Addressing Intent-Impression Misalignment in Marketplaces via Query-Level Multinomial Blending

Beatrice Musizza
Vinted
Berlin, Germany
beatrice.musizza@vinted.com

Thorsten Krause
Vinted
Berlin, Germany
thorsten.krause@vinted.com

Abstract

Understanding user intent is central to effective search and recommendation systems. However, user queries are often ambiguous and can reflect multiple possible needs. This challenge becomes more pronounced in marketplaces where each query may relate to items from several distinct categories. In such settings, uneven supply and legacy relevance patterns can result in overexposure of established categories, while newer or smaller ones are overlooked. Addressing query intent in the context of multiple categories is therefore crucial for ensuring fair and relevant results. This study examines the issue in the context of a large C2C marketplace and addresses it using an online query-aware re-ranking framework that integrates a Query Intent Model with Query-Level Multinomial Blending (MB) to rebalance exposure. The approach predicts each query's category intent and uses the resulting probability distribution to guide exposure across categories in the final ranked list while preserving relevance within each category. In a large-scale online A/B test, Query-Level MB calibrates exposure across categories in response to dynamic user intent, increasing visibility and engagement for target categories without necessitating modifications to the existing retrieval, ranking, or post-ranking stack.

CCS Concepts

• **Information systems → Retrieval models and ranking**.

Keywords

Marketplace Search, Query Intent Modeling, Re-ranking, Exposure Control, Multinomial Blending

ACM Reference Format:
Beatrice Musizza and Thorsten Krause. 2026. Addressing Intent-Impression Misalignment in Marketplaces via Query-Level Multinomial Blending. In *Proceedings of Online & Adaptive Recommender System (OARS) @ ACM RecSys 2026 (OARS@RecSys '26).* ACM, New York, NY, USA, 7 pages.

1 Introduction

In modern search and recommendation systems, ranking algorithms must balance the relevance objective with diverse stakeholder requirements [8, 9]. Because these systems are typically optimized on historical user interactions [1], they can systematically favor items with abundant past engagement [12]. Accurate search results depend on understanding user intent, yet user queries often reflect multiple possible intents. For instance, a query for a phone model may indicate interest in either the device or its accessories. In consumer-to-consumer (C2C) marketplaces, uneven supply across categories exacerbates this challenge. Prioritizing established categories can transfer legacy relevance patterns to newer categories, where such patterns may not generalize effectively.

Take the example of Vinted's Electronics category. Internal analyses revealed that, for many high-intent queries, the distribution of impressions across categories did not correspond to downstream engagement signals such as clicks and transactions. This phenomenon is termed *intent-impression misalignment*. For example, queries like *"iPhone 14"* or *"Nintendo Switch"* frequently resulted in significant visibility for accessories and other non-target categories, whereas user engagement was concentrated on the intended device categories. Although users could often adjust results through additional filtering and sorting, the initial search experience remained inconsistent with inferred intent.

Several families of re-ranking approaches have been developed to address impression-engagement misalignment; however, none proved practical or feasible in the present context: diversity-based rerankers, such as MRR [4] or intent-aware diversification [2], optimize for inter-item dissimilarity or intent coverage rather than calibrating to a target category distribution. For the queries in question, results were often already diverse but misallocated, and these methods do not provide a mechanism to enforce a specific category mix. Additionally, re-scoring individual items would override within-category serving logic established by existing post-ranking interventions. An alternative considered was incorporating the query-intent signal as a feature in the final-stage ranker. However, this approach was ineffective in this context: Electronics represented only a small portion of training interactions, leading the ranker to overgeneralize from other categories, and offline NDCG remained largely unchanged, both overall and for Electronics queries.

