法律限制，这些数据不能用作训练数据。不过，这些数据集将是理想的测试来源，因为评估将直接面向我们的客户。这将提供更有力的证据

[1]请参阅官方网站 https://business.adobe.com/products/journey-optimizer/experimentation-accelerator.html

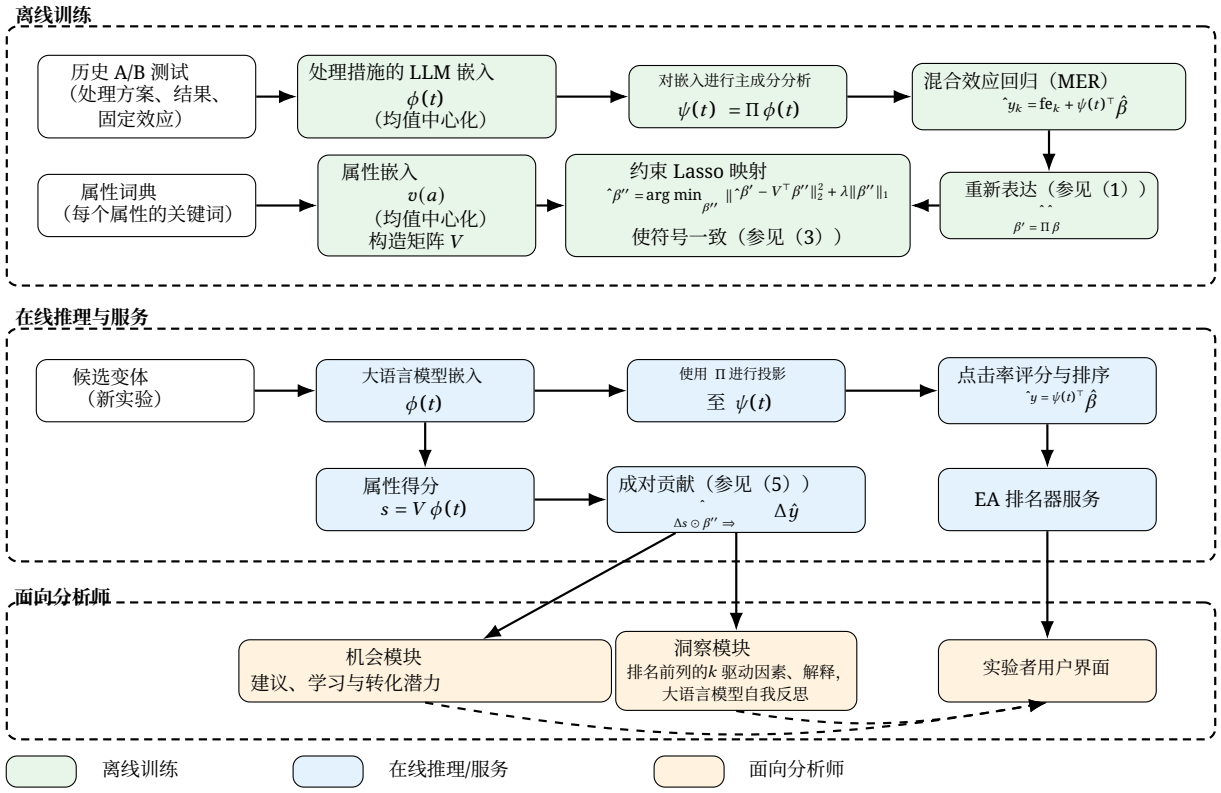

图 1：实验加速器的总体架构：（绿色）离线训练学习主成分分析投影 Π、MER$\hat{}$ $\hat{}$$\hat{}$ 权重 β 和属性空间系数 β''；（蓝色）在线推理复用 (Π, β, V, β'') 来为候选项打分、进行属性评分并计算成对贡献；（橙色）面向分析师的模块将定量归因转化为洞见和机会建议。模型的重新表达见(1)-(3)。

，比基于训练/测试划分的评估等其他方法更有说服力，并能验证迁移学习的质量。

考虑到这些问题，我们使用 Upworthy 数据集 [9] 作为训练数据。该开放数据集包含 Upworthy 于 2013 年 1 月至 2015年4月开展的数千项标题 A/B 测试。数据集包含标题的点击率，与我们的目标结果完全一致。测试数据包括从客户处收集的 65 项实验。

3.2 排序模型构建

本节将介绍完整的排序模型构建流程。设 K_{train} 和 K_{test} 分别表示训练数据和测试数据中的实验数量，$t_{k,i}$ 表示 k个实验中的第 i个处理。该流程总结于算法 1 中。

现在，我们对算法 1 进行一些讨论。

对于步骤 1，需要做出的一个关键选择是采用哪种嵌入方法。处理嵌入可以从许多来源获取。举几个例子，我们考虑上下文变换器（如 BERT、RoBERTa、Llama）、句子变换器（如 SBERT、MiniLM）和 OpenAI 文本嵌入器。我们还可以进一步考虑任务微调模型。在评估部分，我们比较基于不同选择的排序模型的性能

