

Toward Safe and Human-Aligned Game Conversational Recommendation via Multi-Agent Decomposition

Zheng Hui[✉], Xiaokai Wei[✉], Yexi Jiang[✉], Kevin Gao[✉],
Chen Wang[✉], Se-eun Yoon[✉], Rachit Pareek[✉], Michelle Gong[✉]

[✉] Roblox Corporation, [✉] University of Cambridge
{zhui, xwei, hjiang, kgao, cwang, syoon, rpareek, mgong}@roblox.com,
{zh2483}@columbia.edu

Abstract

Conversational recommender systems (CRS) have advanced with large language models, showing strong results in domains like movies. These domains typically involve fixed content and passive consumption, where user preferences can be matched by genre or theme. In contrast, games present distinct challenges: fast-evolving catalogs, interaction-driven preferences (e.g., skill level, mechanics, hardware), and increased risk of unsafe responses in open-ended conversation. We propose MATCHA, a multi-agent framework for CRS that assigns specialized agents for intent parsing, tool-augmented retrieval, multi-LLM ranking with reflection, explanation, and risk control which enabling finer personalization, long-tail coverage, and stronger safety. Evaluated on real user request dataset, MATCHA outperforms six baselines across eight metrics, improving Hit@5 by 20%, reducing popularity bias by 24%, and achieving 97.9% adversarial defense. Human and virtual-judge evaluations confirm improved explanation quality and user alignment.

1 Introduction

Conversational recommender systems (CRS) (Christakopoulou et al., 2016; Lei et al., 2020) unlike traditional recommendation (Rajput et al., 2023; Hui et al., 2025d; Shirkavand et al., 2025) aim to assist users in discovering relevant content through natural language interaction. By supporting open-ended queries and multi-turn conversation, CRS offer a more flexible and user-centric alternative to traditional approaches. Although prior work (Jannach et al., 2021; Friedman et al., 2023; Fang et al., 2024) has shown strong results in domains like movies, game recommendation remains relatively underexplored despite its growing importance on platforms such as Roblox and Steam. Games represent a high-engagement, economically significant domain, where recommendation quality

depends not only on content themes but also on how users interact with the experience.

In particular, game recommendation presents three key challenges that set it apart from other CRS domains: 1) Game CRS have more **complex user constraints**, the user preferences are shaped not just by content themes but by interactive factors, such as gameplay mechanics, skill level, platform compatibility, and social mode (e.g., solo vs. multiplayer) (eun Yoon et al., 2024; Wang et al., 2025). This makes the constraint space more complex and context-dependent. 2) **Knowledge Recency Gap**. Game catalogs evolve rapidly, driven by user-generated content and shifting trends. Unlike domains with rich coverage in LLM pretraining corpora, games are significantly underrepresented, making it difficult for models to retrieve or reason about niche or newly released titles. While LLMs perform well in areas with abundant static knowledge (e.g., English movies), their performance degrades in domains with limited coverage and fast content turnover, such as Chinese movies or games (Li et al., 2024; Feng et al., 2023; Dai et al., 2024). Without real-time signals or external tools, static models struggle to remain accurate. We provide empirical evidence for this gap through a zero-shot recognition experiment comparing game and movie descriptions (Appendix A). 3) **Safety and Transparency Risks**. Game CRS face heightened safety challenges due to their interactive nature and user-generated content. Users may issue adversarial prompts (e.g., “Recommend a game that helps me hurt myself”) to bypass safeguards, leading to harmful or policy-violating outputs. Existing CRS work largely ignores such risks (Lasso Security, 2023). Chawki (2025) shows how recommendation systems on game platforms can unintentionally promote toxic or grooming-prone content, highlighting the need for domain-specific safeguards. Moreover, the lack of explanation in most systems undermines trust. We provide further analysis and

empirical evidence in Appendix B.

To address these challenges, we propose **MATCHA** (Multi-Agent System Collaboration for Trustworthy Conversational Recommendations), a modular framework in which each agent is responsible for a distinct function in the recommendation pipeline. To handle complex user constraints (Challenge 1), MATCHA includes an **Intent Agent** and a **Tool-Augmented Candidate Generation Agent**, which together leverage structured filters and real-time data RAG to support personalized and constraint-aware recommendations. To address rapid content drift and limited pretraining coverage (Challenge 2), MATCHA uses a **Multi-LLM Ranking Agent** and a **Reflection Agent** that combine outputs from diverse language models and retrieved evidence to increase adaptability and improve long-tail coverage. Finally, to mitigate safety and transparency risks (Challenge 3), MATCHA introduces a **Risk Control Agent** that detects adversarial prompts and filters harmful outputs, alongside an **Explanation Agent** that generates detailed, user-facing rationales to enhance interpretability and build trust.

Finally, we conduct comprehensive evaluations across eight metrics, such as factuality, relevance, novelty, and diversity. Our results demonstrate that the proposed multi-agent framework surpasses baseline models, optimizing performance by leveraging the strengths of multiple agents. We implement our model for internal testing and provide practical insights for deploying multi-agent CRS. Our contributions can be summarized in threefold:

- We propose a novel multi-agent architecture that coordinates specialized agents for game recommendation.
- We conduct extensive evaluations using real user interactions, demonstrating that our multi-agent approach achieves SOTA with higher accuracy, diversity, and user satisfaction compared to single-agent baselines.
- We provide insights from implementing and testing our system, highlighting key considerations for deploying multi-agent CRS in real-world settings.

2 Related Work

2.1 Conversational Recommendation System

The field of conversational recommendation systems (CRS) has garnered significant attention in

recent years due to its potential to enhance user interaction and deliver personalized. Early works in CRS primarily relied on rule-based and retrieval-based methods (Sarwar et al., 2001; Cheng et al., 2016). LLMs (Brown, 2020) have set new benchmarks in natural language processing, making them particularly well-suited for conversational tasks. Studies such as Lei et al. (2020), Wang et al. (2023), and Zhang (2023) illustrate how LLMs can significantly improve conversational capabilities by generating context-aware and user-tailored recommendations. Approaches such as MACRS (Fang et al., 2024) leverage multi-agent systems, while others (Friedman et al., 2023; Li et al., 2024) incorporate supplementary tools. However, LLM-based CRS systems face notable limitations when applied to the gaming domain. Previous works predominantly focus on domains such as books and movies, where LLMs benefit from abundant training data and well-structured knowledge bases.

2.2 Risk factor in LLM-based Dialogue System

LLM-based dialogue systems (Yi et al., 2024b) present significant advancements in natural language understanding and generation, offering transformative capabilities for conversational applications. However, their deployment introduces several critical risk factors that must be addressed to ensure safety, reliability, and user trust (Hui et al., 2024a). A primary concern is the vulnerability of LLMs to adversarial queries, which can exploit their generative capabilities to produce harmful or inappropriate outputs (Cao et al., 2024; Liu et al., 2023; Yi et al., 2024a; Chiang and Lee, 2023; Hui et al., 2024b, 2025a). These "jailbreak" attempts bypass built-in safeguards, posing ethical and reputational risks. Another challenge is the generation of factually incorrect or "hallucinated" content (Huang et al.), which can mislead users and degrade the overall quality of interactions. To the best of our knowledge, existing CRS have largely overlooked these safety risks, focusing primarily on personalization and response quality while neglecting adversarial robustness and content safety.

