arXiv:2603.05780v1 [cs.IR] 6 Mar 2026

平衡国内与全球视角：评估双重校准与大语言模型生成的助推在多元化新闻推荐中的应用

Ruixuan Sun*
GroupLens Research，明尼苏达大学
明尼苏达州
美国明尼苏达州明尼阿波利斯
美国

Matthew Zent*
GroupLens Research，大学
明尼苏达州
明尼阿波利斯，明尼苏达州，美国

Minzhu Zhao
GroupLens Research，明尼苏达大学
明尼苏达州
明尼阿波利斯，明尼苏达州，美国
美国

Thanmayee Boyapati
GroupLens Research，明尼苏达大学
明尼苏达州
明尼阿波利斯，明尼苏达州，美国

Xinyi Li
西北大学
埃文斯顿，伊利诺伊州，美国

Joseph A. Konstan
GroupLens Research，明尼苏达大学
明尼苏达州
明尼苏达州明尼阿波利斯，美国
州

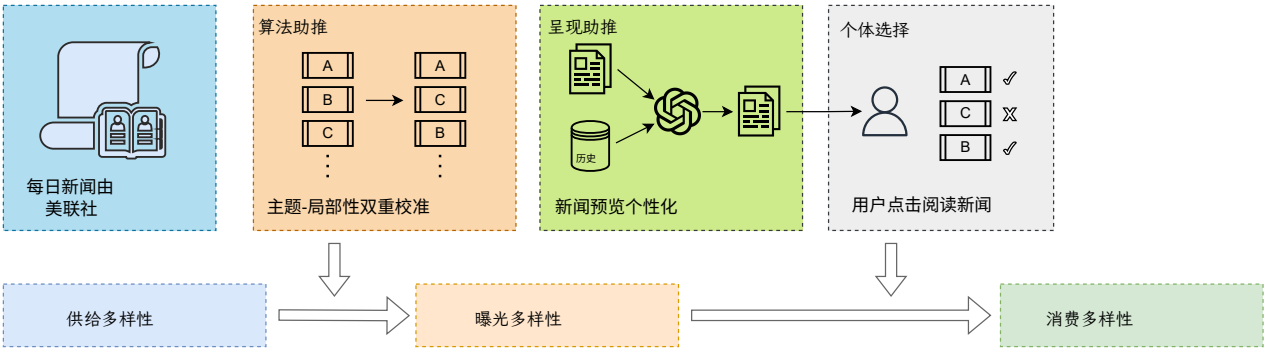


图1：概览　　我们对 Mattis 等人提出的个性化多样性助推框架的应用 [42]　　.

摘要

在本研究中，我们应用了“个性化多样性助推框架”，旨在扩大用户在新闻地域性方面的阅读覆盖范围（即国内新闻和世界新闻）。我们设计了一种新颖的主题-地域双重校准算法助推，以及一种基于大语言模型的新闻个性化呈现助推，并在新闻推荐实验平台 POPROX 上，对120名美国新闻读者开展了为期5周的真实用户研究。通过分析用户交互日志和调查回应，我们发现，算法助推能够有效提升接触多样性和消费多样性，而基于大语言模型的呈现助推效果则不尽相同。用户层面的主题兴趣是预测用户点击量，同时突出新闻文章与先前阅读的文章之间的相关性，其表现优于通用的基于主题的方法和无个性化方法。我们还证明，纵向接触校准后的新闻可能会改变读者的阅读习惯，使其重视同时涵盖国内和世界文章的平衡的新闻摘要。我们的研究结果为新闻推荐系统中助推多样化消费的未来工作提供了方向。

CCS 概念

• **以人为本的计算 → 人机交互 (HCI)**；• **信息系统 → 用户与交互式检索**；**推荐系统**；•**计算方法 → 自然语言生成**。

关键词

新闻推荐、推荐校准、内容个性化

ACM 参考格式：
Ruixuan Sun、Matthew Zent、Minzhu Zhao、Thanmayee Boyapati、Xinyi Li 和 Joseph A. Konstan。2018。平衡国内与全球视角：评估双重校准和大语言模型生成的助推在多元化新闻推荐中的应用。载于请确保输入正确的会议名称

*两位作者对本研究的贡献相同，并通过抛硬币决定作者署名顺序。

我们授予许可，允许为个人或课堂使用而免费制作本作品全部或部分内容的数字副本或纸质副本，但不得为营利或商业利益制作或分发副本，且副本首页必须附上此声明和完整引文。本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。*Conference acronym 'XX，2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2018 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-XXXX-X/18/06https://doi.org/XXXXXXX.XXXXXXX

来自版权确认邮件的标题（*Conference acronym 'XX*）。ACM，美国纽约州纽约市，14 页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

2023年，美联社（AP）向约4400万月活用户分发了超过37.5万篇报道 [2]。为应对信息过载，个性化新闻推荐系统（推荐系统）已无处不在，它们利用用户过去的消费行为，筛选出符合用户独特兴趣的内容 [30, 46]。尽管这种方法成功提升了准确性 [75] 和用户参与度 [33], , 却有可能形成强化狭隘偏好的反馈回路 [37], , 生成不公平的推荐 [66], , 并将用户困在过滤气泡中 [34]。这造成了个性化的短期益处与促进公众充分了解信息这一长期社会目标之间的矛盾。

为解决这一矛盾，先前研究将新闻多样性划分为三个不同阶段：**供给多样性**（出版商提供的内容）、**曝光多样性**（用户可见的内容）和**消费多样性**（用户主动参与的内容）[36, 42, 63]。由于供给多样性由编辑决策决定，推荐系统研究主要侧重于优化曝光多样性，以减轻偏见 [6, 59]。然而，如何在实践中将多样化推荐转化为多样化消费，仍不明确。实证研究表明，即使用户接触到多元观点，也往往会忽略那些看起来无关或难以理解的内容 [21, 29]。为弥合这一差距，近期研究提出了个性化多样性助推的理论框架，其目标不仅是展示多样化内容，也要促进用户消费这些内容 [42, 69]。

在这项工作中，我们基于 Mattis 等人 [42] 的框架，通过实证检验两种新颖的助推干预：双重校准（一种旨在提升接触多样性的算法助推）和大语言模型生成的呈现（一种旨在影响消费的呈现助推）。首先，在接触多样性方面，我们将多样性的范畴从传统的政治和意识形态 [18, 38] 维度扩展到新闻地域性。新闻学研究强调平衡全球意识与本地相关性的重要性 [11, 39], , 但推荐系统研究很少以实现这种平衡为优化目标。我们提出一种双重校准方法，同时平衡个性化的**主题**（如体育、政治、教育等）与多样化的**地域**（即国内和世界），以确保美国用户能够接触到与自身相关的全球和本地视角。

其次，为了将这种接触多样性转化为实际消费，我们借助大语言模型（LLM）个性化新闻呈现。我们改写新闻预览（即标题和副标题），在因双重校准中的局部性组件而被推荐的文章与用户此前的新闻阅读兴趣之间建立联系，同时不改变文章的事实内容。这种方法旨在通过降低用户接触陌生主题或地域时的认知成本，助推用户以更平衡的方式消费新闻。

我们通过在新闻推荐平台 POPROX 上开展为期5周的纵向研究，评估这一方法的效果。基于此，我们回答以下研究问题：

研究问题1 地域性 +主题双重校准 与仅按主题校准的策略相比，对曝光多样性有何影响？

研究问题2 与仅按主题校准的策略相比，双重校准如何影响**消费多样性**？ **研究问题3 大语言模型生成的呈现助推** 如何影响曝光多样性向消费多样性的转化？ **研究问题4 双重校准 + 大语言模型助推** 如何影响用户参与度和主观满意度？

我们在这一领域的贡献有三方面：

(1) 我们提出了一种新颖的多样性感知推荐系统，将概率双重校准（用于算法助推）与基于大语言模型的内容重构（用于呈现助推）相结合，并通过纵向实验评估了该系统。
(2) 我们通过实证展示了双重校准对促进国内与国际新闻多样化消费的影响，并确立了**地域性**作为新闻多样性的一项有价值的主题内维度。(3) 我们指出，基于事件的叙事重构是呈现助推的一种有前景的方向，并提出了一条影响用户选择性的路径。

本文其余部分将首先讨论相关研究。随后，我们介绍 POPROX、我们的校准方法和大语言模型个性化方法，接着说明实验设计。然后，我们分享离线训练和在线随机对照实地研究的结果。最后，我们讨论研究发现、局限性以及未来研究方向。

2 相关工作

2.1 新闻推荐多样性

2.1.1 个性化多样性。 个性化多样性方法利用用户的历史行为，使推荐与其多样化兴趣相一致，以克服标准方法中的刻板归类，例如采用多头自注意力的新闻推荐方法（NRMS） [68]。主题校准 [3, 57, 64] 利用比例，使推荐分布与跨主题的消费模式相一致。深度神经网络方法（即异构图 [73] 和多视图表示学习 [13]）能够随时间推移，针对用户的动态偏好实现更丰富的个性化。

2.1.2 算法多样性助推。 算法多样性助推会有意使推荐分布偏向多样性/公平性目标 [23, 50, 69], , 从而引入准确性权衡 [48]。一些方法利用主题级距离，通过基于分类体系的 [49] 和频率度量 [12] 对推荐进行正则化，以奖励多样化的接触多样性。另一些方法则以主题内多样性度量（即语气 [62], 、情感 [70], 和政治倾向 [21]）为目标，以实现类似目标。在新闻语境之外，研究还探索了使用大语言模型进行基于多样性的重排序，效果不一 [10]。

