ChatShopBuddy：通过强化学习迈向可靠的对话式购物智能体

Yiruo Cheng*高瓴人工智能学院 中国人民大学 中国北京
chengyr@ruc.edu.cn

Kelong Mao
Tianhao Li 京东 中国北京
m
aokelong.1@jd.com

Jiejun Tan Ji-Rong Wen Zhicheng Dou†高瓴人工智能学院 中国人民大学 中国北京
dou@ruc.edu.cn

arXiv:2603.06065v1 [cs.IR] 6 Mar 2026

摘要

对话式购物智能体是大语言模型（LLM）驱动智能体面向消费者的关键应用，但如何有效应用后训练强化学习（RL）来优化此类智能体，仍缺乏充分研究。本文研究真实场景下基于强化学习的购物智能体优化问题。在这些场景中，智能体必须同时满足多个相互依赖的目标，涵盖客观指标（产品正确性）、主观质量（说服力）、结果奖励（最终响应质量）以及过程奖励（工具效率）。针对这一挑战，我们提出了一套完整的方法论。具体而言，我们首先构建了 SmartShopBench，这是一个能够捕捉多样化购物意图的基准测试，并通过分层评估将复杂的质量要求分解为可测量的层级。在此评估框架的基础上，我们设计了层次化奖励建模（HRM），通过条件门控来组织混合奖励类型，从而反映它们之间的逻辑依赖关系。为实现高效训练，我们进一步提出动态对比策略优化（DCPO），根据奖励和推理长度进行动态轨迹选择，在响应质量与操作效率之间取得平衡。大量实验表明，我们训练的强化学习智能体 ChatShopBuddy 持续优于依赖通用推理的大型模型，展现出更优的稳定性，而不仅仅是取得更高的峰值性能。我们的工作为将强化学习应用于真实世界的对话式智能体提供了有价值的指导。数据集和代码已在该代码仓库中公开。

CCS概念

• **应用计算** → **在线购物**；•**计算方法-论** → **智能体**；**强化学习**。

*本工作是Yiruo Cheng在京东实习期间完成的。†通讯作者。

允许免费为个人或课堂使用制作本作品全部或部分的数字副本或纸质副本，前提是不得制作或分发副本
用于营利或商业优势，且复制品必须带有此声明及首页上的完整引文。作者以外的其他方拥有本作品组件的版权，必须予以尊重。注明出处后可以摘录。若要以其他方式复制、重新发布、发布到服务器或分发至列表，则需事先获得特定许可和/或支付费用。请向 permissions@acm.org 申请许可。会议缩写 '*XX*，纽约州伍德斯托克 © 2018 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

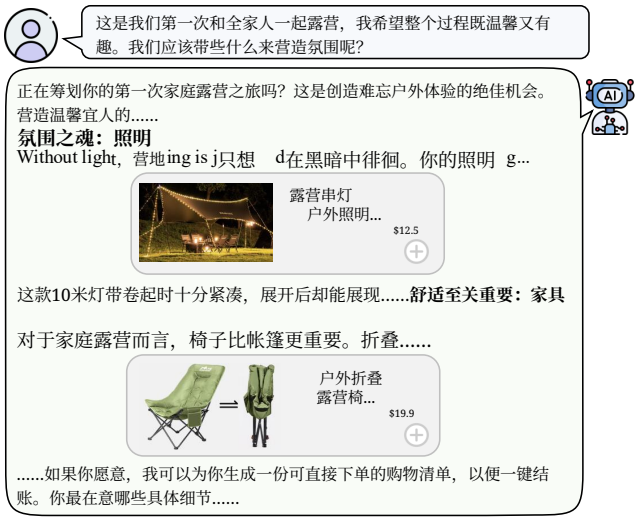


图1：对话式购物代理示例。

关键词

对话式购物代理，强化学习

ACM参考文献格式：
Yiruo Cheng、Kelong Mao、Tianhao Li、Jiejun Tan、Ji-Rong Wen和Zhicheng Dou。2018。ChatShopBuddy：通过强化学习实现可靠的对话式购物智能体。载于请务必根据版权确认邮件输入正确的会议论文集标题（会议缩写'*XX*）。ACM，美国纽约州纽约市，15页。https://doi.org/ XXXXXXX.XXXXXXX

1 引言

对话式购物智能体已成为电子商务领域的一种变革性范式，使用户能够通过自然对话表达复杂且由意图驱动的查询，而非依赖关键词搜索 [8, 32]。这些由大语言模型（LLM）驱动的智能体能够理解含糊的用户需求，例如"我们全家第一次去露营，我希望这次体验温馨又有趣。我们应该带些什么？"，并通过利用外部工具、推理能力和上下文理解，综合生成连贯的购物指导 [2, 42, 43]，，如图 1 所示。当购物智能体从研究原型迈向

实际部署时，它们提供可靠、完整且符合上下文的推荐的能力变得至关重要。

然而，对话式购物智能体在实际部署中面临着重大挑战。智能体必须提供流畅、结构良好的响应，同时确保内容准确、完整，并始终符合用户意图。近期研究揭示了推荐完整性、产品相关性和实用性方面持续存在的严重问题 [30, 42, 46]。强化学习（RL）为有效解决这些局限提供了一条自然途径：通过基于奖励的训练，直接优化智能体，使其表现出期望的行为。

将强化学习应用于对话式购物智能体面临着独特挑战，使其有别于许多强化学习领域。与具有确定答案的数学推理 [37, 50], 、可通过执行进行测试的代码生成 [48, 56],或具有明确成功标准的航班预订 [53], 不同，购物智能体必须同时优化多个难以直接验证的维度。高质量的购物响应不仅要求产品推荐准确无误，还需要有说服力的呈现、结构连贯性，并符合用户主观偏好。此外，实际部署还要求提高运营效率，缩短推理长度并减少不必要的工具调用，以维持可接受的延迟。多维度、难以直接验证的目标与效率约束并存，这对旨在处理单一目标或易于验证任务的标准强化学习方法构成了根本挑战。

在这项工作中，我们研究如何有效地应用强化学习来优化购物智能体，使其具备所需能力。我们发现，多维奖励建模、训练与平衡是核心挑战。这既涵盖产品正确性等客观指标，也包括说服力等主观质量；既涉及工具调用成功等直接验证的信号，也涉及结构连贯性等间接评估的属性。为系统地应对这一挑战，我们的方法分为三个阶段。

我们首先构建**SmartShopBench**，这是一个包含多样化购物查询和分层评估框架的基准测试，可将复杂的质量要求分解为可衡量的层级。一级（L1）评估器验证基本正确性，包括产品相关性和文本忠实度；二级（L2）评估器则评估更高阶的属性，包括结构连贯性和内容深度。在此评估框架的基础上，我们提出**层次化奖励建模（HRM）**，通过门控机制将多维目标综合为统一的训练信号。HRM 明确采用条件结构：必须先满足基本正确性，才会对更高层次的质量和效率给予奖励，从而防止奖励投机，并确保智能体优先考虑可靠性。最后，我们提出**动态对比策略优化（DCPO）**，这是一种能够平衡响应质量与运行效率的强化学习算法。DCPO 针对每个查询对轨迹进行过采样，并根据奖励分数与推理长度的联合效用应用动态对比选择策略，从而明确鼓励简洁、高质量且适合部署的推理路径。

在 SmartShopBench 上开展的大量实验得出了若干关键发现。首先，实验表明，ChatShopBuddy 这一专门针对该任务进行对齐的强化学习训练模型，持续优于依赖通用推理的更大规模模型，这表明定向后训练比单纯扩大规模更为有效。其次，强化学习主要提升的是响应的稳定性与一致性，而非仅仅提高峰值性能，从而弥补了生产系统的一项关键缺口。最后，我们发现，仅鼓励延长推理并不会在领域特定任务上带来稳定收益，这凸显了使训练目标与任务需求对齐的重要性。

我们的主要贡献总结如下：

(1) 我们系统地研究了强化学习在对话式购物智能体中的应用，并提出了**HRM**，一种门控奖励机制，用于协调可靠性、说服力和效率等多个目标，使购物智能体在确保基本正确性的同时达到较高的质量。

(2) 我们提出了**DCPO**，一种注重效率的强化学习算法，它利用动态对比选择策略，同时优化响应质量和推理长度，从而有效降低整体部署延迟。

(3) 大量实验表明，我们经过任务对齐强化学习训练的模型 ChatShopBuddy 始终优于依赖通用推理的更大型模型，在无需过长推理长度的情况下提升了稳定性和效率，并凸显了定向后训练优化的重要性。

2 相关工作

2.1 对话式购物代理

由大语言模型驱动的对话式购物智能体 [30, 32, 43, 46] 正日益成为电子商务领域备受关注的一种新范式，因为它们能够与用户互动，了解需求，推理其偏好与约束条件，并有策略地调用外部工具，生成高质量的推荐，并附上连贯且可信的解释 [2, 42]。尽管近期研究已展现出基础模型在对话式购物场景中的强大零样本或少样本能力 [42, 43], , Shopping- Comp [42] 的结果表明，即使是能力出众的智能体，仍倾向于产生冗余的内部推理，还可能生成流畅却与真实产品属性不符的响应，或无法满足用户意图。这些局限凸显了采用有效后训练方法 [46] 的必要性，以更好地使智能体与用户意图保持一致，提供更可靠的购物指导，并减少不必要的推理。因此，在本研究中，我们提出利用强化学习，从意图对齐、响应质量和推理效率等方面提升对话式购物智能体。

2.2 智能体的强化学习

强化学习（RL） [3, 35]已成为释放自主智能体潜力的强大范式 [11, 22 –24, 29, 45]。通过利用奖励信号优化智能体行为，RL 使智能体能够提升复杂推理能力和自主工具使用能力 [1, 7, 13]。

因此，近期有一系列研究聚焦于奖励客观且可直接验证的场景 [3, 25, 26, 39, 44, 49] 。在这些场景中，奖励可以根据明确的结果获得，例如数学问题中的数值标准答案 [37, 50, 61] , 或代码生成中的编译器反馈 [48, 56, 60]。这种直接可验证性使学习过程

表1：SmartShopBench 中的代表性查询。

类别	示例查询
模糊搜索	“我准备第一次去见女朋友的父母，带什么礼物比较合适？”
搜索-多重-约束	“我想找一台黑色的全自动咖啡机，内置研磨器和奶泡器，预算约为 3000 美元。”
捆绑搜索	“我最近想开始在家自己烘焙。请帮我搭配一套入门级的家用电器。”
搜索-通用	“推荐几款适合用于婴儿衣物。
问答比较	“我想了解雅诗兰黛 Double Wear 粉底液、兰蔻 Teint Idole 粉底液和阿玛尼 Luminous Silk 粉底液之间的区别。” 荐。”
问答咨询	“带独立分区的双槽水槽值得买吗？”

