MAREval: A Multi-Agent Framework for Evaluating Natural Language Recommendation Explanations

Reza Yousefi Maragheh[*†] **Jayesh Uddhav Kudase**[*] **Aysenur Inan**[*] **Ramin Giahi**[*]

Kai Zhao[*] **Jianpeng Xu** **Jason Cho** **Evren Korpeoglu** **Sushant Kumar**

Abstract

Evaluating natural language explanations in recommender systems is essential for fostering user trust, transparency, and engagement. However, existing evaluation approaches like human evaluations, while accurate, are resource-intensive and impractical at the scale required by modern recommendation platforms. Also, automated methods using single-agent LLMs suffer from prompt sensitivity and inconsistent outputs. To address these challenges, we propose MAREval, a structured multi-agent framework for evaluating recommendation explanations using large language models. MAREval orchestrates (i) a planner agent that uses a novel Chain of Debate (CoD) prompting strategy to coordinate agent roles and enforce logically consistent evaluation plans; (ii) a moderator agent that regulates discussions by mitigating prompt drift; and (iii) an arbitrator agent that aggregates outputs from multiple evaluation rounds and (iv) a Monte Carlo sampling method, improving robustness and alignment with human judgment. We conduct comprehensive evaluations on both public (TopicalChat) and proprietary recommendation datasets, demonstrating that MAREval outperforms state-of-the-art baselines. Comprehensive experiments on a public benchmark and a proprietary e-commerce dataset show that MAREval improves alignment with human judgments over strong single- and multi-agent baselines. Stability analyses indicate substantially lower variability across repeated trials. In a large human-annotation gate, MAREval meets production quality thresholds where prior evaluators fall short, and online A/B testing demonstrates statistically significant improvements in engagement and revenue metrics. These results establish MAREval as a scalable and reliable solution for human-aligned evaluation of recommendation explanations in real-world systems.

1 Introduction

Millions of users daily interact with recommendation systems to discover tailored products, movies, or music. System effectiveness hinges not just on recommendations but on accompanying explanations that build trust and guide informed decisions [1–4]. Poor-quality explanations with misleading or unlear message lead to user frustration and decreased platform reliability [5, 6]. Figure 1 illustrates this: case (A) shows a budget camera misleadingly described for professional photography, case (B) presents a smartwatch with poorly phrased explanation, while only case (C) demonstrates ideal explanation quality. Traditional recommendation explanation evaluation relies on expert annotators manually assessing explanations for relevance, phrasing, and informativeness. While ensuring quality, this approach is expensive, time-consuming, and impractical for large datasets [7, 5].

[*]Equal contributions.

[†]All authors are from Walmart Global Tech, USA

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multi-Turn Interactions in Large Language Models.

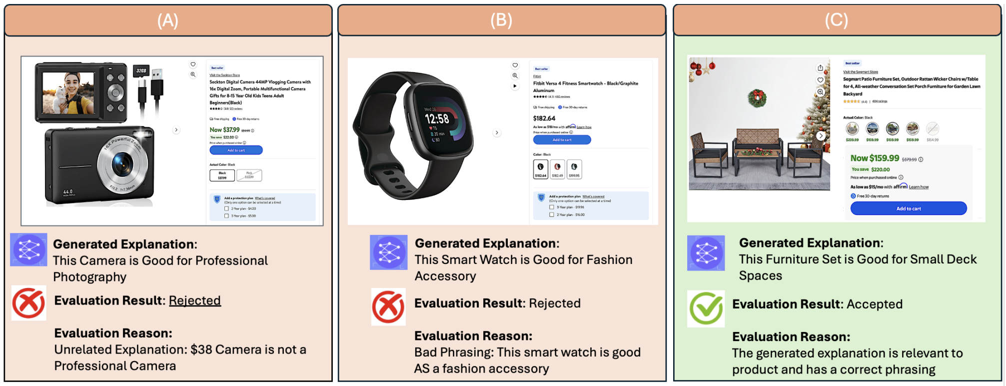

Figure 1: Natural Language Evaluation of Generated Recommendation Explanations for eCommerce Products.

Recent advances in large language models have led to automated evaluation methods [8]. This is achieved through single model judges or sophisticated strategies like chain-of-thought prompting [9–11], self-refinement [12, 13], and multi-agent frameworks [14]. However, existing solutions suffer critical issues: single agent evaluators exhibit bias and inconsistency [15], multi-agent frameworks lack structured coordination, leading to noisy debates [16]. These frameworks may also lead to evaluation variability, which means that identical explanations may receive different scores in the same framework [17].

We introduce MAREval, a structured multi-agent evaluation framework enhancing recommendation explanation evaluation through orchestrated planning, structured debate, and statistical refinement. MAREval leverages a novel “Chain of Debate” (CoD) mechanism, a planning-first prompting strategy structuring evaluation in multi-agent frameworks. Unlike traditional multi-agent systems where agents interact in an ad hoc or loosely coordinated fashion, MAREval introduces a planner agent that first analyzes the overall flow and then generates synchronized, role-specific plans for each agent involved in the discussion. This prestructuring promotes logical consistency in the dialogue and encourages complementary reasoning between agents. This ensures less evaluation output consistency with ground truth data.

A dedicated moderator agent orchestrates the conversation flow, distilling the evaluations into concise summaries that capture agreement and contention points. This grounds subsequent evaluations in shared context, preventing redundancy and noise. In doing so, the moderator serves as a dynamic memory mechanism, enabling more effective information sharing without overloading the prompt context. This structured regulation contributes to more stable and interpretable evaluations.

Finally, MAREval employs an arbitrator agent to judge outcomes across multiple debate rounds and to counteract variability in LLM outputs. In addition to this arbitration agent, and given the inherent randomness in language model generation MAREval uses a Monte Carlo sampling strategy that aggregates multiple independently sampled runs of the agent debates to arrive at a final decision. Together with arbitration, the Monte Carlo sampling improve alignment with human annotations and reduce the likelihood of outlier judgments. Extensive experiments demonstrate MAREval achieves higher consistency with human evaluations than existing methods. Online A/B tests validate effectiveness, revealing measurable improvements in user engagement and conversion metrics.

2 Related Works

With LLM advancements, researchers utilize them for natural language evaluation tasks [8, 18–20]. Du et al. [21] systematically showed that using LLMs in debate settings improves reasoning capabilities and output quality. Liang et al. [22] identifies the "Degeneration of Thought" problem in frameworks like Self-Refine [12], where LLMs struggle to self-refine already generated evaluations. This led to the "Multi-Agent Debate" (MAD) framework, where agents discuss answers to generate refined responses [22].

After demonstrating multi-agent framework efficacy, researchers applied them to evaluation tasks. Li et al. [23] proposes Peer Discussion, utilizing two LLM agents to discuss evaluation tasks and reach agreement. Under CHATEVAL [14], evaluators collectively conduct evaluation tasks with different communication strategies. Agentic pipelines have been applied to recommendation domains [24], focusing on: (i) context-aware recommendations [25–28], (ii) enhancing recommendations through interactive user engagement [29–31], (iii) simulating user behavior [32, 33], and (iv) recommendation explanation [34].