我们的工作结合了这两种方法，将主题校准与非个性化地域校准相结合，从而实现个性化的算法多样性助推。

2.2 新闻呈现助推

新闻呈现助推具有多种不同的形式和目标。已有研究使用推荐前或推荐后的助推，通过社会比较 [4]和

自我反思 [44]。其他研究则通过将推荐与同一报道的替代观点配对来助推，以减轻偏见 [47]。受大语言模型在营销方面说服成效的启发 [5] 和新闻摘要 [24],，近期研究聚焦于改变新闻文章本身以减轻偏见 [51, 52], 使内容去极化 [35], 并应用正向/负向重构 [25]。与我们的工作最为接近，Gao等人 [17] 将生成式新闻推荐与主题级和语义级表示相结合，在推荐文章之间创建个性化叙事。

总体而言，这些算法助推和呈现助推方法中的许多方法都在离线评估中取得了成功。然而，很少有研究开展用户研究，更少有研究在现实情境中评估这些助推干预的影响 [42]。我们的研究建立在对多样性感知新闻推荐系统 [25], 长期效应的有限理解之上；鉴于先前研究表明新闻多样性有益于平台长期增长 [34, 66]，这一点尤为重要。

3 POPROX平台

我们在 POPROX [9], 上开展了这项研究。这是一个实验平台，维护着美国新闻读者群体，并每天向他们分发由美联社（AP）提供的个性化新闻简报。研究人员可以设计不同的算法或界面，以改变用户的新闻简报内容，并利用交互日志和调查数据评估其有效性。POPROX 还提供了一套基础推荐算法和应用程序接口，供实验人员集成和修改。

内部试点。我们开展了两组内部试点，以在实验开始前完善算法设计并验证系统功能[1]。在开发过程中，我们与*24* 名内部用户持续开展试点，这些用户由 POPROX 提供，代码更改会推送到他们的 POPROX 每日新闻简报中。在此过程中，我们通过迭代反馈改进了 4.1 和 4.2，并据此调整阈值和提示词。完成这些优化后，在正式实验启动前，我们开展了为期一周的暗场 试运行，每个处理组安排 8 名用户，以验证实验基础设施。

4 方法

整体方法分为两部分：1）基于主题校准（TC）和主题-局部性双重校准（DC）对推荐结果重新排序；2）对双重校准处理组中经重新排序、但未被基线主题校准算法推荐的新闻进行新闻预览个性化（DC-SC）（见图 2）。以下小节将详细说明各个组成部分。在深入介绍之前，我们先说明关键符号：令 $A = \{a_1 \ldots, a_n\}$ 表示新闻提供商今日发布的文章集合，H 表示用户此前点击过的文章集合。令 I_{10} 和 J_{10} 分别表示经 TC 和 DC 重新排序的排名前十文章集合。个性化新闻预览应用于属于 I但不属于 J的文章，我们将此集合定义为 $S = \{s_1, \ldots\ldots, s_k\}$。

4.1 校准

为了使一组推荐内容在特定维度上符合用户的消费偏好，我们借鉴了 Steck [57] 所正式提出的校准框架。对于新闻，主题校准意味着，如果用户点击过 20 篇政治文章和 80 篇体育文章，其个性化新闻简报中政治文章和体育文章的比例也应分别约为 20% 和 80%。然而，先前研究指出了这种做法的风险，尤其是它可能强化已有的用户偏见。对此，研究人员提出了各种双重优化策略，以在校准推荐与更广泛的系统目标之间取得平衡 [65, 74]。

沿用先前研究的做法 [45, 57],，我们采用 Kullback-Leibler（KL）散度作为校准指标，沿两个维度对推荐结果重新排名：主题（见公式 1）和地域性（见公式 2）。设$p(t \mid u)$表示用户 t曾点击过的主题 u的分布。类似地，$q(t \mid u)$表示用户新闻简报中的主题分布。设 $p(l \mid u)$表示l 新闻提供商今天发布的文章中的地域分布，而 $q(l \mid u)$ 表示用户新闻简报中的地域分布。

$$C_{KL}^{t}(p, q) = KL_t(p \,\|\, q) = \sum_t p(t \mid u) \log \frac{p(t \mid u)}{q(t \mid u)} \tag{1}$$

$$C_{KL}^{l}(p, q) = KL_l(p \,\|\, q) = \sum_l p(l \mid u) \log \frac{p(l \mid u)}{q(l \mid u)} \tag{2}$$

推荐的新闻简报 I_{10}和 J_{10}通过分别优化公式 3 和公式 4，逐步填入今天的文章：

$$i_{n+1} = \arg \max_{i \in A \setminus I_n} (1 - \theta_l - \theta_t) S_{nrms}(i) - \theta_t C_{KL}^{t}(p, \tilde{q}) \tag{3}$$

$$j_{n+1} = \arg \max_{j \in A \setminus J_n} (1-\theta_l-\theta_t) S_{nrms}(j) - \theta_l C_{KL}^{l}(p, \tilde{q}) - \theta_t C_{KL}^{t}(p, \tilde{q}) \tag{4}$$

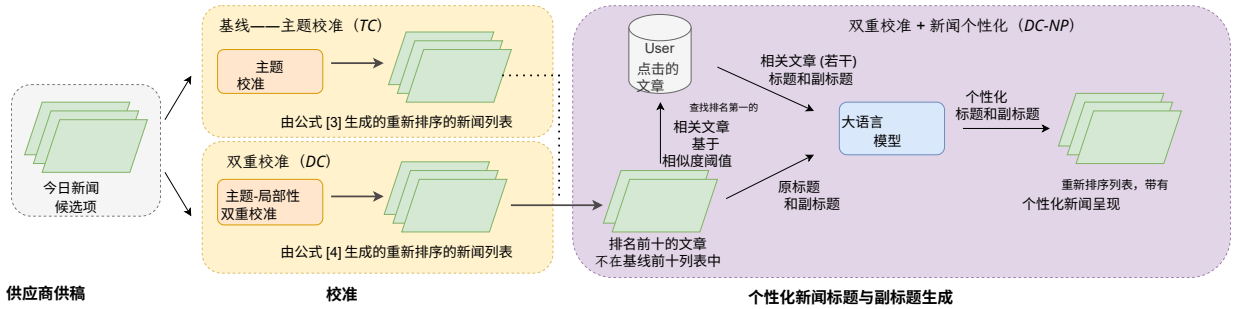
其中 θ_l，$\theta_t \in [0, 1]$和 $S_{nrms}(I) = \sum_{i \in I} s_i$表示由基础 NRMS [68] 提供的 I 中新闻简报的用户偏好预测得分。

在主题校准方面，我们的实现采用了 POPROX 提供的 14 个现有高层主题，这些主题源自美联社的主题元数据，可用于用户偏好建模。[2] 在地域性校准方面，我们进一步调整美联社的主题元数据，以定义文章的地域分布：**国内**、**世界**和**两者皆非**。具体而言，我们利用美联社的一般新闻 类别，将带有美联社标签美国新闻 或华盛顿新闻的内容定义为国内新闻，将带有美联社标签世界新闻的内容定义为世界新闻；若不符合上述任一分类，则归为两者皆非。[3] 文章通常带有多个美联社主题标签，这意味着它们可以涵盖多个主题，也可以同时归类为国内和世界地域。我们承认，这是由内容提供方根据美联社新闻内容的性质所作的选择。这一特点也使地域性成为主题内多样性目标，因为美国、华盛顿或世界新闻标签通常会与其他主题标签交叉。

4.1.1 校准调优。为了高效地找到主题（θ_t）与地域性（θ_l）组合的最优校准项，我们采用了随机搜索 [7]。搜索空间定义为 θ_t与 θ_l在 [0,1] 范围内、步长为 0.05 的笛卡尔积。按照[7],，我们随机抽取了 60 种配置，这使得我们有 95% 的概率在搜索空间的前 5% 中找到解

[1]感谢 POPROX 研究人员的支持。

[2]POPROX 的高层新闻主题标签列表改编自美联社广泛主题元数据的一个子集，[1] 包括：美国新闻、世界新闻、政治、商业、娱乐、体育、健康、科学、技术、生活方式、宗教、气候与环境、教育、奇闻。[3]美联社对受众关注的本地华盛顿新闻使用华盛顿新闻 标签，而非美国新闻 标签 [1]。我们将这两个标签合并，以避免此类新闻被归为两者皆非。实际上，两者皆非通常指不具备地理关联的一般新闻报道。



Figu图2：校准与个性化新闻预览生成流程的方法概览 .

。选取最优对（θ_t^*，θ_l^*）以最大化效用得分 S。该目标函数奖励较高的排序质量（归一化的 $nDCG@10$），同时惩罚校准偏差（归一化的 C_{KL}^t和 C_{KL}^l），其权重分别为（0。4，0。3，0。3）：

$$S(\theta_t, \theta_l) = w_{nDCG} \cdot \widetilde{nDCG} - w_{C_{KL}^t} \cdot \widetilde{C}_{KL}^t - wC_{KL}^l \cdot \widetilde{C}_{KL}^l \quad (5)$$

$$(\theta_t^*, \theta_l^*) = \arg,\max\,(\theta_t, \theta_l) \in \Omega S(\theta_t, \theta_l) \quad (6)$$

4.2 个性化预览生成

如图1 和2所示，我们针对重新排序的新闻生成了个性化新闻预览，以了解这种呈现助推如何影响个体选择。借鉴自我实现助推的概念 [28, 42],，我们的方法旨在帮助用户发现既有偏好之外的新闻：通过呈现不熟悉文章与用户此前的阅读历史或明确表达的兴趣之间的相关联系，推荐因双重校准的局部性组件而新增的文章。我们的目标是通过将相关性说明直接嵌入新闻标题和副标题中，降低认知负荷和注意力障碍，从而克服这些因素对以往自我实现干预效果的限制 [31]。这一流程分为两步：1）识别相关语境文章；2）使用大语言模型，结合原始新闻文章和相关语境，生成新的预览。识别相关语境文章的目的是 H 确保生成的新闻预览更符合用户的主题偏好，或与用户此前读过的事件相关。

为衡量相关性，我们应用句子转换器嵌入来计算余弦相似度。对于每篇文章 $s_i \in S$，我们检查是否存在一篇文章 $h_j \in H$，使得它们嵌入之间的余弦相似度超过经调优的阈值 θ（见第4.2.1节）。这将产生两种可能的改写策略：

