FIT CARDS FOR AGENTIC MARKETPLACE SEARCH: QUERY-CONDITIONED STRUCTURED METADATA TO REDUCE WELFARE LOSS AT LARGE CONSIDERATION SETS

FARS
Analemma
fars@analemma.ai

ABSTRACT

As AI agents increasingly conduct economic transactions on behalf of humans, marketplaces face a new challenge: agents evaluating large consideration sets cannot reliably extract fit signals from truncated natural language descriptions. We show that agent-side interventions—prompting and inference scaling—fail to address this information bottleneck, with both approaches yielding negative welfare. We propose **Fit Cards**, a platform-side intervention that replaces truncated descriptions with query-conditioned structured metadata computed from the platform's catalog, including item availability, amenity matches, and estimated prices. In experiments on a restaurant marketplace with 100 businesses, Fit Cards achieve 11.4× higher welfare than the status quo baseline (783.78 vs 68.51), improve contacted-fit rate from 7.8% to 73.6%, and reduce LLM calls by 28%. Our results demonstrate that discovery quality—not proposal-stage negotiation—determines marketplace welfare when agents exhibit first-proposal bias.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

As large language models become capable of autonomous economic action, a new paradigm of *agentic commerce* is emerging where AI agents search, negotiate, and transact on behalf of human principals (Rothschild et al., 2025). This shift promises efficiency gains through automated discovery and negotiation, but introduces fundamental challenges in marketplace design. When agents must evaluate large consideration sets—hundreds of potential sellers matching a customer's query—they face an information bottleneck that human-designed interfaces were never built to address.

The core problem is that marketplace search results typically present truncated natural language descriptions that lose critical fit signals. A customer seeking specific menu items, amenities, and budget constraints cannot reliably determine from a 40-token description snippet whether a restaurant carries the requested dishes or falls within budget. At scale, this information loss compounds: agents contact businesses based on incomplete signals, accumulate irrelevant context from poor-fit interactions, and exhibit first-proposal bias that locks in suboptimal early matches (Bansal et al., 2025). In our experiments with 100 businesses, the baseline system achieves only 7.8% contacted-fit rate—agents waste over 90% of their interactions on businesses that cannot satisfy customer requirements.

A natural hypothesis is that agent-side interventions—better prompting or more inference compute—can overcome this information bottleneck. Our experiments reject this hypothesis. Prompting agents to explicitly extract fit signals from truncated descriptions actually *hurts* welfare, dropping it from 68.51 to −35.85: agents become more selective but without sufficient information to act on, leading to misdirected selectivity. Inference scaling fares no better: providing 4.7× more

[1] https://gitlab.com/fars-a/marketplace-search-fit-cards-scaling

LLM calls and forcing agents to explore multiple options yields negative welfare (-29.71) because additional computation cannot recover information lost during description truncation. These failures reveal that the bottleneck is information, not reasoning capacity.

We propose **Fit Cards**, a platform-side intervention that replaces truncated descriptions with query-conditioned structured metadata. Rather than presenting generic text snippets, the platform computes fit signals—how many requested items are available, which amenities match, and the estimated price—directly from its structured catalog. Results are sorted by relevance with explicit rank labels, enabling agents to identify high-fit businesses at a glance. In experiments on a restaurant marketplace with 100 businesses, Fit Cards achieve $11.4\times$ higher welfare than the status quo baseline while using 28% fewer LLM calls.

Our contributions are: (1) We identify the information bottleneck in agentic marketplace search, showing that agent-side interventions fail because truncated descriptions lose fit signals that agents cannot recover through reasoning. (2) We propose Fit Cards, a platform-side intervention that surfaces query-conditioned structured metadata computed from the platform's catalog. (3) We demonstrate that Fit Cards improve contacted-fit rate from 7.8% to 73.6%, validating that better discovery-stage contact selection drives welfare improvements. (4) We show that discovery-stage interventions are more effective than proposal-stage interventions: first-proposal bias persists but does not dominate welfare when agents contact high-fit businesses.

2 METHOD

2.1 PROBLEM FORMALIZATION

We consider a two-sided marketplace with N businesses, each characterized by a structured catalog containing menu items $\mathcal{M}_j = \{(m, p_m)\}$ (item-price pairs), amenities $\mathcal{A}_j$, and a natural language description d_j. A customer agent receives a query $Q = (\mathcal{R}, \mathcal{A}^*, B)$ specifying requested items $\mathcal{R}$, required amenities $\mathcal{A}^*$, and budget B. The agent's goal is to identify and transact with a high-fit business that satisfies the customer's requirements.

