

# Shop-R1: 通过强化学习奖励大型语言模型以模拟在线购物中的人类行为

Yimeng Zhang<sup>1</sup> Tian Wang<sup>2</sup> Jiri Gesi<sup>2</sup> Ziyi Wang<sup>3</sup> Yuxuan Lu<sup>3</sup> Jiacheng Lin<sup>4</sup>

Sinong Zhan<sup>5</sup> Vianne Gao<sup>2</sup> Ruochen Jiao<sup>2</sup> Junze Liu<sup>2</sup> Kun Qian<sup>2</sup> Yuxin Tang<sup>2</sup>

Ran Xue<sup>2</sup> Houyu Zhang<sup>2</sup> Qingjun Cui<sup>2</sup> Yufan Guo<sup>2</sup> Dakuo Wang<sup>3</sup>

<sup>1</sup>Michigan State University <sup>2</sup>Store Foundation AI, Amazon <sup>3</sup>Northeastern University

<sup>4</sup>University of Illinois Urbana-Champaign <sup>5</sup>Northwestern University

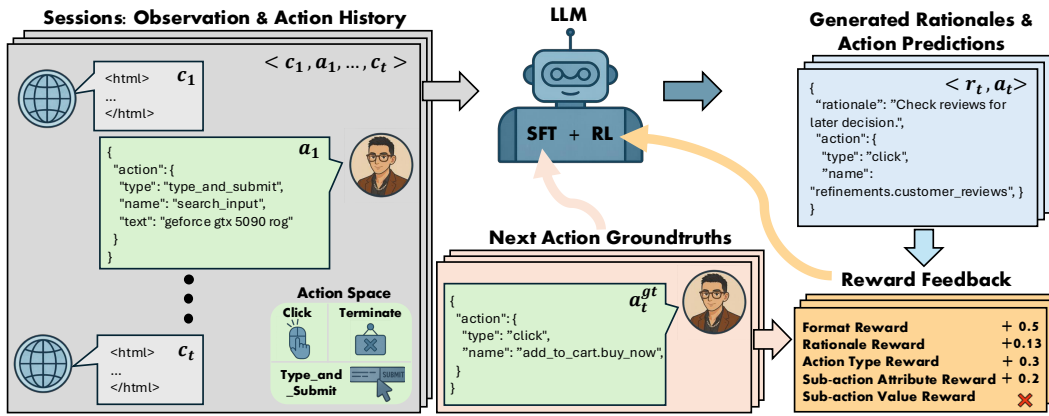


Figure 1: 提出的增强学习框架 *Shop-R1* 的概述，该框架旨在模拟基于网络的购物环境中的真实人类行为。根据带有相应网站观察的动作历史，模型基于历史和最新的网站观察预测下一个动作及其理由。生成的响应从四个角度进行评估：格式正确性，理由的自我确定性，动作类型准确性，以及子动作（属性和值）的准确性。

## Abstract

大型语言模型 (LLMs) 最近在网页环境中生成“可信的人类行为”方面表现出强大的潜力。先前的研究探索了用 LLM 综合推理来增强训练数据，并应用监督微调 (SFT) 以提高推理能力，从而改善下游动作预测。然而，此类方法的性能仍然受限于用于生成推理的模型的推理能力。在本文中，我们引入了 *Shop-R1*，一种旨在增强 LLMs 在模拟线上购物环境中的真实人类行为的推理能力的新颖强化学习 (RL) 框架。具体而言，*Shop-R1* 将人类行为模拟任务分解为两个阶段：推理生成和动作预测，每个阶段由不同的奖励信号引导。对于推理生成，我们利用内部模型信号（如 logit 分布），以自监督的方式指导推理过程。对于动作预测，我们提出了一种具有难度感知缩放的分层奖励结构，以防止奖励欺骗并实现细粒度的奖励分配。该设计评估了高层次动作类型和细粒度子动作细节（属性和值）的正确性，根据难度适当奖励输出。实验结果表明，我们的方法与基线相比，实现了相对超过 65 % 的改进。<sup>1</sup>

<sup>1</sup>代码和模型检查点将在论文接受后发布。

大型语言模型 (LLMs) 在计划、推理和决策任务中表现出色。最近, 研究人员开始利用 LLMs 在基于网络的环境中模拟人类行为, 旨在生成在数字平台上类似用户的现实行动序列。这一能力在诸如电子商务、教育和社交计算等领域有着很有前景的应用。尽管取得了这些进展, 但当前的 LLM 代理在产生与真人一致的行为方面往往不尽如人意。最简单的基准是零样本提示, 其中模型被给予文本指令以模仿某些用户类型, 并以预定义格式输出行动序列。虽然实施简单, 但这种方法缺乏进行高保真行为建模所需的个性化和适应性。为了提高行为准确性和推理连贯性, 最近的工作引入了合成训练数据增强。例如, 他们使用 Claude 3.5 Sonnet 生成理由, 以创建背景、行动、理由三联体。这些三联体随后用于执行监督微调 (SFT), 使模型能够学习行为及其背后的理由。然而, 这种方法面临关键限制: 理由的质量和多样性最终受数据生成过程中所用 LLM 的限制。

由于强化学习提供了一种灵活且有效的训练范式, 特别适合于反馈稀疏且延迟的环境, 并允许对行为输出进行细粒度的控制 [1-5], 我们利用强化学习来模拟人类购物行为, 相较于以前主要关注任务完成的工作 [6, 7]。在这项工作中, 我们提出了 *Shop-RL*, 一种新颖的强化学习框架, 旨在增强 LLMs 用于模拟人类在线购物行为。如图 1 所示, *Shop-RL* 将人类行为模拟任务分解为两个阶段: (1) 理由生成和 (2) 动作预测, 针对每个组件提供量身定制的奖励信号。对于奖励设计, 我们首先引入了一种二进制格式的奖励, 以鼓励模型生成解析友好的结构化响应, 从而促进可靠的下游评估和奖励计算。具体而言, 只有当模型的输出符合预期格式时, 它才会获得非零奖励; 否则, 其将被罚以零奖励。在推理生成方面, 获取真实推理本质上是困难的。尽管像 OPeRA [8] 这样的努力尝试从真实用户那里收集自我报告的推理, 但此类注释可能会遗漏隐性或无意识的决策因素。为了解决这个问题, 我们结合了自我确定性奖励 [9, 10], 通过模型输出分布与均匀分布之间的平均 Kullback-Leibler (KL) 散度 [11] 进行量化。该信号捕捉模型在其生成的推理中的置信度, 提供了一种无需监督的替代方案来获取真实推理。在动作预测方面, 我们通过引入一个分层奖励方案来超越二元奖励信号, 该方案考虑了动作类型和子动作的正确性。这一设计允许智能体对合理但不完美的行为获得部分奖励, 促进更平滑和更稳健的学习。此外, 为了减轻奖励操控并反映不同动作的难度差异, 我们应用了一种难度感知奖励缩放策略, 根据动作的复杂性调整奖励的大小。我们的主要贡献总结如下:

