# 奖励什么，就会学到什么：比较公共场所人机交互中在线语音策略优化的奖励

Sichao Song*†Yuki Okafuji*†
CyberAgent 东京，日本 song_sichao@cyberagent.co.jp
okafuji_yuki_xd@cyberagent.co.jp

Kaito Ariu
CyberAgent 东京，日本
kaito_ariu@cyberagent.co.jp

Amy Koike 威斯康星大学麦迪逊分校 威斯康星州，美国
ekoike@wisc.edu

arXiv:2601.01969v1 [cs.RO] 5 Jan 2026

## 摘要

在开放多样的环境中，为对话服务机器人设计既高效又易于接受的策略并非易事。与固定的手动调优参数不同，在线学习能够适应非平稳条件。本文研究如何在真实环境中调整社交机器人的语音策略。在为期12天的现场部署中，我们记录了超过1,400次公众互动，并将在线策略优化建模为多臂老虎机问题，使用汤普森采样在六种动作中进行选择。这些动作由语速（慢/正常/快）和详略程度（简洁/详细）定义。我们比较了三种互补的二元奖励——$R_u$（用户评分）、$R_c$（对话结束）和 $R_t$（≥2 轮次）——并表明每种奖励都会产生不同的臂分布和互动行为。我们还通过离线评估补充在线结果，并利用视频标注数据分析情境因素（如人群密度、群体规模）。综合来看，我们提炼出一系列实用的设计经验，可用于在真实公共场所人机交互场景中部署语音策略在线优化。

## CCS 概念

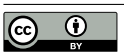
•**以人为本的计算** → **人机交互中的实证研究**；**实地研究**；用户模型。

## 关键词

人机交互（HRI）、对话机器人、汤普森采样、多臂老虎机（MAB）、语音参数优化、实地实验、服务机器人

**ACM 引用格式：**
Sichao Song、Yuki Okafuji、KaitoAriu 和 Amy Koike。2026年。《你奖励什么，就学到什么：比较公共场所人机交互中在线语音策略优化的奖励》。载于第 *21* 届 *ACM/IEEE* 人机交互国际会议论文集（*HRI '26 19，2026*年，*16*12），*3* 月，苏格兰爱丁堡，英国。ACM，纽约，纽约州，美国，页码。https: //doi.org/10.1145/3757279.3785589

*两位作者对本研究的贡献相同。
†同时也与大阪大学合作。



本作品采用知识共享署名 4.0 国际许可协议授权。*HRI '26,* 爱丁堡，苏格兰 英国© 2026年版权所有，版权归所有者/作者所有。ACM 国际标准书号 979-8-4007-2128-1/2026/03 https://doi.org/10.1145/3757279.3785589

## 1 引言

在线学习越来越多地应用于人机交互（HRI），以便在交互过程中实现个性化和策略适应。对于社交机器人而言，这不再局限于固定的人工调试脚本：机器人可以实时学习并调整其动作选择。这种即时适应能带来切实的好处，尤其是在真实环境中：更高的任务成功率、更顺畅的轮流交谈、更好的用户满意度，以及无需持续重新调试所带来的较低操作人员负担。在实践中，人机交互中的在线学习通过多臂老虎机（MAB）或强化学习（RL）实现；这些方法已反复证明，能够有效地根据用户偏好调整机器人行为，应用领域包括偏好学习、说服式推荐、辅导和社会交互控制 [4, 11, 24–26, 29]。

从概念上看，多臂老虎机与完整的强化学习不同：它们在没有显式状态模型的情况下，通过平衡探索与利用，在有限的试验次数内最大化累积奖励。这一特性，加上针对偏好漂移或时变偏好的扩展，使多臂老虎机非常适合用于真实环境中的学习，因为用户具有异质性，且交互条件会随时间发生变化 [8, 15, 22]。

然而，大多数关于将多臂老虎机应用于社交机器人的实证评估，仍处于封闭或半受控环境中。具有代表性的研究包括：用于适应运动偏好的对决老虎机 [29], ；根据显式反馈调整语言风格的辅助陪伴机器人 [26], ；用于个性化教育聊天机器人的上下文老虎机 [4], ；形式化人机辅助的辅助型多臂老虎机框架 [5],；以及用于多用户分配的公平性约束上下文老虎机 [8]。即使是近期针对机器人辅助喂食的人在回路上下文老虎机研究，虽然涉及真实用户，但其时间跨度较短且处于受控环境中，并非持续的公共场所部署 [2]。

相比之下，真正的真实场景研究——在自然环境中持续部署，并面向现场主动参与的用户——仍然十分有限。一项面向活动场景的劝导式饮品顾问研究表明，由多臂老虎机驱动的策略选择能够在快速变化的情境中运行 [25]。更广泛地说，关于“真实场景中的多臂老虎机”的在线实验文献记录了实际应用中的陷阱（非平稳性、延迟奖励、干扰）以及可推广至交互系统的实地验证策略 [22]。这些迹象表明，此类方案具有可行性，同时也凸显了公共场所人机交互在奖励塑形、曝光控制和情境化方面仍存在设计欠账。

综上所述，既有研究表明，基于多臂老虎机的社交机器人个性化在受控环境中能够取得成功，而在真实环境中的学习则必须应对分布偏移、多样化的用户行为，以及对预设交互流程不完全遵循的问题。A

因此，一个核心的开放性问题是：在这些条件下，奖励操作化如何塑造学习行为：哪些可观测信号应当获得奖励（例如任务成功、参与度、用户报告的满意度）？应如何将它们结合起来？在不违反社会规范或公平约束的情况下，应如何调节探索？ [8, 15, 21, 22]

尽管一些聚焦于奖励设计的研究比较了主观（显式）反馈与客观（隐式）反馈，并显示出将二者结合的益处 [21], ，此前的人机交互研究主要在受控环境中探索了这些理念。相关实例包括：在多次交互中利用显式用户评分进行偏好学习 [3]；仅在必要时询问用户澄清问题，同时利用隐式线索完成物体取回 [34]；将任务表现（显式）与任务参与度（隐式）相结合，以推动实时个性化的交互式强化学习 [32]；根据用户反馈调整机器人语言风格的社会感知强化学习 [27]；以及同时基于隐式的动作信息和显式的多模态通信进行联合规划的社会导航 [7]。尽管这些研究清楚地展示了融合反馈模态的益处，但相关演示仍主要局限于封闭环境；不同的奖励定义如何影响长期真实环境交互中的在线学习，仍有待深入探索。

本文面向真实环境中的商业设施，在社交机器人的对话语音策略中，系统比较了汤普森采样（TS）多臂老虎机中的多种奖励设计。已知商业设施中的用户具有多样化的属性，例如年龄；此外，用户在交互过程中的行为也会因其心理状态 [18], （如动机）而有所不同。这使得制定明确的动作选择规则变得困难。TS 直接表示奖励估计中的不确定性，并利用这种不确定性决定何时进行探索、何时进行利用。因此，它非常适合用户属性多样化的商业设施，同时还能避免策略依赖脆弱的人工设计选择规则。

我们在一家购物中心部署了一台服务机器人，为期 12 天（进行了超过 1,400 次公众互动），并调整两个维度——语速（慢/正常/快）和 详略程度（简洁/详细）——同时优化三种能够体现互补构念的二元奖励：$R_u$（用户评分）、$R_c$（对话结束）和 $R_t$ （≥2 轮次）。我们记录了每次交互，并通过视频对其社会情境（例如人群密度和群体规模）进行标注，从而能够在事后分析情境如何调节习得的策略。我们报告了：（i）每种奖励下的 TS 学习结果和后验臂偏好；（ii）用于量化臂×情境交互效应的广义线性模型。在汤普森采样中，臂是一种动作类型。在本实验中，每个臂表示机器人的语速 × 详略程度设置。基于这些发现，我们总结了公共场所人机交互情境在线优化的设计启示。

我们在真实环境中的公共场所部署对话服务机器人，致力于实现两个目标：

- **研究目标一：** 评估不同奖励定义如何引导社交机器人的语音策略进行在线学习。具体而言，我们在语速（慢/正常/快）和详略程度（简洁/详细）的 TS 中，对比 $R_u$（用户评分）、$R_c$（对话结束）和$R_t$（≥2 轮次）。
- **研究目标二：** 利用视频标注和广义线性模型，量化社会情境如何调节结果和臂效果；进而将这些规律转化为可操作的未来指导，助力情境感知在线优化。

本文提出了两项相辅相成的创新。

- 我们在一家购物中心部署了一台社交机器人，为期12天，展示了汤普森采样如何在超过1,400次公众互动中实时调整其语音策略。学习到的臂偏好因奖励设计而异，体现了真实环境中对构念敏感的策略选择。
- 我们提供了带视频标注的分析，利用广义线性模型量化情境调节作用，揭示影响绩效结果的因素。根据结果，我们提炼出针对公共场所人机交互的情境感知在线优化设计经验。

