TAOBAO &TMALL GROUP

E-GenRM: Generative Reward Modeling for Open-Ended E-Commerce Shopping Agents

Ying Sun[1,*], Qiuyu Zhao[1,*], Xiao Wang[2], Xiyu Zhao[1], Gege Shi[1], Zenghui Sun[1], Jinsong Lan[1], Haihong Tang[1], Bo Zheng[1,†]

[1]Taobao & Tmall Group of Alibaba [2]Technische Universität Wien

*Equal contribution †Corresponding author

arXiv:2611.03008v1 [cs.IR] 2 Nov 2026

Abstract

We present **E-GenRM** (**E**-commerce **Gen**erative **R**eward **M**odel), a generative reward model tailored to e-commerce shopping agents. User expectations are shifting from simple product queries to open-ended shopping tasks. However, existing trajectories and reward designs remain largely bounded by current agent usage patterns and known failures, providing limited guidance for further capability improvement. These limitations call for an integrated reward modeling framework for open-ended shopping agent tasks. On the data side, E-GenRM leverages naturally occurring user behavior across the shopping app to construct diverse, long-horizon conversational trajectories beyond established usage patterns. On the modeling side, E-GenRM first learns from behavior-grounded evaluation chains through supervised fine-tuning. It is then optimized through criterion-guided reinforcement learning, which rewards correct preference judgments and evidence-supported factual claims under each criterion. Personalized hard-negative training further improves its sensitivity to user preferences reflected in shopping behavior, beyond explicit requirements. We additionally develop an active evidence retrieval extension that acquires missing e-commerce information for evaluation. On E-GenRM-Bench, E-GenRM-27B achieves 87.40% preference accuracy, outperforming the strongest baseline by 6.97 percentage points. In a 14-day online A/B test on Taobao, a shopping agent post-trained with E-GenRM improves the effective shopping-assistance rate by 1.39 percentage points and user satisfaction by 2.33 percentage points.

Date: November 04, 2026

Keywords: Generative Reward Models, E-Commerce, Shopping Agent

Correspondence: yudan.sy@taobao.com*; zhaoqiuyu.zqy@taobao.com*; bozheng@alibaba-inc.com†

1 Introduction

The effectiveness of reward-guided post-training hinges critically on the quality and verifiability of the reward signal. This relationship has been shown by reinforcement learning with verifiable rewards (RLVR) [1], which has yielded substantial gains in mathematics and coding agents, where correctness can be evaluated using well-defined rules or executable feedback. However, shopping lacks similarly tractable rewards, as evaluation depends on dynamic user intent and heterogeneous product evidence [2, 3, 4, 5]. This complexity makes reliable reward modeling, and hence reward-guided post-training challenging for shopping agents.

As AI capabilities advance and shopping agents gain wider adoption, user expectations for these agents continue to rise. Traditional e-commerce search focuses on matching user queries with products, with performance typically measured by query–item relevance [6]. In practice, users must iteratively refine their intent through repeated queries, filters, and product browsing [3]. Shopping agents offer a fundamentally new interaction paradigm. An ideal shopping agent should interpret complex, open-ended shopping goals and directly follow user instructions to retrieve fine-grained product information or suitable items, reducing the cognitive burden of interpreting and refining search results [7, 3, 4].

This multi-turn interaction paradigm also shifts reward modeling from pointwise relevance to trajectory-level correctness. Relevance to the latest query alone is insufficient, as a response may overlook constraints established earlier in the interaction or recommend products unsupported by reliable evidence. Success must therefore be assessed against

requirements accumulated throughout the trajectory, with each judgment grounded in reliable product evidence [8, 9]. Learning a reward model that meets this standard requires shopping trajectories that expose diverse accumulated requirements and learning signals that provide feedback on individual requirements and decisions. Existing trajectories and reward designs are limited in both respects, which gives rise to two challenges.

Challenge 1: The Capability–Trajectory Feedback Loop. Reliable reward modeling requires diverse trajectories that expose behaviors beyond current shopping-agent usage patterns. However, deployed trajectories are endogenous to agent capabilities: users adapt their requests to supported tasks, causing logged interactions to concentrate on established use cases and known failures. Reward-model improvement is therefore constrained by trajectories generated within existing capability boundaries (Figure 1a).

Challenge 2: Reward Aggregation and Requirement-Level Credit Assignment. Shopping tasks involve interdependent decisions over long trajectories, with reward signals often sparse and delayed. Shopping requirements (such as budget, functionality, and specifications) carry different weights across tasks, which fixed aggregation cannot capture. Scalar rewards further compress per-requirement judgments into a single score, diluting the feedback on each. E-commerce platforms often list several products that meet the user's stated requirements and differ only in unstated preferences, and coarse rewards cannot distinguish among them. This information loss obscures which decisions drive success or failure, providing limited guidance for policy improvement.

To address these challenges, we propose **E-GenRM**, a generative reward model purpose-built for e-commerce. Leveraging high-value shopping trajectories, E-GenRM provides expressive reward signals for post-training shopping agents. Specifically, we collect more than 120 types of user behavior over one month and organize these heterogeneous signals into semantically structured, task-specific interaction histories. As they originate from app-wide shopping activity rather than solely from agent-mediated interactions, these histories capture diverse real-world shopping patterns that can be translated into shopping-agent evaluation trajectories, providing broader, more open-ended supervision beyond current agent usage patterns. A teacher model performs this translation by deriving task-specific criteria from the histories and constructing evidence-grounded evaluation chains for candidate responses (Figure 1b).

Using these evaluation chains, we train E-GenRM through a three-stage framework comprising supervised fine-tuning, criterion-guided reinforcement learning, and targeted hard-negative training (Figure 2). First, E-GenRM learns to derive a task-specific rubric of weighted criteria from behavioral histories and apply it to evidence-grounded evaluation. Second, criterion-guided reinforcement learning rewards correct preference judgments and verifies criterion-level factual claims against the visible evidence. Rather than directly predicting a scalar score, E-GenRM produces evidence-grounded judgments for individual criteria before deriving an overall preference. Factual feedback on these judgments provides partial credit and denser supervision beyond final preference correctness. In the final stage, personalized hard-negative training emphasizes candidates that satisfy explicit requirements but differ in behavior-supported preferences, improving the model's sensitivity to personalized trade-offs.

