arXiv:2603.14449v1 [cs.HC] 15 Mar 2026

Tap-to-Adapt：为语音智能体学习与用户对齐的响应时机

Zihong He、Hai-Ning Liang*、Chen Liang*

香港科技大学（广州） zhe154@connect.hkust-gz.edu.cn
{hainingliang, chenliang2}@hkust-gz.edu.cn*通讯作者

2026年3月17日

摘要

响应时机判断是交互式语音智能体的关键组成部分。尽管轮次建模和语音唤醒已有大量研究，但对于如何使响应时机判断持续与用户意图保持一致，相关研究仍然不足。为此，我们提出 Tap-to-Adapt 框架，用户可通过点击交互自然地唤醒或打断智能体，从而为响应时机模型构建在线学习标签。在该框架中，膨胀时间卷积网络和顺序重放策略发挥了重要作用，这一点已通过数据驱动实验和用户研究得到验证。此外，我们针对 Tap-to-Adapt 框架开发了一套评估与持续数据挖掘系统，并通过涉及 20 名参与者的用户研究收集了约 20,000 个样本。

1 引言

随着大语言模型（LLMs）推理能力的涌现，经大语言模型增强的交互式语音智能体展现出日益卓越的性能，并被广泛应用于移动设备 [Deng et al., 2024, Liu et al.,2025b], 可穿戴设备 [Mishra et al., 2025, Ren et al., 2025], 智能家居系统 [Rivkin et al., 2024, Lin et al., 2025], 车载系统 [Furu- sawa and Saitoh, 2024, Furusawa, 2025], 老年护理 [Yang et al., 2024, Liu et al., 2025a], 等领域。确定是否以及何时响应用户，是交互式语音智能体系统的关键组成部分 [Skantze, 2021, Walker, 2024, Liu et al., 2025a, Ethiraj 等人, 2025]。

一类相关工作聚焦于对语音对话场景中话语起始时机进行建模；然而，这些方法仅反映了对话中的客观轮次初始化，未能捕捉人类期望智能体作出响应的时机 [Skantze, 2017年, Ekstedt 和 Skantze, 2022年, Li 等人, 2022年, Wagner 等人, 2025]。另一类工作涉及语音唤醒任务，这类任务大多依赖关键词检测或说话人识别机制，无法在与用户意图一致的多轮交互中进行响应时机判断或停止响应 [Kë- puska 和 Klein, 2009年, Këpuska, 2011年, Këpuska 和 Bohouta, 2017], , 这与人类-语音智能体交互存在显著不匹配。目前，尚缺乏通过在多轮对话上下文中使响应时机与人类意图对齐来驱动人类-智能体交互的工作。

另一方面，屏幕上的用户点击交互，结合手势和面部信息，有助于系统识别特定用户 [Choi 和 Park, 2015年, Wang 等人, 2021年, Khamis 等人, 2022], 这表明通过目标用户点击行为构建个性化标签是可行的。

基于这些观察，我们提出了面向交互式语音智能体的 Tap-to-Adapt 框架，使用户能够通过简单的点击操作手动激活或中断智能体响应，从而构建“正向”（来自激活）和“负向”（来自中断）的在线学习标签。在这里，点击是一种抽象的交互原型，代表明确的用户干预，并可扩展到实际中的其他交互模态。在此框架中，我们发现，将膨胀时间卷积网络

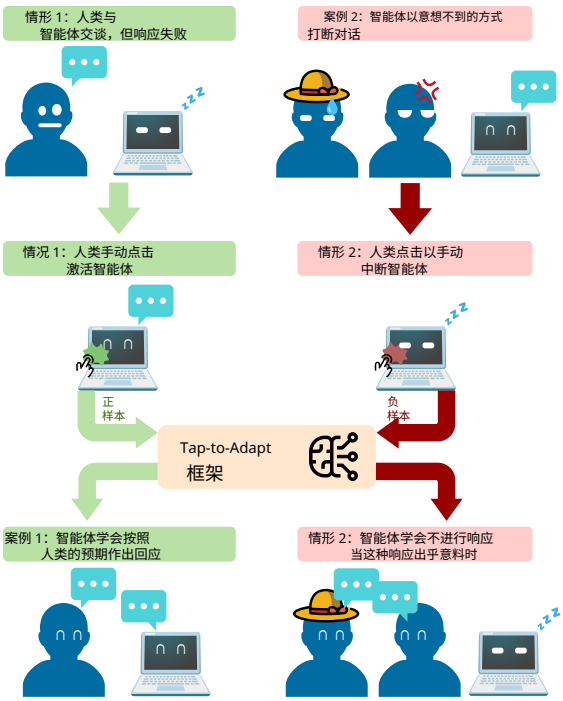


图1：在人类与语音智能体的交互场景中，Tap-to-Adapt 框架通过人类执行的手动激活和打断来构建在线学习样本，驱动智能体学习按照人类的预期做出响应。

编码器 [Lea 等人， 2017]与标签均衡的序列回放在线学习策略 [Rolnick 等人，2019年，Ji 等人，2021年，Liu 等人， 2024] 相结合，可显著提升性能。该模型利用在线学习标签，从语音上下文中学习与用户意图相符的响应时机。

此外，我们提出了一个统一的系统，用于评估 Tap-to-Adapt 框架，并在由受控模拟和用户研究共同驱动的人类与智能体交互场景中构建响应时机判断数据。利用该系统，我们模拟了 1,000 条按时间顺序排列的多场景语音样本，并对 20 名参与者开展用户研究，最终获得了 2,478 条根据交互行为标注的在线学习样本，以及一个从真实交互中收集的大规模时序数据集。我们从后者中提取了 17,416 条多用户语音样本。这些数据有助于弥补人类与智能体交互场景中缺少与用户意图相符的响应时机评估基准这一不足。借助该系统和数据集，我们进一步验证了模型和学习策略的有效性。

总而言之，我们的贡献如下：

1. 我们提出了**Tap-to-Adapt**，这是一种通过自然的用户点击操作构建在线学习监督信号的框架，使语音模型能够在实时语音智能体交互中，主动与用户特定的预期响应时机保持一致。该框架解决了使智能体语音响应时机与用户意图保持一致这一挑战。

2. 在这一框架下，我们发现，将基于扩张 TCN的模型与顺序样本回放在线学习策略相结合，效果显著。该方法无需依赖离线人工构建样本，即可在冷启动、预训练、用户专属权重和多用户共享权重等场景中实现高效适应。