MB offers a promising approach for aligning impressions and engagement while meeting business requirements. MB enables practical control over inter-category-level visibility in rankings without altering intra-category visibility [12]. However, its original formulation assumes a single static target exposure distribution for all queries. In marketplace search, the optimal category mix is highly dependent on query-specific intent (for example, *"iPhone 14"* versus *"iPhone 14 case"*), necessitating a query-conditioned solution. In a C2C marketplace where inventory turns over rapidly and category supply shifts continuously, a serving-time, query-conditioned mechanism is better suited than a statically tuned one.

In this work, we propose a query-aware re-ranking framework that combines a Query Intent Model with *Query-Level MB*. The Query Intent Model predicts a probability distribution over catalog categories for each query, which Query-Level MB then uses to guide category-level exposure in the final ranked list while preserving the original ranking order within each category bucket. The method

OARS@RecSys '26, Minneapolis, Minnesota, USA
2026.

is lightweight, compatible with existing post-ranking rules, and can be deployed on top of a production ranking pipeline. Unlike static exposure-control schemes that fix a single target distribution, our method re-parameterizes the blending distribution per query at serving time, adapting category exposure online to the intent inferred for each incoming request.

The proposed approach is implemented and evaluated at Vinted, one of Europe's largest C2C re-commerce marketplaces with a catalog comprising billions of items. The evaluation focuses on Electronics-related queries to assess the impact of MB on category visibility, user engagement, and downstream transaction metrics. The contributions of this work are as follows: (1) documentation of intent–impression misalignment in a specific category within a large marketplace; (2) introduction of a query-aware re-ranking framework that combines a Query Intent Model with MB, deployed as a self-contained post-ranking layer; and (3) validation of the approach through online A/B testing in Vinted's large-scale production environment.

2 Empirical Motivation

To understand the extent of intent-impression misalignment in our platform, we analyzed search performance in Vinted's Electronics category, focusing on four product types: *mobile phones, tablets, laptops, and consoles*. For each product type, we identified the top 100 queries by search-attributed transactions and compared the distribution of impressions, clicks, and search-attributed transactions across catalog categories.

The analysis revealed a consistent gap between which categories users see and which they actually engage with. In mobile-phone-related queries, phone accessories commanded the largest share of visibility; however, actual devices received a disproportionately higher share of engagement. The phone device click share was 2.2 times its impression share, indicating substantial misalignment. A similar pattern appeared in the video game console category, where actual device items drew a click share over 1.5 times greater than non-devices. Users often compensated for this misalignment through additional refinement actions. Compared with the platform average, high-intent Electronics sessions showed nearly six times higher usage of the price filter, almost double the usage of the category filter, and a 27% relative increase in sorting by *"newest first"*. Because accessories are both more abundant and substantially cheaper than the devices they complement, the elevated price-filter usage is consistent with users manually correcting for this low-price bias.

Taken together, these findings suggest that the desired inventory was underexposed in the initial ranking, motivating a query-aware mechanism for category-level exposure control.

3 Problem Setting

In this section, we introduce our notation and formally define the problem and the objective.

First, we consider a search system that, given a query q, retrieves an initial ranking of items.

$$L(q) = (d_1, d_2, \ldots, d_n),$$

where each item d_i belongs to a catalog category $c(d_i) \in C$. In our setting, C denotes a set of category buckets used for exposure control (e.g., *mobile phones*, *phone cases*, *consoles*, and *laptops*). The initial ranking $L(q)$ is produced by the production ranking pipeline, which can include multiple stages such as retrieval, learned ranking, and post-ranking interventions [11, 14]. Then, for a given category $c \in C$, let

$$L_c(q) = (d_{c,1}, d_{c,2}, \ldots, d_{c,n_c})$$

denote the subsequence of items from category c, preserving their original order in $L(q)$. Thus, each category defines a bucket of items whose internal ranking is inherited from the baseline system. Our goal is to construct a final ranking

$$\tilde{L}(q) = (\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_n)$$

such that category-level exposure, particularly in the top-ranked positions, better reflects query-specific intent. To represent this intent, we associate each query with a target distribution

$$p(q) = \big(p_1(q), p_2(q), \ldots, p_{|C|}(q)\big),$$

where $p_c(q) \geq 0$ and $\sum_{c \in C} p_c(q) = 1$. Each component $p_c(q)$ specifies the desired share of exposure for category c for query q.