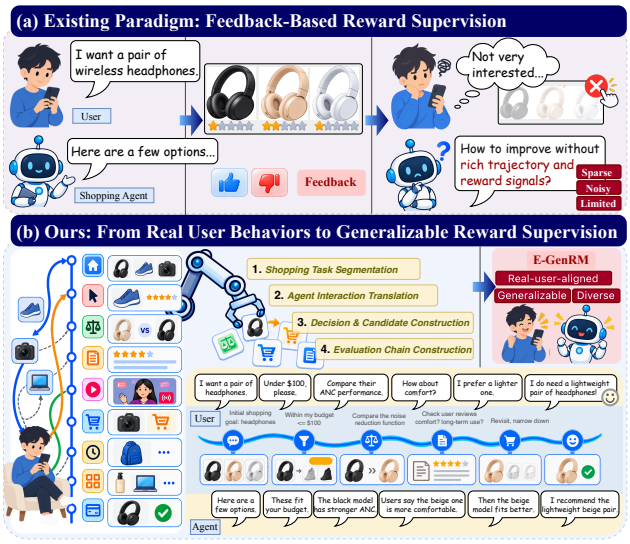


Figure 1 From agent feedback to behavior-grounded reward supervision. (a) Existing methods rely on sparse and limited agent feedback. (b) E-GenRM converts real user behavior into evaluation chains, capturing diverse reward-relevant factors.

Beyond the three-stage training framework, we develop a separately trained active e-commerce information retrieval extension to address cases where decisive information is absent from the evaluator's context, as often occurs for long-tail products or recently updated information. Building on tool-augmented reward modeling [10], its tool-use objective aligns retrieval decisions with whether the context lacks decisive evidence. The retrieved evidence is then incorporated into the same rubric-based evaluation process, expanding information access without altering the core judgment procedure. Our contributions are twofold.

- We propose **E-GenRM**, which derives task-specific evaluation criteria from naturally occurring user behavior and translates diverse real-world shopping patterns into agent-evaluation trajectories, providing open-ended supervision for robust reward modeling. E-GenRM learns context-dependent, evidence-grounded evaluation through supervised fine-tuning, criterion-guided reinforcement learning, and personalized hard-negative training, with an active evidence retrieval extension that acquires missing e-commerce information when needed.
- We construct E-GenRM-Bench from real shopping behavior to evaluate reward models. E-GenRM-27B achieves 87.40% preference accuracy on this benchmark, exceeding

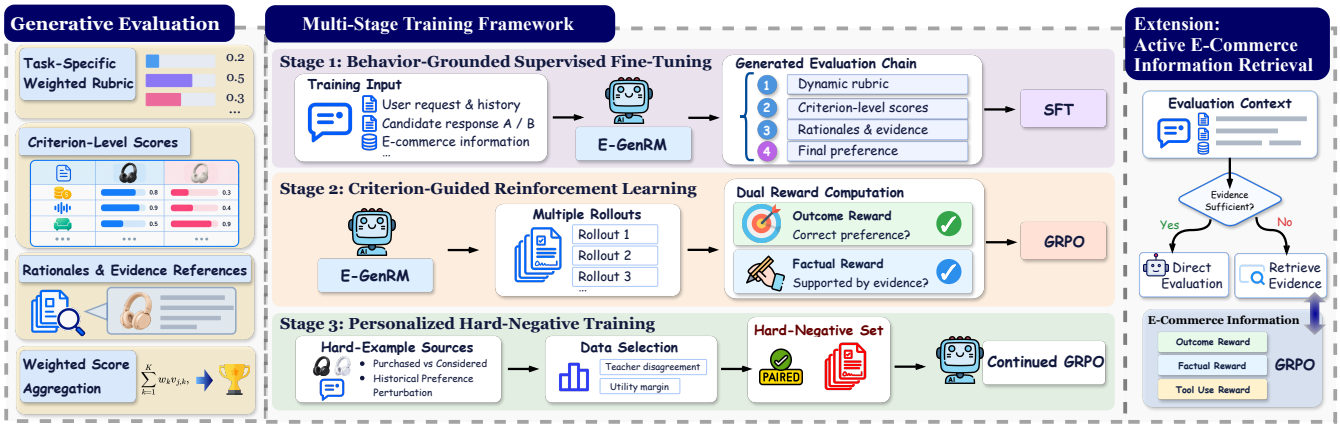


Figure 2 **Overview of E-GenRM, including the evaluation chain, three-stage training, and the active e-commerce information retrieval extension.**

the strongest baseline by 6.97 percentage points. Using E-GenRM for shopping-agent post-training further improves effective shopping assistance and user satisfaction.

2 Related Work

2.1 E-Commerce Shopping Agents

Shopping agents assist users with multi-step tasks through interaction. WebShop enables agents to search, navigate, and select products from user instructions [7], ProductAgent combines dialogue and retrieval to clarify ambiguous demands [3]. ShoppingBench expands the task space to multi-product goals involving budgets and vouchers [4]. More recent benchmarks examine long-horizon interactions: EComAgentBench distributes intent across queries, profiles, and clarification exchanges [8], while Shopping Companion studies cross-session preference inference and product-attribute verification [9]. These studies highlight the importance of interaction history and product evidence in shopping-agent evaluation.

User behavior has also been leveraged to improve agent capabilities. IntentTune uses behavioral signals to resolve underspecified intents [2]; LOFA combines purchase feedback with conversational instructions [11]; and MACS integrates preference tracking with deterministic catalog filtering for reliable execution [12]. Despite this progress, existing shopping agents remain centered on localized capabilities such as product retrieval, intent understanding, and preference tracking, and are commonly evaluated by benchmarks designed around these capabilities. They remain insufficient to address users' open-ended and evolving shopping needs.

2.2 Generative Reward Models

Generative reward models formulate evaluation as text generation, allowing reasoning to precede scores or preferences. GenRM trains verifiers via next-token prediction and supports synthetic verification rationales [13]. RM-R1 generates sample-specific rubrics or reference solutions and combines reasoning distillation with reinforcement learning [14]. Reward Reasoning Models acquire reward reasoning through RL without supervised reasoning traces [15]. RubricRM extends rubric-based evaluation to image generation and editing using input-specific weighted criteria and dimension-level rewards [16]. Together, these methods establish explicit evaluation chains as an alternative to opaque scalar scores.

Personalization and evidence acquisition address two complementary requirements of evaluation. P-GenRM derives personas and scoring rubrics from preference signals and uses user prototypes for test-time scaling [17]. A separate line of work equips reward models with evidence-seeking capabilities. ARM-Thinker learns to use tools to verify visual details and retrieve supporting documents [10], but focuses on evidence grounding rather than recovering evaluation criteria from behavioral histories. Despite recent progress, generative reward modeling remains particularly challenging in shopping due to the complexity of the underlying data, decision rules, and supporting evidence.

3 Problem Formulation

We formulate the evaluation of shopping agent responses in open-ended shopping as context-dependent pairwise reward modeling.

A shopping trajectory contains multiple decision steps at which the agent responds to the user, such as product comparison, product information queries, and the final recommendation. At decision step t, the interaction history $h_t = (u_1, \ldots, u_t)$ ends with the current user request. Each u_i records a user request or prior agent response in natural language.

The evaluator receives context $x_t = (h_t, e_t)$ and candidate responses y_t^A, y_t^B, where e_t is optional e-commerce information for the current task. The goal is to identify the response that better satisfies the user's current requirements and preferences.