更加直接，且性能更易于量化。然而，许多现实世界中的智能体应用运行于开放式环境中，在这些环境里，奖励往往稀疏、主观或难以直接验证 [47, 54], ，这给通过强化学习训练的智能体优化带来了重大挑战。为拓展这一研究范围，一些研究通过引入基于评分标准的奖励 [12, 14, 16, 27, 59]，将强化学习驱动的智能体扩展至开放式领域。在这项工作中，我们进一步探索这一方向，将研究置于真实世界的对话式购物场景中，探讨奖励设计与策略优化，旨在开发能够满足用户意图、建立信任并提供高质量推荐的对话式购物智能体。

2.3 对话式推荐

对话式推荐 [15, 33] 旨在通过多轮交互帮助用户逐步表达偏好，并通过迭代反馈不断明确需求。现有的对话式推荐研究 [6, 17–19, 21, 34, 40, 51, 55, 57]主要关注如何提升推荐准确率，以及如何在闭环对话场景中保持语义连贯性。这些方面固然重要，但它们主要采用以检索为中心的视角，将任务视为选择相关商品并生成连贯对话 [28]。然而，在现实场景中，尤其是在线购物时，用户的需求不止于准确检索商品。用户日益期待有说服力的解释、有意义的比较和透明的决策支持，以便做出明智的选择 [42]。为弥合这一差距，我们采用由大语言模型驱动的对话式购物智能体，并配备专用工具，使系统能够超越纯粹的检索，在真实购物场景中提供透明且有说服力的指导。

3 SmartShopBench

为了系统地研究面向对话式购物智能体的强化学习，我们需要一个评估框架，用以衡量前文所识别的多维质量属性。我们构建了**SmartShopBench**，将多样化的真实购物查询与一个层次化评估框架相结合，该框架旨在将复杂的质量要求分解为可衡量的组成部分。

3.1 数据集构建

我们通过两阶段方法构建该基准测试：首先建立基于真实用户行为的查询类型分类体系，然后利用该分类体系指导查询生成。

3.1.1 查询分类体系。 通过分析生产平台上数千条真实用户查询，我们归纳出六种基本查询类型，以概括对话式购物智能体面临的核心挑战。前四种以**搜索为导向**:

(1) 模糊搜索 查询表达宽泛的需求，但不指定具体产品（例如“旅行时能让我的植物保持健康的东西”），需要智能体创造性地探索产品空间。

(2) 多约束搜索查询同时对品牌、价格、规格及其他特征提出多项要求（例如“100 美元以下、安静且能碎冰的搅拌机”），要求智能体联合满足这些约束。

(3) 捆绑搜索 查询要求提供可协同使用的互补产品（例如“完整的家用咖啡设备套装”），用于测试智能体对兼容性和整体效用的推理能力。

(4) 搜索-通用表示需要准确理解意图并提供相关推荐的常见产品请求。

其余两类为**问答类**:

(5) 问答比较查询要求明确比较产品（例如，“iPhone 15 与 Samsung Galaxy S24”），并清晰阐述各项属性之间的权衡。

(6) 问答咨询查询寻求针对特定场景的实用建议（例如，“Sony WH-1000XM5 适合经常乘飞机旅行吗？”），需要进行情境推理并提供可执行的指导。

表1列出了各类别的示例 .

3.1.2 生成与标注流程。我们从主要电子商务平台收集当前在售的产品，并将其划分为十个代表性类别（服装、电子产品、家居用品等）。对于每种查询类型，我们使用大语言模型生成与这些产品相关的候选查询，并通过提示词指定结构和语义要求。所有查询均经过人工审核，以确保其自然、明确且能代表真实用户需求。最终，我们得到涵盖全部六个类别的1,680条查询，其中1,560条用于训练，120条留作测试。

3.2 分层评估框架

评估对话式购物智能体，需要考察从基本正确性到精细表达等多方面的特质。将所有维度合并为一个分数会掩盖失败原因，而将它们视为彼此独立则会忽视逻辑依赖关系。我们提出一种**分层评估框架**，包含两个层级：第 1 级（L1）验证基本正确性与可靠性，第 2 级（L2）则仅在 L1 通过后评估更高阶的特质。这样可以有效避免系统凭借措辞流畅、实则事实错误的响应获得高分。

*3.2.1 L1*评分器。 L1评分器作为基础层，用于评估智能体是否满足用户的信息需求和任务要求。对话式购物智能体生成的响应包含自然语言文本和结构化产品推荐（以用户界面卡片的形式展示）。L1评分器通过分析以下内容之间的关系来评估正确性:

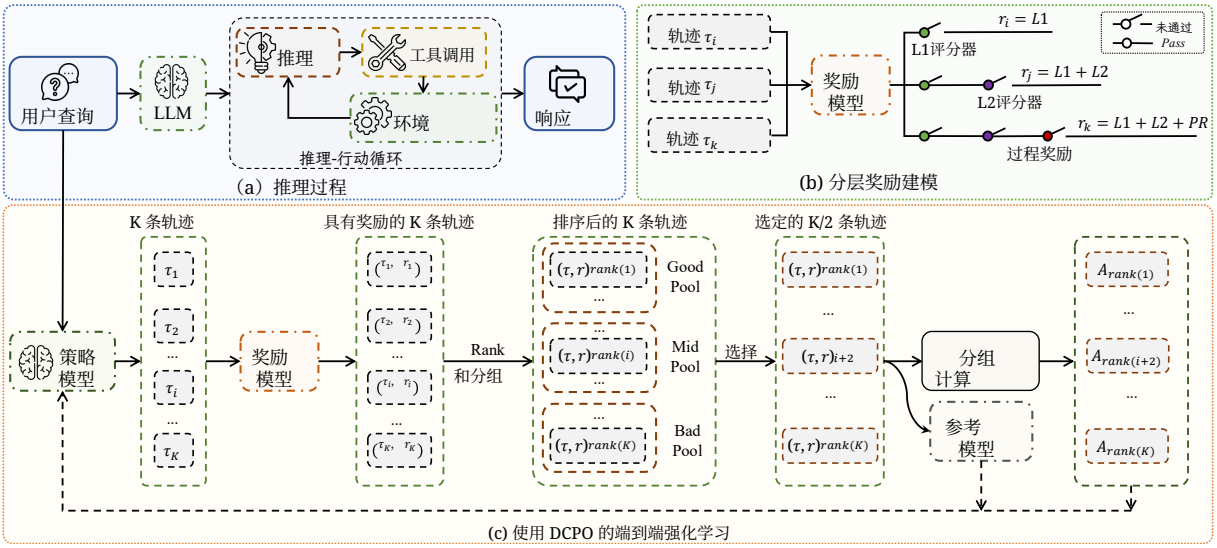


图 2：对话式购物代理概览。图 (a) 展示了购物智能体的推理过程。图 (b) 展示了分层奖励建模。绿色门表示 L1评分器是否通过，紫色门表示 L2评分器是否通过，红色门表示过程奖励的计算。图 (c) 展示了使用 DCPO 的端到端强化学习训练过程。我们使用动态对比选择策略，从总计 K 条轨迹中选取 K/2 条。

用户查询、生成文本和推荐产品这三个核心组成部分。该评估从以下三个互补维度展开：

(1) 产品正确性用于评估推荐产品是否满足用户需求，以及呈现方式是否恰当。此维度包含四项检查：1）产品相关性，验证每件商品是否符合用户请求的范围；2）UI格式验证，确保以恰当的结构进行呈现；3）UI触发评估，确认在适当情况下推荐产品，并在不适当时避免推荐；4）UI 完整性检查文本中提及的所有产品是否都有对应的产品卡片。

(2) 文本相关性 评估文本回复是否回应了用户查询，检查智能体是否存在离题性，即讨论相关但偏离主题的内容，并验证回复是否针对用户需求提供连贯指导。

(3) 描述忠实度用于验证文本中的产品描述是否准确、忠实地反映产品的实际属性，防止出现可能损害用户信任的幻觉特征或误导性描述。

只有当回复通过全部三个维度的检查时，才能通过L1。我们通过混合评估来实现这一点：结合基于规则的验证检查格式和结构，并通过基于大语言模型的评估判断语义正确性。详细的评分标准和说明见附录A.1。

*3.2.2 L2*评分器。 虽然L1确保了基本正确性，但高质量的购物回复还必须具备良好的组织结构和充实的内容。L2评分器通过基于评分标准的评估来衡量这一点，围绕两个互补维度——结构和深度——考察七项指标。

(1) **结构**。结构维度考察响应是否连贯且组织良好，包括响应是否以清晰的问题界定开篇，建立上下文并体现对用户需求的理解；在呈现建议或比较的中间部分保持逻辑一致性；并以可执行的指导收尾，帮助用户做出知情决策。

(2) **深度**。深度维度评估整体内容的充实程度。具体考察回复在讨论多个选项时是否提供有意义的比较分析；在介绍备选方案时是否审慎地进行优先级排序，而非将所有选项一视同仁；是否清晰、有说服力地阐明产品层面的论证，解释特定推荐为何适合用户需求；以及是否意识到用户应考虑的潜在权衡或风险。

每项指标均依据详细的评分标准进行评估，其中通过具体示例界定各个表现等级，从而确保评估一致，同时捕捉合格响应与卓越响应之间细微的质量差异。完整的指标定义和评分标准见附录A.2。

4 方法论

我们介绍了 ChatShopBuddy 方法，旨在通过强化学习训练对话式购物智能体。我们首先介绍智能体框架，该框架将大语言模型转化为能够使用工具的购物智能体；随后介绍 HRM，以构建多维目标；最后介绍 DCPO，以平衡响应质量与运营效率。

4.1 智能体框架

我们的智能体框架为大语言模型配备了外部工具，并通过迭代推理和工具调用来构建其行为。给定用户查询 q，智能体会生成一条由多个

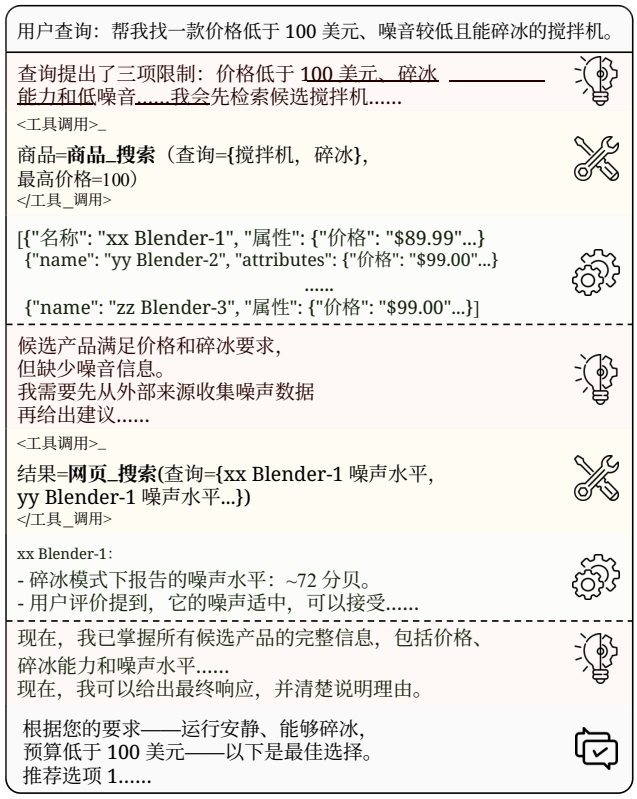


图3：购物智能体的示例轨迹。

推理-行动-观察循环组成的轨迹，最终生成响应。

4.1.1 轨迹结构。 包含 T 个步骤的轨迹可写作 $\tau = (R_1, A_1, O_1, \ldots\ldots, R_T, A_T, O_T, y)$。在每个步骤 t：