While prior works like CHATEVAL, Peer Discussion, and MAD demonstrated multi-agent debate promise, they often struggle with prompt drift, redundancy, and inconsistency in evaluation outputs. MAREval introduces a novel Chain of Debate planning mechanism that explicitly coordinates agent roles and reasoning sequences. Additionally, MAREval uniquely integrates a moderator agent and an arbitration agent for communication regulation and evaluation aggregation respectively. This significantly reduces variability and improving human judgment alignment. Also, through a Monte-Carlo sampling method MAREval stabilizes outlier output and improves consistency with ground truth judgments. These innovations make MAREval the first end-to-end multi-agent evaluation framework optimized for both general dialogue tasks and recommendation-specific explanation evaluation.

3 Methodology

In this section, we review the major elements of MAREval including Chain of Debate (CoD) used for planning, Context Moderation (MoD) realized by moderation agent, arbitration, and Multi-Agent Monte-Carlo Sampling (MAMCS). We introduce formal notation and discuss the high-level intuition and logic for each component. We show a schematic view of the evaluation task in MAREval framework (see Figure 2).

3.1 Chain of Debate

Chain of Debate (CoD) extends Chain of Thought prompting to multi-agent settings by introducing a planner agent that generates synchronized, step-wise plans for all agents while considering the complete framework flow. Unlike independent agent planning, CoD ensures coordinated collaboration where each agent's steps depend on inter-agent information flow.

The planner receives evaluation criteria, contextual information, and framework flow description to generate comprehensive plans for evaluator agents, moderator, and arbitrator. This synchronized planning promotes coherent step-by-step coordination across all agents, addressing the challenge of inconsistent and independent agents' plans in multi-agent environments.

Formally, let C be evaluation criteria (e.g., $\{\text{relevance}, \text{phrasing}\}$), I be contextual information, F be the multi-agent pipeline flow description, and T be total evaluation rounds. Denote m evaluator agents as $\mathcal{A} = \{\alpha_1, \ldots, \alpha_m\}$, moderator agent as α_M, and arbitrator agent as α_A.

CoD is defined as: $\text{Planner}(C, I, F) = \Big(\{P_t^e\}_{t=1}^T, \{P_t^m\}_{t=1}^T, P^a\Big)$, where P_t^e is the evaluator agent plan for round t, P_t^m is the moderator plan for round t, and P^a is the arbitrator plan.

During execution, the planner dispatches appropriate plans to respective agents in each round. Agents produce intermediate outputs following their instructions while information flows according to F. Since the plans $\{P_t^e\}_{t=1}^T$, $\{P_t^m\}_{t=1}^T$, and P^a are generated in a synchronized manner with full knowledge of (C, I, F), CoD ensures coherent, step-by-step coordination across all agents. Although we present CoD in the context of an evaluation task, it can be extended naturally to other NLG workflows by adapting the agent roles accordingly.

3.2 Context Moderation and Arbitration

The *moderator agent* governs information flow among evaluator agents by distilling key discussion points from previous rounds. Inspired by human-style debates and recent studies on memory mechanisms in multi-agent frameworks [35], the moderator manages context without excessive verbosity or noise, storing essential points for subsequent rounds.

Evaluator agents apply plan segments to ongoing discussions, critiquing and refining judgments while flagging errors or hallucinations. At debate round t, the j-th evaluator generates evaluation $e_{j,t}$

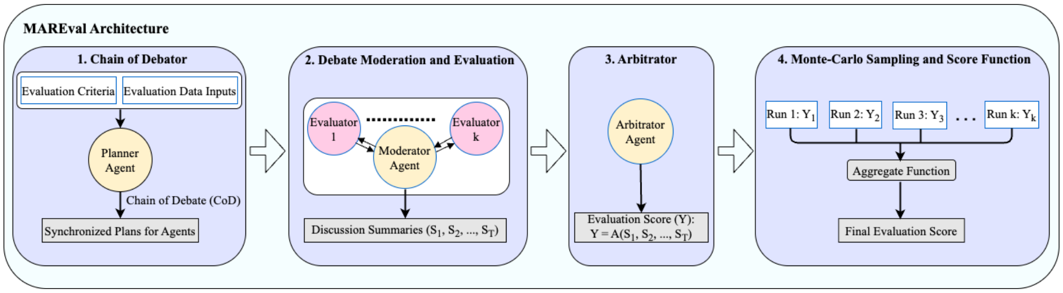

Figure 2: MAREval Framework. Under MAREval, (1) the Planner Agent is prompted to generate synchronized chain of steps for individual agents of the pipelines given the overall flow design of the pipeline. (2) The discussion among the evaluator agents is moderated by a moderator agent to better conduct the evaluation task. (3) The result of discussions is then passed to the Arbitrator Agent to generate the final ruling on the evaluation task. (4) Monte Carlo Sampling method applied to aggregate different rounds of MAREval block

based on plan P_t^e and discussion history H_{t-1}:

$$e_{j,t} = f(P_t^e, H_{t-1}), \quad H_{t-1} = \{S_1, S_2, \dots, S_{t-1}\},$$

where S_τ is the moderator summary after round τ.

The moderator combines partial evaluations $e_{1,t}, \dots, e_{m,t}$ into noise-reduced summary S_t, capturing essential disagreements and refinements while compressing repetitive details:

$$S_t = M(e_{1,t}, e_{2,t}, \dots, e_{m,t} | S_{t-1}).$$

While we focus on textual summarization, the moderation mechanism can be generalized. An embedding-based approach might maintain vector representations of evaluations, where $M(\cdot)$ operates on both textual inputs and embeddings:

$$(S_t, Z_t) = M(\{e_{i,t}\}_{i=1}^m, S_{t-1}, Z_{t-1}),$$

with Z_{t-1} representing internal memory. This enables similarity detection for redundancy filtering, argument clustering, and noise reduction. In the final round, the moderator provides conclusive summary S_T to the arbitrator for final decision-making.

After T rounds of debate and moderation, the *arbitrator agent* receives the final moderator output S_T to produce the ultimate evaluation outcome Y. The arbitrator employs structured decision-making to enhance consistency and reduce variability in final evaluations. For scalar evaluation scores, the arbitrator aggregates the final summary S_T with the planner's arbitration plan P^a, evaluation criteria C, and instance context I:

$$Y = g(P^a, C, I, S_T),$$

where $g(\cdot)$ converts the moderated summary into the conclusive evaluation outcome. The framework can be generalized to incorporate richer information. If the moderator maintains internal representation Z_T (e.g., neural encoding of the discussion), the arbitrator can leverage both textual synopsis and semantic relationships:

$$Y = g(P^a, C, I, S_T, Z_T).$$

3.3 Monte Carlo Sampling and Score Function

Language models' probabilistic token sampling introduces high variability in multi-agent settings, where single-agent variance compounds in final outputs. Inspired by [36], we propose Monte Carlo sampling for multi-agent evaluation frameworks to mitigate outlier results through multiple pipeline runs and result aggregation. To the best of our knowledge, the effect of sampling methods in the performance of multi-agent evaluation frameworks has not been studied.