- 据我们所知, 我们是首个将强化学习引入面向模拟的人类行为建模任务中。我们将人类在线购物行为模拟重新表述为一个包含理由生成和行动预测的两阶段预测问题, 并为每个阶段设计了不同的强化学习目标。
- 我们介绍了 *Shop-RL*, 这是一种具有混合奖励设计的强化学习框架。它将用于推理生成的自我确定信号与用于动作预测的层次化奖励方案结合在一起。为了确保稳定学习和防止奖励作弊, 我们进一步引入了一种格式奖励和难度感知奖励缩放机制。
- 实验表明, 我们提出的训练流程实现了 27.72 % 的准确匹配率, 超过了监督微调 (16.76 %) 超过 65 %, 这表明我们的方法在面向仿真的人类购物行为建模方面表现出了强大的效果。我们进一步进行了全面的消融研究, 以评估我们设计中每个组件的贡献。

## 1 相关工作

用于模拟人类行为的大型语言模型 (LLM)。大型语言模型已经成为在多种现实世界场景中模拟人类行为的强大工具。最近的进展导致了代理系统的发展, 这些系统能够基于静态人物角色和交互历史生成合理的用户行为, 从而在社会科学 [12, 13]、推荐系统 [14] 和用户体验研究 [15] 等上下文中建模行为。这些系统通常根据用户资料 (例如, 偏好、人口统计) 和会话历史 (例如, 点击流、任务序列) 来预测下一个可能的用户行为, 以便进行个性化和上下文感知的模拟。除了行为预测之外, 最近的工作通过引入显性推理过程丰富了这些模拟。诸如 ReAct [16] 和基于反思的模型 [17, 18] 等方法促使大型语言模型在生成行为之前产生中间思维线索, 增强了解释性和决策质量。诸如 WebAgent [19] 和 UX-Agent [15] 等系统进一步通过专注的推理模型将任务分解成子目标, 在复杂环境如网页界面中提供了更好的控制。另一条研究线路探索了在动态环境中模拟多代理交互的基于代理的大型语言模型框架 [20-22]。这些系统通常采用模块化角色 (例如, 规划者、执行者) 和协作推理 [23, 24], 提供了对新兴社会行为和团队动态的见解。尽管最近取得了进展, 但在探索如何利用强化学习 (RL) 进一步增强大型语言模型在人类行为模拟中的应用, 特别是在基于网络的购物环境中, 仍存在显著的差距。

强化学习的奖励设计。奖励设计在强化学习算法的有效性和泛化上起着核心作用, 尤其是在将大型语言模型 (LLMs) 与期望行为对齐的背景下。一个突出的范式是来自人类反馈的

强化学习 (RLHF)，这种方法广泛应用于使用训练于人类偏好数据的奖励模型来微调大型语言模型 [25]。尽管 RLHF 展示了强大的对齐能力，但它常常受到收集可靠人类标注的高成本和有限的可扩展性的限制 [26]。此外，奖励模型本身可能会引入对齐偏差和不准确性，尤其是在有限或噪声较大的偏好比较上训练时 [27]。为了解决这些限制，直接偏好优化 (DPO) [2] 提出了一种更为高效的替代方法，该方法直接根据人类偏好信号优化模型参数，而无需显式的奖励模型。虽然在计算上较轻量，但 DPO 及其变体仍然依赖于人类生成或近似的偏好数据的可用性和质量，这在不同任务和领域中可能不一致。一个补充方向通过具备可验证奖励的强化学习 (RLVR) 出现，特别适合具有确定性正确性标准的领域，如代码生成和数学推理 [28, 29]。RLVR 框架使用基于规则的验证器来根据严格的正确性（例如，精确字符串匹配或功能等价性）自动计算奖励信号，绕过了对人类反馈的需求。这种向自动化客观奖励函数的转变使得能够训练出高度有能力的模型，如 DeepSeek-R1 [28]，并且激发了新的策略优化方法，如 GRPO [30] 及其最近的扩展 [31, 32]。尽管有这些进展，在人类行为的 RL 中，奖励设计仍是一个基本挑战。RLHF 在建模主观任务方面提供了灵活性，但常常面临可扩展性和可靠性问题 [33–36]。相比之下，RLVR 通过依赖于明确定义的评估标准提供了高精度，但仅限于存在此类标准的任务 [29, 37, 38]。为了解决模拟人类在线购物行为的独特挑战，我们提出了一种专门为该领域量身定制的混合奖励框架。

## 2 方法论

在本节中，我们首先在基于网络的购物环境下定义人类行为模拟问题。然后，我们展示了我们提出的 RL 框架 *Shop-R1* 的设计，该框架专门用于在这种环境下模拟人类行为。