- R_t 是智能体的内部推理（思维链文本，用于分析当前情况并规划后续行动）；
- $A_t = \{a_{(1)t}, \ldots\ldots, a_{(m_t)t}\}$ 是智能体发出的工具调用；
- $O_t = \{o_{(1)t}, \ldots\ldots, o_{(m_t)t}\}$ 是工具返回的观察结果。

经过 T 步后，智能体生成最终响应 y，其中包括自然语言解释和结构化产品推荐。智能体遵循策略 π_θ，该策略将当前上下文（查询、推理历史、操作和观察结果）映射为下一步推理和操作。

4.1.2 工具套件。 该智能体可以使用三个工具：

- 网页_搜索：检索网络信息，以回答知识密集型问题；
- Python_执行：执行 Python 代码进行计算（例如，计算组合价格、比较单位成本）；
- 产品_搜索：使用筛选条件查询商品数据库，并返回匹配的产品。

4.1.3 智能体轨迹示例。 考虑查询"帮我找一台100美元以下、能够碎冰的安静搅拌机。"智能体首先对约束条件进行推理（R_1：噪声水平、价格、碎冰），然后使用适当的筛选条件调用产品_搜索（A_1），并接收候选产品（O_1）。注意到缺少噪声规格后，它再次进行推理（R_2），并调用网页_搜索以查找噪声水平（A_2），从而获得补充数据（O_2）。在信息完整后，它进行最终推理（R_3），并生成包含带有理由的推荐和产品卡片的响应 y。图3展示了这一过程。

我们的学习目标是优化策略 π_θ，以最大化期望轨迹奖励：

$$\max_\theta \ \mathbb{E}_{\tau\sim\pi_\theta}\big[r(\tau)\big]. \tag{1}$$

4.2 HRM：分层奖励建模

购物智能体必须满足多个具有自然依赖关系的目标：基本正确性是实现更高层次质量的前提，而只有当响应质量令人满意时，高效的工具使用才有意义。我们通过 HRM 组织奖励，以体现这些逻辑依赖关系。总奖励分解为评估最终响应的结果奖励 $r_{\text{out}}(\tau)$，以及评估工具使用情况的过程奖励$r_{\text{proc}}(\tau)$：

$$r(\tau) = r_{\text{out}}(\tau) + \beta \cdot r_{\text{proc}}(\tau), \tag{2}$$

其中，β控制过程奖励的贡献。两个组成部分均采用分层方式计算。

4.2.1 分层结果奖励。结果奖励基于我们的分层评估框架（第 3.2 节）。L1评分器通过二元指标$c_i(\tau) \in \{0,1\}$评估三个基本维度（产品正确性、文本相关性、描述忠实度）：$G_{\text{L1}}(\tau) = \prod_{i=1}^{3} c_i(\tau)$，仅当所有维度均通过时才等于 1。我们将 G_{L1} 视为硬性可行性门槛：对于 $G_{\text{L1}}(\tau) = 0$ 的响应，奖励为零。对于可行的响应，L2评分器会给出精细的质量分数 $G_{\text{L2}}(\tau) \in [0,1]$，用于衡量结构连贯性和内容深度。最终结果奖励计算如下：

$$r_{\text{out}}(\tau) = \begin{cases} 1 + \alpha \cdot (G_{\text{L2}}(\tau))^k & \text{if } G_{\text{L1}}(\tau) = 1, \\ 0 & \text{if } G_{\text{L1}}(\tau) = 0, \end{cases} \tag{3}$$

其中，α 和 k 是超参数。加法常数创造了一种隐式课程，优先考虑可行性，然后再考虑质量；而五次变换（G_{L2}）^{k} 则强化了末端奖励，避免策略停留在可行性边界。

4.2.2 层级过程奖励。过程奖励用于评估工具使用准确性和效率，仅针对通过 L1 且质量超过 L2 阈值的轨迹进行条件计算 η：

$$r_{\text{proc}}(\tau) = \begin{cases} \text{Score}(\mathcal{A}, O) & \text{if } G_{\text{L1}}(\tau) = 1 \text{ and } G_{\text{L2}}(\tau) \geq \eta, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

其中，Score(A,O)由基于大语言模型的评判器计算，该评判器会评估工具调用是否使用了正确的参数，以及观察结果是否有效地促成了最终响应。此门控机制确保智能体优先关注响应质量，然后才会因工具效率而获得奖励。图2(b)展示了这一层级结构。

表2：我们在 SmartShopBench 上的主要结果。对于 L1评分器，我们报告产品正确性（PC）、文本相关性（TR）、描述忠实度（DF）、Avg@4 和 Pass^4。PC、TR 和 DF 是四次运行的平均值。Avg@4 是四次运行的平均成功率，只有三项指标全部通过才算该次运行成功；Pass^4 则是四次运行中所有响应均通过全部三项指标的比例。对于 L2评分器，我们报告四次运行的均值和标准差。# Tokens 表示推理令牌数。对于开源模型，最佳结果以粗体标出，次佳结果加下划线。闭源模型的结果以灰色显示，供参考。

模型	# 词元	L1评分器					L2评分器	
		PC	TR	DF	Avg@4	通过^4	Avg@4	Std@4
非思考模型								
DeepSeek-V3.2-chat	-	69.80	98.75	66.90	43.33	7.50	0.4550	**0.0058**
GLM-4.7（关闭思考）	-	61.67	99.60	67.90	38.98	5.00	0.4825	0.0096
Kimi-K2.5（关闭思考）	-	31.45	72.70	77.90	21.65	5.00	0.2850	0.0520
Qwen3-235B-A22B-Instruct-2507	-	42.30	99.40	65.42	25.20	2.50	0.5775	0.0126
Qwen3-30B-A3B-Instruct-2507	-	14.57	93.97	42.30	3.92	0.00	0.5625	0.0299
GPT-5.2（关闭思考）	-	80.03	99.80	91.88	66.88	35.00	0.6925	0.0150
Gemini-3-Flash-Preview（最简模式）	-	74.80	100.00	71.22	48.75	9.20	0.5600	0.0163
Doubao-Seed-1.8（关闭推理）	-	73.53	98.75	49.38	30.20	1.70	0.3525	0.0222
推理模型								
DeepSeek-V3.2-reasoner	1,412	86.05	**99.80**	79.38	62.10	19.20	0.6200	0.0271
GLM-4.7	764	64.80	99.60	65.40	37.50	2.50	0.4250	0.0100
Kimi-K2.5	515	70.42	98.95	77.72	47.73	12.50	0.4900	0.0115
Qwen3-235B-A22B-Thinking-2507	1,526	36.88	99.40	24.80	8.32	0.00	**0.7015**	0.0150
Qwen3-30B-A3B-Thinking-2507	2,803	47.30	98.55	31.70	12.70	0.00	0.6275	0.0096
GPT-5.2	N/A	88.15	99.80	99.20	75.85	42.50	0.6750	0.0058
Gemini-3-Pro-Preview	N/A	80.40	99.80	70.85	55.03	15.80	0.6725	0.0050
Doubao-seed-1.8	1,056	76.47	99.38	58.75	39.77	3.30	0.2825	0.0189
ChatShopBuddy-SFT	1,200	86.88	98.55	72.28	60.40	18.30	0.4800	0.0606
ChatShopBuddy-SFT-RL	633	**93.35**	99.60	**84.97**	**75.22**	**34.20**	0.6325	0.0096

4.3 动态对比策略优化

真实场景部署要求尽可能降低推理延迟，我们通过推理长度 $L(\tau)$ 来量化这一指标，即内部推理的总词元数。DCPO 通过动态轨迹选择，明确优化质量-效率权衡。

对于每个查询 q，我们采样 K 条候选轨迹，并按词典序对它们进行排序：首先按奖励 $r(\tau)$降序，然后在并列时按推理长度 $L(\tau)$升序。这会优先选择高奖励轨迹，并在质量相当的情况下偏好简洁的推理。基于此，我们将轨迹划分为三个大小相等的池$\mathrm{P_{good}}$、$\mathrm{P_{mid}}$、$\mathrm{P_{bad}}$（顶部、中部、底部）。我们通过以下方式构建一个约含 $K/2$条轨迹的选定子集$\mathrm{S}(q)$：

- **锚点**：选择τ个最佳项（来自Pgood的最高排序）作为正向参考，并选择τ个最差项（来自Pbad的最低排序）作为负向参考。
- **分层抽样**：用从所有池中随机抽取的、按比例具有代表性的样本填充剩余的 $K/2-2$ 个名额。

我们使用奖励归一化计算 $\mathrm{S}(q)$ 上的优势：

$$\hat{A}(\tau) = \frac{r(\tau) - \mu(q)}{\sigma(q) + \delta}, \tag{5}$$

where δ ensures numerical stability, $\mu(q) = \frac{1}{|\mathcal{S}(q)|}\sum_{\tau\in\mathcal{S}(q)} r(\tau)$ and $\sigma(q) = \sqrt{\frac{1}{|\mathcal{S}(q)|}\sum_{\tau\in\mathcal{S}(q)}(r(\tau)-\mu(q))^2}$. The policy optimization 目标为：

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau\sim\mathcal{S}(q)}\left[\min\left(\rho_\theta(\tau)\hat{A}(\tau), \mathrm{clip}(\rho_\theta(\tau), 1-\epsilon, 1+\epsilon)\hat{A}(\tau)\right)\right] - \lambda\,\mathbb{D}_{\mathrm{KL}}(\pi_\theta\|\pi_{\mathrm{ref}}),$$

其中 $\rho_\theta(\tau) = \frac{\pi_\theta(\tau)}{\pi_{\text{旧}}(\tau)}$ 是重要性采样比率，ϵ 是裁剪阈值，λ 控制KL惩罚。图2(c)展示了DCPO训练过程。