算法 1： 基于文本嵌入的混合效应回归，用于点击率预测

输入： 跨实验的营销文案 $\{t_{k,i}\}$ k；嵌入模型 $\phi(\cdot)$；目标维度 q

输出： 训练得到的系数 β；预测的点击率 $\hat{y}_i$

1 **步骤 1：嵌入生成**。为以下内容生成文本嵌入每条营销文案：

$$\phi(t_{k,\,i}) \leftarrow \text{嵌入}(t_{k,\,i})。$$

2 **步骤 2：降维**。应用主成分分析以获得 q维处理表示：

$\psi(t_{k,i}) = \Pi\phi(t_{k,i})$，其中，Π 是PCA投影矩阵。

3 **步骤 3：模型训练**。对处理表示中的点击率进行回归并加入实验层面固定效应项：

$$\hat{y}_{k,i} = \mathrm{fe}_k + \psi(t_{k,i})^\top \hat{\beta} + \epsilon_{k,i},$$

其中，fe_k 为实验固定效应。

4 **第 4 步：预测**。对于新实验，预测相对点击率：

$$\hat{y}_i = \psi(t_i)^\top \hat{\beta} + \epsilon_i.$$

嵌入，以确定最适合我们场景的选择。

此外，计算处理嵌入时，我们会移除均值分量。例如，假设我们使用基于大语言模型的嵌入。我们收集了一份常见营销材料列表 c_i，共包含 N_c 个样本。最终的处理嵌入计算如下：

$$\phi(t_i) = \text{Embedding}(t_i) - \frac{1}{N_c}\sum_{j=1}^{N_c}\text{Embedding}(c_j)$$

对于步骤 2，主成分分析的性能取决于构建投影矩阵 Π。从技术上讲，Π 可以通过多种数据来源构建：仅使用训练数据、仅使用测试数据，或同时使用训练数据和测试数据。在我们的案例中，我们发现，仅使用测试数据进行训练，实际上能实现最成功的知识迁移。

对于步骤 3，我们基本上采用了主成分回归。重要的是，我们加入了固定效应，以调整训练数据集中实验层面的差异。其原理是，有些实验的点击率通常高于其他实验，因此线性模型需要捕捉这种整体差异。

对于步骤 4，在进行预测时，我们无法再加入固定效应项，因为对于新实验，我们事先并不知道其平均水平。不过，相比点击率的绝对值，我们更关注处理组的相对排名。此处的预测应编码这种顺序信息：$\hat{}y_i$ 与预测点击率呈单调关系，但绝不能将其解读为点击率预测的绝对尺度。

4 语义属性表征

在第 3 节中，我们构建了排序模型。一个重要但尚未探讨的问题是：如何解释该模型的权重？本节通过将学习得到的模型权重与一组可解释属性联系起来，来回答这个问题。

4.1 属性集与嵌入

为了构建更有原则的解释流水线，我们收集了一组包含 122 个营销属性的集合，这些属性反映了营销场景的语义含义。这些属性经与产品经理、营销高管和客户讨论后最终确定。表 1 列出了一些属性示例。在生产环境中，客户可以根据自身业务需求动态更新这组属性。

为了基于属性进行计算分析，我们为这些属性构建嵌入，并将属性集编码为数值词典。考虑 m 个属性 $\mathcal{A} = \{a_1, \ldots\ldots, a_m\}$。对于每个属性，我们都有一组代表性短语（示例见表 1）$\{u_{m,i}\}_{i=1}^{N_{a_m}}$，属性 a_m 的短语总数为 N_{a_m}。属性嵌入的计算方式如下：

$$v(a_m) = \frac{1}{N_{a_m}}\sum_{i=1}^{N_{a_m}}\text{Embedding}(u_{m,i}) - \frac{1}{\sum_{m=1}^{M} N_{a_m}}\sum_{m=1}^{M}\sum_{i=1}^{N_{a_m}}\text{Embedding}(u_{m,i}).$$