3. 我们构建了一套系统化的系统，用于评估 Tap-to-Adapt 框架并构建响应时机数据，结合了受控模拟和用户研究。借助该系统，我们挖掘了约 20,000 个样本，有助于弥补人机交互场景中与用户意图对齐的响应时机数据集方面的空缺。

源代码可在 https://github.com/ZihongHe/tap_获取_adapt。

2 相关工作

2.1 语音回合启动建模

我们的工作旨在学习一种符合用户偏好的语音回合启动模型，以确定语音智能体适宜的响应时机。语音回合启动建模可视为轮替建模的一个子集，指预测语音活动何时开始 [Skantze，2017年，Ekstedt 和 Skantze，2022年，Li 等人，2022年，Wagner 等人， 2025]。已有若干具有代表性的研究。Skantze 等人 [Skantze， 2017]使用长短期记忆循环神经网络，对每一帧预测说话人在短暂的未来时间窗口内开始讲话的概率，并通过预测有效停顿后 1 秒内的说话人来评估模型。Ekstedt 等人 [Ekstedt 和 Skantze， 2022] 进一步采用基于对比预测编码和 Transformer 的模型，捕捉未来语音活动窗口中的时间依赖关系，并在双人语音对话中通过检测轮次切换来验证模型。为解决基于停顿的触发机制的局限性

在轮次发起建模中，这类方法与自然的人类对话有所不同，因为自然对话具有快速轮流发言和话语重叠的特点。Li 等人 [Li 等人， 2022]使用高斯混合模型（GMM）预测对话发起的连续时机，并在语音对话数据上验证了他们的方法。然而，当前的轮次发起模型只与真实语音对话中说话人话语的实际时机保持一致。这些研究在人类与语音智能体的交互方面存在不足，因为它们无法恰当地建模用户期望智能体作出响应的时机。

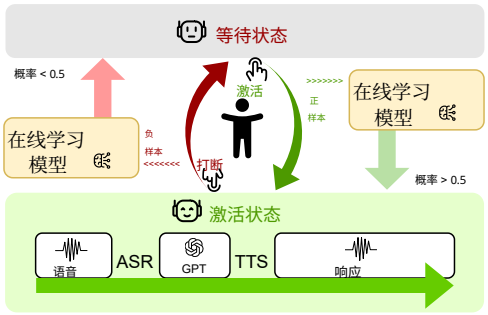


图 2：该图展示了 Tap-to-Adapt 框架的架构。

2.2 语音唤醒系统

语音唤醒是指通过特定语音信号激活系统的任务 [Këpuska and Klein, 2009, Këpuska, 2011, Këpuska and Bohouta,2017]。广义而言，我们的 Tap-to-Adapt 框架属于语音唤醒系统这一类别。这类系统广泛应用于可穿戴设备 [Schmalfuß-Schwarz et al., 2024], 家用语音助手 [Ferro Filho et al., 2024], 老年护理语音助手 [Yan et al., 2024],等领域。早期的语音唤醒系统主要基于隐马尔可夫模型（HMM）、高斯混合模型（GMM）或混合高斯混合模型-隐马尔可夫模型方法。这些方法在数据有限时具有良好的适应性，但对说话人差异性和噪声的鲁棒性较低 [Vroomen and Normandin, 1992, Ghaf- farzadegan et al., 2017, Schröder et al., 2016]。近年来，基于神经网络的解决方案表现出更高的可靠性 [Oord et al., 2016, Shan et al., 2018, Li et al., 2019, Gulati et al., 2020, Jose et al., 2020, Radford et al., 2023]。例如，Jose 等人 [Jose et al., 2020] 使用卷积神经网络（CNN）进行唤醒词检测，并实现精确的起止定位；Sahai 等人 [Sahai et al.,2023] 提出了一种双重注意力唤醒词检测模型，其中一个注意力机制聚焦于唤醒词，另一个则关注更广泛的语音上下文，以提高准确率。然而，大多数语音唤醒系统仍依赖固定唤醒词 [Këpuska and Klein, 2009, Këpuska, 2011, Këpuska and Bohouta, 2017, Jose et al., 2020], ，很少有系统提供与真实用户交互标签对齐的机制 [Mallidi et al., 2018, Norouzian et al., 2019, Dighe et al., 2024],，这限制了它们匹配用户实际意图的能力。

2.3 基于用户反馈的在线语音学习

我们的系统会实时学习用户在交互过程中通过点击屏幕生成的语音数据。在基于用户反馈的在线语音学习领域，已有若干具有代表性的工作。Gavsic 等人 [Gaši´c 等人， 2011] 允许用户在对话系统中提供二元满意度评分，并通过在线学习优化响应策略。Su 等人 [Su 等人， 2015] 挖掘对话过程中自然产生的交互行为，并通过在线强化学习优化模型，从而提升性能；Su 等人 [Su 等人， 2016]则将用户对对话成功与否的主观反馈纳入在线强化学习奖励。Zhou 等人 [Zhou 等人， 2023]提出了一种语音转文本系统，用户可手动纠正识别错误以优化模型；He 等人 [He 等人， 2025]则允许用户在语音识别和说话人分离过程中提供实时语音反馈来纠正错误，例如识别正确的说话人。然而，这些方法要么依赖用户刻意提供的反馈 [Gaši´c 等人，2011年，Su 等人，2016年，Zhou 等人，2023年，He 等人， 2025] ，要么依赖交互过程中产生的隐式信号 [Su 等人， 2015]。目前，针对在线语音学习任务中明确且自然产生的用户反馈信号，相关研究仍然缺乏。

3 方法

3.1 框架概述

遵循 [In- oue 等人，2024年，Castillo-López 等人 的连续预测方法， 2025], Tap-to-Adapt 框架使用最新15秒音频作为输入，以0.2秒为间隔进行连续预测

。具体工作流程如图 2所示。当在线学习模型预测的概率低于0.5时，智能体保持静默；当预测值超过0.5时，则激活智能体，使其进入 激活 状态，并最终完成响应。在 等待 状态下，用户点击会强制使智能体进入激活 状态，并使用该时刻的音频上下文为模型构建一个正在线学习样本。反之，在激活 状态下，用户点击会强制使智能体返回 等待状态，并使用相应的音频上下文为模型构建一个负在线学习样本。

