

Behavior Retrieval plus Response Generation for Interpretable Conversational Personalized Recommendation

Qi Xin

Management Information Systems, University of Pittsburgh, United States of America
qix29@pitt.edu

Corresponding author, qix29@pitt.edu

Abstract— This paper presents an empirical study of interpretable conversational personalized recommendation built from two tightly coupled layers: a behavior retrieval layer for candidate ranking and a response generation layer for grounded recommendation wording. The study is motivated by recent large language model surveys in recommendations, but the implementation remains deterministic so that candidate selection, evidence extraction, and wording can be inspected separately. We used the 2010-2011 partition of Online Retail II for transaction-grounded next-basket recommendation and a retail/service subset of the Bitext customer-support 27K corpus for conversational response generation. The behavior retrieval layer combined popularity, repeat purchase memory, recency-weighted memory, item-to-item collaborative filtering, and user-neighborhood scoring. The response layer compared intent-majority, TF-IDF retrieval, intent-conditioned retrieval, and a constrained fusion verbalizer. On 1,999 test users, the hybrid behavior retriever achieved Hit@10 = 0.7282, Recall@10 = 0.1982, NDCG@10 = 0.2855, and MRR@10 = 0.4467, outperforming the strongest baseline RecencyRepeat by 2.03% on Hit@10 and 6.34% on NDCG@10. On 1,989 Bitext test utterances, intent-conditioned retrieval produced the best standalone verbalization quality with BLEU-4 = 0.1699, ROUGE-L = 0.3431, intent accuracy = 0.9915, and slot F1 = 0.8239. Grounded recommendation responses reached Top1Hit = 0.3252, HistoryGroundRate = 1.0000, and BundleValidity = 0.9980. The results show that retrieval-first conversational recommendation can combine ranking accuracy, grounding faithfulness, and low-overhead response realization without relying on unconstrained generation.

Keywords: Behavior retrieval, conversational recommender systems, explainable recommendation, next-basket prediction, personalized recommendation.



This work is licensed under a [CC BY-NC](https://creativecommons.org/licenses/by-nc/4.0/). Copyright ©2026 by Author. Published by Universitas Riau.

INTRODUCTION

Classical recommender systems were developed to reduce information overload in large item catalogs, and the core technical line from user-based collaborative filtering to item-based similarity and hybrid recommendation remains central to modern e-commerce [1], [2], [3], [4]. Those methods solved an essential retrieval problem: how to map incomplete evidence of user preference into a short-ranked list that contains at least one item the user is likely to choose. In Online Retail settings, however, this retrieval problem is no longer sufficient by itself. Modern recommendation interfaces are conversational; customers expect a recommendation to be accompanied by a rationale, and commercial systems are frequently evaluated not only on

whether an item appears in the top part of the list but also on whether the recommendation fits the basket context, can be explained with evidence, and can be articulated in natural language. That shift is particularly visible in commerce, where a suggestion such as "buy another mug" is less useful than a grounded conversational suggestion such as "this mug matches the two floral kitchen items you purchased recently; a cheaper alternative is also available and these napkins bundle well." The movement from static ranking toward interactive explanation therefore changes both the system architecture and the evaluation protocol. It requires a recommender to retrieve an item and justify the retrieval in a form that respects user history, product relationships, and communicative clarity.

The ranking side of this problem is well grounded in the recommender systems literature. Gain-based ranking metrics such as NDCG were designed precisely because the position of a correct item matters in ranked retrieval [5]. Pairwise ranking objectives such as Bayesian personalized ranking formalized the difference between positive and unobserved feedback [10], while recurrent and neural collaborative approaches extended the literature toward sequential and basket-aware user modeling [11], [12]. Yet most production retail data remain noisy, sparse, and highly repetitive. Customers reorder the same products, revisit narrow category bands, and form baskets whose composition reflects convenience, seasonality, and budget constraints. In such settings, simple retrieval signals-repeat purchase counts, recent purchases, and item co-occurrence-often remain highly competitive, especially when the target problem is next-basket prediction rather than long-horizon content discovery. The present work therefore uses interpretable retrieval signals as the basis of a conversational recommender because they preserve explicit links between user behavior and system output.

The conversational side of the problem introduces a different set of requirements. Conversational recommender systems aim to exchange information with users over multiple turns, update the system's belief about user intent, and present items in a natural and adaptive manner [13], [15], [16]. A conversation can ask clarifying questions, respond to objections, explain substitutes, and suggest bundles. These capabilities make conversational recommenders attractive for product discovery and service automation, but they also create a reproducibility challenge. Many recent systems report strong qualitative examples, yet the exact contribution of the language layer remains difficult to isolate when ranking, explanation, and generation are entangled in large end-to-end models. The present study addresses that issue by separating the problem into two explicit layers. The first layer performs behavior retrieval from transactional evidence. The second layer transforms the retrieved evidence into a customer-facing response. This decomposition preserves interpretability: every response is anchored to an identified candidate item and an explicit set of evidence items from user history.

Explainability research offers an additional reason to prefer this decomposition. Explanations in recommender systems are valuable when they increase transparency, trust, persuasiveness, and user control, but those gains depend on whether the explanation is faithful to the underlying recommendation process [8], [9], [14]. A recommendation that says "you may like this item because you bought similar items" is not especially useful unless the system can point to concrete historical evidence and show that the claim is consistent with the ranking mechanism that produced the item. In commerce, this requirement is even stricter because users make money-based decisions and frequently compare substitutes or complementary products. The explanation should therefore do more than paraphrase popularity. It should identify the historical items that supported the recommendation, distinguish between direct reordering and cross-item association, and clarify whether a bundle or lower-priced alternative is reasonable.

Recent literature on large language models strengthens the relevance of this two-layer design. Surveys of LLM-based recommendations consistently report that large language models are useful as feature encoders, prompt able controllers, and generative interfaces, but they also

note efficiency costs, evaluation instability, and difficulty in guaranteeing faithful grounding [17], [18], [19], [20]. LLM-oriented systems, such as Chat-REC, illustrate the appeal of interactive and explainable recommendations through prompting, candidate compression, and natural language reasoning [21]. At the same time, recommendation remains fundamentally tied to candidate quality. A fluent generator cannot rescue a poor candidate set, and a hallucinated explanation can actively damage user trust. The implication is straightforward: a conversational recommender benefits when the retrieval layer is strong enough to constrain the language layer.

The empirical gap targeted here is not a lack of models, but a lack of coherent end-to-end studies on retail recommendation that report measured results for both ranking and response behavior under one consistent protocol. In much of the literature, candidate generation is studied on one dataset, natural language generation is studied on another, and explanation quality is illustrated by a few examples rather than measured across a full test set. This study therefore uses public datasets whose size supports complete local experimentation. The 2010-2011 partition of Online Retail II provides transaction histories with timestamps and product descriptions [22]. A retail/service subset of Bitext's customer-support 27K corpus provides intent-labeled response pairs for deterministic response modeling [23]. The two corpora are not treated as natively aligned dialogue turns. Instead, each dataset evaluates the part of the pipeline it can support: next-basket retrieval on transaction data and response verbalization on conversational service data.

