Mind the Gap: Bridging Behavioral Silos with LLMs in Multi-Vertical Recommendations

Nimesh Sinha*
nimesh.sinha@doordash.com
DoorDash Inc.
San Francisco, California, USA

Raghav Saboo*
raghav.saboo@doordash.com
DoorDash Inc.
San Francisco, California, USA

Martin Wang
martin.wang@doordash.com
DoorDash Inc.
San Francisco, California, USA

Sudeep Das
sudeep.das2@doordash.com
DoorDash Inc.
San Francisco, California, USA

ABSTRACT

In multi-vertical e-commerce platforms like DoorDash, relatively newer product verticals such as grocery and retail present a significant opportunity for personalization innovation. A key challenge lies in solving the "cold start" problem for users. This paper introduces a novel framework for enhancing recommendation quality by transferring knowledge from data-rich verticals (e.g., restaurants at DoorDash) to data-sparse ones. We leverage Large Language Models (LLMs) to perform generative inference, synthesizing sparse, high-dimensional features that encapsulate latent user affinities. Specifically, we employ a hierarchical Retrieval-Augmented Generation (RAG) pipeline to derive multi-level taxonomic features from user restaurant order histories and search queries. These generated features, encoding both long-term cross-vertical preferences and short-term intent, are integrated into a production Multi-Task Learning (MTL) ranking model. We demonstrate through extensive offline and online evaluation that this approach significantly improves personalization and engagement in emerging business verticals, effectively bridging the behavioral data gap.

KEYWORDS

Large Language Models, Recommender Systems, Cross-Domain Personalization, Cold-Start Problem, Feature Engineering, Multi-Task Learning, Ranking

ACM Reference Format:
Nimesh Sinha, Raghav Saboo, Martin Wang, and Sudeep Das. 2025. Mind the Gap: Bridging Behavioral Silos with LLMs in Multi-Vertical Recommendations. In *Proceedings of the second workshop on Generative AI for E-Commerce 2025, September 22, 2025.* ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

Modern e-commerce ecosystems, characterized by multi-vertical marketplaces like DoorDash, continually expand into new domains such as groceries, retail, and beauty. This expansion introduces a critical personalization challenge: how to provide relevant recommendations in these verticals for users with no engagement history. Concurrently, these platforms possess a wealth of behavioral data from established, high-traffic verticals. This data asymmetry presents a valuable opportunity for strategic knowledge transfer.

This work addresses this challenge by conceptualizing user behavior in established verticals as a source of rich, latent signals that can inform preferences in emerging ones. We posit that large language models are powerful tools for semantic feature engineering [2, 10, 13, 14]. They can distill unstructured data, such as restaurant orders and search queries [7, 9], into structured, interpretable representations of user affinity.

Our core contribution is a novel methodology that employs a hierarchical Retrieval-Augmented Generation (RAG) framework [4, 12] to infer user affinities at multiple levels of a product taxonomy. This structured inference pipeline mitigates the risk of hallucination and enhances the fidelity of the generated features. By injecting these LLM derived features into our production Multi-Task Learning (MTL) ranking model, we effectively enrich the user representation, enabling the ranker to discern nuanced cross-vertical preferences even in the absence of direct historical data. This approach directly confronts the cold-start problem [3, 11, 15] and demonstrates a practical path toward building more holistic and adaptive recommender systems.

2 SYSTEM ARCHITECTURE AND METHODOLOGY

Our production recommendation system employs a multi-stage architecture, prominently featuring Two-Tower Embedding (TTE) models for candidate retrieval [5] and a Multi-Task Learning (MTL) model for fine-grained ranking [1, 6, 8]. This research focuses on augmenting the feature space of the MTL ranker to enhance its ability to model cross-domain preferences. We introduce two novel classes of LLM-synthesized features:

- **Long-Term Cross-Vertical Affinity Profile:** Captures a consumer's latent, historical preferences for taxonomy categories, inferred from their cumulative restaurant order history.
- **Short-Term Intent Profile:** Models a consumer's recent, transient interests in specific taxonomy categories, as indicated by their on-platform search activity.

*Both authors contributed equally to this research.

Genaiecom '25, September 22, 2025, Prague, CZ

2.1 Generative Feature Synthesis via Hierarchical RAG

We employ LLMs to map unstructured user activity (restaurant orders, search queries) to our internal, four-level hierarchical product taxonomy (L1–L4; e.g., Dairy & Eggs → Cheese → Hard Cheeses → Cheddar). Using a 20% sample of consumer data from the preceding three months, we execute the hierarchical RAG process depicted in Figure 1.