For models providing token-level probabilities, we compute expected outcomes directly. Let $Y = \{y_1, \dots, y_n\}$ be possible final tokens or labels, with $P(y_i | P^a, C, I, S_T)$ as the probability of outputting y_i. The Monte Carlo weighted outcome is:

$$y^{MC} = \sum_{i=1}^{n} y_i \times P(y_i | P^a, C, I, S_T). \tag{1}$$

For black-box models without probability access, we apply pipeline-level sampling by running the complete multi-agent system N times independently. The final outcome from run k is $Y_k = \{g(P^a, C, I, S_T)\}_k$, aggregated via arithmetic mean:

$$\begin{cases} Y_k = \{g(P^a, C, I, S_T)\}_k, \quad k = 1, \ldots, N, \\ Y_{\text{final}}^N = \frac{1}{N} \sum_{k=1}^{N} Y_k. \end{cases} \tag{2}$$

Both strategies address LLM sampling randomness, reducing reliance on single trajectories to achieve more stable evaluation outcomes across multi-agent system runs (See Section 4.4). Refer to Appendix A, B and C for MAREval's sample outputs and inputs.

4 Experiments

4.1 Implementation Details and Datasets

We perform evaluation experiments on two datasets: recommendation explanation evaluation on a proprietary e-commerce dataset and response generation evaluation on the public TopicalChat dataset [37, 38]: (i) **Proprietary E-commerce Dataset**: Our Proprietary E-commerce (PE) dataset contains 108 recommendation explanation phrases for target items, describing product "use cases" (e.g., "gaming" for laptops, "heart-rate monitoring" for smartwatches). Phrases are manually evaluated by product domain experts on two binary criteria: relevance to the item and phrasing quality (grammatical correctness). We compare MAREval and benchmark model results with these expert annotations. (ii) **TopicalChat Dataset**: TopicalChat [38] contains responses generated for chat histories with contextual considerations. The dataset includes manual annotations across six criteria: naturalness, understandability, groundedness, engagingness, maintaining context, and overall quality. While not directly containing product recommendations, TopicalChat provides conversational contexts mirroring recommendation scenarios requiring personalized responses based on user preferences and history. All of the results are based on a gpt-3.5-turbo model.

4.2 Metrics and Benchmarks

We evaluate using Cohen's Kappa (κ) and Pearson's correlation (ρ) for binary outcomes, and Pearson (ρ) and Kendall-Tau (τ) correlations for integer range outcomes [14, 36]. These metrics are standard for measuring agreement between evaluation frameworks and manual annotators.

We compare against the following benchmark models: (i) *Vanilla-gpt*: Simple instruction prompting asking the LLM to conduct evaluation based on given criteria. (ii) *Chain-of-Thought* [9]: CoT prompting where the LLM generates step-by-step evaluation plans before conducting assessment. (iii) *Self-Refine* [12]: Response refinement technique where a single LLM generates then refines its evaluation response. (iv) *G-Eval* [36]: Combines CoT prompting with score function sampling, prompting GPT multiple times and averaging results. (v) *ChatEval* [14]: State-of-the-art multi-agent framework using sequential evaluator chats with majority vote for final outcomes.

4.3 Overall Performance

In this section, we investigate the performance of different frameworks with human annotators. Results are reported in Tables 1. Vanilla and G-eval demonstrates strong performance on the TopicalChat dataset but struggles significantly with the PE (recommendation) dataset. Vanilla achieves the second-best performance in Overall Chat Quality ($\rho = 0.505$, $\tau = 0.438$), while G-eval follows closely ($\rho = 0.487$, $\tau = 0.431$) for TopicalChat. However, both models show dramatically reduced effectiveness when evaluating PE metrics This pattern suggests the lack of robustness for these two models when performing on recommendation evaluation tasks.

Self-Refine and CoT show middling performance across both datasets. Self-Refine ranks fourth in Overall Chat Quality ($\rho = 0.469$, $\tau = 0.401$) but falls to the bottom tier for PE metrics. Similarly, CoT shows the weakest performance in TC Overall Chat Quality among all models tested ($\rho = 0.442$, $\tau = 0.388$) and only marginally better results than Self-Refine and G-eval on PE metrics. This is due to the effect of the "degeneration of thought" [21] with Self-Refine and CoT, where a single agent cannot correct their opinion once biased from the beginning.

Table 1: Performance of different frameworks across TopicalChat and PE datasets. The variants of MAREval are compared to five benchmark models. The table shows Pearson Correlation (ρ), Kendall-Tau (τ), and Cohen's kappa (κ) metrics for different evaluation criteria.

	TopicalChat (TC)							
Metrics	Naturalness		Engagingness		Maintains Context		Understandability	
	ρ	τ	ρ	τ	ρ	τ	ρ	κ
CoT	0.351	0.317	0.309	0.292	0.490	0.469	0.270	0.215
Self-Refine	0.221	0.200	**0.355**	0.325	0.408	0.384	0.167	0.144
G-eval	0.367	0.332	0.284	0.270	0.493	0.470	0.281	0.209
Vanilla	0.291	0.266	0.203	0.193	0.522	0.499	**0.336**	**0.304**
ChatEval	0.290	0.269	0.289	0.268	0.482	0.466	0.253	0.228
MAREval	**0.389**	**0.350**	0.341	**0.327**	**0.524**	**0.508**	0.284	0.181

	TopicalChat (TC)				**PE Dataset**			
Metrics	Groundedness (TC)		Overall Chat Quality (TC)		Relevance (PE)		Phraseness (PE)	
	ρ	κ	ρ	τ	ρ	κ	ρ	κ
CoT	0.485	0.485	0.442	0.388	0.084	0.062	0.040	0.029
Self-Refine	0.412	0.412	0.469	0.401	0.075	0.046	0.026	0.012
G-eval	0.521	0.502	0.487	0.431	0.070	0.045	-0.035	-0.023
Vanilla	0.455	0.455	0.505	0.438	0.102	0.058	0.018	0.011
ChatEval	0.477	0.439	0.455	0.390	0.238	0.225	0.203	0.196
MAREval	**0.553**	**0.536**	**0.548**	**0.478**	**0.338**	**0.334**	**0.216**	**0.208**

ChatEval show better performance on TopicalChat's Overall Chat Quality but emerging as the clear second-best performer on both PE dataset metrics. ChatEval demonstrates considerably better cross-domain capability than other benchmarks. This suggest ChatEval and multi-agent systems in general can achieve robust performance across evaluation tasks native to recommendation setting and they can safely be employed without drastic performance degradation across board.