(1) **基于事件的预览：**当至少存在一篇文章$h_j \in H$的余弦相似度高于 θ_{sim}时。(2) **基于主题的预览：**当来自 H的文章中没有任何一篇的余弦相似度高于 θ_{sim}时。

对于基于事件的预览，我们提示 GPT-4o-mini 在改写标题和副标题时，考虑 s_i 与相似度最高的文章 $h_j \in H$ 之间的关联。对于基于主题的改写，我们根据用户的 POPROX 入门阅读兴趣调查和此前点击的文章 H。随后，我们提示大语言模型在改写标题和副标题时融入这些高层次的主题偏好（每种改写策略的示例见图 3）。

与先前工作 [72], 类似，我们采用基于反馈的重新提示，以落实新闻预览的可衡量质量。为满足POPROX API的延迟要求，改写操作并行执行。反馈分为两类：1）用于执行风格规范（例如美联社大写规则）的文章级反馈；2）用于确保单份新闻简报内各标题之间语义多样性的新闻简报级反馈。

4.2.1 阈值调优。为确定最佳相似度阈值 θ_{sim}，我们结合定量调优与内部用户试点。以往研究曾使用大语言模型在离线环境中评估生成式新闻叙事中阈值与连贯性之间的权衡 [17],；我们的方法则旨在通过摘要质量指标和人工评估，在保持连贯性与可信度的同时，尽可能增加合理的事件级改写机会。我们使用 ROUGE-L [32] 进行初步参数调优，并承认它作为改写质量代理指标并不完美；我们通过持续开展内部试点测试来弥补其局限性。[4] 离线测试的多轮结果表明，对于事件级改写，当相似度阈值达到 $\theta_{sim} = 0.3$ 时，摘要质量的提升开始出现边际收益递减。因此，我们将此处的初始值设为 θ_{sim}，以捕捉这一拐点，用于试点评估。在内部试点期间，用户提供了非正式反馈，指出了成功和存在问题的改写案例。例如，有人标记了一则不恰当的案例：该改写将一篇关于 Virginia Giuffre 去世的文章与一篇此前报道 7-Eleven 领导层变动的文章相提并论，并把两者都描述为关于“决心”的故事。结合文章元数据分析这些用户反馈后，我们发现 $\theta_{sim} = 0.4$能够清晰地区分不恰当案例与积极的改写，因此在实验中采用了这一更为保守的阈值。

5 实验设计

在本节中，我们详细介绍了针对我们方法的随机对照试验的实验设计（第4节）。我们的研究持续五周，从2025年5月18日至2025年6月22日。共计120名活跃

[4]ROUGE-L 用于衡量改写后的预览与原始预览之间的相似度，因此有助于确保改写内容准确呈现原始文章。然而，它并未考虑我们的目标：融入来自其他来源的信息性背景，而这会降低相似度。

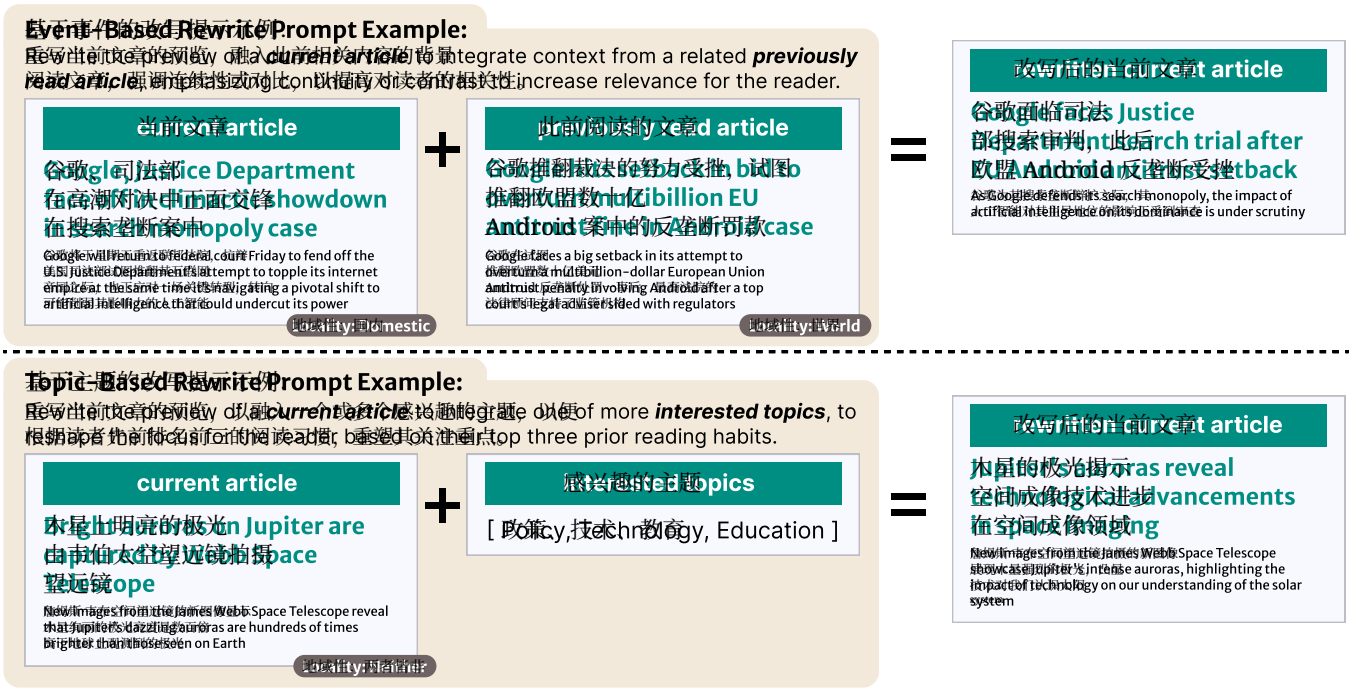


图3：改写策略示例。基于事件的提示（上）整合了与先前点击的文章具有较高余弦相似度的上下文。基于主题的提示（下）则退而采用用户最感兴趣的前3个主题。所示提示词仅用于说明。详细提示词见附录图7。

美国POPROX读者被分配到该实验中，40名参与者被随机分配到三个处理组中的每一组。这些处理措施在实验期间直接影响了用户每日简报中排名前10的新闻文章。

5.1 处理组

我们设计了三种处理条件，以了解双重校准和个性化预览生成相对于主题校准基线的效果。每个处理组都在前一处理定义的基础上构建，以控制算法混杂因素并分离新增组件的效果。

(1) （基线）主题校准（***TC***）：根据公式3 对基础 NRMS 模型应用主题校准。(2) 双重校准（***DC***）：根据公式4，在基线 TC 的基础上额外应用地域性校准。(3) 双重校准 + 新闻个性化（***DC-NP***）：根据第4.2节，在 DC 的基础上应用个性化新闻预览。

5.2 自愿参与调查

为收集用户主观阅读体验的数据，POPROX 每隔 5 周向读者发放问卷。订阅用户会收到一封问卷邮件，询问他们对过去一周阅读体验的多样性、准确性、满意度、控制感及其他方面的看法。我们还设计并发放了研究专用的前后测问卷，供参与者自愿填写。完整的问卷题目见附录表1。所有问卷结果均以描述性结果呈现，不代表统计检验。

5.3 分析方法

为分离实验操纵的影响，我们使用 R 中的 `lme4` 软件包拟合广义线性混合效应模型（GLMM）。我们将参与者在每期新闻简报中点击的文章作为主要关注信号。这一选择受到 POPROX 平台现有功能以及进一步追踪所涉及的伦理影响的限制；我们将在第 8节进一步讨论这一启发式方法的局限性。我们纳入了 67 名至少点击过一篇文章的参与者的数据。对于每个模型，我们将所有连续预测变量转换为 z分数，以提升模型的性能和稳定性。考虑到模型中包含交互项，我们使用 BOBYQA 优化器，并将函数最大评估次数提高到 200,000，以确保模型收敛。我们根据研究问题、理论关注点、平台限制下的数据可用性以及嵌套模型检验来纳入预测变量。这些建模决策使我们能够控制并理解个体、时间以及算法设计方面的差异。

对于接触多样性和消费多样性，我们拟合了两个采用对数连接函数的伽马模型，根据实验期间 AP 的供给多样性，预测截至并包括新闻简报 j 在内的所有推荐文章的 KL $ExposureDivergence_j$和 $ConsumptionDivergence_j$。我们使用 KL散度，因为它在我们的校准模型中是明确优化的目标。由于该指标具有非对称性，我们在整个校准和

评估过程中始终保持相同的方向，将 AP 供给分布作为参考分布，并将用户真实的接触或消费分布作为目标分布（即使用观测到的接触或消费分布近似 AP 供给分布时所损失的信息）。[5] 与 Warncke-Wang 等人类似， [67], 我们在下文报告这些模型的公式，其中 $a_{k[j]}$ 是一个可因用户而异的随机截距 k，$Group_j$ 是实验组， $Week_j$ 是发送新闻简报的实验周， j 而 β（因篇幅所限省略）是各预测变量的估计系数。我们使用嵌套模型检验来验证方程 7 和8中的因素。[6]

$$\log(ExposureDivergence_j) = a_{k[j]} + Group_j + Week_j + Group_j : Week_j \quad (7)$$

$$\log(ConsumptionDivergence_j) = a_{k[j]} + Group_j + Week_j + Group_j : Week_j \quad (8)$$