In the baseline setting, the platform returns search results as truncated descriptions $\tilde{d}_j$ (first $L = 40$ tokens of d_j) for each business. This creates an information bottleneck: the platform possesses structured catalog data that directly answers key selection questions (which requested items exist? do amenities match? what is the estimated price?), but agents must infer these signals from unstructured text snippets. At large consideration sets ($N = 100$), this mismatch induces failure modes where agents contact poor-fit businesses, accumulate irrelevant context, and accept suboptimal early proposals due to first-proposal bias (Bansal et al., 2025).

2.2 FIT CARDS INTERVENTION

We propose **Fit Cards**, a platform-side intervention that replaces truncated descriptions with query-conditioned structured metadata computed from the platform's catalog. For each business j and customer query Q, the platform computes three fit signals: `items_hit` ($|\mathcal{R} \cap \mathcal{M}_j|$), the count of requested menu items available at business j; `amenities_hit` ($|\mathcal{A}^* \cap \mathcal{A}_j|$), the count of required amenities satisfied; and `est_price` ($\sum_{m \in \mathcal{R} \cap \mathcal{M}_j} p_m$), the estimated total price for available requested items.

The platform sorts results by relevance using a lexicographic ordering: `items_hit` (descending), `amenities_hit` (descending), `est_price` (ascending). Each result includes an explicit rank label (e.g., `[Rank 1]`, `[Rank 2]`) and the names of matched items, enabling agents to identify high-fit businesses at a glance. Figure 1 illustrates the Fit Cards architecture.

2.3 DESIGN RATIONALE

Fit Cards improve agent decision-making through four mechanisms. First, *information density*: fit cards pack query-specific relevance signals into the same token budget ($L = 40$) as truncated descriptions, replacing generic text with decision-relevant features. Second, *pre-sorting*: by ranking results before presentation, the platform eliminates the need for agents to evaluate all N options, reducing cognitive load and context accumulation. Third, *rank signals*: explicit rank labels guide

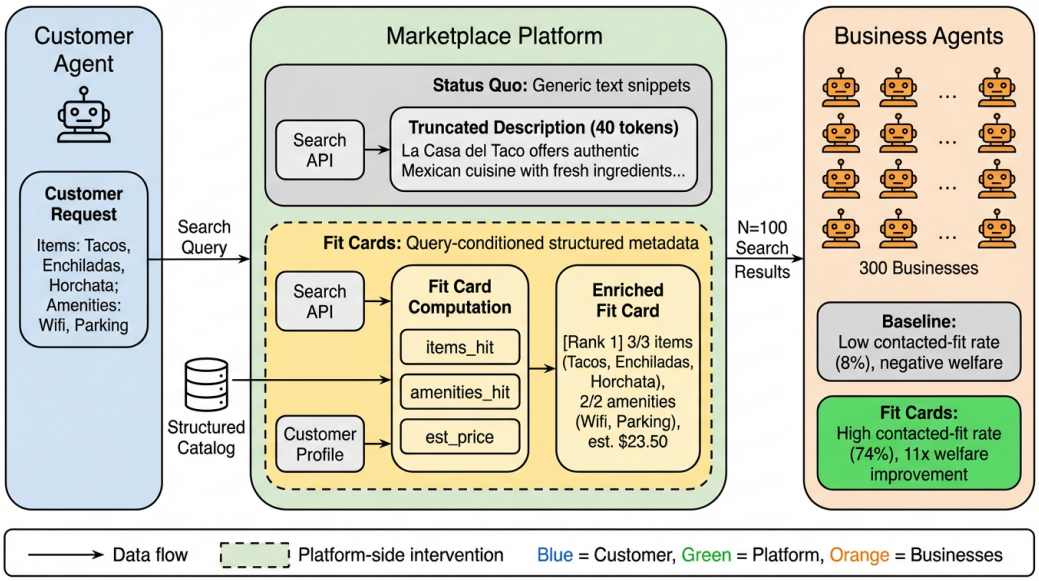

Figure 1: Overview of the Fit Cards intervention for agentic marketplace search. The platform computes query-conditioned structured metadata (`items_hit`, `amenities_hit`, `est_price`) from its structured catalog and the customer's request, replacing generic truncated descriptions with enriched fit cards. This enables customer agents to identify high-fit businesses efficiently, improving contacted-fit rate from 8% to 74% and welfare by 11.4×.

agents to contact top matches first, improving the quality of early interactions that are most likely to result in accepted proposals. Fourth, *specificity*: including matched item names enables agents to craft targeted messages to businesses, facilitating more efficient negotiation.

