arXiv:2603.01024v1 [cs.HC] 1 Mar 2026

SimAB：使用角色画像条件化 AI 智能体模拟 A/B 测试以实现快速设计评估

Tim Rieder*†
苏黎世联邦理工学院
苏黎世，瑞士
timrieder@ethz.ch

Marian Schneider*†
苏黎世联邦理工学院
苏黎世，瑞士
smarian@ethz.ch

Mario Truss
Adobe
德国慕尼黑
mtruss@adobe.com

Vitaly Tsaplin
Adobe
瑞士巴塞尔
tsaplin@adobe.com

Alina Rublea
Adobe
巴塞尔，瑞士
rublea@adobe.com

Sinem Dere
Adobe
巴塞尔，瑞士
ndere@adobe.com

Francisco Chicharro Sanz
Adobe
巴塞尔，瑞士
chicharr@adobe.com

Tobias Reiss
Adobe
巴塞尔，瑞士
reiss@adobe.com

Mustafa Doga Dogan
Adobe Research
巴塞尔，瑞士
doga@adobe.com

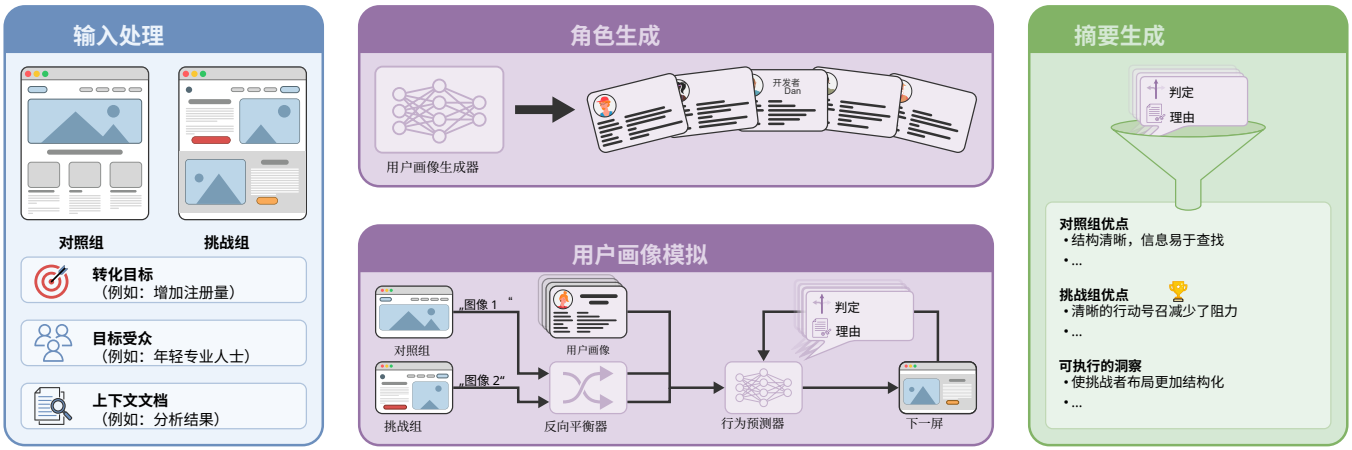

图 1：SimAB 流程将 A/B 测试重新构想为快速模拟。(1) 输入处理：系统接收测试规范，其中包含设计变体、转化目标、目标受众定义以及用于提供背景的文档。(2) 角色生成：系统根据具体背景创建多样化的合成用户群体（例如“营销人员米娅”）。(3) 角色模拟：角色驱动型智能体与设计进行交互。(4) 摘要生成：将个体判定及其理由汇总为可执行的洞察。此工作流程将反馈延迟从数月缩短至数分钟。

摘要

A/B 测试是验证设计决策的标准方法，但它依赖真实用户流量，限制了迭代速度，也使某些实验难以开展。我们提出 SimAB，这是一种利用角色画像条件化 AI 智能体，将 A/B 测试重新构想为快速、保护隐私的模拟的系统。给定网页设计截图和转化目标，SimAB 会生成用户画像，将其部署为表达偏好的智能体，汇总结果并综合理由。通过与实验从业者开展形成性研究，我们发现了一些流量限制妨碍测试的场景，包括低流量页面、多变体比较、微优化以及隐私敏感场景。我们的设计强调速度、早期反馈、可执行的理由以及受众规范。我们使用 47 项结果已知、用于测试网页布局和内容，整体准确率达到 67%，在高置信度案例中提升至 83%。其他实验表明，该方法对命名和位置偏差具有稳健性，并证明了角色设定带来的准确率提升。从业者反馈显示，SimAB 有助于加快评估周期，并快速筛选难以通过传统 A/B 测试评估的设计。

CCS 概念

• 以人为中心的计算 → 用户模型；人机交互；• 计算方法 多智能体系统。

关键词

A/B 测试、合成受众、大语言模型、用户模拟、网页优化、生成式智能体

*贡献相同。作者顺序为随机排列。
†此项工作是在 Adobe Research 实习期间完成的。

1 引言

A/B 测试已成为评估设计决策因果效应的黄金标准。它提供了一种数据驱动的方法，通过同时向真实用户部署原始版本（对照组）和修改后的版本（挑战组），衡量变更带来的影响，并收集这两个变体的用户行为数据。根据收集到的数据，可以依据点击率和年度经常性收入等指标确定表现更好的变体。这种方法常用于 Web 开发，在那里，流量可以轻松分配到不同变体，以量化设计变更对性能指标的影响 [30]。

尽管非常有效，但对真实用户数据的依赖造成了显著瓶颈。要获得具有统计显著性的结果，需要大量用户流量，收集过程往往长达数月；对于流量较低的页面，这一问题尤为严重，它们可能永远无法达到统计显著性。此外，传统测试仅在开发周期结束时验证想法。这意味着，如果测试失败，大量精力会耗费在实际上毫无回报的变体上。这些局限性凸显了对快速评估环境的需求，使测试无需依赖真实用户即可进行。

大语言模型（大语言模型）的最新进展激发了人们使用这些模型模拟人类行为的兴趣[4, 25, 38]。这些合成受众 方法有望在无需等待真实用户的情况下，生成关于设计变体的预测。然而，现有方法要么缺乏严格的验证工作 [10]，要么仅限于评论设计，而不是对其进行验证 [51]。

在这项工作中，我们通过提出 SimAB 来应对这些局限性。SimAB 是一个利用可扩展、经过用户画像条件化且可交互的人工智能智能体生成合成偏好判断的系统，可将达到统计显著性所需的时间缩短至几分钟。SimAB 根据提供的设计变体和测试规范生成多样化的用户画像，将这些用户画像部署为评估设计变体的 AI 智能体，并汇总其判断。我们承认大语言模型存在系统性偏差，例如命名偏差和位置偏差 [71], ，并采用平衡化和中性命名来减轻这些偏差。我们利用一个包含 47 项历史 A/B 测试的定制语料库开展了广泛的验证工作，以评估 SimAB 的准确率，评估用户画像条件化和偏差缓解方法的有效性，并通过与从业者开展反馈会话，验证 SimAB 在现实场景中的有效性。

总而言之，我们的贡献如下：

(1) **一项形成性研究**（第 3 节），通过识别四个痛点，明确了对快速 A/B 测试系统的需求，并据此推导出系统的设计启示。(2) **一个合成 A/B 测试系统**（第 4 节），利用经过用户画像条件化的智能体实现快速 A/B 测试。(3) **全面的实证验证**（第 5 节），表明该系统在 47 项历史 A/B 测试中的准确率为 67%；对于高确定性测试，准确率提升至 >80%。

2 相关工作

本节将讨论为我们的方法提供参考的既有工作，该方法旨在无需依赖大量真实用户流量即可验证和优化设计。我们回顾了有关 A/B 测试和优化平台、基于大语言模型的设计反馈与评估，以及模拟用户的智能体的研究，并将我们的贡献置于这些工作所具备的能力与局限性之中。

2.1 A/B 测试与优化平台

自 A/B 测试在 Web 场景中普及以来，它一直是数据驱动设计优化的基石 [30]。多臂老虎机 [53] 和汤普森采样 [20, 70] 等算法通过动态将用户分配给表现更好的变体，减少了所需流量，但这些优化解决的是效率 问题，而非对真实用户流量的根本依赖。近期研究探索了通过统计框架模拟用户流量 [2] ，以及基于大语言模型的市场研究方法 [6]。然而，这些方法仍需要初始的真实世界观察，或在生成具有外部效度的消费者洞察方面存在局限，因而限制了其适用性和有效性。

2.2 大语言模型用于设计反馈与评估

研究者已探索基于大语言模型的设计反馈和用户界面评估方法 [5, 12, 14–18, 48, 51, 54, 72]。先前研究开发了 Figma 插件等工具，利用 GPT-4 自动对界面模型稿进行启发式评估 [18], 。研究发现，尽管 GPT-4 在评估较差的用户界面设计时表现良好，也有助于发现错误和改进文本，但随着设计不断迭代、质量逐渐提升，其表现会有所下降，因此不太适合专家设计师进行迭代式使用。

研究者已创建数据集，以改进基于大语言模型的自动化用户界面评估；通过少样本和视觉提示技术，针对性数据集的性能提升了 55% [16]。近期研究提出了使用视觉提示和迭代优化的多模态方法，使用户界面评论生成与人类表现之间的差距缩小了 50% [15], ，并将这些进展应用于用户体验评估 [72]。研究者还探索了基于用户画像的方法，用于按需生成反馈 [5, 12, 51] 、支持设计迭代 [54], ，以及根据定性访谈数据提取设计需求 [13]。然而，这些方法本质上模拟的是批评界面的用户体验专家的行为，而非试图达成目标的用户。虽然这些方法有助于发现常见的可用性问题，却无法捕捉真实用户可能具有的复杂且往往不合乎理性的偏好。

2.3 用户模拟智能体

基于大语言模型在设计反馈方面的这些基础应用，研究者开始探索模拟真实用户及其行为。先前研究表明，大语言模型智能体能够浏览真实网站 [24] ，并为自主网页智能体建立了基准 [33, 73]。研究者提出了基于智能体的 A/B 测试系统 [60] 和可用性测试系统 [34], ，不过，这些系统与真实人类对设计干预所产生的行为反应之间的吻合程度仍未得到充分研究。

利用人工智能模拟人类行为已日益受到关注。早期研究提出了生成式智能体，通过记忆检索和行动规划来模拟可信的人类行为 [38], ；后续研究则展示了对真实个体进行生成式智能体模拟，能够以很高的准确率复现调查问卷回答 [40],，以及可提升模拟准确率的微调方法

[31], , 还有推动集体行为研究的大规模模拟 [42]。早期研究探索了为社交计算系统创建填充式原型 [39] , 以及生成用于人机交互（HCI）研究的合成人类数据 [27]。研究人员探索了将大语言模型用作人类参与者的代理，并提出了“算法保真度” [4] 和“硅样本” [25, 47], 等术语，调查了用户建模方法 [56], , 并引入了“合成用户”用户画像 [22]。在用户画像的基础性研究之上 [44], , 早期研究探讨了人们如何在大语言模型中定制和使用智能体角色设定 [23]。数字孪生范式已成为一种强有力的延伸方向，相关研究引入了大规模数据集 [57] , 并在多项研究中评估数字孪生，发现与仅依据人口统计特征构建的用户画像相比，详细的个人信息能提升相关性 [41]。

先前研究已展示了基于大语言模型的智能体如何模拟推荐系统中的复杂多轮用户行为 [11,61, 68, 69],, 涉及电子商务应用，包括购物行为数据集 [63], 、角色设定对齐框架 [35], 、与人类参与者的比较 [55], , 以及用于个性化的强化学习方法 [64]。在人物画像开发与评估方面，先前研究系统性地回顾了生成式人工智能方法 [3], , 提出了生成与人口分布相符的用户画像的方法 [26], , 推出了用于评估用户画像一致性的基准 [45], , 开发了用于角色扮演的可训练智能体 [49], , 创建了评估数据集 [7],, 研究了用于协作式角色设定生成的人机协作工作流 [52], , 并通过多轮强化学习解决一致性问题 [1]。