## 2 方法

### 2.1 汤普森采样（TS）

汤普森采样（TS）是一种处理探索与利用权衡的贝叶斯方法，最早可追溯至 Thompson 于 1933 年提出的方案：按照某个行动为最优的后验概率比例来选择行动 [31]。过去十年间，TS 获得了强有力的理论支持，包括与最佳可实现速率相匹配或接近的有限时间遗憾分析 [13], ，并已广泛应用于在线推荐和广告领域 [6]。全面的教程进一步介绍了 TS 在多臂老虎机文献中的位置，并展示了它如何扩展到上下文、非平稳和受约束设置 [28]。其主要优势包括：(i) 概念和实现简单；(ii) 由后验不确定性驱动探索，无需显式的乐观奖励项；以及 (iii) 能够直接适应现实世界的约束和部署实际情况，例如延迟奖励和批量更新。

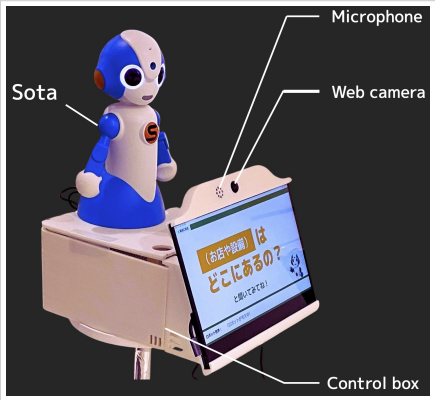
具有贝塔先验的伯努利奖励。在我们的设置中，每次交互都会产生一个二元结果 $r \in \{0, 1\}$（成功/失败），因此我们将臂 $a$ 的成功概率建模为 $\theta_a \sim$ 贝塔分布($\alpha_a$, $\beta_a$)。由于贝塔分布与伯努利似然共轭，在选择臂 $a$ 后观察到奖励 $r$，会通过以下方式更新后验分布：

$$\alpha_a \leftarrow \alpha_a + r, \qquad \beta_a \leftarrow \beta_a + (1 - r).$$

在第 $t$ 轮，汤普森采样会为每个臂独立抽取一个样本 $\tilde{\theta}_a \sim \text{Beta}(\alpha_a, \beta_a)$，并选择 $a_t = \text{argmax}_a \tilde{\theta}_a$。后验不确定性较高的臂在随机意义上更受青睐，从而实现原则性的探索。常见的无信息初始化包括 $\alpha_a$=$\beta_a$=1 （均匀分布）或较小的对称先验。在我们的部署中，我们先进行了一小段均匀预分配，再开始学习，以确保达到最低曝光量（见 §2.2）。算法1展示了伯努利奖励的标准循环。

我们使用汤普森采样在六个臂之间进行优化（慢速/常速/快速 × 简洁/详细）。以下工程设计选择使汤普森采样适用于我们的部署环境：

- **冷启动**。 我们通过均匀分配每个臂一定数量的初始交互，并使用这些观测值为 $(\alpha_a, \beta_a)$ 赋予初始值，来确保达到最低曝光量。
- **可选的非平稳性处理**。 为降低过时证据的权重，可以应用一个简单的遗忘因子 $\lambda \in (0, 1]$ 。




(a) 服务机器人系统“Sota”。




(b) 购物中心的实地场景。

**图1：Sota机器人，专为路线引导而设计，并在购物中心进行实地应用。它配备了集成显示屏。显示屏明确展示了机器人提供方向指引的能力，确保用户了解其功能。**

**算法1**伯努利奖励的汤普森采样

**要求：** 臂 A；贝塔先验 $\{(\alpha_a,\ \beta_a)\}_{a\in\mathcal{A}}$

1: **f或** $t = 1,2,\ \cdots$ **d**

o2：**对于所有** $a \in$ A **执行**3：采样 $\theta_a \sim$ 贝塔分布$(\alpha_a,\ \beta_a)$4：$a_t \leftarrow$ 取最大值$_{a\in\mathcal{A}}\,\theta_a$5：执行 $a_t$ 并观察 $r_t \in \{0,\ 1\}$（可能存在延迟）6：**更新：** $(\alpha_{a_t},\ \beta_{a_t}) \leftarrow (\alpha_{a_t} + r_t,\ \ \beta_{a_t} + 1 - r_t)$

至充分统计量：$\alpha_a \leftarrow \lambda\alpha_a + r, \beta_a \leftarrow \lambda\beta_a + (1-r)$（我们将 $\lambda$=1 作为默认 TS 报告结果）。

- **多个标签**。我们使用三种二元奖励更新算法：一种是基于好/坏判断的第三方判断（$R_c$，$R_t$），另一种是网站上的用户评分（$R_u$）。这些奖励在交互后立即给出，并在收到后应用更新。

总之，这些做法在应对非平稳性以及真实部署中的延迟反馈的同时，保持了 TS 的简洁性。

## 2.2 实验设计

本研究与一家购物中心合作，在其正常营业期间开展。部署和数据处理遵循机构伦理审查要求及场地协议。本研究已获大阪大学研究伦理委员会批准（参考编号：R1–5–9）。对于不愿参与并希望从视频数据中移除其影像的人员，本研究采用选择退出机制。

现场与任务。 我们在日本的一家购物中心开展了现场研究（为便于双盲评审，具体信息已匿名化）。如图 1 所示，社交机器人被安放在二楼和三楼楼层地图旁边。二楼和三楼同时运行两种实验条件；一种条件完成后，便实施另一种条件。机器人连续部署了 12 天，每天约 8 小时，直至完成所有实验条件。机器人的主要任务是路线引导，但它也经过设计，能够处理各种类型的交互，例如闲聊。

奖励设计。 以往研究在奖励设计中比较了来自用户的主观（显式）反馈和客观（隐式）反馈 [21]。然而，在真实环境中，用户行为各不相同：尽管社交机器人的预期角色是向导，它们却常被用作聊天伙伴。因此，难以明确界定客观指标。为此，我们首先为社交机器人对话任务准备了四项客观指标（对话结束、话语重叠、对话冲突和轮次数量） [12]，并开展初步观察，以确定这些指标是否是语音策略在线学习的重要方面。经过初步观察后，我们决定采用三种二元奖励（一个主观奖励和两个客观奖励）开展实验，以反映不同的运营目标：

- $R_u$ **(用户评分)：** 若交互结束后在7点量表上获得的单项满意度评分为 $\geq 6$，则视为成功；否则视为失败。
- $R_c$**(对话结束)：** 若交互达到明确的对话结束惯例（告别语和/或感谢语），则视为成功；否则视为失败。
- $R_t$**(≥ 2 轮次)：** 若对话至少包含两个轮次（用户与机器人之间的交替发言），则视为成功；否则视为失败。接触语句从机器人的接近触发问候语开始计数，成功要求最短序列为 R1（机器人） → U1 → R2 → U2；“轮次”指连续的说话者片段。

$R_u$ 直接反映用户的主观评估，而 $R_c$ 和 $R_t$ 则反映交互的客观方面。对于 $R_u$，我们将分数 $\geq 6$ 映射为成功。根据我们先前的实地研究，分数存在正向偏差；若将 5 视为“正面”，大多数交互都会被归类为成功。采用更严格的 $\geq 6$ 阈值，可以更均衡地划分正面与负面结果，并提高区分能力。其余两项客观奖励依据二元标准确定，以实现结果的均衡划分：对话结束惯例是否发生，以及对话是否至少包含两轮

该例程是否发生，以及对话是否至少包含两轮次。

我们的三种奖励条件（$R_u$，$R_c$，$R_t$）将显式反馈与隐式反馈落实到公共空间部署中，从而能够在受控条件下比较每种定义如何引导同一在线学习流程。不同于以往主要依赖显式评分 [3] 或在实验室式研究中结合显式与隐式信号的 HRI 表述 [7, 27, 32, 34], ，我们的研究在硬件、动作空间和环境均相同的条件下，对这些替代方案进行了并列评估。

数据收集。对于每种奖励条件，我们采用两阶段安排：先进行 30 次交互的冷启动阶段，为后验分布提供初始数据；随后在多臂老虎机运行期间收集 450 次用户交互。因此，每种奖励条件总计获得 480 次交互。

在冷启动阶段，我们进行了简短的均匀预分配：六个臂各拉取五次，收集奖励并更新每个臂的后验分布。然后，我们将所得的贝塔参数用作汤普森采样多臂老虎机的初始状态。这确保了各臂获得均匀的最低曝光量，并降低了早期学习错误的风险。

对于每次交互，我们都记录了时间戳、所选臂（语速、详略程度）、对该臂的评分以及多臂老虎机变量（例如先验/后验参数）。这些日志使我们能够重建学习曲线并汇总后验分布。除了日志之外，我们还在整个部署过程中录制了视频素材，以供标注。