The paper makes four concrete contributions. First, it introduces a behavior retrieval plus response generation architecture, abbreviated BRRG, that decomposes conversational recommendation into a ranking stage and a grounded verbalization stage. Second, it evaluates the selected public datasets under explicit preprocessing and chronological splitting. Third, it adds grounding-oriented response metrics to standard ranking and text-generation metrics, thereby evaluating whether a response remains tied to real purchase evidence. Fourth, it analyzes accuracy, segment robustness, catalog coverage, response quality, explanation faithfulness, and inference-time deployment cost in one consistent system.

METHOD

The study used a two-corpus design because the target task had two irreducible components: transaction-grounded candidate retrieval and conversational response realization. Table 1 summarizes the datasets.

Table 1. Dataset Overview and Role Assignment in the BRRG Pipeline

Attribute	Value	
Dataset	Online Retail II (2010–2011)	Bitext 27K Subset
Primary role	Behavior retrieval	Response generation
Rows used	396,196	19,887
Users/intents	4,334	20
Items/categories	3,656	9
Units	Baskets	Utterance-response pairs
Unit count	15,265	19,887
Span/Categories	2010-12-01 to 2011-12-09	ORDER, REFUND, CONTACT, INVOICE, PAYMENT, FEEDBACK, DELIVERY, SHIPPING, CANCEL

The transaction corpus was the 2010-2011 partition of Online Retail II, a public invoice-level e-commerce dataset describing a United Kingdom-based non-store retailer [22]. The conversational corpus was Bitext's customer-support 27K dataset, from which the retail/service subset retained order, refund, contact, invoice, payment, feedback, delivery, shipping, and

cancellation-related interactions [23]. This design matched each dataset to the evaluation it could support. Online Retail II contained real baskets, products, prices, and timestamps, so it was appropriate for next-basket ranking and grounded explanation analysis. Bitext contained natural-language instructions, responses, categories, and intents, so it was appropriate for response-generation benchmarking.

For Online Retail II preprocessing, the 2010-2011 partition was cleaned into invoice-level baskets before model training. Rows with missing customer identifiers, non-positive quantities, non-positive prices, cancellation invoices, and stock codes or descriptions corresponding to postage, manual adjustments, banking charges, test rows, or other non-merchandise service entries were removed. Line items were then grouped by customer and invoice, quantities were aggregated inside each invoice, the median observed price per stock code was retained as the product price summary, and the most frequent textual description observed for each stock code was used as the canonical description. This process produced 396,196 cleaned line items, 4,334 customers, 3,656 unique stock codes, and 15,265 invoice baskets. Users with at least three baskets were retained for evaluation. For each qualified user, all but the last two baskets formed the training history; the penultimate basket formed the validation target; and the last basket formed the test target. Table 2 reports the resulting counts.

Table 2. Preprocessing, Filtering, and Split Statistics used in All Reported Experiments

Component	Value (Unit)
Online Retail II cleaned line items	396,196
Online Retail II customers	4,334
Online Retail II unique stock codes	3,656
Constructed baskets	15,265
Users with at least 3 baskets	1,999
Training baskets	11,267
Validation cases	1,999
Test cases	1,999
Training item vocabulary	3,541
Bitext filtered pairs	19,887
Bitext filtered intents	20
Bitext filtered categories	9
Bitext training pairs	15,909
Bitext validation pairs	1,989
Bitext test pairs	1,989

The Bitext preprocessing followed the same principle of aligning the dataset with the method. The full corpus contained 26,872 instruction-response pairs across customer-service intents and categories. The filtered retail/service subset contained 19,887 pairs, 20 intents, and 9 categories. A stratified split produced 15,909 training pairs, 1,989 validation pairs, and 1,989 test pairs. Each example preserved the original instruction text, intent label, category label, and reference response. The purpose of this corpus in the pipeline was not to predict next-basket items directly. Its role was to benchmark verbalizers that transform structured recommendation evidence into concise, domain-compatible customer-facing language.

Figure 1 places the two layers of BRRG at the beginning of the method description. User transaction histories are first cleaned and split chronologically. Retrieval baselines then score candidates through popularity, memory, item-to-item collaborative filtering, user-neighborhood transfer, and hybrid fusion. The response side uses the Bitext subset to learn intent-conditioned response retrieval, while the final constrained verbalizer receives only the ranked item and evidence slots exposed by the behavior retrieval layer.

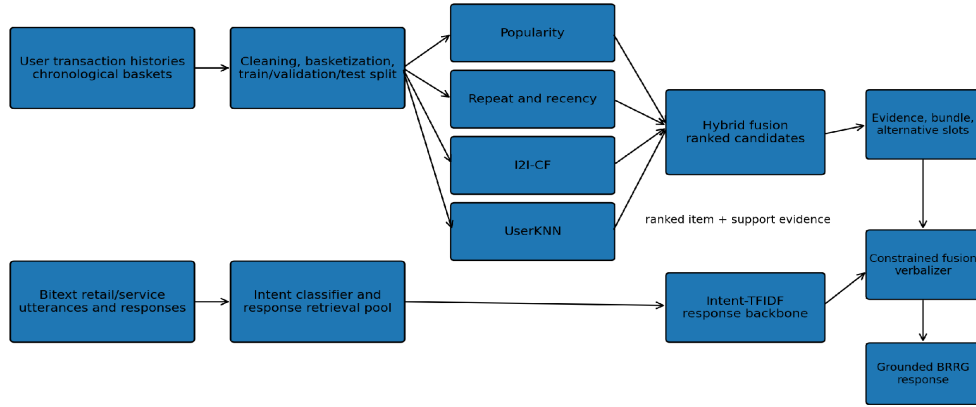


Figure 1. End-to-end BRRG pipeline from user transaction histories to constrained grounded responses

The behavior retrieval layer evaluated seven rankers. POP ranked items by overall training-basket frequency and served as a non-personalized baseline. Repeat ranked items by user-specific historical purchase counts and represented a pure memory model. LastBasket used membership in the user's most recent training basket and captured short-term repetition. RecencyRepeat applied exponentially decayed weights to historical baskets with a decay factor of 0.7, thereby emphasizing recent purchases while keeping older evidence available. I2I-CF computed cosine similarity from the basket-item co-occurrence matrix and scored a candidate by its similarity to items in the user's history. UserKNN represented each user as an item-frequency profile over training baskets and ranked items using nearest-neighbor transfer from similar users. BR-Hybrid combined the five validation-strong signals: popularity, Repeat, RecencyRepeat, I2I-CF, and UserKNN. The hybrid score is shown in (1).