然而，合成数据能否有效用于理解人类偏好，仍存在重大疑虑。先前研究表明，大语言模型的输出反映了特定人口群体的观点 [46], , 发现其在捕捉细微偏好方面存在系统性缺陷 [74], , 量化了人类与智能体之间的行为差异 [62], , 发现了基于生成式人工智能的可用性测试的局限性 [43], , 并指出了验证模拟行为所面临的挑战 [9]。系统性综述对在人机交互研究中使用大语言模型提出了效度与可复现性方面的疑虑 [37],, 研究还表明，以大语言模型取代人类参与者可能会对身份群体造成有害的失真呈现和扁平化 [59]。基于这些考量，我们对所提出的系统进行了全面评估（第 5 节），以确保其能够提供可靠的输出。

总之，现有方法要么依赖真实用户流量（传统 A/B 测试），要么只能提供开放式定性评述，而非验证。SimAB 的不同之处在于，它借助结构化聚合框架运用合成受众，从而能够直接对设计进行成对比较。与以往基于智能体的探索不同，我们不再停留于概念验证模拟，而是融入系统性的偏差缓解，并首次针对包含已知真实结果的大型历史 A/B 测试语料库进行了广泛的实证验证。

3 形成性研究

我们通过一项形成性研究奠定了 SimAB 的设计基础，该研究对与实验团队持续合作中总结出的从业者挑战进行了定性综合。

3.1 对从业者的信息性访谈

基于与四家企业组织中14名从业者开展的参与式设计合作，我们确定了塑造 SimAB 设计的关键背景因素。这些组织涵盖多个行业，包括电信、工业设备制造和软件分发。所有参与企业都已建立成熟的实验项目，并负责高流量的软件产品和营销页面。参与组织的 A/B 测试实践开展时间最长达 15+ 年。参与者（10名男性、4名女性）均担任高级职位，涵盖从产品营销专员、首席产品经理到数字体验平台高级经理等角色。该群体拥有丰富的领域专业知识，平均具备14年的营销经验（最多20+年），其中专门从事 A/B 测试的经验平均为5.8年（最多 10+ 年）。我们与这些组织中的每一家分别开展了1-5次、每次30分钟的线上会议，并收到了他们对我们提供的个性化测试结果的书面反馈。我们详细记录了这些会话的内容，并通过迭代综合分析，提炼出反复出现的挑战和设计启示。

下文将介绍我们的工具应解决的痛点（第 3.2 节），以及工具在解决这些痛点时应遵循的关键设计原则（第 3.3 节）。

3.2 A/B 测试的痛点

我们确定了传统 A/B 测试成本高昂或难以实施的四种主要情境：

低流量页面：许多服务于新兴市场中的利基受众或企业对企业（B2B）细分市场的页面，流量不足，无法在合理的时间内达到统计显著性。根据行业估算，由于流量不足，70%–80 % 的拟议测试在达到显著性之前就被放弃了 [50]。

大量设计变体：当设计变体数量较多时，穷举式 A/B/n 测试将难以实施，因为所需流量会随着变体数量增加而迅速增长（呈线性增长，且通常超线性增长） [28, 29]。因此，从业者觉得自己不得不限制创造力，只测试一两个保守的变更，而不去探索多样的设计假设，只为确保测试能在一个业务周期内完成。

低价值变更：由于预期提升不足以抵消完整 A/B 测试所需的成本和时间，小幅变更往往不会经过测试。从业者知道有多项测试值得开展，也知道这些测试会产生影响并带来提升，但它们的提升幅度太小，不足以证明开展测试是合理的。

测量与抽样限制：即使在流量很高的页面上，获取有效的人工反馈也越来越困难。从业者指出，隐私法规往往将数据收集限制在同意使用 Cookie 的用户中，从而形成一个可能无法代表总体人群的偏差子集。

这些痛点形成了一道障碍，常常迫使团队彻底放弃测试。因此，尽管从业者认识到验证的必要性，他们仍经常凭直觉而非数据来实施变更。

3.3 设计启示与原则

为应对这些挑战并有效融入从业者的工作流程，我们总结出四项关键设计原则，以指导 SimAB 的开发。

(1) **在设计流程中左移**：传统实验反馈到达时，开发周期往往已接近尾声。即便实时结果能够即时呈现，它们也只会在漫长的设计流程末端出现，此时构思、实现和迭代已投入大量资源。因此，SimAB 需要在工作流程的早期、全面实现之前，支持预测性筛选，从而提供切实的价值。

(2) **速度优先于完美准确率**：合成 A/B 测试的替代方案往往是不进行测试，因此从业者更倾向于快速获得方向性指导，而非追求完美准确。系统应在几分钟内而不是几天内得出结论，充当快速筛选机制，在不同方案触达实时流量之前筛除明显失败的方案。

(3) **受众限制与附加上下文**：长期从事 A/B 测试的从业者对其（目标）受众已有一定了解，并希望利用这些知识，更准确地界定受众及其行为（例如从已完成的 A/B 测试中检索相关信息）。此外，我们还发现一个颇具吸引力的应用场景：用户可以限制受众，以同样的机制测试某个产品在新市场中的反响。不过，与其说这是核心功能，不如说优先利用从过往 A/B 测试中获得的知识来提高新测试的准确率更为重要。

(4) **理由与可操作的洞察**：传统 A/B 测试的一大局限在于，它能揭示哪个变体胜出，却无法确定原因。为了提供长期价值并支持快速迭代，SimAB 必须具备可解释性，超越简单的胜负判断，确保系统发挥解释工具的作用，而非沦为黑箱。

4 系统设计

根据形成性研究的洞察，我们设计了 SimAB，这是一个利用角色画像条件化 AI 智能体来模拟用户与设计变体之间的交互及用户对变体的偏好，以测试、验证和迭代早期设计的系统。如图 1 所示，该流程包含四个阶段：(1) 输入处理与上下文检索，(2) 角色生成，(3) 角色模拟，以及 (4) 聚合与摘要生成。我们将介绍各个阶段，并着重说明受形成性研究和验证实验启发而作出的设计决策。在接下来的每个介绍性小节末尾，我们将通过图示演示完整流程，其中以维基媒体基金会[1] 的一项 A/B 测试为例。

4.1 输入处理

用户通过提供最多四项输入来设置实验，这些输入向 SimAB 说明实验内容及相应背景：

(1) **变体图像**（必需）：对照组和挑战组设计的图像。遵循在设计流程中左移（第 3.3 节）的设计原则，SimAB 使用屏幕截图而非完全实现的网页。这使团队能够在构思和早期迭代阶段验证设计方向，此时变更的成本仍然较低，且易于落实。对于简单的实验，每个变体提供一个视口即可。对于超出单个视口的实验，用户可以提供整页截图，供智能体通过模拟滚动进行探索；对于多页面流程（例如结账流程），多张屏幕截图可以表示不同步骤（第 4.3.4 节）。

(2) **转化目标**（必填）：对实验成功指标的自然语言描述。目标可以很简单（“按钮点击次数”），也可以很复杂（“按套餐价值加权的订阅注册数”）。此项说明会纳入智能体提示词，以指导评估标准。

(3) **目标受众**（可选）：对模拟用户分布的限制，符合受众限制（第 3.3 节）原则。说明内容可以包括人口统计特征（“年龄 25–35 岁”）、行为特征（“通过付费搜索访问的用户”）或心理特征细分群体（“对价格敏感、会比较商品的购物者”）。如果未提供此项，则会生成能够代表页面内容广泛潜在受众的用户画像。

(4) **上下文文档**（可选）：其他文档，例如分析报告、过往实验结果或领域知识，其中可能包含与角色生成和模拟相关的信息。此项输入符合附加上下文（第 3.3 节）原则。系统会使用这些文档进行检索增强生成（RAG） [32]，以检索相关信息（第 4.1.1 节）。

对照组图像、挑战组图像以及主要转化目标的示例见图 2。

4.1.1 检索增强生成。 当用户提供背景文档时，SimAB 采用 RAG 流水线，提取用于角色生成和模拟的相关信息。该流程处理两类文档：文本类文档会被分块，并使用嵌入模型进行嵌入。对于表格数据，我们采用两阶段方法：（1）大语言模型根据表结构和实验上下文生成 SQL 查询；（2）检索查询结果，并将其汇总为文本洞见。随后，这些摘要会与文本分块一起进行嵌入。

在角色生成和模拟过程中，我们基于嵌入，通过最近邻搜索检索最相关的前 k 个分块。该搜索的查询依据用户输入构建，并包含为两张变体图像生成的文本描述。可选增强包括假设文档嵌入（HyDE）查询扩展 [19] 以及交叉编码器重排序 [36]，以提高检索精度。

4.2 角色生成

用户画像代表将评估两个变体的模拟用户。每个用户画像由分为四个类别的 13 个属性表示：

(1) **人口统计特征**：姓名、年龄范围、职业、收入水平、教育程度、所在地 (2) **心理特征**：兴趣、目标、痛点、技术熟练度 (3) **行为特征**：在线行为模式、典型浏览情境

[1] https://meta.wikimedia.org/wiki/Fundraising_2011/Test_更新/十月/21

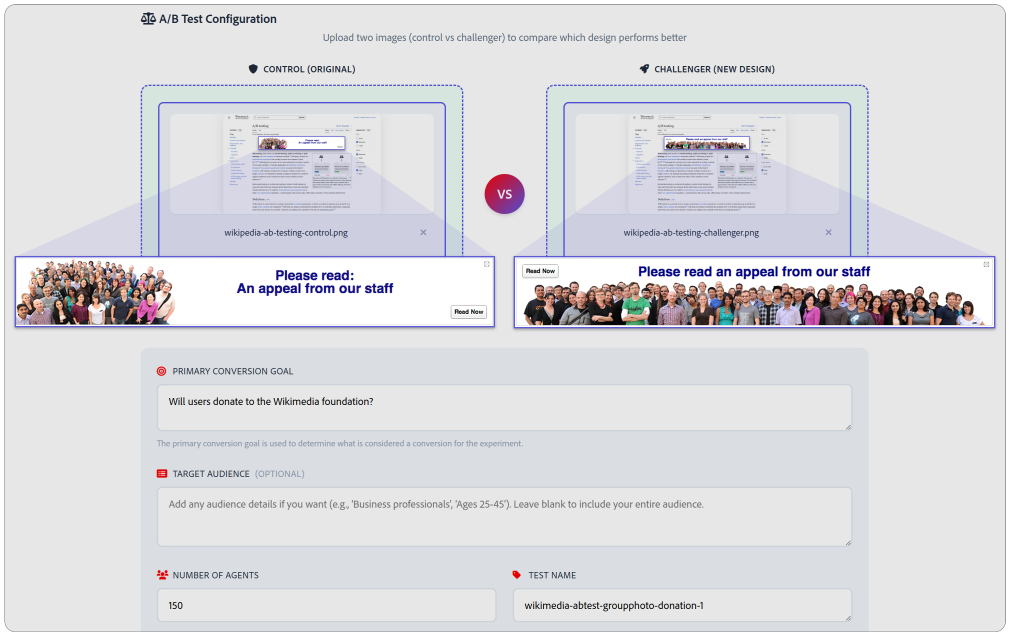

图 2：输入示例。对照组位于左侧；挑战组位于右侧。再加上一个主要转化目标（“用户会向维基媒体基金会捐款吗？”），这些便构成了使用 SimAB 开始 A/B 测试所需的必填输入。本研究使用一个旧的维基百科 A/B 测试创建了该测试，并将其适配到当前的维基百科 A/B 测试页面。该测试比较了两种不同的团队照片位置，以确定哪一种更能促进向维基媒体基金会捐款。