5 实验设置

5.1 数据集与评估指标

5.1.1 数据集。 我们使用 SmartShopBench 进行训练和评估。该基准测试包含 1,680 条查询，涵盖六个不同类别（第 3.1.1 节）：模糊搜索、多约束搜索、捆绑搜索、搜索-通用、问答比较和问答咨询。其中，1,560 条查询用于训练，其余 120 条查询留作测试。

5.1.2 评估协议。 我们采用第 3.2 节中的分层评估框架。为考虑抽样方差，每个响应均进行四次独立评估。

对于 L1评估，我们使用基于规则的检查和大语言模型评估三个维度（产品正确性、文本相关性、描述忠实度）。只有三个维度均通过，运行才算成功。我们报告：(1) 四次运行中各维度的平均通过率，

四次运行，(2) **Avg@4**，四次运行的平均成功率，以及 (3) **Pass^4**，四次运行中所有响应均通过的比例。

对于 L2 评估，我们报告四次独立运行中质量分数的均值和标准差，从而对结构连贯性和内容深度进行比二元正确性更细致的评估。

5.2 基线

我们将多种推理能力各异的基线模型纳入全面比较。开源模型包括 DeepSeek-V3.2 [4], GLM-4.7 [20], Kimi-K2.5 [41], 以及 Qwen3 系列（235B-A22B、30B-A3B）的标准版和思考版 [52]。专有模型包括 GPT-5.2 [31], Gemini-3-Flash-Preview [9], Gemini-3-Pro-Preview [10], 以及 Doubao-seed-1.8 [36]。对于具有扩展推理模式的模型（例如 GPT-5.2），在适用情况下，我们分别评估启用和禁用推理的表现，以衡量审慎推理对购物任务的影响。

5.3 实现

我们使用 Qwen3-30B-A3B-Thinking-2507 [52] 作为基础模型。智能体首先在 480 条高质量标注轨迹上进行监督微调，批大小为 8，学习率为 7.0e-6，共 2 个 epoch；随后使用 DCPO 训练 2 个 epoch，批大小为 16，学习率为 1e-6。在强化学习训练期间，我们针对每个查询采样 $K = 16$条轨迹，并通过动态对比选择选出 $K/2 = 8$ 。训练期间的最大序列长度为 32,768 个 token。推理时，我们使用最大上下文长度 81,920 个 token、温度 1.0 以及 top-p 0.9。

在奖励计算方面，L1 和 L2 评分器以及过程奖励均基于 DeepSeek-V3.2-reasoner [4]。我们在 HRM 中设置 $\alpha = 0.5$, $\beta = 0.05$, $\eta = 0.7$, $k = 5$ 。在工具实现方面，网页_搜索使用智谱的 search_pro API，产品_搜索是来自真实工业部署的商业产品搜索 API。更多实现细节见附录 B。

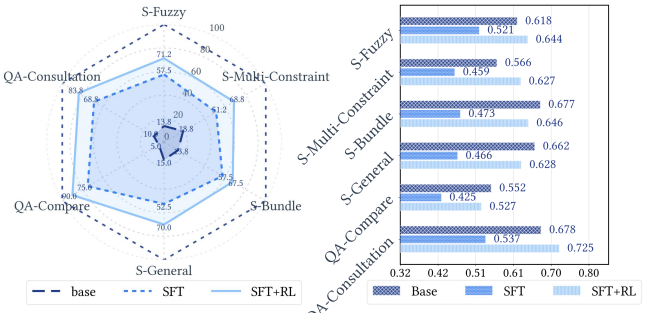
6 实验结果

6.1 主要结果

表2展示了SmartShopBench在L1评分器和L2评分器下的主要结果，从而得出若干清晰的观察结果。

（1）任务对齐强化学习优于模型规模和通用推理。通过显式的任务对齐优化，ChatShopBuddy展现出强劲的整体性能，产品正确性达到93.35，描述忠实

忠实度为 84.97，Avg@4 为 75.22，Pass^4 为 34.20。与规模更大的 DeepSeek-V3.2-reasoner 相比，ShoppingAgent-RL 在几乎所有 L1 指标上均表现更优，产品正确性提高了 +7.30，描述忠实度提高了 +5.59，Avg@4 提高了 +13.12，Pass^4 提高了 +15.00。值得注意的是，在 L2评分器下，其 Avg@4 达到 0.6325，与 DeepSeek-V3.2-reasoner （ 0.6200 ）相当，同时方差低得多（0.0096）。这表明，稳定且高质量的对话式购物表现主要由任务对齐优化驱动，而非模型规模或通用推理能力。



（a）L1评分器在六个类别上 （b）L2评分器在六个类别上

图4：ChatShopBuddy 在六个购物类别中的表现。我们报告四次独立运行的 Avg@4。

（2）强化学习在提升响应质量之外，还能增强稳定性。与监督微调相比，经强化学习优化的 ChatShopBuddy 在任务成功和一致性方面均有显著提升。具体而言，Pass^4——即四次运行中所有响应均通过全部 L1 维度的比例——从 18.30 升至 34.20，表明强化学习显著提高了多次试验中获得完全可靠响应的可能性。同时，L2 平均表现显著提升（从 0.4800 升至 0.6325），四次运行的标准差则从 0.0606 大幅降至 0.0096。这表明，强化学习的主要优势不仅在于提高响应质量的峰值，更在于确保各次运行都能产生一致、可靠的输出；这对于实际部署至关重要。

(3) 扩展推理本身并不总能转化为特定领域任务中的性能提升。 尽管启用思考的模型通常在 SmartShop-Bench 上取得更高的响应质量，但思考带来的益处在 L1 和 L2 评估中并不一致。例如，DeepSeek-V3.2-reasoner（1,412 个推理令牌）通过思考取得了明显提升，Pass^4 提高了 19.20，L2 Avg@4 达到 0.6200。相比之下，GLM-4.7（764 个推理令牌）的 L1 Pass^4 略微下降了 2.50，尽管其 L2 Avg@4 达到相近的 0.4250。同样，Qwen3-235B-A22B-Thinking-2507（1,526 个推理令牌）的 L2 Avg@4 高达 0.7015，但任务性能不稳定，产品正确性为 36.88，L1 Avg@4 为 8.32。这些结果表明，推理并不能在不同评分器下始终提升性能。若缺乏任务对齐优化，更长或更复杂的推理有时会导致过度推理或推理漂移，从而可能损害而非改善任务结果。

6.2 不同购物类别的性能

图4展示了训练过程在六个不同购物意图类别上带来的性能提升。在衡量智能体是否满足用户基本需求的L1评分器下，训练在所有类别上均持续提升性能。监督微调（SFT）相较基础模型取得了显著增益，表明其能有效适应复杂购物任务；而强化学习则进一步提升性能，在问答比较、问答咨询以及多约束搜索意图上的改进尤为突出。

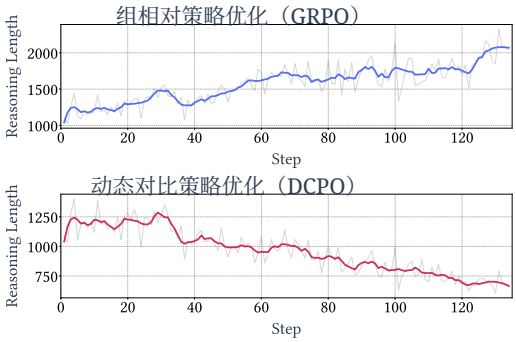
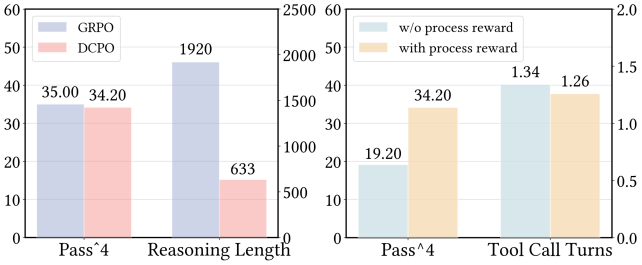


图5：训练过程中 DCPO 与 GRPO 的推理长度对比。



（a）推理长度分析 （b）工具效率分析

图6：运营效率分析。

相比之下，L2评分器评估结构连贯性和内容深度等更高层次的质量。在这一更严格的指标下，仅进行监督微调实际上会导致所有类别的性能下降，这表明尽管它提升了基本任务能力，却并未针对更高层次的响应质量进行优化。然而，强化学习不仅弥补了这一下降，还在每个类别上都超越了基础模型，其中问答咨询和问答比较的提升尤为显著。

总体而言，这些结果揭示了清晰的训练动态：监督微调提升了基础任务成功率，但可能降低高层次响应质量；而强化学习则发挥纠正阶段的作用，使模型同时满足基本用户需求和更高阶的输出标准，从而在所有购物意图上都取得稳健提升。

6.3 运营效率分析

运营效率对于具备推理能力的购物智能体至关重要，因为推理延迟和工具使用效率会影响用户体验。本节将从推理长度和工具使用方面分析效率，并评估能否在不损害任务性能的前提下提升效率。

推理长度效率。 推理长度直接影响在线延迟：更长的推理轨迹会增加推理成本，却不一定能提升响应质量。如图 5 所示，DCPO 在整个训练过程中推理长度明显缩短，而 GRPO 生成的推理序列则逐渐变长，反映出两者在优化

表3：对我们的 ChatShopBuddy 进行消融研究，最佳分数以粗体显示，次佳分数加下划线。

方法	L1评分器		L2评分器	
	Avg@4	通过^4	Avg@4	Std@4
ChatShopBuddy	**75.22**	**34.20**	0.6325	0.0096
不使用 DCPO 训练（监督微调）	60.40	18.30	0.4800	0.0606
不使用层次化奖励	66.45	26.90	**0.6450**	0.0058
不使用过程奖励	66.05	19.20	0.6325	**0.0050**

动态上存在根本差异。训练期间的这种分化也延续到了推理时的行为。如图 6 (a) 所示，使用 DCPO 训练的模型在推理时生成的推理内容更短，同时任务性能与 GRPO 相当。这表明，DCPO 并不依赖越来越冗长的推理来获得高奖励。相反，DCPO 利用动态对比选择策略，鼓励高效且有效的推理模式，减少不必要的计算和推理延迟，同时保持较高的任务成功率。

工具使用效率。 工具调用行为也会影响运行成本，因为过多或冗余的调用会增加执行开销。为考察这一影响，我们引入了一种过程奖励，明确鼓励高效使用工具。有趣的是，引入过程奖励后，任务性能（以 Pass@4 衡量）大幅提升，同时工具调用效率也得到改善。如图 6(b) 所示，模型以更少的工具调用轮数取得了显著更高的 Pass@4 分数，这表明，鼓励更精准地使用工具不仅能提升解答质量，还能减少不必要的计算和运行开销。