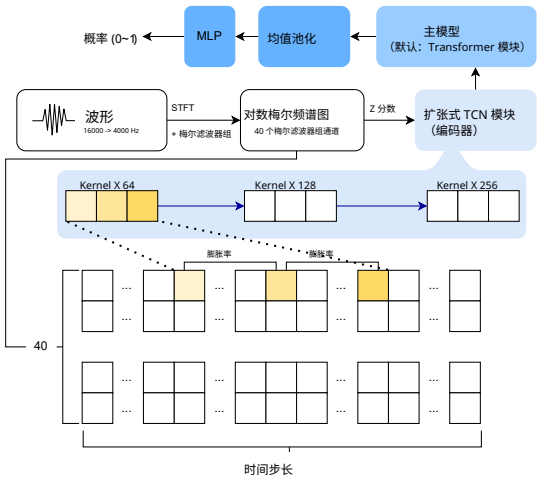


图 3：该图展示了 **Tap-to-Adapt**框架中的**高效的**基于扩张TCN的模型。该模型首先从输入音频中提取梅尔频谱特征，然后将其输入多个采用膨胀机制的时序卷积网络模块，以扩大时间感受野。随后，得到的特征被送入主模型，该模型默认采用 Transformer 模块 [Vaswani 等人， 2017], 由输入投影层、位置编码和 Transformer 编码器组成。

3.2 模型与学习策略

基于扩张TCN的模型 在基于人类与 LLM 智能体语音交互的响应预测中，大尺度和小尺度时间依赖关系都发挥着关键作用。大尺度依赖关系包括对话的先后顺序，例如智能体是否先于人类发言；小尺度依赖关系则涉及话语内容，例如根据人类说话者的最后几个词或语调来判断对话是否应该结束。这表明模型需要捕捉多尺度的时间依赖关系。与标准 TCN 相比，膨胀 TCN 在捕捉语音的长程依赖关系方面效率更高 [Lea 等人， 2017]。结合实验验证，基于扩张TCN的模型作为 Tap-to-Adapt 框架中的在线学习模型表现出色。图3展示了我们为该框架构建的模型。模型输入为从以 16 kHz 采样并降采样至 4 kHz 的原始音频中提取的 Log-Mel 频谱图 [Davis 和 Mermelstein, 1980]。我们进一步使用 Z 分数归一化处理音频输入，以消除振幅变化，使模型能够关注关键特征而非响度 [Fawcett 等人， 2021]。

序列回放在线学习策略对于在线学习方法，我们从先前的工作中汲取灵感 [Rolnick 等人, 2019，Ji 等人，2021年，Liu 等人， 2024] 并采用序列回放在线学习策略，默认配置为标签均衡。具体而言，在每个交互步骤 t，模型首先对传入样本$(x_t,\ y_t)$ 按照严格的先预测后更新协议。评估后，模型会对一个小批量执行一系列序列回放更新，该小批量 $\mathcal{S}_t$ 由存储在回放缓冲区中的若干近期样本组成。为强调近期用户行为，每个样本 $(x_i,\ y_i) \in \mathcal{S}_t$ 都会被赋予一个随样本年龄呈指数衰减的权重：

$$w_i = w_0 \cdot \gamma^{t-i},$$

其中， w_0 是原始样本权重，$\gamma \in (0, 1)$ 是滑动窗口衰减因子。这种加权重放机制确保较新的交互在更新过程中具有更大的影响力，同时较旧的样本仍能发挥作用，从而有效缓解灾难性遗忘。

3.3 样本挖掘与评估

模拟样本构建 Tap-to-Adapt 框架通过预定义的概率驱动场景配置，支持模拟人类与语音智能体的交互场景。该框架支持使用 DataSec 数据集模拟背景声音，该数据集包含 22 个顶级声音类别 [Fredianelli 等人， 2025]。对于每个类别，可以选择时长最长的音频样本，并在预设范围内缩放音量进行播放，以模拟

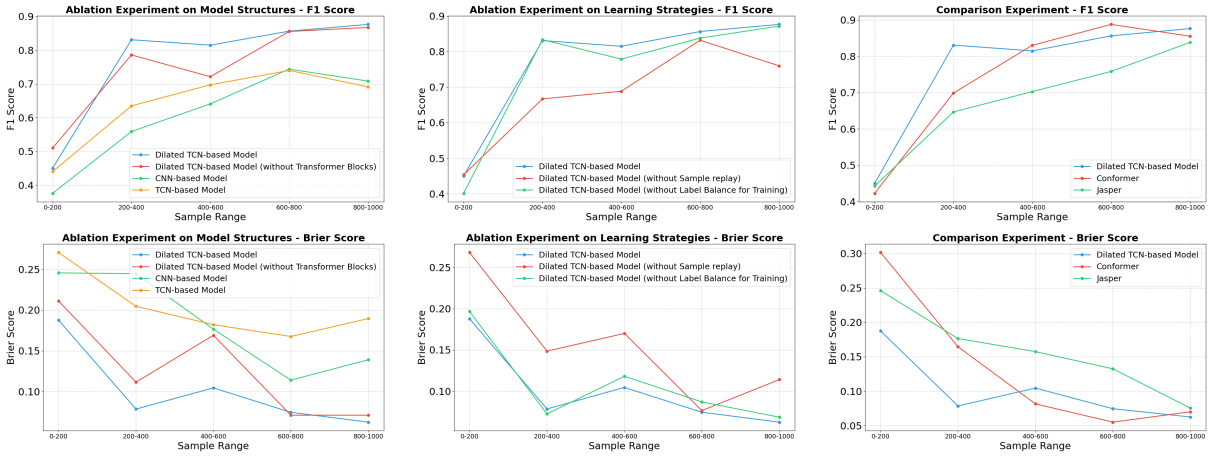


图4：使用**1,000 个模拟样本**对 Tap-to-Adapt 框架进行验证实验的结果。相关模型和在线学习策略详见第3.2节。此处，“不含 Transformer 模块”表示主模型采用简单的多层感知机实现。

真实环境中不同的背景音量。播放结束后，会插入一段时长相同的静音片段，随后随机选择另一个音频样本并重复这一过程。对话主题遵循 [Li 等人所述的配置， 2017]。为保持场景连续性，该框架允许为某些与场景相关的因素分配复现概率权重，使其能够在连续场景中持续存在。这些因素总结于附录 A中。仅当用户正在与智能体交谈，且样本类型为“说话完毕，等待回复”时，才会出现正样本。