(4) **任务特定**：用户画像将在页面上执行的具体任务，以及其访问时的情境（例如，“在重大购买之前进行研究”，或“通过社交媒体广告到达”）

用户画像通过向`GPT-5-mini-2025-08-07`提供变体截图、目标受众规范（如已提供）以及从任何所提供文档中检索得到的背景来生成。维基媒体捐赠测试的生成的人物角色样本见图 4。生成提示明确要求多样性。完整提示见附录 A.1。

4.2.1 批量生成与多样性约束。 为在效率与多样性之间取得平衡，用户画像以 5–10 个为一批生成。在每一批中，先前生成的人物角色会被纳入背景，以促使模型生成彼此不同的个体。然而，跨批次时该背景会丢失，从而可能导致重复。我们的评估（第 5.4.1 节）量化了角色多样性对准确率的影响。

当用户提供已知的受众细分分布（例如，分析数据显示 30% 为“精通科技的创作者”、25% 为“注重预算的浏览者”）时，我们会调整角色生成以匹配这些比例。将细分分布作为受众限制传入后，均方误差和均方根误差显著降低，如图 3 所示。我们注意到，可以采用更精细的方案来对齐任意固定的受众限制。

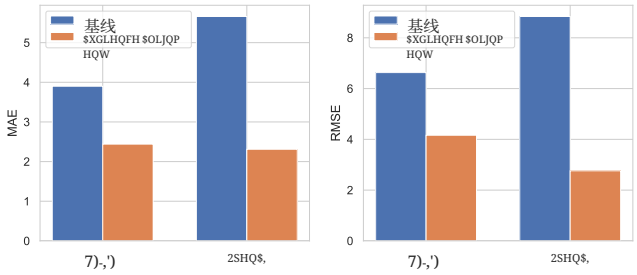

图 3：比较用户提供的受众限制中的目标分布与 SimAB角色生成模块生成的分布。蓝色：未向大语言模型调用提供群体信息的基线；橙色：将群体用作受众限制，以此对齐生成的用户画像。左：均方误差（MSE）；右：均方根误差（RMSE）。词频-逆文档频率（TF-IDF）：评估词语相对于整个语料库在文档中的重要性；使用余弦相似度衡量两个集合中词语的相似度。TF-IDF 和直接向大语言模型提出请求这两种方法，均用于将生成的用户画像归类到最相似的用户画像细分（即受众限制中提供的对齐目标）。值得注意的是，离群值（尤其是 RMSE）显著减少。



图 4：用户画像示例。生成的用户画像包含背景信息，涵盖年龄、教育程度和背景，以及当前背景（当前浏览器会话中的主要目标）。

4.3 角色模拟

每个生成的用户画像都作为人工智能智能体运行，用于评估两个变体。评估采用与大语言模型进行多轮对话的形式，分为三个阶段。

维基百科 A/B 测试捐赠横幅网页的胜出版本见图 5。

4.3.1 背景化。智能体会接收一条系统提示，其中包含：

(1) 完整的用户画像说明 (2) 转化目标，并将其设定为偏好判断的标准。(3) 从所提供文档中检索到的背景信息。(4) 评估指南，涵盖清晰度、阻碍点、信任信号、价值传达以及法律与无障碍合规性。

4.3.2 通过平衡化呈现变体。智能体会在同一个提示中收到两张变体图像，并以中性名称（“图像 1”和“图像 2”）标注。为减轻位置偏差 [71]，我们采用确定性平衡化，在不同智能体之间交替变体的呈现顺序。我们还在第 5.3 节进一步验证了中性命名和平衡化的有效性，以确认这些方法能够成功减轻大语言模型中的偏差。

4.3.3 结构化响应。智能体会生成结构化的 JSON 响应，其中包含：

- **判定**：“图像 1”、“图像 2”或“无”之一
- **理由**：简短说明指出影响偏好的具体元素，符合提供理由与可操作的洞察的设计原则（第 3.3 节）。

“无”选项允许智能体表明该用户画像不会在任一变体上转化，这对于基线转化率较低的页面而言是一个重要信号。

4.3.4 交互式行动。对于超出初始视口或涉及多步骤流程的页面，SimAB 支持交互式探索。智能体可以调用工具函数来调整其视图，然后再做出判定。这些行动既可以是在所提供的（全尺寸）设计屏幕截图上移动视口（宽高比可配置）的滚动行动，也可以是在设计的不同页面之间切换的自定义行动，例如结账按钮将用户导航至结账流程的下一页。对于关键转化元素（例如行动号召元素、定价）出现在首屏以下的页面，这一能力对测试至关重要。如果不进行滚动，智能体可能仅基于首屏内容进行评估，从而遗漏推动真实世界转化的因素。智能体会迭代执行行动，直至达到可配置的上限；若它们表达出足够的信心可以做出判定，则会提前停止。

4.3.5 统计聚合。在每个用户画像给出陈述偏好后，我们使用渐近置信序列 [65]检验聚合合成偏好是否显著异于随机机会。如果合成共识达到显著性阈值，就停止角色模拟并完成实验。我们承认，使用渐近置信序列并不满足所有必要假设，例如，由于用户画像是批量生成的，同一批次内的样本（用户画像）并非相互独立。尽管如此，我们通过实证观察发现，这种方法是用于提前停止的一种非常有效的经验启发式方法。

4.4 摘要与可执行洞察生成

实验完成后，SimAB 会通过最后一次大语言模型调用生成汇总。摘要提示包含所有个体理由，并指示模型：

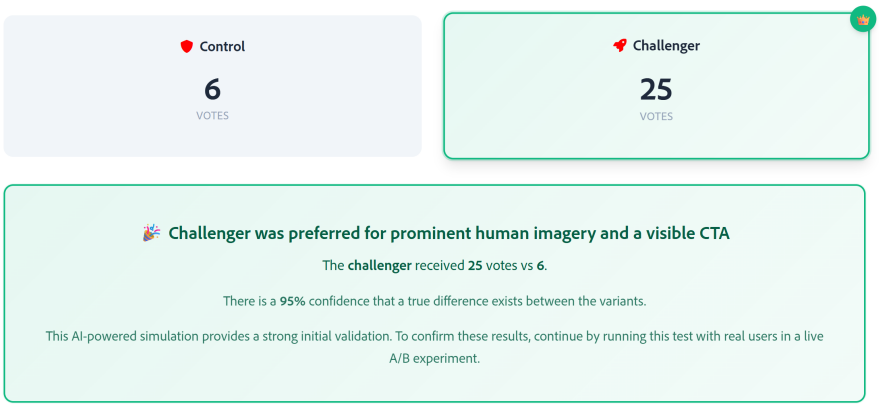

图 5：获胜方案示例。仅在 31 个智能体得出结论后（耗时 1–2 分钟，且全部并行启动），结果便达到统计显著性。该结果与从真实用户处收集的真实数据一致。

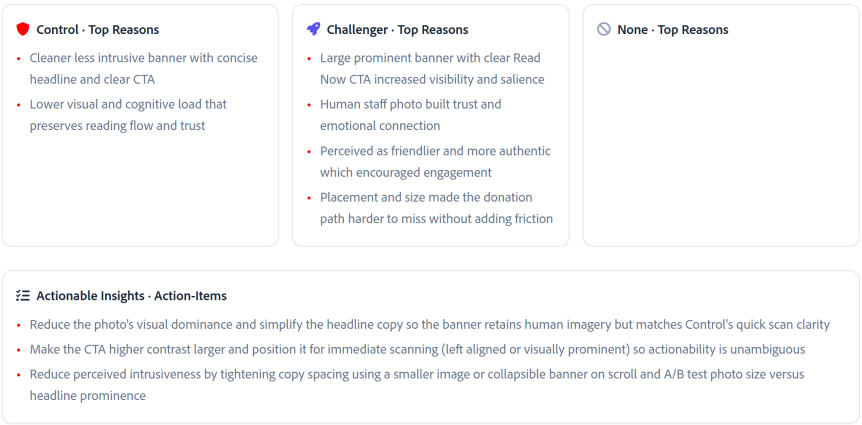

图 6：洞察示例。汇总所有智能体给出的理由，展示偏好对照组/挑战组/无 的主要原因（针对偏好对照组/挑战组/无的智能体）。此外，SimAB 还提出了可立即执行的行动项，以改进和迭代挑战组。

(1) 找出影响对照组和挑战组偏好的 3–5 个最常见主题；(2) 提炼持不同意见的智能体提出的可执行改进建议

汇总结果以及合成人物角色和统计判定可导出为 PDF 报告，以便与利益相关方沟通。

图 6 展示了维基百科 A/B 测试示例的汇总和可执行洞察。

4.5 可复现性

我们使用 GPT-5-mini-2025-08-07 作为生成用户画像的默认模型，并使用 GPT-5-2025-08-07 作为角色模拟的默认模型；只要支持多模态（图像输入），我们也不限定模型。我们最多使用 200 个并发请求；启用提前停止后，足以在 1–5 分钟内完成实验。我们注意到，在我们的用例中无需更高程度的并发，但只需增加并发智能体的数量即可轻松实现。提示模板见附录 A。

5 评估

我们通过五种不同的评估指标来评估 SimAB 的有效性：(1) 针对结果已知的历史 A/B 测试进行回测（第 5.1 节）；(2) 稳健性和一致性评估（第 5.2 节）；(3) 验证偏差缓解策略（第 5.3 节）；(4) 用户画像消融研究（第 5.4 节）；以及 (5) 收集多位实验从业者参与会话后的反馈（第 5.5 节）。

5.1 与历史测试进行回测

为评估 SimAB 的预测是否与历史实验结果一致，我们在包含 47 个有记录结果的历史 A/B 测试的语料库上评估其准确率。该语料库由人工筛选的真实世界 A/B 测试组成，这些测试均达到统计显著性，来源有三个：（1）大型商业 Web 应用；（2）我们组织开发的一款商业桌面应用；（3）维基媒体基金会公开发布的测试 [66, 67]。该语料库涵盖广泛的设计变更，包括针对行动号召元素、标题、布局和图像的变更，这些变更在多个地理区域开展。此外，这些

表 1：SimAB 在由 47 个多样化历史 A/B 测试构成的语料库中取得的混淆矩阵。请注意，其中两项测试未达到统计显著性，因此未计入混淆矩阵。

		真实标签		
		挑战组	对照组	精确率
预测结果	挑战组	22（真正例）	9（假正例）	71%
	对照组	6（假负例）	8 (TN)	
	召回率	79%		准确率：67%

实验的目标各不相同，例如增加注册量、提高用户参与度或降低跳出率。数据覆盖的时间跨度较广：我们组织的测试收集于过去五年，而维基媒体数据集涵盖 2010 年和 2011 年的实验。对于每项测试，我们都使用该测试的对照组和挑战组屏幕截图、转化目标，以及可用时的目标受众限制来配置 SimAB，并运行实验，使用 $\alpha = 0.05$ 检验统计显著性。表 1 展示了混淆矩阵。

结果表明，SimAB 提供了强有力的预测信号：在三分之二的案例中与历史实验结果一致（一致率为 67%），并在历史测试偏向挑战组的案例中有 80% 达成一致（召回率）。尽管这一一致率远高于 50% 的随机基线，表明 SimAB 提供了有用的预测信号，但我们注意到，预测准确率仅是该系统整体价值的一部分。正如从业者反馈（第 5.5 节）所强调的，附带的理由 和可执行的洞察 往往被认为比二元判定本身更有价值。通过阐明偏好背后的推理，SimAB 揭示了具体的设计权衡和可用性障碍，从而引导进一步的迭代与优化，即使在统计预测并不完美的情况下也是如此。

此外，我们发现，将一份拆解网页过去三年表现的分析报告作为上下文文档纳入其中，可将达到统计显著性所需的智能体数量减少 50% 以上，从而显著更快速地结束实验，同时保持相同的准确率。

除了汇总指标之外，我们还发现准确率与达到统计显著性所需的智能体数量之间存在很强的相关性。按达到统计显著性所需的智能体数量对准确率进行分层（图 7）后可以看出，达到统计显著性所需的智能体越少，就越