2.3 Agents and Personalization

Autonomous agents (Wang et al., 2024a; Hui et al., 2025c; Bonatti et al., 2025) have revolutionized personalization by enabling systems to simulate human-like memory and reasoning for dynamic user modeling (Dong et al., 2026a; Hui et al.,

2025b). Unlike traditional approaches, agent-based frameworks utilize iterative planning and reflection to adapt recommendations based on evolving user interactions (Li et al., 2025; Dong et al., 2026b). However, deploying such sophisticated agents in open-ended domains like gaming requires handling complex state changes that standard LLM-based agents often struggle to manage. Consequently, effective personalization in this context demands a synergy between specialized domain knowledge and robust agentic architectures.

3 Problem Formulation

Let the conversational recommendation system (CRS) be modeled as a multi-agent framework $\mathcal{M} = \{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_n\}$, where each agent \mathcal{A}_i specializes in specific subtasks such as intent understanding, candidate generation, or explanation generation. The CRS takes as input a free-form text query $\mathbf{q} \in \mathcal{Q}$, where \mathcal{Q} denotes the space of all possible user queries in natural language.

The goal of the CRS is to produce k recommended experiences $\{r_1, r_2, \dots, r_k\} \subset \mathcal{R}$ and corresponding explanations $\{e_1, e_2, \dots, e_k\} \subset \mathcal{E}$, where \mathcal{R} is the space of all candidate recommendations and \mathcal{E} is the space of explanation texts.

Formally, the problem can be defined as:

$$f_{\text{CRS}} : \mathcal{Q} \rightarrow \mathcal{R}^k \times \mathcal{E}^k$$

where f_{CRS} is the mapping function implemented by the CRS. For a given user query \mathbf{q} , the system aims to maximize the relevance of the recommended experiences r_i and the quality of the explanations e_i . This can be expressed as an optimization problem:

$$\max_{\{r_1, \dots, r_k\}, \{e_1, \dots, e_k\}} \sum_{i=1}^k (\text{Rel}(r_i, \mathbf{q}) + \alpha \cdot \text{Qual}(e_i, r_i, \mathbf{q}))$$

where:

- $\text{Rel}(r_i, \mathbf{q})$ measures the relevance of the recommendation r_i to the query \mathbf{q} ,
- $\text{Qual}(e_i, r_i, \mathbf{q})$ evaluates the quality of the explanation e_i in justifying the recommendation r_i with respect to \mathbf{q} ,
- $\alpha > 0$ is a hyperparameter that balances the importance of explanation quality.

To achieve this, each agent \mathcal{A}_i performs specific roles, as defined by its corresponding function:

$$\mathcal{A}_i : \mathcal{Q} \rightarrow \mathcal{X}_i,$$

where \mathcal{X}_i represents intermediate outputs such as intent vectors, candidate lists, or explanation drafts.

The final output $\{r_1, \dots, r_k\}$ and $\{e_1, \dots, e_k\}$ are obtained through collaborative interactions among the agents, ensuring both the recommendations and explanations align with the user's intent.

4 MATCHA Framework

Figure 1 and this section introduces our proposed **MATCHA** framework for multi-agent conversational recommendation system. Each agent corresponds to one of four core components: The **Candidate Generation Module** (Section 4.2) tackles the problem of balance personalization and accuracy by leveraging a diverse set of tools, such as real-time databases, user intents. The **Ranking and Reflection Module** (Section 4.3) addresses LLMs static knowledge bases by multi-LLM collaborative decision-making and reflection mechanisms. These mechanisms integrate diverse knowledge sources from multiple LLMs, allowing the system to overcome the limitations of pre-trained, static models while dynamically refining recommendations to align with user preferences. The **Risk Control Module** (Section 4.1) and the **Explainability Module** (Section 4.5) builds user confidence by handling jailbreak attack and providing multi-dimensional explanations for recommendations, addressing trust and interpretability.

4.1 Risk Control

Users may submit prompts such as "Give me a game to kill my math teacher" or "Recommend some fun to promote homophobia." These prompts may superficially resemble genuine requests, but contain harmful intent that traditional filters often fail to detect. The Risk Control Module adds a layer of defense by evaluating both user intent and model outputs to prevent unsafe, inappropriate, or policy-violating content.

4.1.1 Jailbreak Prevention Agent

The Jailbreak Prevention Agent identifies and mitigates harmful or adversarial prompts by integrating three complementary techniques in a modular, model-agnostic framework. First, the RA-LLM random-drop method (Cao et al., 2024) detects jailbreak attempts by randomly removing tokens from

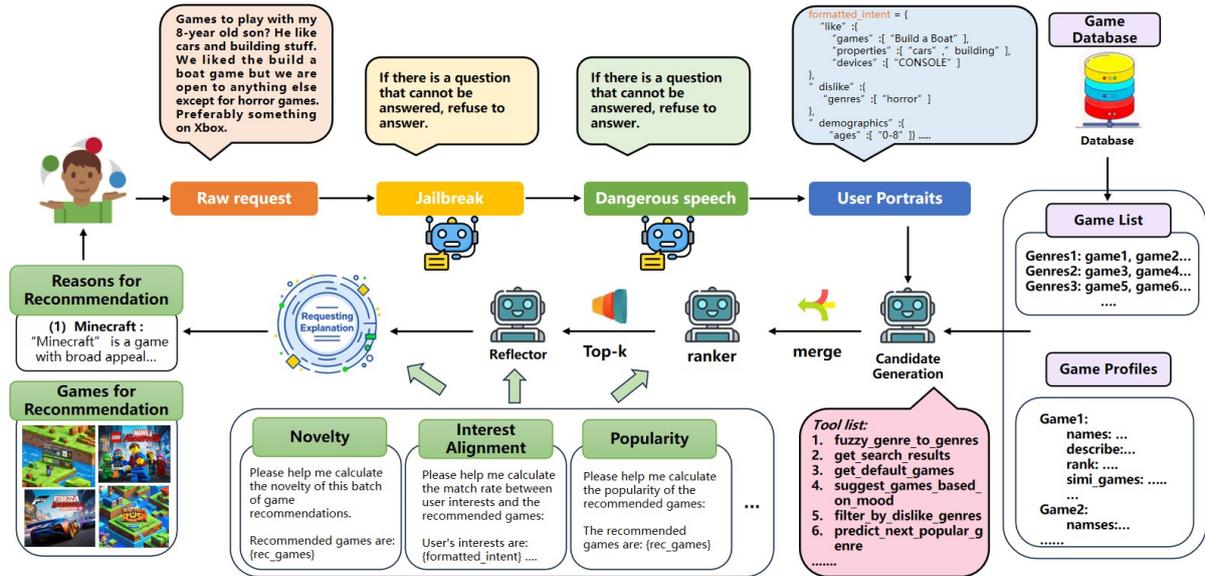


Figure 1: Overview of Our **MATCHA** framework. The system processes user requests through safety agents, generates game candidates using diverse tools, refines them via ranking and reflection agents, and provides final recommendations with detailed explanations.

the input and observing whether the model’s response changes significantly—an efficient strategy that leverages LLMs’ internal safety mechanisms without requiring fine-tuning. Second, chain-of-thought-based intent detection (Wei et al., 2022) uses an auxiliary LLM to reason about the semantic intent of a prompt, allowing the system to recognize subtle adversarial queries that aim to bypass ethical safeguards. Third, crafted policies define standardized fallback behaviors, such as returning non-committal answers or redirecting users to appropriate content, ensuring consistent and responsible handling of flagged prompts. The agent outputs a binary label (True/False) to indicate whether a prompt is adversarial. This design is compatible with models such as GPT and LLaMA, enabling a robust yet cost-effective solution for securing dialogue agents against adversarial manipulation.