对于参与度，我们拟合了两个采用 logit 链接函数的二项模型，以预测每篇推荐文章 i 是否被点击（$Clicked_i$）。令 $Position_i$ 表示 i 在新闻简报中的排名。我们将$PriorTopicInterest_i$和 $PriorLocalityInterest_i$定义为 i的主题和地域向量与 k的历史点击之间的余弦相似度。最后，令$PromptLevel_i$ 表示用于呈现文章i的提示策略。我们使用嵌套模型检验来验证方程9 中的因素，并在方程10中保留这一结构。[7] 请注意， $PromptLevel_i$ 仅适用于双重校准+NP 组。因此，我们拟合了一个单独的模型（方程10）来了解其影响。

$$\begin{aligned}\text{logit}(P(Clicked_i = 1)) = {} & a_{k[i]} + Group_i + Position_i + Week_i \\ & + PriorTopicInterest_i + PriorLocalityInterest_i \\ & + Group_i{:}PriorTopicInterest_i\end{aligned} \quad (9)$$

$$\begin{aligned}\text{logit}(P(Clicked_i = 1 \mid Group_i = \text{Context})) = {} & a_{k[i]} + PromptLevel_i + Position_i + Week_i \\ & + PriorTopicInterest_i + PriorLocalityInterest_i \\ & + PromptLevel_i{:}PriorTopicInterest_i\end{aligned} \quad (10)$$

5.4 伦理考量

机构审查。POPROX 实验平台经第三方伦理审查委员会 AD- VARRA [8]审查并批准，我们的具体研究设计也经当地机构审查委员会（IRB）审查并批准。

生成式人工智能的使用。初步设计和文本由作者完成后，我们使用 ChatGPT、Claude 和 Gemini 协助代码开发、调试，并在润色论文草稿时修改语法和文风。所有生成的内容均由作者审核其准确性，并体现了我们的原创思想。

位置性。 我们的研究团队由 POPROX 研究平台的内部和外部成员组成。作为一个新系统，这使我们在试点测试其实验基础设施时，能够优先获得 POPROX 的支持。在整个研究过程中，我们始终意识到自身的位置性，并在可能的情况下优先维护这些界限，以支持 POPROX 不断发展的外部研究人员合作流程。

6 结果

在平均分配至三个处理组的 120 名参与者中，我们从 633 篇不同的新闻文章对应的 915 份个性化新闻简报中收集了 1697 次文章点击。我们还向参与者发放了七份问卷，供其自愿填写，其中包括五份常规 POPROX 周度问卷，以及两份定制研究问卷，用于了解参与者对双重校准和新闻改写的事前与事后看法。我们总共从 28 名参与者处收集了 94 份问卷。由于调查回应数量有限，无法达到统计显著性，我们主要利用问卷结果识别观察趋势，以补充统计分析并为未来工作提供参考。

6.1 训练/离线结果

我们基于参与者的历史数据训练了三个推荐模型，并按照第*4.1*节和*4.2.1*节所述的校准方法，为DC和DC-NP模型调整 theta 值。具体而言，在正式实验开始前，我们收集了参与者连续四周的新闻阅读推荐和活动数据，用前三周进行训练，最后一周用于评估。图*nDCG*4展示了评估数据上原始（ C^t_{KL} @10 C^l_{KL}）值中随机抽取的 60 组 theta 配对的热力图。根据采用启发式权重（公式6）计算的归一化最终得分，DC和 θ_l DC-NP θ_t 的最终 theta 值设为(0.6,0.25)。

6.2 接触多样性（RQ1）

我们通过比较双重校准和 DC-NP 处理组与 TC 基线，评估了算法助推对曝光多样性的影响；请注意，两个处理组接受了相同的底层算法助推。如图5所示，与 TC 基线相比，两种处理措施均显著降低了 KL 曝光差异：双重校准处理组的差异降低了 ~97%，DC-NP 处理组则降低了 ~92%。TC 组的差异随时间推移而下降（每个标准化周降低~29%）。从调查回应中，我们观察到用户对地域性重要性的看法也因处理组而异。两个处理组在以下问题的前后调查回答中，认同率均未发生变化：“我认为，新闻简报为读者提供多样化的国际与国内新闻主题非常重要。”而对照组的认同率从事前到事后调查降低了 ~14%。

研究问题1。地域性+主题双重校准成功提升了曝光多样性，相比单独进行主题校准，处理效应使供给多样性分布的偏离程度减少了近一半。这些改进主要

[5]由于该指标具有尺度不变性，我们对散度指标进行缩放以确保建模的数值稳定性，并对截距进行反向缩放，以报告原始尺度下的系数。[6]公式 7 的嵌套似然比检验从 o $Group_j$ 开始，并依次加入：$a_{k[j]}$(chisq(2)=835.4***)、 $Week_i$(chisq(1)=835.4***)和 $Group_i xWeek_i$(chisq(2)=74.3***)。对于公式 8：$a_{k[j]}$(chisq(1)=823.2***)、 $Week_i$(chisq(1)=823.2***)和 $Group_i xWeek_i$(chisq(2)=12.6**)。***p<.001，**p<.01。[7]公式 9 的嵌套似然比检验从 $Group_i$开始，并依次加入：$a_{k[i]}$(chisq(1)=1300.4***)、 $Position_i$(chisq(1)=145.3***)、 $Week_i$(chisq(1)=56.5***)、$PriorTopicInterest_i$ (chisq(1)=149.9***)、 $PriorLocalityInterest_i$(chisq(1)=6.5*)和 $Group_i xPriorTopicInterest_i$(chisq(2)=12.9**)。***p<.001，**p<.01，*p<.05。[8] https://www.advarra.com/

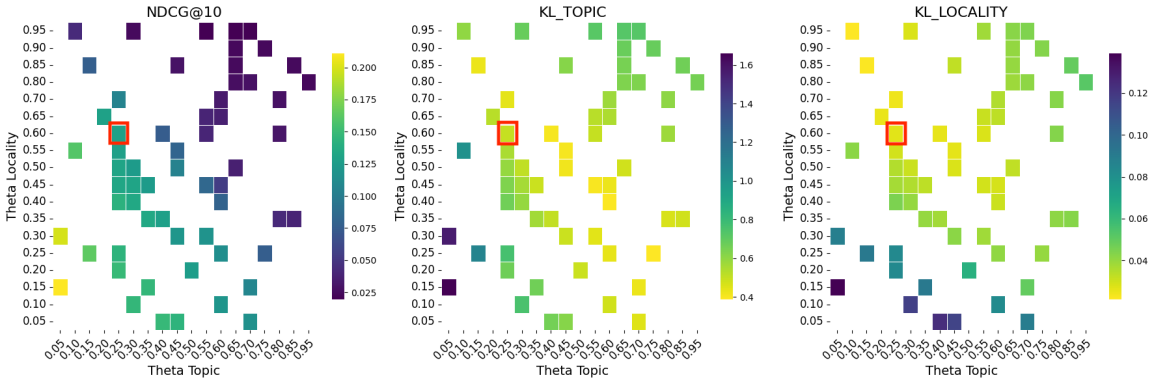


图4：基于实验前一个月评估数据的theta配对调优指标。NDCG@10分数越高越好，theta主题和theta局部性上的KL散度分数越低，表示对齐越好。红框表示基于公式6的最终theta配对选择。

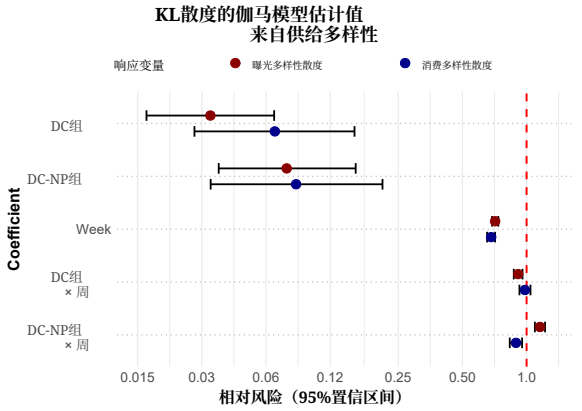


图5：来自公式7（红色）和公式8（蓝色）的相对风险森林图（对数尺度）。数值越小越好。各组与具有中等KL散度的TC 基线进行比较（接触多样性截距：0.06 [0.04，0.10]，消费截距：0.10 [0.06，0.19]）。两种处理措施均显著降低接触多样性差异（双重校准：RR=0.03，p<.001；DC-NP：RR=0.08，p<.001）和消费差异（双重校准：RR=0.07，p<.001；DC-NP：RR=0.08，p<.001）。值得注意的是，双重校准与DC-NP之间的消费差异无显著差异。周对TC组的接触多样性（RR=0.71，p<.001）和消费多样性（RR=0.68，p<.001）有较小但显著的降低效应，处理交互作用带来的偏离较小且显著性不一。

惠及了主题偏好各异的用户，表明在我们的实验人群中，针对这些校准维度进行优化具有实际可行性。

6.3 消费多样性

6.3.1 主题与局部性算法 *nudges*（*RQ2*）。接下来，我们通过比较DC 组与 TC 基线，评估算法助推对消费多样性的影响。图 5 显示与 TC 相比，双重校准使地域消费分歧大幅减少了 ~93%。TC 组的这种分歧也随时间推移而减少（每个标准化周减少 ~32%）。在事后调查中，DC 组用户比 TC 组用户多出 ~33% 认同“他们希望未来的 *POPROX* 新闻简报也能平衡国际与国内新闻。”

RQ2. 地域性+主题双重校准相较于单独的主题校准也提高了消费多样性，其效应量与曝光多样性效应相近。

6.3.2 个性化呈现助推（*RQ3*）。接下来，我们通过将DC-NP组与我们的TC 基线进行比较，观察了双重校准与个性化新闻预览的联合效应。图5 显示，DC-NP处理与地域性消费差异的大幅降低相关（~92%），相较于TC组，但我们发现DC与DC-NP之间没有显著差异。在事后调查中，DC-NP处理组的参与者比对照组更认同（高出16.67%），提高地域性平衡是对POPROX基线算法的改进（“这种平衡旨在根据我们的来源提供方，推荐相近比例的国际与国内新闻。回顾这一平衡，您认为这是一种改进吗？”）。他们还比TC组用户多认同 ~8%，希望未来的新闻简报能够平衡国际与国内新闻。 **RQ3. 我们发现，在双重校准中加入由大语言模型驱动的个性化预览，并不会在算法助推本身的效应之外显著提高消费多样性。** 在下一节中，我们将进一步反思我们的实现方案对参与度的影响，以识别潜在的设计改进，并支持未来关于个性化新闻呈现的研究。