结果有多准确。这一模式与直觉相符：对于差异非常明显的变体，较少的智能体就足以准确检测出偏好。当差异较为细微时，则需要更多智能体，合成信号也会变得更加不确定，这与真实 A/B 测试同样难以应对的情况类似。

我们将高置信度 测试定义为：在不超过 70 个智能体的情况下得出显著结果的测试。筛选高置信度 结果后，我们报告的准确率为 75%，精确率为 75%，召回率为 88%，F1 分数为 81%。所有测试中有 60% 属于高置信度 测试。值得注意的是，在 20 个智能体内达到显著性的极高置信度 测试，其准确率超过 80%，F1 分数超过 90%。然而，达到极高置信度 的代价是，这类测试中达到显著性的比例为 30%，与传统 A/B 测试中通常的比例相当。

这表明一种实用的运行模式：用户根据快速收敛的预测采取行动，同时将较慢、置信度较低的结果视为尚无定论，以提高显著结果的准确率（即便仅限于高置信度 测试，我们报告称，达到显著性的测试占 60%，而传统 A/B 测试的行业估算仅为 20–30% [50]）。除了准确率更高，提前停止还能显著提升速度，直接满足了第 3.3 节所指出的设计要求——“优先考虑速度

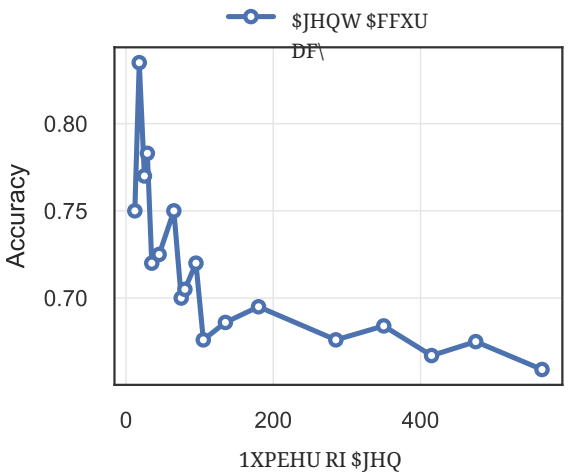

图 7：智能体数量与准确率之间的权衡。我们绘制了以 $\leq x$ 个智能体得出结论的实验子集的累计准确率。快速达到统计显著性的实验，其准确率显著高于需要更多智能体的实验。

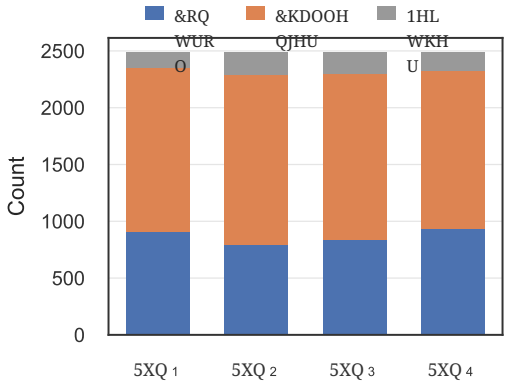

图 8：同一实验的四次独立运行的一致性分析。虽然由于大语言模型的随机性，具体投票数略有波动，但系统始终以显著优势识别出同一获胜方（挑战组）。

而非完美准确性”，从而实现时间敏感型设计环境所必需的关键、高速反馈循环。我们经常观察到，在这 60% 的快速测试中，结果可在 <1 分钟内得出，这足以满足终端用户应用的需求；随着大语言模型调用预计会提速，这一优势将更加切实可见。

5.2 稳健性与一致性

要使 SimAB 在实践中具有实用价值，结果必须具有可复现性和一致性。我们开展了三项实验来进行稳健性分析。

5.2.1 重复实验。 为评估不同运行之间的稳定性，我们对同一项实验（一个差异明显的定价页面测试）执行了四次。正如

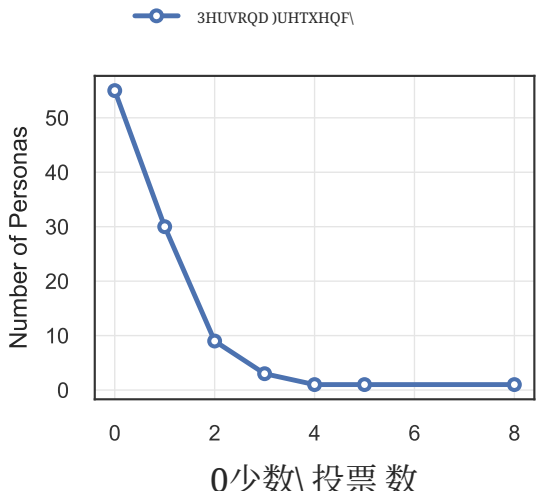

图 9：角色一致性分析。我们在完全相同的输入上让每个用户画像运行 20 次。横轴表示偏离票（少数票）的数量。陡峭的下降趋势表明，各个用户画像都保持着稳定、确定性的偏好，其中大多数在 20 次运行中仅出现 0 或 1 次偏离。

图 8 所示，尽管投票分布存在差异，但四次运行均确定了同一个胜者（挑战组）。在各次运行中，挑战组的偏好率介于 56% 至 62% 之间，处于预期的统计变化范围内。这表明该方法能够抵御大语言模型采样所固有的随机性。

5.2.2 多挑战者传递性。 由于 SimAB 进行两两比较，多变体测试需要运行所有两两比较。一个关键要求是 传递性：如果 A 胜过 B，且 B 胜过 C，那么 A 应胜过 C。非传递结果（循环）会削弱人们对排名的信心。

我们分析了来自与回测语料库（第 5.1 节）相同来源的多变体实验中整理出的 50+ 次两两比较。在所有情况下，结果都完全满足传递性，为每组设计确定了唯一的全序。我们未观察到任何循环偏好实例（例如，$A > B > C > A$）。这种一致性证实，SimAB 适用于迭代式“爬山”工作流：设计师可以根据先前的胜出方案逐步改进挑战组；这也进一步验证了其能够抵御随机（种子）变化的稳健性。

5.2.3 用户画像一致性。为了评估用户画像能否提供一致的回答，我们测试了单个用户画像在面对相同输入时是否会给出确定性的判定。我们让 100 个不同的用户画像分别针对同一对照组—挑战者组运行 20 次，并测量少数票数：即智能体偏离其用户画像多数判定的次数。

图 9 所示的结果表明，用户画像具有较高的一致性。超过 50% 的用户画像表现出完美的稳定性（少数票数为 0），在 20 次试验中始终识别出相同的偏好。此外，85% 的用户画像出现的偏离不超过两次。这证实 SimAB 智能体能够生成内部一致的陈述偏好。

5.3 偏差缓解验证

大语言模型评估偏差已有充分文献记载 [71]。在本节中，我们评估我们的方法能否有效缓解这些偏差。

表 2：对我们的偏差缓解技术——平衡处理和中性名称——进行的验证。该表展示了在使用相同图像进行的测试中两个变体分别获得的投票数，以隔离偏差。

平衡处-理	中性名称	对照组	挑战组	None	p值
✗	✗	126	749	125	0.0000
✗	✓	52	887	61	0.0000
✓	✗	442	434	124	0.8131
✓	✓	460	469	71	0.7930

5.3.1 平衡呈现的有效性。 我们通过开展一项实验验证了系统中的平衡化，其中对照组与挑战组为相同图像。4,756个智能体之后的结果：

- 先呈现图像的票数：2,187
- 后呈现图像的票数：2,191
- “无”票数：378

4票之差在统计上可忽略不计（$\chi^2 = 0.0004$， $p = 0.984$）。作为对照，在未进行平衡化（对照组优先）的情况下，我们观察到挑战组被偏好的次数比对照组多6–17倍，即一种会严重扭曲结果的偏差。在表2中，我们报告了结合平衡化与中性命名的效果。

5.4 角色消融研究

5.4.1 角色多样性消融。 为评估使用用户画像是否会影响准确率，我们比较了三种设置：

- **基线**：标准批量生成（每批 5–10 个），不同批次之间不进行显式去重
- **无多样性**：所有智能体使用同一个固定用户画像 • **高多样性**：使用嵌入[2] 余弦相似度（$\theta = 0.85$）进行去重，并持续重新生成，直到达到多样性要求。重新生成采用角色间生成方式 [8]

结果汇总于图 10。我们发现，移除多样性会使一致率降低 10 个百分点，这表明多样化的合成视角有助于提高与真实实验结果之间的预测一致性。然而，极高的多样性（高多样性条件）并未优于基线，这表明边际收益递减。基线方法通过批量生成自然地产生适度的多样性，似乎在多样性与效率之间取得了有效平衡。

5.4.2 按组分析的角色。 接下来，我们旨在证明，按某种分类划分的不同用户画像群体会呈现不同的投票行为。图 11 展示了按年龄、职业原型、收入水平和技术熟练度进行的划分。在四种划分中，有三种至少有一个群体显著偏离分布；年龄组“60-69”的 p值为 $4{\cdot}10^{-32}$，原型组“教师”的 p值为 $3 \cdot 10^{-7}$，技术熟练度组“低”的 p值为 $7 \cdot 10^{-23}$，表明这些群体与其他群体来自同一分布的可能性很低，说明不同用户画像的行为存在差异。

此外，即使未指定受众限制，SimAB 在获得不同输入图像时也足以生成显著不同的用户画像，这表明用户画像生成会对输入图像作出响应。

[2]嵌入是根据用户画像的文本描述，使用 OpenAI 文本嵌入生成的。

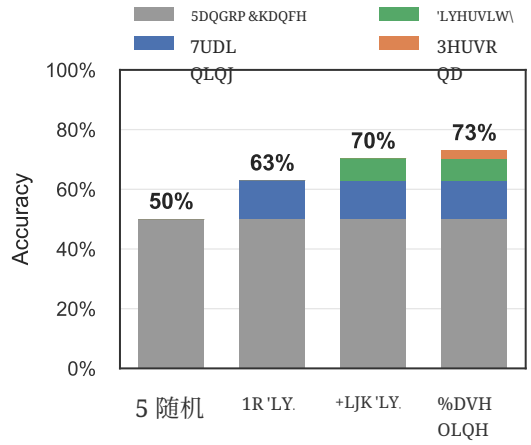

图 10：角色多样性消融研究。我们展示了完全不使用用户画像（即每个实验只采用一个固定用户画像）与使用多个自适应用户画像的效果对比，后者可进一步提高准确率。

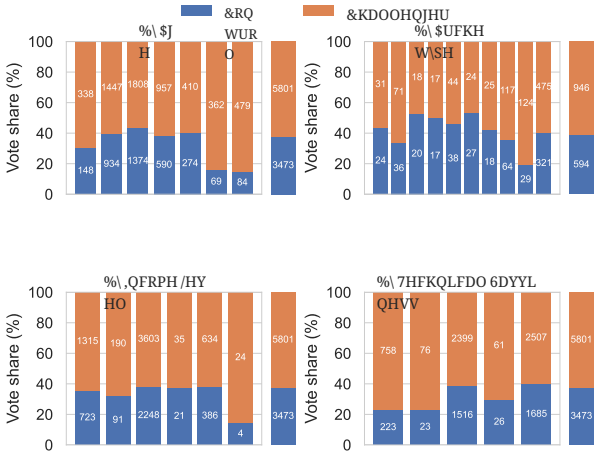

图 11：按不同角色人设划分呈现的投票分布：按年龄、职业原型、收入水平和技术熟练度划分。各组中对照组的得票率以蓝色表示，挑战组的得票率以橙色表示。

5.5 从业者反馈

作为持续参与式设计活动的一部分，我们使用第3节所述来自四家企业组织的同一批14名从业者对 SimAB 进行了评估。我们保留该队列，以直接验证系统是否有效解决了我们在初始会话（第3节）中识别出的具体痛点。