The objective is not to rescore individual items, but to transform the initial ranking $L(q)$ into a final ranking $\tilde{L}(q)$ such that:

- the category composition of the top-ranked results is guided by $p(q)$;
- the original within-category ordering L_c is preserved for all categories.

In other words, we seek a query-aware mechanism that controls how categories are interleaved in the final ranking without reordering items within each category bucket. In the next section, we describe how the Query Intent Model estimates $p(q)$ and how Query-Level MB constructs the final ranking.

4 Method

We present the proposed query-aware exposure-control framework, which comprises two components: a Query Intent Model that estimates a query-specific target exposure distribution over categories, and Query-Level MB, which utilizes this distribution to interleave category buckets in the final ranking while maintaining within-category order.

4.1 Query Intent Model

The original MB application assumed a fixed set of available categories and a predefined exposure share for each category. In contrast, within the search context, both the item categories present in a ranking and the desired exposure shares are query-dependent. Furthermore, the set of possible queries is unbounded, making it infeasible to define exposure shares in advance. Consequently, extending MB to search necessitates a method capable of predicting the desired exposure for any incoming query on an ad hoc basis.

To address this challenge, we developed a Query Intent Model that predicts a click-based category intent distribution $p(q)$ for any incoming query q. Similar models have been successfully applied in industrial contexts to address various problems. For instance, some approaches estimate the likelihood of different user behavioral

intents, such as exploration [3, 10, 15]. Chang et al. [5] predict a distribution over expected item regions for serving international users. Zhou et al. [16] and Lin et al. [13] estimate user intents across categories using laboratory study data to improve relevance predictions. Ahmadvand et al. [3] propose an approach that predicts relevant categories rather than returning a distribution. While overall query intent models are common in large-scale search applications, to our knowledge, there is no existing application of such models to MB.

Consistent with existing approaches, our implementation models query intent as a multi-class classification task. Given our focus on integration with the MB framework, we prioritized identifying a simple architecture that achieves sufficient query intent prediction accuracy, followed by evaluation within the MB framework. The resulting system first tokenizes the input query using the BERT tokenizer [7], then processes the tokens through a recurrent neural network with gated recurrent units [6] to obtain a latent query representation. This representation is linearly transformed to produce a vector of category-level logits s, and the final probability estimate $p(q)$ is obtained by applying the softmax function to s. The model is trained on labels derived from historical search-to-click logs: each query with at least ten clicks is mapped by majority vote to its most-clicked catalog category, with all non-electronics clicks consolidated into a single `NON_ELECTRONICS` class. The training set was composed of historic search interactions, with a heavily skewed distribution in which `NON_ELECTRONICS` accounts for the majority of queries. Because this distribution is heavily imbalanced, aggregate accuracy is correspondingly inflated; we therefore rely on macro-averaged F1 (Table 1), which weights all classes equally, to confirm that the minority electronics categories are learned rather than ignored. During inference, the model dynamically predicts the expected choice distribution.

We evaluated the model on a *golden set* of 13,500 queries selected from frequent and rare electronic queries, as well as queries we previously found to be sensitive to user intent. The set is dominated by no single class: `NON_ELECTRONICS` accounts for 36% of queries and video games 22%, with the remaining categories (phone cases, headphones, mobile phones, smartwatches, and various minor subcategories each below 6%). Each query intent was manually classified by human reviewers. Table 1 shows the query intent model's performance on the golden set. In addition to the quantitative evaluation on the golden set, we conducted extensive qualitative analysis into the query intent model's performance and which queries it could not accurately classify, finding that most mis-classification occurred in the context of brands that are popular in one segment and less popular in another, such as the term *"Spiderman"*, which refers to movies most of the time and sometimes to games or merchandise. Other instances where the model failed to predict the human-annotated label were queries that different annotators labeled differently. Another challenge was queries that could relate to both fashion and non-fashion items such as the query *"snow runner"* which can relate to a video game or running shoes for the winter season. Overall, we concluded that the model's performance is sufficient to proceed with the downstream MB task.