6.4 参与度（研究问题4）

6.4.1 主题与局部性算法 *nudges*。 我们使用点击似然逻辑回归模型，以理解实验设计对参与度的影响。图 6a显示，群体层面效应的差异并不显著；任何群体差异主要由用户层面方差所解释。先前主题和

地域兴趣使点击几率每增加一个标准差分别提高22%和8%。值得注意的是，处理与主题兴趣之间的交互作用显著，这意味着先前主题兴趣对双重校准和DC-NP处理组用户的效应分别增强了28%和21%。**研究问题4。回顾主题校准与双重校准组件，我们发现，在控制位置、实验周以及主题 +地域兴趣后，基线中的用户点击推荐文章的概率为4%；我们没有发现证据表明双重校准处理组的非个性化地域组件对总体参与度产生了负面影响。**

6.4.2 个性化呈现助推。 最后，我们评估双重校准（DC-NP）及其用于个性化每篇文章预览的改写策略对参与度的影响。在实验期间，该算法对总计5,362篇文章采用了不同程度的个性化：531篇文章（10%）采用了基于事件的改写，1,711篇文章（32%）采用了基于主题的改写，3,120篇文章（58%）则未进行改写。与 6.4 中的结果类似，我们发现，图 6a 中双重校准对总体点击参与度没有显著影响。

在双重校准处理组中，图 6b 显示，大多数提示层面的差异可由用户层面方差解释，但事件级提示使点击几率大幅提高了43%，且这一增幅达到边缘显著性（$p<0.05$）。用户层面的主题兴趣是点击的重要预测因素，而位置和周也有显著的负面影响。我们发现主题提示与用户先前主题兴趣之间存在边缘显著的交互效应，这意味着，对于采用主题级改写的文章，先前兴趣的影响更为明显，达 ~40%。

研究问题4。审视 DC-NP 组件后，我们未发现证据表明个性化预览生成对参与度产生了负面影响。此外，按改写策略分析参与度，呈现出更细致的情况：尽管由于安全与合法性约束，较为具体的基于事件的提示词仅在 10% 的情况下得到应用，但即使控制先前主题兴趣等已知算法混杂因素，这些提示词仍有效促进了点击。有趣的是，事后调查的回答显示，DC-NP 组的用户认为自己对新闻简报拥有更多控制权（ 16.67%），高于 DC 组和 TC 组（即，“我觉得自己能够控制 *POPROX News* 选择哪些文章。”）。

7 讨论

在此，我们借助先前研究和多样性助推框架 [41] 的视角，反思实验结果。我们将考察算法助推和呈现助推的作用机制，以及在设计多样化新闻推荐系统时更广泛的启示。

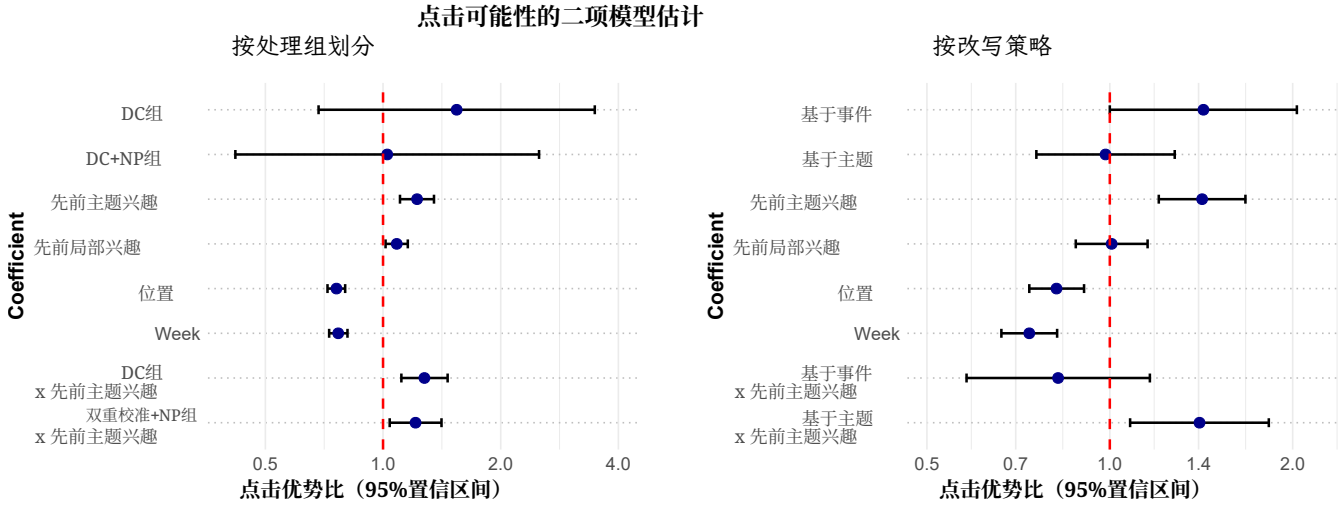
7.1 多样性助推

我们的结果在诸多重要方面既与先前关于新闻多样性的研究一致，也有所不同。与先前研究一样， [27], 我们的发现表明，当多样化内容根据个人偏好进行个性化时，曝光多样性与消费多样性密切相关。我们观察到，将用户的曝光多样性校准至我们的目标，会立即影响消费多样性；与~93相比，双重校准（DC）使地域消费分歧降低了 %。此外，对图6a 进行的嵌套模型检验显示，周与相比我们的基线，我们的处理条件并未出现显著交互作用，这表明这一改善并非以牺牲长期参与度为代价，这与近期一些关于新闻推荐系统的研究结果相呼应 [41, 71]。值得注意的是，TC 基线的地域多样性随着时间推移而提升，且期间没有进行任何操纵。这一模式不同于社交媒体上关于政治新闻消费的讨论 [29], ，因为用户会随着新闻周期自然而然地探索国内与国际新闻。尽管双重校准的即时响应和一致的行为转变可能对某些目标很重要，但我们的结果强调，一次性用户研究可能会高估多样性感知推荐系统的效应量。呼应先前的倡议 [26], ，我们强调有必要继续了解助推新闻多样性的长期影响。

数据驱动的助推近期受到了广泛关注。 [53–55] 回顾助推框架，处理组之间缺乏显著差异，这一点值得进一步思考。我们的大语言模型呈现助推符合多种理论上的助推定义：一种引导认知过程的智能助推 [43], 一种促使人们反思过去选择所产生后果的助推 [60],以及一种帮助用户培养新偏好的自我实现助推 [28, 42]。与先前研究 [40, 41], 一致，我们的汇总结果表明，操控呈现给 DC-NP 组的信息基本无效。考虑到助推框架区分了算法助推与呈现助推，我们注意到，在实践中很难将这些干预措施彼此区分。例如，推荐位置／排名对参与度的影响最大，但这种助推同时作用于接触多样性和选择性，因此无法判定多样化消费究竟源于真正的选择性，还是接触限制。如图 6b所示，基于事件的改写大幅提高了点击可能性，但仅达到边缘显著，同时也呈现出降低用户既有主题兴趣（即其习惯性消费模式）的趋势。未来研究应将呈现助推目标融入排序算法，以增加合理的基于事件的叙事改写数量，并提高此类干预的统计功效。这种方法为影响多样性目标中的选择性维度提供了一个有前景的方向，使用户能够发挥自主性，实现多样化的选择。

7.2 社会影响与新闻完整性

新闻推荐校准可以持续为读者选择不同主题和地域性的文章供其阅读，从而改变他们对事物的发生过程的信息感知和理解 [16, 22]。我们的调查回应揭示了一种规范性转变：处理组中的用户维持或增强了这样一种看法，即新闻简报应提供多样性，而对照组的认同度则有所下降。这表明，长期接触校准后的多样性可能会培养出更具公民意识的读者，他们期待并重视多元新闻视角，这与良性循环假说一致 [22, 58]。与单独校准相比，个性化预览（DC-NP）并未显著改变消费多样性，但在用户自主性方面发挥了重要作用。我们发现，DC-NP组的参与者表示，他们对推荐内容的控制感更强。这表明，大语言模型驱动的个性化产生的社会影响，或许与其说在于改变



（a）方程式中优势比的森林图（对数尺度）：9。相对于 TC 基线的组间比较（截距：0。04 [0.02, 0。07]）。两者皆非双重校准（OR：1.5，p=.36）与 DC-NP（OR：1.03，p=.36）均未提高整体点击可能性，但二者都显著增强了先前主题兴趣的影响（双重校准×主题：OR=1.28，p<.001；DC-NP×主题：OR=1.21，p<.01），表明校准提高了用户对熟悉主题的选择性。位置（OR=0.76，p<.001）和周（OR=0.77，p<.001）对参与度有显著负面影响。

（b）方程式中优势比的森林图（对数尺度）：10。相对于不改写的改写比较（截距：0。04 [0.02, 0。07]）。基于事件的改写显著增加了点击量（OR=1.42，p<.05），同时控制先前主题兴趣（OR=1.42，p<.001）和位置（OR=0.82，p<.001）。基于主题的改写增强了先前兴趣的影响（OR=1.41，p<.01），而事件提示则没有（OR=.82，p=.48），这表明基于事件的改写能够以更高的统计功效影响用户选择性。

人们阅读的内容，不如说在于提高算法的透明度和可理解性 [15, 61]。通过个性化语境解释推荐背后的“原因”，我们从黑箱助推转向用户与模型之间的协作式发现过程 [20, 56]。尽管如此，我们在尝试改写新闻预览时观察到一个关键挑战：如何平衡原始新闻内容的事实准确性与个性化语境的整合。具体而言，为避免过度个性化扭曲原始报道，我们需要经过多轮试点，才能确定语境文章与目标新闻文章之间的相似度阈值。