评估方法经过调整，以适应组织方面的限制。由于组织政策，外部用户采用了“礼宾式”评估模式：他们为当前的设计问题提供素材，我们则执行模拟并与他们分享结果。相比之下，四位内部从业者能够直接访问 SimAB 用户界面，开展独立实验。完成这些评估后，我们召开了评审会议，分析结果。

支持以往无法测试的场景：各组织反复提到，SimAB 对于传统 A/B 测试不切实际或不可行的场景具有重要价值。参与者特别强调了低流量页面（例如利基市场、区域性页面或面向企业客户的内容），以及预期只能带来小幅但具有实际意义的改进的变更，这直接回应了第 3.2 节中提出的低流量页面痛点。

从业者意识到，SimAB 可以用于那些流量不足、无法进行测试的页面，因此对此感到十分兴奋；也就是说，它不仅能够帮助开展现有测试，还能让全新的页面以更加一致且数据驱动的方式得到优化。

促进利益相关者之间的一致性与共识：几位参与者指出，SimAB 有望帮助减少内部决策中的不一致。设计评审通常涉及多位观点和启发式判断各不相同的利益相关者，从而引发主观争论。模拟结果及其附带的理由提供了一个共享的数据驱动参照点，有助于达成共识并提高讨论效率，从而验证了提供理由与可操作的洞察（第 3.3 节）这一设计原则。

接受准确性权衡：参与者对系统的局限性表现出了务实的理解。团队并未期待完美准确性，而是强调透明度、已知的失效模式和恰当的使用方式；他们将 SimAB 视为指导早期决策和优先级排序的工具，同时把实时测试留给高风险候选方案。这种采用模式证实了我们的设计原则是有效的：优先考虑速度而非完美准确性，并在设计流程中左移（第 3.3 节）。从业者成功地将该工具视为一种快速筛选手段，而非真实测试的替代品。

6 讨论

我们的评估表明，合成受众能够为 A/B 测试提供有意义的信号，其准确率会随测试难度而呈现可预测的变化。本节将讨论这些发现的意义，说明适用的用例，承认局限性，并概述未来研究方向。

6.1 设计研究背景下的准确率结果

与从业者对传统 A/B 测试所期待的确定性相比，报告中的部分准确率乍看之下可能并不高。然而，若考虑若干背景因素，这些表现可能比表面看来更有用：

将分层置信度用作筛选器：对于高置信度预测（按照第 5.1 节的定义，测试在 ≤70 个智能体时得出结论），准确率为 75%；对于极高置信度预测（≤20 个智能体），准确率为 >80%，这在很大程度上更具可操作性。实际应用中，用户可以将 SimAB 视为高精确率筛选器：快速得出的结果很可能可靠；需要大量智能体的结果则存在不确定性，值得进行实时测试。这种解读，即快速得出的结果置信度高，缓慢得出的结果不确定，与人们对 A/B 测试的一般直觉相符。

将模拟作为直觉的替代方案：传统 A/B 测试本身也存在不可忽视的错误率，原因包括多重检验、季节性和外部混杂因素。真正相关的比较是 SimAB 与完全不进行测试之间的比较；对于低流量页面、多变体比较和微优化，默认做法正是不进行测试。在 5 分钟内达到 67% 的一致率，或许优于基于直觉的决策，不过要进行直接比较，还需要进一步研究。

解释性理由的价值：即使判定不正确，生成的理由通常也能揭示合理的设计考量。

几位从业者提到，他们会利用理由来识别潜在问题，即使他们不同意总体判定。

6.2 用例

基于我们的评估和部署经验，我们确定了 SimAB 最能发挥显著价值的四种主要用例：

预检筛选：最明确的用例是在开展实时实验之前，快速筛选设计变体。SimAB 可以迅速淘汰明显不理想的选项，从而缩小变体空间，减少更昂贵的实时测试所需的工作量。一位部署合作伙伴将此描述为“快速扼杀糟糕的想法”。如果 SimAB 强烈否定某个变体，我们可以认为，这项模拟 A/B 测试的结果很可能与未来真实用户的反馈一致，从而为设计、营销和产品从业者节省时间和资源。

测试“不可测试项”：由于页面流量低、受众细分群体高度特定、需要探索新市场以及变体空间庞大（例如有 5 种潜在设计），许多设计优化机会在结构上无法通过传统测试进行评估。对于过去往往只能依靠直觉的情形，SimAB 能提供方向性信号。在这些只能靠猜测的场景中，67% 的准确率尤其有价值。

隐私优先测试：与传统 A/B 测试不同，SimAB 无需收集用户数据。这为数据治理要求严格（例如，符合《通用数据保护条例》）的组织，或在隐私敏感场景（例如，医疗保健、金融）中进行测试，提供了一种保护隐私的替代方案。无需观察真实用户行为即可模拟用户反馈，这代表着隐私保护方面的一项实质性改进。

快速设计迭代：传统 A/B 测试只有在开发完成后才能提供反馈。SimAB 可在实现之前评估模型图和原型，从而将反馈前移至设计流程中。这使迭代节奏从“设计 → 开发 → 测试（数月） → 学习”转变为“设计 → 测试（数分钟） → 改进 → 开发”。

6.3 局限性与边界条件

我们指出了若干限制 SimAB 适用性的关键局限：

超越屏幕截图的行为复杂性。 SimAB 评估的是静态截图，并可选择进行模拟滚动。因此，对于这些因素是主要驱动因素的测试，仍有必要开展实时实验，因为它无法捕捉：

- 性能效应：页面加载时间、响应能力和感知速度都会显著影响转化，但这些因素在屏幕截图中不可见。• 交互动态：悬停状态、动画、微交互和复杂的表单行为都会影响用户体验，而静态图像无法传达这些影响。
- 时间效应：重复访问、随时间变化的品牌认知以及网络效应，都会在多个会话中逐渐显现。

新颖的交互模式：大语言模型评估反映的是从训练数据中学到的模式。对于真正新颖的界面，例如与现有网页惯例截然不同的设计，智能体可能无法准确预测人类反应。设计偏离训练数据分布越远，预测的可靠性就越低。我们注意到，训练分布与测试时设计之间的这种偏移，可能既源于领域变更，也源于人类行为随时间发生的变化。

细微的效应量：我们的准确率分层结果表明，SimAB 在差异明显的测试中表现最佳。对于细微的优化，即使人类也难以表达偏好（也就是传统 A/B 测试同样难以应对的“棘手案例”），合成评估可能无法提供可靠的信号。这并非我们方法独有的问题，而是针对几乎相同的选项进行偏好 elicitation 时存在的根本局限。不过，需要指出的是，随着模拟运行时间延长至 2–10 分钟，我们有 95% 的测试达到了显著性，远高于传统 A/B 测试通常达到的 30% [50]。

表征效度：合成人物角色继承了底层语言模型的偏见。Santurkar 等人 [46] 指出，大语言模型的输出过度反映了西方受教育人群的观点。我们的人物角色可能无法准确代表全球各地、不同文化背景或边缘化的用户群体。从业者应批判性地评估生成的人物角色是否充分覆盖了预期受众，尤其是服务于多元人群的产品。

操纵手段的自动化：长期以来，A/B 测试一直引发人们对“暗黑模式”的担忧，即利用心理偏差、违背用户利益以促成转化的设计 [21]。SimAB 的速度和低成本使说服技巧更容易快速迭代。尽管该工具本身是中性的，但负责任的部署需要制定伦理准则，明确哪些优化是合适的。

推广到其他领域：我们的 A/B 测试语料库主要由营销和电子商务网页构成。可能需要进一步测试和开发，才能推广这些结果，并使系统适用于其他领域。

6.4 未来方向

我们的工作为未来研究开辟了多个方向。

领域自适应微调反映的是未经领域自适应的零样本性能。在特定领域内，利用历史 A/B 测试结果进行微调，可大幅提高该领域的准确率。拥有大量实验档案的组织可以训练针对其用户进行校准的领域专用评估器。

多模态与动态评估。除了屏幕截图，评估视频录像、交互式原型，甚至实时网站，也能捕捉当前尚未涵盖的行为动态。集成 Browser Use [58] 等工具以实现完整的页面交互，有望弥合静态评估与动态评估之间的差距。

校准的不确定性量化。我们目前的置信度估计是启发式的。开发有原则的不确定性量化方法（针对判定正确性的校准概率估计），将帮助用户做出更好的决策。

面向大语言模型的界面评估。随着人工智能智能体与网站的交互日益频繁（用于购物、信息检索等），针对大语言模型“用户”进行优化，可能会变得与针对人类进行优化同样重要。SimAB 的架构可自然扩展到这一新兴用例，评估设计对人工智能交互模式的支持效果。

7 结论

传统 A/B 测试依赖真实用户流量，这造成了一个根本瓶颈，限制了整个网络的实验能力。本文介绍了 SimAB，这一系统将 A/B 测试重新定义为利用角色画像条件化 AI 智能体进行的快速、保护隐私的模拟。通过与实验从业者开展形成性研究，我们确定了三种流量限制导致无法进行测试的情形：低流量页面、多变体比较和微优化。这些洞察为系统设计提供了依据，其核心是速度、可执行的理由以及灵活的受众规范。我们的技术贡献包括一套完整的合成 A/B 测试流程：根据网页内容生成多样化的用户画像、由平衡设计 AI 智能体进行评估、通过序贯统计聚合实现有效的提前停止，以及综合理由以形成可执行的洞察。我们证明，平衡化能有效消除位置偏差（在 4,756 次相同图像评估中，票数差为 4 票，$p > 0.99$），而角色多样性可将准确率提高 10 个百分点。对 47 项结果已知的历史 A/B 测试进行评估后发现，总体一致率为 67%；对于快速得出结论的高置信度 和极高置信度 预测，一致率分别提高至 75% 和 >80%。在 50+ 次多变体比较中实现了完全传递性，证实了排名的一致性。与四家企业合作伙伴的部署验证了其实用

价值：团队认可其快速筛选、解释性理由，以及测试此前无法实施的情形的能力。

我们将 SimAB 定位为现场实验的补充工具，而非替代品；它拓展了可供测试的范围。分层准确率结果表明，一种切实可行的运作方式是：将快速收敛的结果视为较强的方向性信号；将收敛缓慢的结果视为较弱的假设，需通过现场验证。这种“预检”思路，即快速淘汰不佳创意、筛选变体空间，并将设计流程中的反馈前移，体现了 SimAB 的核心价值主张。我们认为，SimAB 代表着设计评估由谁开展这一点正在发生转变。传统 A/B 测试需要大量流量、技术基础设施以及组织层面的耐心。合成 A/B 测试使实验更加普及，让小型组织、新兴市场和个人设计师也能进行数据驱动设计。

随着生成式人工智能改变内容创作和用户交互模式，对快速评估的需求只会不断增长。SimAB 表明，合成受众提供的预测信号与历史实验结果吻合的频率高于随机水平。我们希望这项工作能推动 AI 辅助评估领域的进一步研究，并助力实现一个人人都能进行数据驱动设计的未来。

