Toward Agentic Reconciliation: The Case for Multi-Stakeholder Negotiation in Tourism Recommender Systems

Ashmi Banerjee[1], Adithi Satish[1], Fitri Nur Aisyah[1], Wolfgang Wörndl[1] and Yashar Deldjoo[2]

[1]*Technical University of Munich, Munich, Germany*
[2]*Polytechnic University of Bari, Bari, Italy*

Abstract

Large language models (LLMs) are rapidly transforming recommender systems by enabling conversational, natural-language–driven recommendation generation. While these capabilities enhance usability and expressiveness, they also introduce new risks in domains where recommendations have real-world societal impact, such as tourism. In this position paper, we argue that monolithic LLM-based recommenders are ill-suited for balancing competing stakeholder objectives and instead advocate for agentic, moderator-mediated architectures as a principled alternative. We use the Collab-REC framework as a case study to demonstrate how role-specialized agents and transparent aggregation mechanisms can support grounded, auditable, and sustainability-aware recommendations. Building on these insights, we outline key research challenges at the intersection of LLMs, multi-stakeholder recommendation, and responsible AI, and argue that agentic recommender systems offer a promising pathway toward aligning generative personalization with broader societal goals.

This position paper (extended abstract) is submitted to the second LLM4Good Workshop on Sustainable and Trustworthy Large Language Models for Personalization (LLM4Good '26) [1], co-located with UMAP 2026, to present and discuss findings from our work *"Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism"*. A detailed description of the framework and empirical results is available in our accompanying arXiv preprint.[2]

Keywords

LLMs, Multi-Agent Systems, Tourism Recommender Systems, Multi-Stakeholder Fairness

1. Introduction and Context

Recommender systems are increasingly expected to support not only individual user satisfaction but also platform-level and societal objectives. This is particularly evident in tourism, where recommendations shape the physical movement of people across destinations, affecting local economies, environmental sustainability, and residents' quality of life [1]. As a result, tourism recommendation constitutes a complex multi-stakeholder problem involving travelers seeking personalized and feasible itineraries, platforms optimizing engagement, and destinations aiming to manage demand sustainably and mitigate overtourism [2, 3].

Recent advances in large language models (LLMs) have accelerated a paradigm shift from retrieval-based recommenders to conversational, generative systems that allow users to express nuanced travel intents in natural language, including constraints on budget, accessibility, seasonality, and sustainability [4]. While this improves usability and expressiveness, it also amplifies risks. Most LLM-based recommenders operate as monolithic end-to-end pipelines in which candidate generation, reasoning, and ranking are entangled within a single opaque model. This design limits transparency, complicates auditing of trade-offs, and often results in persistent popularity bias and hallucinated or infeasible

[1]https://recsys-lab.at/llm4good26
[2]Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2025. Collab-REC: An LLM-based Agentic Framework for Balancing Recommendations in Tourism. arXiv preprint arXiv:2508.15030 [cs.IR] (2025).

Joint Proceedings of the ACM UMAP Workshops 2026, UMAP 2026, June 8–11, 2026, Gothenburg, Sweden

*Corresponding author.
ashmi.banerjee@tum.de (A. Banerjee); adithi.satish@tum.de (A. Satish); fitri.aisyah@tum.de (F. N. Aisyah); woerndl@tum.de (W. Wörndl); yashar.deldjoo@poliba.it (Y. Deldjoo)

destinations [5, 6].

We argue that these limitations stem from a structural mismatch between monolithic LLM architectures and the inherently multi-objective, multi-stakeholder nature of tourism recommendation. When a single model is tasked with balancing competing objectives within a single prompt, it is prone to *objective collapse*, implicitly prioritizing statistically dominant patterns in its training data—typically globally popular destinations. Empirical observations in the Collab-REC framework demonstrate that even when prompted for sustainable or less-visited alternatives, monolithic LLMs frequently revert to high-traffic hubs, reinforcing spatial concentration rather than alleviating it [4].

Tourism Recommendation as a Socio-Technical System Unlike purely digital domains, tourism recommendations produce real-world externalities that extend beyond the user–platform interaction. Recommending already popular destinations can exacerbate congestion and strain local infrastructure, whereas aggressively promoting lesser-known locations may reduce user satisfaction if practical constraints are ignored. Consequently, tourism recommendation can be viewed as a socio-technical optimization problem requiring explicit trade-offs between personalization, platform goals, and sustainability [7, 8].

