推荐系统冷启动探索的动态先验汤普森采样

Zhenyu Zhao
zzhao@roblox.com
Roblox 圣马特奥，加利福尼亚州，美国

David Zhang
david.zhang@roblox.com
Roblox 圣马特奥，加利福尼亚州，美国

Ellie Zhao
ezhao@roblox.com
Roblox 圣马特奥，加利福尼亚州，美国

Ehsan Saberian
esaberian@roblox.com
Roblox 圣马特奥，加利福尼亚州，美国

arXiv:2602.00943v1 [cs.LG] 1 Feb 2026

摘要

冷启动探索是在线优化和推荐系统中的核心瓶颈：平台必须为新引入或数据稀疏的项目分配流量，以估计其价值，但过度探索会直接损害用户体验。在实践中，汤普森采样通常采用均匀的 Beta(1，1) 先验，隐含地将新项目的成功概率设为 50%。在优质内容基准率低得多的大规模市场中，这种乐观先验会导致系统性地过度探索表现不佳的项目。批量模型更新和流水线延迟等生产环境的现实情况会加剧这一问题：新上线项目可能连续数小时实际上处于“无数据”状态，因此分配主要受先验支配，在纳入反馈前浪费大量曝光。我们提出了动态先验汤普森采样，这是一种有原则的先验设计，可显式控制新臂胜过当前 incumbent 获胜者的概率。我们的核心贡献是为先验均值提供了一个二次闭式解，确保在引入时满足 $P(X_j > Y_k) = \varepsilon$，使探索强度与运营基准保持一致，同时保留汤普森采样的贝叶斯更新。与强制探索启发式方法不同，我们的方法内嵌于汤普森采样决策规则中，并能随后验分布的演变平滑调整。在蒙特卡洛验证、离线批量仿真以及一项面向服务数百万用户的缩略图个性化系统的大规模在线实验中，动态先验实现了精确的探索控制并提升了效率。在线实验中，该方法显著提高了成功指标“合格播放完成率”，同时减少了遗憾曝光。这些结果表明，在批量部署约束下，精心设计的先验能够大幅降低冷启动成本，从而更快地发现优质内容，避免表现不佳的内容长时间过度曝光。

允许为个人或课堂使用而免费制作本作品的全部或部分数字副本或纸质副本，前提是这些副本不得用于营利或商业用途，且每份副本的首页均须附上本声明和完整引文。对于本作品中由作者以外的其他方拥有的组成部分，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，须事先获得特别许可和/或支付费用。请联系 permissions@acm.org 申请许可。*ACM KDD* 端到端客户旅程优化研讨会 '*25*，© 2025年 版权归所有者/作者所有。出版权已授权给 ACM。
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM https://doi.org/10.1145/nnnnnnn.nnnnnnn

关键词

多臂老虎机、动态先验、汤普森采样、探索-利用、推荐系统、冷启动

ACM 参考文献格式：

Zhenyu Zhao、David Zhang、Ellie Zhao 和 Ehsan Saberian。2025年。面向推荐系统冷启动探索的动态先验汤普森采样。载于 *2025*年*8*月论文集（*ACM KDD*端到端客户旅程优化研讨会 '*25*）。ACM，纽约州纽约，美国，11页。https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

1.1 问题背景与部署挑战

冷启动探索是大规模推荐系统中最持久的挑战之一 [7, 11]。现代平台持续引入新内容，每个项目都必须获得足够的曝光，以便在尽量降低将流量从已有成熟赢家转移所产生的机会成本的同时，估计其质量。汤普森采样（TS）通过从贝叶斯后验分布中采样，并选择抽样结果最大的臂，为探索与利用之间的权衡提供了一种概念上简洁优雅的解决方案 [2, 10]。

尽管理论上颇具吸引力，生产环境中的汤普森抽样部署往往取决于一个看似不起眼、实则影响重大的设计选择：新臂的先验。常见的默认设置是均匀 Beta(1，1) 先验，它为尚未见过的项目赋予 50% 的成功概率。然而，在许多真实系统中，“获胜”内容的基准率远低于 50%，因此均匀先验会高估新项目质量。这种不匹配会导致对表现不佳的项目进行可预见且系统性的过度探索，从而降低参与度，并加剧反馈回路动态：曝光带来数据，数据又带来曝光 [8, 11]。

批量更新问题：均匀先验假设与实际运行状况之间的差距，在一种普遍存在的生产约束下尤其代价高昂：**存在流水线延迟的批量更新**。与理想化的老虎机场景不同，在后验会在每次交互后更新的情况下，大规模系统通常每隔几小时批量更新一次，以满足计算资源、数据处理流水线和稳定性要求 [3, 9]。在两次更新之间，新引入的臂可能会一直处于

从策略的角度看"实际上没有数据"的状态，即使交互数据不断累积也是如此。因此，先验会在较长时间内主导决策规则，而过于乐观的先验可能会在反馈纳入模型之前浪费大量曝光。

一个典型的部署流程说明了这一问题：(1) 在时间 t_0 引入一个新的缩略图，其先验为 Beta(1，1)。(2) 用户产生反馈，但学习管道会异步汇总数据。(3) 策略仅在下一个批处理时间 $t_0 + \Delta t$ 更新（通常晚 2–8 小时）。(4) 在整个 $[t_0, t_0 + \Delta t)$ 期间，新臂仍由其先验支配，因此乐观默认值会继续分配过多流量。(5) 如果该臂表现不佳，系统将在整个 Δt 内承受本可避免的用户影响和机会成本。

这种批量部署的现实，使一个看似无害的建模选择转变为首要的产品问题：在平台规模下，即使冷启动分配中只有轻微的校准偏差，也可能造成大量的绝对展示浪费，并带来可衡量的参与度损失。

后果：在批量更新下过于乐观的冷启动先验可能：(1) 因长时间曝光弱势项目而在批次周期内压低参与度指标；(2) 当劣质素材尽管持续累积负面反馈仍获得持续流量时，降低创作者收入与信任；(3) 在数百万并行老虎机实例上浪费曝光；(4) 将流量分散到被过度探索的劣质表现者上，从而延迟真正优质内容的发现 [8]。除即时奖励之外，这些效应还会放大早期噪声并形成自我实现的曝光模式，从而强化生态系统层面的偏差。

空白：现有补救措施并不完整。一些方法在选择先验时忽略了现有最优项性能分布，而元学习方法需要大量离线训练，且可能无法跨内容类别或产品界面迁移 [1, 6]。强制探索策略可以保证最低曝光，但会引入剧烈的策略不连续性，并需要人工调度与调参，这在非平稳性和延迟反馈下尤其脆弱 [3, 9]。关键在于，已有工作并未提供一种简单、内在的机制，能够（i）将冷启动探索校准到当前生产基线，并且（ii）在批量部署约束下保持可预测性。

符号说明。我们用 k 表示当前表现最佳的臂，用 ε 表示目标探索概率，用 r 表示先验强度参数，并用 j 索引新臂。其余符号依上下文定义。

1.2 我们的方法与关键创新

我们提出动态先验汤普森采样，这是一种先验设计方法，通过利用当前优胜臂的观测表现分布，**直接控制新引入臂的探索概率**。具体而言，我们为新臂 j 计算一个先验，使其汤普森抽样值以目标概率 ε 超过当前最佳臂 k 的抽样值：$P(X_j > Y_k) = \varepsilon$。