6.4 消融研究

我们在表3中进行消融研究，以验证训练框架中各组件的有效性。

（1）DCPO训练的重要性：移除DCPO训练会导致所有指标出现最显著的性能下降。具体而言，L1 Avg@4从75.22降至60.40，Pass^4从34.20降至18.30，同时L2 Avg@4大幅下降且方差增大。这凸显了DCPO训练在使模型与任务导向目标对齐并确保性能稳定方面的核心作用。

（2）层次化奖励的有效性：移除层次化奖励会导致L1性能明显下降，表明可靠性减弱。尽管L2 Avg@4略有上升，但整体性能仍不及完整模型。这表明层次化奖励信号对于生成既可靠又与任务对齐的响应至关重要。

(3) 过程奖励的作用：移除过程奖励会导致 Pass^4 明显下降，反映出多次运行之间的一致性降低。与此同时，L2 Avg@4 仍保持在相近水平，且方差更低，这表明过程级奖励主要提升了决策过程中的多步可靠性和稳定性，而非直接提升更高层次的响应质量。

7 结论

我们推出了 ChatShopBuddy，这是一款针对真实世界对话式购物任务、通过强化学习训练的任务对齐型智能体，能够有效应对在确保效率的同时优化多个难以间接验证的目标这一挑战。我们推出了 SmartShopBench，用于分层评估；提出了分层奖励建模（HRM），通过反映各目标逻辑依赖关系的条件门控机制，整合正确性、说服力和效率；并开发了动态对比策略优化（DCPO），以动态选择简洁、高质量的推理轨迹，兼顾响应质量与运营效率。大量实验表明，ChatShopBuddy 始终优于依赖通用推理的更大模型，在稳定性、效率和整体用户满意度方面表现更佳。我们的工作为设计可靠、有说服力且高效、可用于实际部署的对话式购物智能体提供了实用指导。

参考文献

[1] Yifei Chen、Guanting Dong 和 Zhicheng Dou。2025年。通过自演化偏好学习实现有效的工具集成推理。*CoRR* abs/2509.23285（ 2025 2509 23285 10 48550 2509 23285 ）。arXiv：。doi：。/ARXIV。。[2] Preetam Prabhu Srikar Dammu、Omar Alonso 和 Barbara Poblete。2025年。满足主观产品需求的购物智能体。载于第十八届 *ACM* 网络搜索与数据挖掘国际会议论文集，*WSDM 2025*，德国汉诺威，*2025*年*3*月*10*日至*14*日，Wolfgang Nejdl、Sören Auer、Meeyoung Cha、Marie-Francine Moens 和 Marc Najork（编）。ACM，1032–1035。doi:10.1145/3701551.3704124[3] DeepSeek-AI。2025年。DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*CoRR* abs/2501.12948（2025年）。arXiv：2501.12948 doi:10.48550/ARXIV.2501.12948[4] DeepSeek-AI。2025年。DeepSeek-V3.2：推动开放式大语言模型的前沿。*CoRR* abs/2512.02556（2025年）。arXiv：2512.02556 doi:10.48550/ARXIV.2512. 02556[5] Guanting Dong、Yifei Chen、Xiaoxi Li、Jiajie Jin、Hongjin Qian、Yutao Zhu、Hangyu Mao、Guorui Zhou、Zhicheng Dou 和 Ji-Rong Wen。2025年。Tool-Star：通过强化学习赋能基于大语言模型的多工具推理器。*CoRR* abs/2505 16410 2025 2505 16410 10 48550 2505 16410。（）。arXiv：。[6] 2025 doi：。/ARXIV。。Hanwen Du、Bo Peng 和 Xia Ning。。SAPIENT：通过战略规划和蒙特卡洛树搜索掌握多轮对话式推荐。载于*2025*年计算语言学协会美洲分会会议：人类语言技术，*NAACL 2025*——第*1*卷：长篇论文，美国新墨西哥州阿尔伯克基，*2025*年*4*月*29*日至*5*月*4*日，Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，2629–2648。doi:10.18653/V1/2025.NAACL-LONG.133[7] Jiaxuan Gao、Wei Fu、Minyang Xie、Shusheng Xu、Chuyi He、Zhiyu Mei、Banghua Zhu 和 Yi Wu。2025年。超越十轮：通过大规模异步强化学习释放长程智能体搜索能力。*CoRR* abs/2508.07976（2025年）。arXiv：2508.07976 doi:10.48550/ARXIV.2508.07976[8] Google Cloud。2025年。对话式商务智能体开发者指南。https: // cloud.google.com/retail/docs/conversational-commerce-dev-guide。[9] Google Cloud。2025年。Gemini 3 Flash Preview。https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-flash。[10] Google Cloud。2025年。Gemini 3 Pro Preview Model。https://docs.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro。[11] Zhibin Gou、Zhihong Shao、Yeyun Gong、Yelong Shen、Yujiu Yang、Minlie Huang、Nan Duan 和 Weizhu Chen。2024年。ToRA：用于数学问题求解的工具集成推理智能体。载于第十二届国际学习表示会议，*ICLR 2024*，奥地利维也纳，*2024*年*5*月*7*日至*11*日。OpenReview.net。https://openreview.net/forum?id=Ep0TtjVoap[12] Anisha Gunjal、Anthony Wang、Elaine Lau、Vaskar Nath、Bing Liu 和 Sean Hendryx。2025年。将评分标准作为奖励：超越可验证领域的强化学习。*CoRR* abs/2507.17746（2025年）。arXiv：2507.17746 doi:10.48550/ARXIV。2507.17746[13] Yuxuan Huang、Yihang Chen、Haozheng Zhang、Kang Li、Meng Fang、Linyi Yang、Xiaoguang Li、Lifeng Shang、Songcen Xu、Jianye Hao、Kun Shao 和 Jun Wang。2025年。深度研究智能体：系统性考察与路线图。*CoRR* abs/2506.18096（2025年）。arXiv：2506.18096 doi:10.48550/ARXIV.2506.18096[14] Zenan Huang、Yihong Zhuang、Guoshan Lu、Zeyu Qin、Haokai Xu、Tianyu Zhao、Ru Peng、Jiaqi Hu、Zhanming Shen、Xiaomeng Hu、Xijun Gu、Peiyi Tu、Jiaxin Liu、Wenyu Chen、Yuzhuo Fu、Zhiting Fan、Yanmei Gu、Yuanyuan Wang、Zhengkai Yang、Jianguo Li 和 Junbo Zhao。2025年。基于评分标准的强化学习锚点。*CoRR* abs/2508.12790（2025）。arXiv:2508.12790 doi:10.48550/ARXIV.2508.12790[15] Dietmar Jannach、Ahtsham Manzoor、Wanling Cai 和 Li Chen。2021。对话式推荐系统综述。*ACM Computing Surveys (CSUR)* 54，5（2021），1–36。[16] Yongnan Jin、Xurui Li、Feng Cao、Liucun Gao 和 Juanjuan Yao。2025。基于几何投影参考约束的多维度评分标准导向奖励模型学习。*CoRR* abs/2511.16139（2025）。arXiv:2511.16139 doi:10.48550/ARXIV.2511.16139[17] Noriaki Kawamae。2025。单次多视图视觉对话式推荐。载于第*31*届*ACM SIGKDD*知识发现与数据挖掘会议论文集，第*2*卷，*KDD 2025*，加拿大安大略省多伦多，*2025*年*8*月*3*日至*7*日，Luiza Antonie、Jian Pei、Xiaohui Yu、Flavio Chierichetti、Hady W. Lauw、Yizhou Sun 和 Srinivasan Parthasarathy（编）。ACM， 1118 1127 10 1145 3711896 3737070 – 。[18] doi: . / . Taeho Kim、Hyeongjun Jang、Juwon Yu、Taeuk Kim、Hyunyoung Lee、Ji-Hui Im 和 Sang-Wook Kim。2025。ESPRESSO：面向高质量对话式推荐系统的有效段落检索方法。载于*AAAI-25*，由人工智能促进协会赞助，*2025*年*2*月*25*日至*3*月*4*日，美国宾夕法尼亚州费城，Toby Walsh、Julie Shah 和 Zico Kolter（编）。AAAI出版社， 11942 11949 10 1609 39 11 33300 – 。doi: . /AAAI.V I 。[19] Heejin Kook、Junyoung Kim、Seongmin Park 和 Jongwuk Lee。2025。通过对比用户偏好增强基于检索的对话式推荐。载于*2025*年美洲计算语言学协会分会会议：人类语言技术论文集，*NAACL 2025*——第*1*卷：长篇论文，美国新墨西哥州阿尔伯克基，*2025*年*4*月*29*日至*5*月*4*日，Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会， 7692 7707 10 18653 1 2025 392 – 。doi: . /V / .NAACL-LONG。Z Lab。. GLM- 。：提升编码能力。[20] 2025 4 7 https://z.ai/blog/glm- 4.7。[21] Chuang Li、Yang Deng、Hengchang Hu、Min-Yen Kan 和 Haizhou Li。2025。ChatCRS：将外部知识和目标引导融入基于大语言模型的对话式推荐系统。载于计算语言学协会论文集：*NAACL 2025*，美国新墨西哥州阿尔伯克基，*2025*年*4*月*29*日至*5*月*4*日，Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，295–312。doi:10.18653/V1/2025.FINDINGS-NAACL.17 [22] Kuan Li、Zhongwang Zhang、Huifeng Yin、Liwen Zhang、Litu Ou、Jialong Wu、Wenbiao Yin、Baixuan Li、Zhengwei Tao、Xinyu Wang、Weizhou Shen、Junkai Zhang、Dingchu Zhang、Xixi Wu、Yong Jiang、Ming Yan、Pengjun Xie、Fei Huang 和 Jingren Zhou。2025。WebSailor：驾驭网页智能体的超人类推理能力。*CoRR* abs/2507.02592（2025）。arXiv:2507.02592 doi:10.48550/ARXIV.2507. 02592[23] Xiaoxi Li、Guanting Dong、Jiajie Jin、Yuyao Zhang、Yujia Zhou、Yutao Zhu、Peitian Zhang 和 Zhicheng Dou。2025。Search-o1：智能体式搜索增强型大型推理模型。*CoRR* abs/2501.05366（2025）。arXiv:2501.05366 doi:10.48550/ ARXIV.2501.05366[24] Xiaoxi Li、Jiajie Jin、Guanting Dong、Hongjin Qian、Yutao Zhu、Yongkang Wu、Ji-Rong Wen 和 Zhicheng Dou。2025。WebThinker：赋予大型推理模型深度研究能力。*CoRR* abs/2504.21776（2025）。arXiv:2504.21776 doi:10.48550/ARXIV.2504.21776[25] Xuefeng Li、Haoyang Zou 和 Pengfei Liu。2025。ToRL：扩展工具集成强化学习。*CoRR* abs/2503.23383（2025）。arXiv:2503.23383 doi:10.48550/ARXIV.2503.23383[26] Jiate Liu、Yiqin Zhu、Kaiwen Xiao、Qiang Fu、Xiao Han、Wei Yang 和 Deheng Ye。2023。RLTF：基于单元测试反馈的强化学习。*Trans. Mach. Learn. Res.* 2023（2023）。https://openreview.net/forum?id=hjYmsV6nXZ [27] Tianci Liu、Ran Xu、Tony Yu、Ilgee Hong、Carl Yang、Tuo Zhao 和 Haoyu Wang。2025。OpenRubrics：面向奖励建模和大语言模型对齐的可扩展合成评分标准生成。*CoRR* abs/2510.07743（2025）。arXiv:2510.07743 doi:10.48550/ARXIV.2510.07743[28] Yuanxing Liu、Weinan Zhang、Yifan Chen、Yuchi Zhang、Haopeng Bai、Fan Feng、Hengbin Cui、Yongbin Li 和 Wanxiang Che。2023。对话式推荐系统与大语言模型在电子商务售前对话中相得益彰。载于计算语言学协会论文集：*EMNLP 2023*，新加坡，*2023*年*12*月*6*日至*10*日，Houda Bouamor、Juan Pino 和 Kalika Bali（编）。计算语言学协会，9587–9605。doi:10. 18653/V1/2023.FINDINGS-EMNLP.643[29] Junyu Luo、Weizhi Zhang、Ye Yuan、Yusheng Zhao、Junwei Yang、Yiyang Gu、Bohan Wu、Binqi Chen、Ziyue Qiao、Qingqing Long、Rongcheng Tu、Xiao Luo、Wei Ju、Zhiping Xiao、Yifan Wang、Meng Xiao、Chenwu Liu、Jingyang Yuan、Shichang Zhang、Yiqiao Jin、Fan Zhang、Xian Wu、Hanqing Zhao、Dacheng Tao、Philip S. Yu 和 Ming Zhang。2025。大语言模型智能体：方法论、应用与挑战综述。*CoRR* abs/2503.21460（2025）。arXiv:2503.21460 doi:10.48550/ARXIV.2503.21460[30] Yougang Lyu、Xiaoyu Zhang、Lingyong Yan、Maarten de Rijke、Zhaochun Ren 和 Xiuying Chen。2025。DeepShop：深度研究购物智能体基准测试。*CoRR* abs/2506.02839（2025）。arXiv:2506.02839 doi:10.48550/ARXIV.2506. 02839[31] OpenAI。2025。推出 GPT-5.2。https://openai.com/index/introducing-gpt- 5-2/。