$$Score(u, i) = 0.1 \cdot Pop(i) + 0.2 \cdot Repeat(u, i) + 0.3 \cdot RecencyRepeat(u, i) + 0.2 \cdot I2I - CF(u, i) + 0.2 \cdot UserKNN(u, i) \quad (1)$$

All recommendation scores were computed on the training histories only. For I2I-CF, a basket-item matrix was built and cosine similarity was derived between item columns. For UserKNN, a user-item matrix from training baskets was used to find users with similar purchase distributions and transfer their item preferences to the target user. For memory models, purchase counts were extracted directly from the chronological basket history. The output of the behavior retrieval layer was not merely a ranked list. For the top-ranked candidate, the retained context also included the top-20 recommendation list, the user's basket history length, target basket size, and per-user metric contributions. These artifacts supported segment analysis, list-quality analysis, and grounded explanation construction.

The response generation layer evaluated four verbalization strategies on the Bitext subset. IntentMajority predicted the intent with a linear classifier but returned the majority response template for that intent. TFIDF ignored the intent label and retrieved the nearest training instruction in a TF-IDF vector space using word unigrams and bigrams. Intent-TFIDF first predicted the intent with a LinearSVC classifier and then restricted retrieval to the predicted-intent pool, combining semantic narrowing with lexical retrieval. BRRG-Gen was an exploratory constrained verbalizer that fused a retrieved Bitext response pattern with recommendation-

specific slots. The text representation used a TF-IDF vectorizer with word 1-2 grams, min_df = 2, max_features = 10,000, and lowercase normalization. The intent classifier used LinearSVC(random_state = 42). BLEU-4 [6], ROUGE-L [7], intent accuracy, and slot F1 were measured against the held-out reference responses. The best standalone verbalizer on validation and test splits was Intent-TFIDF, so it was used as the response backbone in the final BRRG pipeline.

The grounded explanation module transformed recommendation evidence into conversational justifications. Given a target user and the highest-ranked candidate from BR-Hybrid, the module identified up to two supporting history items by scoring the similarity between the candidate and the user's historical items with the same item-similarity matrix used by I2I-CF. Direct repeats received an additional bonus, which allowed the explanation generator to distinguish reorders from related items. The module then searched the top-ranked recommendation list for a bundle item with positive similarity to the candidate and for a cheaper lexical alternative that overlapped with the candidate description. The final response contained up to four components: the recommended item, evidence items from history, an optional bundle, and an optional lower-priced alternative. The wording used deterministic templates to keep the response faithful to the retrieved evidence.

The recommendation layer, response layer, and explanation layer were evaluated separately because each stage optimizes a different objective. For recommendation, Hit@K, Recall@K, NDCG@K, and MRR@K were reported at K = 5, 10, and 20. Let U denote the evaluated users, Target(u) the held-out basket for user u, TopK(u) the top-K recommendation list, and FirstRank(u) the rank of the first relevant item. The ranking metrics are defined in (2)-(5).

$$\text{Hit@K} = \text{mean over } u \text{ in } U \text{ of } 1[\text{TopK}(u) \text{ intersects Target}(u)] \quad (2)$$

$$\text{Recall@K} = \text{mean over } u \text{ in } U \text{ of } |\text{TopK}(u) \text{ intersects Target}(u)| / |\text{Target}(u)| \quad (3)$$

$$\text{NDCG@K} = \text{mean over } u \text{ in } U \text{ of } \text{DCG@K}(u) / \text{IDCG@K}(u) \quad (4)$$

$$\text{MRR@K} = \text{mean over } u \text{ in } U \text{ of } 1/\text{FirstRank}(u), \text{ with } \text{FirstRank}(u) \leq K \quad (5)$$

For response generation, BLEU-4 and ROUGE-L measured surface similarity to reference responses, intent accuracy measured whether the verbalizer preserved the task intent, and slot F1 measured entity-level correctness. For explanation grounding on Online Retail II, Top1Hit measured whether the item described in the final response appeared in the actual next basket; HistoryGroundRate measured whether every cited evidence item truly appeared in the user's history; EvidenceRecall measured agreement between cited evidence and the highest-similarity support items; BundleValidity measured whether a mentioned bundle item came from the top-20 recommendation list; AlternativeValidity measured whether the named alternative was cheaper than the candidate; and Actionability measured whether the text gave an explicit next-step cue.

To quantify whether the observed ranking gains were stable, a paired bootstrap comparison was run between BR-Hybrid and RecencyRepeat on the 1,999 test cases. The bootstrap used 2,000 samples with replacement and computed the mean per-user difference in Hit@10, Recall@10, NDCG@10, and MRR@10. User-segment analyses by history length and target-

basket size, catalog coverage and personalization statistics, and response confusion analysis were then computed from the same evaluation artifacts. Table 3 lists the fixed implementation settings.

Table 3. Fixed Implementation Settings and Hyperparameters used for the Reported Experiments

Module	Parameter	Setting
Retail preprocessing	Invoice filter	Positive quantity, positive price, non-null customer, non-cancel invoice, non-service stock codes
Basket construction	Basket key	CustomerID + Invoice
Retail split	Per-user chronological split	All but last two baskets for training; penultimate for validation; last for test
Eligibility	Minimum baskets per user	3
Repeat scorer	Base score	Historical purchase frequency per item
RecencyRepeat scorer	Decay	0.7
I2I-CF	Similarity	Cosine similarity on basket-item co-occurrence
UserKNN	Profile	User-item interaction vector from training baskets
Hybrid ranker	Weights	pop 0.1, repeat 0.2, recency 0.3, i2icf 0.2, userknn 0.2
Ranking evaluation	Metrics	Hit@K, Recall@K, NDCG@K, MRR@K for K={5,10,20}
Bitext filter	Categories	ORDER, REFUND, CONTACT, INVOICE, PAYMENT, FEEDBACK, DELIVERY, SHIPPING, CANCEL
Text retrieval	Vectorizer	TF-IDF word 1-2 grams, min_df=2, max_features=10000
Intent classifier	Model	LinearSVC(random_state=42)
Response evaluation	Metrics	BLEU-4, ROUGE-L, intent accuracy, slot F1
Explanation evaluation	Metrics	Top1Hit, HistoryGroundRate, EvidenceRecall, BundleValidity, AlternativeValidity, Actionability

RESULT AND DISCUSSION

The data characteristics explain why simple memory signals were strong but not sufficient. Figures 2 and 3 show that purchase history length and basket size were both highly skewed. Many customers had only a few baskets, but a meaningful tail of repeat customers produced deeper histories that supported personalized retrieval. Basket size was likewise variable, with a concentration around modest basket sizes and a long right tail of large wholesale-style orders. On the retail side, 396,196 cleaned line items were condensed into 15,265 baskets, and 1,999 customers met the minimum-history threshold for chronological evaluation. On the Bitext side, the filtered subset retained 19,887 utterance-response pairs across 20 intents and 9 categories. These counts made the ranking and response summaries meaningful while keeping the experiments reproducible on standard local hardware.