关键算法思想：我们推导出先验均值 q_j 的**二次闭式解**，使其在贝塔后验分布的正态近似下满足该概率约束。由此可实现高效计算，并在大样本下具有可预测的行为

量，并使有效 冷启动探索率成为可直接调节的量。与均匀先验不同，均匀先验嵌入了隐含且往往不切实际的成功率基线，而动态先验则使探索强度与当前运行状态相一致；与强制探索启发式方法不同，动态先验仍内在于汤普森采样决策规则，并随后验分布的更新而平滑过渡 [2]。

这对批量部署为何重要：当更新出现延迟时，先验会在数小时内占主导；动态先验会明确设计这一"先验主导"窗口，以避免可能表现较差者被长期过度曝光，同时保留足够的概率质量以发现真正强势的新项目。这使得冷启动行为在校准偏差代价最高的运行阶段既更安全（浪费有界），也更可解释（由 ε 控制）。

我们的贡献：

(1) 我们推导出用于计算动态先验的闭式解，可在汤普森采样中实现探索概率控制

$$P\ X_j > (\ ),\quad Y_k\ = \varepsilon$$

并针对生产系统中的批量更新约束进行了专门设计。(2) 我们进行了广泛的蒙特卡洛验证，结果表明，在跨越四个数量级的样本量范围以及多种成功率基线下，实际探索概率与设计值的偏差均不超过 0.01。(3) 我们通过批量模拟证明，与均匀先验相比，动态先验可将累积奖励最多提高 9.5%，且其表现可与固定时域强制探索基线 [3, 9] 相媲美。(4) 我们在一项服务数百万用户的大规模在线实验中验证了其生产可行性：在真实世界的延迟和批量更新条件下，合格播放完成率最高提升 +0.20%，遗憾曝光相对减少 21% [4, 5]。

2 汤普森采样背景

汤普森采样（TS）是一种用于多臂老虎机问题的贝叶斯算法，它会维护各臂成功率的后验分布，并从这些分布中进行采样以作出决策 [2, 10]。在伯努利奖励场景中，每个臂 i 都会维护其成功率的贝塔后验分布 Θ_i：

$$\Theta_i \sim \mathrm{Beta}(\alpha_i, \beta_i),$$

其中，α_i 和 β_i 分别表示成功次数和失败次数，$n_i = \alpha_i + \beta_i - 2$ 表示观测总数，观测成功率为 $\hat{p}_i = \frac{\alpha_i - 1}{n_i}$。

在每个决策轮次 t，TS 都会从各后验分布中进行采样：

$$\theta_i^{(t)} \sim \mathrm{Beta}(\alpha_i, \beta_i) \quad \text{for all arms } i,$$

并选择采样值最高的臂：

$$i_t = \arg\max_i \theta_i^{(t)}.$$

在观测到来自所选臂 i_t 的奖励 $R_t \in \{0, 1\}$ 后，后验分布更新：

$$\alpha_{i_t} \leftarrow \alpha_{i_t} + R_t, \qquad \beta_{i_t} \leftarrow \beta_{i_t} + (1 - R_t).$$

设 k 表示当前表现最佳的臂（最高观测成功率为 $\hat{p}_k$），并具有 n_k 个观测值。当新臂 j 在没有历史数据（$n_j = 0$）的情况下被引入时，先验的选择会显著影响上下文推荐场景 [7] 中的探索行为与系统性能。

3 方法：动态先验推导

3.1 设计理论依据与实验问题

我们的实验设计旨在回答三个基本问题：(1) 我们能否推导出一个可保证精确控制探索概率的闭式解？(2) 这一理论保证能否转化为实际性能提升？(3) 该方法在各种不同的运行条件下表现如何？

我们方法的核心洞见是，探索强度应根据现有内容的性能基线进行调整，而不是假设先验固定不变。我们通过一个有原则的数学框架实现这一点，该框架可直接控制概率 $P(X_j > Y_k) = \varepsilon$。

3.2 数学框架

直观方法。与其使用可能与实际运行环境不匹配的固定先验，不如根据当前表现最佳的臂 k 的数据分布，动态计算新臂 j 的先验。关键思路是设置新臂的先验，使其与最佳臂相比时，汤普森采样会以预先指定的目标概率 ε 向其分配流量。

设臂 k 是当前表现最佳的臂，拥有 n_k 个观测值，观测成功率为 $\hat{p}_k$。对于新臂 j，我们定义先验 $\text{Beta}(\alpha_{prior,\,j},\ \beta_{prior,\,j})$，其中：

$$\alpha_{prior,j} = n_k r q_j, \quad \beta_{prior,j} = n_k r(1 - q_j).$$

其中，r 控制先验强度（有效样本量 $n_k r$），而 q_j 是我们需要推导的先验均值。

3.3 核心约束与推导

我们要求 $P(X_j > Y_k) = \varepsilon$，其中：- $X_j \sim \text{Beta}(n_k r q_j,\ n_k r(1-q_j))$ 表示新臂 j- $Y_k \sim \text{Beta}(n_k \hat{p}_k + n_k r q_j,\ n_k(1-\hat{p}_k) + n_k r(1-q_j))$ 表示最佳臂 k，并加入先验影响以进行公平比较

当 n_k 较大时，为提高计算效率，我们将两个分布都近似为正态分布：

$$X_j \approx \mathcal{N}\left(q_j, \frac{q_j(1-q_j)}{n_k r}\right), \quad Y_k \approx \mathcal{N}\left(\frac{\hat{p}_k + r q_j}{1+r}, \frac{\mu_{win}(1-\mu_{win})}{n_k(1+r)}\right)$$

其中 $\mu_{win} = \frac{\hat{p}_k + r q_j}{1+r}$。

约束 $P(X_j - Y_k > 0) = \varepsilon$ 推导得：

$$\frac{\mu_{new} - \mu_{win}}{\sqrt{\sigma_{new}^2 + \sigma_{win}^2}} = \Phi^{-1}(\varepsilon) = z_\varepsilon$$

两边平方并代入表达式，得到关于 q_j 的二次方程：

$$A_q q_j^2 + B_q q_j + C_q = 0$$

令 $T_{z\varepsilon} = z_\varepsilon^2$（其中 $z_\varepsilon = \Phi^{-1}(\varepsilon)$ 为正态分位数）且 $C_{nk} = n_k r(n_k + n_k r)$。系数为：

$$A_q = C_{nk} + T_{z\varepsilon}(1+r)^2(n_k + n_k r) + T_{z\varepsilon} n_k r r^2 \tag{1}$$

$$B_q = -2C_{nk}\hat{p}_k - T_{z\varepsilon}(1+r)^2(n_k + n_k r) - T_{z\varepsilon} n_k r r(1 + r - 2\hat{p}_k) \tag{2}$$

$$C_q = C_{nk}\hat{p}_k^2 - T_{z\varepsilon} n_k r \hat{p}_k(1 + r - \hat{p}_k) \tag{3}$$