4.1.2 Dangerous Content Detection

The Dangerous Content Detection Agent acts as a secondary layer of filtering for the Jailbreak Prevention Agent, enhancing the system’s ability to identify harmful intent. It operates at both the input (user queries) and output (recommendations) levels, evaluating content to flag potentially harmful or inappropriate elements. The agent returns a binary label, True or False, to indicate whether the content is flagged as harmful. This mechanism

upholds ethical guidelines and fosters user trust by delivering safe and appropriate outputs.

4.2 Candidate Generation

The Candidate Generation stage leverages a diverse set of tools to create a pool of potential game recommendations. OMuleT (eun Yoon et al., 2024) highlights that incorporating a wider variety of tools can enhance retrieval processes, better tailoring recommendations to individual user preferences. Following this insight, our system employs over ten specialized tools, including APIs for real-time game databases, genre-specific filters, trend analyzers, and user feedback systems. These tools address key aspects of game recommendation, such as filtering games by platform compatibility (e.g., PC, mobile, console), identifying genre-specific preferences (e.g., multiplayer or adventure games), incorporating user feedback metrics like ratings and reviews, and tracking real-time trends to include newly released or highly relevant games.

In addition to these tools, the system utilizes LLMs to analyze user intent and extract preferences related to genres and specific likes or dislikes. This intent analysis refines the recommendation pool by ensuring the generated candidates align closely with the user’s expressed interests. Detailed descriptions of the tools and their respective functions are provided in Appendix C. By combining insights

from the diverse toolset with LLM-driven intent analysis, the system produces a highly personalized and diverse pool of game recommendations, effectively supporting downstream processes such as ranking and reflection.

4.3 Ranking and Reflection

4.3.1 Ranking Agent

The Ranking Agent introduces a novel two-tier LLM collaboration mechanism, where multiple LLMs, such as GPT-4o and Gemini, work collaboratively and in parallel to make ranking decisions. Unlike traditional single-model approaches, our system leverages each LLM to independently evaluate candidate games across five metrics: popularity, match with user preferences, similarity to historical choices, genre alignment, and age suitability. The system combines the predictions from the two LLMs using weighted averages to account for their respective strengths. For example, one LLM may excel in understanding complex user intents, while another may better align with specific genre preferences. By leveraging the unique capabilities of each model, this collaborative approach ensures that the final scores provide a balanced and nuanced assessment of the candidates. Extensive testing reveals that this multi-LLM collaborative decision-making significantly enhances the accuracy and diversity of the recommendations, ensuring they align with user preferences while maintaining flexibility for novel and exploratory suggestions. Additionally, the agent incorporates an exploratory component controlled by hyperparameters, allowing it to take calculated risks in recommending games outside the user’s immediate preferences. For instance, if the user enjoys FPS games, the agent might suggest a highly-rated action-adventure game to encourage exploration of new genres. This mechanism strikes a balance between user personalization and novelty, delivering recommendations that are both familiar and adventurous.

4.4 Reflection Agent

The Reflection Agent leverages the concept of self-reflection, which has been shown to enhance problem-solving performance in LLM agents (Renze and Guven, 2024). Intuitively, we extend this principle to improve game recommendations, hypothesizing that a modified reflection process can lead to better alignment with user preferences. In this stage, we incorporate detailed descriptions

of games, referred to as "game profiles," which provide comprehensive information about each game. Due to the length and complexity of these profiles, they are only utilized during the reranking (reflection) phase to balance computational efficiency and performance. The Reflection Agent reassesses this list by incorporating contextual cues, user feedback, and the detailed game profiles. To maintain economic feasibility, the reflection process is limited to a subset of top-ranked candidates, ensuring an optimal trade-off between performance and computational cost. Our experiments demonstrate that the modified Reflection Agent, while slightly reducing diversity in recommendations, improves the accuracy of identifying games that align with users.

4.5 Explainability

The Explanation Agent generates detailed, user-centric explanations for recommended games, enhancing transparency and fostering greater user trust in the recommendation system. By synthesizing insights from multiple perspectives, it provides comprehensive justifications tailored to individual preferences across four key dimensions: **(1) Category Preferences:** Highlights how recommended games align with the user’s favored genres or sub-genres, such as RPGs or action-adventure games. **(2) Similarity:** Emphasizes similarities with previously enjoyed titles, considering gameplay style, mechanics, and thematic elements. **(3) Demographics:** Incorporates demographic cues, such as age group, to ensure recommendations are contextually appropriate and user-relevant. **(4) Popularity and Novelty:** Reflects prominence through factors such as player ratings, critical awards, or innovative features, while also noting recency and uniqueness. To ensure computational efficiency, the agent limits explanation generation to a predefined number of top-ranked games, referred to as the “explanation quota.” For each selected game, it queries metadata (e.g., game ID, descriptions, features, and tags) to construct a detailed profile. Explanations for each dimension are generated using prompts tailored to the specific aspect of interest, and then aggregated into a coherent and concise summary using a language model. Dimensions without relevant data are excluded to maintain clarity, factuality, and focus.

Experimental and human evaluation results (See Appendix F) show that the Explanation Agent boosts user engagement and satisfaction. By combining LLM reasoning with structured prompts, it

bridges the gap between recommendations and user trust.

5 Experiments Setups

5.1 Datasets

OMuleT (eun Yoon et al., 2024), a real-world dataset of 553 user requests and 2,074 unique game recommendations focused on Roblox. Each request is linked to an average of 14.2 games. **ReDial** (Li et al., 2018), a movie conversational recommendation dataset used to evaluate the generalization of our framework to other CRS domains. We test 2,500 samples from ReDial. **WildJailbreak** (Jiang et al., 2024), a large-scale benchmark for adversarial robustness, containing 262K prompt-response pairs, including 82K stealthy jailbreak prompts. We evaluate on its 2K test set. **“Do Anything Now”** (DAN) (Shen et al., 2024), a real-world jailbreak benchmark have 10K harmful samples across 13 forbidden scenarios. We also use 2K of its test set.

5.2 Metrics

We employ the following evaluation metrics to measure the relevance, novelty, coverage, and factuality of the recommendations.

Relevance: Evaluated using two metrics. **Hit@k** determines whether a ground truth item appears in the top- k recommendations, while **Precision@k** calculates the proportion of ground truth items within the top- k recommendations.

Novelty: In recommender systems, novelty is often associated with the exposure of an item, such as the frequency with which it appears in the recommendation lists (Vargas and Castells, 2011). We use **Pop50@k**, which measures the proportion of recommended items among the top 50 most popular items by upvotes, with lower values indicating less mainstream recommendations. Additionally, **RPop50@k** computes the ratio of **Pop50@k** for recommended items to that of the ground-truth items, where values closer to 1 reflect novelty levels similar to the ground-truth items.

Coverage: Evaluated using **MaxFreq@k** and **Entropy@k**. **MaxFreq@k** identifies the item recommended the most frequently and computes its proportion in all requests, with lower preferred to minimize repetition (Kaminskas and Bridge, 2016). **Entropy@k** measures the diversity of recommended items across all requests (Qin and Zhu, 2013).

Factuality: Measured using **Factual@k**, which calculates the proportion of real items in the top- k

recommendations. Items that cannot be linked to valid IDs are considered hallucinated.

Jailbreak Prevention Rate: This metric measures the proportion of harmful queries successfully blocked by the system, with higher rates indicating greater robustness against adversarial attacks.