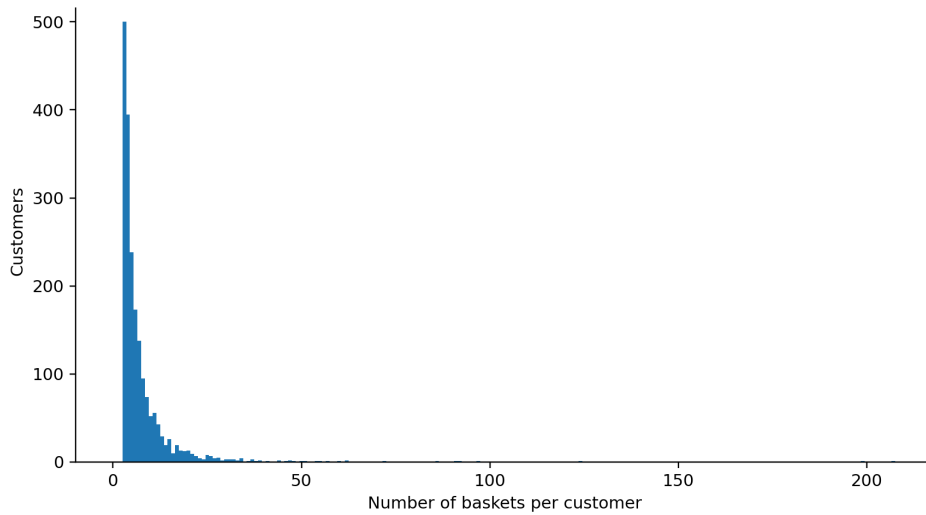


Figure 2. Distribution of customer purchase-history length measured as baskets per customer

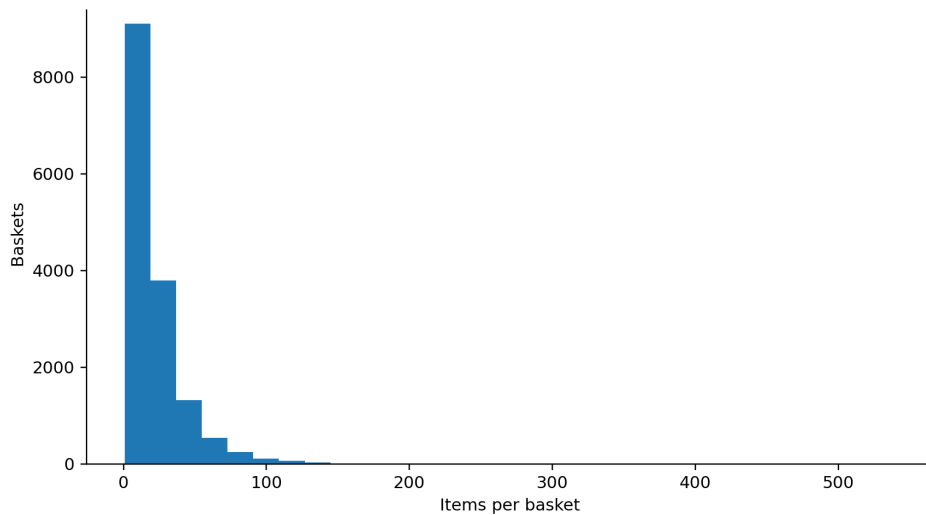


Figure 3. Distribution of basket size in the cleaned retail corpus

Table 4 reports the main next-basket ranking results at $K = 10$, and Table 5 expands the comparison to $K = 5, 10$, and 20 . The strongest overall model was BR-Hybrid, which achieved $\text{Hit@10} = 0.7282$, $\text{Recall@10} = 0.1982$, $\text{NDCG@10} = 0.2855$, and $\text{MRR@10} = 0.4467$. The strongest baseline was RecencyRepeat with $\text{Hit@10} = 0.7137$ and $\text{NDCG@10} = 0.2685$. Repeat and RecencyRepeat were already competitive, confirming that retail repurchase behavior carried substantial predictive power. Nevertheless, the hybrid layer added measurable gains by combining repetition with collaborative evidence. Compared with RecencyRepeat, BR-Hybrid improved Hit@10 by 2.03% and NDCG@10 by 6.34%. Compared with POP, Hit@10 increased from 0.3639 to 0.7282, roughly doubling practical hit probability.

Figure 4 visualizes the per-model ranking comparison for Hit@10 and NDCG@10 . This pattern is consistent with recent next-basket recommendation studies, which show that repeat and explore items should be distinguished when interpreting NBR performance [24] and that repeat consumption contributes substantially to recall-oriented performance in grocery-style recommendation settings [25]. The present result therefore does not contradict prior neural or

repeat-aware NBR research; instead, it shows that in the Online Retail II setting, simple memory signals remain strong while a transparent hybrid of repeat, recency, item-to-item, and user-neighborhood evidence can still add stable early-ranking gains. Because previous work also notes that NBR results are difficult to compare directly across different datasets, preprocessing rules, and evaluation protocols [26], the main contribution here is not a universal numerical superiority claim, but a reproducible and interpretable comparison showing that BR-Hybrid improves the strongest local baseline under the same chronological split.

Table 4. Main Next-basket Recommendation Results on the Online Retail II Test Set at K=10

Model	Hit@10	Recall@10	NDCG@10	MRR@10
BR-Hybrid	0.7282	0.1982	0.2855	0.4467
RecencyRepeat	0.7137	0.1903	0.2685	0.4248
Repeat	0.7077	0.1811	0.2645	0.4286
UserKNN	0.5961	0.1373	0.2044	0.3477
LastBasket	0.5846	0.1321	0.1836	0.3017
I2I-CF	0.5110	0.0920	0.1563	0.2661
POP	0.3639	0.0348	0.0688	0.1661