The process operates via cascaded inference. First, the model identifies broad, top-level (e.g., L1, L2) taxonomic affinities from the input signals. These initial, high-confidence classifications then constrain the generative search space for subsequent, more granular retrievals at lower taxonomy levels (e.g., L3). This iterative refinement strategy enhances precision and relevance by preventing the model from generating plausible but incorrect subcategories. For our MTL ranker, we concentrate on L2 and L3 affinities, as L1 is often too general and L4 suffers from excessive sparsity.

2.2 Prompt Engineering and Inference Control

To structure the model's input, we concatenate historical restaurant names and ordered items in chronological order, prioritizing recent behavior. Search queries are similarly sequenced. This temporal ordering helps the model capture evolving preferences. The prompt is enriched with contextual information, including the target taxonomy structure and anonymized consumer profile attributes.

To ensure deterministic and high-fidelity output, we set the inference temperature to 0.1. Critically, the prompt instructs the model to return a confidence score for each generated affinity and to only output taxonomies exceeding a confidence threshold of 0.8. This acts as a self-correction mechanism, filtering out low-confidence or spurious associations.

Prior to implementing the aforementioned improvements in the prompt, the extracted taxonomies from restaurant orders included some irrelevant categories, resulting in poorer alignment with the actual cuisine. For example, as shown in Table 1, a consumer placing orders from an Indian restaurant was previously associated with less relevant affinities, such as "Sandwiches" which do not accurately capture the nuances of Indian cuisine. After applying the improved prompt engineering techniques, the model instead produced more specific and appropriate taxonomies, such as "Specialty Breads (Naan)" thereby significantly enhancing the relevance of the taxonomic associations.

Prompt Example

```
CONTEXT_SETUP = "You are a recommendation
engine. Given a consumer's order history and an
allowed L3 taxonomy, infer up to 50 relevant
L3 categories. Focus on the cuisine of the
restaurant and use only categories present in
the provided taxonomy (case-insensitive match,
output exact taxonomy spelling). Do not invent
categories. Assign a confidence score in [0,1];
include only categories with confidence >=
0.80. Sort by confidence descending; break
ties alphabetically. If no category meets the
threshold, return an empty list. Output must be
strict JSON"
OPERATING_RULES = "Map items to the most
specific applicable L3. If a dish could map to
multiple categories, choose the best-supported
one. If the result is ambiguous and confidence
<0.80, exclude it."
TAXONOMY_INFO = "The allowed L3 categories are:
[L3 taxonomy list]"
USER_HISTORY = "The consumer has ordered the
following dishes from the restaurants in
chronological order (store || dish): [restaurant
name || dish name], ..."
prompt = CONTEXT_SETUP + OPERATING_RULES + " "
+ USER_HISTORY + " " + TAXONOMY_INFO
```

2.3 Model Selection and Cost Optimization

A cost-performance analysis of various models, including GPT-4o and GPT-4o-mini, revealed that GPT-4o-mini delivered comparable output quality for this specific task at a substantially lower computational cost. To further optimize, we implemented a prompt-caching strategy. The static portion of the prompt (containing instructions and taxonomy) is cached, and only the dynamic user history portion is appended for each inference call. This token-level optimization, combined with a just-in-time feature materialization strategy (updating affinities only upon new user actions), reduced overall computational costs by approximately 80%.

2.4 Feature Materialization and Serving

A daily batch pipeline computes and updates the affinity features. These features are materialized to our data lake for model training and simultaneously propagated to a low-latency online feature store for real-time inference during serving.

2.5 Qualitative Feature Evaluation

Table 2 provides illustrative examples of the taxonomic affinities generated. To quantitatively assess the quality of the features, we conducted two studies, one based on human evaluation and the other based on LLM as a judge with the more powerful model GPT-4o, scoring the relevance of personalization on a 3-point scale. As shown in Table 3 for human evaluation and Table 4 for LLM as a judge evaluation, features derived from search queries exhibit

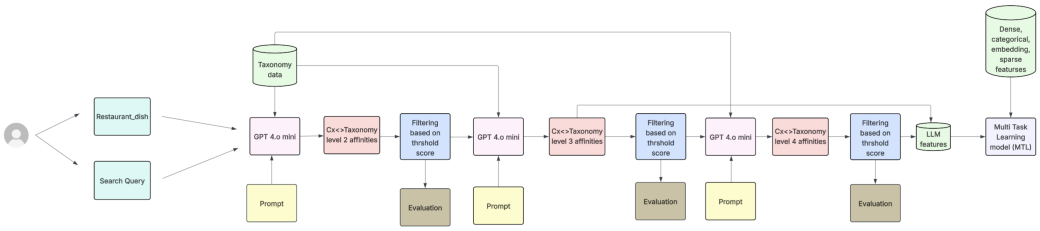

Figure 1: System flowchart illustrating the hierarchical RAG pipeline for generating user affinity features and their integration into the MTL ranking model.