确保 $q_j < \hat{p}_k$（当 $\varepsilon < 0.5$ 时）的保守解为：

$$\boxed{q_j = \frac{-B_q - \sqrt{B_q^2 - 4A_q C_q}}{2A_q}}$$

算法 1 具有理论保证

要求： 历史数据 $\{(n_i,\ s_i)\}_{i=1}^m$，目标概率 ε，强度 r

确保： 新臂 j 的先验参数 $(\alpha_j,\ \beta_j)$

$k \leftarrow \arg\max_i \hat{p}_i$ ▷ 寻找最佳臂

2：通过公式 3.3 计算 q_j

如果 $q_j \le 0$ 或 $q_j \ge \hat{p}_k$ **则**

$q_j \leftarrow \varepsilon \cdot \hat{p}_k$ ▷ 回退策略

返回 $(\alpha_j, \beta_j) = (n_k r q_j, n_k r(1-q_j))$

保证：$|P(X_j > Y_k) - \varepsilon| \le O(1/\sqrt{n_k})$ —

4 理论验证：经验探索概率

4.1 验证设计与理论依据

任何探索策略的一个根本问题是，它在实践中能否实现其设计目标。我们进行了大量蒙特卡洛验证，以确认我们的动态先验在各种运行条件下都能达到目标探索概率 ε。这项验证直接检验了我们方法的核心主张：我们能够实现精确的探索概率控制。

我们评估经验概率 $P(\theta_{new} > \theta_{best}) \approx \varepsilon$，其中$\theta_{new}$和 θ_{best} 分别是从相应后验分布中抽取的样本。我们的验证涵盖：

参数空间：

- 已有臂成功率：$p \in \{0.005, 0.01, 0.02, 0.03, 0.05\}$
- 样本量：$n \in \{10_4, 10_5, 10_6, 10_7\}$ • 目标探索概率：$\varepsilon \in \{0.01, 0.03, 0.05, 0.1\}$ • 先验强度参数：$r \in \{0.01, 0.05\}$
- 蒙特卡洛样本：每种配置 100,000 个

验证过程： 对于每个参数组合，我们：(1) 使用公式 3.3 计算动态先验均值 q，(2) 从两个后验分布中生成蒙特卡洛样本，(3) 计算新臂优于最佳臂的经验概率，以及 (4) 将其与目标值 ε进行比较。

4.2 理论验证结果

图 1 表明，在所有测试配置中都具有很强的理论一致性，这证明我们的闭式解正确实现了预期的探索行为。

关键验证结果：

- **精确度**：在测试的全部 800 种参数组合中，平均绝对偏差为 < 0.01
- **统计严谨性**：95% 的测量值与目标值的偏差在两个标准误差以内
- **规模稳健性**：样本量跨越四个数量级（10_4 至 10_7）时，性能保持一致
- **基线独立性**：性能不受现有臂成功率影响，证实了其适应性

这项验证提供了有力证据，表明从业者可以放心地设定目标探索概率，并确信这些概率在部署时能够实现，从而弥补现有方法中的一项关键不足：探索强度难以预测或控制。

5 仿真评估

5.1 实验设计与问题

我们的仿真研究探讨了三个关键问题：（1）理论探索概率控制能否转化为实际性能提升？（2）我们的方法与最先进的替代方案相比表现如何？（3）分配模式揭示了哪些有关探索行为的见解？

我们对批处理推荐系统进行建模，以反映真实部署中的约束，包括系统延迟和非平稳内容引入。

仿真环境：

- 初始配置：5个臂，成功率（0.05, 0.06, 0.07 0.08 0.09,,）
- 新臂引入：一个故意设置的弱臂，在批次 p = 新引入
- 批次结构：共10个批次，每批次10,000拉取次数
- 统计稳健性：每种设置10次独立重复试验

这个故意表现较差的新臂（$p_{\text{new}} = 0.01$）用作压力测试：每种方法能否在限制过度探索的同时仍保持发现能力？

比较的方法：

(1) **动态先验（所提出的方法）**：探索目标为 $\varepsilon \in \{0.01,$ 0.03 0.05 0.10 $r \in$ 0.01 0.001 ,,}，先验强度为 {,} (2) **均匀先验（基线）**：标准 Beta(1,1) 先验 (3) **固定时域强制探索** [3, 9]：以概率 $\alpha \in \{0.01, 0.03, 0.05, 0.10\}$ 强制选择新臂，持续 $K \in \{1,2\}$ 个批次，然后恢复为标准 Thompson 采样

这种强制探索方法不同于经典的 ε-贪心方法，它会在预定时段内专门将探索力度集中于新臂，从而能够更直接地与我们的自适应方法进行比较。

5.2 性能结果与分析

执行摘要：我们的动态先验方法在各种参数设置下都能获得最高的累积奖励；其中，最佳配置（$\varepsilon = 0.01$，$r = 0.001$）的累积奖励达到 8797.67，高于均匀先验的 8034.38，提升了**9.5%**，彰显了有原则的探索控制在实践中的价值。

表 1 展示了包含统计指标的全面结果。实验发现揭示了几个重要见解：

- **统计显著性**：不重叠的置信区间以高度置信度证实了动态先验优于均匀基线 • **参数敏感性**：当新臂确实为次优时，保守探索（$\varepsilon = 0.01$）表现最佳，验证了探索强度控制的重要性
- **竞争性表现**：最佳强制探索配置接近动态先验的表现，但在不同参数组合之间缺乏一致性
- **稳健性**：动态先验在不同 r 值下均保持出色表现，表明其具备灵活的实际部署能力

5.3 分配行为分析

图3–5所揭示的分配模式，为理解不同探索策略之间的根本差异提供了洞见：

动态先验行为：图3展示了受控的探索过程，随着证据不断积累，探索迅速减弱，从而实现最佳的探索—利用平衡。该方法表现出自适应行为：随着新臂表现不佳逐渐显现，探索强度自然降低。

均匀先验的局限性：图4显示，由于采用了乐观的 Beta(1，1)先验假设，与低成功率运行环境不匹配，导致对表现不佳的新臂持续过度分配。

强制探索模式：图5展示了探索与利用阶段之间明显的时间边界，这与我们的动态先验方法平滑的自适应行为形成对比。

6 在线评估

为验证动态先验框架，我们在一个全球游戏平台上开展了一项为期20天的大规模在线实验。评估重点关注“飞行中插入”用例，即在不中断现有臂服务的情况下，将新的缩略图（作为新臂）添加到活跃的多臂老虎机实例中。

6.1 实验设计与因果有效性

不同于标准的用户随机化A/B测试，本实验采用了**游戏级随机化** 以防止数据泄漏，因为探索训练数据会在单个游戏上下文中的各处理组之间共享。

- **对照组（C）：** 当前生产逻辑。引入新缩略图后，算法会进行硬重置，丢弃现有臂的所有后验历史数据，并使用 Beta(1，1)先验重新初始化。
- **处理组1（T_1）：** 采用保守探索的动态先验（$\varepsilon = 3\%$）和先验强度（$r = 0.1$）。
- **处理组2（T_2）：** 采用适度探索的动态先验（$\varepsilon = 10\%$）和先验强度（$r = 0.1$）。