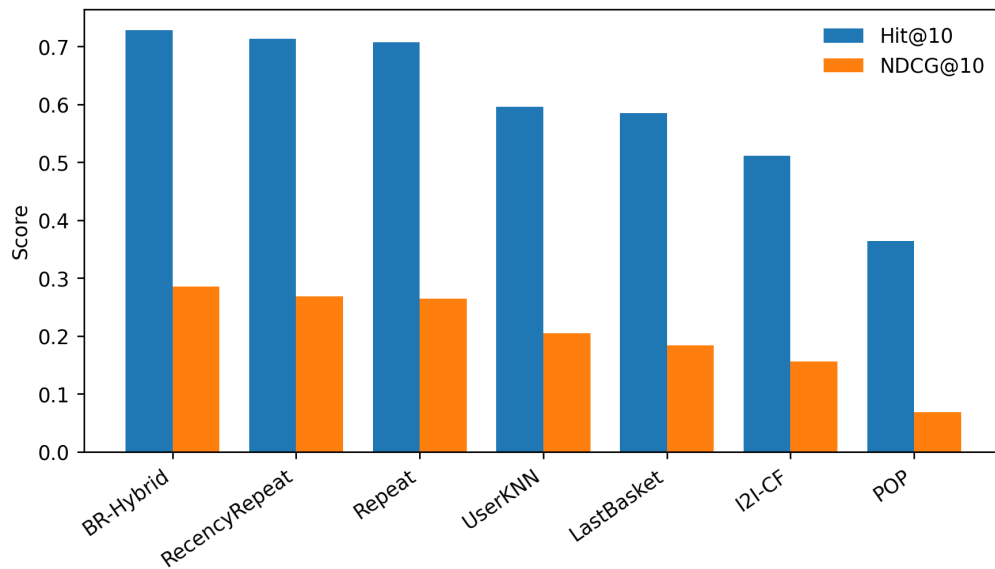


Figure 4. Hit@10 and NDCG@10 across ranking models on the Online Retail II test set

The relatively low Recall@10 of the pure I2I-CF baseline follows from transaction-data sparsity. The cleaned basket-item space contained 15,265 baskets and 3,656 unique items, giving 55,808,840 possible basket-item cells. Even using 396,196 cleaned line items as a conservative upper bound for nonzero cells, the implied density is only 0.710%, and aggregation of duplicate item lines inside baskets makes the effective binary co-occurrence density lower. In this setting, item-item cosine similarity has limited shared-basket evidence for long-tail products and tends to emphasize frequent co-purchases. It can therefore place related but non-target complements in the top-10 list while missing items that the customer simply repeats. This explains why I2I-CF achieved a lower Recall@10 than Repeat and RecencyRepeat. The hybrid model mitigated this weakness by allowing repeat and recency memory to dominate when collaborative neighborhoods were sparse, while still using I2I-CF and UserKNN to surface useful complements and substitutes when co-occurrence evidence was available. This finding also helps position BRRG relative to prior NBR studies. Earlier work has shown that many apparent gains in NBR come from correctly modeling repeat behavior rather than from unconstrained exploration [24], [25]. The present results support that conclusion, but add a system-level extension: the same repeat and collaborative evidence

used to improve ranking is retained as explicit support for downstream explanation generation. Thus, the novelty is not simply another ranking model for sparse baskets; it is the preservation of the ranking evidence chain for interpretable conversational response construction.

Table 5. Detailed Ranking Results Across Multiple Cutoffs on the Online Retail II Test Set

Model	H@5	H@10	H@20	R@5	R@10	R@20	N@5	N@10	N@20
BR-Hybrid	0.6061	0.7282	0.8148	0.1312	0.1982	0.2785	0.2900	0.2855	0.2915
RecencyRepeat	0.5906	0.7137	0.8038	0.1326	0.1903	0.2684	0.2757	0.2685	0.2760
Repeat	0.5916	0.7077	0.7993	0.1261	0.1811	0.2590	0.2740	0.2645	0.2707
UserKNN	0.4830	0.5961	0.7107	0.0898	0.1373	0.1970	0.2124	0.2044	0.2071
LastBasket	0.4615	0.5846	0.6627	0.0922	0.1321	0.1737	0.1822	0.1836	0.1865
I2I-CF	0.3879	0.5110	0.6356	0.0541	0.0920	0.1486	0.1607	0.1563	0.1607
POP	0.2558	0.3639	0.4610	0.0217	0.0348	0.0559	0.0719	0.0688	0.0692

Figure 5 and Table 6 provide a more granular view of ranking depth. BR-Hybrid reached Hit@5 = 0.6061, Hit@10 = 0.7282, and Hit@20 = 0.8148, while Recall@20 rose to 0.2785. The curves show that the hybrid advantage persisted as K increased, indicating that the model did not merely move one good item upward by chance. Table 6 adds a paired bootstrap comparison against RecencyRepeat. The mean NDCG@10 improvement was 0.0170 with a 95% bootstrap interval of [0.0102, 0.0233], and the probability that the Hit@10 improvement was positive was 0.967. The raw margin over RecencyRepeat was not large in absolute terms, but the bootstrap analysis showed that the improvement was stable on NDCG@10, MRR@10, and Recall@10.

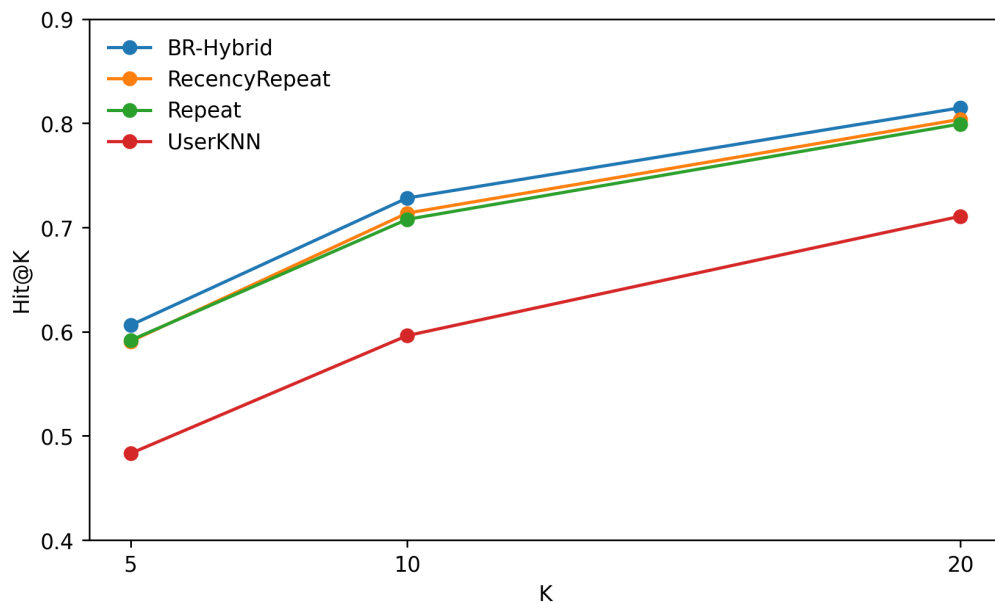


Figure 5. Hit@K curves for the strongest ranking models

Table 6. Paired Bootstrap Comparison between BR-Hybrid and RecencyRepeat on the Online Retail II Test Set

Metric	Mean diff.	95% CI low	95% CI high	P(diff>0)
Hit@10	0.0145	-0.0005	0.0295	0.9670
NDCG@10	0.0170	0.0102	0.0233	1.0000
MRR@10	0.0219	0.0086	0.0355	1.0000
Recall@10	0.0079	0.0021	0.0135	0.9995

Segment analysis clarified where the gains came from. Table 7 reports Hit@10 grouped by purchase-history length. BR-Hybrid achieved Hit@10 = 0.6820 for users with exactly two historical baskets, 0.7184 for users with three to four baskets, and 0.7621 for users with at least five baskets. The strongest improvement over simpler baselines appeared in the long-history segment, where collaborative cues had more opportunity to reinforce repeated-purchase memory. RecencyRepeat slightly exceeded BR-Hybrid in the medium-history group, reaching 0.7294 versus 0.7184, which indicates that when history is short but not extremely short, recency alone can capture much of the usable signal.