[32] OpenAI。2025。在 ChatGPT 中推出购物研究。https://openai.com/index/chatgpt-shopping-research/。[33] Qiyao Peng、Hongtao Liu、Hua Huang、Qing Yang 和 Minglai Shao。2025。大语言模型驱动的推荐系统智能体综述。*CoRR* abs/2502.10050（ 2025 2502 10050 10 48550 2502 10050）。arXiv: 。doi: 。/ARXIV. 。[34] Mathieu Ravaut、Hao Zhang、Lu Xu、Aixin Sun 和 Yong Liu。2024。作为语言处理任务的参数高效对话式推荐系统。载于第*18*届计算语言学协会欧洲分会会议论文集，*EACL 2024*——第*1*卷：长篇论文，马耳他圣朱利安，*2024*年*3*月*17*日至*22*日，Yvette Graham 和 Matthew Purver（编）。计算语言学协会， 152 165 2024 9 – 。[35] https://aclanthology.org/ .eacl-long。John Schulman、Filip Wolski、Prafulla Dhariwal、Alec Radford 和 Oleg Klimov。2017。近端策略优化算法。*CoRR* abs/1707.06347（2017）。arXiv: 1707 06347 1707 06347 。http://arxiv.org/abs/ 。字节跳动 Seed。[36] . Seed . 模型卡：迈向通用化现实世界智能体能力。PDF。2025 1 8 https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/research/Seed-1.8-Modelcard.pdf[37] Maohao Shen、Guangtao Zeng、Zhenting Qi、Zhang-Wei Hong、Zhenfang Chen、Wei Lu、Gregory W. Wornell、Subhro Das、David Daniel Cox 和 Chuang Gan。2025。Satori：行动链思维强化学习通过自回归搜索增强大语言模型推理能力。载于第四十二届国际机器学习会议，*ICML 2025*，加拿大不列颠哥伦比亚省温哥华，*2025*年*7*月*13*日至*19*日。OpenReview.net。https://openreview.net/forum?id=j4FXxMiDjL[38] Guangming Sheng、Chi Zhang、Zilingfeng Ye、Xibin Wu、Wang Zhang、Ru Zhang、Yanghua Peng、Haibin Lin 和 Chuan Wu。2025。HybridFlow：灵活高效的基于人类反馈的强化学习框架。载于第*20*届欧洲计算机系统会议论文集，*EuroSys 2025*，荷兰鹿特丹，*2025*年*3*月*30*日—— *3* *2025* 1279 1297 10 1145 3689031 3696075 *4*月。ACM，– 。doi: 。/ 。Joykirat Singh、Raghav Magazine、Yash Pandya 和 Akshay Nambi。2025。通过强化学习实现大语言模型的智能体推理与工具集成。*CoRR* abs/2505 01441 2025 2505 01441 10 48550 2505 01441 。（）。arXiv: 。doi: 。/[39] ARXIV. 。Manato Tajiri 和 Michimasa Inaba。。通过直接偏好优化改进逼真的对话式推荐文本生成。*CoRR* abs/2025 2508 19918 2025 2508 19918 10 48550 2508 19918 。（）。[40] arXiv: 。doi: 。/ARXIV. 。Kimi 团队、Tongtong Bai、Yifan Bai、Yiping Bao、S. H. Cai、Yuan Cao、Y. Charles、H. S. Che、Cheng Chen、Guanduo Chen、Huarong Chen、Jia Chen、Jiahao Chen、Jianlong Chen、Jun Chen、Kefan Chen、Liang Chen、Ruijue Chen、Xinhao Chen、Yanru Chen、Yanxu Chen、Yicun Chen、Yimin Chen、Yingjiang Chen、Yuankun Chen、Yujie Chen、Yutian Chen、Zhirong Chen、Ziwei Chen、Dazhi Cheng、Ming- han Chu、Jialei Cui、Jiaqi Deng、Muxi Diao、Hao Ding、Mengfan Dong、Mengnan Dong、Yuxin Dong、Yuhao Dong、Angang Du、Chenzhuang Du、Dikang Du、Lingx- iao Du、Yulun Du、Yu Fan、Shengjun Fang、Qiulin Feng、Yichen Feng、Garimugai Fu、Kelin Fu、Hongcheng Gao、Tong Gao、Yuyao Ge、Shangyi Geng、Chengyang Gong、Xiaochen Gong、Zhuoma Gongque、Qizheng Gu、Xinran Gu、Yicheng Gu、Longyu Guan、Yuanying Guo、Xiaoru Hao、Weiran He、Wenyang He、Yunjia He、Chao Hong、Hao Hu、Jiaxi Hu、Yangyang Hu、Zhenxing Hu、Ke Huang、Ruiyuan Huang、Weixiao Huang、Zhiqi Huang、Tao Jiang、Zhejun Jiang、Xinyi Jin、Yu Jing、Guokun Lai、Aidi Li、C. Li、Cheng Li、Fang Li、Guanghe Li、Guanyu Li、Haitao Li、Haoyang Li、Jia Li、Jingwei Li、Junxiong Li、Lincan Li、Mo Li、Wei- hong Li、Wentao Li、Xinhang Li、Xinhao Li、Yang Li、Yanhao Li、Yiwei Li、Yuxiao Li、Zhaowei Li、Zheming Li、Weilong Liao、Jiawei Lin、Xiaohan Lin、Zhishan Lin、Zichao Lin、Cheng Liu、Chenyu Liu、Hongzhang Liu、Liang Liu、Shaowei Liu、Shudong Liu、Shuran Liu、Tianwei Liu、Tianyu Liu、Weizhou Liu、Xiangyan Liu、Yangyang Liu、Yanming Liu、Yibo Liu、Yuanxin Liu、Yue Liu、Zhengying Liu、Zhongnuo Liu、Enzhe Lu、Haoyu Lu、Zhiyuan Lu、Junyu Luo、Tongxu Luo、Yashuo Luo、Long Ma、Yingwei Ma、Shaoguang Mao、Yuan Mei、Xin Men、Fan- qing Meng、Zhiyong Meng、Yibo Miao、Minqing Ni、Kun Ouyang、Siyuan Pan、Bo Pang、Yuchao Qian、Ruoyu Qin、Zeyu Qin、Jiezhong Qiu、Bowen Qu、Zeyu Shang、Youbo Shao、Tianxiao Shen、Zhennan Shen、Juanfeng Shi、Lidong Shi、Shengyuan Shi、Feifan Song、Pengwei Song、Tianhui Song、Xiaoxi Song、Hongjin Su、Jianlin Su、Zhaochen Su、Lin Sui、Jinsong Sun、Junyao Sun、Tongyu Sun、Flood Sung、Yunpeng Tai、Chuning Tang、Heyi Tang、Xiaojuan Tang、Zhengyang Tang、Jiawen Tao、Shiyuan Teng、Chaoran Tian、Pengfei Tian、Ao Wang、Bowen Wang、Chensi Wang、Chuang Wang、Congcong Wang、Dingkun Wang、Dinglu Wang、Dongliang Wang、Feng Wang、Hailong Wang、Hai ming Wang、Hengzhi Wang、Huaqing Wang、Hui Wang、Jiahao Wang、Jinhong Wang、Jiuzheng Wang、Kaixin Wang、Linian Wang、Qibin Wang、Shengjie Wang、Shuyi Wang、Si Wang、Wei Wang、Xiaochen Wang、Xinyuan Wang、Yao Wang、Yejie Wang、Yipu Wang、Yiqin Wang、Yucheng Wang、Yuzhi Wang、Zhaoji Wang、Zhaowei Wang、Zhengtao Wang、Zhexu Wang、Zihan Wang、Zizhe Wang、Chu Wei、Ming Wei、Chuan Wen、Zichen Wen、Chengjie Wu、Haoning Wu、Junyan Wu、Rucong Wu、Wenhao Wu、Yuefeng Wu、Yuhao Wu、Yuxin Wu、Zijian Wu、Chenjun Xiao、Jin Xie、Xiaotong Xie、Yuchong Xie、Yifei Xin、Bowei Xing、Boyu Xu、Jianfan Xu、Jing Xu、Jinjing Xu、L. H. Xu、Lin Xu、Suting Xu、Weixin Xu、Xinbo Xu、Xinran Xu、Yangchuan Xu、Yichang Xu、Yuemeng Xu、Zelai Xu、Ziyao Xu、Junjie Yan、Yuzi Yan、Guangyao Yang、Hao Yang、Junwei Yang、Kai Yang、Ningyuan Yang、Ruihan Yang、Xiaofei Yang、Xinlong Yang、Ying Yang、Yi Yang、Yi Yang、Zhen Yang、Zhilin Yang、Zong- han Yang、Haotian Yao、Dan Ye、Wenjie Ye、Zhuorui Ye、Bohong Yin、Chengzhen Yu、Longhui Yu、Tao Yu、Tianxiang Yu、Enming Yuan、Mengjie Yuan、Xiaokun Yuan、Yang Yue、Weihao Zeng、Dunyuan Zha、Haobing Zhan、Dehao Zhang、Hao Zhang、Jin Zhang、Puqi Zhang、Qiao Zhang、Rui Zhang、Xiaobin Zhang、Y. Zhang、Yadong Zhang、Yangkun Zhang、Yichi Zhang、Yizhi Zhang、Yongting Zhang、Yu Zhang、Yushun Zhang、Yutao Zhang、Yutong Zhang、Zheng Zhang、Chenguang Zhao、Feifan Zhao、Jinxiang Zhao、Shuai Zhao、Xiangyu Zhao、Yikai Zhao、Zijia Zhao、Huabin Zheng、Ruihan Zheng、Shaojie Zheng、Tengyang Zheng、Junfeng Zhong、Longguang Zhong、Weiming Zhong、M. Zhou、Runjie Zhou、Xinyu Zhou、Zaida Zhou、Jinguo Zhu、Liya Zhu、Xinhao Zhu、Yuxuan Zhu、Zhen Zhu、Jingze Zhuang、Weiyu Zhuang、Ying Zou和Xinxing Zu。2026。Kimi K2.5：视觉智能体智能。arXiv:2602.02276 [cs.CL] https://arxiv.org/abs/2602.02276[42] Huaixiao Tou、Ying Zeng、Cong Ma、Muzhi Li、Minghao Li、Weijie Yuan、He Zhang和Kai Jia。2025年。ShoppingComp：大语言模型真的准备好进入你的购物车了吗？*CoRR* abs/2511.22978（2025年）。arXiv:2511.22978 doi:10.48550/ARXIV.2511.22978[43] Jiangyuan Wang、Kejun Xiao、Qi Sun、Huaipeng Zhao、Tao Luo、Jiandong Zhang和Xiaoyi Zeng。2025年。ShoppingBench：面向基于大语言模型的智能体的真实世界意图驱动的购物基准。*CoRR* abs/2508.04266（2025年）。arXiv:2508.04266 doi:10.48550/ARXIV.2508.04266[44] Junqiao Wang、Zeng Zhang、Yangfan He、Yuyang Song、Tianyu Shi、Yuchen Li、Hengyuan Xu、Kunyu Wu、Guangwu Qian、Qiuwu Chen和Lewei He。2024年。通过强化学习增强代码大语言模型的代码生成：综述。*CoRR* abs/2412.20367（2024年）。arXiv:2412.20367 doi:10.48550/ARXIV.2412.20367[45] Lei Wang、Chen Ma、Xueyang Feng、Zeyu Zhang、Hao Yang、Jingsen Zhang、Zhiyuan Chen、Jiakai Tang、Xu Chen、Yankai Lin等。2024年。基于大语言模型的自主智能体综述。计算机科学前沿18, 6（2024年），186345。[46] Pei Wang、Yanan Wu、Xiaoshuai Song、Weixun Wang、Gengru Chen、Zhongwen Li、Kezhong Yan、Ken Deng、Qi Liu、Shuaibing Zhao、Shaopan Xiong、Xuepeng Liu、Xuefeng Chen、Wanxi Deng、Wenbo Su和Bo Zheng。2026。ShopSimulator：评估与探索面向购物助手的强化学习驱动的大语言模型智能体。arXiv: 2601 18225 [ ] 2601 18225 . cs.AI https://arxiv.org/abs/ 。Pengkai Wang、Qi Zuo、Pengwei Liu、Zhijie Sang、Congkai Xie和Hongxia Yang。2025年。InfiMed-ORBIT：通过基于评分标准的增量训练使大语言模型对齐开放式复杂任务。*CoRR* abs/2510.15859（2025年）。arXiv:2510.15859 doi:10.48550/ARXIV.2510.15859[48] Sijie Wang、Quanjiang Guo、Kai Zhao、Yawei Zhang、Xin Li、Xiang Li、Siqi Li、Rui She、Shangshu Yu和Wee Peng Tay。2025年。CodeBoost：利用强化学习从代码片段中提炼知识以增强代码大语言模型。*CoRR* abs/2508.05242（2025年）。arXiv:2508.05242 doi:10.48550/ARXIV.2508.05242[49] Yinjie Wang、Ling Yang、Ye Tian、Ke Shen和Mengdi Wang。2025年。通过强化学习协同演化大语言模型代码生成器与单元测试器。*CoRR* abs/2506.03136（2025年）。arXiv:2506.03136 doi:10.48550/ARXIV.2506.03136[50] Yiping Wang、Qing Yang、Zhiyuan Zeng、Liliang Ren、Lucas Liu、Baolin Peng、Hao Cheng、Xuehai He、Kuan Wang、Jianfeng Gao、Weizhu Chen、Shuohang Wang、Simon Shaolei Du和Yelong Shen。2025年。用单个训练样本实现大语言模型推理的强化学习。*CoRR* abs/2504.20571（ 2025 2504 20571 10 48550 2504 20571 ）。arXiv: 。doi: 。/ARXIV. 。Haozhe Xu、Xiaohua Wang、Changze Lv和Xiaoqing Zheng。2025年。超越单一标签：通过大语言模型驱动的数据增强改进对话式推荐。载于计算语言学协会第*63*届年会论文集（第*1*卷：长文），*ACL 2025*，奥地利维也纳，*2025*年*7*月*27*日至*8*月*1*日，Wanxiang Che、Joyce Nabende、Ekaterina Shutova和Mohammad Taher Pilehvar（编）。计算语言学协会，15573–15590。https://aclanthology.org/2025.acl-long.758/[52] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jian Yang、Jiaxi Yang、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、Shuang Luo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou和Zihan Qiu。2025年。Qwen3技术报告。*CoRR* abs/2505 09388 2025 2505 09388 10 48550 2505 09388 。（）。arXiv: 。doi: 。/[53] ARXIV. 。Shunyu Yao、Noah Shinn、Pedram Razavi和Karthik Narasimhan。2024年。 *τ*-bench：面向真实世界领域中工具—智能体—用户交互的基准测试。*CoRR* abs/2406 12045 2024 2406 12045 10 48550 2406 12045 。（）。arXiv: 。doi: 。/ARXIV. 。Li S. Yifei、Allen Chang、Chaitanya Malaviya和Mark Yatskar。。ResearchQA：利用调查挖掘的问题与评分标准，大规模评估跨越75个领域的学术问答。*CoRR* abs/2509.00496（2025年）。arXiv:2509.00496 doi:10.48550/ARXIV.2509.00496[55] Wei Yuan、Shutong Qiao、Tong Chen、Quoc Viet Hung Nguyen、Zi Huang和Hongzhi Yin。2026。将以视觉为中心的文本理解融入对话式推荐系统。arXiv:2601.13505 [cs.IR] https://arxiv.org/abs/ 2601.13505