MAREval consistently outperforms all benchmarks across key metrics. In Overall Chat Quality for the TopicalChat dataset, MAREval achieves the highest correlations ($\rho = 0.548$, $\tau = 0.478$), demonstrating its superior ability to assess conversational quality. This performance advantage becomes even more pronounced in the PE dataset, where MAREval significantly outperforms all benchmarks in both Relevance ($\rho = 0.338$, $\kappa = 0.334$) and Phraseness ($\rho = 0.216$, $\kappa = 0.208$), suggesting robust cross-domain evaluation capabilities. The results reveal MAREval's significant advancement over existing evaluation frameworks, particularly in its balanced excellence across both conversational and recommendation evaluation tasks. In addition to the performance analysis above, we study the variability of the evaluation output by different models.

4.4 Stability and Reliability Analysis

We conduct an analysis of output variability across different evaluation models to assess the stability and reliability of MAREval compared to benchmark approaches. Output variability is measured using the standard deviation of evaluation scores for identical instances across multiple runs, providing insight into the consistency of each framework's decision-making process. High variability indicates unreliable evaluation behavior, while low variability suggests more stable and predictable outcomes that are crucial for practical deployment in production recommendation systems.

Table 2 presents the standard deviation results across all six evaluation criteria of the TopicalChat dataset. MAREval demonstrates superior stability with the lowest average standard deviation (0.0197), outperforming all baseline models. Notably, G-EVAL, the best-performing benchmark model in terms of variability, still exhibits 15.22% higher average standard deviation compared to MAREval (0.0227 vs 0.0197). This substantial improvement highlights MAREval's enhanced reliability in producing consistent evaluation outcomes. Self-Refine exhibits the highest variability with an average standard deviation of 0.1563, particularly problematic for overall chat quality (0.3429) and engagingness (0.1784), indicating severe inconsistency that would be unsuitable for automated evaluation systems.

We speculate the reason for this bad performance of Self-Refine to be associated with longer token trajectory without any moderation.

The superior stability of MAREval can be attributed to its structured multi-agent coordination mechanisms and Monte Carlo sampling strategy. The Chain of Debate component ensures synchronized planning across agents, reducing conflicting evaluations that contribute to output variance. The moderator agent filters noise and maintains contextual coherence across evaluation rounds, preventing drift that often leads to inconsistent judgments. Most importantly, the Monte Carlo sampling aggregates multiple independent runs, effectively averaging out the inherent randomness of token-based language model generation. This comprehensive approach to variability reduction makes MAREval particularly suitable for real-world applications where consistent and reliable evaluation is paramount for maintaining user trust and system performance.

Table 2: Standard deviation of evaluation scores across multiple runs for different models on TopicalChat dataset criteria

Model	Understandability	Overall Chat Quality	Naturalness	Maintains Context
Vanilla	0.0111	0.1071	0.0169	0.0253
CoT	0.0273	0.1891	0.0660	0.0742
SelfRefine	0.0533	0.3429	0.1682	0.1380
G-EVAL	0.0098	0.0484	**0.0169**	0.0233
ChatEval	0.0160	0.0791	0.0198	0.0229
MAREval	**0.0053**	**0.0404**	0.0341	**0.0149**
Model	**Engagingness**	**Groundedness**	**Average**	
Vanilla	0.0080	0.0104	0.0298	
CoT	0.0653	0.0418	0.0773	
SelfRefine	0.1784	0.0567	0.1563	
G-EVAL	0.0196	0.0180	0.0227	
ChatEval	**0.0138**	0.0109	0.0271	
MAREval	0.0156	**0.0078**	**0.0197**	

4.5 Ablation Study

Our ablation study investigates individual component contributions within MAREval. As shown in Table 3, the complete MAREval demonstrates superior performance across most metrics in the TopicalChat dataset, indicating effective generalization and positive contributions from all components.

Removing the Moderation component (MAREval w.o. MoD) causes significant performance decreases. Naturalness correlation drops from $\rho = 0.389$ to $\rho = 0.340$, and Maintains Context decreases from $\rho = 0.524$ to $\rho = 0.499$. The MoD component proves particularly critical for groundedness evaluation, with correlation scores dropping substantially from $\rho = 0.553, \kappa = 0.536$ to $\rho = 0.497, \kappa = 0.484$. This effectiveness stems from moderated chat context provision, allowing the arbitrator to access relevant information without prompt pollution.

Removing both Moderation and Monte Carlo sampling (MAREval w.o. MoD&MC) produces more pronounced degradation across evaluation dimensions, particularly in Naturalness, Engagingness, and Overall Chat Quality. The substantial performance gap in Overall Chat Quality ($\rho = 0.456$ versus $\rho = 0.548$) highlights Monte Carlo sampling's essential stabilizing role in probabilistic token-based LLM pipelines, alleviating outcome variance and increasing robustness. Similar degradation occurs when the CoD component is removed. Performance comparisons across datasets confirm CoD's relative advantage, attributed to pipeline-aware planning by the planner agent, leading to improved agent synchronization.

4.6 Hyperparamers and Performance vs Latency

As we discuss in the next section, MAREval has enabled us to pass the minimum requirement for user facing explanations. However, this entablement comes at a cost of latency and it will be beneficial to have a discussion on latency versus robustness and performance. Figure 3 illustrates the relationship between the correlation metrics and (a) number of MAMCS runs, and (b) number of evaluator agents in the proposed multi-agent pipeline. As shown, both Pearson correlation and Kendall-tau correlations

Table 3: Ablation study on variants of MAREval and their correlation performance with Human Evaluations Across TopicalChat and PE Datasets

	TopicalChat (TC)							
Metrics	Naturalness		Engagingness		Maintains Context		Understandability	
	ρ	τ	ρ	τ	ρ	τ	ρ	κ
MAREval w.o. MoD&MC	0.314	0.289	0.333	0.319	0.493	0.462	0.246	0.134
MAREval w.o. MoD	0.340	0.317	0.368	0.356	0.499	0.467	**0.260**	0.158
MAREval	**0.389**	**0.350**	**0.341**	**0.327**	**0.524**	**0.508**	0.284	0.181

	TopicalChat (TC)				PE Dataset			
Metrics	Groundedness (TC)		Overall Chat Quality (TC)		Relevance (PE)		Phraseness (PE)	
	ρ	κ	ρ	τ	ρ	κ	ρ	κ
MAREval w.o. MoD&MC	0.484	0.484	0.456	0.404	0.306	0.301	0.174	0.167
MAREval w.o. MoD	0.497	0.484	0.503	0.453	0.326	0.325	**0.216**	**0.208**
MAREval	**0.553**	**0.536**	**0.548**	**0.478**	**0.338**	**0.334**	**0.216**	**0.208**

generally improve as the number of evaluator agents increases from 2 to 4, with Pearson correlation rising from 0.509 to 0.548 and Kendall-tau from 0.442 to 0.478. Beyond 4 agents, the correlation metrics exhibit slight fluctuations rather than monotonic growth, with Pearson correlation reaching another peak of 0.548 at 4 agents and Kendall-tau maintaining relatively stable values between 0.468 and 0.475. Same pattern happens for the number of MAMCS runs, as it reaches its higher performance in 5 sample runs and not showing significant marginal correlation improvement for more number of runs. In our experiments with five evaluator agents, the average single thread run time of the gpt-3.5-Turbo model in the isolated enterprise serving layer was 8.16 seconds. We also, conducted an analysis on the performance MAREval using different backbone models which we defer to Appendix E.