致谢

感谢 David Arbour 对本研究提出宝贵的讨论意见和反馈。

参考文献

[1] Marwa Abdulhai、Ryan Cheng、Donovan Clay、Tim Althoff、Sergey Levine 和 Natasha Jaques。2025。《通过多轮强化学习持续模拟人类用户画像》。[2] Anish Agarwal、Devavrat Shah 和 Dennis Shen。2024。《使用合成干预进行合成 A/B 测试》。*arXiv* 预印本（2024）。[3] Danial Amin、Joni Salminen、Farhan Ahmed、Sonja M. H. Tervola、Sankalp Sethi 和 Bernard J. Jansen。2025。《生成式人工智能如何用于人物画像开发？：对 52 篇研究文章的系统综述》。arXiv:2504.04927 doi:10.48550/arXiv.2504.04927[4] Lisa P. Argyle、Ethan C. Busby、Nancy Fulda、Joshua R. Gubler、Christopher Rytting 和 David Wingate。2023。《由一及众：使用语言模型模拟人类样本》。*Political Analysis* 31，3（2023），337–351。doi:10.1017/ pan.2023.2[5] Karim Benharrak、Tim Zindulka、Florian Lehmann、Hendrik Heuer 和 Daniel Buschek。2024。《由作者定义的 AI 用户画像：按需生成反馈》。载于*2024*年 *CHI* 人机交互系统大会（美国纽约州纽约，2024-05-11）*(CHI '24)*。计算机协会，1 18 10 1145 3613904 3642406 – 。doi: . / 。[6] James Brand、A. Israeli 和 Donald Ngwe。2024。。《使用 GPT 进行市场研究》。载于*ACM* 经济学与计算大会。[7] Aurélien Bück-Kaeffer、Je Qin Chooi、Dan Zhao、M. P. Touzel、Kellin Pelrine、J. Godbout、Reihaneh Rabbany 和 Zachary Yang。2025。《BluePrint：用于 LLM 角色画像评估与训练的社交媒体用户数据集》。[8] Xin Chan、Xiaoyang Wang、Dian Yu、Haitao Mi 和 Dong Yu。2024。《利用 1,000,000,000 个用户画像扩展合成数据生成》。doi:10.48550/arXiv.2406. 20094[9] Chaoran Chen、Bingsheng Yao、Ruishi Zou、Wenyue Hua、Weimin Lyu、Yanfang Ye、Toby Jia-Jun Li 和 Dakuo Wang。2025。《迈向 RPA 评估指南：基于大语言模型的角色扮演智能体综述》。doi:10.48550/arXiv.2502.13012 arXiv:2502.13012 [cs]。[10] J unzhe Chen 等。2024。《AgentA/B：借助交互式大语言模型智能体实现自动化、可扩展的网页 A/B 测试》。https://arxiv.org/abs/2400.00000[11] Luyu Chen、Quanyu Dai、Zeyu Zhang、Xueyang Feng、Mingyu Zhang、Pengcheng Tang、Xu Chen、Yue Zhu 和 Zhenhua Dong。2025。《RecUserSim：用于评估对话式推荐系统的逼真、多样化用户模拟器》。载于万维网大会。[12] Yoonseo Choi、Eun Jeong Kang、Seulgi Choi、Min Kyung Lee 和 Juho Kim。2025。*Proxona*：利用 *LLM* 驱动的受众角色画像，支持创作者进行意义建构与创意构思。arXiv: 2408 10937 [ ] 10 48550 2408 10937 . cs doi: . / arXiv. 。[13] Stefano De Paoli。2026。。《用户画像、创意构思与大语言模型：一项事后研究》。208 2026 103690 10 1016 2025 103690 ( ), 。doi: . / j.ijhcs. 。[14] Shiying Ding、Xinyi Chen、Yan Fang、Wenrui Liu、Yiwu Qiu 和 Chunlei Chai。2023。《DesignGPT：设计中的多智能体协作》。载于*2023*年第*16*届国际计算智能与设计研讨会（*ISCID*）。204–208。doi:10. 1109/ISCID59865.2023.00056 ISSN：2473-3547。[15] Peitong Duan、Chin-Yi Cheng、Bjoern Hartmann 和 Yang Li。2025。用于生成设计评论的迭代式优化视觉提示。（2025）。arXiv: 2412 16829 [ ] 10 48550 2412 16829 . cs doi: . /arXiv. . Peitong Duan、Chin-Yi Cheng、Gang Li、Bjoern Hartmann 和 Yang Li。2024。UICrit：借助界面评论数据集增强自动化设计评估。载于第*37*届 *ACM* 用户界面软件与技术年度研讨会论文集（美国纽约，2024-10-11）（*UIST '24*）。计算机协会，1 17 10 1145 3654777 3676381 – 。doi: . / . Peitong Duan、Jeremy Warner 和 Bjoern Hartmann。2023。探索利用大语言模型生成界面设计反馈。载于第*36*届 *ACM* 用户界面软件与技术年度研讨会增刊论文集（*UIST '23* 增刊）。计算机协会，美国纽约，1–3。doi:10.1145/3586182.3615810[18] Peitong Duan、Jeremy Warner、Yang Li 和 Bjoern Hartmann。2024。利用大语言模型为界面模型稿生成自动反馈。载于*2024*年 *CHI* 人机交互系统大会论文集（美国纽约，2024-05-11）（*CHI '24*）。计算机协会，1–20。doi:10.1145/3613904.3642782[19] Luyu Gao、Xueguang Ma、Jimmy Lin 和 Jamie Callan。2023。无需相关性标签的精确零样本稠密检索。doi:10.48550/arXiv.2212.10496[20] Aditya Gopalan、Shie Mannor 和 Yishay Mansour。2014。用于复杂在线问题的汤普森采样。载于第*31*届国际机器学习大会论文集（机器学习研究论文集，第*32*卷）。PMLR，100–108。[21] Colin M. Gray、Yubo Kou、Bryan Battles、Joseph Hoggatt 和 Austin L. Toombs。2018。用户体验设计的阴暗（模式）面。载于*2018*年 *CHI* 人机交互系统大会论文集。doi: 10 1145 3173574 3174108 . / .[22] (Eric) Heng Gu、Senthil Chandrasegaran 和 Peter Lloyd。2025。. 合成用户：设计师与基于角色设定的聊天机器人互动所得的洞察。用于工程设计、分析与制造的人工智能39（2025），e2。doi:10.1017/S0890060424000283[23] Juhye Ha、Hyeon Jeon、Daeun Han、Jinwook Seo 和 Changhoon Oh。2024。CloChat：了解人们如何在大语言模型中自定义、互动和体验角色。载于*2024*年 *CHI* 人机交互系统大会论文集（美国纽约，2024-05-11）（*CHI '24*）。计算机协会，1–24。doi:10.1145/3613904.3642472[24] Hongliang He、Wenlin Yao、Kaixin Ma、Wenhao Yu、Yong Dai、Hongming Zhang、Zhenzhong Lan 和 Dong Yu。2024。WebVoyager：利用大型多模态模型构建端到端网页智能体。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），Lun-Wei Ku、Andre Martins 和 Vivek Srikumar（编）。计算语言学协会，泰国曼谷，6864 6890 10 18653 1 2024 371 – 。doi: . / v / .acl-long。[25] John J. Horton。2023。. 作为模拟经济代理的大语言模型：我们能从 Homo Silicus 中学到什么？*NBER* 工作论文（2023）。doi:10. 3386/w31122[26] Zhengyu Hu、Zheyuan Xiao、Max Xiong、Yuxuan Lei、Tianfu Wang、Jianxun Lian、Kaize Ding、Ziang Xiao、Nicholas Jing Yuan 和 Xing Xie。2025。面向基于大语言模型的社会模拟、与总体对齐的角色生成。[27] Perttu Hämäläinen、Mikke Tavast 和 Anton Kunnari。2023。评估大语言模型生成合成人机交互研究数据的能力：案例研究。载于*2023*年 *CHI* 人机交互系统大会论文集（美国纽约，2023-04-19）（*CHI '23*）。计算机协会，1 19 10 1145 3544548 3580688 – 。doi: . / . Kevin Jamieson 和 Robert Nowak。2014。. 固定置信度设定下多臂老虎机的最佳臂识别算法。载于信息科学与系统。[29] Emilie Kaufmann、Olivier Cappé 和 Aurélien Garivier。2016。多臂老虎机模型中最佳臂识别的复杂性。机器学习研究期刊17（2016），1–42。[30] Ron Kohavi、Alex Deng、Brian Frasca、Toby Walker、Ya Xu 和 Nils Pohlmann。2013。大规模在线受控实验。第*19*届 *ACM SIGKDD* 国际知识发现与数据挖掘大会论文集（ 2013 10 1145 2487575 2488217 ）。doi: . / . Akaash Kolluri、Shengguang Wu、Joon Sung Park 和 Michael S. Bernstein。2025。针对社会科学实验中的人类行为预测微调大语言模型。arXiv:2509.05830 doi:10.48550/arXiv.2509.05830[32] Patrick Lewis、Ethan Perez、Aleksandra Piktus、Fabio Petroni、Vladimir Karpukhin、Naman Goyal、Heinrich Küttler、Mike Lewis、Wen-tau Yih、Tim Rock- täschel 等。2020。面向知识密集型自然语言处理任务的检索增强生成。神经信息处理系统进展33（2020），9459–9474。[33] Yuxuan Lu、Jing Huang、Yan Han、Bennet Bei、Yaochen Xie、Dakuo Wang、Jessie Wang 和 Qi He。2025。大语言模型智能体能否模拟多轮人类行为？来自真实在线客户行为数据的证据。[34] Yuxuan Lu、Bingsheng Yao、Hansu Gu、Jing Huang、Jessie Wang、Laurence Li、Jiri Gesi、Qi He、Toby Jia-Jun Li 和 Dakuo Wang。2025。UXAgent：用于网页设计的基于大语言模型智能体的可用性测试框架。载于*CHI* 扩展摘要。doi:10.1145/3706599.3719729