真实样本挖掘 Tap-to-Adapt 框架利用用户交互行为和状态信息，在实际使用过程中进行真实样本挖掘。具体而言，它会收集：（1）用户最近一次手动激活 Tap-to-Adapt 模型时对应的音频输入，作为正样本；（2）用户打断时处于激活状态起始时刻的音频输入，作为负样本；（3）模型预测值超过 0.5 且用户未打断时刻的音频输入，作为正样本；（4）从所有模型预测值低于 0.5 的样本中随机选取其他负样本，但排除用户手动激活前紧邻的 20 个样本，以扩充负样本集。

4 实验

4.1 验证实验

实验方法： 语音智能体系统运行于一台支持触摸屏的 x64 计算机上，以提供交互界面。遵循 [Coucke 等人， 2019]，使用 “Snips” 作为智能体的名称，但每次交互开始时无需通过该名称触发系统。根据第 3.3节中的模拟样本构建方法，构建了 1000 个按顺序排列的模拟数据样本，并按照数据的时间顺序进行序列化学习；每个样本在学习前都会先进行预测，以模拟真实交互场景。我们通过消融研究考察了基于扩张TCN的模型和序列回放在线学习策略（第3.2节）在用户对齐的语音响应时机预测任务中的有效性，并进一步将基于扩张TCN的模型与具有代表性的现有模型 Conformer [Gulati 等人、 2020]以及 Jasper [Li 等人， 2019]。

评估指标： 每 100 个样本计算一次 F1 分数和 Brier分数，其中 Brier分数衡量预测概率与实际结果之间的均方误差 [Glenn 等人， 1950]。

评估结果： 图4 展示了用于评估模型和学习策略有效性的实验结果。可以看出，引入膨胀时间卷积网络和基于回放的学习策略带来了最显著的性能提升。此外，基于扩张TCN的模型整体性能优于

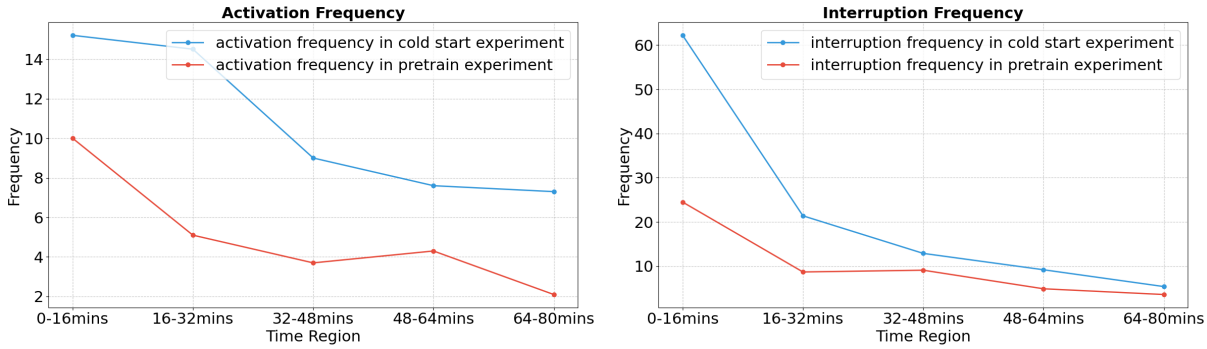


图5：在一致的环境条件下，持续参与者互动中平均激活次数和中断频率随时间变化的趋势。

	趋势导向问卷							总体问卷						
设置	Q1	Q2	Q3	Q4	Q5	Q6	Avg	Q1	Q2	Q3	Q4	Q5	Q6	Avg
冷启动	5.9	5.9	6.4	6.0	6.2	6.2	6.1	5.4	5.5	5.1	6.2	5.6	5.6	5.6
预训练	5.9	5.5	5.4	5.8	6.3	5.9	5.8	5.6	5.3	5.2	6.2	5.8	5.8	5.7

表1：面向趋势的问卷和总体问卷的用户研究评分结果。Q1–Q6分别对应系统响应准确性、喜好度、认知负荷、烦扰度、习惯性和速度。

Conformer 和 Jasper。相比之下，将 Transformer 模块用作主模型的内部配置，以及在学习策略中使用标签平衡机制，对性能的影响有限。

4.2 可用性实验

实验设置： 实验设置采用与第 4.1 节相同的硬件配置。我们告知参与者数据使用方式和潜在风险，并取得了他们签署的同意书。我们招募了20名参与者（男女比例均衡，为1:1），并将他们随机分为两组。第一组在冷启动设置下（使用未训练的初始模型）进行在线学习实验，第二组则基于由第一组的激活和中断样本构建的预训练模型进行在线学习。

参与者的年龄介于18至30岁之间；研究表明，这一人群对智能代理的接受度相对较高 [Ho等人，2022年，Hail等人，2024]，参与者的平均年龄为24岁。每位参与者的实验持续80分钟。在整个实验过程中，参与者持续与智能体互动；不过，每隔6分钟，参与者会与实验人员进行2分钟的轻松交谈。这种交替设计让智能体学会在背景对话期间避免不恰当的介入。同时，这些2分钟的交谈也作为定期的微休息，有助于缓解参与者的疲劳。当智能体响应过晚、过早或时机不当时，用户会通过点击屏幕手动纠正，启动或中断智能体；与此同时，在线学习机制利用根据这些基于点击的标签构建的样本，持续使智能体的响应时机与用户意图保持一致。在特定音量范围内播放背景音乐，设置参照第3.3节关于模拟样本构建的说明。参与者可以从 3.3 中选择对话主题 [Li 等人， 2017] 以发起对话。

实验结束后，问卷分为两种：（i）趋势导向问卷（详见附录B.1），用于了解参与者感知到的系统随时间推移的改进情况；（ii）总体问卷（详见附录B.2），用于反映参与者对系统的整体印象。两份问卷均依据 SASSI 框架 [Honeand Graham设计， 2000]，涵盖六个维度，每个维度包含六个问题。所有条目均采用7点李克特量表评分（1：非常不同意，7：非常同意）。

评估指标： 问卷评估了根据 SASSI 衍生出的六个维度 [Honeand Graham， 2000]：（1）系统响应准确性，（2）喜好度，（3）认知负荷，（4）烦扰度，（5）习惯性，以及（6）速度。