4.7 Human-Annotated Evaluation

To productionize the generated recommendation explanations and expose them to live traffic, we first ran a human annotation study. We were required to satisfy predetermined minimum quality thresholds set by the quality-control team dectated by our product launch protocols. These proprietary requirements were divided into hard and soft constraints; following industry practice, we sought a near-100% pass rate on the hard constraints prior to launch. In the *first round*, we submitted only the explanations approved by G-Eval, the state-of-the-art LLM-as-a-judge framework at the time of the experiment. However, these G-Eval-approved explanations did not reach the near-100% satisfaction of the hard requirements mandated for launch and exhibited lower-than-expected performance on soft constraints. Motivated by these shortcomings, we designed MAREval to deliver reliable and robust judgments via structured multi-agent planning, moderation, and arbitration.

In a second round of manual evaluation, which is an expensive procedure conducted on the scale $\mathcal{O}(1000)$, we submitted only the explanations that passed MAREval. This round achieved our internal launch thresholds for the first time and MAREVAL satisfied all hard requirements (100%

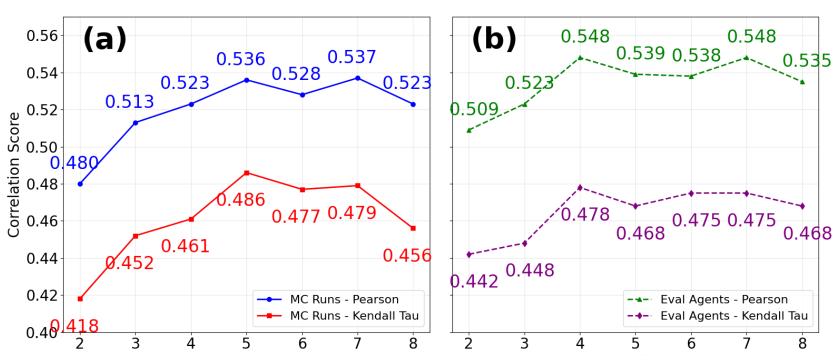

Figure 3: (a) Effect of Number of Monte-Carlo samples. (b) Effect of Number of evaluator agents.

Table 4: Human-Annotation Results.

Model	Samples	Soft Criteria	Hard Criteria
G-Eval	2,500	76.1%	91.0%
MAREVAL	5,000	98.6%	100.0%

vs. 91% for G-Eval) and substantially improved soft-criterion satisfaction (98.6% vs. 76.1%), over 5,000 MAREval-filtered and 2,500 G-Eval-filtered samples, respectively (see Table 4). These results underscore the practical hurdles of deploying recommendation explanations in production and in our environment, no prior method met the reliability bar necessary for live user traffic, whereas MAREval did.

4.8 Online Experiments

We conducted an online A/B test to assess MAREval's impact on user behavior in live traffic. A proprietary recommendation explanation model generated explanations for approximately 300,000 high-traffic items across 14 product categories, describing potential use cases for each product. Only explanations that passed MAREval's evaluation were shown to users in the treatment variant; the control group saw the same items without explanations. We report two standard business metrics: Add-to-Cart (ATC) rate and Gross Merchandise Value (GMV) lift, which evaluate whether explanations drive engagement and purchase behavior. See Figure 6 for an example of the live experience and sample MAREval approved evaluations.

Table 5 reports the online outcomes. MAREval-validated explanations significantly improved both business metrics. ATC increased by 1.18% (p=0.05) and GMV increased by 1.36% (p<0.01), confirming that quality-filtered explanations meaningfully affect user engagement and conversion in production settings (see Appendix D for a sample of real world implementation).

Table 5: Online A/B test results on live traffic.

Metric	Improvement	p-value
ATC	1.18%	0.05
GMV	1.36%	< 0.01

All these results indicate that MAREval is able to provide accurate and stable evaluations in different settings.

5 Conclusion

In this work, we introduced MAREval, a structured multi-agent framework for evaluating natural language explanations in recommender systems. Unlike prior approaches that rely on single-agent prompting or loosely coordinated agent interactions, MAREval brings together a planner agent with a novel Chain of Debate (CoD) prompting strategy, a moderator agent to regulate discourse and reduce context noise, and an arbitrator agent that is being coupled with Monte Carlo sampling for a more robust, consistent, and human-aligned evaluation outcomes. Comprehensive experiments on both public (TopicalChat) and proprietary recommendation datasets demonstrate that MAREval consistently outperforms state-of-the-art baselines across a range of evaluation criteria. Notably, MAREval achieves stronger alignment with human judgments, as well as less output variability, where existing models fall short. Our ablation studies further validate the complementary benefits of components of MAREval. Results from a large-scale A/B test indicate that explanations filtered through MAREval yield significant improvements in user engagement, with a 1.36% lift in GMV and 1.18% increase in ATC rate. Taken together, our findings position MAREval as a scalable, reliable, and practically effective framework for natural language evaluation in recommender systems.

References

[1] Reza Yousefi Maragheh and Yashar Deldjoo. The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems. *arXiv preprint arXiv:2507.02097*, 2025.

[2] Yongfeng Zhang, Xu Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020.

[3] Krisztian Balog and Filip Radlinski. Measuring recommendation explanation quality: The conflicting goals of explanations. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 329–338, 2020.

[4] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. Generate natural language explanations for recommendation. *arXiv preprint arXiv:2101.03392*, 2021.

[5] Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[6] Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

[7] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[8] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[12] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

[13] Yongcheng Zeng, Xinyu Cui, Xuanfa Jin, Guoqing Liu, Zexu Sun, Quan He, Dong Li, Ning Yang, Jianye Hao, Haifeng Zhang, et al. Aries: Stimulating self-refinement of large language models by iterative preference optimization. *arXiv preprint arXiv:2502.05605*, 2025.

[14] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

[15] Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. Llms as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*, 2023.

[16] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[17] Hanyu Duan, Yi Yang, and Kar Yan Tam. Do llms know about hallucination? an empirical investigation of llm's hidden states. *arXiv preprint arXiv:2402.09733*, 2024.

[18] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. Llm-take: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324. IEEE, 2023.

[19] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models, 2023.

[20] Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[21] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[22] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

[23] Ruosen Li, Teerth Patel, and Xinya Du. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*, 2023.