Averaging	F1 Score	Precision	Recall
Weighted	98.13	98.25	98.12
Micro	98.12	98.12	98.12
Macro	89.24	91.65	93.69

Table 1: Query intent model performance on the golden evaluation set in percent.

4.2 Query-Level Multinomial Blending

To enforce the predicted intent distribution $p(q)$ at serving time, we propose *Query-Level MB*. While prior applications of MB have typically relied on static, global target exposure [12], our approach dynamically parameterizes the blending distribution based on the query-specific exposure returned by the Query Intent Model introduced in the previous section. In this section, we present our implementation of Query-Level MB, and discuss the compatibility with existing business rules.

4.2.1 Implementation. We assume that for some incoming query q, the serving system retrieved the baseline ranking $L(q)$ and the Query Intent Model computed the target exposure distribution $p(q)$. The first step of the MB procedure is to partition the baseline ranking $L(q)$ into category-specific buckets $L_c(q)$. The final ranking $\tilde{L}(q)$ is then constructed sequentially by sampling, at each position, a category c according to the target distribution $p(q)$ and appending the next available item from the corresponding bucket $L_c(q)$. Because items are always drawn from the front of their category bucket, the original within-category ordering induced by $L(q)$ is preserved.

Formally, let $A_t(q) \subseteq C$ denote the set of categories with non-empty buckets before position t is filled. For each position t, a category is sampled according to

$$\Pr(c_t = c \mid q) = \frac{p_c(q)}{\sum_{c' \in A_t(q)} p_{c'}(q)} \quad \text{for } c \in A_t(q).$$

Let

$$m_c(t) = \sum_{\tau < t} \mathbb{I}[c_\tau = c]$$

denote the number of items from category c that have already been placed before position t. If category c_t is selected at position t, the next item in the final ranking is

$$\tilde{d}_t = d_{c_t, m_{c_t}(t)+1}.$$

Thus, the algorithm always selects the next not-yet-placed item from the chosen category bucket. The procedure continues until the desired ranking length is reached or all buckets are exhausted.

Algorithm 1 summarizes the serving-time procedure, and Figure 1 illustrates the same mechanism on a simplified example.

4.2.2 Compatibility with Existing Business Rules. A critical challenge in industrial search systems is that the baseline ranking $L(q)$ is rarely the strict output of a single semantic relevance model. Instead, $L(q)$ may already reflect a combination of model scores and post-ranking interventions, such as sponsored item insertions, policy constraints, or other business rules. To preserve the integrity of these downstream business rules, our blending mechanism does not operate on raw model scores. Rather, it treats the items' final positions in $L(q)$ as the proxy for overall utility and only controls how

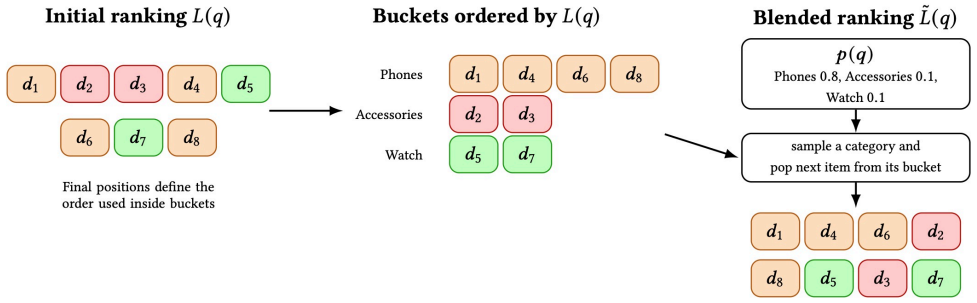

Figure 1: Query-Level MB partitions $L(q)$ into buckets ordered by baseline position and interleaves them according to $p(q)$.