The evaluator R_θ generates an evaluation chain τ_t:

$$\begin{aligned} \tau_t &= \big(C_t, \{(v_{j,k}, r_{j,k}, \mathcal{E}_{j,k})\}_{j,k}\big), \\ \tau_t &\sim R_\theta(\cdot \mid x_t, y_t^A, y_t^B). \end{aligned} \tag{1}$$

The task-conditioned rubric $C_t = \{(c_k, w_k)\}_{k=1}^K$ is shared by both candidates. Each criterion c_k has weight $w_k \geq 0$, with $\sum_k w_k = 1$. For $j \in \{A, B\}$ and $k = 1, \ldots, K$, $v_{j,k}$ is the score of candidate j under c_k. The rationale $r_{j,k}$ explains this score, and $\mathcal{E}_{j,k}$ lists its evidence references to task requirements and available information.

Scores are measured on a comparable numerical scale across criteria, with higher values indicating better performance. We deterministically compute an overall score for each candidate through weighted aggregation and predict the preference by comparing the resulting scores:

$$s_t^j = \sum_{k=1}^{K} w_k v_{j,k}, \qquad \hat{\ell}_t = \arg\max_{j \in \{A,B\}} s_t^j. \tag{2}$$

4 Method

E-GenRM learns the evaluation chain defined in Section 3. Section 4.1 converts real shopping trajectories into agent evaluation data. Section 4.2 trains E-GenRM in three stages (Figure 2). Section 4.4 introduces a separate retrieval extension that acquires e-commerce information when the current context is insufficient.

4.1 From Shopping Trajectories to Agent Evaluation Data

Training a shopping evaluator requires supervision about the requirements and trade-offs that drive a user's decision, but existing data do not directly provide such supervision. Platform behavior records what the user did, but rarely identifies why [2, 11]. Agent-generated trajectories reflect the tasks and behaviors that existing agents already produce [4], and their context is often incomplete. The central data problem is therefore to turn naturally occurring, heterogeneous behavior into agent evaluation data comprising complete decision contexts, candidate responses, and preference labels.

Our central premise is that the evaluation standard of a shopping task emerges from the organization of behavior over time. The sequence, repetition, and relation among events provide information that is unavailable from an isolated action. Such information reflects a task-specific intent that can change across sessions, rather than a fixed user profile. A purchase marks the end of a shopping task. Intermediate behavior before purchase (such as comparing products and querying product information) corresponds to decision steps at which the agent responds, and subsequent behavior indicates which response better fits the user. Accordingly, we trace each shopping trajectory backward from the purchase and construct the evaluation data in three steps.

Trajectory Collection and Semantic Translation. We collect more than 120 types of user behavior over one month and group platform events into shopping tasks with a consistent intent, each ending at a purchase decision. On average, a shopping trajectory spans 3.8 sessions and contains 162 behavioral events. Current-session behavior is restricted to the decision-time cutoff, and prior sessions provide historical context. We then translate the behavior into the interaction history defined in Section 3. User-side behavior is rendered as user requests. Content returned by the platform is rendered as prior agent responses, with search and filtering represented as tool calls and their returned results. Of these behavior types, 126 serve as context and 18 are constructed into agent decision steps. Post-decision events and the purchase outcome are stored apart from the decision-time context and are used only to construct labels and check data quality.

Decision-Step Extraction and Response Construction. Tracing backward from the purchase, we decompose each shopping trajectory into multiple decision steps, yielding 4.6 steps per trajectory on average. The response type at each decision step is determined by the user behavior that triggers it. For example, when a user compares three products, we translate this behavior into a request to view their differences. The agent should present these differences, and its inferred preference should agree with the product the user later purchases.

We take the response consistent with the user's subsequent behavior as the preferred response and construct the other candidate by perturbing it along a single dimension (such as omitting a historical preference, contradicting product facts, or recommending directly when clarification is needed). Each candidate pair therefore differs in a known aspect, and its preference label is determined jointly by the user's subsequent behavior and this controlled difference.

At the final recommendation step, the candidate pair consists of the purchased product and an alternative considered in the same decision episode. The alternative shares the purchased product's leaf category and differs in an observable attribute. Tasks in which both products were purchased are excluded. Translation and response construction are performed by an LLM. We randomly sample 500 training decision steps for e-commerce annotators to review the translation and labels, yielding a pass rate of 94.2%.

Criteria Induction and Evaluation-Chain Construction. We extend preference-conditioned evaluation-chain construction [17] to heterogeneous shopping behavior. At each decision step, a teacher examines the translated interaction history, together with candidate responses and product evidence. The teacher distinguishes explicit requirements from inferred preferences while reconciling redundant or conflicting signals. This synthesis yields a task-specific rubric C_t comprising weighted evaluation criteria. The teacher then evaluates both candidates under the rubric and produces criterion-level scores $v_{j,k}$, rationales $r_{j,k}$, and evidence references $\mathcal{E}_{j,k}$. We retain chains whose criteria are supported by the history, cover salient decision factors, and agree with the

cited product facts, and whose preference agrees with the label. E-GenRM receives only the decision-time context and its available evidence. The filtered chains form $\mathcal{D}_{\text{eval_chain}}$ for training in Section 4.2.

4.2 Multi-Stage Training

4.2.1 *Stage 1: Behavior-Grounded Supervised Fine-Tuning*

Stage 1 fine-tunes the model on $\mathcal{D}_{\text{eval_chain}}$ from Section 4.1. Each example $(x_t, y_t^A, y_t^B, \tau_t)$ corresponds to a decision step. Its input is the decision context x_t and the candidate pair defined in Section 3, and its target is the filtered teacher evaluation chain τ_t from Section 4.1. The model thus learns to derive the rubric from the visible context and then assess both candidates under each criterion. Current explicit requirements take precedence over conflicting historical inferences, keeping the learned rubric tied to the active task. Following next-token training for generative verification [13], we supervise the complete evaluation chain τ_t and optimize

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x_t, y_t^A, y_t^B, \tau_t)\sim\mathcal{D}_{\text{eval_chain}}}\left[\log p_\theta(\tau_t \mid x_t, y_t^A, y_t^B)\right]. \tag{3}$$

4.2.2 *Stage 2: Criterion-Guided Reinforcement Learning*

In Stage 1, E-GenRM learns to infer task-relevant criteria and express them as a rubric. Although SFT exposes the model to evidence-based demonstrations, self-generated assessments admit diverse reasoning paths and surface forms, so its token-level imitation objective only indirectly reflects whether these assessments are supported by evidence. This is particularly evident when high-level shopping requirements involve multiple product attributes. This motivates directly optimizing model-generated assessments against verifiable criterion-level feedback [9].