先验强度 $r = 0.1$ 经过策略性校准，以平衡先验稳定性与后验响应性。由于

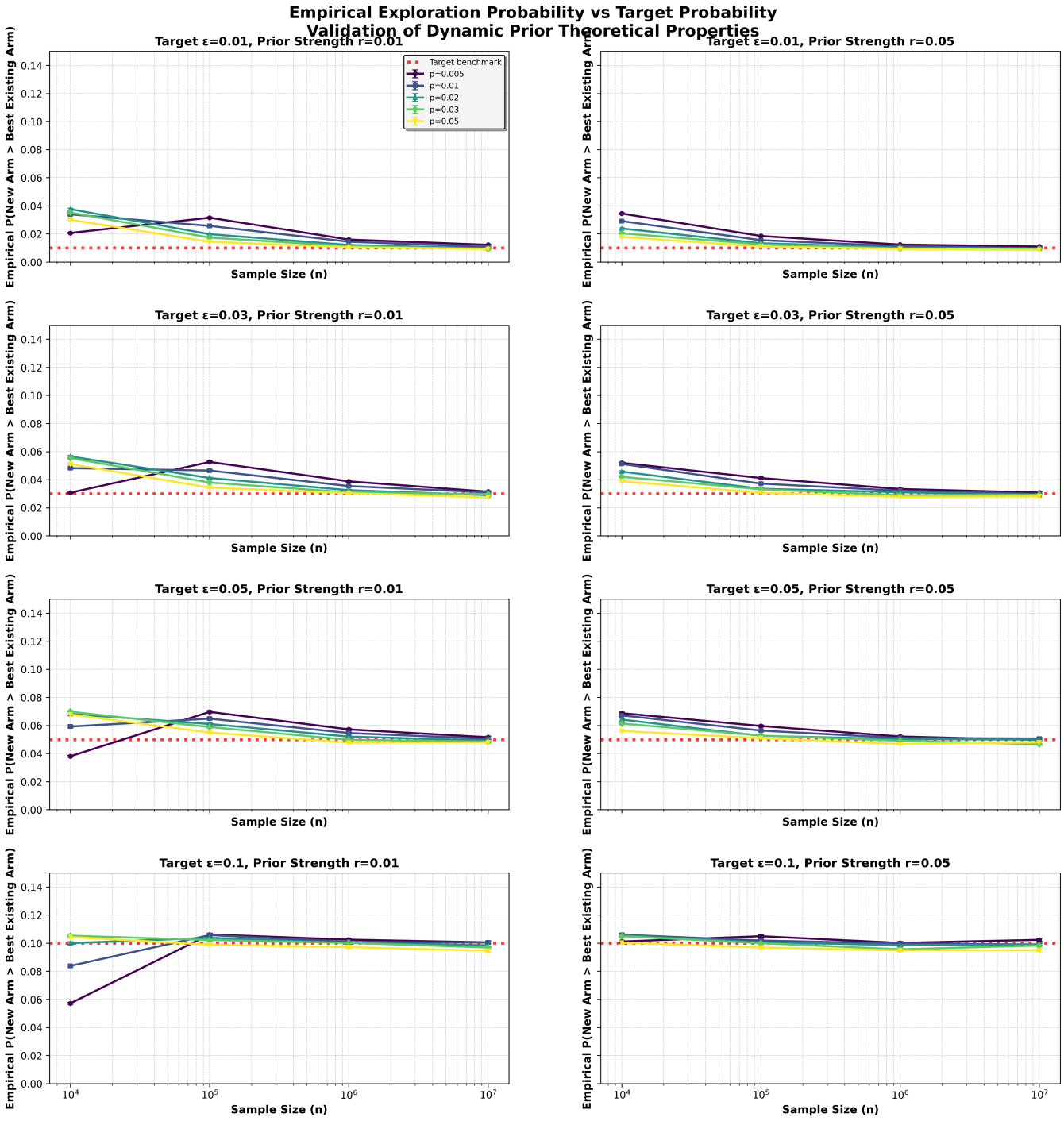

图 1：（理论验证；蒙特卡洛）在 800 种配置下，经验探索概率与目标值 ε 的关系。每个子图对应一个目标概率 ε 和先验强度 r，颜色表示现任成功率 p。红色虚线标示目标值，误差条表示蒙特卡洛标准误差。两者高度吻合，表明闭式动态先验在广泛的运行工况下都能实现预期的探索概率。

表 1：（仿真研究）最终累积奖励及其 95%置信区间（最优配置）。

方法	参数1	参数2	最终奖励	标准误差	置信区间下限	置信区间上限
动态先验	$\varepsilon = 0.01$	$r = 0.001$	8797.67	68.37	8663.65	8931.68
固定时域强制探索	$\alpha = 0.03$	$K = 1$	8761.67	14.44	8733.37	8789.96
固定时域强制探索	$\alpha = 0.01$	$K = 1$	8743.67	87.62	8571.92	8915.41
动态先验	$\varepsilon = 0.05$	$r = 0.01$	8739.33	46.63	8647.93	8830.74
固定时域强制探索	$\alpha = 0.05$	$K = 1$	8727.33	66.75	8596.51	8858.16
动态先验	$\varepsilon = 0.01$	$r = 0.01$	8725.67	21.61	8683.32	8768.01
动态先验	$\varepsilon = 0.05$	$r = 0.001$	8706.00	82.40	8544.50	8867.50
动态先验	$\varepsilon = 0.1$	$r = 0.01$	8705.67	54.14	8599.56	8811.77
均匀先验	参数无关		8034.38	7.75	8019.18	8049.57