## 2.3 服务机器人系统

我们使用了一款小型人形机器人“Sota”（Vstone株式会社），身高28厘米，外形稚气可爱。每只手臂有两个自由度（DoF），能够做出简单的手势（例如指向手势，用来表示“那个方向”）。Sota配有三个面部LED灯（眼睛和嘴巴），用于表达情绪提示；它还能转动身体，以实现注视行为。在语音方面，我们使用了Google Cloud语音转文本（STT）和文本转语音（TTS）API。对话内容生成和对话状态管理由OpenAI API（GPT-4系列）提供支持。安装在正面的显示屏会显示简短提示和指引，其中包括设施地图。

如图 1a 所示，Sota 下方的控制箱内装有一台运行行为控制器的迷你电脑。我们实现了一个完全自主的对话系统，由四个基本组件构成：识别器、对话管理器、动作管理器和模态管理器。具体而言，识别器使用 180° 鱼眼摄像头和 PoseNet [14] 识别最近的访客并推断其到访状态；对话管理器是一个有限状态控制器，在检测到访客时致以问候，并在访客靠近时接收语音，同时将大语言模型生成的内容基于精心整理的设施知识库和固定的“Sota”人设进行约束；动作管理器根据对话历史和知识库生成任务响应（例如路线引导）；模态管理器则触发约 40 种词语—手势映射（包括指向手势），以强化口头指令。

汤普森采样集成. 为实现语音策略的在线优化，我们实现了一个TS模块，在每次交互开始时使用汤普森采样，根据伯努利奖励 [6]选择臂（语速 × 详略程度）。我们设计了一个系统，使机器人的行为通过交互和评估实时更新。使用TS的学习流程如图2所示。奖励信号来自两个互补来源：（i）现场用户自我报告，用于 $R_u$；以及（ii）第三方成功/失败判定，用于 $R_c$/$R_t$。

在情形（i）中，用户会在交互结束后立即填写问卷，并使用问卷平板电脑通过 7 点李克特量表对交互满意度进行评分。评分达到 6 分或以上视为成功，否则视为失败；结果会立即发送至机器人系统。在情形（ii）中，监控界面会实时传输机器人的摄像头画面。第一、第二和第四作者分别实时独立观察交互，并在交互结束后根据预先制定的评估标准按下“成功”或“失败”。通过多数表决得出的成功或失败标签会立即发送至机器人系统。

发送给机器人系统的奖励结果会立即触发 TS，更新该次交互所用臂的后验分布。随后，根据更新后的后验分布确定机器人的语音策略，并立即反映到机器人系统中。用户下次交互时，系统便会执行更新后的行为。这种在线方法平衡了探索与利用，使语音策略能够在真实环境中实时进行设备端自适应。

关于语音策略的具体设置，对于语速，我们使用 Google Cloud TTS 的语速乘数：慢=0.80、正常=1.20和快=1.60；对于详略程度，我们在 GPT 对话提示词中加入了每轮指令：简洁 = “请用简短精炼的句子作答。聚焦要点，避免不必要的展开。”；详细 = “请用较长、较详细的文字作答。包含具体示例和补充说明。”部署前，作者在设备端试用了全部六种模式（臂），并一致认为这些设置自然、安全，且足以在我们的实地实验中彼此区分。

## 2.4 评估与分析

老虎机性能。对于每种奖励条件（$R_u$/$R_c$/$R_t$），我们报告总体成功率和各臂的后验贝塔分布，以揭示学习到的臂偏好。

通常，评估老虎机算法时，会将累积奖励与均匀采样的结果进行比较；均匀采样会随机执行所有动作，作为基线。在本实验中，除了在三种奖励条件下运行机器人外，我们还在均匀采样条件下进行了实验。然而，三种奖励条件下实验的大部分数据是在工作日收集的，而均匀采样条件下的数据则主要是在节假日收集的。因此，我们认为，由于用户属性和行为模式存在显著差异，难以进行公平比较。因此，本研究仅比较了各奖励条件下学习到的语音策略（如 ROs 中所述），与均匀采样的比较结果则报告于附录中。

视频标注。除了分析对不同奖励作出响应而习得的行为外，我们还进行了事后分析，以考察各种情境因素如何调整习得的语音策略，从而探讨未来的指导方向。所有 1,400+ 次交互均进行了录像并编码，以获取

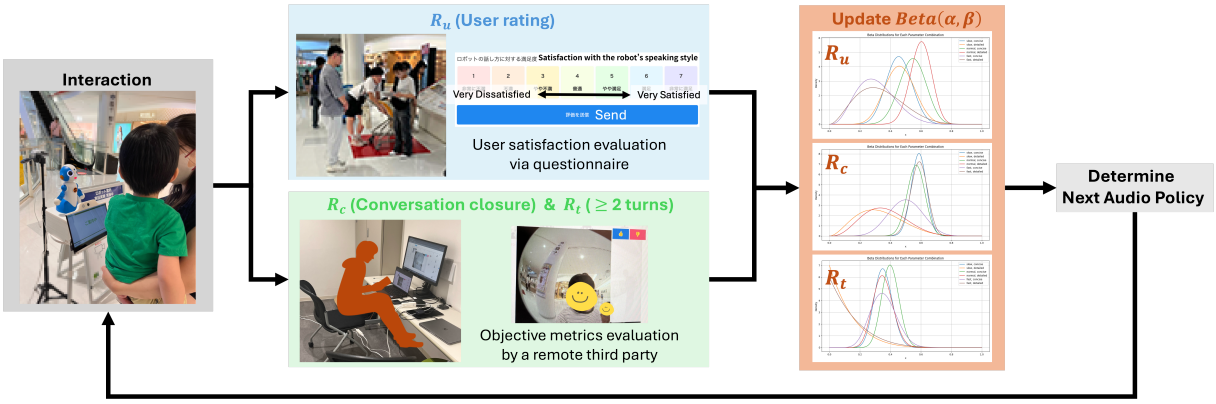



**图2：该图展示了机器人如何使用汤普森采样（TS）改变其语音策略。交互结束后，会提供三种奖励条件。在 $R_u$（用户评分）条件下，与机器人交互的用户通过问卷对机器人进行评价；而在 $R_c$（对话结束）和 $R_t$ ($\geq 2$ 轮次)条件下，则由远程第三方评估交互的客观指标。获得的评价作为奖励实时发送至机器人系统，机器人的下一步行为则由TS决定，并立即体现出来。**

上下文变量。我们将以下因素纳入上下文回归变量：

- **组**（独自/结伴）：目标用户是独自前来还是与同伴一同前来。人类群体的存在及其特征会系统性地影响人们在真实环境中与机器人互动的程度及对机器人的评价 [33]。
- **人群**（存在/不存在）：交互期间机器人附近是否有旁观者。旁观者的存在会改变人们提供帮助及进行交互的意愿（旁观者效应） [20]。
- **询问方向**（是/否）：用户是否询问了路线指引信息。寻路/对话需求（例如询问方向）会显著影响公共空间人机交互中的对话策略和结果 [10]。
- **动机**（功能/实验/好奇心/教育）：用户与机器人交互的表面原因。用户动机的差异会导致行为差异 [18]。

在我们的开放世界、多方场景中，我们有意将人口统计协变量（年龄、性别）排除在广义线性模型之外。当组内角色和受话者不断变化时，稳定地识别主要说话者并不可靠，因此按个人划分的人口统计信息在方法论上并不稳健。因此，我们建模的是可观察且有理论依据的上下文调节变量——组、人群、询问方向、动机——而非个人的人口统计信息。

编码过程主要由四人负责，他们提前审阅并确定了编码标准。由于组/人群/询问方向是情境和交互的客观指标，不同编码员之间的编码标准差异很小。然而，由于动机属于用户的内部信息，编码标准更可能因编码员而异。因此，对于动机，约有 10% 的数据由多名编码员进行核查，Cohen's $\kappa$ 系数为 0.53。

对于二元结果奖励，针对每种奖励条件，使用所有这些经交互编码的自变量拟合广义线性模型（广义线性模型）：

结果 ~臂 + 人群 + 组 + 询问_方向 +动机+ 臂 × 人群 + 臂 × 组+ 臂 × 询问_方向 + 臂 × 动机。

基于这些分析，我们讨论交互情境和用户行为可能如何影响通过汤普森采样获得的语音策略的学习结果。广义线性模型的结果对于未来开发真实环境在线学习框架（例如多臂老虎机）具有重要意义。