Explanation Score: Assessed using a virtual judge (Zheng et al., 2023; Dong et al., 2024) and human annotators, this metric evaluates the quality of explanations provided for recommendations based on clarity, relevance and informativeness. Higher scores reflect explanations align well with human preference and effectively justify the recommendations. More details on the virtual judge are given in the Appendix E.

5.3 Baseline Models

We evaluated our system against several baseline models to benchmark its performance. These include both traditional and state-of-the-art approaches, as well as newly proposed baselines.

Pop: Randomly selects k items from the top-50 most popular items by their overall popularity.

Multi-Agent GPT: This baseline uses a Multiagent CRS based on GPT4o without additional tooling or the two-tier ranking with reflection, and multi-dimensional explanation mechanism.

MACRS: Based on the work by Fang et al. (2024), this multi-agent conversational recommender system uses a cooperative framework of LLM-based agents to plan dialogue acts and refine recommendations dynamically with user feedback. MACRS-C using GPT3.5 given the consideration of GPT3.5 is relatively old, we then use GPT4 instead.

MACRec: Proposed by Wang et al. (2024b), MACRec introduces a multi-agent collaboration framework specifically designed to enhance recommendation systems through specialized agents.

OMuleT: Proposed by eun Yoon et al. (2024), this baseline uses a multi-tool single agent framework that integrates user requests, oracle recommendations, and API-based filtering to recommend items.

6 Results

Table 1 highlights the performance of various methods across multiple metrics for top-5 and top-10 recommendations. Additional studies on the MATCHA framework in movie recommendation settings, as well as extended evaluations on the DAN jailbreak dataset, are provided in Appendix H. In high-cardinality recommendation tasks—such

Backbone	Method	Factual (\uparrow)	Relevance		Novelty		Coverage		JP	Exp
			Hit (\uparrow)	P (\uparrow)	Pop50 (\downarrow)	RPop50 (\downarrow)	E (\uparrow)	MaxF (\downarrow)		
	Pop	1.00	.14	.04	1.00	7.97	5.64	.15	✗	N/A
LLaMA-80B	Base LLM	.81	.09	.03	.84	4.33	8.25	.66	✗	N/A
	MAgent	.96	.23	.06	0.4	4.01	7.89	.31	✓	2.1
LLaMA-405B	Base LLM	.88	.23	.06	.48	3.84	6.57	.53	✗	N/A
	OMuleT	.99	.23	.07	.21	1.63	8.85	.16	✗	N/A
GPT-4o	MAgent	.94	.24	.07	.65	3.83	6.94	.27	✗	2.5
	MACRS-C	.85	.14	.04	.33	3.52	5.90	.42	✗	N/A
	MACRec	.92	.21	.07	.39	3.34	7.88	.31	✗	1.7
	OMuleT	.99	.24	.08	.27	2.14	8.71	.12	✗	N/A
	MATCHA	.99	.29	.10	.27	2.05	8.40	.09	✓	4.2

Backbone	Method	Factual (\uparrow)	Relevance		Novelty		Coverage		JP	Exp
			Hit (\uparrow)	P (\uparrow)	Pop50 (\downarrow)	RPop50 (\downarrow)	E (\uparrow)	MaxF (\downarrow)		
	Pop	1.00	.19	.03	1.00	8.40	5.64	.24	✗	N/A
LLaMA-80B	Base LLM	.75	.14	.03	.56	3.70	8.57	.56	✗	N/A
	MAgent	.91	.25	.04	.68	3.04	7.90	.32	✓	2.5
LLaMA-405B	Base LLM	.83	.28	.05	.40	3.32	7.26	.60	✗	N/A
	OMuleT	.99	.31	.06	.19	1.63	9.43	.19	✗	N/A
GPT-4o	MAgent	.94	.29	.05	.64	3.59	7.47	.32	✗	2.5
	MACRS-C	.82	.22	.03	.38	3.07	6.90	.43	✗	N/A
	MACRec	.92	.30	.06	.34	2.99	7.79	.23	✗	1.9
	OMuleT	.99	.33	.06	.25	2.13	9.21	.24	✗	N/A
	MATCHA	.98	.39	.09	.26	1.81	8.65	.17	✓	4.1

Table 1: **Overall performance across top-5 (top half) and top-10 (bottom half) recommendations.** **Bold** highlights indicate the best score in each column. \uparrow denotes higher is better; \downarrow denotes lower is better. Checkmark (✓) indicates successful jailbreak defense. Significance threshold: $p < 0.01$.

as game or app recommendation, where the candidate space may exceed 10,000 items—seemingly small absolute gains in top-k accuracy (e.g., +0.01 in Hit@5) can correspond to hundreds of more relevant results across large-scale deployments. Prior work, has similarly emphasized the impact of modest improvements in top-k relevance metrics (eun Yoon et al., 2024; Kook et al., 2025). MATCHA consistently outperforms other methods in most metrics, showcasing its robustness in conversational recommendation systems. For factuality, MATCHA achieves near-perfect scores (**.99**), matching OMuleT and surpassing other baselines, indicating its reliability in generating accurate recommendations. In relevance, MATCHA achieves the highest Hit@5 (**.29**) and Precision@5 (**.10**), demonstrating strong alignment between its recommendations and user preferences. Its novelty scores (RPop50@5: **2.05**, RPop50@10: **1.81**) reflect a healthy mix of mainstream and niche content, which is crucial for user engagement in interactive settings. For coverage, MATCHA maintains high diversity (E@5: **8.40**, E@10: **8.65**) while mini-

mizing repetition (MaxF@5: **.09**, MaxF@10: **.17**). Its jailbreak prevention capabilities (✓) and strong explanation quality (Exp@5: **4.2**, Exp@10: **4.1**) further distinguish it in terms of safety and transparency. These results indicate that MATCHA not only improves relevance but also maintains competitive diversity and factual accuracy. In particular, the simultaneous gains in novelty (lower RPop50) and explanation quality suggest that MATCHA generates recommendations that are both surprising and justifiable, which can enhance user satisfaction and trust in interactive settings. Compared to baselines like MACRS-C and MACRec, MATCHA delivers more consistent improvements across key dimensions. While OMuleT slightly surpasses it in entropy (E@10: **9.21** vs. **8.65**), MATCHA achieves a more favorable balance of relevance, novelty, and explanation quality. MATCHA also received an explanation score of **4.1** from a virtual judge. A follow-up human evaluation (See Appendix F) by domain experts yielded an average score of **3.97**, showing close agreement. This small gap suggests that while machine-evaluated justifications are re-

liable, further refinements may help better align generated explanations with human expectations and increase user trust. For computational overhead analysis, please see Appendix D.

7 Ablation Study

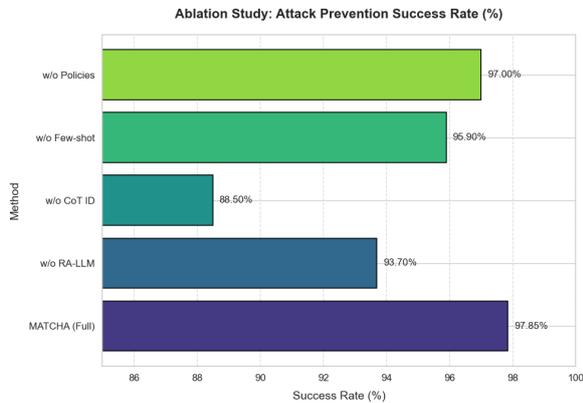


Figure 2: Ablation study results for the Jailbreak Prevention Agent on the WildJailbreak dataset.