问题陈述。在网络购物的背景下，一个用户会话由一系列多步骤的动作  $a_{1...t...N}$  组成，通常以搜索查询开始，并以产品购买或终止行为（例如关闭浏览器）结束。根据 [39] 的设置，动作空间包括三种主要的动作类型：‘type\_and\_submit’，‘click’和‘terminate’。有关动作空间的更多细节可见 App. A。每个动作  $a_t$  都配有相应的理由  $r_t$ ，它捕捉了用户在该时间步的潜在动机或理由。模型还接收上下文信息，即观察空间，表示网络环境的当前状态。该上下文被编码为简化的 HTML 结构，如在 Lu et al. [40] 中介绍的，这保留了基本布局和内容元素，同时舍弃了非信息性组件，如脚本和样式。人类在线购物行为模拟的任务被形式化地定义为学习一个函数  $f$ ，该函数在给定累积上下文和动作历史的条件下预测下一个理由和动作：

$$f(c_{1...t}, a_{1...t-1}, r_{1...t-1}) = r_t, a_t, \quad (1)$$

，其中  $f$  表示被训练用于通过生成下一步理由  $r_t$  和动作  $a_t$  来模拟用户行为的模型，其条件基于之前的上下文  $c_{1...t}$ ，过去的动作  $a_{1...t-1}$ ，以及之前的理由  $r_{1...t-1}$ 。这些理由是使用 LLMs 生成的，并在冷启动的监督微调阶段作为监督信号。需要注意的是，在随后的 RL 阶段中不使用生成的理由。

通过 SFT 冷启动。按照 Guo et al. [28] 的方法，我们通过对注释轨迹进行监督微调 (SFT)，初始化行为模拟模型  $f$ ，其中每个推理是由 Claude 3.5 Sonnet [41] 通过 Amazon Bedrock 生成的，而不利用任何用户资料信息。这个 SFT 阶段作为后续强化学习 (RL) 的冷启动，将模型建立在现实的推理和行动模式之上。在此阶段，模型被训练以联合生成推理和相应的行动。训练目标是最大化在输入查询  $q_t = c_{1...t}, a_{1...t-1}, r_{1...t-1}$  条件下，真实推理-行动对的似然性：

$$L_{\text{sft}} = - \sum_{t=1}^N \log p(r_t, a_t | q_t), \quad (2)$$

这种监督初始化在训练过程中早期帮助模型内化上下文、推理和行动之间的结构依赖关系方面起着至关重要的作用。通过提前将模型建立在这些模式之上，我们显著增强了后续 RL 阶段的稳定性和样本效率。更重要的是，它为高质量长文本输出的构成提供了明确的信号，例如正确命名被点击的按钮或指定有意义的搜索查询。这些能力在仅通过 RL 获得时是非常困难的，尤其是考虑到稀疏且延迟的奖励结构。

**Shop-R1**。为了更好地在模拟人类行为的环境中指导政策优化，我们将每一步分解为两个子任务：推理生成和动作预测。每个子任务都被分配了量身定制的奖励以提高一致性和可解释性。为了确保从模型输出中解析预测的推理和动作的简便性和正确性，我们引入了一种二进制格式奖励，鼓励模型以结构化的 JSON 格式生成响应。该格式遵循具有两个键值的字典模式：‘rationale’和‘action’。对于推理生成，我们采用自我确定性评分 [9, 10]，该评分量化了模型对其生成的推理的信心水平。具体来说，我们计算模型在整个输出序列中词汇上的

**Table 1:** 具有难度感知奖励缩放（DARS）的分层奖励机制。响应如果符合有效的 JSON 格式，则获得 0.5 的格式奖励；否则，不获得格式奖励。一个有效的响应进一步可以获得 (i) 正确的动作类型，(ii) 所需子动作属性的存在，以及 (iii) 任何长文本值预测的部分奖励，其奖励等于 DARS 因子乘以其与真实值的 ROUGE-L 相似度。

| Action Type     | Type Reward | Sub-action Attribute Reward                                       | Text-Similarity Value Reward                             |
|-----------------|-------------|-------------------------------------------------------------------|----------------------------------------------------------|
| terminate       | 0.3         | None                                                              | None                                                     |
| click           | 0.3         | +0.2 (if name $\neq \emptyset$ )                                  | +DARS $\times$ ROUGE-L(name)                             |
| type_and_submit | 0.3         | +0.1 (if name $\neq \emptyset$ ) +0.1 (if text $\neq \emptyset$ ) | +0.1 $\times$ ROUGE-L(name) +DARS $\times$ ROUGE-L(text) |

预测分布与均匀分布之间的 KL 散度：

$$s(r_t | q_t) = \frac{1}{N|V|} \sum_{j=1}^N \sum_{i=1}^{|V|} p_{ij} \log \left( \frac{p_{ij}}{U_i} \right), \quad (3)$$

，其中  $N$  是生成的推理中的词元数量  $r_t$ ， $p_{ij}$  是在位置  $j$  的词元  $i$  的预测概率， $U_i = \frac{1}{|V|}$  是词汇  $V$  上的均匀分布。 $s(\cdot)$  的较高值表明模型推理的确定性和一致性更高。对于动作预测，我们用一种层次化的奖励机制取代脆弱的二元信号，该机制既奖励粗粒度的动作类型又奖励其细粒度的子动作，以稳定训练并阻止恶劣的奖励黑客政策。该层次化机制使奖励景观更加密集：它扩展了利润丰厚的轨迹集合，使代理从通常阻碍政策搜索的“无奖励”高原中解救出来，并使奖励黑客行为变得不经济。具体而言，无论是简单还是困难的动作，一旦其高级类型正确，都能获得相同的粗略级别奖励；只有更复杂的动作可以通过其长文本子组件解锁额外的收益。因此，单纯地反复执行琐碎的“结束”动作不再能获得有竞争力的收益，而执行完整的（“点击”，“输入并提交”）序列成为最有利可图的策略。具体来说，“点击”动作包含一个子动作，指定要点击的按钮名称；正确预测的组件会获得部分奖励。同样，“输入并提交”包含子动作，提供预期的文本内容。相比之下，“结束”没有子动作，只在动作类型层面进行评分。预测准确度使用任务特定的指标：离散动作类型使用精确匹配标准，而自由形式的子动作使用 ROUGE-L 评估。基于文本的子动作，比如按钮标签或搜索查询，会根据其与真实情况的 ROUGE-L 相似度获得一个软奖励，但仅当该相似度超过预定阈值（例如 0.75）时才会得到。因为长文本子动作相当困难，现代网页可以显示数千个候选元素，我们引入了一个难度感知奖励缩放（DARS）因子，放大对正确预测这些组件的奖励。这防止了代理反复选择琐碎的“结束”动作以获取简单分数的奖励作弊行为。Tab. 1 中总结了所提出的分层奖励方案。将这些组件结合在一起，*Shop-RL* 的目标是最大化从多个来源获取的组合奖励信号，同时通过与一个参考策略的 KL 散度进行正则化：

$$\max_{\pi_\theta} \mathbb{E}_{r,a \sim \pi_\theta(q)} [v(a) + \alpha s(r) + -\beta \text{KL}(\pi_\theta(r, a | q) \| \pi_{\text{ref}}(r, a | q))], \quad (4)$$