Criterion-Level Factual Feedback. We optimize complete evaluation chains with Group Relative Policy Optimization (GRPO) [18], allowing the model to generate both the rubric and its candidate assessments. Training examples are drawn from decision steps in Section 4.1 whose users are disjoint from those in Stage 1. Each example contains only the decision context, the candidate pair, and the preference label, without a teacher evaluation chain. The reward combines final preference correctness with criterion-level feedback from a fact verifier. For each generated criterion, the verifier checks the factual claims in each candidate assessment against the visible evidence (such as product facts, returned tool results, and the interaction history). A claim citing an evidence reference absent from the visible input receives no factual credit, and a numerical comparison whose operands are addressable in the structured data is verified directly. This feedback evaluates the link from each criterion to the evidence used to assess it, providing partial credit for supported judgments within a chain.

For criterion k, let $F_{j,k} \in \{0, 0.5, 1\}$ mark whether the assessment of candidate $j \in \{A, B\}$ is unsupported, partially supported, or supported by visible evidence. The criterion-level reward is

$$r_k = \frac{F_{A,k} + F_{B,k}}{2}, \qquad R_{\text{crit}} = \frac{1}{K}\sum_{k=1}^{K} r_k. \tag{4}$$

We combine the criterion-level signal with a reward for the final pairwise decision,

$$R = R_{\text{out}} + \lambda R_{\text{crit}}, \tag{5}$$

where R_{out} rewards a correct final preference and λ balances the two objectives. GRPO compares alternative evaluation chains for the same decision step, encouraging chains that make the right choice for the right evidence-based reasons.

4.2.3 *Stage 3: Personalized Hard-Negative Training*

As Stage 2 improves evidence-based judgment, the remaining errors concentrate on cases where both candidates satisfy the explicit shopping requirements. Verifying their price, functionality, and intended use then establishes only that both are suitable, and which one better fits this user depends on personal preferences supported by repeated interactions (such as material, style, or brand). Among the candidate pairs constructed in Section 4.1, such cases mainly correspond to final recommendation steps that compare the purchased product with a considered alternative, and to perturbed candidates that alter a historical preference. We construct a targeted hard-negative set from these two types of candidate pairs for this decision boundary.

Hard-Example Selection. Hard-negative training has been used to refine personalized reward models [17]. We select hard examples using two complementary signals. The first signal is teacher disagreement. We retain pairs for which roughly half of the teachers' judgments disagree with the preference label, as these pairs expose ambiguous reasoning. The second signal is utility margin. We retain pairs with similar aggregated utilities, as these pairs lie near the decision boundary. All selected pairs satisfy explicit task constraints and are mixed with standard examples to preserve coverage while emphasizing fine-grained preference distinctions.

Continued Training. From the Stage 2 model, we continue training on this mixture with the same outcome and criterion-level objectives. This stage focuses learning on user-specific trade-offs while retaining the evidence-based judgment learned in Stage 2.

4.3 E-GenRM-Bench Construction

E-GenRM-Bench evaluates whether a reward model can identify the agent response that better fits the user at a shopping decision step. Each evaluation item follows the interface in which a reward model evaluates shopping agent responses, comprising a decision-time context, two candidate responses, and a preference label.

Expert Annotation. Two e-commerce experts independently evaluate each candidate pair and determine the superior response. They are given the decision-time context, both responses, product evidence, and the user's subsequent behavior, allowing the annotation to reflect both response quality and the user's observed choice. Disagreements are resolved by a third expert; ties are excluded, and the adjudicated judgment serves as the reference label.

Evidence Separation and Evaluation Groups. The construction pipeline separates decision evidence, preference labels, and reference criteria, with teacher-generated criteria retained as references for criterion-recall evaluation. For evaluation, a fixed manifest assigns each held-out decision step to one group. Decision steps containing behavior or leaf categories unseen during training are assigned to OOD. Among the remaining decision steps, those with high query or comparison complexity are assigned to Hard, and all others to Standard. The group assignments do not depend on evaluated-model predictions.

4.4 Active E-Commerce Information Retrieval

The evaluation capability trained in the preceding stages relies only on the information available to the evaluator, whereas some e-commerce decisions depend on long-tail or rapidly changing facts (such as detailed compatibility, current inventory, or recent product feedback). These facts are often absent from the context and cannot be reliably recovered from model knowledge, which leaves the evaluator without decisive evidence. Retrieval can improve answers to e-commerce brand and category questions [26], and tool-augmented reward models likewise acquire external evidence before judging a response [10]. We therefore add an evidence-acquisition policy to E-GenRM while keeping its rubric generation and judgment procedure unchanged.

Evidence Acquisition. We extend the evaluation trajectory with retrieval actions and their observations, following the ReAct [27] and adaptive evidence acquisition [28]. At each step, the evaluator can request product-encyclopedia or review evidence, or complete its judgment using the available context. Let $\mathcal{E}_{\text{tool}}(\tau)$ collect the facts returned along trajectory τ. These facts augment the initial evidence e_t and support the same criterion-level assessments and preference aggregation defined in Section 3.

Retrieval-Need Supervision. To supervise evidence acquisition, the extension data pairs two teacher evaluations of each task: one uses the original context, and the other additionally uses facts returned by the available tools. We assign a retrieval-need label $n = 1$ when the original context lacks a decisive fact and the augmented context supplies evidence needed to establish or correct the preference. Tasks whose original evidence already supports the judgment receive $n = 0$.

Tool-Use Reward. Let $c(\tau) \in \{0, 1\}$ indicate whether trajectory τ invokes a tool. We retain the core evaluation reward $R_{\text{out}} + \lambda R_{\text{crit}}$ and align the retrieval decision with the need label through

$$R_{\text{use}}(\tau) = \begin{cases} +1, & n = 1,\ c(\tau) = 1,\ R_{\text{out}}(\tau) = 1, \\ -1, & c(\tau) \neq n, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The positive reward couples a needed retrieval action with a correct final judgment. The combined objective is

$$R_{\text{ext}}(\tau) = R_{\text{out}}(\tau) + \lambda R_{\text{crit}}(\tau) + R_{\text{use}}(\tau). \tag{7}$$

Continuing GRPO with this objective trains evidence acquisition alongside rubric-based preference evaluation.

5 Experiments

5.1 Experimental Setup

E-GenRM-Bench. Following the construction in Section 4.3, E-GenRM-Bench contains 373 expert-annotated decision steps with 2,387 teacher-generated reference criteria and decision-time contexts collectively containing 45,133 behavior events that may inform reward judgments. For expert annotation, at most two decision steps are sampled from each trajectory. The selected steps cover 231 users and 264 product categories and form three mutually exclusive groups, Standard (262), Hard (55), and OOD (56).

Shopping Reasoning Bench. SRB contains 525 expert-authored shopping missions and 10,863 expert rubric items [31]. Each method generates criteria from the current and preceding user messages, without access to the expert rubrics used for scoring. We report the Full set and the official Hard subset of 108 missions included in it.