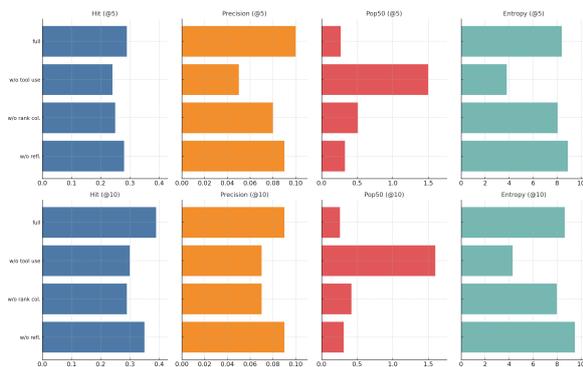


Figure 3: Ablation study results demonstrating the impact of removing the Reflection, multi-LLM collaboration for decision-making, and tool-used.

7.1 Jailbreak Prevention Module

We conducted ablation studies to evaluate the contribution of key components in the Jailbreak Prevention Agent. Figure 2 demonstrates that the full MATCHA system achieves the highest Attack Prevention Success Rate of **97.85%**. Removing RA-LLM reduces the success rate to **93.7%**, highlighting its role in detecting adversarial prompts effectively. Disabling CoT intent detection drops the success rate further to **88.5%**, showing its importance in identifying subtle harmful patterns. Few-shot prompting enhances robustness, as its removal leads to a decline in performance to **95.9%**. These results confirm that the combination of RA-LLM,

intent detection, and few-shot techniques is critical for mitigating adversarial queries.

7.2 Ranking, Reflection and Candidate Generation Modules

Figure 3 evaluates the impact of removing the Reflection Agent, multi-LLM collaboration, and tool-based candidate generation on the system’s overall performance. Disabling the Reflection Agent reduces relevance, as reranking with refined user preferences improves alignment between recommendations and user expectations. Removing multi-LLM collaboration diminishes ranking accuracy, demonstrating its role in leveraging diverse LLM knowledge to overcome static knowledge limitations and enhance personalization. Excluding tools significantly lowers coverage and novelty, emphasizing their importance in generating diverse and high-quality candidates. Together, these findings highlight the necessity of each module in achieving a balance of relevance, novelty, and robustness within the MATCHA framework.

8 Deployment

Our application is built on a full-stack server using Streamlit, which simplifies the creation of an interactive user interface and the management of backend operations. The deployment is hosted in an internal data center, using HashiCorp’s Nomad and Consul for cluster orchestration, deployment, and configuration management. To enhance usability, we have improved the feedback mechanism, allowing users to provide responses on the quality of recommendations directly within the application. This feedback is integrated into the system’s evaluation pipeline, facilitating continuous improvement of the recommendation framework. The application is accessible through the internal VPN, enabling widespread testing and iteration based on diverse user interactions. An example of the system in action is shown in Figure 4, with additional examples provided in the Appendix G.

9 Conclusion & Discussion

This work introduces MATCHA, a multi-agent conversational recommendation framework leveraging LLMs in combination with specialized tools. Extensive experiments show that MATCHA delivers accurate, diverse, and user-aligned game recommendations while maintaining strong guarantees for safety and transparency. In addition to propos-

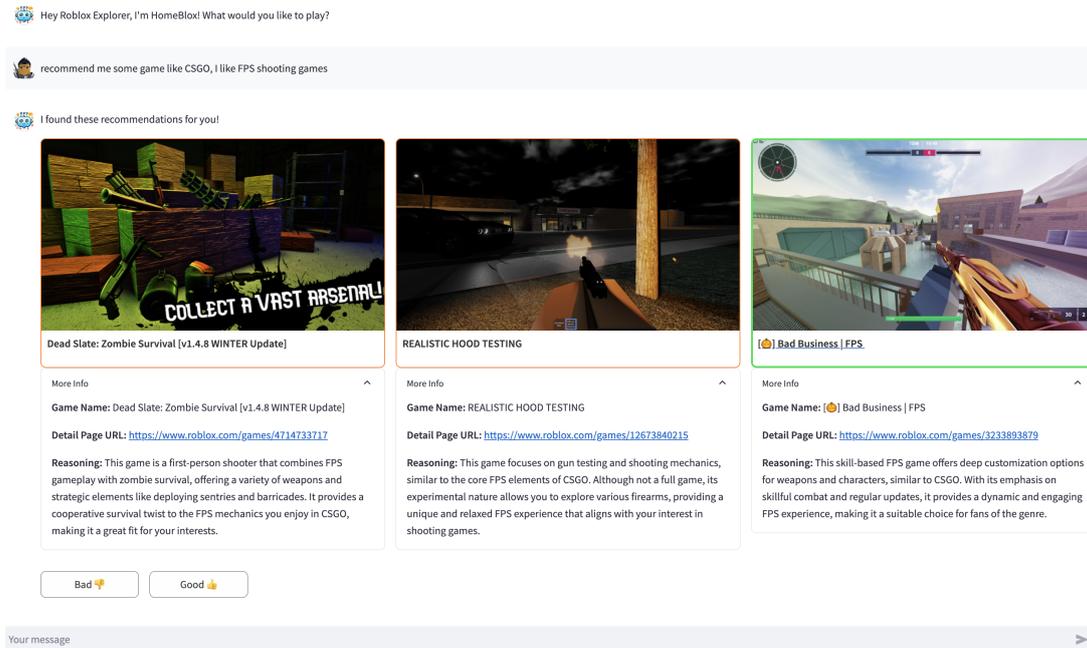


Figure 4: Demonstration of the application interface showcasing recommendations with reasons.

ing the core framework, this paper also provides a detailed analysis of system design, computational cost, and key challenges unique to game recommendation. The system has already been deployed internally, offering valuable feedback on performance, robustness, and usability under real-world setting. Future work will enhance explanation quality, better capture nuanced human preferences, explore additional domains, and improve the trade-off between diversity, relevance, and personalization.

10 Ethical Considerations

The MATCHA framework weaves in safeguards to help ensure it’s used responsibly. Its Risk Control Module works to filter out adversarial inputs and harmful content. To protect user privacy, inputs are anonymized and handled according to data regulations. MATCHA also takes steps to reduce bias by using multiple language models and generating a wide range of responses, although some bias from the original models may still be present. To make the system more transparent and easier to trust, the Explainability Module provides clear, multi-perspective explanations behind its decisions.

11 Limitations

While MATCHA shows strong performance, several limitations exist. (1) Limited Domain Coverage: The system is designed for game recommendations and may require adaptation for other domains. (2) Static Knowledge: Despite using tools

and multi-LLM collaboration, reliance on external databases may limit real-time updates. (3) Computational Cost: Multi-LLM collaboration and reflection increase latency and resource requirements. (4) Human Preference Gaps: Minor discrepancies remain between system outputs and nuanced human preferences. (5) Ethical Safeguards: Emerging jailbreak methods and new harmful content require continuous updates to maintain robustness. Future work will address these issues to enhance adaptability, efficiency, and fairness.