Table 1: Improvements in the results with prompt engineering techniques

Signal Type	Consumer Input History	Before (LLM Output)	After (LLM Output)
Restaurant Orders	Royal Spice: Naan; Butter Chicken; Vegetable Samosa	Sandwiches, Burgers & Wraps, Entrees, Appetizers & Sides, Chicken	Specialty Breads, Naan, Vegetable Sides, Chicken

higher personalization scores, which aligns with the explicit nature of search intent compared to the implicit signals from order history.

2.6 Multi-Task Learning (MTL) Architecture

Our MTL ranker is designed to jointly optimize for multiple objectives (e.g., click-through rate, add-to-cart, purchase). The total loss $\mathcal{L}$ is a weighted sum of individual task losses $\mathcal{L}_t$:

$$\mathcal{L} = \sum_{t=1}^{T} \alpha_t \, \mathcal{L}_t(\hat{y}_t, y_t),$$

where $\hat{y}_t$ is the model's prediction for task t and α_t is a task-specific weight. We augment the model's input feature space by concatenating our LLM-generated features with existing feature vectors:

- $\mathbf{u}_{\text{LLM}} \in \mathbb{R}^d$: The new sparse features representing user affinities, derived from both restaurant orders and search queries.
- $\mathbf{u}_{\text{eng}} \in \mathbb{R}^p$: The existing user engagement feature vector (e.g., historical interactions).
- $\mathbf{i}_{\text{eng}} \in \mathbb{R}^q$: The item feature vector (e.g., category, brand, price).

The augmented user and item vectors are defined as:

$$\mathbf{u}_{\text{aug}} = [\mathbf{u}_{\text{eng}}; \mathbf{u}_{\text{LLM}}], \quad \mathbf{i} = [\mathbf{i}_{\text{eng}}].$$

Variable-length categorical features, such as the lists of taxonomy IDs in $\mathbf{u}_{\text{LLM}}$, are handled by mapping each ID to a dense vector via a shared embedding table. The embeddings corresponding to a list are then aggregated into a fixed-size representation using mean pooling. This technique efficiently handles jagged input tensors and promotes parameter sharing. The final concatenated feature vector feeds into a shared MLP trunk, ϕ, followed by task-specific prediction heads:

$$\mathbf{z} = \phi([\mathbf{u}_{\text{aug}}, \mathbf{i}]), \quad \hat{y}_t = \sigma(\mathbf{w}_t^\top \mathbf{z} + b_t). \tag{1}$$

Here, σ is an activation function (e.g., sigmoid), and $\mathbf{w}_t, b_t$ are the weights and bias of the prediction head for task t.

3 EXPERIMENTAL EVALUATION

We conducted rigorous set of offline and online experiments to measure the impact of LLM-generated features on our item-ranking model trained on three months of data. Offline evaluation was performed on a 15-day holdout using conversion as the ground-truth label, and results are reported with the metrics described in the next section.

3.1 Experimental Setup

We compare the performance of two models using standard evaluation metrics: Area Under the Receiver Operating Characteristic Curve (AUC-ROC) for classification performance and Mean Reciprocal Rank (MRR) for ranking quality.

- **Baseline Model**: The production MTL item ranking model, trained exclusively on historical user engagement and item attribute features.
- **Proposed Model**: An identical MTL architecture augmented with the LLM-derived user affinity features ($\mathbf{u}_{\text{LLM}}$).

3.2 Offline Evaluation and Cohort Analysis

We evaluated performance across the general user population and on two specific cohorts critical to our business: "cold-start" consumers (new to non-restaurant verticals) and "power" consumers (highly active in these verticals). The results, summarized in Figures 2 and 3, demonstrate a consistent and significant performance uplift.

Key findings include:

- **Overall Population:** The proposed model achieved a 4.4% relative lift in AUC-ROC and a 4.8% lift in MRR, indicating a broad improvement in ranking quality.
- **Cold-Start Consumers:** This cohort benefited most from the restaurant order signal, with the combined signals yielding a 4.0% lift in AUC-ROC and a 1.1% lift in MRR. This

Table 2: Illustrative examples of LLM-generated category recommendations from consumer signals.