## 3 结果

### 3.1 多臂老虎机性能

臂缩写（语速 ×详略程度）：SC =慢速-简洁，SD =慢速-详细，NC = 正常-简洁，ND =正常-详细，FC = 快速-简洁，FD = 快速-详细。

概述。表1报告了各奖励条件下的描述性变量（计数、成功率和选择率）。在每种条件下，每个臂都进行了五次冷启动，但 $R_t$ 条件除外；由于数据缺失，该条件下并非所有臂都进行了五次冷启动。正如表1中的全部数据值所示，成功率高的臂选择率也高，而成功率低的臂选择率也低，这体现了汤普森采样的探索-利用特性。因此，尽管 $R_t$ 条件存在冷启动数据缺失，但这不太可能对总体数据产生显著影响。

图3展示了后验贝塔概率，总结了基于一半数据和全部数据学习到的臂后验概率。展示一半数据有助于我们了解学习过程。分布的期望值越大，表示在每种奖励条件下的成功率越高。选择次数越多，

**表1：六种音频策略组各奖励的描述性统计（$R_u$，$R_c$，$R_t$）：[SC = 慢速简洁、SD = 慢速详细、NC = 正常–简洁、ND = 正常–详细、FC = 快速简洁、FD = 快速详细]。对于每种奖励定义，分别列出冷启动样本和全面部署样本的数据：试验次数 n / 成功次数 / 成功率，以及汤普森采样（TS）的选择比例（选择率）。**

**(a) $R_u$ (用户评分)**

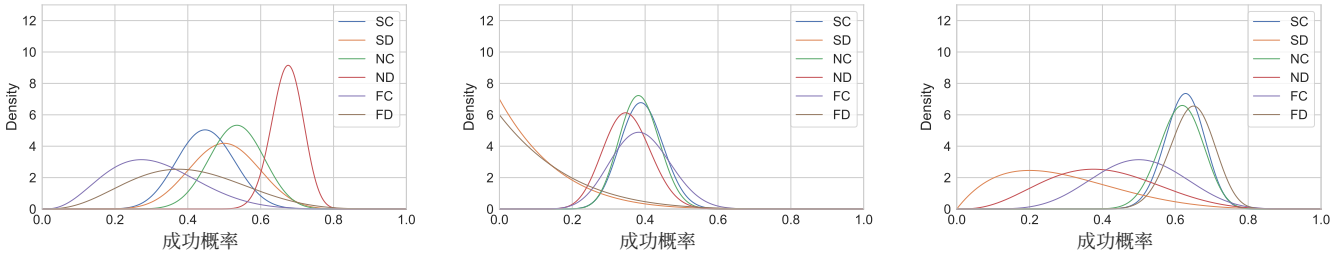
| | 冷启动 | | 全部数据 | | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | *n* | 成功 | 成功率 % | 选择率 % |
| SC | 5 | 3 | 44 | 18 | 40.9 | 9.2 |
| SD | 5 | 4 | 39 | 19 | 48.7 | 8.1 |
| NC | 5 | 3 | 98 | 62 | 63.3 | 20.4 |
| ND | 5 | 4 | 240 | 146 | 60.8 | 50.0 |
| FC | 5 | 3 | 16 | 6 | 37.5 | 3.3 |
| FD | 5 | 1 | 43 | 24 | 55.8 | 9.0 |

**(b) $R_c$ (对话结束)**

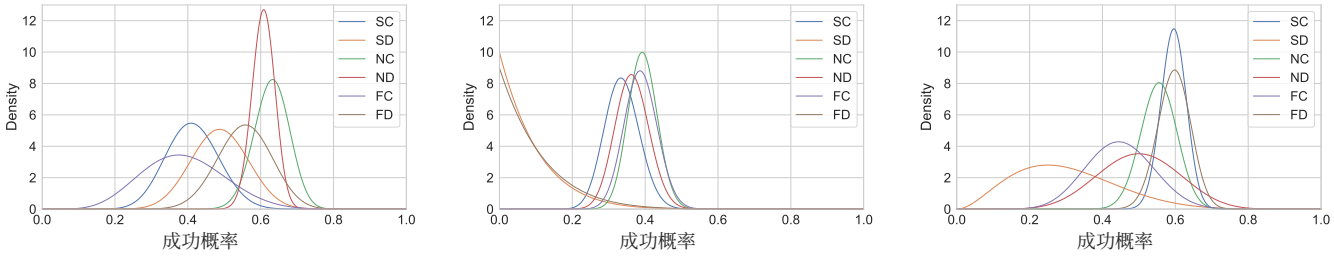
| | 冷启动 | | 全部数据 | | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | *n* | 成功 | 成功率 % | 选择率 % |
| SC | 5 | 3 | 96 | 32 | 33.3 | 20.0 |
| SD | 5 | 0 | 9 | 0 | 0.0 | 1.9 |
| NC | 5 | 4 | 148 | 58 | 39.2 | 30.8 |
| ND | 5 | 3 | 105 | 38 | 36.2 | 21.9 |
| FC | 5 | 3 | 114 | 44 | 38.6 | 23.8 |
| FD | 5 | 0 | 8 | 0 | 0.0 | 1.7 |

**(c) $R_t$ (≥2 轮次)**

| | 冷启动 | | 全部数据 | | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | *n* | 成功 | 成功率 % | 选择率 % |
| SC | 4 | 4 | 198 | 118 | 59.6 | 42.4 |
| SD | 3 | 0 | 8 | 2 | 25.0 | 1.7 |
| NC | 3 | 3 | 99 | 55 | 55.6 | 21.2 |
| ND | 3 | 1 | 18 | 9 | 50.0 | 3.9 |
| FC | 3 | 2 | 27 | 12 | 44.4 | 5.8 |
| FD | 3 | 3 | 117 | 70 | 59.8 | 25.1 |




**(a) $R_u$ （用户评分），半数据结果** **(b) $R_c$ （对话结束），半数据结果** **(c) $R_t$ （≥2 轮次），半数据结果**




**(d) $R_u$ （用户评分），全数据结果** **(e) $R_c$ （对话结束），全数据结果** **(f) $R_t$ （≥2 轮次），全数据结果**

**图3：后验贝塔密度 SD = 慢速详细 各奖励条件下各臂成功概率的后验贝塔密度，分别使用一半和全部训练数据。[SC = 慢速简洁，SD = 慢速详细，NC = 正常–简洁，ND = 正常–详细，FC = 快速简洁，FD = 快速详细]**

分布的方差越小，这与表1中的结果一致。

*$R_u$ (用户评分).* 共分析了480条数据。选择率最高的臂是ND（50.0%），而成功率最高的臂是NC（63.3%）。这可以从图3所示的学习过程中看出。在前半部分数据中，ND的成功率高于其他臂，因此被选中的概率较高。然而，在学习全部数据后，NC的成功率最高。因此，在数据的后半部分，NC开始更频繁地被选中，从而取得最高成功率。此外，在NC和ND之后，其他臂中选择率和成功率最高的依次是FD、SD、SC和FC。与其他奖励条件相比，所有臂的方差都较小，这表明在所有奖励条件下，各臂的选择分布都最为均匀。

*$R_c$ (对话结束)。* 共分析了 480 条数据。选择率和成功率最高的臂是 NC（选择率：30.8%，成功率：39.2%）。虽然 NC 表现最佳，但 SC、ND 和 FC 的选择率与成功率相近，这表明从学习早期阶段开始，学习效果就大致与 NC 相当。此外，SD 和 FD 在冷启动时的成功率为 0.0%；此后，尽管它们偶尔会被探索性地选择，但在整个数据中，其成功率始终为 0.0%。

*$R_t$ （≥ 2 轮次）。* 共分析了 467 个数据集。SC 的选择率最高（42.4%），FD 的成功率最高（59.8%）。这可能是由于各臂的成功率在学习过程中发生了变化，与 $R_u$ 条件类似。继这两个臂之后，NC 臂的选择率和成功率最高，其余各臂的选择率和成功率则较低。

## 3.2 用于调节的广义线性模型

针对奖励条件下的每种二元结果，我们拟合了包含预测变量和臂与上下文交互作用的二项逻辑回归广义线性模型。我们报告了各臂在 $R_u$/$R_c$/$R_t$ 方面的部分统计摘要，如表 2 所示；完整的广义线性模型输出见附录。

编码和基线。 臂采用处理编码，以 SC（慢速简洁）为基线。crowd_T= 1 表示有人群（基线：无人群）；grp_single_T= 1表示单个用户（基线：组）；