Limitations of Monolithic LLM Recommenders Despite their flexibility, monolithic LLM-based recommenders exhibit several systematic shortcomings in such settings [9]:

- **Objective collapse:** Multiple goals specified within a single prompt are implicitly collapsed into a single dominant objective, often popularity.
- **Hallucination and lack of grounding:** Models may generate out-of-catalog destinations or fabricate attributes such as sustainability indicators, leading to infeasible or misleading recommendations.
- **Lack of auditability:** End-to-end generation obscures intermediate reasoning steps, making it difficult to explain or verify the trade-offs underlying a recommendation.

Agentic Architectures for Multi-Stakeholder Recommendation To address these challenges, we advocate for agentic recommender architectures in which role-specialized agents independently generate recommendations from distinct stakeholder perspectives, followed by a moderator component that evaluates, grounds, and aggregates these proposals using transparent decision policies. This modular design decouples generation from decision-making, enables explicit representation of stakeholder objectives, and allows intermediate artifacts—such as rejected candidates and per-objective scores—to be logged and audited. Such properties align with emerging principles in responsible AI that emphasize transparency, controllability, and human oversight.

2. Collab-Rec: Framework

System Design and Workflow We illustrate our approach through Collab-Rec, a multi-agent, moderator-mediated framework for city-trip recommendation (Figure 1). The system instantiates three LLM-based agents representing *personalization*, *popularity*, and *sustainability* objectives. Each agent independently generates candidate destinations from a fixed catalog of 200 European cities, ensuring diverse stakeholder perspectives are present before aggregation. This multi-agent setup prevents the early-exit behavior observed in monolithic LLM recommenders, where less dominant objectives—such as sustainability—are often ignored in favor of popular, high-probability destinations [4].

A deterministic moderator coordinates the agents by validating, scoring, and iteratively refining the collective recommendation set. It enforces grounding by mapping outputs to a structured destination catalog and penalizing hallucinated or invalid entries. Per-agent diagnostics—including constraint satisfaction, reliability across rounds, and invalid-output rate—are aggregated through a transparent

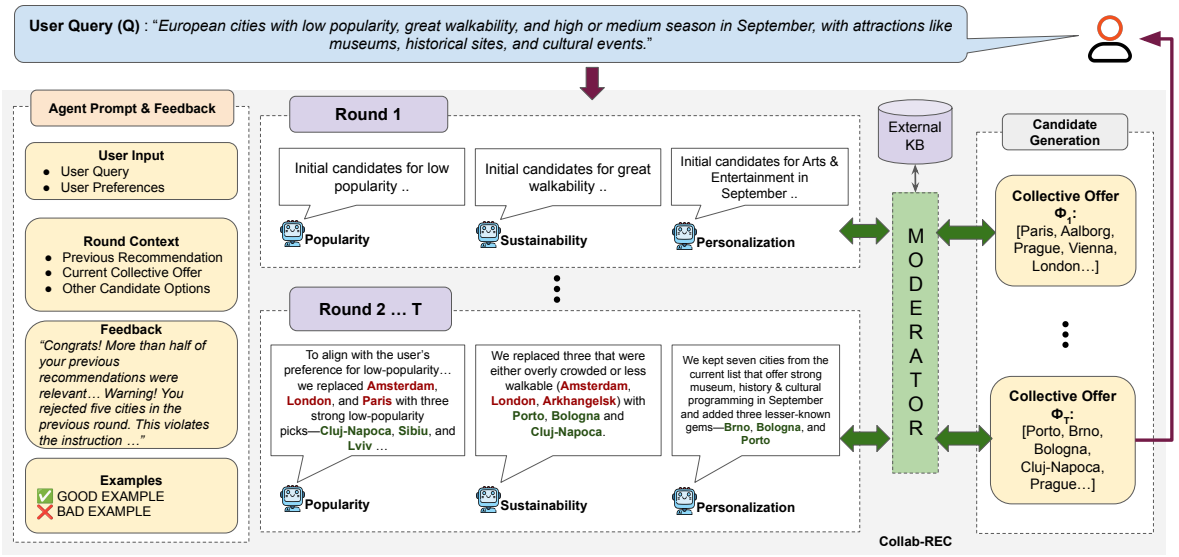

Figure 1: Overview of the Collab-Rec workflow to generate city trip recommendations using multiple LLM agents. The non-LLM **Moderator** evaluates and combines the agent proposals, iteratively refining the final recommendation set, which is then communicated to the user.