，其中  $\pi_{\text{ref}}$  表示固定的参考策略， $v(a_t)$  表示行动预测的奖励， $\alpha$  和  $\beta$  是控制相应正则化项强度的超参数。

### 3 实验

#### 3.1 实验设置

数据集和模型。我们的研究基于一个专有的语料库，其中包括从一个领先的全球电子商务平台收集的 52,137 个真实购物会话。每个会话记录了人类客户与网站界面的多轮互动。我们为每个记录的动作添加了由 Claude 3.5 Sonnet 自动生成的自然语言理由（提示细节见附录 B）。所提供的观察上下文被格式化为简化的 HTML [40]，保留了基本的结构元素，同时过滤掉了无关的内容，如脚本、样式信息和用户特定的数据。对于 SFT 数据集，我们保持每个会话的完整性。模型需要生成包含理由和结构化动作预测的助手响应。对于 RL 数据集，我们将一个会话转换为 <上下文, 动作> 对的序列。上下文是 (i) 所有先前观察到的上下文和 (ii) 已经采取的动作的拼接；目标仅为下一个动作。因为每个会话都是从主页开始的，所以在第一次预测步骤之前，总会至少有一个观察到的 <context, action> 对，这消除了第一次操作的开放世界歧义。为了给模型在更困难的行为上提供稍微丰富的监督，两个复杂操作（点击和输入\_并提交）各自比简单的终止操作多出现约 10%。这种轻微的偏斜防止了学习者对琐碎案例的过拟合，同时仍然保持接近平衡的覆盖率，从而支持每类的公平和信息性评估。所有实验都微调了公开可用的 Qwen-2.5-3B-Instruct 模型。默认的 3B 参数主干提供了良好的计算性能平衡。

用于比较的基线。我们评估了我们的方法与几种基线方案的对比：(a) 零样本提示，即模型仅根据指令提示生成输出而无需额外训练；(b) RL（二进制），即基础模型直接通过强化学习进行优化，仅使用稀疏的二进制奖励信号；(c) 仅有监督微调，即通过使用大型语言模型生成的推理数据进行有监督微调对模型进行训练；(d) 监督微调 + RL（二进制），这扩大了监督微调，结合基于精确动作匹配的二进制奖励进行强化学习；以及 (e) Shop-R1，我们提出的用于模拟导向的人类行为建模任务的混合奖励设计的 RL 框架。

训练设置。我们的代码库基于 verl [42] 开发，所有实验均在 NVIDIA A100 GPU (80 GB) 上进行。我们使用了 PyTorch [43] 中的完全分片数据并行 (FSDP) 以最大化训练效率。默认的策略优化算法是 Group Relative Policy Optimization (GRPO) [30]。输入序列被填充或截断到最大上下文长度为 32k 个标记，默认的采样温度为 0.6。我们将每个设备的批量大小设置为 1，从而得到一个全局批量大小为 64。对于监督微调 (SFT)，我们以学习率  $2 \times 10^{-5}$  训练 4 个周期；对于强化学习 (RL)，我们以学习率  $1 \times 10^{-7}$  训练 500 步。默认情况下，我们将 DARS 因子设置为 1000，并使用  $\alpha = 0.005$  和  $\beta = 0.001$  来对相应的奖励项进行加权。

评估指标。我们为预测用户操作的准确性评估应用了一个完全匹配标准。只有当每个相关组件与真实数据完全匹配时，一个预测才被认为是正确的。例如，在“点击”操作的情况下，特定的子类型（例如点击过滤器、搜索区域或其他 UI 元素）和选择的目标都必须与真实标签一致。类似地，对于“输入并提交”的操作，模型应重现输入文本的相似含义。此外，我们单独报告粗粒度操作类型的准确性和 F1 分数。将这些分数与完全匹配的准确性进行比较，以突出剩余错误是源于高层操作类型的错误分类还是细粒度标签（按钮名称或查询文本）的错误。

### 3.2 实验结果

与基线的性能对比。主要的性能对比结果如 Tab. 2 所示。首先，零样本提示的表现较差：Qwen-2.5-3B-Instruct 在没有任何任务特定调整的情况下仅实现了 0.32 % 的精确动作准确率，这证实了仅靠通用指令调整无法恢复长时序网页行为。其次，仅靠稀疏二元奖励进行的强化学习仍无法为代理提供有意义的指导。当我们在这一信号下从零开始训练策略时，它仅达到 1.01 % 的精确匹配动作准确率和 6.17 % 的类型准确率。第三，简单的一轮 SFT 更为有效，将性能提升到 16.76 % 的精确匹配准确率和 22.25 % 的类型准确率。这证实了密集的教师强制轨迹对于注入结构知识（上下文  $\rightarrow$ 、理由  $\rightarrow$  动作）和展示长文本字段（按钮标签或搜索查询）的形态至关重要，是二元信号所无法传达的。第四，在 SFT 之后附加一个二元奖励的强化学习阶段仅能带来混合的结果：精确匹配动作准确率实际上下降到 16.55 %，而类型级别的 F1 上升到 28.07 %。因此，代理更好地学习了猜测粗略意图，但仍难以复现驱动精确匹配指标的长文本值。换句话说，策略在猜测粗略动作类型方面变得更好，但在复现需要精确匹配的细粒度长文本值方面略有下降。由于缺乏更细粒度的奖励分配，二元目标无法推动模型超越 SFT 已经达到的水平，在某些方面甚至有所倒退。用这一目标进行训练不仅容易导致不稳定性，并且相比使用更丰富、结构化的奖励进行优化，其收敛速度要慢得多。我们提出的 Shop-R1 框架在很大程度上缩小了这一差距。通过结合层次奖励、自信信号、格式奖励和难度感知缩放，它达到了 27.72 % 的精确动作准确率（较 SFT 相对提高 65 %），并将动作类型准确率和 F1 分别提高到 36.40 % 和 31.28 %。粗略（类型级别）和细粒度（精确匹配）指标的同步提升表明，Shop-R1 不仅更频繁地识别正确的意图，还能够更高保真度地再现长文本值（按钮标签、文本查询）。