未来的研究应探索更为稳健的保障机制，以确保个性化引子不会以牺牲新闻完整性为代价，例如“验证链” [14] 和“大语言模型评审” [19]。尽管我们当前的方法依赖启发式相似度阈值，但未来的迭代可以采用辅助事实核查模型，验证在改写过程中核心事件细节和新闻实体是否保持不变。此外，探索基于事件的改写与基于主题的改写之间的权衡，有助于确定能够最大化参与度的最佳个性化粒度，同时避免给计算资源或新闻生态系统的事实边界带来过大压力 [8]。未来的研究还应进一步完善这些呈现助推及其事实保障措施，最终支持算法在不同新闻文章之间建立更多新颖的关联。

8 局限性

我们承认本研究存在三个主要局限。首先，新闻主题分类体系和地域性由美联社编辑和其自动标注算法所定义。尤其是，少量新闻就其地域性被归类为"两者皆非"，这有时会限制平衡的地域性校准。未来研究可以进一步探索更细粒度的新闻地理校准，例如地方、区域、国家、洲际和世界层面的新闻。

其次，由于 POPROX 是一个相对较新的平台，我们的研究受到了新兴工具典型生命周期的影响。这主要表现为两个方面：高度活跃用户的样本量相对较小；随着新订阅用户最初的新鲜感逐渐消退，用户流失或参与度下降的趋势也日益明显。这些因素，加上响应变量的平稳性以及实验数据的时间窗口设置，可能会限制我们将纵向研究结果推广到规模更大、更成熟的新闻环境中的可推广性。

最后，POPROX 的生态情境给数据收集带来了挑战。作为一个遵循伦理规范的研究平台，POPROX 除 API 调用（即电子邮件打开行为和阅读行为）外，不会收集其他行为日志，这限制了我们追踪更细致参与行为的能力。此外，由于平台完全依赖参与者的内在动机，他们自愿参与每周问卷和研究专项调查。因此，调查回应数量有限，制约了我们在某些感知指标上实现被试间统计显著性的能力。尽管这些调查结果为日志分析提供了有价值的观察趋势作为补充，但未来研究或可采用混合激励模型，在不损害新闻消费体验自然情境的前提下，确保获得更稳定的纵向反馈。

9 结论

在本研究中，我们通过对国内和国际新闻报道中的两类助推进行实证评估，阐明接触多样性与消费多样性之间的关系。通过在 POPROX 平台上对 120 名美国读者开展为期 5 周的纵向研究，我们证明，主题局部性双校准算法助推成功提升了国内和国际新闻报道的接触多样性与消费多样性。对于基于大语言模型的呈现助推，我们的研究发现揭示了个性化策略影响中的重要细节：高度具体、基于事件的方法优于基于主题的方法和不采用个性化的方法，但缺少合理进行基于事件的改写的机会，限制了其对消费多样性的影响。这些结果表明，基于事件的语境化是未来研究中影响用户面对多样化推荐时选择性的一条有前景的方向。总体而言，本研究为考察多样性助推对用户情境化阅读行为的长期影响这一有限领域贡献了新的成果。

参考文献

[1] [n . d.]. 美联社分类元数据。https://developer.ap.org/ap-media-api/agent/ AP Classification Metadata.htm 出版物标题：美联社分类元数据。

[2] 2024 美联社数字概览。https://www.ap.org/about/annual-report/2023- report/2023-ap-by-the-numbers/ 出版物标题：美联社。[3] Himan Abdollahpouri、Masoud Mansoury、Robin Burke 和 Bamshad Mobasher。2020。流行度偏差、校准与公平性之间的联系。载于第14届ACM推荐系统会议论文集。726–731。[4] Elena Agapie 和 Sean Munson。2015。社交线索与阅读政治新闻报道的兴趣。载于AAAI国际网络与社交媒体会议论文集，第9卷。562–565。[5] Kholoud Aldous、Joni Salminen、Ali Farooq、Soon-gyo Jung 和 Bernard Jansen。2024年。将ChatGPT用于内容营销：通过生成式人工智能增强用户在跨平台内容创作中的社交媒体互动。载于第35届ACM超文本与社交媒体会议论文集。376–383。[6] Christine Bauer、Chandni Bagchi、Olusanmi A. Hundogan 和 Karin van Es。2024年。《价值何在？关于新闻推荐系统的系统性文献综述》。2，3（2024年）。https://doi.org/10.1145/3654805 地点：美国纽约州纽约市 出版商：计算机协会。[7] James Bergstra 和 Yoshua Bengio。2012。用于超参数优化的随机搜索。《机器学习研究杂志》13，1（2012），281–305。[8] Sébastien Bubeck、Varun Chandrasekaran、Ronen Eldan、Johannes Gehrke、Eric Horvitz、Ece Kamar、Peter Lee、Yin Tat Lee、Yuanzhi Li、Scott Lundberg 等人。2023。通用人工智能的曙光：使用GPT-4的早期实验。arXiv预印本 arXiv:2303.12712（2023）。[9] Robin Burke、Joseph Konstan 和 Michael Ekstrand。2024年。使用POPROX开展推荐系统用户研究。载于第18届ACM推荐系统会议论文集。1277–1278。[10] Diego Carraro 和 Derek Bridge。2025。使用大语言模型重新排序以提升推荐多样性。《ACM推荐系统汇刊》4，2（2025年11月）。https://doi.org/10.1145/3700604 地点：美国纽约州纽约市 出版商：计算机协会。[11] Stefano Castriota、Marco Delmastro 和 Mirco Tonin。2023。全国性还是地方性信息疫情？COVID-19期间意大利的新闻需求。《国际健康经济学与管理杂志》23，4（2023年12月），507–536。https://doi.org/ 10.1007/s10754-023-09350-3 地点：美国。[12] Abhijnan Chakraborty、Saptarshi Ghosh、Niloy Ganguly 和 Krishna P Gummadi。2019。在非个性化新闻推荐中优化时效性—相关性—多样性权衡。《信息检索期刊》22（2019），447–475。[13] Chao Chang、Feiyi Tang、Peng Yang、Jingui Zhang、Jingxuan Huang、Junxian Li 和 Zhenjun Li。2025。用于个性化新闻推荐的多视图知识表示学习。《科学报告》15，1（2025年1月），1152。https://doi.org/10.1038/s41598-025-85166-0[14] Shehzaad Dhuliawala、Mojtaba Komeili、Jing Xu、Roberta Raileanu、Xian Li、Asli Celikyilmaz 和 Jason Weston。2024。验证链可减少大语言模型中的幻觉。载于计算语言学协会发现论文集：ACL 2024。3563–3578。[15] Nicholas Diakopoulos。2025。前瞻性算法问责与新闻媒体的作用。载于《跨学科计算机伦理学：Deborah G. Johnson与算法问责》。施普林格，125–143。