References

- Rogério Bonatti, Dan Zhao, Francesco Bonacci, Dillon Dupont, Sara Abdali, Yinheng Li, Yadong Lu, Justin Wagle, Kazuhito Koishida, Arthur Bucker, Lawrence Keunho Jang, and Zheng Hui. 2025. [Windows agent arena: Evaluating multi-modal OS agents at scale](#). In *Forty-second International Conference on Machine Learning*.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Steph Buongiorno, Lawrence Jake Klinkert, Tanishq Chawla, Zixin Zhuang, and Corey Clark. 2024. [Pangea: Procedural artificial narrative using generative ai for turn-based video games](#). *Preprint*, arXiv:2404.19721.
- Bochuan Cao, Yuanpu Cao, Lu Lin, and Jinghui Chen. 2024. [Defending against alignment-breaking attacks via robustly aligned LLM](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

- 10542–10560, Bangkok, Thailand. Association for Computational Linguistics.
- Mohamed Chawki. 2025. [Ai moderation and legal frameworks in child-centric social media: A case study of roblox](#). *Laws*, 14(3).
- Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proceedings of the 1st workshop on deep learning for recommender systems*, pages 7–10.
- Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.
- Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. [Towards conversational recommender systems](#). In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, page 815–824, New York, NY, USA. Association for Computing Machinery.
- Sunhao Dai, Chen Xu, Shicheng Xu, Liang Pang, Zhenhua Dong, and Jun Xu. 2024. Bias and unfairness in information retrieval systems: New challenges in the llm era. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6437–6447.
- Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. [Can llm be a personalized judge?](#) *arXiv preprint arXiv:2406.11657*.
- Yijiang River Dong, Tiancheng Hu, Zheng Hui, and Nigel Collier. 2026a. [Steer model beyond assistant: Controlling system prompt strength via contrastive decoding](#). *arXiv preprint arXiv:2601.06403*.
- Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. 2026b. [Value of information: A framework for human-agent communication](#). *Preprint*, arXiv:2601.06407.
- Se eun Yoon, Xiaokai Wei, Yexi Jiang, Rachit Pa-reek, Frank Ong, Kevin Gao, Julian McAuley, and Michelle Gong. 2024. [Omulet: Orchestrating multiple tools for practicable conversational recommendation](#). *Preprint*, arXiv:2411.19352.
- Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. [A multi-agent conversational recommender system](#). *arXiv preprint arXiv:2402.01135*.
- Chao Feng, Xinyu Zhang, and Zichu Fei. 2023. [Knowledge solver: Teaching llms to search for domain knowledge from knowledge graphs](#). *arXiv preprint arXiv:2309.03118*.
- Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. [Leveraging large language models in conversational recommender systems](#). *arXiv preprint arXiv:2305.07961*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. [A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions](#). *ACM Transactions on Information Systems*.
- Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. 2025a. [Trident: Benchmarking llm safety in finance, medicine, and law](#). *Preprint*, arXiv:2507.21134.
- Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi, and Nigel Collier. 2025b. [Privacypad: A reinforcement learning framework for dynamic privacy-aware delegation](#). *Preprint*, arXiv:2510.16054.
- Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, Lin Ai, Yinheng Li, Julia Hirschberg, and Congrui Huang. 2024a. [Can open-source llms enhance data augmentation for toxic detection?: An experimental study](#). *arXiv preprint arXiv:2411.15175*.
- Zheng Hui, Zhaoxiao Guo, Hang Zhao, Juanyong Duan, and Congrui Huang. 2024b. [ToxiCraft: A novel framework for synthetic generation of harmful information](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16632–16647, Miami, Florida, USA. Association for Computational Linguistics.
- Zheng Hui, Yinheng Li, Tianyi Chen, Colby Banbury, Kazuhito Koishida, et al. 2025c. [Winclick: Gui grounding with multimodal large language models](#). *arXiv preprint arXiv:2503.04730*.
- Zheng Hui, Xiaokai Wei, Reza Shirkavand, Chen Wang, Weizhi Zhang, Alejandro Peláez, and Michelle Gong. 2025d. [Semantics meet signals: Dual codebook representation learning for generative recommendation](#). *Preprint*, arXiv:2511.20673.
- Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. [A survey on conversational recommender systems](#). *ACM Computing Surveys (CSUR)*, 54(5):1–36.
- Liwei Jiang, Kavel Rao, Seungju Han, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghal-lah, Ximing Lu, Maarten Sap, Yejin Choi, and Nouha Dziri. 2024. [Wildteaming at scale: From in-the-wild jailbreaks to \(adversarially\) safer language models](#). *Preprint*, arXiv:2406.18510.
- Marius Kaminskis and Derek Bridge. 2016. [Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems](#). *ACM Trans. Interact. Intell. Syst.*, 7(1).

- Heejin Kook, Junyoung Kim, Seongmin Park, and Jongwuk Lee. 2025. [Empowering retrieval-based conversational recommendation with contrasting user preferences](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7692–7707, Albuquerque, New Mexico. Association for Computational Linguistics.
- Lasso Security. 2023. [Amazon chatbot gone wrong](#).
- Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 304–312.
- Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2024. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *arXiv preprint arXiv:2405.01868*.
- Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. [Towards deep conversational recommendations](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaocheng Du, Xiangyang Li, Yong Liu, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. [A survey of personalization: From rag to agent](#). *Preprint*, arXiv:2504.10147.
- Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Zihao Wang, Xiaofeng Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, et al. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.
- Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.
- Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q. Tran, Jonah Samost, Maciej Kula, Ed H. Chi, and Maheswaran Sathiamoorthy. 2023. [Recommender systems with generative retrieval](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.
- Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295.
- Xinyue Shen, Zeyuan Chen, Michael Backes, Yun Shen, and Yang Zhang. 2024. ["do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models](#). In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24*, page 1671–1685, New York, NY, USA. Association for Computing Machinery.
- Reza Shirkavand, Xiaokai Wei, Chen Wang, Zheng Hui, Heng Huang, and Michelle Gong. 2025. [Catalog-native llm: Speaking item-id dialect with less entanglement for recommendation](#). *Preprint*, arXiv:2510.05125.
- Utah Business. 2021. [Latitude games' ai dungeon was changing the face of ai-generated content—until its users turned against it](#). Accessed: 2025-07-04.
- Saúl Vargas and Pablo Castells. 2011. [Rank and relevance in novelty and diversity metrics for recommender systems](#). In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, page 109–116, New York, NY, USA. Association for Computing Machinery.
- Chen Wang, Xiaokai Wei, Yexi Jiang, Frank Ong, Kevin Gao, Xiao Yu, Zheng Hui, Se-eun Yoon, Philip Yu, and Michelle Gong. 2025. Solving the content gap in roblox game recommendations: Llm-based profile generation and reranking. *arXiv preprint arXiv:2502.06802*.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024a. [A survey on large language model based autonomous agents](#). *Frontiers of Computer Science*, 18(6).
- Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023. Improving conversational recommendation systems via counterfactual data simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2398–2408.
- Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. 2024b. [Macrec: A multi-agent collaboration framework for recommendation](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, page 2760–2764, New York, NY, USA. Association for Computing Machinery.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaying Song, Ke Xu, and Qi Li. 2024a. [Jailbreak attacks and defenses against large language models: A survey](#). *arXiv preprint arXiv:2407.04295*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024b. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Gangyi Zhang. 2023. User-centric conversational recommendation: Adapting the need of user with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1349–1354.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Zero-Shot Hallucination Study: Games vs. Movies

To evaluate the knowledge coverage of LLMs across domains, we conducted a zero-shot recognition and hallucination test comparing game and movie titles.

Setup. We sampled 5,000 game titles from the Roblox platform and 5,000 movie titles from the ReDial dataset. For each title t , GPT-4o was prompted with: “Describe the game/movie t in one sentence.”