值得注意的是，我们在此采用了与第 3 节处理嵌入生成步骤类似的去均值方法。最终得到的嵌入集合保存为矩阵 V，我们称之为属性嵌入字典。

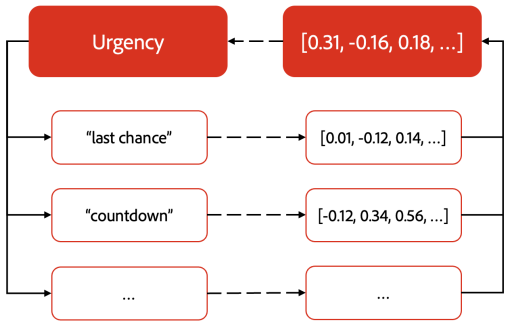

图 2：属性表征

4.2 属性得分

属性得分表示处理中的营销属性水平。处理 $t_{k,i,a}$ 在属性 a 方面的属性得分如下：

$$s_{k,i,a} = \phi(t_{k,i})^\top v(a), \quad a \in \mathcal{A}.$$

其中，$v(a)$ 是作为属性 a 嵌入的 p 维向量。用更简洁的形式表示：

$$s_{k,i} = V\phi(t_{k,i}),$$

其中$s_{k,i}$是处理 $t_{k,i}$ 的属性得分向量。

4.3 影响系数

属性得分可以与排序模型权重结合，以定量衡量每个属性的重要性。我们将这个指数称为影响系数。下面的算法2总结了该系数的计算流水线。

我们补充一些说明。步骤 4 很重要。对于定义 (2) 中的系数 $\hat{}\beta''$，下面的引理表明，它可以重新表示为一个最小二乘问题的解：

引理 4.1（重新表示 $\hat{}\beta''$）。对于定义 (2) 中的影响系数 $\hat{}\beta''$，有

$$\hat{\beta}'' = \arg\min_{\beta''} \|\hat{\beta}' - V^\top\beta''\|_2^2. \tag{4}$$

这启发我们进一步对最小二乘问题添加正则化，以降低共线性，并施加符号约束[10]，从而确保业务影响一致性：

$$\hat{\beta}'' = \arg\min_{\beta''} \|\hat{\beta}' - V^\top\beta''\|_2^2 + \lambda\|\beta''\|_1, \text{ s.t. } \beta'' \cdot \text{sign}(V^\top\hat{\beta}') \geq 0.$$

$\hat{}$ 该约束保证 β''_a 始终与 $\hat{}(V\beta')_a$ 符号相同（或恰好为零），确保无论新增多少属性，“正/负属性”的解释方向都保持一致。此外，这个带符号约束的 Lasso 是一个凸问题，可以通过现有的 Python 实现求解，例如 sklearn.linear_model.Lasso，并将参数 positive = 设为 TRUE。

5 生产环境中的下游应用

在本节中，我们将介绍上述语义排序模型的若干下游应用，该模型已在生产环境中部署。这些应用主要涉及构建若干

表 1：营销属性示例，包含描述、示例文案和一些代表性短语

属性名称	描述	示例文案	代表性短语
行动_导向	行动 _导向型语言能够激励客户通过直接、具有时效性的语言促使他们立即采取行动，旨在激发互动并提高转化。	“立即行动——点击此处开始今天就开始！”	立即行动、马上采取行动、开始行动今天，.....
对_错失_的_恐惧（错失恐惧）	“错失恐惧”属性运用紧迫性，独家性和病毒式传播吸引力，促使客户立即采取行动，并感受到自己是热门或限量事物的一部分。	“加入当下最热门的潮流，这是今天人人都在谈论的话题！”	错过，潮流，加入，热门，...
…	…	…	…

算法2： 属性空间分解与正则化排序器估计

输入： PCA投影矩阵 Π，属性嵌入字典 V，学习得到的PCA排序器系数 $\hat{}\,\beta$，嵌入 $\phi(t_i)$**输出：** 属性空间系数 $\hat{}\,\beta''$

步骤 1：重新参数化 PCA排序器模型。
通过逆转 PCA 步骤，在原始嵌入空间中计算等效系数：

$$\hat{\beta}' \leftarrow \Pi^\top \hat{\beta}, \qquad \hat{y}_i = \phi(t_i)^\top \hat{\beta}'. \tag{1}$$

步骤 2：将嵌入分解为属性嵌入和干扰项。
将每个 $\phi(t_i)$ 投影到属性空间及其正交补：

$$\begin{aligned}\phi(t_i) &= V^\top (VV^\top)^{-1} V\phi(t_i) + (I - V^\top (VV^\top)^{-1} V)\phi(t_i) \\ &= V^\top (VV^\top)^{-1} s_i + \tilde{\phi}_i.\end{aligned}$$

此后，我们用 $V^+ = V^\top (VV^\top)^{-1}$ 表示 V 的伪逆。

第 3 步：组合表示。
将分解代入排序模型：

$$\hat{y}_i = s_i^\top V^{+\top} \hat{\beta}' + \tilde{\phi}_i^\top \hat{\beta}' \triangleq s_i^\top \hat{\beta}'' + \tilde{\phi}_i^\top \hat{\beta}', \tag{2}$$

其中 $\hat{}\,\beta'' = V^{+\top} \hat{}\,\beta'$。

第 4 步：正则化属性系数估计。
通过符号约束 Lasso 问题估计 $\hat{}\,\beta''$：

$$\begin{aligned}\beta'' = \arg\min_{\beta''} \ & \|\hat{\beta}' - V^\top \beta''\|_2^2 + \lambda \|\beta''\|_1, \\ \text{使得 } & \beta'' \cdot \operatorname{sign}(V^\top \hat{}\,\beta') \geq 0,\end{aligned} \tag{3}$$