Signal Type	Consumer Input History	Generated L3 Taxonomy Affinities (LLM Output)
Restaurant Orders	Taco Bell: Cantina Chicken Crispy Taco; Cheese Quesadilla Royal Spice: Cheese Naan; Butter Chicken Starbucks: White Chocolate Mocha	Tacos, Chicken, Cheese, Naan, Specialty Breads, Coffee
Search Queries	Protein bar, drink, pop tart, protein, yogurt, healthy snacks	Cereal & Granola Bars, Packaged Snacks, Yogurt, Juices & Smoothies, Protein Supplements, Nutrition Shakes, Energy Drinks, Sour Cream & Dips

Table 3: Human Evaluation of LLM-Generated Feature Personalization. N=1000 samples per signal.

Signal Source	Not Personalized	Partially Personalized	Highly Personalized
Restaurant Orders	17.7%	29.3%	53.0%
Search Queries	6.8%	22.5%	70.7%

Table 4: LLM Evaluation of LLM-Generated Feature Personalization (GPT-4o). N=1000 samples per signal.

Signal Source	Not Personalized	Partially Personalized	Highly Personalized
Restaurant Orders	15.6%	27.8%	56.6%
Search Queries	8.2%	30.2%	61.6%

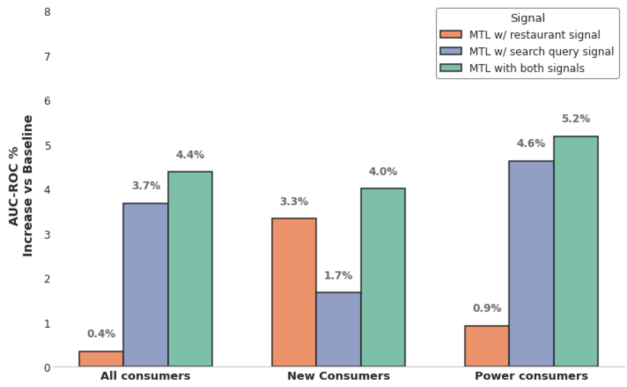

Figure 2: Relative improvement (%) in AUC-ROC for the Proposed Model over the Baseline across different consumer cohorts.

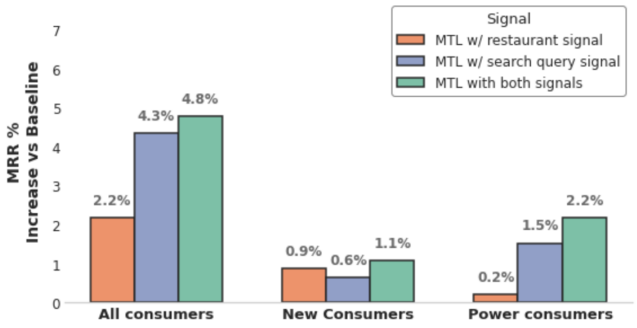

Figure 3: Relative improvement (%) in MRR for the Proposed Model over the Baseline across different consumer cohorts.

validates our hypothesis that historical taste preferences can be effectively transferred across verticals.

- **Power Consumers:** This group saw the largest gains from the search query signal, which captures short-term intent. The model achieved a 5.2% lift in AUC-ROC and a 2.2% lift in MRR, showcasing its ability to adapt to recent user needs.

3.3 Online Shadow Deployment

To validate our offline findings in a production environment, we conducted an online evaluation by shadowing live traffic. The results, shown in Figure 4, were consistent with the offline analysis. For the general population, the LLM-augmented model demonstrated a 4.3% improvement in AUC-ROC and a 3.2% increase in MRR over the baseline, confirming the real-world efficacy of our approach.

4 CONCLUSION AND FUTURE WORK

This work successfully demonstrates the efficacy of leveraging LLMs as a semantic bridge to transfer knowledge from data-rich to

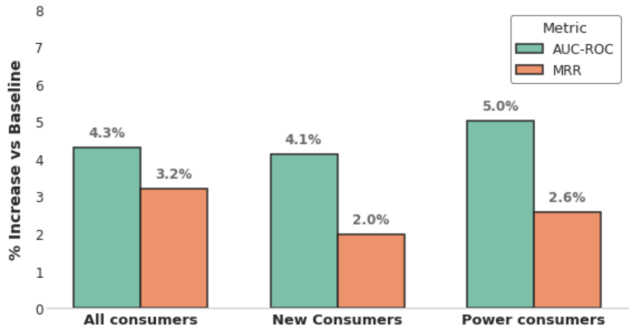

Figure 4: Relative improvement (%) in online shadow traffic metrics for the Proposed Model versus the Baseline.

data-sparse domains within a multi-vertical marketplace. By employing a hierarchical RAG framework to synthesize high-fidelity user affinity features, we significantly enhanced the performance of our production MTL ranking model, particularly for cold-start users.