Table 2: 在不同大小的模型中，不同微调方法下的模拟准确性。有三个互补的指标：精确动作准确率（所有子字段必须与标签匹配）；动作类型准确率，以及动作类型的 F1，以区别粗糙意图图分类中的错误与长文本参数中的错误。

| Model                  | Settings            | Exact Action | Action Type |         |
|------------------------|---------------------|--------------|-------------|---------|
|                        |                     | Acc.         | Acc.        | F1      |
| Qwen-2.5-3B-Instruct   | Zero-shot prompting | 0.32 %       | 15.33 %     | 16.15 % |
|                        | RL (Binary)         | 1.01 %       | 6.17 %      | 9.92 %  |
|                        | SFT                 | 16.76 %      | 22.25 %     | 24.52 % |
|                        | SFT + RL (Binary)   | 16.55 %      | 23.74 %     | 28.07 % |
|                        | Shop-R1 (Ours)      | 27.72 %      | 36.40 %     | 31.28 % |
| Qwen-2.5-1.5B-Instruct | Zero-shot prompting | 0.53 %       | 3.94 %      | 6.16 %  |
|                        | SFT                 | 10.86 %      | 23.58 %     | 29.02 % |
|                        | Shop-R1 (Ours)      | 24.11 %      | 34.54 %     | 29.19 % |
| Qwen-2.5-0.5B-Instruct | Zero-shot prompting | 6.76 %       | 12.88 %     | 15.55 % |
|                        | SFT                 | 9.90 %       | 17.72 %     | 21.61 % |
|                        | Shop-R1 (Ours)      | 27.72 %      | 31.83 %     | 21.20 % |

如 Tab. 3 所示，我们按动作类型分解准确性。零样本提示显示了经典的“意图-内容”分离。例如，它可以猜测需要“点击” (38.7 % 类型准确性)，但几乎从未命名准确的 UI 目标 (0.58

Table 3: 每种动作类型的精确动作准确性和动作类型准确性: “click”, “type\_and\_submit” 和 “terminate”, 针对不同的模型和微调方法。

| Models               | Settings            | Exact Action Acc. Per Action Type |                 |           | Action Type Acc. Per Action Type |                 |           |
|----------------------|---------------------|-----------------------------------|-----------------|-----------|----------------------------------|-----------------|-----------|
|                      |                     | click                             | type_and_submit | terminate | click                            | type_and_submit | terminate |
| Qwen-2.5-3B-Instruct | Zero-shot prompting | 0.58 %                            | 0.15 %          | 0.00 %    | 38.7 %                           | 1.62 %          | 0.00 %    |
|                      | SFT                 | 4.93 %                            | 3.84 %          | 49.80 %   | 8.55 %                           | 15.36 %         | 49.80 %   |
|                      | SFT + RL (Binary)   | 8.12 %                            | 3.25 %          | 45.51 %   | 17.25 %                          | 13.88 %         | 45.51 %   |
|                      | Shop-RL (Ours)      | 7.39 %                            | 7.53 %          | 81.84 %   | 10.29 %                          | 28.66 %         | 81.84 %   |
| Qwen-1.5-3B-Instruct | Zero-shot prompting | 1.01 %                            | 0.15 %          | 0.39 %    | 10.00 %                          | 0.44 %          | 0.39 %    |
|                      | SFT                 | 4.49 %                            | 7.83 %          | 23.44 %   | 15.07 %                          | 32.35 %         | 23.44 %   |
|                      | Shop-RL (Ours)      | 3.62 %                            | 8.12 %          | 72.85 %   | 6.52 %                           | 34.12 %         | 72.85 %   |
| Qwen-0.5-3B-Instruct | Zero-shot prompting | 0.43 %                            | 0.15 %          | 24.02 %   | 12.90 %                          | 4.43 %          | 24.02 %   |
|                      | SFT                 | 3.19 %                            | 7.68 %          | 21.88 %   | 5.94 %                           | 26.59 %         | 21.88 %   |
|                      | Shop-RL (Ours)      | 0.72 %                            | 3.99 %          | 97.07 %   | 1.01 %                           | 17.87 %         | 97.07 %   |

% 准确)。即使 SFT 能提升性能，但增益仍然不均匀，这表明单靠教师强制并不能为模型提供足够的信号来预测高熵参数，例如搜索查询。在 SFT 之后附加一个稀疏的二进制 RL 阶段仍然不能提升这些更难的情况。Shop-RL 调整了这些激励措施，达到了更高的准确匹配率，这表明模型不再仅仅满足于选择正确的类型，而是学习识别正确的组件和查询文本。总结来说，密集和结构化的反馈是必不可少的：它克服了无奖励平台，并使奖励黑客行为不经济。

### 3.3 消融研究与分析

模型规模。Tab. 2 和 Tab. 3 展示了一个一致的缩放趋势。在零样本场景下，3B 骨干网络在粗略动作类型准确率上已经超越了其 1.5B 和 0.5B 的对应版本  $\times 4 \sim 5$  倍，这表明在网站购物环境下，较大的模型拥有更强的开箱即用的人类行为模拟先验。经过 SFT 后，所有规模的模型都有所提升，但对于两种较小的骨干网络相比 3B 模型的提升更为显著，表明示范学习弥补了容量的不足。Shop-RL 将每一个骨干网络提升到其最佳运行点，但增益的形态随规模而变化。3B 变体达到最高整体数据，并在两种复杂的动作类型上均匀分布其改进。相比之下，0.5B 模型几乎完全通过过度预测最简单的“终止”操作（97.07 % 精确）而忽略了更具语义要求的类别来达到相当的精确匹配准确率（27.72 %）。1.5B 骨干网络介于两者之间，在“点击”和“输入并提交”操作上恢复适度准确性，同时保持对“终止”操作的强烈但不压倒性的偏向。简而言之，扩展主要增强了模型处理长文本、高熵动作的能力；较小的网络仍然可以通过利用高回报的终止分支来匹配整体准确性，但这样做是以牺牲行为多样性为代价的。这些发现强调，尽管 Shop-RL 显著缓解了容量限制，但对于面向模拟的网页购物动作预测任务的真正掌握仍然有赖于更大的骨干网络。