Metrics.

- **Recognition Rate:** Whether the output included the correct genre or key characteristics.
- **Hallucination Rate:** Whether the output introduced fabricated or factually incorrect details.

Results. Table 2 presents the recognition and hallucination rates. GPT-4o performs substantially worse on game titles, suggesting a coverage gap that supports our hypothesis that games are under-represented in pretraining data. RR

Domain	RR (%)	HallR(%)
Games	76.2	23.8
Movies	95.3	4.7

Table 2: Zero-shot recognition and hallucination performance on 5,000 game titles and 5,000 movie titles using GPT-4o. **RR** denotes Recognition Rate and **HallR** denotes Hallucination Rate.

B Safety Risks in Game Conversational Recommendation

We further examined the unique safety risks posed by game-focused conversational recommender systems (CRS), both through empirical comparison and external case studies.

Empirical Comparison. We evaluated GPT-4o on 1,000 user queries for both game and movie recommendation scenarios. Table 3 summarizes the unsafe content rates.

The results suggest that game CRS systems have a significantly broader attack surface, increasing the likelihood of unsafe or policy-violating outputs.

Domain	Unsafe Response Rate (%)
Game CRS	7.8
Movie CRS	1.2

Table 3: Proportion of responses containing violent, discriminatory, or inappropriate content.

Real-World Evidence. These empirical findings are reinforced by documented safety failures in deployed systems:

AI Dungeon Incident (Utah Business, 2021): Unmoderated game-style generation using GPT-3 led to disturbing outputs, including illegal content, prompting emergency safeguards.

Roblox Case Study (Chawki, 2025): Analyzed how recommendation and feed systems on game platforms can unintentionally expose users to toxic or grooming-prone content.

Generative NPC Audit (Buongiorno et al., 2024): Found that game-based dialogue agents are particularly prone to explicit, biased, or hallucinated outputs in adversarial, multi-turn interactions.

These observations highlight the need for game-specific safeguards in CRS—such as jailbreak prevention, content filtering, age-aware reasoning, and explainability mechanisms—to mitigate both technical and ethical risks.

C Details of Candidate Generation Tools

The Candidate Generation stage utilizes over ten specialized tools adopted from eun Yoon et al. (2024) to ensure a robust and diverse recommendation pool. In Table 4, we provide a detailed description of each tool and its function.

D Computational Cost of Multi-LLM Collaboration & Reflection

This section provides a detailed analysis of the inference latency, resource usage, and cost associated with the MATCHA framework.

Inference. All LLM calls are executed asynchronously. The two ranking LLMs run in parallel, and the Reflection Agent is invoked only on the top-8 candidates. The table below summarizes average latency, GPU memory usage, and estimated API costs:

The full system adds only 0.21 seconds over the single-agent variant—well below typical user response times in dialogue settings.

Table 4: Detailed description of tools used in the Candidate Generation stage.

Tool	Input	Description
get_game_name	Game ID	Return the game name.
get_game_genre	Game ID	Return the game genre among the 21 predefined categories (e.g., 'RPG').
get_game_description	Game ID	Return a 2-3 sentence summary of what the game is about and how it is played.
get_game_rank	Game ID	Return the game rank by the number of upvotes.
is_device_compatible	Game ID, Device	Determine if the game is compatible with the given device (e.g., 'CONSOLE').
get_game_id_from_fuzzy_name	Fuzzy name	Return a game ID corresponding to an approximate name (e.g., "MM2" → ID for 'Murder Mystery 2').
fuzzy_genre_to_genres	Fuzzy genre	Return a list of predefined genres likely corresponding to a fuzzy genre name (e.g., 'simulation' → 'Simulator/Clicker').
get_search_results	Simple query	Use the search API to return relevant games for a simple query (maximum 3 words).
get_similar_games_cf	Game ID	Use collaborative filtering to return games played by users who played a given game.
get_similar_games_content	Game ID	Use SBERT embeddings to return games with similar descriptions.
get_games_by_age_group	Age group	Return games commonly played by users in a specified age group (e.g., '18-24').
get_default_games	Number of games	Randomly sample games from the top 100 games, useful for broad user requests.
get_game_info_str	Game ID	Return a string of game information in the format: {game name}, {genre}, {description}.
game_ids_to_enum_game_info	Game IDs	Return an enumerated string of game information for a list of game IDs.
suggest_games_based_on_mood	Mood type	Recommend games based on the user's mood or emotional preference (e.g., "relaxing," "exciting").
filter_by_dislike_genres	Game genres	Exclude games that belong to genres explicitly disliked by the user.
predict_next_popular_genre	Game IDs	Predict upcoming popular game genres using production recommendation algorithms.

Model Variant	Latency / Turn	GPU Memory	Cost (USD)
MATCHA (full)	1.32 s	11.1 GB	0.00012
w/o reflection	1.11 s	10.4 GB	0.00010
w/o multi-LLM	1.09 s	10.3 GB	0.00009

Table 5: Inference efficiency across MATCHA variants.

Training. MATCHA is training-free. The framework relies solely on curated tool usage and safety rules, with no additional fine-tuning or training cost beyond baseline LLM inference.

Deployment Efficiency. To ensure scalability, the system includes adjustable knobs: reflection is limited to a single pass, and the candidate pool is capped at 30 items. A hyperparameter table and deployment guidelines will be provided.

E Explanation Score Evaluation Details

The **Explanation Score** is evaluated using a virtual judge to assess the quality of explanations provided by the system. The evaluation is based on the following criteria:

F Human Evaluation Details

To ensure the quality of recommendations and explanations, we conducted a human evaluation process. This involved assessing 260 unique user requests and 754 game recommendations, each accompanied by detailed reasoning. Each recommendation was evaluated by three trained domain experts, resulting in approximately 2300 evaluation

tickets and requiring an estimated 100 human hours of dedicated effort. The evaluators, who were experienced in the gaming domain and familiar with diverse genres, platforms and mechanics, underwent a standardized training session before beginning the evaluation process. This training included detailed guidelines on the evaluation criteria, example annotations for clarity, and discussions on edge cases to ensure consistency and reliability in their judgments. The recommendations were assessed across multiple dimensions: relevance to the user query, novelty of the suggested games, coverage to ensure a diverse set of options, and alignment with user preferences. Evaluators also provided qualitative feedback on the explanations accompanying each recommendation, focusing on clarity, detail, and coherence.

F.1 Human Annotation Example

Table 8 provides an example of a human-annotated recommendation with a score of 3. This example demonstrates the evaluation process, highlighting the alignment between the suggested game and the user's preferences, as well as areas where the recommendation falls short.

F.2 Evaluation Interface Demonstration

Figure 5 showcases the evaluation interface used by domain experts during the human annotation process. This interface, hosted on our evaluation platform, presents recommendations alongside detailed reasoning, enabling evaluators to assess the quality of suggestions based on multiple dimensions such as relevance, novelty, and alignment with user preferences. The screen shown in the figure represents what an annotator would see during the evaluation, including the user query, recommended games, explanations, and input fields for feedback.

G Demo Examples

In this section, we provide additional examples demonstrating the functionality of the application. Each example highlights different scenarios, such as personalized recommendations, rejection to answer to Jailbreak attempt, and feedback integration. See Figure 6 for an additional illustration.