完成问卷后，我们进行了半结构式访谈，访谈内容包括参与者对智能体系统的整体印象，以及他们对问卷中显著评分背后动机的看法。

评估结果： 实验结果见图 5。图中显示，与冷启动设置相比，预训练模型降低了参与者的激活程度和中断频率；两种设置下的交互频率都逐渐降至较低水平，表明模型能够有效适应。进一步的统计分析证实，激活程度和中断频率均随时间发生显著变化（弗里德曼检验，$p < 0.01$），且在两种设置下，从第一个交互时段到最后一个交互时段均显著下降（威尔科克森符号秩检验，$p < 0.01$）。

平均问卷评分见表 1。进一步采用单样本 t 检验进行统计分析，结果证实，两组的趋势导向评分和总体评估评分均显著高于中性中点 4（$p < 0.001$）。此外，冷启动组内的威尔科克森符号秩检验显示，其趋势导向评分显著高于总体评估评分（$p = 0.0078$）。

对实验观察结果和访谈记录的分析揭示了以下几项发现：

发现 *1*（相互适应）。参与者通过激活和中断信号调整系统的同时，也调整了自己的语速和停顿时长，以更好地适应智能体，从而共同提升了对话流畅性。

发现 *2*（接受度驱动的个性化）。参与者对智能体过早或延迟的响应表现出不同程度的容忍度。当智能体过早响应时，一些参与者会停止说话，等待智能体完成响应；而另一些参与者则会在说完话语后耐心等待延迟的响应。这些容忍度差异导致模型拟合结果呈现个体化差异。此外，我们发现，针对容忍度较高的参与者训练的模型，能够更快地适应他们偏好的响应模式。

发现 *3*（培养效应）。参与者普遍感受到系统逐步改善，冷启动组尤为明显。一些参与者表示，他们获得了成就感，认为系统可以通过训练得到提升，并感到与智能体的亲密度有所增加，因此觉得该系统比传统语音助手更具特色。

发现 *4*（同步性的形成）。随着时间推移，模型预测值超过 0.5（从而触发智能体激活）的情况越来越多地与参与者在屏幕上的点击时序相吻合，这表明参与者与智能体的预期响应时间逐渐趋于一致。

我们获得了 2,478 个在线学习样本，其真实标签源自参与者的激活和中断行为。随后，我们将可用性实验期间收集的大量交互音频数据应用《真实样本挖掘》（第 3.3 节）中概述的方法，进一步扩展了数据集。这一过程产生了 17,416 个带标签音频样本。

4.3 扩展实验

基于用户研究中累积的数据，我们从两个方面开展了扩展实验。首先，我们使用挖掘得到的 17,416 个样本，复现第 4.1节中的显著结论，并评估用户特定和用户共享权重设置下的性能。其次，我们探究了我们的方法是否从音频信号中隐式学习了语义信息，以判断它能否在不依赖辅助语言转换或增强模块的情况下，有效实现语言理解。

挖掘样本上的性能 我们在以下设置下进行了实验：

设置 *1*，用户特定权重训练：针对每位参与者，按照时间顺序依次训练其对应的音频样本。在训练每个样本之前，先生成预测并计算评估指标，以模拟真实的冷启动在线学习场景。最终结果报告为所有参与者的平均值。

设置 *2*，用户共享权重训练：合并所有挖掘得到的音频样本，同时确保合并前后每位参与者的样本时间顺序保持不变。随后，在不同参与者的样本之间随机交错排列，并用于串行训练。评估流程与设置 1 相同，用以模拟多个用户共享模型权重的在线学习场景。

实验结果如图6所示。为了使比较更具针对性，我们仅保留了验证实验中与基于膨胀时间卷积网络的



图6：第4.3 节中的实验结果采用了验证实验中选出的组别子集，这些组别与基于扩张TCN的模型相比呈现出较为显著的差异。上排报告了**F1 分数**和**Brier分数**在**用户特定权重训练**下的结果，下排则报告了相应的**F1 分数**和**Brier分数**在**用户共享权重训练**下的结果。此项聚焦比较突显了基于扩张TCN的模型和序列回放在线学习策略的有效性。

模型相比表现差距较为明显的实验组。结果表明，无论采用用户专属权重训练还是用户共享权重训练设置，挖掘数据都仍然可以学习。此外，这也证实了第3.2节所述模型和学习策略的有效性。

语义学习分析我们开展了额外实验，结果表明，所提出的方法能够从音频信号中有效学习语言语义信息。我们使用自动语音识别转录了20位参与者的响应，并通过 *paraphrase-multilingual-MiniLM-L12-v2*[1]对句子进行编码。模型预测和理想正例构成了语义子空间。对于一个10步的时间窗口 t，我们计算了平均成对余弦相似度$C_{\text{pred}}(t)$和 $C_{\text{true}}(t)$，并得到离散增长率：

$$\Delta C_{\text{pred}}(t) = C_{\text{pred}}(t) - C_{\text{pred}}(t-1), \quad (1)$$

$$\Delta C_{\text{true}}(t) = C_{\text{true}}(t) - C_{\text{true}}(t-1). \quad (2)$$

然后，推导出两条轨迹之间的增长差异为 $D(t)$ $= \Delta C_{\text{pred}}(t) - \Delta C_{\text{true}}(t)$。

对所有窗口计算了 $D(t)$序列的累积和，并使用线性回归估计趋势斜率，并将其定义为语义学习指数（SLI）。通过自助法重采样评估 SLI 的显著性，以获得 95%置信区间（CI）。结果为：SLI $= 0.0116$，95%置信区间 $= [0.0088, 0.0148]$。这证实，我们的方法能够在理想子空间之外形成稳定的语义结构。

[1]https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

5 未来工作

Tap-to-Adapt 框架的未来研究可以聚焦于以下方向。

- **用户特定标签：**结合点击、手势和面部线索，确保标签来源于目标用户；
- **持续适应：**持续适应用户不断变化的交互习惯、环境和设备；
- **意图增强：** 利用点击以外的信号（如面部表情或语义线索）构建用户意图标签；
- **自适应停止：** 探索自适应语音停止时机；
- **视觉输入：** 融入视觉信号，以丰富信号来源；