其中，λ 控制稀疏性，并通过 K 折交叉验证进行选择。

用于从不同方面衡量实验质量的关键指标。

5.1 洞察生成

如果某项实验已结束并报告了显著的点击率差异，营销人员可能希望更全面地了解和总结为何某些处理组优于其他组。我们将已完成实验的这一目标称为“洞察生成”。我们的解决方案是构建一个实验级指标（我们称之为洞察指数），用来定量衡量某个属性对处理组间差异的贡献程度，以及该贡献的方向。

为了构建该指数，假设两个处理方案 t_i，t_j 的属性得分差异为

$$\Delta_{i,j} s_a = s_{i,a} - s_{j,a}, \quad a \in \mathcal{A},$$

它衡量了实验组 t_i 和 t_j 在属性集方面表现力水平的差异。为了进一步了解这一差异对点击率差异的贡献程度，我们采用算法 1 中的预测步骤 (2)，并将预测的点击率差异表示为：

$$\Delta_{i,j}\hat{y} = (\Delta_{i,j}\vec{s})^\top V^{+\top} \hat{\beta}' + (\Delta_{i,j}\tilde{\phi})^\top \hat{\beta}' \triangleq (\Delta_{i,j}\vec{s})^\top \hat{\beta}'' + (\Delta_{i,j}\tilde{\phi})^\top \hat{\beta}'.$$

从上式可以看出，对于属性 a，实验组 t_i 和 t_j 之间的差异会导致预测点击率产生 $\Delta_{i,\,j} s_a \cdot \hat{}\,\beta''_a$ 的差异。直观而言，该数值考虑了影响系数，这些系数会重新调整$\Delta_{i,\,j} s_a$的重要性。基于这一解释，我们将这个特殊数值定义为属性 a 的洞察指数，并将其记为

$$R_a^{\text{Ins}} = \Delta_{i,j} s_a \cdot \hat{\beta}''_a. \tag{5}$$

R_a^{Ins}的符号现在表示属性 a 对预测相对点击率差异 $\Delta_{i,\,j} \hat{}\,y$的贡献方向。 R_a^{Ins}的大小则表示属性 a贡献的强度。

借助洞察指数，我们可以开展下游分析以提取洞察，例如绘制每个属性 a的 $\Delta_{i,j} s_a \cdot \hat{}\,\beta''_a$分布，或者找出最 k正面和负面的属性，并提示它们对客户的影响。例如，与大语言模型结合后，我们得到方框1中的洞察提取示例输出。

方框1：洞察提取示例

Surprising Statistic 有助于洞察：表现最佳的处理方案，例如“几秒钟内创作惊艳艺术作品并完成复杂编辑”，凭借出人意料的速度和能力吸引注意力并推动互动；排名靠后的处理方案缺乏这类引人注目的说法，因此吸引力较弱，也不那么令人感兴趣。

5.2 机会生成

我们希望根据排序模型为实验提供机会。具体来说，我们要计算每个处理方案在一组营销属性上的属性得分，并将这些得分与排序模型权重结合，为新实验生成机会。与洞察生成类似，关键步骤是构建机会指数。核心思路是找出新实验中正在测试的处理方案所缺失的重要属性。我们构建一个综合指数，用于衡量单个实验中属性的重要性及其表达度。

为衡量重要性，我们取影响系数 β''_a 的排名 R_a^{Imp}：

$$R_a^{\text{Imp}} = \text{Rank}(\beta''_a).$$

R_a^{Imp} 越小，属性 a 的重要性越高。为衡量属性的表达程度，我们取平均属性得分的排名R_a^{Exp}：

$$R_a^{\text{Exp}} = \text{Rank}\left(\max_{i\in\{1,\dots,I\}} s_{i,a}\right).$$

R_a^{Exp} 越小，表示属性 a 在变体中的探索程度越高。现在，我们构建一个机会指数，将这两项指标结合起来：

$$R_a^{\text{Opp}} = R_a^{\text{Imp}} - R_a^{\text{Exp}}.$$

R_a^{Opp} 越小，表示属性 a 越重要，且在当前变体中的体现越少。这启发了一种直接的应用方式：我们可以选取若干个 R_a^{Opp} 最小的属性，并在后续实验迭代中针对这些属性提出改进建议。框 3 给出了一个机会生成示例。

框 2：机会生成示例

统一品牌声音：确保信息传递与品牌形象保持一致。例如：“您日常省钱的可靠伙伴。”

社会认同：通过展示他人的积极体验来建立信任。例如：“超过 10,000 位满意客户都不会错！”