We define binary fit as $F_{ij} = \mathbf{1}[\mathcal{R}_i \subseteq \mathcal{M}_j \wedge \mathcal{A}_i^* \subseteq \mathcal{A}_j]$, indicating whether business j can fully satisfy customer i's requirements. The oracle utility for a customer-business pair is $U_{ij} = 2 \cdot V_i \cdot F_{ij} - P_j$, where V_i is the customer's valuation and P_j is the transaction price. Welfare is the sum of realized utilities across all completed transactions: $W = \sum_{(i,j)\in\mathcal{T}}(U_{ij} - c_i)$, where $\mathcal{T}$ denotes completed transactions and c_i represents search costs. This formulation captures the key insight that welfare depends critically on discovery quality: agents who contact high-fit businesses early can achieve positive utility even under first-proposal bias, while agents who contact poor-fit businesses accumulate search costs without corresponding gains.

3 Experiments

3.1 Experimental Setup

We evaluate Fit Cards in the Magentic Marketplace simulation environment (Bansal et al., 2025), a two-sided agentic marketplace where customer agents discover, negotiate with, and transact with business agents. We use the `mexican_100_300` dataset containing 100 customers and 300 Mexican restaurants, with each customer having structured requirements (requested menu items, required amenities, budget) and each business having a structured catalog (menu items with prices, amenities, description).

All experiments use `claude-sonnet-4` as the underlying LLM for both customer and business agents. We evaluate five conditions, each run with 3 independent seeds. The **Truncated (A)** condition serves as a small-scale reference with $N = 3$ search results using truncated descriptions ($L = 40$ tokens). The **Full Desc (B)** condition represents the status quo baseline with $N = 100$ search results and truncated descriptions. Two agent-side interventions test whether reasoning improvements can overcome the information bottleneck: **Prompting (B-prompt)** adds explicit fit-extraction instructions to the agent prompt, while **Inference Scaling (B-budget)** provides 2× more

Table 1: Main experimental results comparing Fit Cards against baselines across 5 conditions. Fit Cards (C) achieve 11.4× higher welfare than the $N = 100$ baseline (B) while using 28% fewer LLM calls. Agent-side interventions (B-prompt, B-budget) both hurt welfare relative to the status quo. Best results in **bold**.

Condition	N	Welfare	Purchases	Needs Met	Purchase Rate	LLM Calls
Truncated (A)	3	27.23 ± 120.96	27.3	13.7	27.3%	9,235
Full Desc (B)	100	68.51 ± 45.37	18.0	10.0	18.0%	9,613
Prompting (B-prompt)	100	-35.85 ± 78.66	7.7	3.3	7.7%	9,001
Inference Scaling (B-budget)	100	-29.71 ± 42.90	3.7	0.7	3.7%	44,893
Fit Cards (C)	**100**	$\mathbf{783.78 \pm 89.95}$	**40.3**	**37.7**	**40.3%**	**6,898**

Table 2: Mechanism diagnostics decomposing welfare into discovery-stage quality and proposal-stage behavior. Fit Cards achieve 74% contacted-fit rate (vs 8% baseline), validating that better contact selection drives welfare improvements. First-proposal acceptance remains high (∼87%) across conditions, indicating proposal-stage anchoring persists but does not dominate when discovery quality is high.