采样温度。Fig. 2 表明 Shop-RL 对采样温度具有鲁棒性，但三个评估指标以不同方式反应，揭示了温度采样如何通过决策层次传播。动作类型准确率在整个温度范围内几乎保持不变（36%），因为该指标汇总所有预测：一个方向的小错分在很大程度上会被另一个方向的修正抵消，使总体命中率保持不变。相比之下，随着温度升高，F1 分数稳步下降（31.28%  $\rightarrow$  28.36%）；类别平均化会惩罚任何不对称的混乱增加。有趣的是，从默认的  $\tau = 0.6$  到  $\tau = 0.7$  的适度提升将精确匹配准确率提高到了其峰值 28.63%：一丝随机性帮助生成的响应逃离局部最大值，偶尔能组装完整的长文本论证，而贪婪解码可能错过。当  $\tau > 0.8$  时，添加的熵不再发现新的正确完成；相反，它破坏细粒度字段的速度比修复它们的速度更快，因此精确匹配达到平台期，而 F1 继续侵蚀。这种模式是预料中的，因为 SFT 阶段已经将模型锚定于数据集特定行为，优先考虑忠实模拟而非创造力。综合来看，这些趋势表明温度在 0.6-0.8 范围

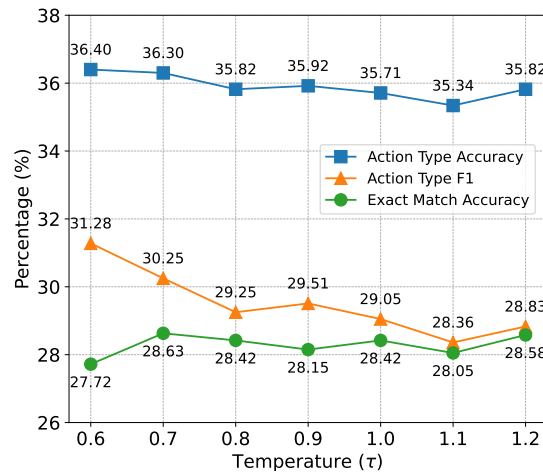


Figure 2: 采样温度消融研究。

提供最佳平衡，保持稳健的意图分类，最大化严格的精确匹配，并避免采样器过于探索时出现的指标退化。

Table 4: 不同训练组件配置的消融研究，使用精确匹配动作准确率和动作类型准确率/F1 进行评估。

| Model                | Training Scheme Components |               |                  |              |               | Exact Action | Action Type |         |
|----------------------|----------------------------|---------------|------------------|--------------|---------------|--------------|-------------|---------|
|                      | SFT                        | Format Reward | Rationale Reward | Reward Scale | Action Reward | Acc.         | Acc.        | F1      |
| Qwen-2.5-3B-Instruct | ✗                          | ✓             | ✓                | ✓            | hierarchical  | 4.63 %       | 36.56 %     | 21.92 % |
|                      | ✓                          | ✗             | ✓                | ✓            | hierarchical  | 2.87 %       | 3.19 %      | 5.04 %  |
|                      | ✓                          | ✓             | ✗                | ✓            | hierarchical  | 26.93 %      | 37.25 %     | 33.74 % |
|                      | ✓                          | ✓             | ✓                | ✗            | hierarchical  | 27.83 %      | 27.20 %     | 11.70 % |
|                      | ✓                          | ✓             | ✓                | ✓            | binary        | 27.41 %      | 27.46 %     | 12.11 % |
|                      | ✓                          | ✓             | ✓                | ✓            | hierarchical  | 27.72 %      | 36.40 %     | 31.28 % |

训练组件。Tab. 4 清楚地表明，*Shop-RI* 的每个元素针对的是不同的病理。去掉 SFT 的预热启动会削弱代理，尽管拥有所有 RL 信号，但精确匹配下降到 4.63 %，这强调了监督学习先验在掌握长文本论点形状是不可或缺的。省略格式奖励甚至更具破坏性，精确度降至 2.87 %，类型级指标降到 6 % 以下，因为无法解析的 JSON 输出得不到任何奖励，使学习者缺乏梯度信号。关闭自确定性（理由）奖励后，粗略的意图预测依然强劲，但精确匹配落后完整系统 0.8 %，这意味着对生成理由的明确反馈主要提高了动作的长文本部分，而不是它的顶级标签。禁用难度感知的奖励缩放或恢复到二进制作奖励会导致不同的失败模式：模型精确度仍可达到大约 27 %，但类型级 F1 降低到 11–12 %。检查表明，如果没有缩放或分层奖励，代理会偏向于简单高奖励的“终止”动作，很少涉及更困难的“点击”或“输入并提交”情况，这是一种经典的奖励滥用模式。完整的配置结合了所有信号，并展示了最佳平衡，表明每个组件都是必要的：SFT 注入语言先验，格式奖励保障可解析性，自确定性项提高长文本精度，而分层难度缩放奖励防止退化策略，同时促进细粒度动作的准确性。

整个会话与最新步骤的上下文。Tab. 5 隔离了在动作历史中包括每个访问页面的简化 HTML 的影响。去掉这一结构提示将精确匹配的准确率从 27.72 % 降低到 14.74 %，相对损失近 50 %，而粗略的动作类型准确率下降得更温和。这种显著的差异表明，虽然模型仍然可以仅从对话记录推断出下一个可能的交互动作类型，但在没有页面详细上下文的情况下，难以生成细粒度的参数，如精确的按钮标签或查询字符串。有趣的是，类别平衡的 F1 得分略有上升，这表明只有最新步骤上下文的变体通过更均匀地在动作类型中分布概率质量进行补偿，然而，这种重新分配并未转化为正确的长文本完成。简而言之，即便是一个令牌效率高的简化 HTML 视图的提供对于高保真模拟至关重要：它将语言模型锚定在具体的 UI 功能上，这是精确重现所需的。虽然它在上下文窗口中施加了大量的额外负担，但这种成本因其对准确模拟的必要性而变得合理。