[35] Saab Mansour、Leonardo Perelli、Lorenzo Mainetti、George Davidson 和 Stefano D'Amato。2025。PAARS：角色对齐的代理型零售购物者。载于第1届智能体语言模型研究研讨会（REALM 2025）论文集。[36] Rodrigo Nogueira 和 Kyunghyun Cho。2019。使用 BERT 进行段落重排序。arXiv 预印本 arXiv:1901.04085（2019）。[37] Rock Yuren Pang、Hope Schroeder、Kynnedy Simone Smith、Solon Barocas、Ziang Xiao、Emily Tseng 和 Danielle Bragg。2025。理解 CHI 的大语言模型化：通过系统性文献综述剖析大语言模型对 CHI 的影响。载于2025年 CHI 人机交互系统大会论文集（美国纽约，2025-04-25）（CHI '25）。计算机协会，1–20。doi:10.1145/3706598.3713726[38] Joon Sung Park、Joseph O'Brien、Carrie Jun Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。2023。生成式智能体：人类行为的交互式拟像。载于第36届 ACM 用户界面软件与技术年度研讨会论文集。ACM，美国加利福尼亚州旧金山，1–22。doi:10.1145/3586183.3606763[39] Joon Sung Park、Lindsay Popowski、Carrie Cai、Meredith Ringel Morris、Percy Liang 和 Michael S. Bernstein。2022。社会拟像：为社会计算系统创建填充式原型。载于第35届 ACM 用户界面软件与技术年度研讨会论文集（美国纽约，2022-10-28）（UIST '22）。计算机协会，1–18。doi:10.1145/ 3526113.3545616[40] Joon Sung Park、Carolyn Q Zou、Aaron Shaw、Benjamin Mako Hill、Carrie Cai、Meredith Ringel Morris、Robb Willer、Percy Liang 和 Michael S Bernstein。2024。对1000人的生成式智能体模拟。arXiv 预印本 arXiv:2411.10109（2024）。[41] Tianyi Peng、George Gui、Daniel J. Merlau、Grace Jiarui Fan、Malek Ben Sliman、Melanie Brucks、Eric J. Johnson、Vicki Morwitz 等。2025。数字孪生大型研究揭示其优势、弱点及进一步改进的机遇。arXiv: 2509 19088 [ ] 10 48550 2509 19088 . cs.AI doi: . /arXiv. .[42] J. Piao、Yuwei Yan、Jun Zhang、Nian Li、Junbo Yan、Xiaochong Lan、Zhihong Lu、Zhiheng Zheng、Jing Yi Wang、Di Zhou、Chen Gao、Fengli Xu、Fang Zhang、Ke Rong、Jun Su 和 Yong Li。2025。AgentSociety：大规模模拟大语言模型驱动的生成式智能体，助力深入理解人类行为与社会。[43] Ali Ebrahimi Pourasad 和 Walid Maalej。2024。生成式人工智能会让可用性测试过时吗？arXiv: 2411 00634 10 48550 2411 00634 . doi: . /arXiv. .[44] John Pruitt 和 Jonathan Grudin。2003. 用户画像：实践与理论。载于2003年面向用户体验的设计大会论文集（美国纽约，2003-06-06）（DUX '03）。计算机协会，1–15。doi:10.1145/997078.997089[45] Vinay Samuel、Henry Peng Zou、Yue Zhou、Shreyas Chaudhari、A. Kalyan、Tan- may Rajpurohit、A. Deshpande、Karthik Narasimhan 和 Vishvak Murahari。2024。PersonaGym：评估角色智能体与大语言模型。载于自然语言处理实证方法大会。[46] Shibani Santurkar、Esin Durmus、Faisal Ladhak、Cinoo Lee、Percy Liang 和 Tatsunori Hashimoto。2023。语言模型反映谁的观点？载于第40届国际机器学习大会论文集。[47] Marko Sarstedt、Susanne J. Adler 和 Christian M. Ringle。2024。将大语言模型用作硅样本：实用指南与最佳实践。《心理学方法》（2024）。doi:10.1002/mar.21982[48] Albrecht Schmidt、Passant Elagroudy、Fiona Draxler、Frauke Kreuter 和 Robin Welsch。2024。用 ChatGPT 在以人为本的设计中模拟人类：借助人工智能重新设计交互设计。《Interactions》 31, 1 2024 24 31 10 1145 3637436 ( ), – 。doi: . / Yunfan Shao、Linyang Li、Junqi Dai 和 Xipeng Qiu。2023. Character-LLM：用于角色扮演的可训练智能体。载于自然语言处理实证方法大会。[49] Disha Sharma。2025。A/B 测试的统计显著性：如何以及何时结束测试。https://www.convert.com/blog/a-b-testing/experiments-statistical-significance-speed/。访问日期：2026年1月13日。[50] Donghoon Shin、Daniel Lee、Gary Hsieh 和 Gromit Yeuk-Yin Chan。2025。PosterMate：用于海报设计的受众驱动的协作式角色智能体。载于第38届 ACM 用户界面软件与技术年度研讨会论文集（美国纽约，2025-09-27）（UIST '25）。计算机协会， 1 20 10 1145 3746059 3747769 – 。doi: . / . Joongi Shin、Michael A. Hedderich、Bartłomiej Jakub Rey、Andrés Lucero 和 Antti Oulasvirta。2024。理解用于生成用户画像的人机协作工作流。载于2024年 ACM 交互系统设计大会论文集（美国纽约，2024-07-01）（DIS '24）。计算机协会，757–781。doi:10.1145/3643834.3660729[51] Aleksandrs Slivkins。2019。多臂老虎机导论。《机器学习基础与趋势》。doi: 10 1561 2200000068 . / Sangho Suh 、 Michael Lai 、 Kevin Pu 、 Steven P. Dow 和 Tovi Grossman。2025。StoryEnsemble：利用人工智能和前向—后向传播，在设计流程中实现动态探索与迭代。载于第38届 ACM 用户界面软件与技术年度研讨会论文集（美国纽约，2025-09-27）（UIST '25）。计算机协会，1–36。doi:10.1145/ 。3746059 3747772[52] Lu Sun、Shihan Fu、Bingsheng Yao、Yuxuan Lu、Wenbo Li、Hansu Gu、Jiri Gesi、Jing Huang、Chen Luo 和 Dakuo Wang。2025。大语言模型智能体遇上智能体人工智能：大语言模型智能体能否模拟客户以评估基于智能体人工智能的购物助手? arXiv:2509.21501 doi:10.48550/arXiv.2509.21501[53] Zhaoxuan Tan 和 Meng Jiang。2023。大语言模型时代的用户建模：当前研究与未来方向。*IEEE*数据工程简报（2023）。[54] Olivier Toubia、George Z. Gui、Tianyi Peng、Daniel J. Merlau、Ang Li 和 Haozhe Chen。2025。数据库报告：Twin-2K-500：一个基于 2,000 多人对 500 多个问题的回答构建数字孪生的数据集。*Marketing Science* 44, 6 2025 1446 1455 10 1287 2025 0262 （），–。doi:。/mksc.。2024[55] Browser Use。。Browser Use：人工智能浏览器自动化。https://github.com/browser-use/browser-use[56] Angelina Wang、Jamie Morgenstern 和 John P. Dickerson。2025。取代人类参与者的大语言模型可能会有害地误表征并简化身份群体。arXiv: 2402 01908 [ ] 10 48550 2402 01908 . cs doi:。/arXiv.。[57] Dakuo Wang、Ting-Yao Hsu、Yuxuan Lu、Hansu Gu、Limeng Cui、Yaochen Xie、William Headean、Bingsheng Yao、Akash Veeragouni、Jiapeng Liu、Sreyashi Nag 和 Jessie Wang。2025。AgentA/B：借助交互式大语言模型智能体实现自动化、可扩展的网页 A/B 测试。arXiv: 2504 09723 [ ] 10 48550 2504 09723 . cs.HC doi:。/arXiv.。[58] Lei Wang、Jingsen Zhang、Hao Yang、Zhi-Yuan Chen、Jiakai Tang、Zeyu Zhang、Xu Chen、Yankai Lin、Hao Sun、Ruihua Song、Xin Zhao、Jun Xu、Zhicheng Dou、Jun Wang 和 Ji-Rong Wen。2025。基于大语言模型的智能体进行用户行为模拟。*ACM*信息系统汇刊43，2（2025），55:1–55:37。doi:10.1145/3708985[59] Zhefan Wang、Ning Geng、Zhiqiang Guo、Weizhi Ma 和 Min Zhang。2025。面向任务的对话中，人类与智能体的比较。载于 *2025* 年亚太地区 *ACM SIGIR* 信息检索研究与发展年度国际会议论文集。[60] Ziyi Wang、Yuxuan Lu、Wenbo Li、Amirali Amini、Bo Sun、Yakov Bart、Weimin Lyu、Jiri Gesi、Tian Wang、Jing Huang、Yu Su、Upol Ehsan、Malihe Alikhani、Toby Jia-Jun Li、Lydia Chilton 和 Dakuo Wang。2025。OPeRA：用于评估大语言模型模拟人类在线购物行为的数据集，涵盖观察、用户画像、理由和行动。[61] Ziyi Wang、Yuxuan Lu、Yimeng Zhang、Jing Huang 和 Dakuo Wang。2025。Customer-R1：通过基于强化学习的大语言模型智能体对在线购物中的人类行为进行个性化模拟。[62] Ian Waudby-Smith、David Arbour、Ritwik Sinha、Edward H Kennedy 和 Aaditya Ramdas。2024。时间一致中心极限定理与渐近置信序列。*The Annals of Statistics* 52，6（2024），2613–2640。[63] 维基媒体基金会。2010。2010年筹款：示例测试2。https://meta. wikimedia.org/wiki/Fundraising_2010/Report/Example_Test_2。访问日期：2026- 01-19。[64] 维基媒体基金会。2011。2011年筹款：测试更新（10月21日）。https: //meta.wikimedia.org/wiki/Fundraising_2011/Test_Updates/October/21。访问日期：2026-01-19。[65] An Zhang、Leheng Sheng、Yuxin Chen、Hao Li、Yang Deng、Xiang Wang 和 Tat-Seng Chua。2023。论推荐系统中的生成式智能体。载于*ACM*信息检索研究与发展年度国际会议。[66] Junjie Zhang、Yupeng Hou、Ruobing Xie、Wenqi Sun、Julian McAuley、Wayne Xin Zhao、Leyu Lin 和 Ji-rong Wen。2023。AgentCF：利用自主语言智能体为推荐系统进行协同学习。载于万维网大-会。[67] Ben Zhao。2025。使用汤普森采样优化在线广告中的点击率。*ITM Web of Conferences* 73（2025），01012。doi:10.1051/ itmconf/20257301012[68] Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023。使用 MT-Bench 和 Chatbot Arena 评判大语言模型评审。载于神经信息处理系统进展。[69] Qingxiao Zheng、Minrui Chen、Pranav Sharma、Yiliu Tang、Mehul Oswal、Yiren Liu 和 Yun Huang。2025。EvAlignUX：通过大语言模型支持的指标探索推进用户体验评估。载于*2025*年*CHI*人机交互大会论文集（*CHI '25*）。计算机协会，纽约，纽约州，美国， 1 25 10 1145 3706598 3714045 –。doi:。/。[70] Shuyan Zhou、Frank F. Xu、Hao Zhu、Xuhui Zhou、Robert Lo、Abishek Sridhar、Xianyi Cheng、Yonatan Bisk、Daniel Fried、Uri Alon 和 Graham Neubig。2024。WebArena：用于构建自主智能体的真实网页环境。载于国际学习表征会议。[71] Lixi Zhu、Xiaowen Huang 和 Jitao Sang。2024。你的模拟器有多可靠？分析当前基于大语言模型的对话式推荐用户模拟器的局限性。载于*2024*年万维网大会附属论文集。doi:10.1145/3589335.3651489

A 提示词

本节详细介绍 SimAB 中使用的核心提示。虽然我们的实验评估和提示主要面向电子商务环境，但初步观察表明，这些提示也能有效泛化至其他领域，例如营销活动或捐赠横幅。不过，由于这些领域缺乏历史真实数据，我们未将其纳入定量分析。

A.1 用户画像生成提示词模板

以下模板展示了用于生成用户画像的提示词结构。大括号中的变量会针对每次评估动态填充。

SYSTEM_PROMPT = """ 你是一位用户研究和人物画像开发专家，负责为电子商务 A/B 测试创建逼真的用户画像。G

enera te divers以及代表可能访问该网站的不同用户类型的用户画像。"" "

限制_模板 = """此外，上述字段还存在一些必须遵守的限制。**重要**：{ restrictions } """

用户_提示词_图像_模板 = """分析图像中的网站，并生成{ num_个用户画像 }个可能访问该网站的详细目标受众用户画像。

请以 JSON 格式为每个用户画像提供以下信息：- name：描述性用户画像名称 - 年龄_范围：年龄范围（例如，“25 -35”、“18 - 24”、“45 -60”）- 职业：职业角色或背景 - 收入_水平：收入档次（例如，“低”、“中等”、“高”、“中上”）- 教育程度：教育水平 - 所在地：地理位置或类型 - 兴趣：主要兴趣和爱好 - 目标：主要目标和动机 - 痛点_：挑战和挫折 - 技术_熟练度：对技术的适应程度（低/中等/高）- 在线_行为：通常如何使用互联网 - 任务：列出他们希望在该网站上执行的 3-5 项具体任务 - 背景：当前会话背景——用户从何处而来以及当下的具体情况

{ maybe_限制 }

生成具有显著多样性的用户画像，涵盖以下方面：隐私取向（严格 <-> 宽松）、对初始阻力的容忍度（低 <-> 高）、时间敏感度（急迫 <-> 悠闲）以及对个性化的接受程度（低 <-> 高）。

- 用户画像通过不同渠道接触产品（社交媒体、广告、浏览、大语言模型，即自然流量、自有流量和付费流量）。大多数人对你的产品感兴趣，但并非所有人都已做好立即转化的准备，也就是说，浏览网站会影响他们的行为。

- 用户画像涵盖随意浏览、产品调研和购买意向等不同阶段；年龄、风险承受度、决策果断程度和技术素养也各不相同。

- 用户画像 m ight reject a ds, cookies, 提示 ions if they 已经结束 ly 激进 ive或 fishy.
- 不要假设各用户画像的行为是一致的；确保该群体反映出上文所述的全部态度范围，并反映你实际用户群的分布。

- 从更广阔的背景来思考，不要生成这样的用户画像：其所有痛点、任务和目标都仅仅精确针对该产品或产品类别，而应与该产品相关。确保用户画像拥有超出该产品之外的生活与需求。

请聚焦于真正会访问此类网站的真实用户画像。请基于所提供的实际内容进行分析，

considering the products, services, messag ing, and 目标 audience indicated by the website内容 .