[56] Huaye Zeng、Dongfu Jiang、Haozhe Wang、Ping Nie、Xiaotong Chen和Wenhu Chen。2025年。ACECODER：通过自动化测试用例合成精通代码生成强化学习。载于计算语言学协会第*63*届年会论文集（第*1*卷：长文），*ACL 2025*，奥地利维也纳，*2025*年7月*27*日至*8*月*1*日，Wanxiang Che、Joyce Nabende、Ekaterina Shutova和Mohammad Taher Pilehvar（编）。计算语言学协会，12023–12040。https://aclanthology.org/2025.acl-long.587/[57] Yongsen Zheng、Mingjie Qian、Guohua Wang、Yang Liu、Ziliang Chen、Mingzhi Mao、Liang Lin和Kwok-Yan Lam。2025年。HyperCRS：超图感知的多粒度偏好学习，以打破对话式推荐系统中的过滤气泡。载于计算语言学协会论文集，*ACL 2025*，奥地利维也纳，*2025*年7月*27*日至*8*月*1*日，Wanxiang Che、Joyce Nabende、Ekaterina Shutova和Mohammad Taher Pilehvar（编）。计算语言学协会，5597–5608。https://aclanthology.org/2025.findings-acl.292/[58] Yaowei Zheng、Richong Zhang、Junhao Zhang、Yanhan Ye、Zheyan Luo、Zhangchi Feng和Yongqiang Ma。2024年。LlamaFactory：100+ 语言模型的统一高效微调。载于第*62*届年会论文集

计算语言学协会（第*3*卷：系统演示）。计算语言学协会，泰国曼谷。http://arxiv.org/ abs/2403.13372[59] Yang Zhou、Sunzhu Li、Shunyu Liu、Wenkai Fang、Jiale Zhao、Jingwen Yang、Jianwei Lv、Kongcheng Zhang、Yihe Zhou、Hengtong Lu、Wei Chen、Yan Xie 和 Mingli Song。2025年。《突破探索瓶颈：用于通用大语言模型推理的评分标准支架式强化学习》。*CoRR* abs/2508.16949 (2025年)。arXiv:2508.16949 doi:10.48550/ARXIV.2508.16949[60] Speed Zhu、Jianwei Cai、Guang Chen、Lulu Wu、Saiyong Yang 和 Wiggin Zhou。2025年。《DRIVE：竞争性代码生成中采用可验证奖励的强化学习数据整理最佳实践》。*CoRR* abs/2511.06307 (2025年)。arXiv:2511.06307 doi:10.48550/ARXIV.2511.06307[61] Yuxin Zuo、Kaiyan Zhang、Shang Qu、Li Sheng、Xuekai Zhu、Biqing Qi、Youbang Sun、Ganqu Cui、Ning Ding 和 Bowen Zhou。2025年。《TTRL：测试时强化学习》。*CoRR* abs/2504.16084 (2025年)。arXiv:2504.16084 doi:10.48550/ARXIV.2504.16084

附录

分层评估框架

A.1 L1评分器

一级（L1）评估器从三个维度评估响应：产品正确性、文本相关性和描述忠实度。提示词见图8。

产品正确性确保推荐的产品不仅满足用户需求，而且在预期的 UI格式中准确、一致地呈现。该维度包含四项互补的检查，共同验证产品推荐的相关性、正确展示、适当触发和完整性，如图7所示：

(1) UI触发：卡片必须在适当的时间显示。(2) UI格式：卡片必须遵循规定的结构。(3) UI相关性：产品必须满足用户的核心需求。(4) UI 完整性：提及的产品必须附有相应的说明。