Algorithm 1 Query-Level MB

(1) **Input:** initial ranking $L(q)$, category function $c(\cdot)$, target distribution $p(q)$, ranking length n.
(2) **Output:** blended ranking $\tilde{L}(q)$.
(3) Partition $L(q)$ into category buckets $\{L_c(q)\}_{c \in C}$, preserving the original order within each bucket.
(4) Initialize $\tilde{L}(q) \leftarrow [\ ]$.
(5) For $t = 1, \ldots, n$:
  (a) Set $A_t(q) \leftarrow \{c \in C : L_c(q) \neq \emptyset\}$.
  (b) If $A_t(q) = \emptyset$, stop and return $\tilde{L}(q)$.
  (c) Set $Z_t(q) \leftarrow \sum_{c' \in A_t(q)} p_{c'}(q)$.
  (d) Set $\pi_c(q) \leftarrow p_c(q)/Z_t(q)$ for all $c \in A_t(q)$.
  (e) Sample $c_t \sim \text{Categorical}(\{\pi_c(q)\}_{c \in A_t(q)})$.
  (f) Select the next not-yet-placed item from $L_{c_t}(q)$ and append it to $\tilde{L}(q)$.
(6) Return $\tilde{L}(q)$.

category buckets are interleaved. In this way, query-level exposure can be adjusted while preserving both within-category relevance and pre-existing within-category serving logic.

5 Experimental Setup

The real-world impact of the proposed query-aware exposure-control framework was evaluated by deploying the Query Intent Model and Query-Level MB in a large-scale online A/B test at Vinted.

The control group received the standard search experience, with ranked lists generated by the production retrieval pipeline, followed by deep-funnel relevance re-rankers and existing post-ranking business rules, without explicit intent-based category blending. The treatment group received search results processed through the standard pipeline, with the final slate additionally processed using Query-Level MB. The category exposure proportions for MB were generated dynamically by the Query Intent Model for each query.

The scope of queries affected by the treatment was dynamically determined by the Query Intent Model, which was trained specifically on the Electronics category. For any given search query, the model outputs intent probabilities for specific Electronics categories, and a `NON_ELECTRONICS` class. MB was applied based on a confidence threshold. If the Query Intent Model assigned a probability greater than 80% to the `NON_ELECTRONICS` class, the MB step was bypassed. In these cases, category blending was expected to have minimal effect on the final slate and could introduce unnecessary serving cost or noise when the model was not sufficiently confident. For all other queries, MB was applied using the predicted probabilities over the Electronics categories. The threshold was set as a conservative operating point. Misapplying MB degrades the slate, while bypassing MB preserves the baseline performance.

The evaluation focused on whether the framework increased visibility and engagement for the Electronics category and its higher-value subcategories, including mobile phones and consoles. Visibility was measured by impression share, and engagement was measured by click share. Global guardrail metrics were monitored to verify that the intervention did not reduce user satisfaction or introduce friction within the broader marketplace ecosystem. Statistical significance of metric changes was assessed using a threshold of $p < 0.05$.

The experiment was conducted over a three-week period on live production traffic using a randomized A/B testing framework.

6 Results and Discussion

In this section, we present and analyze the effects of our Query-Level MB implementation on category visibility, user engagement, and transaction-level guardrails. We also highlight practical insights from the experiment and outline directions for future research.

6.1 Experiment Results

The online effects of Query-Level MB are reported for the subset of relevance-sorted search traffic where the Query Intent Model identified Electronics-related intent. Unless otherwise specified, the MB treatment is compared against the control condition. All lifts reported in this section are statistically significant unless explicitly stated otherwise. Table 2 summarizes the results.