作为附加上下文，你拥有以下从文档中检索到的、可能与角色生成相关的片段：{ context }

将响应作为有效的 JSON 角色对象数组返回。"""

A.2 角色模拟提示模板

以下模板展示了结构 用于角色模拟的提示词结构。花括号中的变量会在每次评估时动态填充。

SimAB：使用角色画像条件化 AI 智能体模拟 A/B 测试，以实现快速设计评估

系统提示：你是一位人工智能专家，能够通过特定用户画像的视角分析电子商务网站，预测其转化潜力。

你将看到两个版本的电子商务网页，需要预测哪个版本更可能促使你的用户画像产生转化，也就是为公司带来更多收入。如果该用户画像对两个页面都不感兴趣、不愿购买，你也可以选择两个版本都无法促成购买。

{ EvaluationPrompts . 转化_目标_模板 (转化_目标)}

{ EvaluationPrompts . 价值_假设 (假设)}

你应该从具备以下特征的用户的角度评估这些页面，这些特征是**你的主要特征**：{ persona_字符串 }

这两个版本分别标记为“{ LLMPref. 对照组}”和“{ LLMPref. 挑战组}”。呈现顺序是随机的，标签并不代表质量。请严格根据你看到的内容做出预测。

用户画像会浏览并感兴趣，但并不总是立即有需求。- 可能只是随意浏览，并花时间做出购买决策。- 如果网站不够吸引人，尤其是在内容美感和用户友好性方面表现欠佳，他们可能会离开。- 如果他们离开，会说明原因。

请从即时转化因素和长期客户满意度两个方面评估这些电子商务页面：

即时转化因素： - 定价和后续步骤是否清
晰 - 完成购买是否便捷 - 尽量减少阻力和困惑

- 清晰的定价和配送信息 - 移动端友好性和响应式设计 - 搜索和筛选功能 -
整体购物体验和转化潜力

长期信任与满意度因素：
- 对费用和承诺保持透明 - 清晰的取消和退款政策 - 法律
合规与信任信号 - 清晰的订阅管理

** 关键成功因素（按重要性排序）：**
1. ** 转化可能性 **：该用户画像是否真的会完成购买？2. ** 认知负荷 **：复杂性会阻碍还是促成决策？3. ** 信任 vs. 阻力 **：额外细节是建立必要的信任，还是造成分析瘫痪？4. ** 用户画像 - 特定需求 **：这个特定用户真正关心什么？

注意： 收入 /年度经常性收入/和 诸如此类都由转化 * 价格决定。也就是说，要确定收入 /AR R，
你需要乘 ly将转化乘以所购产品的价格。

决策框架： 1. Wh
at is this 用户画像的 首要 关注点? （价格, 功能, 合规性, 速度, 等等。)
2. 他们的 决策 风格? （快速／冲动，还是 审慎／分析型）
3. 什么会让 make 他们放弃 the 购买? （复杂性, 不清楚 定价, lack 详情）
4. 哪个 版本 更符合 他们的具体关注点，并 风格?
转化心理学指南：
- 只有当用户画像需要更多信息才能感到有信心时，更多信息才更好 - 当用户画像对价格敏感、时间紧迫或排斥技术时，简洁更胜一筹 - 当用户画像规避风险、注重合规性或正在做出企业决策时，复杂性更胜一筹 - 考虑选择瘫痪：过多选项会降低转化 - 考虑认知负荷：复杂布局可能让某些用户画像不堪重负

- 在评估中考虑色彩的心理效应
示例场景：
- 注重预算的用户通常更喜欢清晰、简单的定价，而非详细的功能比较 - 时间紧迫的用户可能会放弃复杂的决策树，转而选择显而易见的选项 - 非技术用户可能会被详细的功能列表弄得不知所措 - 个人买家可能会觉得团队功能是无关的累赘 - 学生和初创企业通常更重视速度和较低的承诺，而非全面信息

{ EvaluationPrompts . 检索增强生成_背景_模板 (检索增强生成_背景)}

请考虑到，某个版本若能建立起足以克服略微增加的复杂性的信任，或能防止购买后的后悔与取消订单，则可能实现更好的转化。

注意：请记住，你是一个人类用户画像，高度重视视觉与主观推理，并遵循你的用户画像原则。但仍切勿忽视购物体验的任何其他方面。请选择对该特定用户画像而言更可能发生转化的版本；若该用户画像不会从任一版本进行购买，则选择 '{ LLMPref .NONE}'。

注意：对用户画像而言“更好”，是基于该用户画像并不了解两个网站版本、而只了解其中单一版本这一假设。这意味着，举例来说，若用户也会购买更贵的版本，因为它确实值这个价格，那么更便宜的版本对公司而言未必是更好的选择。

重要：部分屏幕截图可能包含已隐去的价格或未知价格（XX.XX）。请将其替换为该市场中合理的价格。它们并非在遮掩，也并非不可信，而是尚未填写（草稿）或已被隐去。它们不应使你对网站的看法变好或变坏。

用户提示：请仔细分析这两个电商页面，考虑给定的用户画像是否会在任一页面购买，并从产品展示、可信度、购物便捷性、价格透明度和整体用户体验等方面进行评估。请选择更有可能促成转化的版本（必须严格为“{ LLMPref . CONTROL }”或“{ LLMPref . CHALLENGER }”）；如果该用户画像不会在任何一个版本中购买，则选择“{ LLMPref . NONE }”。请根据符合该用户特征的具体电商因素进行推理。

重要：预测哪个版本能让这一特定用户画像带来更多实际转化。请考虑该用户画像更可能因复杂性还是信息不足而放弃购买。不同的用户画像对详细程度与简洁性的容忍度不同。

A.3 摘要生成提示模板

以下模板展示了用于摘要生成的提示结构。花括号中的变量会在每次评估时动态填充。

分析 以下 评估两个电商网站版本的人工智能智能体的反馈 .
Each 智能体 选择对照组, 挑战组, 或无（会 不购买 来自任一）。

对照组被选中 { control_count } 次，挑战组被选中 { challenger_count } 次，无被选中 { none_count } 次。

重要：获胜者是 {winner }。

请找出他们在决策过程中的主要模式和主题，并提供可执行的洞察，以改进挑战组。

{ feedback_text }

提供：。
1
简要总结智能体为何更偏好某个版本。只写最主要的一个原因，不要开场白或结束语。用一个句子，不要用逗号，也不要包含从句。例如：对照组胜出，因为……或挑战组更受青睐…… 2. 智能体选择对照组的首要 1-2 个原因（如果对照组落败）或 3-5 个原因（如果对照组胜出）（如有） 3. 智能体选择挑战组的首要 1-2 个原因（如果挑战组落败）或 3-5 个原因（如果挑战组胜出）（如有） 4. 智能体选择不购买任一版本的首要 1-2 个原因（如有）。如果只有少数智能体选择「无」，则留空（[]）。 5. 本次实验的首要 1-3 条可执行的洞察（如有）。如果没有智能体选择对照组，可以留空。如果有智能体选择对照组，则提供可用于改进挑战组的可执行洞察。也就是说，哪些方面可以改进，让挑战组更受青睐；即使挑战组以并非压倒性的优势胜出，也要说明「哪些方面可以改进，让结果更加显著？」。

重点关注电子商务因素，例如：产品展示、可信度、易用性、价格透明度、导航、移动端友好性和转化优化。

A.4 检索增强生成提示模板

以下模板展示了 用于输入处理中的检索增强生成部分的提示结构。花括号中的变量会在 for
每次评估。

HYDE_系统_提示词 =("你是一位专业撰稿人，负责生成简短、客观的段落，" "这些段落*可能合理地回答*给定的查询。"

```
    "写一个单一的，连贯的段落（3-5 句话）that 直接"
    "and 自信地回答查询，仿佛 you 知道正确答案。”
    “不要 ask 提问，list 选项，或提及不确定性。"
    "请勿加入元评论或指示。"
)
HYDE_用户_提示词_模板 = "查询：{query }"

用户画像_RAG_查询 _模板 = "哪些用户画像可能会对具有以下文本描述的网站感兴趣
    ：{ description }？" 实验_检索增强生成_查询_模板 ="对于以下详情，哪些信息有助于评估 A/B 测试：转化目标：{ conversion_goal }。
变体的图像描述：{ image_descriptions }" SQL_查询_系统_提示词 = ( "你是一名专精于 A/B 测试的 SQL 分析助手。" "你的工作流程分为两
步：\n" "1. 首先，根据提供的实验上下文和数据集摘要，编写一条 DuckDB SQL 查询（使用名为 'df' 的表），以呈现有关用户行为、业务指
标和当前趋势的相关背景。此步骤的输出应*仅包含* SQL 查询（不含注释或其他文字）。\n" "2. 查看查询结果后，编写一段简短易懂的摘要，
为评估实验的用户画像提供背景信息。\n" "查询步骤指南：\n" "- 使用 DuckDB 语法编写清晰、高效的 SQL 查询。" "- 着重查询能揭示用户行
为模式、业务趋势或相关背景的聚合、分群和筛选条件。\n" "- 查询有助于了解当前状况及实验重要性的指标和维度。\n" "- 始终从 df 查询。
\n" "- 使用描述性列别名。\n" "- 对探索性查询或结果量较大的查询，酌情加入 LIMIT 子句。\n" "- 输出*仅限* SQL 查询，不要附加解释。\n"
"摘要步骤指南：\n" "- 面向将评估此实验的非技术用户画像/用户，而非数据科学家或实验设计师。\n" "- 提供业务背景、用户行为洞察和当前
状况信息。\n" "- 着重说明“正在发生什么”，而非“如何设计测试”。\n" "- 使用通俗易懂的语言，避免使用“CV”“协变量调整”“统计功效”“样本
量”等统计术语。\n" "- 简洁明了（最多 3–5 句话），同时包含有用信息。\n" "- 帮助用户画像了解数据所显示的情况为何与此实验相关。\n" "-
不要提供实验设计建议、统计方法或技术指导。\n" "- 将洞察表述为可观察到的事实或趋势，而非可用于测试设计的行动建议。\n\n" '良好摘要
示例：“数据显示，过去一年每周订阅收入大幅增长，平均每周收入从接近零增至约 6 万美元。新增订阅者数量与收入增长密切相关，最近几
周的周环比波动为 10%。页面访问流量也与订阅量相关，这表明访客数量和转化都对增长很重要。”\n\n' '不良摘要示例：“将新增订阅总数作
为主要指标。考虑到 CV 为 66%，应延长测试时长。考虑按周进行区组随机化，以避免上升趋势造成混杂。使用协变量调整来降低方差。”
\n\n' "系统会始终先提示你编写查询，然后在查看查询结果后提示你编写摘要。" )

SQL_查询_用户_提示_模板 =( " 根据以下背景，编写 { num_queries } 条 SQL 查询，以提供有关该实验的洞察。\n" " 实验上下
文：\n" "- 主要转化目标：{ conversion_goal }\n" "- 假设：{ hypothesis }\n" " 数据集描述：\n" "{ dataset_description }" " 受
众：\n" "{ persona_restrictions }" )
```

表3：为创意软件定价页面生成的示例用户画像。

属性	用户画像1	用户画像2	用户画像3
Name	“创意总监 Carlos”	“自由设计师 Priya”	“学生创作者 Alex”
Age	38–45	28–32	19–22
职业	广告公司的创意总监	自由平面设计师	大学生，设计专业
收入	高（$120k+）	中等（$50–70k）	较低（兼职工作）
目标	团队许可证、可靠支持	高性价比个人方案	学生折扣、基本功能
痛点	复杂的采购流程	收入不稳定，需要灵活性	预算有限，被各种选项
精通技术	High	High	中等
背景	为15人团队评估工具	在年度续订前进行比较	为课堂项目进行研究

B 用户画像生成示例

表3展示了为软件定价页面评估生成的三个用户画像示例。