描述忠实度 用于验证产品的文字描述是否准确反映其实际属性。大语言模型评判器会检测幻觉特征或误导性陈述，以免损害用户信任。

文本相关性 用于衡量文本响应是否直接回应用户查询。大语言模型评判器会识别别任何偏题或无关内容，并核实响应是否针对用户请求提供连贯指导。

A.2 L2评分器

L2评分器评估响应的结构连贯性和内容深度。评估通过基于评分标准的评估进行，涵盖分为两个核心维度的七项指标：

结构： 评估逻辑框架，并判断响应是否遵循系统化的决策过程。

(1) 核心决策轴：明确指出驱动决策的关键因素，并解释其因果作用。(2) 逻辑一致性：确保响应始终连贯一致。(3) 可执行的下一步：提供具体指导，帮助用户做出决策。

深度： 评估所提供的比较洞察、战略优先级排序和风险意识的深度。

(1) 路径区分：说明不同备选方案或结果之间的差异。(2) 路线优先级排序：说明为何某个选项或类别优于其他选项或类别。(3) 产品层级比较：说明为何相较于其他备选方案选择了某个特定产品。(4) 风险缓解：指出潜在风险、权衡或次优选择。

L2评分器的提示词见图9。

B 实现细节

我们使用 Qwen3-30B-A3B-Thinking-2507 [52] 作为基础模型。智能体首先在由 DeepSeek-V3.2-reasoner [4]，标注的 480 条高质量轨迹上进行监督微调，批量大小为 8，学习率为 7.0e-6，共训练 2 个 epoch；随后使用 DCPO 训练 2 个 epoch，批量大小为 16，学习率为 1e-6。在强化学习训练期间，我们为每个查询采样 $K = 16$ 条轨迹，并通过动态对比选择选取 $K/2 = 8$。训练期间的最大序列长度为 32,768 个 token。推理时，我们使用 81,920 个 token 的最大上下文长度、1.0 的温度和 0.9 的 top-p。

在奖励计算方面，L1 和 L2 评分器以及过程奖励均基于 DeepSeek-V3.2-reasoner [4]。我们在 HRM 中设置了 $\alpha = 0.5$, $\beta = 0.05$, $\eta = 0.7$, $k = 5$。对于工具实现，web_search 使用智谱的 search_pro API，而 product_search 则基于我们在真实工业部署中使用的商业产品搜索 API 实现。在训练流程中，我们使用 LLaMA Fac- tory [58]进行监督微调。强化学习通过 VeRL [38] 框架进行，从而支持多节点分布式训练。

C 人工校准分析

为验证基于大语言模型的评分器的可靠性，我们通过在一个共用的评估子集上比较大语言模型判断与人工标注结果，进行人工校准。人工标注者使用与大语言模型评分器相同的评分标准评估响应。

对于 L1 评分器评估，我们使用 Cohen's Kappa 衡量人工判断与大语言模型判断之间的一致性。对于 L2 评分器评估，我们通过计算七个评估维度上人工评分与大语言模型评分之间的 Spearman 相关性，衡量二者的一致程度。

结果显示，大语言模型评分器与人工判断在 L1评分器和 L2评分器上的一致性都很高。最终一致率分别达到 $\kappa = 0.79$（L1评分器）和 $\rho = 0.71$（L2评分器）。总体而言，校准结果支持将大语言模型评分器用作可靠的人工评估的代理。

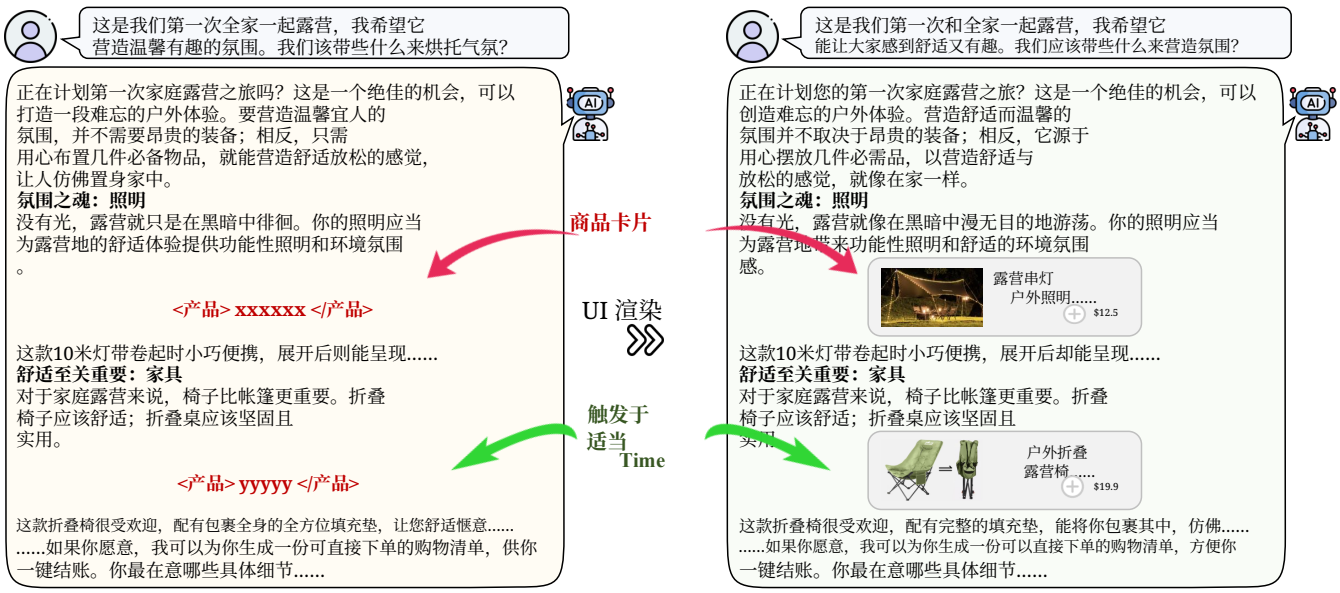


图7：我们的对话式购物代理概述。

L1评分器的提示词

你是一位专业的评估专家，负责评估 AI 智能体的表现。你将收到：

- 用户的原始查询
- 智能体的完整执行轨迹（包括所有工具调用及其结果）
- 智能体的最终响应（包括文本和用户界面卡片）

你的任务是从三个维度评估智能体的表现。

用户界面卡片格式说明

用户界面卡片格式为： <tag>id(s)</tag>

可用的卡片类型：

- <product>：电子商务产品（标识符以PD_开头）

多个商品 ID 可用逗号分隔，表示捆绑商品卡片推荐：

- 示例： <product>PD_123,PD_456</product>表示在同一张卡片中展示两个商品。

评估标准

(1) **描述忠实度**
[描述忠实度的标准]

(2) **UI 完整性（UI 卡片完整性和时效性）**
[UI 完整性的评判标准]

(3) **文本相关性**
[文本相关性的评判标准]

输入内容：

- **用户查询：** {query}
- **智能体执行轨迹：** {trajectory}

输出格式要求： 智能体的评估必须是一个有效的 JSON 对象，格式如下：

```
{
  "description_faithfulness":
    "is_pass": true/false    ,
    "reason":  "提供支持评估结论的详细推理"
  ,
  "ui_completeness":
    "is_pass": true/false    ,
    "推理":     "请提供支持此评估的详细推理"
  ,
  "文本-相关性":
    "is_pass": true/false    ,
    "推理":     "请提供详细推理以支持您的评估"

}
```

重要说明：

- 请在理由字段中提供具体示例，以支持您的评估
- 保持客观，关注具体问题
- 区分次要问题与严重失败
- 仅输出 JSON；JSON结构之外不要添加任何文本

图8：L1评分器的提示词。

L2评分器的提示词

您是一名电子商务智能体评估专家，负责根据给定的评分标准，严格且客观地评估智能体响应的质量。您的任务是**不要**给出总体评分或总结，而是：**逐项评估智能体的响应是否满足每一条评分标准**。

评估原则（必须遵守）：（1）**按评分标准独立评估：**• 每条评分标准都必须单独评估。• 当前评分标准的评估不得受其他评分标准通过或未通过的影响。（2）**严格遵循评分标准的字面含义：**

• 仅评判评分标准明确要求的内容 .
• 不得假设或添加评分标准未明确要求的能力。• 如果评分标准的要求未得到明确满足，必须标记为未通过。

(3) **基于证据的判断：**评估只能基于以下内容：
• 用户查询
• 智能体响应文本
• 智能体响应中包含的任何用户界面卡片及其官方信息。不得引入外部知识或个人偏好。

(4) **用户界面卡片使用规则：**
• 如果某项评分标准涉及规格、价格、品牌、目标用户或其他事实细节，应优先采用用户界面卡片中的官方信息 .
• 如果智能体的文本与用户界面卡片相冲突，应以用户界面卡片中的信息作为权威信息源，并将该评分标准标记为未通过。

(5) **强证据与弱证据：**许多评分标准评估的是某项决策逻辑是否被明确采用，而不仅仅是被提及。做出判断时，必须清晰区分这一点：强证据（可通过）：相关内容被清晰、核心且可感知地表达，通常表现为：

• 在给出推荐之前或在总结中明确陈述选择原则、决策维度或权衡• 明确说出"我将优先考虑/聚焦/首先考虑……"

• 将这一点作为独立的评估维度加以展开（至少用一两句清晰的话说明）弱证据（必须未通过）：
相关内容仅以下列形式出现：• 在单个产品描述中顺带提及（例如，把"无毒"当作一行卖点）

• 仅出现在 UI 标签、评价标签或标题关键词中，且未被解释为决策逻辑
• 只有模糊提示或零散的关键词，用户难以将其视为主要选择标准。如果评分标准要求"明确陈述／清晰提出／作为原则"，但证据较弱，则必须标记为未通过。(6) **示例限制：**每项评分标准可以包含通过_示例（理想结构）和／或未通过_示例（不佳结构）：• 如果智能体的行为更接近未通过_示例（不佳结构），则应标记为未通过。• 只有当智能体的行为明显更接近通过_示例（理想结构）时，才应标记为通过。• 如果评分标准未提供示例，则必须严格按照评分标准的字面含义进行判断。**输入内容：**

• **用户查询：**{query}
• **智能体响应：**{response}
• **智能体响应中的用户界面卡片（XML 标签）及其官方信息：**{ui_卡片_info}
• **评分标准（结构化 JSON，按顺序评估；不要更改顺序）：**{rubrics_json}
• **结构示例：**{structure_whole_example}

输出要求（严格）：请严格按照以下JSON数组格式输出评估结果：
• 数组的长度必须与评分标准的数量完全一致 .
• 数组中的第N个对象对应于第N条评分标准的评估结果。• 请勿添加、删除、合并或重新排序评分标准。

每个JSON项的格式：{ "is_pass": true或false, "reason": "说明其为何满足或不满足该评分标准" } 现在，请根据上述规则评估每条评分标准并输出结果。

图9：L2评分器的提示词。