Metrics. E-GenRM-Bench accuracy (Acc.) measures agreement between model judgments and expert preference labels. A decision step is counted as correct only if the model selects the preferred response under both candidate orderings; order-inconsistent or invalid outputs are counted as incorrect. Avg. denotes micro-averaged accuracy over all decision steps. For ablations, lexical reference-criterion recall (RR-L) measures coverage of teacher-generated reference criteria and is macro-averaged over decision steps. On SRB, semantic F1 (Sem-F1) scores generated–reference criterion pairs with BERTScore [32] and computes F1 from their maximum-weight one-to-one matching. We adapt ROUGE-L [33] to compute required-weighted recall (RL), which measures lexical recovery with greater weight on required items. SRB scores are averaged first over turns within each mission and then over missions.

Baselines. On E-GenRM-Bench, we compare E-GenRM with LLM-as-a-Judge methods and reward models. LLM-as-a-Judge baselines are evaluated on five base models, including two open-source models (Qwen3.5-9B and Qwen3.5-27B) and three proprietary models (GPT-5.6-Sol, Gemini-3.8-Flash, and Qwen3.8-Max). Direct produces the output directly, whereas CoT [34] reasons explicitly before producing it. CheckEval [19] uses a checklist. Dataset-Specific

Table 1 **Comparison with LLM-as-a-Judge methods on E-GenRM-Bench (%). Bold/underline: best/second-best per column. Dark/light red: best/second-best per metric across Tables 1 and 2. Marks exclude E-GenRM + retrieval. —: not applicable.**

Method	Open-source models								Proprietary models											
	Qwen3.5-9B				Qwen3.5-27B				GPT-5.6-Sol				Gemini-3.8-Flash				Qwen3.8-Max			
	Standard	Hard	OOD	Avg.	Standard	Hard	OOD	Avg.	Standard	Hard	OOD	Avg.	Standard	Hard	OOD	Avg.	Standard	Hard	OOD	Avg.
LLM-as-a-Judge																				
Direct	56.49	54.55	48.21	54.96	62.21	45.45	57.14	58.98	70.61	61.82	69.64	69.17	79.01	67.27	80.36	77.48	58.40	49.09	50.00	55.76
CoT	53.82	52.73	48.21	52.82	59.92	52.73	55.36	58.18	70.61	63.64	66.07	68.90	80.15	**69.09**	73.21	77.48	60.69	49.09	55.36	58.18
CheckEval [19]	51.15	41.82	50.00	49.60	53.05	47.27	50.00	51.74	54.58	50.91	55.36	54.16	68.70	65.45	71.43	68.63	58.40	41.82	**58.93**	56.03
Dataset-Specific Rubric [20]	38.93	49.09	41.07	40.75	60.69	47.27	58.93	58.45	**76.72**	**69.09**	**73.21**	**75.07**	**82.82**	65.45	**83.93**	**80.43**	62.98	49.09	57.14	60.05
Instance-Specific Rubric [20]	65.65	63.64	67.86	65.68	56.87	38.18	57.14	54.16	54.20	56.36	55.36	54.69	55.34	52.73	58.93	55.50	56.87	47.27	51.79	54.69
Rubrics on Trial [21]	56.11	54.55	48.21	54.69	66.79	56.36	53.57	63.27	63.36	63.64	53.57	61.93	55.34	50.91	64.29	56.03	**63.36**	**56.36**	**58.93**	**61.66**
Ours																				
E-GenRM	**82.06**	**67.27**	**75.00**	**78.82**	**85.88**	**69.09**	**80.36**	**82.57**	NOTE:											
Δ vs. same-scale best	+3.82	−3.64	−3.57	+1.88	+17.18	+12.73	+12.50	+16.35	Compared to the best open-source baseline of comparable size (≤10B for 9B; 27–32B for 27B), selected per metric.											
Δ vs. best baseline overall	−0.76	−3.64	−8.93	−1.61	+3.06	−1.82	−3.57	+2.14	Compared to the best baseline among all methods in Tables 1 and 2, selected per metric.											
E-GenRM + retrieval	—	—	—	—	90.08	76.36	85.71	87.40	Retrieval is evaluated with Qwen3.5-27B (Section 5.4).											
Δ vs. best baseline overall					+7.26	+5.45	+1.78	+6.97	Compared to the best baseline among all methods in Tables 1 and 2, selected per metric.											

Table 2 **Comparison with reward models on E-GenRM-Bench (%). Notation follows Table 1.**

Method	Base model	Standard	Hard	OOD	Avg.
Bradley–Terry RM [22]	Qwen3.5-9B	59.16	50.91	58.93	57.91
Bradley–Terry RM	Qwen3.5-27B	68.70	52.73	67.86	66.22
Skywork-Reward-V2 [23]	Qwen3-4B	75.57	69.09	67.86	73.46
Skywork-Reward-V2	Qwen3-8B	75.95	65.45	78.57	74.80
Skywork-Reward-V2	Llama-3.1-8B	78.24	**70.91**	76.79	76.94
RM-R1 [14]	DeepSeek-Distilled-Qwen-7B	35.88	38.18	42.86	37.27
RM-R1	DeepSeek-Distilled-Qwen-32B	49.62	49.09	39.29	47.99
RM-R1	Qwen2.5-7B-Instruct	39.69	25.45	35.71	37.00
RM-R1	Qwen2.5-32B-Instruct	31.68	21.82	21.43	28.69
GRAM-R^2 [24]	Llama-3.1-8B	36.64	32.73	30.36	35.12
RewardAnything [25]	Qwen3-8B	51.15	54.55	44.64	50.67
Ours					
E-GenRM	Qwen3.5-9B	82.06	67.27	75.00	78.82
E-GenRM	Qwen3.5-27B	**85.88**	69.09	**80.36**	**82.57**
E-GenRM + retrieval	Qwen3.5-27B	90.08	76.36	85.71	87.40

Rubric [20] constructs criteria shared across the dataset, Instance-Specific Rubric [20] constructs task-specific criteria, and Rubrics on Trial [21] refines criteria through synthetic comparisons. Scalar reward models include Bradley–Terry RM [22], trained on the same shopping preference pairs, and released Skywork-Reward-V2 checkpoints [23]. Generative reward models include RM-R1 [14] and GRAM-R^2 [24], which generate reasoning-based judgments, and RewardAnything [25], which evaluates candidates against supplied principles. On SRB, the LLM-as-a-Judge baselines generate rubrics instead of preference judgments, and we additionally compare the query-conditioned rubric-generation components of RLCF [29] and CARMO [30]. We evaluate released reward checkpoints through their preference-judgment interfaces and restrict the SRB comparison to methods adapted for standalone rubric generation.

Implementation Details. We train Qwen3.5-9B and Qwen3.5-27B following Section 4.2. SFT uses LoRA [35] with rank 64 and scaling factor 128 for 1 epoch, a learning rate of 5×10^{-5}, a global batch size of 32, and 3% warmup. RL updates all parameters at 10^{-6} with an optimizer batch size of 16. Each RL rollout batch contains 16 prompts with four responses per prompt. The three stages use 4,492, 1,503, and 747 examples, respectively, and Stage 3 includes 135 hard examples. Training and evaluation users are disjoint. GPT-5.6-Sol serves as the teacher that generates evaluation chains and as the fact verifier that provides criterion feedback. We set the factual-feedback weight to $\lambda = 0.25$.