框 3：转化与学习潜力示例

建议：提升统一品牌声音**转化潜力**：中 ⋆历史实验表明，该属性对排名预测的影响中等**学习潜力**：高⋆ 当前和历史实验对该属性的探索都不充分。

5.3 影响度与新颖性量化

在本节中，我们将介绍洞察指数和机会指数。这两项综合指标结合了属性得分和排序模型权重。

根据排序模型，我们指出了影响系数 $\vec{\beta}''$ 的关键作用。我们建议将影响系数划分为若干等级，例如强、中、弱，以便报告所选属性在推动转化方面的影响度量。

基于属性分数，衡量某个属性对客户而言有多新颖是很有用的。换句话说，对于一个属性，我们可以衡量它在当前实验和历史实验中的表达程度。如果两者都较低，则意味着当前和过去的实验都尚未充分探索该属性。为了定量衡量新颖性，令 $\bar{s}_{a,\text{history}}$ 表示属性 a 在一组历史实验中的平均属性分数。我们可以定义以下新颖度指数：

$$\begin{aligned} R_a^{\text{Novel}} &= \text{Rank}(\max_{i\in\{1,\dots,I\}} s_{i,a}) + \text{Rank}(\bar{s}_{a,\text{history}}) \\ &= R_a^{\text{Exp}} + \text{Rank}(\bar{s}_{a,\text{history}}). \end{aligned}$$

较低的新颖度指数表明属性 a 尚未得到充分探索，并且可能蕴含许多机会。

在我们的实验智能体中，我们为机会模块实现了影响度量和新颖度度量，为实验提供了有参考价值的定量检查。

6 使用大语言模型进行文本生成

6.1 洞察

该组件将第 3 至 5.1 节中的定量归因转化为简洁、以文本为依据的解释。它分为三个阶段：属性选择、基于提示的解释和质量筛选。

ˆ属性选择。 根据属性空间（第 5.1 节）中的影响系数 β'' 和各处理方案的属性得分s_i，我们按结果（点击率）对实证表现最佳和最差的处理方案进行对比： $\Delta s = s_{\text{best}} - s_{\text{worst}}$。然后计算逐属性贡献 $c_a = (\Delta s)_a \, \hat{}\beta''_a$。我们应用正值筛选，仅保留贡献严格为正的属性（$c_a > 0$），按 c_a 降序排列其余属性，并选出排名前-k的属性。向大语言模型呈现时，每个属性的ˆ ˆ效应方向取自 $\text{sign}(\beta''_a)$（若 $\beta''_a > 0$，则为“正向”，否则为“负向”）。这遵循公式（2）：带符号的得分差乘以学习得到的系数，可量化最佳处理方案相较于最差处理方案带来的方向性提升。

提示与输出格式。我们将每个实验序列化为一个有序的处理方案列表（按点击率从最佳 → 到最差排列），并附上简洁的文本差异；同时传入所选属性及其效应方向（来自 $\text{sign}(\hat{}\beta''_a)$）。提示采用成熟的提示工程技术，以引导模型给出简洁且以文本为依据的理由：(i) 角色条件设定和明确的任务约束，(ii) 少样本示例，用于展示所需风格，(iii) 结构化输出要求，以限定条目数量并确保句子简短且有证据支撑，以及 (iv) 引用优先的依据约束，要求从所提供的处理方案中引用短语。与自由形式先验相比，这一设计优先考虑确定性、易读性，以及对可观察文本的忠实度，

该模板和约束借助基于大语言模型的提示设计器进行调优，并采用留出集上的接受度信号（第 B 节）。

后处理与自我反思。 我们会进行轻量级规范化（移除辅助标记、合并空白字符），并使用基于规则的解析器生成从属性到解释的映射。为进行质量控制，我们会针对每个属性执行自我反思步骤，提示模型依据简明的评分标准（忠实引用所提供的处理方案、正确归因、用一个简洁段落阐述理由）逐一评判候选内容，并输出符合架构约束的验收决定；仅保留通过验收的候选内容。这种内部自我反思筛选机制不同于 B 节中的外部评估。自我反思提示采用标准的提示工程要素：角色条件设定、明确标准、少样本评分示例和固定格式输出；其约束根据留出验收信号进行调整，因此，当低质量候选内容被拒绝时，最终返回的洞见数量可能少于 k。

6.2 机会

该组件通过识别影响力大但表达不足的属性，并生成简洁且具有说明性的建议，为未来的迭代提出实验级机会。

选择缺失的高影响力属性。 从全局影响系数 β'' 出发，我们仅保留影响严格为正的属性（增强该属性会提高预测点击率）。我们还可以应用绝对影响阈值（相对于最大绝对影响值），以剔除影响可忽略不计的属性。对于当前实验，我们衡量每个属性的探索程度（例如，各处理方案中的最高观测得分），并据此形成两个排名：（i）重要性和（ii）表达度。机会得分是排名之差（重要性减去表达度），我们会选择重要性高但表达度低的属性。必要时，我们会退而选取影响为正且排名最高的属性。此流程具体实现了第 5.2 节中的指数。