[16] Seth Flaxman、Sharad Goel 和 Justin M Rao。2016年。《过滤气泡、回音室与在线新闻消费》。《舆论季刊》80，S1（2016年），298–320。[17] Shen Gao、Jiabao Fang、Quan Tu、Zhitao Yao、Zhumin Chen、Pengjie Ren 和 Zhaochun Ren。2024年。生成式新闻推荐。载于《2024年 ACM Web Conference 会议论文集》（美国纽约州纽约市，2024年）（WWW '24）。计算机协会，3444–3453。https://doi.org/10.1145/3589334.3645448 活动地点：新加坡，新加坡。[18] Kiran Garimella、Tim Smith、Rebecca Weiss 和 Robert West。2021年。在线新闻消费中的政治极化。15，1（2021年），152–162。https://doi.org/10.1609/icwsm.v15i1.18049[19] Jiawei Gu、Xuhui Jiang、Zhichao Shi、Hexiang Tan、Xuehao Zhai、Chengjin Xu、Wei Li、Yinghan Shen、Shengjie Ma、Honghao Liu 等。2024年。关于大语言模型评审的综述。《The Innovation》（2024年）。[20] Jeffrey Heer。2019年。能动性与自动化：将人工智能融入交互式系统的设计。《美国国家科学院院刊》116，6（2019年），1844–1850。[21] Lucien Heitz、Juliane A. Lischka、Alena Birrer、Bibek Paudel、Suzanne Tolmei- jer、Laura Laugwitz 和 Abraham Bernstein。2022年。多样化新闻推荐对民主的益处：一项用户研究。10，10（2022年），1710–1730。https://doi.org/10.1080/21670811.2021.2021804 出版商：Routledge。[22] Natali Helberger。2021年。新闻推荐系统的民主角色。载于《算法、自动化与新闻》。Routledge，14–33。[23] Jonathan Hendrickx、Annelien Smets 和 Pieter Ballon。2021年。新闻推荐系统与新闻多样性是同一回事吗？来自小型媒体市场的案例研究。《新闻与媒体》2，3（2021年），515–528。https: //doi.org/ 10 3390 2030031 . /journalmedia Kung-Hsiang Huang、[24] Philippe Laban、Alexander Fabbri、Prafulla Kumar Choubey、Shafiq Joty、Caiming Xiong 和 Chien-Sheng Wu。2024年。《拥抱分歧，获得更丰富的洞见：多文档摘要基准，以及从新闻文章中总结多样化信息的案例研究》。载于《计算语言学协会北美分会2024年会议论文集：人类语言技术》（第1卷：长篇论文），Kevin Duh、He- lena Gomez 和 Steven Bethard（编）。计算语言学协会，墨西哥城，墨西哥，570 593 10 18653 1 2024 32 – 。[25] 2024 https://doi.org/ . /v / .naacl-long。Jia Hua Jeng。。借助推荐系统弥合新闻中的观点差异。载于第18届 ACM推荐系统会议论文集（意大利巴里）（Rec- Sys '24）。计算机协会，美国纽约州纽约市，1283–1289。https://doi.org/ 10 1145 3640457 3688008 . / . Mathias Jesse 和 Dietmar Jannach。[26] 2021年。借助推荐系统进行数字助推：综述与未来方向。《人类行为中的计算机报告》3（ 2021 100052 10 1016 2020 100052 ），。https://doi.org/ . / [27] 2022 j.chbr. . Pascal Jürgens 和 Birgit Stark。。绘制接触多样性：算法策展对新闻消费的不同影响。《传播学刊》72，3（2022年3月），322–344。https://doi.org/10.1093/joc/jqac009 _电子版: https://academic.oup.com/joc/article-pdf/72/3/322/43919230/jqac009.pdf。[28] Bart P Knijnenburg、Saadhika Sivakumar 和 Daricia Wilkinson。2016年。促进自我实现的推荐系统。载于第10届 ACM推荐系统会议论文集。11–14。[29] Ro'ee Levy。2021年。社交媒体、新闻消费与极化：一项田野实验的证据。111，3（2021年），831–70。https://doi.org/10.1257/aer. 20191777[30] Miaomiao Li 和 Licheng Wang。2019年。个性化新闻推荐技术综述。7（2019年），145861–145879。出版商：电气电子工程师学会。[31] Q Vera Liao 和 Wai-Tat Fu。2014年。现在能听见我说话吗？通过来源立场指标缓解回音室效应。载于第17届 ACM计算支持的协同工作与社交计算会议论文集。184–196。[32] Chin-Yew Lin。2004年。ROUGE：用于摘要自动评估的软件包。载于《文本摘要的拓展》。计算语言学协会，西班牙巴塞罗那，74–81。[33] Jiahui Liu、Peter Dolan 和 Elin Rønby Pedersen。2010年。基于点击行为的个性化新闻推荐。载于第15届国际智能用户界面会议论文集（美国纽约州纽约市，2010年）（IUI '10）。计算机协会，31–40。https://doi.org/10.1145/1719970. 1719976 活动地点：香港，中国。[34] Ping Liu、Karthik Shivaram、Aron Culotta、Matthew A. Shapiro 和 Mustafa Bilgic。2021年。政治类型学与新闻推荐算法中的过滤气泡之间的相互作用。载于《2021年 Web会议论文集》（美国纽约州纽约市，2021年）（WWW '21）。计算机协会，3791–3801。https://doi.org/10.1145/3442381.3450113 活动地点：卢布尔雅那，斯洛文尼亚。[35] Ruibo Liu、Lili Wang、Chenyan Jia 和 Soroush Vosoughi。2021年。利用属性感知词嵌入降低新闻文章的政治极化。载于《AAAI国际网络与社交媒体会议论文集》，第15卷。385–396。[36] Felicia Loecherbach、Judith Moeller、Damian Trilling 和 Wouter van Atteveldt。2020年。媒体多样性统一框架：系统性文献综述。《数字新闻学》8，5（2020年），605–642。[37] Philipp Lorenz-Spreen、Stephan Lewandowsky、Cass R. Sunstein 和 Ralph Hertwig。2020年。行为科学如何促进真理、自主性与

在线民主话语。4，11（2020年），1102–1109。https://doi.org/10.1038/s41562-020-0889-7[38] Katharina Ludwig、Alexander Grote、Andreea Iana、Mehwish Alam、Heiko Paul- heim、Harald Sack、Christof Weinhardt 和 Philipp Müller。2023年。被算法所分化？基于内容和情感的新闻推荐对情感、意识形态及感知极化的（有限）影响。41，6（2023年），2188–2210。https://doi.org/10.1177/08944393221149290 出版商：SAGE Publications Inc.[39] GREGORY J. MARTIN 和 JOSHUA McCRAIN。2019年。地方新闻与国家政治。《美国政治学评论》113，2（2019年），372–384。https://doi.org/10.1017/S0003055418000965 版本：2019/02/19 出版商：剑桥大学出版社。[40] Nicolas Mattis、Tim Groot Kormelink、Philipp K. Masur、Judith Moeller 和 Wouter van Atteveldt。2025年。助推新闻读者：运用混合方法理解界面助推何时以及如何影响新闻选择。《数字新闻学》13，9（2025年10月），1647–1667。https://doi.org/10.1080/21670811.2024. 2350464 出版商：Routledge。[41] Nicolas Mattis、Lucien Heitz、Philipp K Masur、Judith Moeller 和 Wouter van Atteveldt。2025年。面向新闻推荐系统的助推：突出安排文章位置可提高环境新闻的选择率、参与度和回忆率，但降低复杂度并无此效。《传播学刊》（2025年），jqaf019。出版商：牛津大学出版社。[42] Nicolas Mattis、Philipp Masur、Judith Möller 和 Wouter van Atteveldt。2024年。助推读者接触多样化新闻：通过推荐系统设计促进多样化新闻消费的理论框架。26，7（2024年），3681–3706。https: //doi.org/10.1177/14614448221104413 出版商：SAGE Publications。[43] Cristina Mele、Tiziana Russo Spena、Valtteri Kaartemo 和 Maria Luisa Marzullo。2021 年。智能助推：认知技术如何助力构建价值共创的选择架构。《商业研究杂志》129（2021年），949–960。https: //doi.org/10 1016 2020 09 004 . /j.jbusres. . . Sean Munson、[44] Stephanie Lee 和 Paul Resnick。2013年。通过浏览器小部件鼓励读者阅读多样化的政治观点。载于《AAAI国际网络与社交媒体会议论文集》，第7卷。419–428。[45] Jinoh Oh、Sun Park、Hwanjo Yu、Min Song 和 Seung-Taek Park。2011年。基于个人流行度倾向的新颖性推荐。载于2011年电气电子工程师学会第十一届国际数据挖掘会议（2011年）。507–516。https://doi.org/10.1109/ICDM.2011.110[46] Shumpei Okura、Yukihiro Tagami、Shingo Ono 和 Akira Tajima。2017年。面向数百万用户的基于嵌入的新闻推荐。载于第23届 ACM SIGKDD 国际知识发现与数据挖掘会议论文集（美国纽约州纽约市，2017年）（KDD '17）。计算机协会，1933–1942。https://doi.org/10.1145/3097983.3098108 活动地点：加拿大新斯科舍省哈利法克斯。[47] Souneil Park、Seungwoo Kang、Sangyoung Chung 和 Junehwa Song。2009年。NewsCube：通过提供新闻的多个方面来缓解媒体偏见。载于 SIGCHI 人机交互系统会议论文集。443–452。[48] Kenny Peng、Manish Raghavan、Emma Pierson、Jon Kleinberg 和 Nikhil Garg。2024年。调和推荐中的准确性—多样性权衡。载于《2024年 ACM Web Conference 会议论文集》（WWW '24）。计算机协会，美国纽约州纽约市，1318–1329。https://doi.org/10.1145/3589334. 3645625 活动地点：新加坡，新加坡。[49] Junyang Rao、Aixia Jia、Yansong Feng 和 Dongyan Zhao。2013年。基于分类体系的个性化新闻推荐：新颖性与多样性。载于《网络信息系统工程——WISE 2013》，Xuemin Lin、Yannis Manolopoulos、Divesh Srivastava 和 Guangyan Huang（编）。施普林格·柏林·海德堡，德国柏林、德国海德堡，209–218。[50] Shaina Raza 和 Chen Ding。2022年。新闻推荐系统：近期进展、挑战与机遇综述。《人工智能评论》55，1（2022年1月），749–800。https://doi.org/10.1007/s10462-021-10043-x 地点：美国 出版商：克鲁维尔学术出版社。[51] Shaina Raza、Deepak John Reji 和 Chen Ding。2024年。Dbias：检测新闻文章中的偏见并确保公平性。《国际数据科学与分析杂志》17，1（2024年），39–59。[52] Qin Ruan、Jin Xu、Susan Leavy、Brian Mac Namee 和 Ruihai Dong。2024年。重写偏见：通过自动改写缓解新闻推荐系统中的媒体偏见。载于第32届 ACM 用户建模、自适应与个性化会议论文集。67–77。[53] Armindokht H. Sadeghian 和 Ali Otarkhani。2024年。数据驱动的数字助推：系统性文献综述与未来议程。《行为与信息技术》43，15（2024年11月），3834–3862。https://doi.org/10.1080/0144929X.2023. 2286535 出版商：泰勒-弗朗西斯出版集团。[54] Donghee Shin。2025年。工程公平：设计多样性感知人工智能以反映人类。《人工智能与社会》40，7（2025年10月），5657–5666。https://doi.org/ s00146-025-02314-w[55] Donghee Shin 和 Shuhua Zhou。2024年。价值与多样性感知新闻推荐系统：算法把关能否助推读者阅读多样化新闻？《新闻与大众传播季刊》（2024年6月），10776990241246680。https://doi.org/10.1177/10776990241246680 出版商：SAGE Publications Inc.[56] Ben Shneiderman. 2020. 以人为中心的人工智能：可靠、安全且值得信赖。《国际人机交互杂志》36, 6 (2020), 495–504.[57] Harald Steck. 2018. 校准推荐。载于《第12届ACM推荐系统会议论文集》（美国纽约州纽约市，2018）（RecSys '18）。计算机协会，154–162。https://doi.org/10.1145/3240323. 3240372 举办地：加拿大不列颠哥伦比亚省温哥华。[58] Jesper Strömbäck and Adam Shehata. 2010. 媒体倦怠还是良性循环？探索新闻媒体接触、政治新闻关注和政治兴趣之间的因果关系。《欧洲政治研究杂志》49, 5 (2010), 575–597.[59] Ruixuan Sun, Avinash Akella, Ruoyan Kong, Moyan Zhou, and Joseph A Kon- stan. 2024. 在线电影推荐系统中的交互式内容多样性与用户探索：一项田野实验。《国际人机交互杂志》40, 22 (2024), 7233–7247。[60] Cass R Sunstein. 2014. 助推：极简指南。《隐私研究手册》37 (2014), 583–588. [61] Nava Tintarev and Judith Masthoff. 2015. 推荐解释：设计与评估。载于《推荐系统手册》。施普林格，353–382。[62] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. 2018. 相同，却又不同：新闻观点的算法多样化。载于第26届用户建模、自适应与个性化会议（UMAP '18）附属出版物。计算机协会，美国纽约州纽约市，7–13。https://doi.org/10.1145/3213586.3226203 举办地：新加坡，新加坡。[63] Richard Van der Wurff. 2011. 受众能从新闻媒体中接触到多元观点吗？新闻媒体接触的多样性，以及个人特征如何决定所接收内容的多样性。《欧洲传播学杂志》26, 4 (2011), 328–342.[64] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. 使命驱动型推荐系统：评估新闻推荐中的多样性。载于《2021年人类信息交互与检索会议论文集》。173–183.[65] Chenyang Wang, Yankai Liu, Yuanqing Yu, Weizhi Ma, Min Zhang, Yiqun Liu, Haitao Zeng, Junlan Feng, and Chao Deng. [n. d.]. 面向质量感知的负责任推荐的双侧校准。载于《第17届ACM推荐系统会议论文集》（美国纽约州纽约市，2023）（RecSys '23）。计算机协会，223–233。https://doi.org/10.1145/3604915. 3608799 举办地：新加坡，新加坡。[66] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. 推荐系统公平性综述。《ACM信息系统汇刊》41, 3（2023年2月）。https://doi.org/10.1145/3547333 地点：美国纽约州纽约市 出版商：计算机协会。[67] Morten Warncke-Wang, Rita Ho, Marshall Miller, and Isaac Johnson. 2023. 借助新成员主页提升同侪生产社区的参与度。《ACM人机交互汇刊》7, CSCW2，文章280（2023年10月），26 10 1145 3610071 页数。https://doi.org/ 。 / Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, [68] Yongfeng Huang, and Xing Xie. 2019. 基于多头自注意力的神经新闻推荐。载于《2019年自然语言处理实证方法会议暨第9届国际自然语言处理联合会议论文集（EMNLP-IJCNLP）》。6389–6394.[69] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. 个性化新闻推荐：方法与挑战。41, 1 (2023)。https://doi.org/ 10.1145/3530257 地点：美国纽约州纽约市 出版商：计算机协会。[70] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. SentiRec：感知情感多样性的神经新闻推荐。载于《亚太计算语言学协会第1届会议暨第10届国际自然语言处理联合会议论文集》，Kam-Fai Wong, Kevin Knight, and Hua Wu（编）。计算语言学协会，中国苏州，44 53 10 18653 1 2020 6 – 。[71] https://doi.org/ 。 /v / .aacl-main. Xudong Yu, Muhammad Haroon, Ericka Menchen-Trevino, and Magdalena Woj- cieszak. 2024. 助推推荐算法能提升YouTube上的新闻消费和多样性。《PNAS Nexus》3, 12 (2024), pgae518. [72] Matthew Zent, Digory Smith, and Simon Woodhead. 2025. PIIvot：面向问题锚定式辅导对话的轻量级自然语言处理匿名化框架。载于《2025年自然语言处理实证方法会议论文集》，Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng（编）。计算语言学协会，中国苏州，27467 27476 10 18653 1 2025 1397 – 。[73] https://doi.org/ 。 /v / .emnlp-main. Guangping Zhang, Dongsheng Li, Hansu Gu, Tun Lu, and Ning Gu. 2024. 用于新闻推荐的、具有个性化和自适应多样性的异构图神经网络。《ACM Web汇刊》18, 3（2024年5月）。https: //doi.org/10.1145/3649886 地点：美国纽约州纽约市 出版商：计算机协会。