multi-objective scoring function. The resulting collective recommendation set is then broadcast back to the agents for further negotiation. This multi-round, moderator-mediated process supports iterative repair of constraint violations and encourages exploration of feasible, less popular alternatives. Empirical evaluations demonstrate that structured agentic interaction increases diversity and reduces popularity concentration while maintaining recommendation feasibility.

A critical design choice is the use of a non-LLM moderator as a trust anchor. While an LLM could theoretically act as coordinator, this introduces additional stochasticity and hallucination risk [10]. The deterministic moderator enforces hard inventory constraints and applies objective diagnostics—success, reliability, and hallucination penalties—ensuring that the generative creativity of the agents is bounded by the operational realities of an industrial travel catalog [9]. This architecture operationalizes multi-stakeholder negotiation in a transparent, auditable, and sustainability-aware manner.

Moreover, to maintain practical deployment efficiency, the framework employs a patience-based early stopping protocol that captures the majority of quality gains while significantly mitigating the linear latency and token costs associated with multi-round coordination.

Experimental Setup We evaluate Collab-Rec on 900 popularity-stratified tourism queries from the SynthTRIPs dataset [4], grounded against a structured catalog of 200 European cities enriched with destination metadata.

The study benchmarks six LLM backbones spanning both proprietary and open-source families, including *Claude-4.5-Sonnet*, *Gemini-2.5-Flash*, *GPT-OSS-20B*, *Gemma-3-12B*, *Gemma-3-4B*, and *Olmo3-7B*. Performance is compared against three baseline configurations: non-LLM recommenders (RandRec, TopPop), a single-agent single-iteration setup (SASI), and a single-round multi-agent configuration (MASI).

We assess grounded recommendation quality using moderator success—i.e., the proportion of recommendations that satisfy user constraints under catalog validation, alongside diversity and concentration measures (Gini index, entropy, and catalog coverage) and agent-level behavior metrics such as reliability and hallucination rate.

3. Research Questions and Key Findings

Our study was guided by five research questions investigating the effectiveness of multi-round agentic coordination: whether iterative negotiation improves grounded recommendation quality (RQ1), reduces popularity bias and increases long-tail coverage (RQ2), stabilizes agent reliability and hallucination behavior across rounds (RQ3), introduces manageable latency and token overhead (RQ4), and how sensitive the framework is to its scoring components through ablation analysis (RQ5).

The results demonstrate that COLLAB-REC, an agentic, moderator-mediated recommendation framework, effectively balances user preferences, destination popularity, and sustainability while remaining deployable in practice.

- **Relevance, Diversity, and Iteration:** Multi-round coordination significantly improves grounded relevance within the first 3–4 rounds, after which gains plateau, while diversity metrics such as long-tail coverage and catalog entropy continue to improve in later rounds [11]. This reveals a controllable "diversity budget" where additional computation can be traded for more equitable destination distribution.
- **Negotiation as a Robust Alternative to Passive Ranking:** By framing recommendation as a convergent multi-agent negotiation rather than a one-shot ranking task, the system preserves consensus items and incrementally repairs constraint violations, leading to more stable and grounded outputs than single-shot generation [12, 13].
- **Systematic Mitigation of Popularity Bias:** Iterative coordination shifts recommendations away from high-traffic hubs toward long-tail destinations, increasing catalog coverage (e.g., 81.5% for Claude vs. 66% in single-shot baselines) and reducing concentration, while moderator grounding eliminates hallucinated suggestions.
- **Operational Feasibility through Early Stopping:** Although multi-round loops introduce additional latency and token usage, a patience-based early stopping protocol captures nearly all relevance gains by round 4 (RQ4), reducing cumulative latency to practical levels (e.g., $\sim$87 seconds for proprietary models) while retaining most quality improvements.
- **Robustness, Ablation Sensitivity, and Small-Model Gains:** Ablation studies (RQ5) show that structured moderator feedback and scoring components are critical for performance stability. Smaller open-source models, such as Olmo-7b, benefit disproportionately from these signals, achieving competitive performance despite lower parameter scale [14], demonstrating the scalability and modularity of the agentic architecture.