**表 2：不同奖励定义下的部分广义线性模型效应（仅列显著预测变量）。基线：臂 =SC，人群 = no，组 = group，询问方向 = no，动机 =功能。系数为广义线性模型估计值（对数优势）。**

| | $R_u$（用户评分） | | | | |
|---|---|---|---|---|---|
| 预测变量 | $\beta$ | SE | $z$ | $p$ | OR |
| grp_single_T | -1.54 | 0.77 | -1.99 | 0.05* | 0.21 |
| ND:grp_single_T | 1.95 | 0.90 | 2.16 | 0.03* | 7.03 |
| NC:motivation_education | 2.76 | 1.41 | 1.96 | 0.05* | 15.76 |
| ND:教育动机 | 2.65 | 1.25 | 2.12 | 0.03* | 14.18 |
| FD:教育动机 | 2.91 | 1.45 | 2.00 | 0.05* | 18.30 |
| | $R_c$ (对话结束) | | | | |
| 预测变量 | $\beta$ | SE | $z$ | $p$ | OR |
| NC:grp_single_T | 2.68 | 1.31 | 2.04 | 0.04* | 14.59 |
| | $R_t$ (≥ 2轮次) | | | | |
| 预测变量 | $\beta$ | SE | $z$ | $p$ | OR |
| ND：人群_T | -1.73 | 0.62 | -2.80 | 0.01* | 0.18 |
| FC：人群_T | -1.39 | 0.52 | -2.69 | 0.01* | 0.25 |
| 标准差:教育动机 | 2.33 | 1.16 | 2.01 | 0.04* | 10.26 |
| FC:教育动机 | 1.89 | 0.93 | 2.04 | 0.04* | 6.63 |

* 表示显著差异。

ask_directions_T= 1 表示询问了方向（基线：否）；动机的基线是功能。

$R_u$ **(用户评分)。** 整体拟合：空模型偏差为 = 1184.7（自由度 = 869），降至残差偏差 = 1103.2（自由度 = 828）；赤池信息准则为 = 1187.2。部分策略臂系数的标准误较大，表明在罕见的策略臂—情境分层中存在部分分离；我们着重关注稳定的情境对比和交互效应。

主效应。 单用户交互的评分低于群体交互（$p$<0.05，优势比=0.21）。人群、询问方向或动机均无总体主效应。

交互效应。 在单用户情境中，相较于SC基线，ND更受青睐（$p$<0.05, 优势比=7.03）。对于教育动机访客，NC、ND和FD均会提高获得高评分的几率（$p$<0.05；优势比

=15 76,. . 和 . ）。

14 18, 18 30

$R_c$ **(对话结束)。** 整体拟合：空模型偏差为 = 592.5（自由度 = 449），降至残差偏差 = 524.4（自由度 = 415）；赤池信息准则为 = 594.4。若干交互项为奇异值（NA），表明策略臂—情境单元数据稀疏。

主效应。 没有主效应达到显著差异。

交互效应。 单个显著交互效应NC×grp_single_T（$p$<0.05，优势比=14.59）表明，在一对一交互中，相较于 SC，Normal–Concise 提高了结束几率。没有其他策略臂—情境交互达到显著水平。

$R_t$ **（≥2 轮次）。** 整体拟合：空模型偏差 = 1218.7（自由度 = 891）降至残差偏差 = 1156.3（自由度 = 852）；赤池信息准则 = 1236.3。两个交互项为奇异项（NA）。

主效应。 没有主效应协变量达到显著水平。

交互效应 在人群场景中（相较于没有人群的 SC），ND 和 FC 处于相对劣势：ND×人群 ($p$<.01, 优势比=0.18) 和 FC×人群 ($p$<.01, 优势比=0.25)。对于以教育为动机的用户（相较于以功能为动机的用户），SD 和 FC 的表现优于 SC：SD=教育 ($p$<.05, 优势比=10.26) 和 FC×教育 ($p$<.05, 优势比=6.63)。

# 4 讨论

## 4.1 结果总结

基于为期12天、涉及超过1,400次交互的真实环境部署，我们得出了三项关键结论。

首先，各臂之间的数据分布并不均匀。随着后验分布逐渐集中，策略会偏向更有前景的臂，而较少采样表现占主导地位的臂；此外，观测成功率较高的臂会被更频繁地选中，这与汤普森采样的预期行为一致。

其次，臂的偏好会受到成功定义方式的影响；根据成功被定义为感知交互满意度（$R_u$）、对话结束（$R_c$）还是对话持续性（$R_t$），在线学习会收敛到不同的语音策略。结果表明，在采样充分的臂中，对于 $R_u$，最佳臂为NC（正常语速–简洁）和ND（正常语速–详细）；对于 $R_c$，NC表现最佳；对于 $R_t$，最佳臂为SC（慢速–简洁）和FD（快速–详细）。

第三，社会情境可能会影响结果，并与语音策略产生交互作用。

- 在 $R_u$条件下，单用户交互的评分低于群体交互（主效应）。ND尤其适合单个用户。在教育动机驱动的访问中，NC/ND/FD的评分往往更高。
- 在 $R_c$下，按对话结束计算的奖励总体上不受情境影响，但单用户交互中特别偏好正常语速–简洁表达。• 在 $R_t$下，拥挤场景中的 ND 和 FC 表现不佳。相比之下，以教育为目的的访问更偏好 SD 和 FC，而非其他设置。

综合来看，(i) 多臂老虎机的表现符合预期（成功率较高的臂获得了更多抽取次数）；(ii) 对“成功”的定义决定了哪种策略会成为最优策略；(iii) 社会情境会调节这些影响。在实践中，运营者应选择与运营目标（满意度、对话结束或对话持续性）相符的奖励定义，并部署情境感知策略（例如，对于满意度，单个用户采用 ND，并在教育动机案例中考虑 NC/ND/FD；对于单个用户，若要最大限度地促成对话结束，则采用正常–简洁表达；若要延长对话，则在人群中避免使用 ND/FC，并在教育动机访客中采用 SD/FC）。

## 4.2 语音策略与奖励设计 n

我们的结果表明，奖励信号的定义会从根本上改变学习行为，从而产生不同的最优策略。这说明不存在单一的“最佳”语音策略，而是存在针对特定运营目标而最优的策略。观察到的策略差异与相关报告一致：显式用户反馈和隐式用户反馈的组合比例不同，可能系统性地改变学习行为 [3, 32]。我们的贡献在于：在公共场所部署中展示了这一效应，同时直接比较了同一机器人和动作集的多种奖励定义。

对于 $R_u$（用户评分）奖励，多臂老虎机收敛到了正常语速–简洁（NC）和正常语速–详细（ND）策略。这表明

在这一公共场景中，要使用户满意度最大化，交互风格应当规范且可预测。正常语速让人熟悉，也易于理解；既不会像语速过慢那样可能令人沮丧，也不会像语速过快那样增加认知负担。简洁与详细两种风格各占一半，可能反映了用户对信息密度的不同偏好，但两者都处于舒适、不过分的范围内。广义线性模型进一步从两个方面细化了这一图景。首先，与群体交互相比，单用户交互的总体评分较低，但正常语速–详细（ND）对单个用户尤其有效，这与一种观点相符：在没有其他观众等待时，人们会欣赏更周全、看似更个性化的回答。其次，以教育为目的的访问会提升用户对正常语速–简洁（NC）、正常语速–详细（ND）和快速语速–详细（FD）的正面评价。这表明，已有信息获取动机的用户，无论是面对清晰简短的解释（NC）、更完整的分步回答（ND），还是快速且信息密集的讲述（FD），都乐于处理其中的内容。

以顺利结束交互为前提的 $R_c$（对话结束）奖励，明显偏向正常-简洁（NC）策略。这一奖励体现了任务效率。NC 臂直接提供信息，不添加多余细节，使用户能够达成目标并顺利结束交互。在这一奖励条件下，慢速-详细（SD）和快速-详细（FD）臂的成功率为 0.0%，这很能说明问题；详细的回答可能会不必要地延长交互，或引出妨碍简单结束对话的话题，从而未能满足奖励条件。广义线性模型进一步揭示了这一点：NC 在一对一交互中尤为有效。

最后，作为参与度替代指标的 $R_t$ （≥ 2 轮次）奖励，对慢速-简洁（SC）和快速-详细（FD）呈现出更复杂的双峰偏好。这表明，促成长时间交互存在两种截然不同的机制。SC 策略节奏从容，可能会让机器人显得更谨慎或更容易接近，或许会促使用户提出澄清问题，或让他们觉得不那么匆忙，从而延长对话。相反，FD 策略可能通过另一种途径提升参与度：新颖性和信息密度。快速而详细的回答可能更有趣，并提供更多可供用户接续的话题线索，从而激发好奇心并引出后续问题 [23]。广义线性模型的结果也佐证了这一点：在拥挤场景中，ND 和 FC 相对处于劣势。

## 4.3 情境感知在线优化的设计启示

在本节中，我们认为，奖励设计——而非单纯的参数调优——主要决定了最终形成何种语音策略。这一观点不仅适用于语速和详略程度，也适用于更广泛的行动空间（如注视、手势和显示提示），由此推动优化方式从平均情形优化转向情境敏感控制。