Table 7. Hit@10 by History-length Segment on the Online Retail II Test Set

Model	2 Baskets	3-4 Baskets	5+ Baskets
BR-Hybrid	0.6820	0.7184	0.7621
I2I-CF	0.4540	0.5142	0.5416
LastBasket	0.5700	0.6044	0.5785
POP	0.3740	0.3766	0.3487
RecencyRepeat	0.6460	0.7294	0.7413
Repeat	0.6480	0.6899	0.7552
UserKNN	0.5940	0.6187	0.5808

Table 8 reports Recall@10 by target-basket size. BR-Hybrid was strongest in every target-size segment, but recall declined monotonically as target baskets became larger because recovering a larger fraction of a large basket is intrinsically harder.

Table 8. Recall@10 by Target-basket-size Segment on the Online Retail II Test Set

Model	1-5 items	6-10 items	11-20 items	21+ items
BR-Hybrid	0.3759	0.2539	0.1485	0.0846
I2I-CF	0.1091	0.1440	0.0885	0.0574
LastBasket	0.2500	0.1746	0.1009	0.0527
POP	0.0430	0.0427	0.0323	0.0271
RecencyRepeat	0.3699	0.2421	0.1409	0.0772
Repeat	0.3546	0.2198	0.1343	0.0766
UserKNN	0.2508	0.1825	0.1045	0.0609

Table 9 examines recommendation-list quality beyond simple accuracy. POP covered only 0.0028 of the item catalog at top-10 depth because it repeatedly returned the same globally popular items. BR-Hybrid increased catalog coverage to 0.4578 while maintaining the best ranking accuracy. The top-10 recommendation lists of BR-Hybrid also remained highly personalized, with a personalization score above 0.97. Its intra-list diversity was lower than LastBasket, reflecting a tighter focus on coherent basket continuations rather than diversity for its own sake. In retail cart building, that focus is often useful because restocking and complementary purchases tend to be semantically concentrated.

Table 9. Catalog Coverage, List Diversity, and Personalization of Top-10 Recommendation Lists

Model	Coverage@10	Unique items	Mean popularity	ILD@10	Pers.@10
RecencyRepeat	0.6665	2360	261.1636	0.9124	0.9909
Repeat	0.6560	2323	287.5403	0.9085	0.9896
LastBasket	0.6326	2240	175.6603	0.9187	0.9580
BR-Hybrid	0.4578	1621	417.3177	0.8566	0.9736
I2I-CF	0.3869	1370	438.0217	0.7743	0.9361
UserKNN	0.2200	779	528.0265	0.7990	0.9330
POP	0.0028	10	909.9000	0.8407	0.0000

The conversational verbalization results are reported in Table 10 and visualized in Figures 6 and 7. Intent-TFIDF achieved the best standalone response quality with BLEU-4 = 0.1699, ROUGE-L = 0.3431, intent accuracy = 0.9915, and slot F1 = 0.8239. The high intent accuracy shows that the intent classifier was not the main bottleneck; the more meaningful differences appeared in lexical fidelity and slot preservation. Pure TF-IDF retrieval performed well, but intent conditioning improved ROUGE-L and slot F1 by removing semantically close but functionally mismatched responses. The exploratory BRRG-Gen fusion verbalizer achieved BLEU-4 = 0.1548, below the intent-conditioned retrieval baseline. This result indicates that recommendation-specific grounding and lexical overlap with Bitext reference responses are not identical objectives. The final system therefore used Intent-TFIDF as the response backbone and delegated product-specific grounding to the explanation composer. Compared with prior conversational recommender research, this result supports evaluating the language component separately from the ranking component. Jannach [27] notes that CRS evaluation is difficult because such systems combine multiple learning components with a natural-language interface, and therefore require assessment of both component-level quality and overall system behavior. The BRRG results follow that recommendation by reporting text metrics, intent accuracy, slot preservation, and grounding metrics separately. In contrast to LLM-augmented systems such as Chat-REC, which use prompting to improve interactivity and explainability [21], BRRG keeps response realization retrieval-conditioned and deterministic. This design sacrifices some open-ended stylistic flexibility, but it makes the link among intent, retrieved product evidence, and final wording easier to audit.

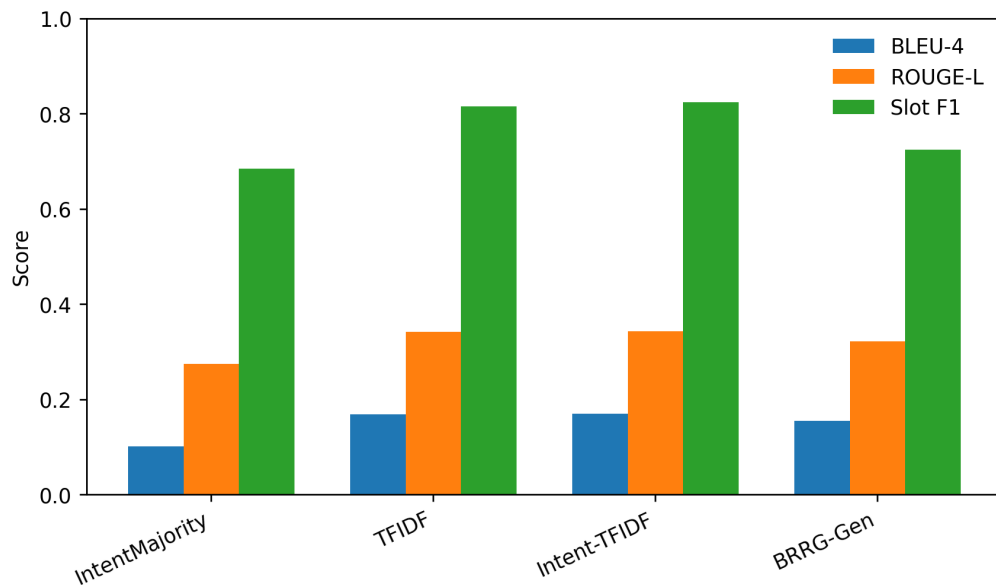


Figure 6. Response-generation comparison on the Bitext retail/service subset

Table 10. Response-generation Results on the Bitext Retail/Service Subset

Model	BLEU-4	ROUGE-L	Intent acc.	Slot F1
IntentMajority	0.1017	0.2740	0.9915	0.6841
TFIDF	0.1687	0.3415	0.9718	0.8160
Intent-TFIDF	0.1699	0.3431	0.9915	0.8239
BRRG-Gen	0.1548	0.3224	0.9915	0.7249