5.2 Main Results

Tables 1 and 2 report preference accuracy on E-GenRM-Bench against LLM-as-a-Judge methods and reward models, respectively, and Table 3 evaluates rubric generation on SRB, an external benchmark held out from training and model selection. In Avg. Acc., E-GenRM outperforms all same-scale baselines at both scales, and E-GenRM-27B further surpasses every proprietary-model configuration. Specifically, E-GenRM-27B attains 82.57%, exceeding the strongest baseline, Gemini-3.8-Flash with Dataset-Specific Rubric, by 2.14 percentage points. E-GenRM-9B attains 78.82%, exceeding the best same-scale open-source baseline, Skywork-Reward-V2-Llama-3.1-8B, by 1.88 points. With active e-commerce information retrieval (Section 5.4), E-GenRM-27B reaches 87.40%, widening its lead over the strongest baseline to 6.97 points.

Comparison with LLM-as-a-Judge Methods. Under the same base model, E-GenRM outperforms every prompting baseline on all subsets. Relative to the strongest prompting baseline, E-GenRM-9B and E-GenRM-27B improve Avg. Acc. by 13.14 and 19.30 points, respectively. On the Hard subset, where a decision step requires aggregating multiple queries or comparing multiple product attributes, E-GenRM-27B gains 12.73 points. The largest gain of E-GenRM-27B emerges on the OOD subset, where it reaches 80.36% versus 58.93% for the strongest prompting baseline, Dataset-Specific Rubric, a margin of 21.43 points. Whereas Dataset-Specific Rubric applies a single rubric across the entire dataset, E-

Table 3 **Rubric generation on SRB (%). Sem: semantic F1; RL: required-weighted ROUGE-L recall; F/H: Full/Hard. Bold/underline: best/second-best per column. Dark/light red: best/second-best per metric. —: not applicable.**

Method	Open-source models								Proprietary models											
	Qwen3.5-9B				Qwen3.5-27B				GPT-5.6-Sol				Gemini-3.8-Flash				Qwen3.8-Max			
	Sem-F	Sem-H	RL-F	RL-H	Sem-F	Sem-H	RL-F	RL-H	Sem-F	Sem-H	RL-F	RL-H	Sem-F	Sem-H	RL-F	RL-H	Sem-F	Sem-H	RL-F	RL-H
Prompted rubric construction																				
Direct	35.38	39.81	12.27	12.10	62.50	61.13	24.75	21.25	63.30	60.48	**26.68**	23.33	67.19	65.14	24.17	**22.51**	**68.84**	67.71	22.65	19.86
CoT	30.22	33.47	10.48	10.14	62.23	59.76	25.18	21.90	66.35	64.12	24.34	22.79	70.08	70.68	22.56	20.51	68.40	**67.77**	23.70	21.74
Dataset-Specific Rubric [20]	50.56	48.41	15.66	14.45	53.79	51.54	15.05	13.59	53.69	51.49	14.87	14.27	65.22	62.90	13.62	12.85	53.62	51.40	15.71	14.87
Instance-Specific Rubric [20]	48.37	51.68	16.72	15.65	63.08	61.98	25.47	21.64	**67.03**	**64.81**	24.29	22.86	71.40	70.99	22.99	21.63	63.68	62.67	22.30	20.21
Rubrics on Trial [21]	18.44	22.71	6.66	7.10	60.16	58.13	24.48	21.27	66.11	63.89	24.43	22.62	**72.78**	**73.39**	24.30	22.45	64.20	61.57	22.60	19.45
Rubric generation from reward methods																				
RLCF [29]	54.22	54.57	21.28	19.20	60.53	58.82	24.88	21.64	65.75	63.83	24.36	22.63	72.48	71.78	24.40	21.89	65.12	61.90	23.98	21.00
CARMO [30]	52.55	55.14	20.83	19.49	61.05	59.70	25.90	22.75	64.55	62.28	25.78	**24.40**	70.69	69.11	**24.47**	21.33	65.57	63.11	**27.58**	**24.89**
E-GenRM	**72.27**	**71.69**	**24.47**	**21.33**	**72.53**	**72.56**	**28.67**	**25.38**	NOTE:											
Δ vs. same-scale best	+18.05	+16.55	+3.19	+1.84	+9.45	+10.58	+2.77	+2.63	Compared to the best baseline of comparable size (≤10B for 9B; 27–32B for 27B), selected per metric.											
Δ vs. best baseline overall	−0.51	−1.70	−3.11	−3.56	−0.25	−0.83	+1.09	+0.49	Compared to the best baseline among all methods, selected per metric.											

GenRM generates a rubric conditioned on the context of each decision step, which is consistent with its gains on unseen behavior and product categories.

Comparison with Reward Models. E-GenRM-9B and E-GenRM-27B achieve higher Avg. Acc. than every reward model in Table 2, and E-GenRM-27B also leads on Standard and OOD. The Bradley–Terry RM shares the shopping preference data and base model with E-GenRM-27B but outputs only a scalar score, whereas E-GenRM first generates task-specific criteria and then evaluates the candidates against each criterion. Under this controlled comparison, E-GenRM-27B exceeds it by 16.35 points. For generative reward models, E-GenRM-9B exceeds the strongest one. These results suggest shopping evaluation calls for training data and objectives tailored to e-commerce challenges, such as inducing evaluation criteria from real shopping behavior and verifying candidate claims on product attributes against facts.

Rubric Generation on SRB. The rubric generation ability of E-GenRM transfers to the external benchmark. With either base model, E-GenRM outperforms all methods built on the same base model across all four metrics. Because SRB criteria are generated solely from user messages, this result indicates that the rubric generation learned in Stage 1 is not tied to the behavioral inputs of E-GenRM-Bench. Across all methods, E-GenRM-27B achieves the highest lexical recall, outperforming the strongest baselines by 1.09 points on Full and 0.49 on Hard, while trailing the best proprietary models in semantic F1 by only 0.25 points on Full and 0.83 on Hard.

5.3 Ablation Studies

Table 4 ablates the training procedure on Qwen3.5-9B by stage.

Behavior-Grounded SFT. Compared with label-only SFT, evaluation-chain supervision improves Avg. Acc. by 5.36 points. The gain is largest on OOD (12.50 points) and smallest on Standard (3.82 points), which is consistent with evaluation chains teaching the model to derive criteria from the visible context rather than merely fit preference labels.