我们的研究发现，尤其是社会情境对臂表现的显著调节作用，表明人机交互中的有效在线优化必须超越简单的多臂老虎机框架。虽然多臂老虎机学到了合理的平均策略，但更大的机遇在于根据每次交互不断变化的情境动态调整策略。我们并非要规定固定规则，而是提出了下一代情境感知学习系统需要考虑的关键因素。

首先，情境特征的丰富程度至关重要。我们的研究结果指出，人群密度、群体规模和用户意图是关键变量。然而，这些仅仅是起点。未来的系统可以纳入范围广泛得多的情境信号，例如用户的情绪表达 [30]，甚至是与该用户过往互动的记忆 [16]。挑战与机遇在于开发稳健的实时感知能力，以捕捉这些细微特征，并以对学习算法有意义的方式表示它们。这使设计重点从单纯选择行动，转向深入理解行动发生的情境。

其次，我们应探索更复杂的策略学习模型。虽然上下文老虎机是顺理成章的下一步，但公共场所人机交互具有动态且往往难以预测的特性，可能需要更先进的方法。例如，能够处理非平稳性的模型至关重要——所谓非平稳性，是指随着商场人流模式变化，一天之中最优策略也可能随之改变。这对于长期部署尤为重要 [19]。此外，随着机器人能力不断增强，其行动将开始影响交互的后续状态，而传统的多臂老虎机模型无法描述这一情况。这表明，未来的发展方向是迈向完整的强化学习（RL），使机器人学习的不仅是即时行动与奖励之间的关联，还包括长期的交互策略 [17]。

第三，奖励信号本身的定义与传递也提供了一个可供探索的设计空间。本研究比较了三种不同的即时奖励。然而，在长期交互中，成功或许更适合通过随时间变化的指标来定义，例如重复参与，或在多次接触中成功完成任务。这需要能够处理延迟或稀疏奖励的框架 [1, 19]。此外，更直接地将人类反馈纳入奖励函数，例如通过学习偏好或纠正意见，或许能让机器人行为更符合用户需求且更具个性化 [9]。

这项工作是一个起点，表明情境不仅是一个调节因素，更是真正自适应、智能的社交互动赖以构建的基础。目标不仅是找到单一的最佳策略，还要创建一个能够灵活地在适当时机生成恰当策略的系统。

## 4.4 局限性

本研究在一家商业购物中心开展，使用一个机器人平台和六种预先定义的语音策略（语速×详略程度）。如果采用其他因素，例如声音类型、声学环境或任务设置，研究结果可能会有所不同。为简化研究并提高样本效率，我们还选择了二元奖励。尽管这种做法具有可解释性，但可能会压缩细微差异。未来的研究应探索替代阈值，并与连续结果或综合结果进行比较（例如，将对话持续性与满意度相结合）。此外，由于某些臂—上下文组合很少出现，一些广义线性模型系数无法估计。如果可行，更长时间的部署可能会减少这种稀疏性。最后，我们的证据主要基于描述性汇总和广义线性模型。若要提出更有力的结论，可以采用离策略评估，并与已建立的基线进行比较。

# 参考文献

[1] Marcin Andrychowicz、Filip Wolski、Alex Ray、Jonas Schneider、Rachel Fong、Peter Welinder、Bob McGrew、Josh Tobin、Pieter Abbeel 和 Wojciech Zaremba。2018。《事后经验回放》。arXiv:1707.01495 [cs.LG] https://arxiv.org/ abs/1707.01495[2] Rohan Banerjee、Rajat Kumar Jenamani、Sidharth Vasudev、Amal Nanavati、Katherine Dimitropoulou、Sarah Dean 和 Tapomayukh Bhattacharjee。2025。询问还是不询问：人在回路中的上下文老虎机及其在机器人辅助喂食中的应用。载于 *2025 IEEE*机器人与自动化国际会议（*ICRA*）。IEEE，1378–1384。[3] Kim Baraka 和 Manuela Veloso。2015。持续运行的机器人针对用户时间偏好的自适应交互。载于 社会机器人学，Adriana Tapus、Elisabeth An- dré、Jean-Claude Martin、François Ferland 和 Mehdi Ammi（编）。施普林格国际出版公司，Cham，61–71。[4] William Cai、Josh Grossman、Zhiyuan Jerry Lin、Hao Sheng、Johnny Tian-Zheng Wei、Joseph Jay Williams 和 Sharad Goel。2021。用于个性化教育聊天机器人的老虎机算法。机器学习 110，9（2021），2389–2418。[5] Lawrence Chan、Dylan Hadfield-Menell、Siddhartha Srinivasa 和 Anca Dragan。2019。辅助型多臂赌博机。载于 *2019*年第*14*届*ACM/IEEE*人机交互国际会议（*HRI*）。IEEE，354–363。[6] Olivier Chapelle 和 Lihong Li。2011。汤普森采样的实证评估。神经信息处理进展 24（2011）。[7] Yuhang Che、Allison M. Okamura 和 Dorsa Sadigh。2020。通过显式和隐式机器人与人类通信实现高效且可信的社会导航。*IEEE*机器人学汇刊 36，3（2020年6月），692–707。doi:10.1109/tro.2020. 2964824[8] Yifang Chen、Alex Cuellar、Haipeng Luo、Jignesh Modi、Heramb Nemlekar 和 Stefanos Nikolaidis。2020。公平上下文多臂老虎机：理论与实验。载于 人工智能不确定性会议。PMLR，181– 190。[9] Paul Christiano、Jan Leike、Tom B. Brown、Miljan Martic、Shane Legg 和 Dario Amodei。2023。基于人类偏好的深度强化学习。arXiv: 1706 03741 [ ] 1706 03741 。stat.ML [10] https://arxiv.org/abs/ 。Marlena R. Fraune、Selma Šabanović 和 Takayuki Kanda。2019。人类群体的存在、群体特征和群体规范会影响自然情境中的人机交互。《机器人与人工智能前沿》第6卷—2019（2019）。doi:10.3389/frobt.2019.00048[11] Yuan Gao、Wolmet Barendregt、Mohammad Obaid 和 Ginevra Castellano。2018。当机器人个性化没有帮助时：来自机器人支持学习研究的启示。载于 *2018*年第*27*届电气电子工程师学会机器人与人类交互通信国际研讨会（*RO-MAN*）。IEEE，705–712。[12] Koji Inoue、Yuki Okafuji、Jun Baba、Yoshiki Ohira、Katsuya Hyodo 和 Tatsuya Kawahara。2025。面向现实场景对话机器人的抗噪轮流发言系统：一项实地实验。arXiv:2503.06241 [cs.RO] https://arxiv.org/abs/2503. 06241[13] Emilie Kaufmann、Nathaniel Korda 和 Rémi Munos。2012。汤普森采样：渐近最优的有限时间分析。载于 算法学习理论国际会议。施普林格，199–213。[14] Alex Kendall、Matthew Grimes 和 Roberto Cipolla。2015。PoseNet：用于实时六自由度相机重定位的卷积网络。载于 *IEEE*计算机视觉国际会议论文集。2938–2946。[15] Chanwoo Kim、Joonhyeok Lee、Eunwoo Kim 和 Kyungjae Lee。2024。用于机器人行为个性化的时变偏好老虎机。《应用科学》-巴塞尔 14，23（2024）。[16] Jens Kober、J. Andrew Bagnell 和 Jan Peters。2013。机器人学中的强化学习：综述。《国际机器人研究杂志》32，11（2013），1238–1274。doi:10.1177/0278364913495721[17] Jens Kober 和 Jan Peters。2014。机器人学中的强化学习：综述。施普林格国际出版公司，Cham， 9 67 10 1007 978 3 319 03194 1_2 – 。 [18] doi: 。/ - - - - Amy Koike、Yuki Okafuji、Kenya Hoshimure 和 Jun Baba。 。是什么驱使你进行交互？：用户动机在真实环境中的机器人交互中的作用。载于 *2025*年第*20*届*ACM/IEEE*人机交互国际会议（*HRI*）。IEEE，183–192。 2025