Table 5: 将整个会话上下文或仅使用最新步骤作为输入时模型性能的比较。

| Settings      | Exact Action | Action Type |         |
|---------------|--------------|-------------|---------|
|               | Acc.         | Acc.        | F1      |
| whole-session | 27.72 %      | 36.40 %     | 31.28 % |
| latest-step   | 14.74 %      | 30.46 %     | 33.48 % |

在这项工作中，我们引入了 *Shop-RI*，一个新颖的强化学习框架，专为使用 LLMs 在基于网络的环境中模拟真实的人类行为而设计。通过将任务分解为推理生成和行动预测两个子问题，并为每个子问题配备精心设计的结构化奖励信号，*Shop-RI* 解决了之前方法仅依赖监督微调或稀疏二进制奖励的主要局限性。我们结合自我确定性评分、分层信用分配、格式正则化和难度感知缩放的混合奖励方案，在模型规模上显著提高了精确匹配的准确性和稳健性。大量实验表明，*Shop-RI* 不仅远远超越现有的基准，而且缓解了常见的问题，如奖励作弊和过度依赖简单操作。这些发现凸显了结构化 RL 框架在使语言代理能够执行细粒度、可解释和高保真度的行为模拟方面的潜力，为未来交互系统中更现实和个性化的虚拟用户建模铺平了道路。

## References

- [1] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, “A survey of reinforcement learning from human feedback,” *arXiv preprint arXiv:2312.14925v2*, 2024.
- [2] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 53 728–53 741, 2023.

- [3] T. Mu, A. Helyar, J. Heidecke, J. Achiam, A. Vallone, I. Kivlichan, M. Lin, A. Beutel, J. Schulman, and L. Weng, “Rule based rewards for language model safety,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 108 877–108 901, 2024.
- [4] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon *et al.*, “Constitutional ai: Harmlessness from ai feedback,” *arXiv preprint arXiv:2212.08073*, 2022.
- [5] A. Glaese, N. McAleese, M. Trębacz, J. Aslanides, V. Firoiu, T. Ewalds, M. Rauh, L. Weidinger, M. Chadwick, P. Thacker *et al.*, “Improving alignment of dialogue agents via targeted human judgements,” *arXiv preprint arXiv:2209.14375*, 2022.
- [6] S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried *et al.*, “Webarena: A realistic web environment for building autonomous agents,” *arXiv preprint arXiv:2307.13854*, 2023.
- [7] Q. Dong, L. Dong, Y. Tang, T. Ye, Y. Sun, Z. Sui, and F. Wei, “Reinforcement pre-training,” *arXiv preprint arXiv:2506.08007*, 2025.
- [8] Z. Wang, Y. Lu, W. Li, A. Amini, B. Sun, Y. Bart, W. Lyu, J. Gesi, T. Wang, J. Huang, Y. Su, U. Ehsan, M. Alikhani, T. J.-J. Li, L. Chilton, and D. Wang, “Opera: A dataset of observation, persona, rationale, and action for evaluating llms on human online shopping behavior simulation,” in *arxiv*, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:279244562>
- [9] Z. Kang, X. Zhao, and D. Song, “Scalable best-of-n selection for large language models via self-certainty,” *arXiv preprint arXiv:2502.18581*, 2025.
- [10] X. Zhao, Z. Kang, A. Feng, S. Levine, and D. Song, “Learning to reason without external rewards,” *arXiv preprint arXiv:2505.19590*, 2025.
- [11] I. Csiszár, “I-divergence geometry of probability distributions and minimization problems,” *The annals of probability*, pp. 146–158, 1975.
- [12] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative agents: Interactive simulacra of human behavior,” in *Proceedings of the 36th annual acm symposium on user interface software and technology*, 2023, pp. 1–22.
- [13] J. S. Park, C. Q. Zou, A. Shaw, B. M. Hill, C. Cai, M. R. Morris, R. Willer, P. Liang, and M. S. Bernstein, “Generative Agent Simulations of 1,000 People,” Nov. 2024.
- [14] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, X. Huang, Y. Lu, and Y. Yang, “Recmind: Large language model powered agent for recommendation,” *arXiv preprint arXiv:2308.14296*, 2023.
- [15] Y. Lu, B. Yao, H. Gu, J. Huang, J. Wang, L. Li, J. Gesi, Q. He, T. J.-J. Li, and D. Wang, “UXAgent: An LLM Agent-Based Usability Testing Framework for Web Design,” Feb. 2025.
- [16] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “React: Synergizing reasoning and acting in language models,” in *International Conference on Learning Representations (ICLR)*, 2023.
- [17] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, “Reflexion: Language agents with verbal reinforcement learning,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.
- [18] J. S. Park, J. O’Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, “Generative Agents: Interactive Simulacra of Human Behavior,” in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, ser. UIST ’23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 1–22.
- [19] I. Gur, H. Furuta, A. V. Huang, M. Safdari, Y. Matsuo, D. Eck, and A. Faust, “A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis,” in *The Twelfth International Conference on Learning Representations*, Oct. 2023. [Online]. Available: <https://openreview.net/forum?id=9JQtrumvg8>