Before analyzing aggregate lifts, we verified that Query-Level MB effectively translated the predicted intent distribution into realized exposure. For treated queries, the Query Intent Model predictions and observed category exposure in the top three search results were strongly correlated, with a Pearson correlation coefficient of 0.89 over category shares. This result validates the serving mechanism by confirming that categories with higher predicted intent received proportionally higher exposure, and that the candidate pool generally contained sufficient items per category for MB to approximate the target distribution.

Table 2: Relative changes in impression and click shares under MB.

Metric (vs control, stat-sig)	Lift
Visibility (Impression share)	
Target category, top 24	+5.5%
Target category, overall	+3.5%
Target subcategory (exact match)	+18.5%
Target subcategory (complement)	−11.1%
Engagement (click share)	
Target category, top 24	+6.0%
Target category, overall	+4.4%
Target subcategory (exact match)	+3.8%
Target subcategory (complement)	−5.2%

Query-Level MB increased the visibility of Electronics items, particularly in the most prominent positions of the ranking. Overall, Electronics impression share increased by +3.5% relative to the control group. The shift was targeted within Electronics: mobile-phone impression share increased by +18.5%, while phone-case impression share decreased by 11.1%. This indicates that MB changed the exposure structure in the intended direction, increasing visibility for higher-intent device categories while reducing exposure for less intent-aligned accessory categories.

The visibility gains also translated into higher engagement with Electronics. Overall, Electronics click share increased by +4.4% relative to the control group. Within the phone subcategory, mobile-phone click share increased, while phone-case click share decreased by 5.2%. These results suggest that Query-Level MB shifted user engagement toward the categories predicted to better match query intent.

Transaction-level guardrails remained stable throughout the three-week experiment. We did not observe statistically significant changes in guardrail metrics. By design, MB has limited impact on non-electronics inventory. Queries confidently classified as non-electronics bypass blending, and within treated queries, non-electronics items retain exposure proportional to their predicted intent share, losing visibility only when the query is predicted to target Electronics categories. Consistent with this, overall Buyer Transactions per Active User remained stable, indicating no measurable degradation outside the Electronics category.

Query-Level MB increased click share for target device categories while leaving Electronics and overall Buyer Transactions per Active User statistically unchanged. We read this as evidence that the intervention redistributes engagement toward under-exposed, intent-aligned categories rather than generating additional purchases. A transaction lift was neither the objective nor a precondition for success: the goal was to serve intent-aligned inventory more prominently without degrading the downstream funnel, which the stable guardrails confirm.

Moreover, the pattern is robust and generalizes across markets with different baseline Electronics shares, affirming that the effect of Query-Level MB is not driven by a single country.

6.2 Discussion and Limitations

We studied intent-impression misalignment in a large multi-category marketplace, where abundant accessory inventory can receive disproportionate exposure even when user intent is directed toward higher-value device categories. To address this problem, we proposed a query-aware exposure-control framework that combines a Query Intent Model with Query-Level MB. The method uses the predicted query-specific intent distribution to interleave category buckets while preserving the original within-category ranking and existing post-ranking business logic.

In a large-scale online A/B test on Electronics-related search traffic, the intervention increased Electronics visibility and engagement, especially in the top-ranked positions. The gains were not uniform across categories: MB shifted exposure and clicks away from accessories such as phone cases and toward more intent-aligned device categories, such as mobile phones. These effects were observed across multiple markets, suggesting that query-conditioned exposure control can generalize beyond a single country or traffic segment. Similar increases in engagement and alignment with target exposure metrics have been previously observed when applying MB under static target exposure [12]. We therefore conclude that MB retains its robustness and effectiveness even when applied at a query level where target exposure is estimated based on logged user feedback.