[19] Tor Lattimore and Csaba Szepesvári. 2020. 老虎机算法. Cambridge University Press. doi:10.1017/9781108571401[20] Guiying Liu, Tatjana Korbanka, Jasmin Timm, Naomi Veit, Laura Maier, Markus Huff, and Frank Papenmeier. 2025. 人机交互中的旁观者效应：旁观者的存在如何抑制对社交机器人的帮助。(Sep. 2025 10 23668 21242 ). doi: . / psycharchives. [21] Marcos Maroto-Gómez, María Malfaz, José Carlos Castillo, Álvaro Castro- González, and Miguel Ángel Salichs. 2024. 根据明确和隐式用户反馈，为辅助型社交机器人个性化选择活动。《国际社交机器人学杂志》(2024), 1–19. [22] David Issa Mattos, Jan Bosch, and Helena Holmström Olsson. 2019. 真实环境中的多臂老虎机：在线实验中的陷阱与策略。《信息与软件技术》113 (2019), 68–81. [23] Zi Haur Pang, Yahui Fu, Divesh Lala, Mikey Elmers, Koji Inoue, and Tatsuya Kawahara. 2025. 自主对话机器人的外观是否会影响现实会议交互中的用户言语行为？载于计算机系统中的人为因素会议（*CHI EA '25*）扩展摘要集。美国计算机协会，文章 198，8 页。doi:10.1145/3706599.3720179[24] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. 2016. 机器人通过多模态深度强化学习获得社会智能。载于 *2016*年*IEEE-RAS*第*16*届国际人形机器人会议（*Humanoids*）。电气电子工程师学会，745–751。[25] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Ilhan Aslan, and Elisabeth André. 2018. Drink-o-mender：自适应机器人饮品顾问。载于第三届人类与食物交互多感官方法国际研讨会论文集。1–8。[26] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. 基于明确的人类反馈，为辅助型机器人健康伴侣自适应调整语言风格。载于第*12*届*ACM*辅助环境相关普适技术国际会议论文集。247–255。[27] Hannes Ritschel, Andreas Seiderer, Kathrin Janowski, Stefan Wagner, and Elisabeth André. 2019. 基于明确的人类反馈，为辅助型机器人健康伴侣自适应调整语言风格。载于第*12*届*ACM*辅助环境相关普适技术国际会议论文集（希腊罗得岛）（*PETRA '19*）。美国计算机协会，纽约，纽约州，美国，247–255。doi:10.1145/3316782.3316791[28] Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. 2018. 汤普森采样教程。《基础与趋势》® 机器学习11, 1 (2018), 1–96。[29] Sebastian Schneider and Franz Kummert. 2017. 探索具身性与对战式老虎机学习在人机交互偏好适应中的应用。载于*2017*年第*26*届电气电子工程师学会机器人与人类交互通信国际研讨会（*RO-MAN*）。电气电子工程师学会，1325–1331。[30] Matteo Spezialetti, Giuseppe Placidi, and Silvia Rossi. 2020. 人机交互中的情绪识别：最新进展与未来展望。《机器人与人工智能前沿》第 7 2020 10 3389 2020 532279 卷（）。doi: . / frobt. [31] 。William R Thompson. 。基于两组样本的证据，一个未知概率大于另一个未知概率的似然。《生物计量学》《生物计量学》25, 3/4 (1933), 285–294。[32] Konstantinos Tsiakas, Maher Abujelala, and Fillia Makedon. 2018. 任务投入作为社会辅助机器人和认知训练的个性化反馈。《技术》 1933 6, 2 2018 2227 7080 6 2 49 （）。 https://www.mdpi.com/ - / / / Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühn- lenz, Dirk Wollherr, and Martin Buss. 2010. 问路机器人：路人支持机器人的意愿。载于第五届*ACM/IEEE*人机交互国际会议论文集（日本大阪）（*HRI '10*）。IEEE出版社，23–30。[34] David Whitney, Eric Rosen, James MacGlashan, Lawson L. S. Wong, and Stefanie Tellex. 2017. 通过社会反馈减少物体取送交互中的错误。载于*2017*年*IEEE*机器人与自动化国际会议（*ICRA*）。1006–1013。doi:10.1109/ICRA.2017.7989121 [33]

# A 补充分析

本附录提供：（1）对比汤普森采样和均匀采样在所有音频策略组中的描述性统计数据和奖励（$R_u$、$R_c$、$R_t$）（表3）；（2）按奖励类别展示两种分配方案的累积奖励图（图4）；以及（3）完整的广义线性模型结果，包括不显著项和诊断结果（表4）。

**表3：按奖励（$R_u$、$R_c$、$R_t$）分层，对比汤普森采样与均匀采样的描述性统计，涵盖六个音频策略组（SC = 慢速简洁、SD = 慢速详细、NC = 正常–简洁、ND = 正常–详细、FC = 快速简洁、FD = 快速详细）。（a）$R_u$（用户评分）**

| | 汤普森采样 | | | 均匀采样 | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | 成功率 % | *n* | 成功 | 成功率 % |
| SC | 44 | 18 | 40.9 | 74 | 31 | 41.9 |
| SD | 39 | 19 | 48.7 | 72 | 29 | 40.3 |
| NC | 98 | 62 | 63.3 | 78 | 49 | 62.8 |
| ND | 240 | 146 | 60.8 | 76 | 43 | 56.6 |
| FC | 16 | 6 | 37.5 | 76 | 56 | 73.7 |
| FD | 43 | 24 | 55.8 | 74 | 53 | 71.6 |

**(b) $R_c$ (对话结束)**

| | 汤普森采样 | | | 均匀采样 | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | 成功率 % | *n* | 成功 | 成功率 % |
| SC | 96 | 32 | 33.3 | 74 | 20 | 27.0 |
| SD | 9 | 0 | 0.0 | 72 | 21 | 29.2 |
| NC | 148 | 58 | 39.2 | 78 | 19 | 24.4 |
| ND | 105 | 38 | 36.2 | 76 | 33 | 43.4 |
| FC | 114 | 44 | 38.6 | 76 | 38 | 50.0 |
| FD | 8 | 0 | 0.0 | 74 | 35 | 47.3 |

**(c) $R_t$ (≥2 轮次)**

| | 汤普森采样 | | | 均匀采样 | | |
|---|---|---|---|---|---|---|
| Arm | *n* | 成功 | 成功率 % | *n* | 成功 | 成功率 % |
| SC | 198 | 118 | 59.6 | 74 | 30 | 40.5 |
| SD | 8 | 2 | 25.0 | 72 | 37 | 51.4 |
| NC | 99 | 55 | 55.6 | 78 | 37 | 47.4 |
| ND | 18 | 9 | 50.0 | 76 | 49 | 64.5 |
| FC | 27 | 12 | 44.4 | 76 | 51 | 67.1 |
| FD | 117 | 70 | 59.8 | 74 | 56 | 75.7 |