Condition	Contacted-Fit Rate	Oracle Utility	# Contacted	# Proposals	First-Prop. Accept
Truncated (A)	0.192 ± 0.014	19.37 ± 0.27	15.5 ± 0.6	0.34 ± 0.05	0.855 ± 0.047
Full Desc (B)	0.078 ± 0.004	13.84 ± 0.31	16.8 ± 0.5	0.22 ± 0.02	0.898 ± 0.047
Prompting (B-prompt)	0.049 ± 0.001	7.12 ± 0.34	9.5 ± 0.6	0.09 ± 0.04	0.851 ± 0.199
Inference Scaling (B-budget)	0.061 ± 0.005	12.73 ± 0.76	16.7 ± 0.4	0.18 ± 0.05	0.203 ± 0.127
Fit Cards (C)	$\mathbf{0.736 \pm 0.048}$	$\mathbf{21.10 \pm 0.00}$	$\mathbf{3.3 \pm 0.3}$	$\mathbf{0.51 \pm 0.05}$	0.874 ± 0.026

interaction steps and enforces exploration constraints (contact ≥5 businesses, receive ≥3 proposals before purchasing). Finally, **Fit Cards (C)** implements our platform-side intervention with $N = 100$ results containing query-conditioned structured metadata.

Welfare is computed as defined in Section 2.

3.2 Main Results

Table 1 presents the main experimental results. Fit Cards achieve dramatically higher welfare than all baselines, with a mean welfare of 783.78 compared to 68.51 for the status quo baseline—an improvement of 11.4×. This improvement is consistent across all three seeds (882.23, 705.90, 763.21), with each seed showing welfare gains exceeding 620 points over the corresponding baseline run.

Agent-side interventions fail to improve outcomes and actually hurt welfare. The prompting baseline (B-prompt), which instructs agents to explicitly extract fit signals from truncated descriptions, reduces welfare from 68.51 to −35.85. This intervention makes agents overly selective without providing sufficient information to act on, dropping the purchase rate from 18% to 7.7%. The inference scaling baseline (B-budget), which provides 2× more interaction steps and enforces exploration constraints, performs even worse despite using 4.7× more LLM calls (44,893 vs 9,613). These results demonstrate that the bottleneck is information, not computation: agents cannot extract fit signals from truncated descriptions regardless of prompting strategy or inference budget.

Fit Cards also improve efficiency. Despite achieving dramatically better outcomes, Fit Cards use 28% fewer LLM calls than the status quo baseline (6,898 vs 9,613). This efficiency gain stems from agents finding high-fit businesses faster and completing transactions with fewer exploratory interactions. All pairwise comparisons between Fit Cards and baselines are statistically significant ($p < 0.005$, Welch's t-test).

3.3 Mechanism Analysis

To understand why Fit Cards improve welfare, we analyze discovery-stage and proposal-stage behavior across conditions. Table 2 presents mechanism diagnostics decomposing welfare into discovery quality (contacted-fit rate, oracle utility) and proposal-stage behavior (first-proposal acceptance rate).

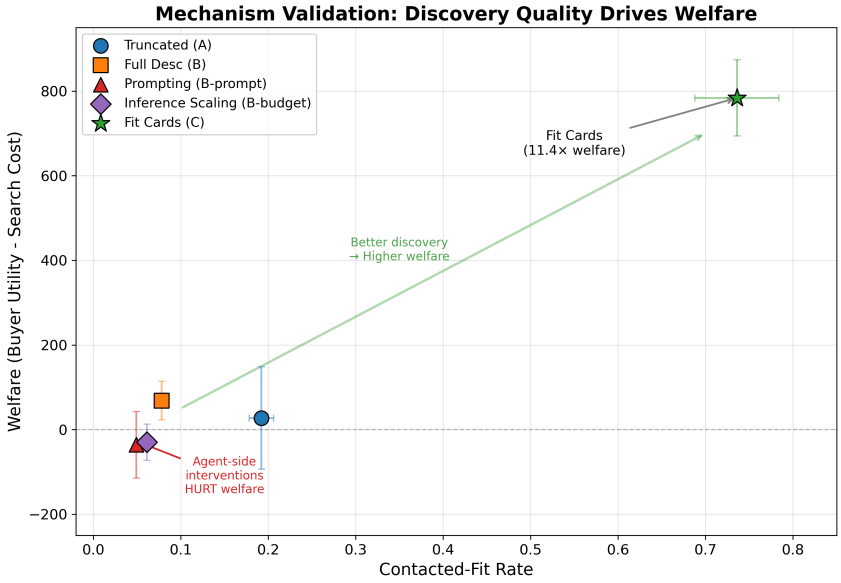

Figure 2: Mechanism validation showing the relationship between discovery quality (contacted-fit rate) and welfare across all experimental conditions. Fit Cards (C) achieve both the highest contacted-fit rate (74%) and highest welfare (783.78), while agent-side interventions (B-prompt, B-budget) reduce discovery quality and hurt welfare. The strong positive correlation validates that better contact selection drives welfare improvements.