提示与输出格式。 大语言模型会接收（i）实验的处理方案文本和（ii）选定的缺失属性。总体而言，提示采用角色条件设定、少量示例以及简单的结构约束，以生成简洁、结构清晰且以所提供文本为依据的建议。

解析与丰富化。我们对结构化输出进行基于规则的解析，以确保格式一致，并剔除格式错误的区块，从而保证质量。我们还为每个机会补充以下信息：(i) 学习潜力，反映相对于实验和历史数据的新颖程度；以及 (ii) 转化潜力，反映所学影响系数的相对大小。两者均以粗略等级报告。

7 上线后评估

7.1 从 Upworthy 数据集迁移

在本节中，我们基于算法 1，评估一组真实的 $N_{\text{test}} = 65$ Adobe 客户实验的迁移学习结果。每个实验包含两个或更多实验组，且均已完成，并确认处理组之间存在统计显著差异。换言之，实验组的排名已通过数据验证。对于实验 i，将变体的排名记为 m as$R_{i,\,m}$，将基于算法 1 得出的预测排名记为 $\hat{}R_{i,\,m}$。我们考虑两项指标：（i）平均斯皮尔曼秩相关系数，其定义为

$$\rho = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} \rho_i, \quad \text{where } \rho_i = \frac{\text{Cov}(R_{i,\cdot}, \hat{R}_{i,\cdot})}{\sqrt{\text{Var}(R_{i,\cdot})\text{Var}(\hat{R}_{i,\cdot})}}.$$

这用于衡量算法 1 的排名预测一致性。(ii) Top-1 准确率，用于检查实验中排名预测完全正确的比例：设 $m_i^\star$和$\hat{}m_i$ 分别为实验 i中的真实最佳变体和估计最佳变体，我们定义

$$\text{Top-1 Accuracy} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} 1(m_i^\star = \hat{m}_i)$$

最终结果见表 2，同时给出了置信区间，该区间基于各项指标在验证实验中的标准差计算得出。与市场参与者随机猜测、并不依赖预测模型做出决策的原始基线相比，迁移学习在总体秩相关和 Top-1 准确率方面均取得了相当显著的结果。

表 2：性能比较。

指标	ρ（秩相关系数）	Top-1 准确率
迁移学习	0.5144 ± 0.1658	0.7021 ± 0.1180
随机猜测基线	0	0.4293

7.2 留一法：嵌入的选择

本节报告了关于构建预测模型时如何选择嵌入的一些数值结果。我们比较了三个嵌入系列：(i) MiniLM [13],，一种紧凑型 Transformer 编码器，广泛用于生成句子级表示；(ii) Llama 嵌入 [4],，从现代大语言模型编码器中提取，以捕捉更丰富、跨度更长的语义；以及 (iii) 属性分数，这是一种 LLM-as-a-judge 方法，用于根据我们引入的属性集直接为各个变体打分，评分范围为 1 至 5。我们报告预测排名与观测到的留出结果之间的斯皮尔曼秩相关，并对留一折的结果取平均，同时给出置信区间；置信区间基于留出实验中各项指标的标准差计算。我们还纳入了随机猜测基线，其预期相关性为零。

表 3 总结了留一法的性能。Llama 模型的平均相关性最高（0.727 ± 0.116），表明其语义几何与结果相关信号之间的对齐程度更强。

表 3：留出数据集上的斯皮尔曼秩相关。

嵌入	留出实验	随机猜测
MiniLM	0.454 ± 0.155	0
Llama	0.727 ± 0.116	0
属性分数	0.576 ± 0.142	0

7.3 洞察生成与评估

本节评估自动生成的有关创意变体和实验结果的见解质量。我们基于一组 Adobe 客户实验，按照所提出的框架生成见解。随后，我们进行人工标注并报告接受率。此外，我们还在附录 B 中报告大语言模型（ GPT-o3 ）的评估结果。

人工评估。 人工评估结果见表 4。我们请经验丰富的营销人员和工程师评估每条呈现的洞见是否结构清晰、准确且合理。根据这一评估方案，GPT-4o 总体上生成了最受认可的洞见（生成的 52 条中有 46 条获认可；88.46%）；LLaMA-70B 的接受率同样很高（88.46%），但生成的洞见总数少得多（生成的 26 条中有 23 条获认可）。GPT-4o-mini 生成的候选内容多得多，但接受率低得多（分别为 42.04% 和 33.33%），这表明尽管产出量更高，其输出的噪声也更多。

表 4：不同语言模型在洞察生成与质量方面的比较，比较指标包括以下总数：(1) 已审查的实验；(2) 包含洞见的实验；(3) 生成的洞见；(4) 高质量洞见。

	GPT-4o	GPT-4o-mini	LLaMA-70B
已审查的实验	65	65	65
包含洞见的实验	37	60	22
生成的洞见	52	157	26
高质量洞见	46	66	23
接受率（%）	88.46	42.04	88.46