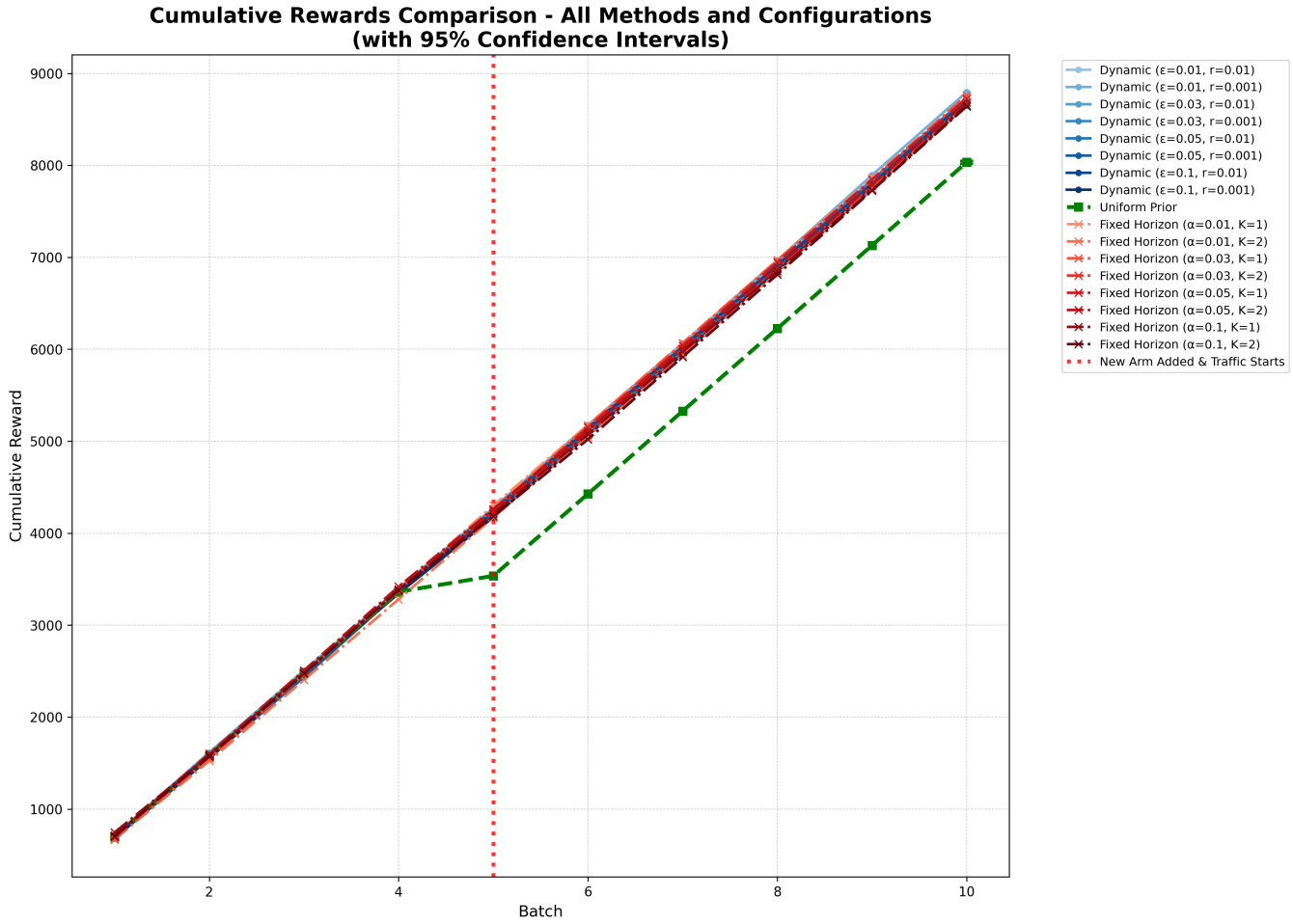

图2：（仿真研究）所有方法和配置的累积奖励轨迹。第5批次处的垂直虚线表示新臂的引入。DynamicPrior 配置始终优于基线方法，误差棒表示最终奖励的95%置信区间。

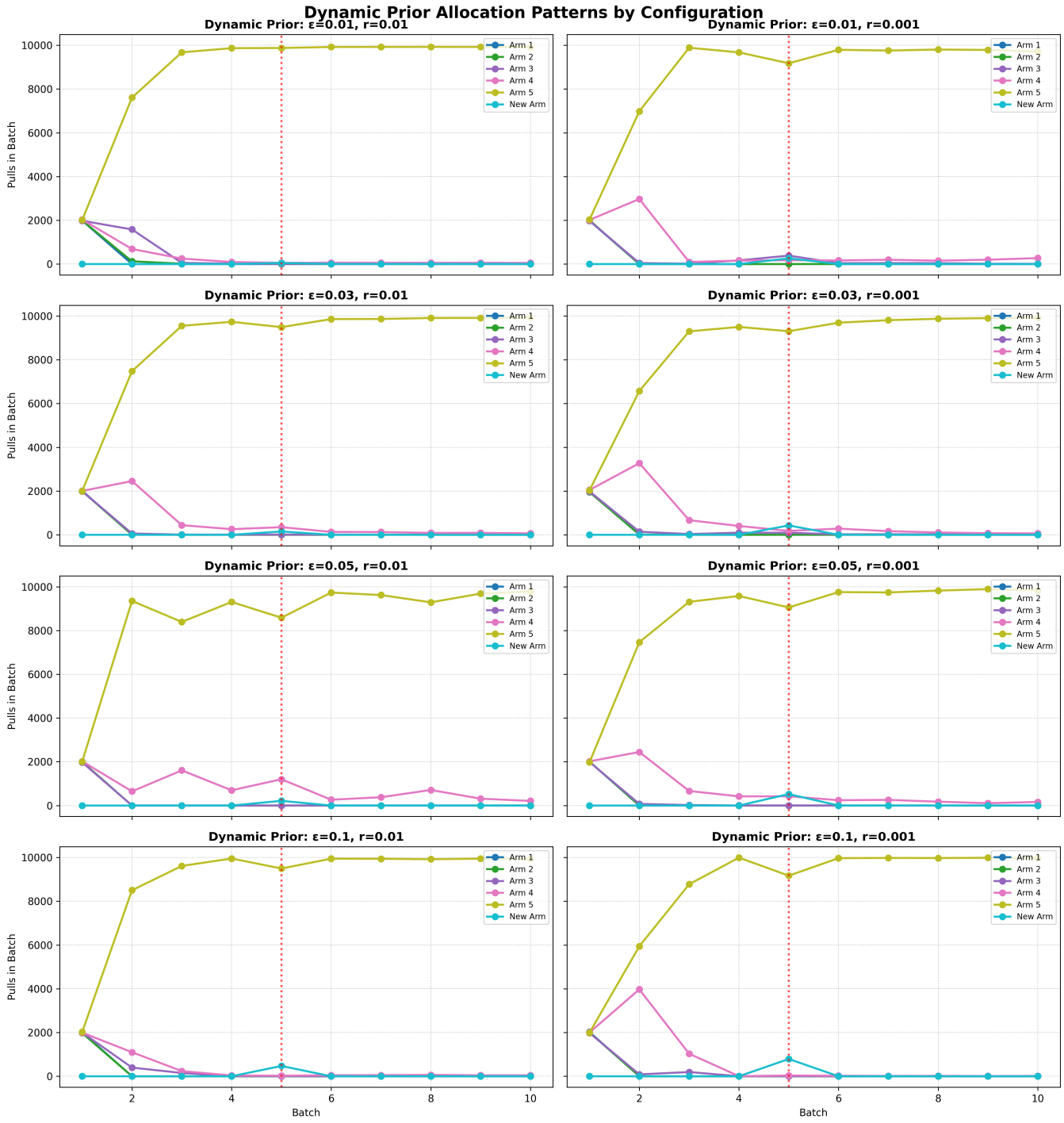

图 3：（仿真研究）DynamicPrior 在不同参数组合下的分配模式。每个子图展示了 10 个批次中，五个初始臂和新臂之间的每批次流量分布。随着证据不断积累，分配会平稳地调整，在减少分配给较弱的新臂的流量的同时，保持校准探索。