- **多阶段系统：** 采用边云协同以提升效率；
- **任务泛化：**将所提出的 Tap-to-Adapt 框架扩展到新任务，例如生成式对话。

另一方面，必须恰当使用该框架，以防范用户偏好泄露或不当收集用户交互音频等风险。

6 结论

为了解决当前针对交互式语音智能体如何根据用户意图建模响应时机的研究不足问题，我们提出了 Tap-to-Adapt，这是一种通过用户自然的点击激活和中断行为来学习响应时机的在线学习框架。在该框架中，我们发现，结合顺序重放在线学习策略的基于扩张TCN的模型能够显著提升性能。此外，我们还提出了专为该任务设计的模拟样本构建与真实样本挖掘系统。借助这一框架，我们在约 20,000 个样本上开展了用户研究、消融实验和对比实验，以验证该任务的可行性以及我们方法的有效性。我们的框架为任务探索、模型与训练机制优化以及数据集扩充奠定了基础。

局限性

在第4节中，我们结合数据驱动的消融实验和对比实验以及用户研究，验证了所提出的 Tap-to-Adapt 框架及其相关模型与训练策略的可行性。然而，目前的样本规模仍不足以支持更细致的结论，例如通过预训练实现的响应时机预测模型的性能上限，或这些模型在持续演变的交互场景中的拟合行为。

参考文献

Galo Castillo-López、Gaël de Chalendar 和 Nasre- dine Semmar。《口语对话系统中轮次交替建模最新进展综述》。载于第*15*届国际口语对话系统技术研讨会论文集，第254–271页，2025年。

Hyunsoek Choi 和 Hyeyoung Park。采用人脸和手势的多模态用户认证系统。生物医学研究国际，2015年(1)：343475，2015年。

Alice Coucke、Mohammed Chlieh、Thibault Gisselbrecht、David Leroy、Mathieu Poumeyrol 和 Thibaut Lavril。使用膨胀卷积和门控实现高效的关键词检测。收录于*ICASSP 2019-2019 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6351–6355 页。IEEE，2019年。

Steven Davis 和 Paul Mermelstein。连续语音句子中单音节词识别的参数表示比较。*IEEE*声学、语音与信号处理汇刊，28(4):357–366，1980年。

Shihan Deng、Weikai Xu、Hongda Sun、Wei Liu、Tao Tan、Liujianfeng Liujianfeng、Ang Li、Jian Luan、Bin Wang、Rui Yan 等。Mobile-bench：一项用于基于大语言模型的移动智能体的评估基准。载于第*62*届计算语言学协会年会论文集（第*1*卷：长论文），第8813–8831页，2024年。

PranayDighe、Yi Su、VineetGarg、Sameer Dharur、Xiaochuan Niu、Ahmed H Abdelaziz、Saurabh Adya、Ahmed Tewfik 等。使用大语言模型进行后续对话的设备定向语音检测。*arXiv* 预印本 *arXiv:2411.00023*，2024年。

Erik Ekstedt和Gabriel Skantze。语音活动预测：轮流发言事件的自监督学习。*arXiv*预印本 *arXiv:2205.09812*，2022年。

Vignesh Ethiraj、Ashwath David、Sidhanth Menon和Divya Vijay。利用流式自动语音识别、量化大语言模型和实时文本转语音，实现面向电信的低延迟端到端语音智能体。*arXiv*预印本 *arXiv:2508.04721*，2025年。

Timothy J Fawcett、Chad S Cooper、Ryan J Longe- necker 和 Joseph P Walton。用于声学惊跳波形分类的机器学习、波形预处理和特征提取方法。*MethodsX*，8:101166，2021年。

Costa Ferro Filho、Elisa Ayumi Masasi de Oliveira、Iago Alves Brito、Pedro Martins Bittencourt 等。与说话人识别集成的唤醒词的实现与应用：一项案例研究。*arXiv* 预印本，第 arXiv–2407 页，2024年。

Luca Fredianelli、Francesco Artuso、Geremia Pom- pei、Gaetano Licitra、Gino Iannace 和 Andac Akbaba。用于声音事件分类与检测的环境噪声数据集。*Scientific Data*，12(1):1712，2025年。

Toru Furusawa。用于代理选择和车辆控制的本地大语言模型车载人工智能代理系统。载于*2025*年计算、网络与通信国际会议（*ICNC*），第259–264页。IEEE，2025年。

Toru Furusawa 和 Masatoshi Saitoh。面向车辆的多个大语言模型语音交互式人工智能代理演示。载于*2024*年 *IEEE* 智能计算国际会议（*SMART-COMP*），第249–251页。IEEE，2024年。

Milica Gaši´c、Filip Jurˇcí´cek、Blaise Thomson、Kai Yu 和 Steve Young。通过与人类受试者实时交互，对口语对话系统进行在线策略优化。载于*2011*年 *IEEE* 自动语音识别与理解研讨会，第312–317页。IEEE，2011年。

Shabnam Ghaffarzadegan、Hynek Boˇril 和 John HL Hansen。利用小型数据库和生成模型采样训练深度神经网络进行耳语语音识别。国际语音技术期刊，20(4)：1063–1075，2017年。

W Brier Glenn 等。以概率形式表达的预测验证。月度天气评论，78(1)：1–3，1950年。

Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu 等。Conformer：用于语音识别的卷积增强型 Transformer。*arXiv* 预印本 *arXiv:2005.08100*，2020年。

GAM Taufiq Hail、Shafiz Affendi Mohd Yusof、Am- mar Rashid、Ibrahim El-Shekeil 和 Abdalwali Lutfi。探讨影响 Z世代接受和采用人工智能及云端应用程序和工具以提升学业成绩的因素。新兴科学期刊，8(3):815–836，2024年。

Xinlu He、Yiwen Guan、Badrivishal Paurana、Zilin Dai 和 Jacob Whitehill。通过人类反馈交互式实时纠正说话人分离结果。*arXiv* 预印本 *arXiv:2509.18377*，2025年。

Manh-Tung Ho、Peter Mantello、Nader Ghotbi、Minh-Hoang Nguyen、Hong-Kong T Nguyen 和 Quan-Hoang Vuong。《情感人工智能时代重新思考技术接受：调查 Z 世代（Zoomer）对无意识数据收集的态度》。《*Technology insociety*》，70:102011，2022年。