7.4 机会生成评估

我们评估不同语言模型自动生成的机会的质量。正文中报告人工接受率，附录 B 中报告大语言模型（GPT-o3）的评估结果。

人工评估。 表 5 总结了人工标注结果。GPT-4o 和 GPT-4o-mini 生成的已接受机会数量最多，且接受率均较高（分别为 86.15% 和 84.10%），这表明每个实验提出三个方案时，覆盖面较广。LLaMA-70B 的人工接受率略低（82.11%）；由于标注预算有限，评估仅涵盖 41 个实验。

8 讨论

实验加速器通过在测试前对变体进行排名、借助简单的属性视图解释成功原因，并将差距转化为具体的创意建议，帮助团队明智地利用稀缺流量。在接下来的步骤中，我们计划将这一方法从文本扩展到图像/视频、音频和页面布局，并扩展到转化或收入等结果。我们希望记录该流水线用于更广泛用例时的潜在局限性：(i) 当从 Upworthy 迁移到客户数据时，如果存在领域偏移，迁移学习是需要重点检查的重要环节。尽管我们在本案例中见证了这一成功，但始终应谨慎验证这种可迁移性，尤其是针对

表 5：不同语言模型在机会生成和质量方面的比较，比较指标包括：(1) 已审查的实验总数，(2) 有机会的实验数，(3) 已生成的机会数，(4) 高质量机会数。

	GPT-4o	GPT-4o-mini	LLaMA-70B
已审查的实验	65	65	41
有机会的实验	65	65	41
已生成的机会	195	195	123
高质量机会	168	164	101
接受率 (%)	86.15	84.10	82.11

新客户。(ii) 当前方法的语言覆盖范围仅限于英语。应谨慎验证其向其他语言的泛化能力。(iii) 嵌入的表达能力对该流水线至关重要，但根据我们的观察，不同文本嵌入器之间的性能差异很大。在实际生产环境中，同样有必要检查多种嵌入模型，以便为特定的目标客户群最终确定最合适的模型。

参考文献

[1] 2020.》. https://www.cs.columbia. edu/~kapil/documents/cikm20creatives.pdf. [2] Count Bayesie. 2023. 用 GPT 替代 A/B测试。博客，网址：https://www.countbayesie.com/blog/2023/3/23/replacing-an-ab-test-with-gpt. [3] Paul Ellickson. 2023. 处理嵌入。http://paulellickson.com/TreatmentEmbeddings.pdf. [4] Aaron Grattafiori 等. 2024. Llama 3 模型家族。arXiv: 2407 21783 [ ] 2407 21783 . cs.AI [5] https://arxiv.org/abs/ 。Jinlan Fu、See-Kiong Ng、Zhengbao Jiang 和 Pengfei Liu. 2023. GPTScore：按需评估。arXiv:2302.04166 [cs.CL] https://arxiv.org/abs/2302. 04166[6] Kevin Lee. 2024. caLM：面向创意优化的内容感知语言模型。https://leekevin.io/caLM.pdf. [7] Benjamin Letham、Brian Karrer、Guilherme Ottoni 和 Eytan Bakshy. 2019. 含噪实验下的约束贝叶斯优化。贝叶斯分析 14, 2 2019 495 519 10 1214 18 1110 ( ), –。doi: 。/ -BA Dugang Liu、[8] Yang Qiao、Xing Tang、Liang Chen、Xiuqiang He 和 Zhong Ming. 2023. 面向多媒体推荐中点击率预测的先验引导准确率—偏差权衡学习。发表于第*31*届*ACM*国际多媒体会议论文集。995–1003.[9] J Nathan Matias、Kevin Munger、Marianne Aubin Le Quere 和 Charles Ebersole. 2021. Upworthy研究档案：美国媒体中32,487项实验的时间序列。*Scientific Data* 8, 1 (2021), 195. [10] Nicolai Meinshausen. 2013. 高维回归的符号约束最小二乘估计。(2013). [11] Lei Shi、David Arbour、Raghavendra Addanki、Ritwik Sinha 和 Avi Feller. 2025. 利用语义相似性开展人工智能生成处理方案的实验。arXiv:2510.21119 [stat.ME] https://arxiv.org/abs/2510.21119[12] Balaji Vasan Srinivasan、Anandhavelu Natarajan、Ritwik Sinha、Vineet Gupta、Shriram Revankar 和 Balaraman Ravindran. 2013. 你的 Facebook 帖子会引发互动吗？发表于首届用户参与度优化研讨会论文集（*CIKM '* *13* 25 28 10 1145 2512875 2512881)。ACM，–。doi: 。/ 。[13] Wenhui Wang、Furu Wei、Li Dong、Hangbo Bao、Nan Yang 和 Ming Zhou. 2020. MINILM：面向预训练Transformer任务无关压缩的深度自注意力蒸馏。发表于第*34*届神经信息处理系统大会论文集（温哥华，BC省，加拿大）*(NIPS '20)*。Curran Associates Inc., Red Hook, NY, USA, 文章485，13页。[14] Zikun Ye、Hema Yoganarasimhan 和 Yufeng Zheng. 2024. LOLA：用于内容实验的大语言模型辅助在线学习算法。arXiv:2406.02611 [cs.LG] https://arxiv.org/abs/2406.02611[15] Guorui Zhou、Xiaoqiang Zhu、Chenru Song、Ying Fan、Han Zhu、Xiao Ma、Yanghui Yan、Junqi Jin、Han Li 和 Kun Gai. 2018. 用于点击率预测的深度兴趣网络。发表于第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集（*KDD '18*）。ACM，1059–1068。doi:10.1145/3219819.3219823