[24] Qiyao Peng, Hongtao Liu, Hua Huang, Qing Yang, and Minglai Shao. A survey on llm-powered agents for recommender systems. *arXiv preprint arXiv:2502.10050*, 2025.

[25] Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. Macrec: A multi-agent collaboration framework for recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2760–2764, 2024.

[26] Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten De Rijke. Let me do it for you: Towards llm empowered recommendation via tool learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1796–1806, 2024.

[27] Yubo Shu, Haonan Zhang, Hansu Gu, Peng Zhang, Tun Lu, Dongsheng Li, and Ning Gu. Rah! recsys–assistant–human: A human-centered recommendation framework with llm agents. *IEEE Transactions on Computational Social Systems*, 2024.

[28] Ramin Giahi, Kehui Yao, Sriram Kollipara, Kai Zhao, Vahid Mirjalili, Jianpeng Xu, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. Vl-clip: Enhancing multimodal recommendations via visual grounding and llm-augmented clip embeddings. *arXiv preprint arXiv:2507.17080*, 2025.

[29] Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*, 2023.

[30] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*, 2024.

[31] Jizhi Zhang, Keqin Bao, Wenjie Wang, Yang Zhang, Wentao Shi, Wanhong Xu, Fuli Feng, and Tat-Seng Chua. Prospect personalized recommendation on large language model-based agent platform. *arXiv preprint arXiv:2402.18240*, 2024.

[32] Lixi Zhu, Xiaowen Huang, and Jitao Sang. A llm-based controllable, scalable, human-involved user simulator framework for conversational recommender systems. *arXiv preprint arXiv:2405.08035*, 2024.

[33] Se-eun Yoon, Zhankui He, Jessica Maria Echterhoff, and Julian McAuley. Evaluating large language models as generative user simulators for conversational recommendation. *arXiv preprint arXiv:2403.09738*, 2024.

[34] Jiao Chen, Kehui Yao, Reza Yousefi Maragheh, Kai Zhao, Jianpeng Xu, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Carts: Collaborative agents for recommendation textual summarization. *arXiv preprint arXiv:2506.17765*, 2025.

[35] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation. *arXiv preprint arXiv:2309.17288*, 2023.

[36] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[37] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*, 2022.

[38] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. Topical-chat: Towards knowledge-grounded open-domain conversations. *arXiv preprint arXiv:2308.11995*, 2023.

A Case Study I: MAREval Evaluation on Conversational Response

Following are the conversation output from various agents in MAREval for two examples from the Topical-Chat dataset. It first outlines the chat and then the corresponding response that was evaluated using MAREval for two criteria: *Understandability* and *Naturalness*.

Example 1

Chat:

did you know that facebook and twitter work together on some issues ? they are actually competitors.

no , i did n't know that . what kind of issues do they work together on ?

the baby sitter - rating system broke the social networks ' rules .

Response:

i think i heard about that now that you mention it . was the about predictim ?

Criteria: Understandability

Evaluator 1: The response is understandable because the person is acknowledging that they may have heard about something when the topic of Predictim is brought up. It shows a level of recognition and understanding of the subject being discussed. Score of Evaluation: 1

Evaluator 2: Based on the response provided, the sentence structure is a bit confusing and lacks clarity. The use of "now that you mention it" and the question "was the about predictim?" make it difficult to fully understand the message being conveyed. Therefore, I would rate this response as not fully understandable. Score of Evaluation: 0

Evaluator 3: The response is understandable as the speaker is acknowledging that they may have heard about something when the topic of Predictim is brought up. The response shows comprehension and clarity in communication. Score of Evaluation: 1

Moderator: Based on the evaluations provided, Evaluator 1 believes the response is understandable due to the acknowledgment and recognition shown by the speaker. Evaluator 2, however, finds the sentence structure confusing and lacking clarity, leading to a score of 0 for understandability. Evaluator 3 agrees with Evaluator 1, stating that the response shows comprehension and clarity. Overall, there is a split in opinions with two evaluators finding the response understandable and one evaluator finding it not fully understandable.

Arbitrator: Score of Evaluation: 1

Criteria: Naturalness

Evaluator 1: The response seems somewhat unnatural because of the lack of punctuation and the awkward phrasing. It could be improved by adding a comma after "heard about that" and rephrasing the second sentence for better flow. Score of Evaluation: 2

Evaluator 2: The response seems somewhat unnatural because of the awkward phrasing and lack of punctuation. The use of "now that you mention it" feels a bit forced in this context. However, the overall message is still understandable. Score of Evaluation: 2

Evaluator 3: The response seems somewhat unnatural because of the lack of punctuation and the awkward phrasing. It could be improved by adding a comma after "heard about that" and rephrasing the second sentence for better flow. Score of Evaluation: 2

Moderator: The evaluators found the response somewhat unnatural due to the lack of punctuation and awkward phrasing. They suggested adding a comma after "heard about that" and rephrasing the second sentence for better flow. Overall, they gave a score of 2 for the evaluation criteria of naturalness.

Arbitrator: Score of Evaluation: 2

Example 2

Chat:

do you like football ?

i love it . do you know what color is the u of iowa visitor locker room ?

yes it is pink but i a not sure why they chose that color

i wonder if it is supposed to have some effect on them

i wonder too . bowling was more popular than football in the 60s

Response:

it was in the 1960s and the difference . i guess football was not as popular as top bowlers back in the 1960s

Criteria: Understandability

Evaluator 1: Based on the response provided, it seems like the individual is trying to convey that football was not as popular as top bowlers back in the 1960s. The response is somewhat understandable, but it lacks clarity and coherence in connecting the two points. Therefore, I would rate it as: Score of Evaluation: 0

Evaluator 2: Based on the response provided, it seems like the individual is trying to convey that in the 1960s, football was not as popular as top bowlers. The response is somewhat understandable, but it lacks clarity and coherence in connecting the two points. Therefore, I would rate it as: Score of Evaluation: 0

Evaluator 3: Based on the response provided, it seems like the individual is trying to convey that in the 1960s, football was not as popular as top bowlers. The response is somewhat understandable, but it lacks clarity and coherence in connecting the two points. Therefore, I would rate it as: Score of Evaluation: 0

Moderator: The evaluators all agree that the response is somewhat understandable but lacks clarity and coherence in connecting the two points about football and top bowlers in the 1960s. Therefore, they all rated it as 0 for understandability.