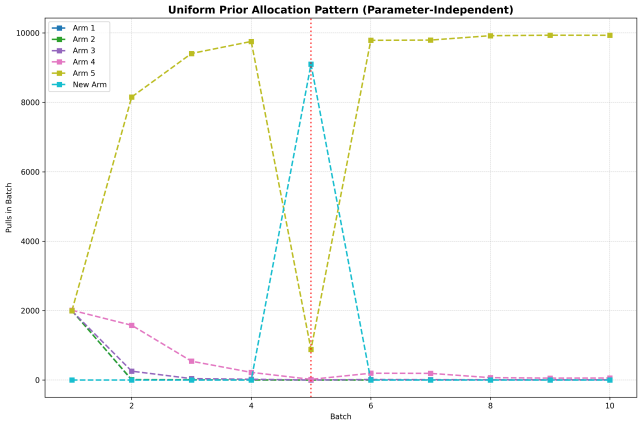

图 4：（仿真研究）均匀先验分配模式显示，由于对弱势新臂的乐观假设，分配给它的流量持续过多。此现象说明了均匀先验与低成功率运营环境之间的不匹配，这正是采用动态先验的原因。

算法基于历史数据的 7 天（168 小时）滑动窗口运行，有效样本量 $n_k \times r$约代表 16.8 小时的观测值。这一强度提供了稳健的缓冲，可防止先验被短期噪声过早覆盖——鉴于我们的生产流水线存在 3–7 小时的数据 SLA 延迟，这一点至关重要——同时确保随着新证据不断累积，模型仍能在一天内适应并更新。后续分析中的实证观测证实了这种平衡：获胜臂通常会在 24 小时内获得多数流量，而次优臂也会在相近的时间内受到抑制。

为确保游戏级分配下的因果有效性，我们采用按游戏曝光加权的**双重差分（DiD）**回归，并控制游戏特定因素和时间混杂因素。主要成功指标为**合格播放完成率（QPTR）**，定义为当会话时长超过预设时间阈值时，某次曝光对应一次播放完成。实验前的 A/A 测试确认各臂之间不存在显著偏差（$p > 0.05$）。

6.2 主要结果：参与度与效率

两个处理组的 QPTR 均取得了统计显著的提升，验证了保持后验连续性能够降低探索成本这一假设。表 2 总结了关键绩效指标。

表 2：（在线实验）飞行中插入的 QPTR 提升及收敛指标。

Arm	QPTR 提升	p-值	遗憾曝光量	收敛
C	–	–	64.1%	基线
$\mathcal{T}_1$	+0.20%	0.034	52.1%	Slow
$\mathcal{T}_2$	+0.19%	0.016	50.6%	Fast

探索效率：我们的分析中有一项关键指标是遗憾曝光——分配给非获胜臂的曝光占比，在收敛后回溯计算。$\mathcal{T}_2$ 与对照组相比，遗憾曝光**相对降低了 21.0%**（50.6% 对比 64.1%）。这一效率源于模型利用先验知识的能力；如果现有臂已经是明显的获胜臂，算法便能避免生产重置所导致的浪费性"重新发现"阶段。

6.3 DeepDive：探索动态

为了选择最优超参数 ε，我们分析了统计效率与生态健康之间的权衡。

1. 新缩略图突围率：我们测量了新引入的缩略图成功成为获胜臂的频率。$\mathcal{T}_1$（71.65%）与$\mathcal{T}_2$（71.73%）未显示出统计学上的显著差异（$p = 0.984$），这表明低至3%的初始 ε 足以识别高表现内容。

2. 开发者感知：作为面向开发者平台，创作者满意度是一项关键目标。定性分析表明，$\mathcal{T}_1$ 中3%的初始探索率可能被开发者感知为"探索不足"，从而在查看老虎机流量仪表板时产生挫败感。这主要是因为3%的分配在仪表板图表中几乎难以辨认，而10%的分配则能提供清晰可见的探索信号。如图8和图9所示，$\mathcal{T}_2$ 通过为新创意素材提供更充分、更可观察的"启动"窗口，更贴近开发者的预期。

6.4 生产上线

基于在复杂场景下更快的收敛速度和更高的开发者满意度评分，我们将 $\mathcal{T}_2$ r T0.1

（$\varepsilon = 10\%$）推向生产环境。此配置保持了动态先验带来的参与度提升，同时确保算法对创作者社区而言依然稳健且透明。图 6 至图 9 展示了曝光分配如何根据不同实验水平下的实证表现进行动态调整。

7 结论与未来方向

本文针对大规模推荐系统中冷启动探索这一关键挑战展开研究，尤其关注批量生产更新和反馈延迟所带来的约束。我们提出了一种用于汤普森采样的动态先验框架，以自适应贝叶斯方法取代无信息均匀先验。通过推导二次闭式解，确保新臂能够以预先指定的目标概率 ε与成熟的赢家竞争，我们为汤普森采样框架内的精确探索概率控制提供了首个内在机制。