Kate S Hone 和 Robert Graham。朝着构建一种用于语音系统界面主观评估的工具（SASSI）迈进。《*Natural Language Engineering*》，6(3-4):287–303，2000年。Koji Inoue、Bing’er Jiang、Erik Ekstedt、Tatsuya Kawahara 和 Gabriel Skantze。利用语音活动投射进行多语言轮次交替预测。载于*2024*年计算语言学、语言资源与评估联合国际会议论文集（*LREC-COLING 2024*），第11873–11883页，2024年。Zhong Ji、Jiayi Liu、Qiang Wang 和 Zhongfei Zhang。协调式经验回放：一种用于持续学习的协调式经验保留方法。《*Knowledge-Based Systems*》，234:107589，2021年。Christin Jose、Yuriy Mishchenko、Thibaud Senechal、Anish Shah、Alex Escott 和 Shiv Vitaladevuni。使用 CNN 准确检测唤醒词的起始和结束位置。载于*INTERSPEECH*，2020年。Veton Këpuska。唤醒词语音识别。《*Speech Technologies*》，第237–262页，2011年。Veton Këpuska 和 Gamal Bohouta。改进唤醒词和通用语音识别系统。载于*2017*年 *IEEE* 第*15*届国际可信、自主与安全计算会议、第*15*届泛在智能与计算国际会议、第*3*届大数据智能与计算国际会议及网络科学与技术大会（*DASC/PiCom/DataCom/CyberSciTech*），第318–321页。IEEE，2017年。

Veton Z Këpuska 和 TB Klein。一种新型唤醒词语音识别系统、唤醒词识别任务、技术与评估。《非线性分析：理论、方法与应用》，71(12):e2772–e2789，2009年。

Mohamed Khamis、Karola Marky、Andreas Bulling 和 Florian Alt。以用户为中心的多模态身份验证：利用视线与触摸输入保护手持移动设备。《行为与信息技术》，41(10):2061–2083，2022年。

Colin Lea、Michael D Flynn、Rene Vidal、Austin Reiter 和 Gregory D Hager。用于动作分割与检测的时序卷积网络。载于*IEEE*计算机视觉与模式识别会议论文集，第156–165页，2017年。

Jason Li、Vitaly Lavrukhin、Boris Ginsburg、Ryan Leary、Oleksii Kuchaiev、Jonathan M Cohen、HuyenNguyen 和RaviTeja Gadde。Jasper：端到端卷积神经声学模型。*arXiv*预印本 *arXiv:1904.03288*，2019年。

Siyan Li、Ashwin Paranjape 和 Christopher D Manning。我什么时候可以开口？预测语音对话代理的启动点。*arXiv*预印本 *arXiv:2208.03812*，2022年。

Yanran Li、Hui Su、Xiaoyu Shen、Wenjie Li、Ziqiang Cao 和 Shuzi Niu。DailyDialog：一个人工标注的多轮对话数据集。*arXiv*预印本 *arXiv:1710.03957*，2017年。

Yun-Wei Lin、Yi-Bing Lin、Yi-Feng Wu 和 Pei-Hsuan Shen。VoiceTalk：一种用于创建语音控制智能家居应用的无代码方法。《*IEEE*计算机学会开放期刊》，2025年。

Chao Liu、Mingyang Su、Yan Xiang、Yuru Huang、Yiqian Yang、Kang Zhang 和 Mingming Fan。通过设计支持打断和反馈语的大语言模型驱动的语音代理，助力老年人实现自然对话。载于*2025*年 *CHI* 人机交互系统会议论文集，第 1–22 页，2025a。

Guangyi Liu、Pengxiang Zhao、Yaozhen Liang、Liang Liu、Yaxuan Guo、Han Xiao、Weifeng Lin、Yuxiang Chai、Yue Han、Shuai Ren 等。手机自动化中的 LLM 驱动 GUI 智能体：进展与前景调研。*arXiv* 预印本 *arXiv:2504.19838*，2025年b。

Lei Liu、Li Liu 和 Yawen Cui。无先验均衡重放：面向长尾持续学习的不确定性引导的蓄水池采样。载于第*32*届 *ACM*国际多媒体会议论文集，第 2888–2897 页，2024年。

Sri Harish Mallidi、Roland Maas、Kyle Goehner、Ariya Rastrow、Spyros Matsoukas 和 Björn Hoffmeister。面向设备的语句检测。*arXiv* 预印本 *arXiv:1808.02504*，2018年。

Ayushi Mishra、Yang Bai、Priyadarshan Narayanasamy、Nakul Garg 和 Nirupam Roy。Sing：面向新一代可穿戴设备的大语言模型中的空间上下文。载于第四十二届国际机器学习会议，2025年。Atta Norouzian、Bogdan Mazoure、Dermot Con- nolly 和 Daniel Willett。探索用于将语音话语分类为面向系统和非面向系统的注意力机制。载于 *ICASSP 2019-2019 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第7310–7314页。IEEE，2019年。Aaron van den Oord、Sander Dieleman、Heiga Zen、Karen Simonyan、Oriol Vinyals、Alex Graves、Nal Kalchbrenner、Andrew Senior 和 Koray Kavukcuoglu。WaveNet：一种原始音频生成模型。*arXiv*预印本 *arXiv:1609.03499*，2016年。Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。载于国际机器学习会议，第28492–28518页。PMLR，2023年。Zhiwei Ren、Junbo Li、Minjia Zhang、Di Wang、Xiaoran Fan 和 Longfei Shangguan。迈向传感器在环大语言模型智能体：基准测试与启示。载于第*23*届*ACM*嵌入式网络传感器系统会议论文集，第254–267页，2025年。Dmitriy Rivkin、Francois Hogan、Amal Feriani、Ab- hisek Konar、Adam Sigal、Xue Liu 和 Gregory Dudek。通过自主大语言模型智能体实现人工智能物联网智能家居。*IEEE*物联网期刊，2024年。

David Rolnick、Arun Ahuja、Jonathan Schwarz、Timothy Lillicrap 和 Gregory Wayne。用于持续学习的经验回放。神经信息处理系统进展，32，2019年。

Saumya Y Sahai、Jing Liu、Thejaswi Muniyappa、Kanthashree M Sathyendra、Anastasios Alexan- dridis、Grant P Strimel、Ross McGowan、Ariya Rastrow、Feng-Ju Chang、Athanasios Mouchtaris 等。用于语音识别中高效唤醒词检测的双重注意力神经转导器。载于*ICASSP 2023-2023IEEE International Confer-ence on Acoustics, Speechand Signal Processing(ICASSP)*，第 1–5 页。IEEE，2023年。