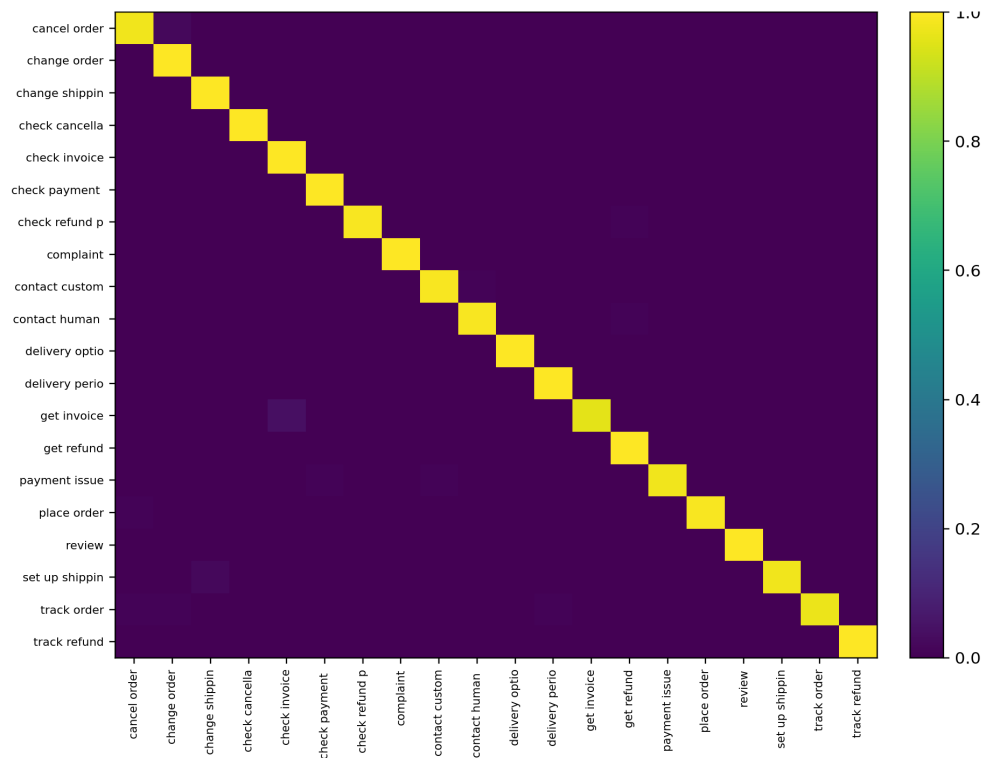


Figure 7. Row-normalized confusion matrix for the intent classifier on the Bitext test split

Grounded explanation quality is reported in Table 11 and summarized in Figure 8. The full BRRG explanation module reached Top1Hit = 0.3252, clearly above POP-Generic and I2I-CF-1Evidence, while maintaining HistoryGroundRate = 1.0000 and EvidenceRecall = 1.0000. Every cited evidence item in the BRRG responses came from the user's actual history, and the cited items matched the highest-similarity support items identified by the retrieval layer. BundleValidity reached 0.9980, showing that the bundling statement was almost always consistent with the model's own recommendation list. AlternativeValidity reached 0.6493, indicating that roughly two thirds of generated alternative statements named an actual lower-priced option with lexical similarity to the candidate. The remaining weakness is therefore price-aware and semantic alternative selection, not historical grounding. Relative to earlier explainable recommendation studies, which emphasize transparency, persuasiveness, trust, and faithfulness [8], [9], [14], the main contribution of BRRG is a measured evidence-control mechanism rather than only a more fluent explanation template. Many conversational and LLM-augmented recommenders illustrate explanations through selected examples, whereas the present evaluation checks whether cited history items, bundle items, and alternative items are supported by the retrieval artifacts for all evaluated test users. This makes the discussion more directly comparable to prior explainability work while highlighting the novelty of joint reporting: BRRG evaluates item correctness and explanation faithfulness under the same Online Retail II protocol.

Table 11. Explanation Faithfulness and Actionability on the Online Retail II Test Set

Model	Top1Hit	Ground	Ev. recall	Bundle	Alt.	Action	Avg chars
BRRG	0.3252	1.0000	1.0000	0.9980	0.6493	1.0000	385.9065
I2I-CF-1Evidence	0.1696	1.0000	1.0000	0.0000	0.0000	0.0000	158.2311
POP-Generic	0.0910	0.0000	0.0000	0.0000	0.0000	0.0000	120.0000