We also note that the three-week window and the granularity of the transaction guardrails limit our ability to detect small transaction effects, and transaction impact may also lag exposure changes. We therefore frame our contribution as improving engagement alignment for less-exposed categories without harming transactions, rather than as a transaction-driving intervention.

Nonetheless, we encountered several practical challenges during implementation. First, Query-Level MB depends on an accurate Query Intent Model. If the model assigns probability mass to incorrect categories, MB may reduce ranking quality by increasing exposure to mismatched buckets. We address this risk by bypassing MB when the model confidently classifies a query as `NON_ELECTRONICS`. A systematic sensitivity analysis of this threshold remains for future work, and future deployments should continue to monitor calibration, query drift, and category-specific failure modes.

Second, the method is constrained by the candidates produced by the upstream retrieval stage [12]. Since Query-Level MB only reorders items present in the retrieved list $L(q)$, the target distribution $p(q)$ can be fully realized only if sufficient candidate items are available for each category. Otherwise, the renormalization step in Algorithm 1 redistributes probability mass across the remaining categories, aligning results with intent as much as the retrieved pool allows.

Third, the current Query Intent Model is trained on hard category labels, so each training example is associated with a single target category. This can produce overly concentrated intent distributions for exploratory or multi-intent queries. For example, *"Nintendo Switch"* may express intent for the console, but also for games or accessories; if the model assigns nearly all of its probability mass to consoles, Query-Level MB may reduce the useful category variety more than desired. Moreover, when the Query Intent Model is substantially less complex than the final-stage ranker, the ranker

can sometimes return more aligned exposures itself than when combined with MB. For example, for the query *"iPhone 14 Pro"*, a personalized ranker could know that a user is also interested in phone cases based on their recent related searches, while a non-personalized query-intent model would still return mostly phones, potentially negatively impacting transactions. Finally, Query-Level MB is stochastic: repeated issuances of the same query may produce different category interleavings. In our context, result slates already vary across users and over time due to personalization and rapid C2C inventory turnover, while within-category order is always preserved. Deterministic variants, such as quota-based interleaving that matches $p(q)$ in aggregate, can be used where reproducibility is required. The stochastic formulation, however, offers the advantage of closed-form propensities for off-policy evaluation [12], enabling offline tuning of intent distributions and the bypass threshold without additional A/B cycles.

These challenges suggest several directions for future work. The present Query Intent Model is quite simple (it uses a GRU on BERT token representations) and has been trained on logged clicks without taking position bias into account. There is therefore scope for improvement, for example by investigating more advanced model architectures or by applying debiasing methods. On the one hand, it might be possible to include additional context such as market-specific supply conditions, user preferences, or session-level refinement behaviour in order to make exposure control more adaptive and so that the system can better tell the difference between general exploratory queries and those with a strong purchase intent. On the other hand, future work could involve the development of uncertainty-aware blending and fallback strategies in order to minimise the risk of over-correcting when the query intent is unclear. Thirdly, training the model with softer target distributions could result in smoother intent estimates and would allow the blended ranking to keep some of the useful variability for exploratory queries. Lastly, a promising approach would be to jointly optimise retrieval, ranking, and blending in order to ensure that intent-aligned categories are better represented in the candidate set before the post-ranking exposure control is applied.

Overall, our findings demonstrate that query-conditioned blending offers a practical and deployable mechanism for controlling category exposure in broad marketplaces. By integrating lightweight intent prediction with order-preserving bucket interleaving, this approach enhances the alignment between displayed results and inferred user intent. It also functions as a self-contained post-ranking layer that can be launched, tuned, and rolled back independently of the existing retrieval and ranking stack.

Acknowledgments

We thank Amir Masoud Sefidian for his work on the Query Intent Model, Justina Bartulevičienė for her help in analyzing the experiment results, and Benediktas Kazanavičius and Monika Narkutė for their contributions to the MB implementation. We are also grateful to Daniele Lubrano and Aleksas Kateiva for their feedback and suggestions during the development of this work. Finally, we thank Gediminas Žiulius and Vaidotas Daukša for their support during the project.