- [20] X. Ma, Z. Zhang, and H. Zhao, “Coco-agent: A comprehensive cognitive mllm agent for smartphone gui automation,” *arXiv preprint arXiv:2402.11941*, 2024.
- [21] Z. Wang, H. Xu, J. Wang, X. Zhang, M. Yan, J. Zhang, F. Huang, and H. Ji, “Mobile-agent-e: Self-evolving mobile assistant for complex tasks,” *arXiv preprint arXiv:2501.11733*, 2025.
- [22] OpenAI. (2025) Introducing operator. [Online]. Available: <https://openai.com/index/introducing-operator/>
- [23] C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, and M. Sun, “ChatDev: Communicative Agents for Software Development,” Jun. 2024.
- [24] Q. Luo, Y. Ye, S. Liang, Z. Zhang, Y. Qin, Y. Lu, Y. Wu, X. Cong, Y. Lin, Y. Zhang, X. Che, Z. Liu, and M. Sun, “RepoAgent: An LLM-Powered Open-Source Framework for Repository-level Code Documentation Generation,” Feb. 2024.
- [25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
- [26] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, “Llama 2: Open foundation and fine-tuned chat models,” *arXiv preprint arXiv:2307.09288*, 2023.
- [27] L. Gao, J. Schulman, and J. Hilton, “Scaling laws for reward model overoptimization,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 835–10 866.
- [28] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, “Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning,” *arXiv preprint arXiv:2501.12948*, 2025.
- [29] Y. Su, D. Yu, L. Song, J. Li, H. Mi, Z. Tu, M. Zhang, and D. Yu, “Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains,” *arXiv preprint arXiv:2503.23829*, 2025.
- [30] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, “Deepseekmath: Pushing the limits of mathematical reasoning in open language models,” *arXiv preprint arXiv:2402.03300*, 2024.
- [31] Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu *et al.*, “Dapo: An open-source llm reinforcement learning system at scale,” *arXiv preprint arXiv:2503.14476*, 2025.
- [32] Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin, “Understanding r1-zero-like training: A critical perspective,” *arXiv preprint arXiv:2503.20783*, 2025.
- [33] D. Alsagheer, A. Kamal, M. Kamal, and W. Shi, “Governance challenges in reinforcement learning from human feedback: Evaluator rationality and reinforcement stability,” *arXiv preprint arXiv:2504.13972*, 2025.
- [34] H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi *et al.*, “Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback,” *arXiv preprint arXiv:2309.00267*, 2023.
- [35] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire *et al.*, “Open problems and fundamental limitations of reinforcement learning from human feedback,” *arXiv preprint arXiv:2307.15217*, 2023.
- [36] T. Moskovitz, A. K. Singh, D. Strouse, T. Sandholm, R. Salakhutdinov, A. D. Dragan, and S. McAleer, “Confronting reward model overoptimization with constrained rlhf,” *arXiv preprint arXiv:2310.04373*, 2023.
- [37] Y. Mroueh, “Reinforcement learning with verifiable rewards: Grpo’s effective loss, dynamics, and success amplification,” *arXiv preprint arXiv:2503.06639*, 2025.

- [38] X. Wen, Z. Liu, S. Zheng, Z. Xu, S. Ye, Z. Wu, X. Liang, Y. Wang, J. Li, Z. Miao *et al.*, “Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms,” *arXiv preprint arXiv:2506.14245*, 2025.
- [39] Y. Lu, J. Huang, Y. Han, B. Bei, Y. Xie, D. Wang, J. Wang, and Q. He, “LLM Agents That Act Like Us: Accurate Human Behavior Simulation with Real-World Data,” Apr. 2025.
- [40] Y. Lu, B. Yao, H. Gu, J. Huang, J. Wang, Y. Li, J. Gesi, Q. He, T. J.-J. Li, and D. Wang, “Uxagent: A system for simulating usability testing of web design with llm agents,” *arXiv preprint arXiv:2504.09407*, 2025.
- [41] Anthropic, “Claude 3.5 sonnet technical overview,” <https://www.anthropic.com/news/claude-3-5-sonnet>, June 2024, <https://www.anthropic.com/news/claude-3-5-sonnet>.
- [42] G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu, “Hybrid-flow: A flexible and efficient rlhf framework,” *arXiv preprint arXiv: 2409.19256*, 2024.
- [43] Y. Zhao, A. Gu, R. Varma, L. Luo, C.-C. Huang, M. Xu, L. Wright, H. Shojanazeri, M. Ott, S. Shleifer *et al.*, “Pytorch fsdp: experiences on scaling fully sharded data parallel,” *arXiv preprint arXiv:2304.11277*, 2023.

A

## Appendix

### A 系统提示

```
<IMPORTANT>
Your task is to predict the next action and provide rationale for the action based
on the previous actions and context.
You need to pretend that you are a user, browsing amazon.com and searching for a
product to purchase.
The history action (with details described below) and context will be provided to
you.
You need to predict the next action and provide rationale for the action.
</IMPORTANT>

# Action Space
An action is represented in JSON format, and there are three primary types of
actions:
#### 1. 'type_and_submit':
Type text into an input field and immediately submit the form. Equivalent to typing
text into an input and pressing enter key.
{
  "type": "type_and_submit",
  "name": "input_name",
  "text": "search_text"
}

#### 2. 'click':
Click on a button or clickable element identified by 'name'.
{
  "type": "click",
  "name": "clickable_name"
}

#### 3. 'terminate':
When you are unsatisfied with the current search result and you don't want to buy
anything, use 'terminate' to indicate that you want to close the browser window
and terminate the task.
{
  "type": "terminate"
}

# Context
Your context will be an **simplified version** of the raw HTML of the amazon page
you are looking at. Some interactable elements will be added a unique "name"
attribute, which you can use to identify the element to interact with (click or
type_and_submit).

# Rationale
The rationale is a first-person sentence of what you are thinking when you make the
action. It should be a short sentence that explains why you are making the
action.

# Output Format
You need to predict the next action and provide rationale for the action. Your
output should follow a strict JSON form:
{
  "rationale": "<rationale>", // rationale goes here, a string
  "action": {
    // action goes here
    "type": "<type>",
```

```
    }, ...  
  }  
<IMPORTANT>  
OUTPUT A SINGLE JSON OBJECT, NOTHING ELSE.  
</IMPORTANT>
```

## B 推理综合提示

You will be given a customer's shopping journey on one of the largest e-commerce platforms globally.  
You will be given the context (what the user is looking at), the action (what the user did), and your job is to predict the user's rationale for the action.  
The rationale should follow

Here is an example:  
{example}

For each action in the input, output a rationale.

If the action is "terminate", it means that you didn't find any desired product and you decided to leave the website by closing the browser window.