Jan Schmalfuß-Schwarz、David Gollasch、Christin Engel、Meinhardt Branig 和 Gerhard We- ber。用于支持 ANC 的耳机的定制唤醒词集成：为办公室中的噪声敏感员工提升辅助技术。载于*Mensch und Computer 2024* 会议论文集，第 533–537 页。2024年。

Jens Schröder、JörnAnemüller 和 Stefan Goetze。针对 DCASE 2016挑战赛任务 3 的声学事件检测，对基于高斯混合模型、隐马尔可夫模型和深度神经网络的方法进行性能比较。载于*DCASE*，第 80–84 页，2016年。

Changhao Shan、Junbo Zhang、Yujun Wang 和 Lei Xie。用于小型关键词检测的基于注意力的端到端模型。*arXiv* 预印本 *arXiv:1803.10916*，2018年。

Gabriel Skantze。使用长短期记忆循环神经网络构建口语对话中轮次交替的通用连续模型。载于第*18*届*SIGdial*话语与对话年会论文集，第220–230页，2017年。

Gabriel Skantze。对话系统和人机交互中的轮次交替：综述。计算机语音与语言，67:101178，2021年。

Pei-Hao Su、David Vandyke、Milica Gasic、Dongho Kim、Nikola Mrksic、Tsung-Hsien Wen 和 Steve Young。向真实用户学习：使用神经网络评估对话成功率，以进行口语对话系统中的强化学习。*arXiv* 预印本 *arXiv:1508.03386*，2015年。

Pei-Hao Su、Milica Gasic、Nikola Mrkši´c、Lina M Rojas Barahona、StefanUltes、David Vandyke、Tsung-Hsien Wen 和 Steve Young。在线用于口语对话系统策略优化的主动奖励学习。载于 第*54*届计算语言学协会年会论文集（第*1*卷：长篇论文），第2431–2441页，2016年。

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need。神经信息处理系统进展，30，2017年。

Louis C Vroomen 和 Yves Normandin。基于鲁棒说话人无关隐马尔可夫模型的关键词检出器。在语音识别与理解：最新进展、趋势和应用中，第 95–100 页。施普林格，1992年。

Nicolas Wagner、Matthias Kraus、WolfgangMinker、David Griol 和 Zoraida Callejas。多用户对话界面综述。应用科学，15(13):7267，2025年。

Will Walker。使用语音打断上下文与人工智能代理进行自然对话交互。2024年。

Leran Wang、Md Shafaeat Hossain、J oshua Pul- frey 和 Lisa Lancor。Zoom触屏手势在身份验证和身份识别中的有效性及其随时间的变化。计算机与安全，111:102462，2021年。

Zhengxu Yan、Victoria Dube、Judith Heselton、Kate Johnson、Changmin Yan、Valerie Jones、Julie Blaskewicz Boron 和 Marcia Shade。理解老年人与智能语音助手的语音交互：一种加入人工输入的新型改进基于规则的自然语言处理模型。数字健康前沿，6:1329910，2024年。

Ziqi Yang、Xuhai Xu、Bingsheng Yao、Ethan Rogers、Shao Zhang、Stephen Intille、Nawar Shara、Guodong Gordon Gao 和 Dakuo Wang。Talk2care：一种用于医疗服务提供者与老年人之间沟通的基于大语言模型的语音助手。*ACM* 交互、移动、可穿戴与普适技术论文集，8(2):1–35，2024年。

Lillian Zhou、Yuxin Ding、Mingqing Chen、Harry Zhang、Rohit Prabhavalkar、DhruvGuliani、Giovanni Motta 和 Rajiv Mathews。礼物

反馈：通过联邦学习从用户纠正中学习，以提升自动语音识别模型质量。发表于*2023*年*IEEE* 自动语音识别与理解研讨会（*ASRU*），第 1–7 页。IEEE，2023年。

模拟样本构建的场景因素

当用户正在对智能体说话，且样本类型为“说话完毕，等待回复”时，将其定义为正样本。模拟所使用的因素如下所述：

- **与智能体的对话主题：**“日常生活”、“校园生活”、“文化与教育”、“态度与情感”、“人际关系”、“旅游”、“健康”、“工作”、“政治”或“金融”。重复前一个对话主题的概率设为50%。
- **与智能体的距离：**“30厘米”、“60厘米”或“90厘米”。重复前一个距离的概率为50%。
- **包含部分对话上下文的音频：**“是”或“否”。
- **用户说话对象：**“智能体”或“附近的其他人”。重复前一个说话对象的概率为50%。
- **样本类型：**“说完话，期待回复”、“说完话，不期待回复”、“说话时停顿”、“说话时不停顿”或“没有说话”。“说完话，期待回复”这一类型出现的频率是其他任何类型的四倍。

B 问卷

B.1 趋势导向问卷

1. **系统响应准确性**
   随着交互次数增加，你是否觉得系统的响应时机越来越符合你的预期?
2. **喜好度**
   随着交互次数的增加，你是否觉得自己对系统的好感逐渐提升，并越来越认为值得继续使用它?
3. **认知负荷**
   随着交互次数的增加，你是否觉得交互所需的注意力和脑力逐渐减少，从而让整个过程变得更轻松、更省力?
4. **烦扰度**
   随着交互次数增加，你是否觉得与系统交互时不那么令人沮丧，也不再那么容易产生失败感?
5. **习惯性**
   随着交互次数增加，你是否逐渐更清楚该对系统说什么，能够预判其响应时机，并且在交互过程中不再那么困惑?
6. **速度**
   随着交互次数的增加，您是否觉得与该系统的交互变得越来越高效?

B.2 总体问卷

1. **系统响应准确性**
   您在多大程度上同意该系统的响应时机符合您的预期?
2. **喜好度**
   您对该系统的印象有多好？您在多大程度上认为值得继续使用该系统?
3. **认知负荷**
   您在多大程度上同意，互动过程中所需的注意力和脑力投入很少，因此整个过程相对轻松省力?
4. **烦扰度**
   您在多大程度上同意，与该系统互动不会令人沮丧，也几乎不会让人产生挫败感?
5. **习惯性**
   您在多大程度上能够清楚地理解该对系统说什么，预判其响应时机，并在互动过程中减少困惑?
6. **速度**
   您在多大程度上认同与该系统交互是高效的?