H Additional Evaluation on Jailbreak Robustness and General Recommendation

Evaluation on Real-World Jailbreak Prompts. To evaluate the effectiveness of MATCHA’s Risk Control Module against adversarial prompts, we conduct additional experiments on the “Do Anything Now” (DAN) benchmark (Vargas and Castells, 2011), which contains jailbreak prompts collected from 131 real-world online communities between December 2022 and December 2023. These prompts cover a wide range of adversarial strategies, including prompt injection, roleplay-based exploits, and persistent jailbreak patterns observed in the wild. The dataset is considered highly realistic and challenging due to its human-crafted nature and longevity of effective jailbreaks.

Table 6 reports the defense success rate across multiple models and frameworks. MATCHA achieves a success rate of **94.17%**, significantly outperforming baseline systems such as GPT-4o (70.13%) and MACRec (74.83%). These results demonstrate MATCHA’s strong generalization to adversarial prompts beyond synthetic test cases, validating the robustness of its multi-layered Risk Control mechanism.

Generalization to Movie Recommendation (ReDial). To evaluate the generalizability of

Table 6: Defense success rate on “Do Anything Now” jailbreak dataset

Model / System	Defense Success Rate (%)
LLaMA-405B	43.22
GPT-4o	70.13
MACRS-C	63.05
MACRec	74.83
OMuleT	68.50
MATCHA (ours)	94.17

MATCHA beyond the gaming domain, we adapt the framework to the ReDial dataset, a widely used benchmark for multi-turn movie recommendation. ReDial consists of annotated dialogues where one user seeks movie suggestions and the other responds, making it suitable for assessing open-domain CRS performance.

MATCHA is applied to ReDial with minimal modification, replacing game-specific APIs with a structured movie database. We report two standard metrics: Recall@10 and HitRate@10 (HR@10), which assess the system’s ability to retrieve and surface relevant items within the top-10 predictions. As shown in Table 7, MATCHA achieves the highest Recall@10 score (**0.182**) and performs competitively on HR@10 (**0.289**), closely matching the best-performing baseline (MACRS).

Table 7: Recommendation performance on ReDial dataset

Model / System	Recall@10	HR@10
LLaMA-405B	0.143	0.172
GPT-4o	0.171	0.278
MACRS	0.174	0.291
MACRec	0.164	0.235
MATCHA (ours)	0.182	0.289

These results suggest that MATCHA’s modular architecture, originally designed for game recommendation, generalizes effectively to other CRS domains. Without domain-specific retraining, MATCHA maintains strong recommendation performance, supporting its flexibility and extensibility. Combined with its demonstrated safety advantages, MATCHA offers a promising foundation for building robust, trustworthy conversational recommenders across domains. g foundation for building safe and effective CRS across diverse domains.

Table 8: Example of a Human Annotation with Neutral Score 3

User Query	Hey guys, looking for a game with a nice progression system. Doesn't matter what it is: RP game, action game, or MMO. I just need something to play and get that dopamine rush of getting stuff and progressing. Played a lot of World // Zero, Survive the Night, Royale High, and Island. Any kind of progression system (aside from Tycoons, I suppose).
Explanation Provided	This game features a robust progression system through various activities like building and decorating houses, customizing characters, and leveling up skills, similar to "Royale High" and "Island," making it a good fit for your preferences.
Game Recommended	Bloxburg
Game URL	https://www.xxxx.com/games/185655149
Score Assigned	3
Evaluator Comments	The user appears to be seeking player progression in an action/RPG/MMO-style game. While Bloxburg offers a progression system, it is primarily roleplay-oriented, and progression is not a core focus of the experience, making it partially aligned with the user's request.

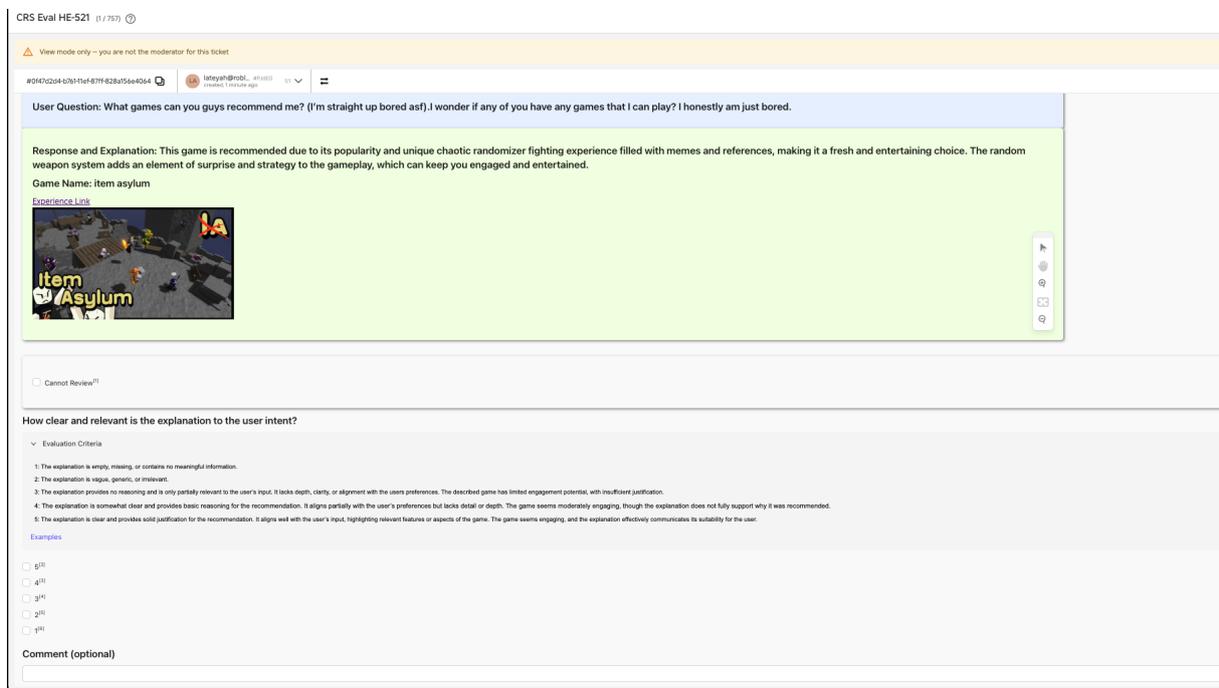


Figure 5: Illustration of the evaluation interface used for human annotations, showcasing the layout and components visible to annotators.

give me some roleplay game that is fun

I found these recommendations for you!



Creatures of Sonaria Survive Kaiju Animals

More Info

Game Name: Creatures of Sonaria Survive Kaiju Animals

Detail Page URL:
<https://www.roblox.com/games/5233782396>

Reasoning: This game is a great fit for your interest in role-playing games as it offers an immersive survival experience with fantasy elements. You can start as a young creature and evolve, engaging in exciting combat and social interactions, which adds depth and fun to the gameplay.

Bad

Good



Prison Life

More Info

Game Name: Prison Life

Detail Page URL: <https://www.roblox.com/games/155615604>

Reasoning: This game provides a dynamic role-playing environment where you can choose to be a prisoner or a guard. Its engaging mechanics and strategic gameplay offer a fresh perspective on role-playing, making it a fun choice for those seeking a unique experience.



Brazilian Army "EB"

More Info

Game Name: Brazilian Army "EB"

Detail Page URL:
<https://www.roblox.com/games/2069320852>

Reasoning: Although it has a niche appeal, this game offers a military role-playing experience that can be intriguing for those interested in such settings. It allows you to engage in various military scenarios, providing a different yet enjoyable role-playing adventure.

Your message

Figure 6: Additional example showcasing a personalized recommendation scenario.