我们的理论验证和离线模拟表明，动态先验能够在多样化的运行条件下实现高精度控制。此外，我们大规模的在线

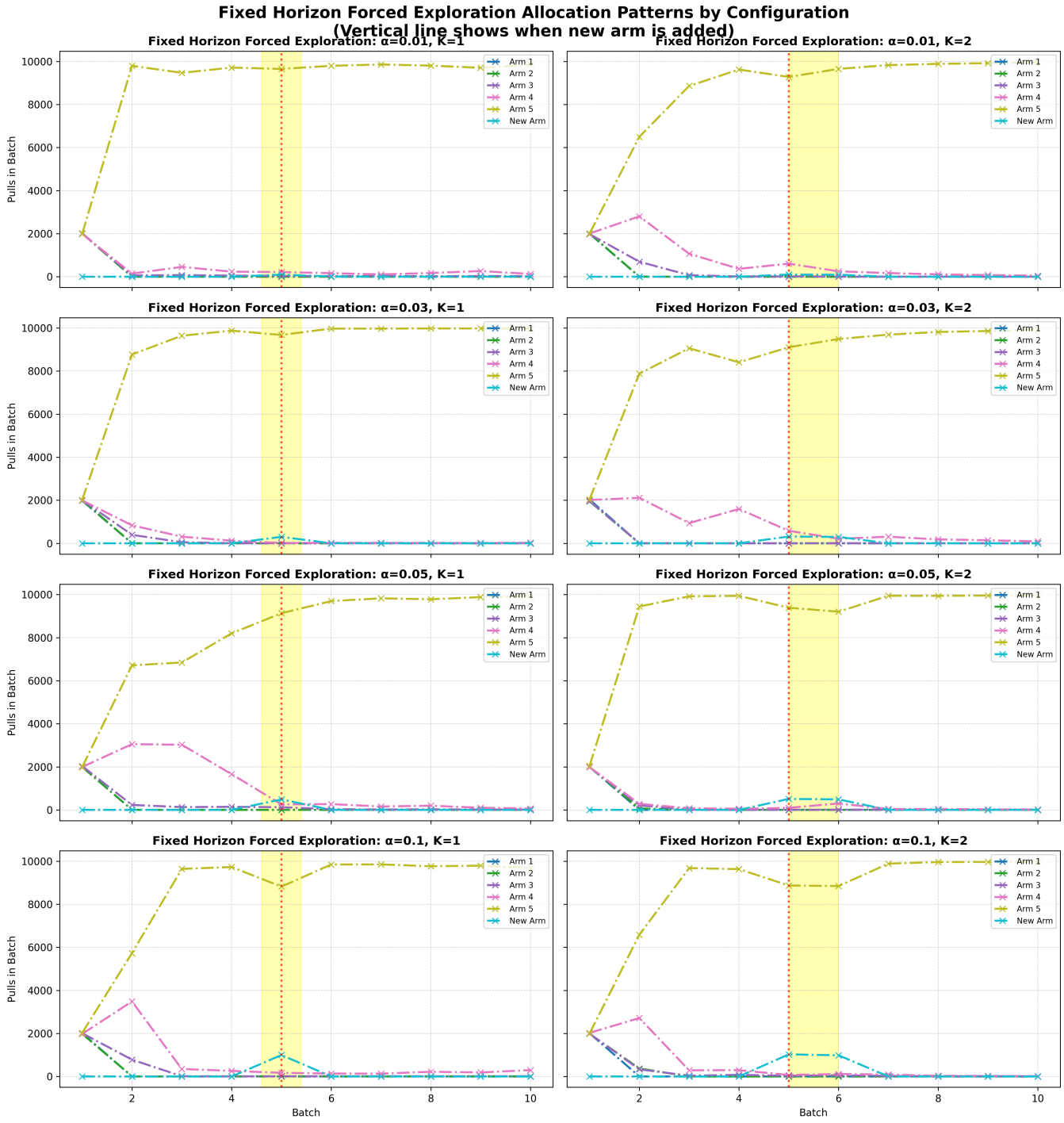

图 5：（仿真研究）不同参数组合下固定时域强制探索的分配模式。黄色区域表示强制探索时段，此时新臂会获得保证流量。由此产生的明显相变，与动态先验平滑的、由后验驱动的自适应形成对比。

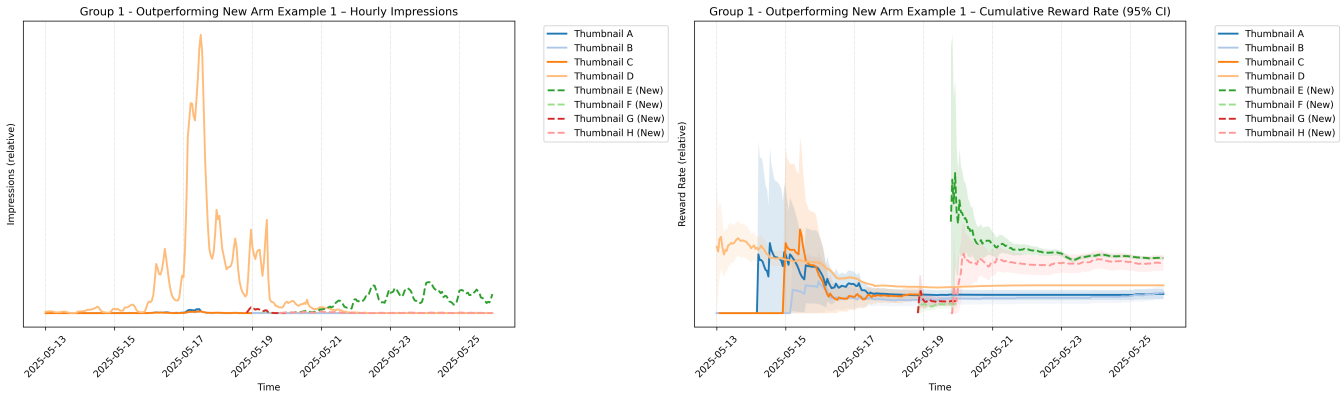

图 6：（在线实验；示例 $\mathcal{T}_1$）新臂最终胜出。新加入的缩略图（虚线）获得受控探索流量 $\varepsilon = 3\%$，随后随着证据不断积累而获得更高的流量份额，最终成为主导选择。

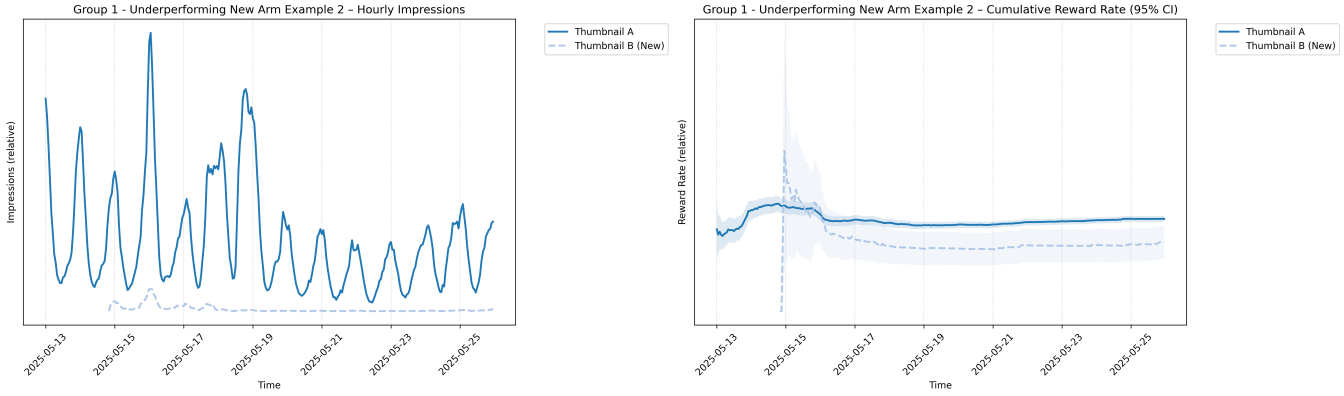

图 7：（在线实验；示例 $\mathcal{T}_1$）新臂表现不佳。新缩略图（虚线）最初获得探索流量，但随着证据表明其表现不如现有选项，分配给它的流量份额迅速下降。

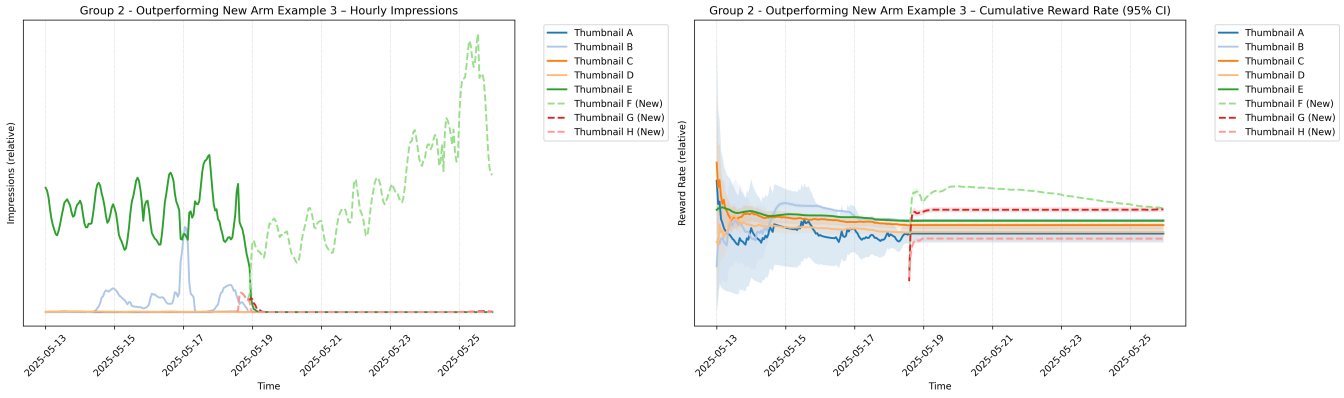

图 8：（在线实验；示例 $\mathcal{T}_2$）更高探索率带来更快发现。与 $\mathcal{T}_1$，相比，更高的 $\varepsilon = 10\%$会向新臂分配更多初始流量，从而加快识别其卓越性能。