A 引理 4.1 的证明

引理 4.1 的证明。根据定义 2，$\hat{}\beta''$ 表示为：

$$\hat{\beta}'' = V^{+\top}\hat{\beta}',$$

其中 $V^{+\top}$ 是 Penrose-Moore伪逆。检验最小二乘问题 (4)，我们知道一阶最优性条件给出

$$VV^{\top}\hat{\beta}'' = V\hat{\beta}'.$$

这进一步得出

$$\hat{\beta}'' = (VV^{\top})^{-1}V\hat{\beta}' = V^{+\top}\hat{\beta}'.$$

□

B 用于自动评估的 LLM-as-a-judge

我们使用外部LLM评判者来验证洞察解释和机会建议。评判者只能看到序列化处理内容和候选输出，并返回一个二元决策（可接受=1，有问题=0）以及简要理由。

提示和评分标准。 简洁的提示包含一个完整示例和一份精简的评分标准。对于洞察，评审会检查解释是否正确说明了表现差异，并引用处理方案中的短语。对于机会，评审会检查建议是否指出了确实缺失且价值高的属性，解释它为何有帮助，并附上一到两个简短的说明性示例。

模型与解析。 评审使用推理模型“o3”。回复遵循简单的逐项格式，我们使用基于规则的解析器提取分数和理由。

大语言模型评估（*GPT-o3* 评判者）。为大规模人工评审提供补充，我们使用 GPT-o3 作为大语言模型评判者，并采用相同的评分标准。表6汇总了所有实验中多轮洞察生成的 GPT-o3 评判质量的均值和标准差。评判者比人类更严格（整体接受率更低），但相对排序保持一致：GPT-4o 和 LLaMA-70B 的接受率最高；由于生成量更大，GPT-4o 最终获得接受的洞察总数显著更多。较小的模型（GPT-4o-mini、LLaMA-8B）会生成许多候选内容，但只有少数能通过更严格的筛选。

用于机会生成的大语言模型评估（*GPT-o3* 评判者）。为扩大评估规模，我们还使用 GPT-o3 作为自动评判者，并采用相同的评分标准。表7报告了所有实验中多轮机会生成的 GPT-o3 评判质量的均值和标准差。各模型的接受率普遍较高，整体排序也与人工评审基本一致：GPT-4o 和 LLaMA-70B 表现相当，位居前列；GPT-4o-mini 和 LLaMA-8B 的表现则稍逊一筹。请注意，在此配置下，LLaMA-70B 和 LLaMA-8B 生成了 130 个机会（平均每个实验两个），而 GPT-4o 系列模型生成了全部 195 个机会（每个实验三个）；因此，即使接受率相近，最终接受的机会总数也会受到影响。

表6：以 GPT-o3 为评判者的洞见生成评估。

语言模型	实验数量		洞见数量		接受率 (%)
	已审阅	含洞见	已生成	高质量	
GPT-4o	65	45.250 ± 0.986	49.250 ± 2.511	26.250 ± 2.421	53.87 ± 6.70
GPT-4o-mini	65	61.000 ± 0.471	145.250 ± 3.140	24.250 ± 1.590	16.87 ± 1.35
LLaMA-70B	65	23.000 ± 0.000	27.750 ± 0.553	16.500 ± 1.106	59.53 ± 4.34
LLaMA-8B	65	53.750 ± 0.289	93.000 ± 2.550	18.000 ± 2.550	19.75 ± 3.27

表7：使用 GPT-o3 作为评审者的机会生成评估。

语言模型 # 实验		机会数量		接受率（%）
	已审阅	已生成	高质量	
GPT-4o	65	195.000 ± 0.000	158.250 ± 1.658	81.15 ± 0.85
GPT-4o-mini	65	195.000 ± 0.000	147.250 ± 1.443	75.51 ± 0.74
LLaMA-70B	65	130.000 ± 0.000	104.250 ± 3.032	80.19 ± 2.33
LLaMA-8B	65	130.000 ± 0.000	101.500 ± 1.202	78.08 ± 0.92