Table 4 **Qwen3.5-9B ablations on E-GenRM-Bench (%). Each entry reports Acc./RR-L. Blocks share an initialization; w/o removes the named component. –: unavailable RR-L.**

Variant	Standard Acc./RR-L	Hard Acc./RR-L	OOD Acc./RR-L	Avg. Acc./RR-L
Stage 1: Behavior-grounded SFT *(init.: Qwen3.5-9B)*				
Evaluation-chain SFT	**70.61/20.26**	**61.82/18.32**	**66.07/20.83**	**68.63/20.05**
w/o evaluation chain (label-only SFT)	66.79/—	56.36/—	53.57/—	63.27/—
Stage 2: Criterion-guided RL *(init.: Stage 1 model)*				
Criterion-guided GRPO	**75.95/27.79**	**65.45/24.06**	**67.86/24.33**	**73.19/26.74**
w/o GRPO (continued SFT)	72.90/24.90	63.64/23.51	66.07/22.17	70.51/24.30
w/o outcome reward	60.31/12.97	40.00/12.11	39.29/8.67	54.16/12.23
w/o factual feedback	72.14/22.05	65.45/24.45	66.07/15.67	70.24/21.51
Stage 3: Personalized hard-negative training *(init.: Stage 2 model)*				
(Ours) E-GenRM	**82.06/15.45**	**67.27/13.84**	**75.00/17.00**	**78.82/15.42**
w/o personalized hard negatives	77.86/21.60	61.82/21.15	66.07/22.67	73.73/21.69

Criterion-Guided Reinforcement Learning. Criterion-guided GRPO outperforms continued SFT by 2.68 points. To reward components, the outcome reward has the larger effect. Removing it reduces Avg. Acc. by 19.03 points, with drops of 25.45 and 28.57 points on Hard and OOD. Removing factual feedback reduces Avg. Acc. by 2.95 points and Avg. RR-L by 5.23 points, including an 8.66-point RR-L drop on OOD. Factual feedback therefore affects criterion recall more than accuracy, which is consistent with the fact verifier providing feedback at the criterion level.

Personalized Hard-Negative Training. Compared with ordinary continued RL from the same Stage 2 model, personalized hard-negative training improves Avg. Acc. by 5.09 points and OOD accuracy by 8.93 points, while Avg. RR-L decreases by 6.27 points. In these hard negatives, both candidates satisfy the explicit requirements and must be distinguished by personal preferences. This is consistent with the model exploring evaluation criteria beyond the teacher references to improve accuracy.

Table 5 **Active e-commerce information retrieval of E-GenRM-27B on E-GenRM-Bench (%). B is the call budget, and a view is one candidate order. Bold marks the better accuracy within each matched pair.**

Configuration			Accuracy (%) ↑				Tool usage	
Retrieval	Tool training	B	Standard	Hard	OOD	Avg.	Call rate (%)	Calls per view
None	No	0	85.88	69.09	**80.36**	82.57	0.00	0.000
None	Yes	0	**86.26**	**72.73**	**80.36**	**83.38**	0.00	0.000
Autonomous	No	1	88.55	**80.00**	82.14	86.33	95.84	0.958
Autonomous	Yes	1	**90.08**	76.36	**85.71**	**87.40**	22.12	0.221
Fixed	Yes	2	90.46	76.36	82.14	87.13	100.00	2.000

Table 6 **Offline evaluation of the Shopping Agent post-trained with E-GenRM-27B or the Expert Judge (expert-calibrated Gemini-3.8-Flash). Prod., Quality, and Pers. are sub-scores, and Overall is their mean.**

Training	Evaluator	Sub-scores ↑			Overall ↑
		Prod.	Quality	Pers.	
None	—	0.693	0.781	0.486	0.653
GRPO	Expert Judge	0.759	0.858	0.528	0.715
	E-GenRM-27B	**0.777**	**0.879**	**0.529**	**0.728**
DPO	Expert Judge	**0.760**	0.806	0.493	0.686
	E-GenRM-27B	0.747	**0.864**	**0.505**	**0.705**

5.4 Effect of Active E-Commerce Information Retrieval

Table 5 evaluates active e-commerce information retrieval of our best-performing variant, E-GenRM-27B, using product-encyclopedia search and review search as tools. We compare E-GenRM-27B with and without tool-use training, each evaluated without retrieval and with autonomous retrieval. The autonomous conditions share the same retrieval instruction and a one-call budget. As a reference, fixed retrieval supplies the tool-trained model with the results of two programmatic calls before judgment.

Effect of Retrieval. Autonomous retrieval improves Avg. Acc. by 3.76 points for E-GenRM-27B and by 4.02 points for the tool-trained model. This is consistent with the motivation in Section 4.4 that some shopping decisions depend on e-commerce information absent from the context.

Effect of Tool-Use Training. Under autonomous retrieval, tool-use training improves Avg. Acc. by 1.07 points and lowers the call rate to 22.12% (95.84% without training). This is consistent with R_{use} penalizing calls when the context is sufficient. Compared with fixed retrieval, the tool-trained autonomous policy achieves comparable accuracy (87.40% vs. 87.13%) with 0.221 calls per view.

5.5 Application to a Deployed Shopping Agent

We use E-GenRM-27B to post-train the Taobao Omni-Shop Agent, which serves over 100,000 user queries daily. We apply DPO [36] and GRPO [18], comparing E-GenRM-27B supervision with an expert-calibrated Gemini-3.8-Flash Judge. For DPO, each evaluator ranks candidate responses to construct preferred–rejected pairs; for GRPO, it evaluates sampled responses to provide reward signals. Within each algorithm, we hold the policy initialization, task pool, and update budget fixed across supervision sources, and include the policy before post-training as a baseline.

Table 7 **Online A/B test results of Shopping Agent over 14 days. Δ is the change relative to the control; rates use percentage points (pp), and P99 latency uses relative change.**

Metric	Δ	p-value
Effective assistance rate	+1.39 pp	0.0433
Satisfaction rate	+2.33 pp	0.0082
Card coverage	+7.47 pp	0.0067
P99 latency ↓	−17.88%	—

Offline Evaluation. A proprietary e-commerce evaluation model scores product suitability (Prod.), response quality (Quality), and personalized preference alignment (Pers.); Overall denotes their arithmetic mean (Table 6). E-GenRM supervision improves Overall score by 0.075 under GRPO and 0.052 under DPO, exceeding the expert Judge by 0.013 and 0.019. The most pronounced improvement is in Quality, where E-GenRM leads the expert Judge by 0.021 and 0.058. These gains are consistent with E-GenRM's evaluation chain, which provides policy training with shopping-specific supervision that reflects both task requirements and personalized trade-offs.

Online A/B Results. In a 14-day A/B test on a Taobao Shopping Agent, the agent post-trained with E-GenRM significantly improves effective shopping assistance, user satisfaction, and product-card coverage, and reduces P99 latency by 17.88% (Table 7).

6 Conclusion

In this paper, we presented E-GenRM, an integrated data and modeling framework for e-commerce reward modeling in support of open-ended shopping-agent post-training. Its central insight is to derive evaluation criteria from naturally occurring shopping behavior and translate these diverse real-world behavioral patterns into shopping-agent trajectories, thereby providing more open-ended training data for reward models robust to diverse agent behaviors. E-GenRM constructs agent evaluation data and E-GenRM-Bench from real shopping trajectories, learns evidence-grounded evaluation through three-stage training, and acquires missing e-commerce information through an active e-commerce information retrieval extension. we offered a reference for reward design and agent post-training in related e-commerce applications. Experiments show that E-GenRM outperforms all same-scale baselines on E-GenRM-Bench and generalizes to unseen behavior, product categories, and an external benchmark. In an online A/B test, the shopping agent post-trained with E-GenRM improves the effective shopping-assistance rate and user satisfaction.