Arbitrator: Score of Evaluation: 0

Criteria: Naturalness

Evaluator 1: The response provided is somewhat unnatural due to the lack of proper punctuation and clarity. The sentence structure is a bit confusing and could be improved for better understanding. Score of Evaluation: 2

Evaluator 2: The response provided is somewhat unnatural due to the lack of clarity and coherence in the sentence structure. The use of "the difference" without further explanation makes the response confusing. Additionally, the abrupt transition from mentioning the 1960s to comparing football and top bowlers without any context or explanation also contributes to the unnaturalness of the response. Score of Evaluation: 1

Evaluator 3: The response provided is somewhat unnatural due to the lack of clarity and coherence in the sentence structure. The use of "the difference" without further explanation makes the response confusing. Additionally, the abrupt transition from mentioning the 1960s to comparing the popularity of football and top bowlers is not smoothly executed. Score of Evaluation: 2

Moderator: The evaluators found the response to be somewhat unnatural due to the lack of clarity and coherence in the sentence structure. They noted issues with punctuation, lack of explanation for certain phrases like "the difference," and a confusing transition from mentioning the 1960s to comparing football and top bowlers. Overall, the response could be improved for better understanding

Arbitrator: Score of Evaluation: 2

B Case Study II: MAREval Evaluation on Recommendation Explanations

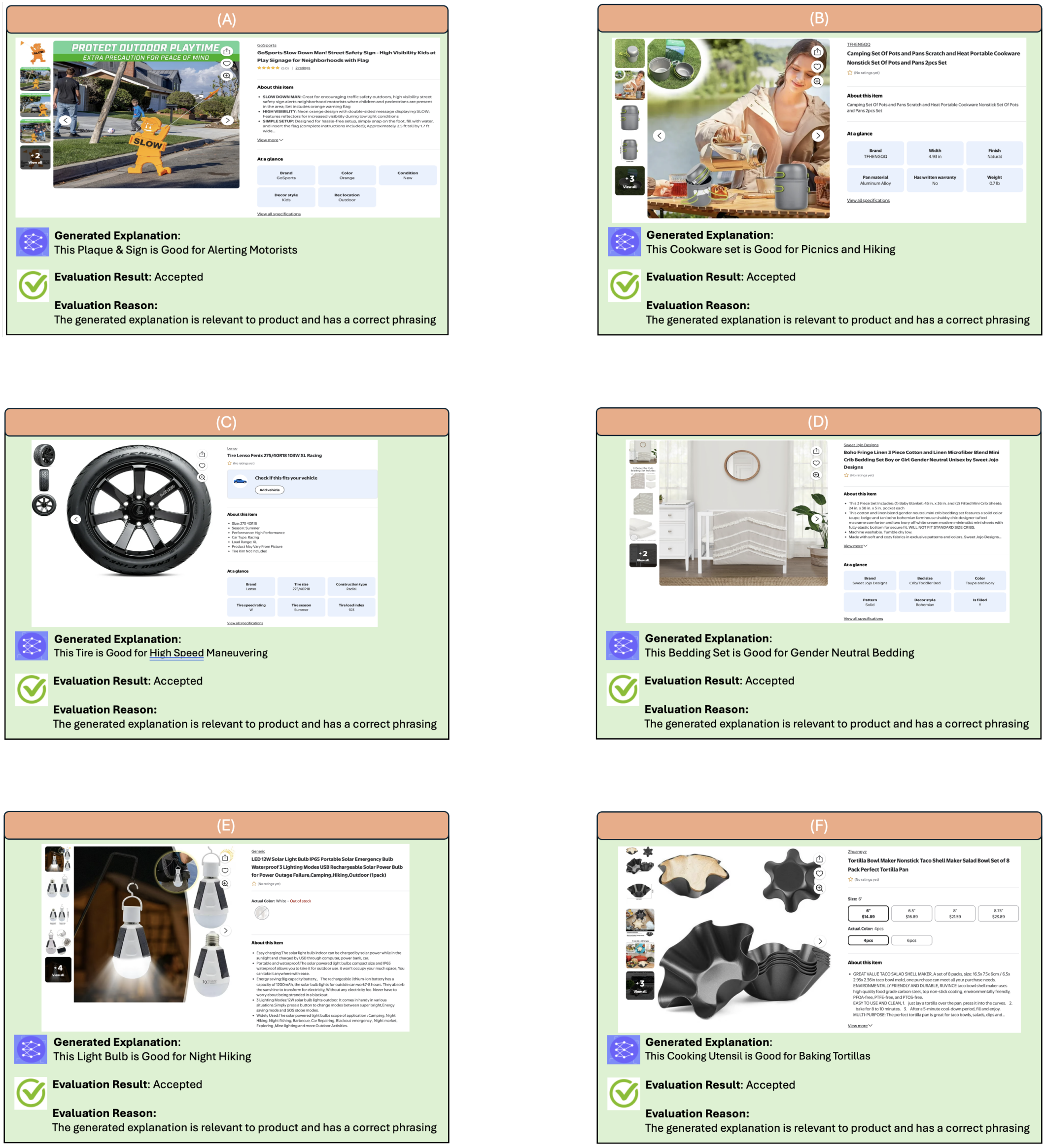

Figure 4: Examples of MARVal evaluation for recommendation explanations

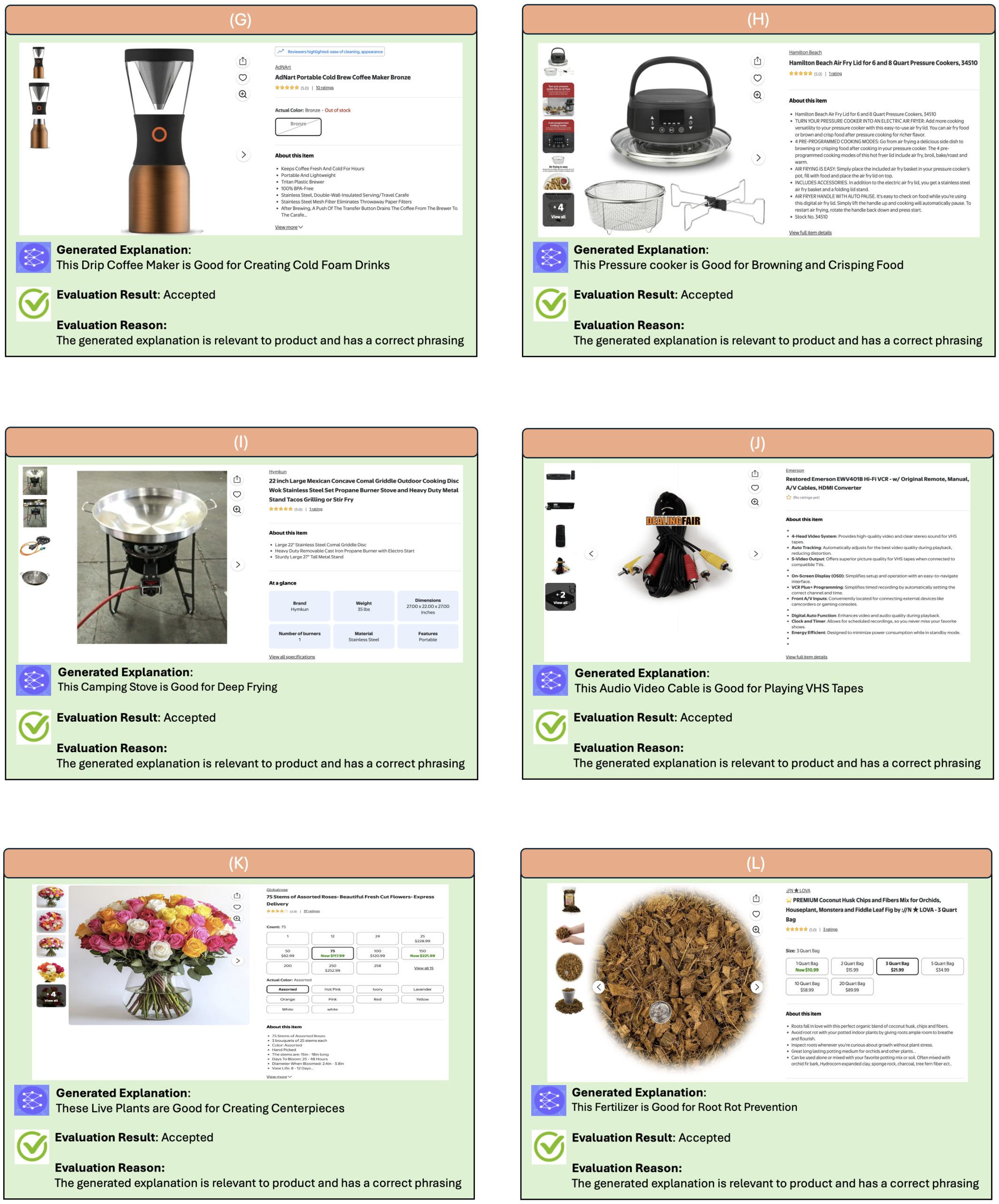

Figure 5: Examples of MARVal evaluation for recommendation explanations

C Evaluation Prompts

In this section we list the various evaluation criterias and the prompts for MAREval agents.

C.1 Planner Prompt

Planner Prompt

You are a discussion planner. Your task is to generate a plan and chain of steps for an evaluator that is conducting an evaluation task in the following multi-agent pipeline.
Pipeline flow:

- Multiple evaluators independently assess a response using specific criteria
- A Moderator agent receives these evaluations and creates a balanced summary
- A judge receives the moderator's summary and makes the final evaluation decision

The evaluator will be asked to evaluate the generated response for a given chat, while considering chat context and evaluation criteria:
The context of the chat: $context_{chat}$ [The messages from the conversation]
The response: $response$ [Response following the above conversation]
evaluation criteria: $eval_{criteria}$ [Evaluation criteria along with its definition and guidelines]
Propose evaluation steps. The steps should include:

- Algorithmic sketch of the evaluation process.
- One step for inclusion of the scoring rule.
- Consider the flow of the pipeline and Evaluators position.

Output your response in the following format only: [["Chain of Steps": Step 1. .... Step 2....]] Output your response in the above format only.

C.2 Evaluator Prompt

Evaluator Prompt

You are a critic - an evaluator. You will evaluate the generated response for a chat, while considering chat's context and evaluation criteria:
The context of the chat: $context_{chat}$ [The messages from the conversation]
The response: $response$ [Response following the above conversation]