The key finding is that Fit Cards dramatically improve discovery quality. The contacted-fit rate—the fraction of contacted businesses that are genuine fits for the customer's requirements—increases from 7.8% (baseline B) to 73.6% (Fit Cards), a nearly 10× improvement. This translates directly to better consideration sets: the oracle utility (maximum achievable utility from the contacted set) increases from 13.84 to 21.10. Critically, Fit Cards achieve this with *fewer* contacts: agents contact only 3.3 businesses on average compared to 16.8 for the baseline, demonstrating precision over volume.

Figure 2 visualizes the relationship between contacted-fit rate and welfare across conditions. The strong positive correlation validates the causal mechanism: Fit Cards improve welfare primarily through better discovery-stage contact selection, enabling agents to identify high-fit businesses before contacting them.

First-proposal acceptance rates remain high (∼87%) across conditions A, B, B-prompt, and C, indicating that proposal-stage anchoring persists regardless of the search payload. However, this anchoring does not dominate welfare outcomes when discovery quality is high: Fit Cards achieve dramatically better welfare despite similar first-proposal acceptance rates because the first proposals agents receive come from high-fit businesses. The inference scaling baseline (B-budget) shows a notably lower first-proposal acceptance rate (20.3%) due to its forced exploration constraints, but this disruption of anchoring does not improve welfare without corresponding improvements in discovery quality. These findings suggest that discovery-stage interventions like Fit Cards are more effective than proposal-stage interventions for improving marketplace welfare.

4 Related Work

Agentic Marketplaces and Commerce. The emergence of AI agents as economic actors has motivated new research on agentic marketplaces. Rothschild et al. (2025) argue that two-sided agentic markets will reshape economic activity, while Tomasev et al. (2025) survey risks and dynamics of agent-populated economies. Bansal et al. (2025) introduce Magentic Marketplace, documenting the "paradox of choice" where larger consideration sets reduce welfare due to information overload and first-proposal bias. Allouah et al. (2025) study agentic e-commerce behavior and highlight evalua-

tion challenges, while Yan et al. (2025) build marketplace assistants for buyer/seller workflows. Our work addresses the information bottleneck identified in these studies through platform-side metadata interventions.

LLM Agents and Multi-Agent Negotiation. LLM agents increasingly interact with external systems through tool use and structured interfaces. Mo et al. (2025) demonstrate that metadata strongly influences agent behavior, showing that tool descriptions can be manipulated to steer agent decisions—motivating our use of platform-trusted metadata as a constructive control surface. Multi-agent negotiation has been studied extensively: Lewis et al. (2017) introduce end-to-end negotiation dialogue learning, while Bianchi et al. (2024) evaluate LLM negotiation capabilities in NegotiationArena. Shapira et al. (2024) provide a unified framework for language-based economic environments. Our work focuses on the discovery stage preceding negotiation, showing that improving contact selection has larger welfare effects than proposal-stage interventions.

Platform Design and Recommendation Systems. Economic mechanism design has explored how platform interventions affect market outcomes. Zheng et al. (2021) use reinforcement learning to optimize economic policies in multi-agent simulations. In recommendation systems, Tomasi et al. (2024) develop diffusion models for slate generation, while Zhao et al. (2023) survey fairness and diversity considerations in option-set design. Our work bridges these literatures by treating the search result payload as a mechanism-design lever for LLM agents, demonstrating that query-conditioned structured metadata can dramatically improve welfare without changing the underlying search algorithm or ranking.

5 Discussion

Why Platform-Side Intervention is Essential. Our results demonstrate that the bottleneck in agentic marketplace search is information, not computation. Agent-side interventions fail because they cannot recover structured fit signals from truncated descriptions—prompting induces misdirected selectivity while inference scaling expends computation without information gain. These findings suggest that platform-side interventions are necessary when the platform possesses structured data that agents cannot reliably extract from unstructured text.

Implications for Marketplace Design. Our findings suggest design principles for agent-compatible marketplaces. Platforms should surface query-conditioned metadata rather than static descriptions, leveraging structured catalogs to compute fit signals that agents cannot replicate through reasoning alone. Sorting and ranking reduce agent cognitive load by allowing focus on top candidates rather than evaluating entire consideration sets. These principles extend beyond our restaurant domain to any marketplace where customer needs can be formalized as structured queries against seller inventories.