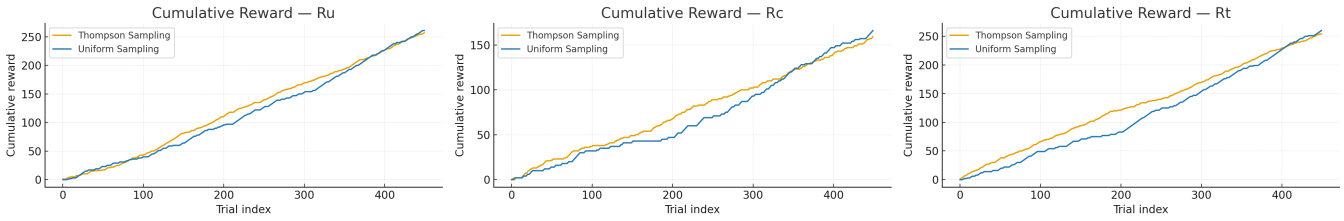



**图4：累积奖励曲线（汤普森采样与均匀采样）。**

**表4：跨奖励的广义线性模型完整结果（修订版）。基线：臂=SC、人群=no、组=group、询问=no、动机=功能。**

| | $R_u$（用户评分） | | | | | $R_c$（对话结束） | | | | | $R_t$（≥ 2轮次） | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 预测变量 | $\beta$ | SE | $z$ | $p$ | OR | $\beta$ | SE | $z$ | $p$ | OR | $\beta$ | SE | $z$ | $p$ | OR |
| （截距） | -14.74 | 1455.40 | -0.01 | 0.99 | 0.00 | -0.72 | 0.96 | -0.76 | 0.45 | 0.48 | -0.52 | 0.73 | -0.72 | 0.47 | 0.59 |
| SD | -16.60 | 2520.82 | -0.01 | 0.99 | 0.00 | -0.99 | 1.14 | -0.87 | 0.38 | 0.37 | -0.71 | 0.65 | -1.09 | 0.27 | 0.49 |
| NC | 16.33 | 1455.40 | 0.01 | 0.99 | — | -15.15 | 1741.84 | -0.01 | 0.99 | 0.00 | 0.52 | 1.20 | 0.43 | 0.66 | 1.68 |
| ND | 14.49 | 1455.40 | 0.01 | 0.99 | — | 18.75 | 2797.44 | 0.01 | 0.99 | — | 1.92 | 1.53 | 1.26 | 0.21 | 6.85 |
| FC | 14.32 | 1455.40 | 0.01 | 0.99 | — | -17.62 | 1940.86 | -0.01 | 0.99 | 0.00 | 0.72 | 1.33 | 0.54 | 0.59 | 2.05 |
| FD | 15.96 | 1455.40 | 0.01 | 0.99 | — | 0.58 | 0.97 | 0.59 | 0.55 | 1.78 | 0.48 | 0.94 | 0.51 | 0.61 | 1.61 |
| crowd_T | 0.57 | 0.61 | 0.92 | 0.36 | 1.76 | -0.28 | 0.90 | -0.31 | 0.76 | 0.76 | 0.35 | 0.26 | 1.33 | 0.18 | 1.42 |
| grp_single_T | **-1.54** | **0.77** | **-1.99** | **0.05*** | **0.21** | -0.62 | 0.91 | -0.68 | 0.50 | 0.54 | -0.06 | 0.39 | -0.16 | 0.87 | 0.94 |
| ask_directions_T | 14.87 | 1455.40 | 0.01 | 0.99 | — | -0.47 | 0.72 | -0.65 | 0.52 | 0.63 | 0.59 | 0.69 | 0.85 | 0.40 | 1.80 |
| 实验 | -0.40 | 0.51 | -0.78 | 0.44 | 0.67 | 0.55 | 0.69 | 0.79 | 0.43 | 1.73 | -0.17 | 0.29 | -0.61 | 0.54 | 0.84 |
| 好奇心 | -0.83 | 1.31 | -0.63 | 0.53 | 0.44 | -0.75 | 1.34 | -0.56 | 0.58 | 0.47 | 0.41 | 0.71 | 0.58 | 0.56 | 1.51 |
| 教育 | -2.11 | 1.19 | -1.77 | 0.08 | 0.12 | -16.37 | 1978.09 | -0.01 | 0.99 | 0.00 | -0.80 | 0.48 | -1.67 | 0.10 | 0.45 |
| 标准差：人群_T | 0.05 | 0.82 | 0.07 | 0.95 | 1.05 | 0.19 | 1.27 | 0.15 | 0.88 | 1.20 | -1.40 | 0.80 | -1.75 | 0.08 | 0.25 |
| NC：人群_T | -0.23 | 0.72 | -0.32 | 0.75 | 0.80 | -15.93 | 1233.21 | -0.01 | 0.99 | 0.00 | -0.42 | 0.43 | -1.00 | 0.32 | 0.65 |
| ND：人群_T | -0.79 | 0.67 | -1.18 | 0.24 | 0.46 | -0.46 | 1.25 | -0.37 | 0.72 | 0.63 | **-1.73** | **0.62** | **-2.80** | **0.01*** | **0.18** |
| FC：人群_T | 0.18 | 0.82 | 0.22 | 0.83 | 1.19 | -0.36 | 1.03 | -0.35 | 0.73 | 0.70 | **-1.39** | **0.52** | **-2.69** | **0.01*** | **0.25** |
| FD：人群_T | -0.62 | 0.75 | -0.82 | 0.41 | 0.54 | 0.79 | 1.03 | 0.77 | 0.44 | 2.20 | -0.30 | 0.42 | -0.71 | 0.48 | 0.74 |
| 标准差:grp_single_T | 0.16 | 1.15 | 0.14 | 0.89 | 1.17 | 1.15 | 1.40 | 0.82 | 0.41 | 3.15 | -0.62 | 1.01 | -0.61 | 0.54 | 0.54 |
| NC:grp_single_T | 0.98 | 0.93 | 1.05 | 0.29 | 2.67 | **2.68** | **1.31** | **2.04** | **0.04*** | **14.59** | -0.05 | 0.69 | -0.07 | 0.94 | 0.95 |
| ND:grp_single_T | **1.95** | **0.90** | **2.16** | **0.03*** | **7.03** | 0.51 | 1.33 | 0.39 | 0.70 | 1.67 | 0.15 | 1.02 | 0.15 | 0.88 | 1.16 |
| FC:grp_single_T | 1.02 | 1.10 | 0.93 | 0.36 | 2.77 | 0.68 | 1.20 | 0.57 | 0.57 | 1.98 | -0.04 | 0.77 | -0.05 | 0.96 | 0.96 |
| FD:grp_single_T | 0.71 | 1.70 | 0.42 | 0.68 | 2.04 | — | — | — | — | — | -0.80 | 0.69 | -1.16 | 0.25 | 0.45 |
| 标准差:ask_directions_T | 16.88 | 2520.82 | 0.01 | 1.00 | — | — | — | — | — | — | — | — | — | — | — |
| NC:问路_T | -16.24 | 1455.40 | -0.01 | 0.99 | 0.00 | 14.32 | 1741.84 | 0.01 | 0.99 | — | -0.57 | 1.13 | -0.51 | 0.61 | 0.56 |
| ND:ask_directions_T | -14.47 | 1455.40 | -0.01 | 0.99 | 0.00 | -17.59 | 2797.44 | -0.01 | 1.00 | 0.00 | -0.97 | 1.47 | -0.66 | 0.51 | 0.38 |
| FC:ask_directions_T | -13.83 | 1455.40 | -0.01 | 0.99 | 0.00 | 18.87 | 1940.86 | 0.01 | 0.99 | — | 0.24 | 1.27 | 0.19 | 0.85 | 1.28 |
| FD:ask_directions_T | -15.24 | 1455.40 | -0.01 | 0.99 | 0.00 | — | — | — | — | — | 0.03 | 0.87 | 0.04 | 0.97 | 1.03 |
| 标准差:实验 | -0.59 | 0.74 | -0.80 | 0.43 | 0.55 | 0.93 | 1.18 | 0.78 | 0.43 | 2.53 | 1.31 | 0.70 | 1.87 | 0.06 | 3.70 |
| NC:实验 | 0.71 | 0.63 | 1.12 | 0.26 | 2.03 | 0.29 | 1.09 | 0.26 | 0.79 | 1.33 | 0.16 | 0.47 | 0.33 | 0.74 | 1.17 |
| ND:实验 | 0.65 | 0.58 | 1.12 | 0.26 | 1.92 | -1.01 | 0.94 | -1.07 | 0.29 | 0.36 | -0.15 | 0.65 | -0.22 | 0.82 | 0.86 |
| FC:实验 | 0.83 | 0.79 | 1.05 | 0.29 | 2.29 | -0.15 | 0.89 | -0.17 | 0.86 | 0.86 | -0.28 | 0.55 | -0.51 | 0.61 | 0.75 |
| FD:实验 | 0.09 | 0.78 | 0.12 | 0.90 | 1.10 | -0.36 | 1.02 | -0.36 | 0.72 | 0.70 | 0.41 | 0.45 | 0.92 | 0.36 | 1.51 |
| 标准差:好奇心 | 15.98 | 1455.40 | 0.01 | 0.99 | — | — | — | — | — | — | — | — | — | — | — |
| NC:好奇心 | 1.39 | 1.84 | 0.75 | 0.45 | 4.01 | — | — | — | — | — | -0.08 | 1.16 | -0.07 | 0.94 | 0.92 |
| ND:好奇心 | 2.26 | 1.65 | 1.37 | 0.17 | 9.54 | -17.97 | 2797.44 | -0.01 | 0.99 | 0.00 | -2.51 | 1.95 | -1.28 | 0.20 | 0.08 |
| FC:好奇心 | -1.10 | 1.82 | -0.61 | 0.54 | 0.33 | — | — | — | — | — | 14.16 | 362.66 | 0.04 | 0.97 | — |
| FD:好奇心 | -0.10 | 1.73 | -0.06 | 0.96 | 0.91 | — | — | — | — | — | 1.49 | 1.41 | 1.06 | 0.29 | 4.44 |
| 标准差:教育 | 17.26 | 650.87 | 0.03 | 0.98 | — | 18.15 | 1978.09 | 0.01 | 0.99 | — | **2.33** | **1.16** | **2.01** | **0.04*** | **10.26** |
| NC:教育 | **2.76** | **1.41** | **1.96** | **0.05*** | **15.76** | 35.95 | 3426.15 | 0.01 | 0.99 | — | 1.35 | 0.69 | 1.96 | 0.05 | 3.88 |
| ND:教育 | **2.65** | **1.25** | **2.12** | **0.03*** | **14.18** | 17.50 | 1978.09 | 0.01 | 0.99 | — | 1.21 | 0.99 | 1.22 | 0.22 | 3.35 |
| FC:教育 | 1.42 | 1.41 | 1.01 | 0.31 | 4.14 | 17.49 | 1978.09 | 0.01 | 0.99 | — | **1.89** | **0.93** | **2.04** | **0.04*** | **6.63** |
| FD:教育 | **2.91** | **1.45** | **2.00** | **0.05*** | **18.30** | 17.16 | 1978.09 | 0.01 | 0.99 | — | 1.26 | 0.71 | 1.78 | 0.07 | 3.54 |

* 表示 $p < .05$。当 $|\beta| \gtrsim 8$ 时，由于估计值条件不适，优势比省略（—）。