Our future research agenda will proceed along several promising vectors. First, we will investigate more advanced prompting paradigms (e.g., Chain-of-Thought, self-correction) and explore domain-specific fine-tuning of smaller, more efficient LLMs to further improve feature quality and reduce costs. Second, we plan to integrate these generative features earlier in the recommendation funnel, specifically within the candidate retrieval stage, to create a more synergistic, end-to-end personalized system. Finally, we will explore the temporal dynamics of these affinities to build more adaptive, session-aware recommendation models.

5 AUTHOR BIO

Nimesh Sinha is a Senior Machine Learning Engineer at DoorDash, where he specializes in personalizing consumer experiences for the company's new verticals business. Previously, he built advanced search and personalization models for Walmart's e-commerce platform. Nimesh has also contributed to machine learning initiatives at Bird and Barnes & Noble Education. He holds a master's degree in Data Science from the University of San Francisco and an integrated master's in Applied Physics from IIT-ISM Dhanbad.

REFERENCES

[1] Olivier Chapelle, Pannagadatta Shivaswamy, Srinivas Vadrevu, Kilian Weinberger, Ya Zhang, and Belle Tseng. 2010. Multi-task Learning for Boosting with Application to Web Search Ranking. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1189–1198.

[2] Rachel M. Harrison, Anton Dereventsov, and Anton Bibin. 2023. Zero-shot Recommendations with Pre-trained Large Language Models for Multimodal Nudging. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1535–1542.

[3] Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S. Yu. 2025. Large Language Model Simulator for Cold-Start Recommendation. In *Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM 2025)*. https://doi.org/10.48550/arXiv.2402.09176 arXiv:2402.09176, accepted by WSDM 2025.

[4] Haoyu Huang, Yongfeng Huang, Junjie Yang, Zhenyu Pan, Yongqiang Chen, Kaili Ma, Hongzhi Chen, and James Cheng. 2025. Retrieval-Augmented Generation with Hierarchical Knowledge. *arXiv preprint arXiv:2503.10150* (2025). https://doi.org/10.48550/arXiv.2503.10150

[5] Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based Retrieval in Facebook Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2553–2561.

[6] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-Task Learning with Multi-Gate Mixture-of-Experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1930–1939.

[7] Luyi Ma, Nimesh Sinha, Parth Vajge, Jason H. D. Cho, Sushant Kumar, and Kannan Achan. 2021. Event-based Product Carousel Recommendation with Query-Click Graph. In *2021 IEEE International Conference on Big Data (Big Data)*. 1–7. https://doi.org/10.48550/arXiv.2402.03277 arXiv:2402.03277.

[8] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2018)*. https://doi.org/10.48550/arXiv.1804.07931 arXiv:1804.07931, accepted by SIGIR-2018.

[9] Nimesh Sinha. 2020. SPIR: Some Practical Item-based Recommendations. In *Proceedings of the Industry-Related Symposium (IRS 2020)*. https://irsworkshop.github.io/2020/publications/paper_13_%20Sinha_SPIR.pdf

[10] Arpita Vats, Vinija Jain, Rahul Raja, and Aman Chadha. 2024. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. *arXiv preprint arXiv:2402.18590* (2024). https://arxiv.org/abs/2402.18590

[11] Rongyao Wang, Veronica Liesaputra, and Zhiyi Huang. 2025. A Survey on LLM-based News Recommender Systems. *arXiv preprint arXiv:2502.09797* (2025). https://arxiv.org/abs/2502.09797

[12] Shu Wang, Yixiang Fang, Yingli Zhou, Xilin Liu, and Yuchi Ma. 2025. ArchRAG: Attributed Community-based Hierarchical Retrieval-Augmented Generation. *arXiv preprint arXiv:2502.09891* (2025). https://doi.org/10.48550/arXiv.2502.09891

[13] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. Llmrec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM 2024)*. 806–815.

[14] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A Survey on Large Language Models for Recommendation. *World Wide Web* 27, 5 (2024), 60.

[15] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, et al. 2024. Recommender Systems in the Era of Large Language Models (LLMs). *IEEE Transactions on Knowledge and Data Engineering* (2024). to appear.