References

[1] Aman Agarwal, Kenta Takatsu, Ivan Zaitsev, and Thorsten Joachims. 2019. A General Framework for Counterfactual Learning-to-Rank. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 5–14. doi:10.1145/3331184.3331202

[2] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying Search Results. In *International Conference on Web Search and Data Mining (WSDM)* (international conference on web search and data mining (wsdm) ed.). Association for Computing Machinery, Inc. https://www.microsoft.com/en-us/research/publication/diversifying-search-results/

[3] Ali Ahmadvand, Surya Kallumadi, Faizan Javed, and Eugene Agichtein. 2020. JointMap: Joint Query Intent Understanding For Modeling Intent Hierarchies in E-commerce Search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 1509–1512. doi:10.1145/3397271.3401184

[4] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. Association for Computing Machinery, New York, NY, USA, 335–336. doi:10.1145/290941.291025

[5] Yi Chang, Ruiqiang Zhang, Srihari Reddy, and Yan Liu. 2011. Detecting Multilingual and Multi-Regional Query Intent in Web Search. *Proceedings of the AAAI Conference on Artificial Intelligence* 25, 1 (Aug. 2011), 1134–1139. doi:10.1609/aaai.v25i1.8074

[6] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1724–1734. doi:10.3115/v1/D14-1179

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[8] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A Deep Relevance Matching Model for Ad-hoc Retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (Indianapolis, Indiana, USA) *(CIKM '16)*. Association for Computing Machinery, New York, NY, USA, 55–64. doi:10.1145/2983323.2983769

[9] Malay Haldar, Prashant Ramanathan, Tyler Sax, Mustafa Abdool, Lanbo Zhang, Aamir Mansawala, Shulin Yang, Bradley Turnbull, and Junshuo Liao. 2020. Improving Deep Learning for Airbnb Search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 2822–2830. doi:10.1145/3394486.3403333

[10] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management* 44, 3 (2008), 1251–1266. doi:10.1016/j.ipm.2007.07.015

[11] Hang Li, Tie-Yan Liu, and ChengXiang Zhai. 2009. Learning to rank for information retrieval (LR4IR 2009). *SIGIR Forum* 43, 2 (Dec. 2009), 41–45. doi:10.1145/1670564.1670571

[12] Jan Malte Lichtenberg, Giuseppe Di Benedetto, and Matteo Ruffini. 2024. Ranking Across Different Content Types: The Robust Beauty of Multinomial Blending. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 823–825. doi:10.1145/3640457.3688059

[13] Yiu-Chang Lin, Ankur Datta, and Giuseppe Di Fabbrizio. 2018. E-commerce Product Query Classification Using Implicit User's Feedback from Clicks. In *2018 IEEE International Conference on Big Data (Big Data)*. 1955–1959. doi:10.1109/BigData.2018.8622008

[14] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. doi:10.1561/1500000019

[15] Yuyan Wang, Cheenar Banerjee, Samer Chucri, Fabio Soldo, Sriraj Badam, Ed H. Chi, and Minmin Chen. 2025. Beyond Item Dissimilarities: Diversifying by Intent in Recommender Systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto ON, Canada) *(KDD '25)*. Association for Computing Machinery, New York, NY, USA, 2672–2681. doi:10.1145/3690624.3709429

[16] Ke Zhou, Ronan Cummins, Martin Halvey, Mounia Lalmas, and Joemon M. Jose. 2012. Assessing and Predicting Vertical Intent for Web Queries. In *Advances in*

Information Retrieval, Ricardo Baeza-Yates, Arjen P. de Vries, Hugo Zaragoza, B. Barla Cambazoglu, Vanessa Murdock, Ronny Lempel, and Fabrizio Silvestri (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 499–502.