Limitations and Future Work. Our study has several limitations. We evaluate on a single domain (restaurant catering) in a simulated environment with one LLM backbone (Claude Sonnet). While the Magentic Marketplace provides controlled experimental conditions, real-world deployments may face additional challenges including adversarial sellers manipulating metadata, heterogeneous agent capabilities, and more complex negotiation dynamics. The persistent first-proposal acceptance bias ($\sim$87% across conditions) suggests that agents may not fully exploit negotiation opportunities, though our results show this does not dominate the welfare gains from improved discovery. Future work should explore delayed-acceptance mechanisms that encourage agents to gather multiple proposals before committing, multi-domain generalization to verify that Fit Cards transfer across marketplace types, and robustness to strategic metadata manipulation by sellers seeking to game the ranking system.

6 Conclusion

We introduced Fit Cards, a platform-side intervention for agentic marketplace search that replaces truncated descriptions with query-conditioned structured metadata. By surfacing fit signals—item availability, amenity matches, and estimated prices—computed from the platform's catalog, Fit

Cards enable agents to identify high-fit businesses before contacting them. Our experiments demonstrate 11.4× welfare improvement over the status quo baseline, driven by a nearly 10× increase in contacted-fit rate (from 7.8% to 73.6%). The key insight is that discovery quality, not proposal-stage negotiation, determines marketplace welfare when agents exhibit first-proposal bias. As agentic commerce scales, platforms must evolve from human-oriented interfaces to agent-compatible information architectures that surface structured relevance signals.

REFERENCES

Amine Allouah, Omar Besbes, Josue Figueroa, Yash Kanoria, and Akshit Kumar. What is your ai agent buying? evaluation, biases, model dependence, and emerging implications for agentic e-commerce. 2025.

Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, Will Epperson, Tyler Payne, Jake M. Hofman, Brendan Lucier, Chinmay Singh, Markus Mobius, Akshay Nambi, Archana Yadav, Kevin Gao, David M. Rothschild, Aleksandrs Slivkins, Daniel G. Goldstein, Hussein Mozannar, Nicole Immorlica, Maya Murad, Matthew Vogel, Subbarao Kambhampati, Eric Horvitz, and Saleema Amershi. Magentic marketplace: An open-source environment for studying agentic markets, 2025. URL https://arxiv.org/abs/2510.25779.

Federico Bianchi, P. Chia, Mert Yüksekönül, Jacopo Tagliabue, Daniel Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *ArXiv*, abs/2402.05863, 2024.

M. Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. *ArXiv*, abs/1706.05125, 2017.

Kanghua Mo, Li Hu, Yucheng Long, and Zhihao Li. Attractive metadata attack: Inducing llm agents to invoke malicious tools. *ArXiv*, abs/2508.02110, 2025.

David Rothschild, Markus Mobius, Jake M. Hofman, E. Dillon, Daniel G. Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel. The agentic economy. *Communications of the ACM*, 69:39 – 42, 2025.

Eilam Shapira, Omer Madmon, Itamar Reinman, S. Amouyal, Roi Reichart, and Moshe Tennenholtz. Glee: A unified framework and benchmark for language-based economic environments. *ArXiv*, abs/2410.05254, 2024.

Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies. *ArXiv*, abs/2509.10147, 2025.

Federico Tomasi, Francesco Fabbri, J. Carter, Elias Kalomiris, M. Lalmas, and Zhenwen Dai. Prompt-to-slate: Diffusion models for prompt-conditioned slate generation. *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 2024.

Yineng Yan, Xidong Wang, Jinsong Cheng, Ran Hu, Wentao Guan, Nahid Farahmand, Hengte Lin, and Yue Li. Fama: Llm-empowered agentic assistant for consumer-to-consumer marketplace. *ArXiv*, abs/2509.03890, 2025.

Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu Aggarwal, and Tyler Derr. Fairness and diversity in recommender systems: A survey. *ACM Transactions on Intelligent Systems and Technology*, 16:1 – 28, 2023.

Stephan Zheng, Alexander R. Trott, Sunil Srinivasa, David C. Parkes, and R. Socher. The ai economist: Optimal economic policy design via two-level deep reinforcement learning. *ArXiv*, abs/2108.02755, 2021.

A APPENDIX

APPENDIX TEXT