[74] Han Zhang, Ziwei Zhu, and James Caverlee. [n.d.]. 新闻推荐中过滤气泡与极化的演变。载于信息检索进展：第*45*届欧洲信息检索会议，*ECIR 2023*，爱尔兰都柏林，*2023*年*4*月*2*日至*6*日，会议论文集，第二部分（柏林，海德堡，2023）。施普林格出版社，685–693。https://doi.org/10.1007/978-3-031-28238-6_60 举办地：爱尔兰都柏林。[75] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN：用于新闻推荐的深度强化学习框架。载于《*2018*年万维网大会论文集》（日内瓦共和国和州，瑞士，2018年）*(WWW '18)*。国际万维网会议指导委员会，167–176。https://doi.org/10.1145/ 3178876.3185994 举办地：法国里昂。

A 附录A.1 提示详情

基于事件的提示词	基于主题的提示词
“你是一名美联社编辑，负责以事实性语气改写一则新闻预览。” “你会收到一条[[主要_新闻]]，包含正文和过去的[[相关新闻_，包含标题、副标题和正文。用户可能对[[主要_新闻]]不感兴趣。用户此前阅读过过去的[[相关_新闻]]。" "创作一条[[主要_新闻]] 标题和副_标题，以吸引他们的兴趣，并使用以下规则。\n" "规则：\n" "1. ***明确***整合[[相关_新闻]] 标题，副_标题和正文融入[[主要_新闻]] 标题和副_标题。它们之间的关联应当有意义，而不是仅仅在一笔带过。\n" "2. 重新构思[[主要_新闻]] 正文中的一个元素，并在[[主要_新闻]] 标题和副_标题中突出自然的递进、对比，或[[主要_新闻]] 正文与[[相关_新闻]]之间更深层的背景正文。突出说明[[主要_新闻]] 正文如何承接、挑战或拓展读者先前的理解。\n" "3. [[主要_新闻]] 副_标题不应以标点符号结尾。\n" "4. ***仅专有名词应在[[主要_新闻]] 标题中大写。***\n" "5. [[主要_新闻]] 标题长度应约为{}个词。\n" "6. [[MAIN_新闻]] SUB_标题应约为 {} 个词长。\n" "7. [[MAIN_新闻]] 标题和 SUB_标题应保持中立并准确描述[[MAIN_新闻]] 正文。\n" "8. 你的回复应仅包含可由 json.loads() 解析的 JSON，格式为「 '\'{{"标题": "[REWRITTEN_HEADLINE]", "SUB_HEADLINE": "[改写后_副标题]"}}\'.	"你是一名美联社编辑，负责以事实性语气改写新闻预览。" "为你提供了一份用户广泛感兴趣的[[INTERESTED_TOPICS]] 列表，其中包含主题，以及美联社定义和一则带有正文的[[MAIN_NEWS]] 。用户可能对[[MAIN_NEWS]]不感兴趣。" "撰写一个能吸引读者的[[MAIN_NEWS]] 标题和副_标题的内容，并遵循以下规则。\n" "规则：\n" "1. ***隐式地***将用户过去的一种或多种阅读习惯融入[[INTERESTED_TOPICS]]，并将其融入[[MAIN_NEWS]] 标题中，且以自然的方式重新塑造新闻预览的焦点以契合用户的兴趣。\n" "2. 重新构思[[主要_新闻]] 正文中的一个元素，以及[[主要_新闻]] 标题和副标题，以强调一个角度或段落，直接吸引用户感兴趣的[[主题_，让这则新闻显得尤为相关。突出意想不到的联系或独到见解在[[主要_新闻]] 正文与用户广泛的[[感兴趣的_主题]].\n" "3. [[主要_新闻]] 副_标题不应以标点符号结尾。\n" "4. ***只有专有名词应在 [[主_新闻]] 中大写、】【标题。***\n" "5. [[主要_新闻]] 标题应约为 {} 个单词。\n" "6. [[主要_新闻]] 副_标题应约为 {} 个词长。\n" "7. [[主要_新闻]] 标题和副_标题应保持中立并准确描述[[主要_新闻]] 正文。\n" "8. 你的回复只能包含可由 json.loads() 解析的 JSON，格式如下为 " '\'{{"标题": "[改写_标题]", "副_标题": "[改写_副标题]"}}\'.

图 7：基于事件和主题的新闻改写提示。

A.2 事前与事后调查问题

ID	问题文本	回答选项	阶段
Q1	我认为，新闻简报提供多样化的国际与国内新闻主题供读者阅读非常重要。	五点李克特量表（非常不同意 [SD] – 非常同意 [SA]）	前测与后测
Q2	我认为，人们阅读多样化的国际与国内新闻主题。	五点李克特量表（非常不同意–非常同意）	前测与后测
Q3	我认为，将新闻标题和描述的文本个性化，使其符合我的兴趣，能改善我的新闻阅读体验。	五点李克特量表（非常不同意–非常同意）	前测与后测
Q4	我认为，使用人工智能（AI）个性化新闻预览以符合我的兴趣，是准确且值得信赖的。	五点李克特量表（非常不同意–非常同意）	前测与后测
Q5	请回想您最近收到的新闻简报，思考其中文章的相关性符合您兴趣的文章。大约有多少篇文章与您的兴趣相关?	没有，1–3 篇，4–6 篇，7–9 篇，全部	前测与后测
Q6	回想一下您最近收到的新闻简报中国际与国内文章的平衡。我注意到.....	国际新闻多得多，国际新闻稍多，没有差异 , 国内倾向稍多 , 很多国内倾向更多 , 我不确定	前测与后测
Q7	*[如果问题6 ≠ "不确定"]* 这种平衡旨在推荐相似的根据我们的来源提供方，国际与国内新闻的比例。回想一下这种平衡，您认为这是一种改进吗?	五点李克特量表（非常不同意–非常同意）	仅后测
Q8	我希望未来的 POPROX新闻简报也能平衡国际和国内新闻。	五点李克特量表（非常不同意–非常同意）	仅后测
Q9	我觉得自己可以控制 POPROX News 选择哪些文章。	五点李克特量表（非常不同意–非常同意）	POPROX 原生版本，前测与后测

表 1：调查问题和回答量表