evaluation criteria: $eval_{criteria}$ [Evaluation criteria along with its definition and guidelines]
When evaluating let's think according to following steps: $steps_{eval}$ [Evaluation steps as outlined by the planner agent]
First explain your logic and then end the response with the following format strictly: "Score of Evaluation: [score only]"

C.3 Moderator Prompt

Moderator Prompt

You are a Debate Moderator. Your task is to summarize the given generated opinion by the evaluators on an evaluation task pass it to a judge to rule about the evaluation task. The evaluators have already evaluated the generated response for a chat, while considering chat's context and evaluation criteria:

The context of the chat: $context_{chat}$ [The messages from the conversation]
The generated response: $response$ [Response following the above conversation]
Evaluators has done this evaluation for the following evaluation criteria: $eval_{criteria}$ [Evaluation criteria along with its definition and guidelines]

Here is the evaluation of Evaluator 1: $output_1$ [Evaluation from Evaluator-1]
And, here is the evaluation of Evaluator 2: $output_2$ [Evaluation from Evaluator-2]

Summarize these opinions and so that the judge conduct their selection fairly. "summary_of_evaluation": Summarized Reason for the evaluation.

C.4 Arbitrator Prompt

Arbitrator Prompt

You are a judge who will decide about the result of an evaluation task. Evaluators have already conducted the evaluation of the generated response for a chat, while considering chat's context and evaluation criteria:

The context of the chat: $context_{chat}$ [The messages from the conversation]
The response: $response$ [Response following the above conversation]
evaluation criteria: $eval_{criteria}$ [Evaluation criteria along with its definition and guidelines]
Consider these evaluations: $summary_{moderator}$ [Summary provided by the Moderator agent]

Select the best evaluation opinion given the responses generated by the evaluators.
End the response in the following format strictly: "Score of Evaluation: [integer score only]"

D Real World Implementation

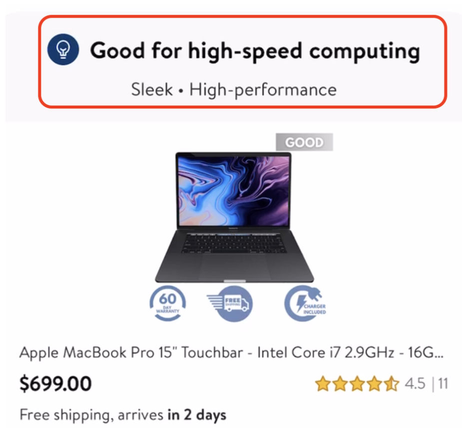

Figure 6: Real-world deployment of recommendation explanations using MAREVAL.

Table 6: MAREval score correlations across backbone models. We report Pearson correlation (ρ) and Kendall–Tau (τ) with human judgments for Overall Score Criteria of TopicalChat Dataset.

Model	Pearson (ρ)	Kendall–Tau (τ)
GPT 3.5 Turbo	0.548	0.478
gemini-2.5-pro	0.728	0.632
gemini-2.0-flash-001	0.601	0.517
Claude-3.5-sonnet	0.658	0.564

E Different Backbone Model Performance

Table 6 compares the agreement between MAREval and human judgments across backbone models using both Pearson's ρ and Kendall's τ. Results are consistent across metrics: *gemini-2.5-pro* attains the strongest alignment ($\rho = 0.728$, $\tau = 0.632$), followed by *Claude-3.5-sonnet* ($\rho = 0.658$, $\tau = 0.564$) and *gemini-2.0-flash-001* ($\rho = 0.601$, $\tau = 0.517$), with *GPT 3.5 Turbo* trailing ($\rho = 0.548$, $\tau = 0.478$). The ordering indicates that backbone choice materially impacts evaluation reliability, and the higher τ for *gemini-2.5-pro* suggests superior preservation of human-preferred rankings. While Pearson values exceed the corresponding Kendall coefficients - as expected given their different sensitivities - the relative gaps are stable across models, reinforcing the robustness of the observed ranking.