4. Conclusion

As LLMs become increasingly embedded in user-facing recommendation pipelines, their societal impact will grow correspondingly. We argue that responsible deployment in high-impact domains such as tourism requires moving beyond monolithic generative architectures toward modular, agentic designs that support explicit trade-off management and grounded validation.

Despite their promise, agentic recommender systems introduce new challenges. Multi-round coordination increases latency and computational cost, and the design of aggregation policies requires normative choices that may reflect platform incentives. Ensuring stable and robust agent behavior across rounds also remains an open problem, particularly when agents rely on stochastic generation. Further research is needed to explore adaptive objective weighting, human-in-the-loop moderation, and standardized evaluation benchmarks that capture both user-centric and societal metrics.

We hope this position paper stimulates further discussion on how the recommender systems and LLM communities can jointly develop architectures that align personalization with sustainability and broader social good.

GenAI Usage Disclosure

We used ChatGPT (OpenAI) to assist with the formulation of code snippets during development. We also used Grammarly to identify grammar inconsistencies and to improve readability. All generated suggestions were critically reviewed and edited by the authors to ensure correctness and originality, and the authors take full responsibility for the content of this manuscript.

Acknowledgments

We thank the Google AI and machine learning Developer Programs team for supporting this work with Google Cloud credits.

References

[1] R. Dodds, R. Butler, The phenomena of overtourism: A review, International Journal of Tourism Cities 5 (2019) 519–528. doi:10.1108/IJTC-06-2019-0090.

[2] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, User Modeling and User-Adapted Interaction 30 (2020) 127–158. doi:10.1007/s11257-019-09256-1.

[3] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: Toward more impactful recommender systems research, Ai Magazine 41 (2020) 79–95. doi:10.1609/aimag.v41i4.5312.

[4] A. Banerjee, A. Satish, F. N. Aisyah, W. Wörndl, Y. Deldjoo, Synthtrips: A knowledge-grounded framework for benchmark data generation for personalized tourism recommenders (2025) 3743–3752. URL: https://doi.org/10.1145/3726302.3730321. doi:10.1145/3726302.3730321.

[5] X. Li, S. Wang, S. Zeng, Y. Wu, Y. Yang, A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges, Vicinagearth 1 (2024) 9.

[6] S. K. Sakib, A. B. Das, Challenging fairness: A comprehensive exploration of bias in llm-based recommendations, in: 2024 IEEE International Conference on Big Data (BigData), IEEE, 2024, pp. 1585–1592.

[7] G. Balakrishnan, W. Wörndl, Multistakeholder recommender systems in tourism, Proc. Workshop on Recommenders in Tourism (RecTour 2021) (2021).

[8] H. Abdollahpouri, R. Burke, Multistakeholder recommender systems, in: Recommender systems handbook, Springer, 2021, pp. 647–677. doi:10.1007/978-1-0716-2197-4_17.

[9] Y. Deldjoo, N. Mehta, M. Sathiamoorthy, S. Zhang, P. Castells, J. McAuley, Toward holistic evaluation of recommender systems powered by generative models, SIGIR'25 (2025).

[10] C. Jiang, J. Wang, W. Ma, C. L. Clarke, S. Wang, C. Wu, M. Zhang, Beyond utility: Evaluating llm as recommender, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 3850–3862.

[11] L. Jost, Entropy and diversity, Oikos 113 (2006) 363–375.

[12] J. Chen, S. Saha, M. Bansal, ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 7066–7085. URL: https://aclanthology.org/2024.acl-long.381/. doi:10.18653/v1/2024.acl-long.381.

[13] S. Erlich, N. Hazon, S. Kraus, Negotiation strategies for agents with ordinal preferences, arXiv preprint arXiv:1805.00913 (2018).

[14] T. Olmo, A. Ettinger, A. Bertsch, B. Kuehl, D. Graham, D. Heineman, D. Groeneveld, F. Brahman, F. Timbers, H. Ivison, et al., Olmo 3, arXiv preprint arXiv:2512.13961 (2025).