References

[1] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024.

[2] Rachith Aiyappa, Ishita K. Khan, Chester Palen-Michel, Jayanth Yetukuri, Samarth Agrawal, Mehran Elyasi, and Shuang Zhou. Intenttune: Using user demand and personalization to resolve "unknown" query intents for e-commerce search. *CoRR*, abs/2607.01530, 2026.

[3] Jingheng Ye, Yong Jiang, Xiaobin Wang, Yinghui Li, Yangning Li, Pengjun Xie, and Fei Huang. Productagent: Benchmarking conversational product search agent with asking clarification questions. In Saloni Potdar, Lina Maria Rojas-Barahona, and Sébastien Montella, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 383–398. Association for Computational Linguistics, 2025.

[4] Jiangyuan Wang, Kejun Xiao, Qi Sun, Huaipeng Zhao, Tao Luo, Jiandong Zhang, and Xiaoyi Zeng. Shoppingbench: A real-world intent-grounded shopping benchmark for llm-based agents. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 33521–33529. AAAI Press, 2026.

[5] Yiwen Tang, Qiuyu Zhao, Zenghui Sun, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Revision:reflective intent mining and online reasoning auxiliary for e-commerce visual search system optimization. In *42nd IEEE International Conference on Data Engineering, ICDE 2026, Montreal, QC, Canada, May 4-8, 2026*, pages 3764–3771. IEEE, 2026.

[6] Chandan K. Reddy, Lluís Màrquez, Fran Valero, Nikhil Rao, Hugo Zaragoza, Sambaran Bandyopadhyay, Arnab Biswas, Anlu Xing, and Karthik Subbian. Shopping queries dataset: A large-scale ESCI benchmark for improving product search. *CoRR*, abs/2206.06588, 2022.

[7] Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. Webshop: Towards scalable real-world web interaction with grounded language agents. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[8] Zeyao Du, Tong Li, Yanci Zhang, and Haibo Zhang. Ecomagentbench: Benchmarking shopping agents on long-horizon tasks with distributed hidden intent. *CoRR*, abs/2606.17698, 2026.

[9] Zijian Yu, Kejun Xiao, Huaipeng Zhao, Tao Luo, and Xiaoyi Zeng. Shopping companion: Benchmarking and training llm agents for long-horizon preference-grounded e-commerce tasks, 2026.

[10] Shengyuan Ding, Xinyu Fang, Ziyu Liu, Yuhang Zang, Yuhang Cao, Xiangyu Zhao, Haodong Duan, Xiaoyi Dong, Jianze Liang, Bin Wang, Conghui He, Dahua Lin, and Jiaqi Wang. Arm-thinker: Reinforcing multimodal generative reward models with agentic tool use and visual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pages 22195–22205. Computer Vision Foundation, 2026.

[11] Haobo Zhang, Kelong Mao, Sulong Xu, Simiu Gu, and Zhicheng Dou. Learning from online user feedback for shopping agents. *CoRR*, abs/2608.11604, 2026.

[12] Juli Huang, Hannah Clay, Sajjad Beygi, Thomas Sarda, Negin Golrezaei, and Amin Saberi. MACS: A hybrid multi-agent framework for reliable conversational e-commerce recommendation. *CoRR*, abs/2608.14068, 2026.

[13] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

[14] Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. RM-R1: reward modeling as reasoning. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026.

[15] Jiaxin Guo, Zewen Chi, Li Dong, Qingxiu Dong, Xun Wu, Shaohan Huang, and Furu Wei. Reward reasoning models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 150477–150510. Curran Associates, Inc., 2025.

[16] Zijian Kan, Wei Wang, Long Luo, Bing Zhao, Xuan Ren, WeiXu Qiao, Wenbo Li, Hu Wei, and Lin Qu. Rubricrm: Generative reward modeling via dynamic rubrics for image generation and editing. *CoRR*, abs/2608.26956, 2026.

[17] Pinyi Zhang, Ting-En Lin, Yuchuan Wu, Jingyang Chen, Zongqi Wang, Hua Yang, Bing Zhao, Fei Huang, Yongbin Li, and Kai Zhang. P-genrm: Personalized generative reward model with test-time user-based scaling. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026.

[18] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

[19] Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 15771–15798. Association for Computational Linguistics, 2025.

[20] Zijie Wang and Eduardo Blanco. Generating and refining dynamic evaluation rubrics for llm-as-a-judge. *CoRR*, abs/2605.30568, 2026.

[21] Haocheng Yang, Licheng Pan, Yuan Lu, Xiaoxi Li, Zhichao Chen, Chunyuan Zheng, Shuai Shao, Haoxuan Li, Hao Wang, and Kun Zhang. Rubrics on trial: Evolving rubrics from a single query via synthetic pairwise evidence, 2026.

[22] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[23] Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and Yang Liu. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In Carl Vondrick, Bharath Hariharan, Colin Raffel, Lerrel Pinto, Diyi Yang, and Aleksandra Faust, editors, *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. proceedings.iclr.cc, 2026.

[24] Chenglong Wang, Yongyu Mu, Hang Zhou, Yifu Huo, Ziming Zhu, Jiali Zeng, Murun Yang, Bei Li, Xiaoyang Hao, Chunliang Zhang, Fandong Meng, Jingbo Zhu, and Tong Xiao. Gram-r^2: Self-training generative foundation reward models for reward reasoning. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor, editors, *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 33395–33403. AAAI Press, 2026.

[25] Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. Rewardanything: Generalizable principle-following reward models. *CoRR*, abs/2506.03637, 2025.

[26] Haibin Chen, Kangtao Lv, Chengwei Hu, Yanshi Li, Yujin Yuan, Yancheng He, Xingyao Zhang, Langming Liu, Shilei Liu, Wenbo Su, and Bo Zheng. Chineseecomqa: A scalable e-commerce concept evaluation benchmark for large language models. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy, editors, *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pages 5311–5321. ACM, 2025.

[27] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[28] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[29] Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

[30] Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. CARMO: dynamic criteria generation for context aware reward modelling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, volume ACL 2025 of *Findings of ACL*, pages 2202–2261. Association for Computational Linguistics, 2025.

[31] Shuxian Fan, Seonwoo Min, Youna Hu, Botao Xia, Jayakrishnan Unnikrishnan, Rowan Musselmann, Yifan Gao, Qingyu Yin, Priyanka Nigam, and Bing Yin. Shopping reasoning bench: An expert-authored benchmark for multi-turn conversational shopping assistants. *CoRR*, abs/2606.12608, 2026.

[32] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[33] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[34] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[35] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.