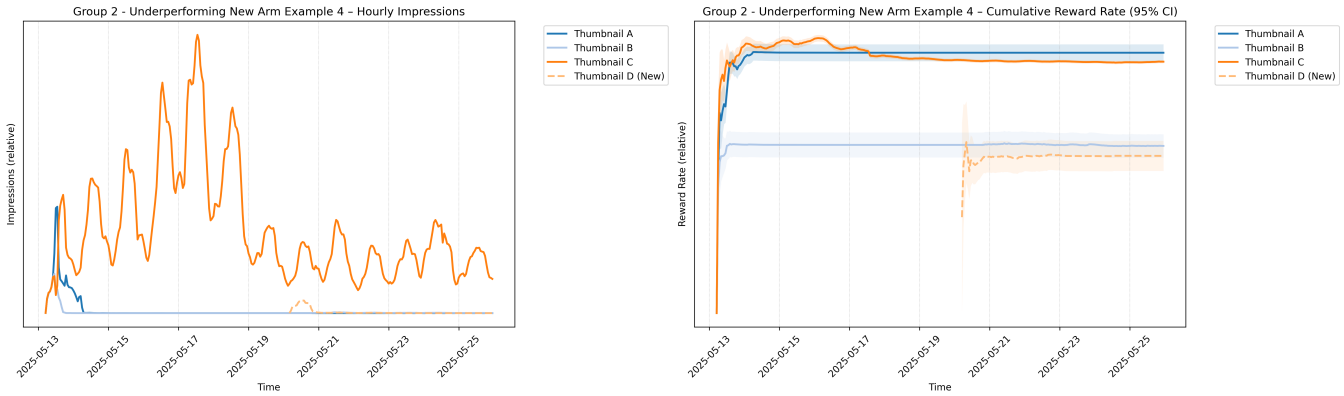

图 9：（在线实验；示例 $\mathcal{T}_2$）对较弱的新臂进行稳健抑制。尽管初始分配比例较高（10%），但在观察到表现不佳后，系统迅速减少其流量，这表明即使在更激进的探索策略下，也能有效控制遗憾。

在一个面向数百万用户的生产环境缩略图个性化系统上进行评估，证实了我们方法的实际效用。采用动态先验后，合格通关率（QPTR）在统计学上显著提升了 +0.19%，遗憾曝光则相对减少了 21.0%。

除了量化收益之外，我们的研究还凸显了平台生态系统中利益相关者协同的重要性。该算法使调整探索率成为可能，从而在算法效率与开发者可观测性之间取得平衡，并确保发现过程对内容创作者保持透明且令人满意。

未来的研究将探索把这一方法与上下文老虎机架构相结合，以利用高维侧信息，同时维持本研究确立的精确探索保证。最后，我们计划研究多目标动态先验，以同时优化参与度、公平性和长期生态多样性。

参考文献

[1] Hamsa Bastani、Mohsen Bayati、Mohammad Khosravi 和 Khashayar Razavi。2021。上下文老虎机的近乎无探索算法。《管理科学》67, 3 (2021), 1329–1349。[2] Olivier Chapelle 和 Lihong Li。2011。汤普森采样的实证评估。载于神经信息处理系统进展。2249–2257。[3] Qian Han、Aditya Baveja、Alekh Agarwal 和 Sandeep Singh。2024。延迟反馈老虎机问题中的强制探索。载于*AAAI*人工智能会议论文集，第38卷。13897–13905。[4] Ron Kohavi、Alex Deng、Brian Frasca、Toby Walker、Ya Xu 和 Nils Pohlmann。2013。大规模在线对照实验。载于第*19*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集（美国伊利诺伊州芝加哥）（*KDD '13*）。ACM，纽约，NY，美国，1168–1176。doi:10. 1145/2487575.2488217[5] Ron Kohavi、Diane Tang 和 Ya Xu。2020。值得信赖的在线对照实验：*A/B*测试实用指南。剑桥大学出版社。[6] Branislav Kveton、Zeyu Wen、Csaba Szepesvári 和 Cheng Wen。2021。元汤普森采样。载于第*38*届国际机器学习会议论文集（《机器学习研究论文集》，第*139*卷）。5865–5875。[7] Lihong Li、Wei Chu、John Langford 和 Robert E Schapire。2010。个性化新闻文章推荐的上下文老虎机方法。载于第*19*届万维网国际会议论文集。661–670。[8] Masoud Mansoury、Hossein Abdollahpouri、Mykola Pechenizkiy、Bamshad Mobasher 和 Robin Burke。2020。推荐系统中的反馈回路与偏差放大。载于第*29*届*ACM*信息与知识管理国际会议论文集。2145–2148。

[9] Yuning Mao、Chi Yu、Zhi Wang 和 Chao Wang。2019。上下文老虎机的批量汤普森采样。载于*IEEE*国际数据挖掘会议论文集。410–419。[10] W. R. Thompson。1933。根据两组样本的证据，一个未知概率超过另一个未知概率的可能性。*Biometrika* 25 (1933), 285–294。[11] Xinyang Yi、Liang Hong、Ed Chi、Shuang Kang、Jun Guo、B. Bělohrad、Yifei Shen 和 Pei Xu。2023。在线匹配：面向大规模推荐的实时老虎机系统。载于第*17*届*ACM*推荐系统会议论文集。481–491。

符号汇总

关键符号：

- i, j, k：臂索引（通用臂 i、新臂 j、表现最佳的臂 k）
- Θ_i：臂 i真实成功率的随机变量
- $\theta_{(t)\,i}$：在第 t 轮从 Θ_i 的后验分布中抽取的样本
- α_i，β_i：臂 i 的贝塔后验分布参数
- n_i：臂 i的观测值数量
- $\hat{}p_i$：臂 i的观测成功率
- r：相对于最佳臂经验的先验强度因子
- ε：新臂的目标探索概率
- z_ε：正态分位数 $\Phi^{-1}(\varepsilon)$
- $T_{z\varepsilon}$：正态分位数的平方 z_ε^2
- q_j：新臂的派生先验平均成功率 j
- n_k：当前最佳臂的观测值数量 k
- $\hat{}p_k$：当前最佳臂的观测成功率 k
- X_j j：新臂成功率的随机
- Y_k k
变量：最佳臂成功率的随机变量