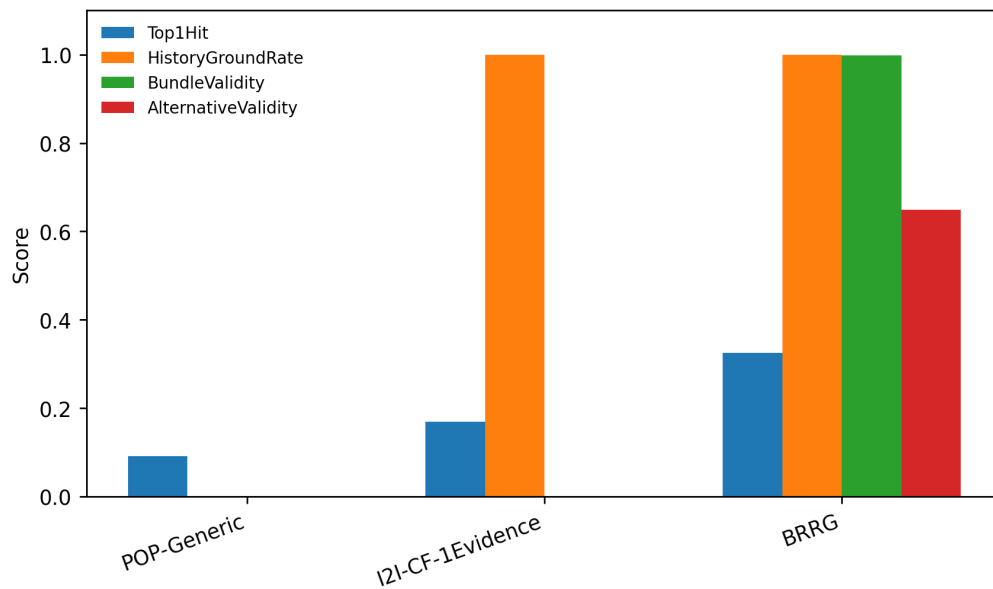


Figure 8. Grounding quality of recommendation explanations

Computational overhead is important because the proposed system is designed for interactive recommendation. Table 12 compares the inference path of deterministic BRRG with a fully generative LLM pipeline. BRRG does not decode the recommendation response token by token. Its online path consists of sparse score lookup, top-K merging, evidence-slot selection, and bounded verbalization. A fully generative LLM pipeline would instead serialize the user history, candidate list, and product evidence into a prompt and then generate the complete response autoregressively. That design can be more flexible in style, but its latency grows with prompt length and generated response length, and deployment usually depends on GPU capacity or an external model service. The deterministic design therefore offers more predictable latency, easier caching, and lower variance for real-time cart-facing use.

Table 12. Inference-time Comparison between Deterministic BRRG and a Fully Generative LLM Pipeline

Inference design	Online retrieval work	Response realization	Generative model call	Latency implication
Deterministic BRRG	Precomputed popularity, repeat/recency maps, sparse I2I-CF/UserKNN lookup, and top-K merge	Intent-conditioned retrieval plus bounded constrained verbalizer	None	Stable CPU-bound inference; response length is fixed by template slots
Fully generative LLM pipeline	Serialize user history, candidate list, and product evidence into a prompt	Autoregressive decoding of the complete recommendation response	One completion per turn	Latency grows with prompt length and generated tokens; deployment depends on GPU or external API capacity

Taken together, the results show a consistent pattern across all evaluation layers. The best candidate ranker was not the most complex single model but the hybrid that combined interpretable behavioral signals. The best response verbalizer on the conversational corpus was not the most constrained fusion generator but the intent-conditioned retrieval model. The best end-to-end recommendation explanation was the model that kept these two findings separate: a strong retrieval backbone selected the item, and a grounded response composer verbalized why the item fit the user's history and what neighboring options were available. This

separation of responsibilities is a substantive systems result because it improves accuracy, preserves grounding, and keeps inference predictable.

The first major implication is architectural. In retail recommendation, the strongest system was retrieval-first rather than generation-first. This finding agrees with prior NBR studies showing that repeat items and recent purchase histories explain a large part of next-basket performance [24], [25], while extending them by connecting the same retrieval evidence to conversational response generation. The ranking gains came primarily from combining repeated purchases by the same user, recency-weighted continuity in the user's basket stream, and collaborative transfer from similar items and similar users. This combination improved not only raw hit rates but also early ranking quality, as shown by the NDCG and MRR gains over RecencyRepeat. Commerce behavior is repetitive but not perfectly repetitive: a user often buys the same item again, but not always at the same time and not always in isolation. Collaborative signals help at the edges of this repetition pattern by surfacing related items, substitutes, and complements that pure memory models miss. Compared with LLM-centered conversational recommendation, the practical contribution is therefore a constrained architecture in which generation serves ranked behavioral evidence instead of replacing it.

The second implication is methodological. Response quality should not be treated as a vague by-product of recommendation quality. The Bitext benchmark showed that intent-conditioned retrieval was the most reliable backbone for response wording under deterministic evaluation. The Online Retail explanation benchmark showed that grounding constraints matter most after the candidate is selected. A single generator can make responses more stylistically diverse, but stylistic fluency is not the same as faithfulness to the ranked evidence. Assigning lexical response selection and product-grounded explanation to different modules made the final system easier to evaluate and more faithful in practice.

The third implication concerns explainability. Prior work has argued that explanations should be transparent, persuasive, and faithful [8], [9], [14]. The present results provide a concrete operationalization of that claim. A good retail explanation names the recommended item, points to evidence in the customer's history, and offers a neighboring action such as a bundle or lower-priced alternative. BRRG satisfied these conditions at scale because the explanation composer reused the evidence chain from retrieval rather than inventing a rationale after the fact.

Several limitations remain. First, the transaction study used the 2010-2011 partition of Online Retail II rather than a multi-domain live commerce platform. This choice supported reproducible preprocessing and dense invoice behavior, but it limited the study to one retailer and one historical period. Second, the conversational corpus was a customer-support dataset rather than a native recommendation dialogue corpus. This made the response-generation experiments realistic for commerce support language, but it did not provide user turns that directly mention product preferences. Third, the explanation evaluation was automatic rather than human rated. Automatic grounding metrics were appropriate for measuring faithfulness to behavior evidence, but they do not fully measure perceived helpfulness or persuasive effect. These limitations define a clear next step: evaluate the same grounded response protocol on a dataset or deployment setting that natively couples' recommendation dialogue with transaction history.

CONCLUSION

This paper presented BRRG, a behavior retrieval plus response generation framework for interpretable conversational personalized recommendation. The final pipeline was simple enough to audit, strong enough to outperform competitive baselines, and expressive enough

to produce grounded conversational outputs that included reasons, bundles, and alternatives. For ranking, BR-Hybrid was best overall. For verbalization, Intent-TFIDF was the best standalone response backbone. For grounded recommendation messaging, BRRG converted retrieval evidence into faithful customer-facing text with full historical grounding. The findings support a practical design rule for commerce recommendation: retrieve first, explain with evidence second, and let response generation serve the ranked evidence rather than replace it. That rule is consistent with classical recommender systems, modern conversational recommender research, and current LLM-oriented recommendation studies, while offering predictable inference behavior for real-time deployment.

REFERENCES

- [1] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in Proc. ACM Conf. Computer Supported Cooperative Work, 1994, pp. 175-186.
- [2] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," *Data Min. Knowl. Discov.*, vol. 5, no. 1-2, pp. 115-153, 2001.
- [3] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in Proc. 10th Int. Conf. World Wide Web, 2001, pp. 285-295.
- [4] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Model. User-Adapt. Interact.*, vol. 12, no. 4, pp. 331-370, 2002.
- [5] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422-446, 2002.
- [6] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, 2002, pp. 311-318.
- [7] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Proc. ACL Workshop Text Summarization Branches Out, 2004, pp. 74-81.
- [8] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," in Proc. ACM Conf. Computer Supported Cooperative Work, 2000, pp. 241-250.
- [9] N. Tintarev and J. Masthoff, "A survey of explanations in recommender systems," in Proc. IEEE 23rd Int. Conf. Data Eng. Workshop, 2007, pp. 801-810.
- [10] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in Proc. 25th Conf. Uncertainty in Artificial Intelligence, 2009, pp. 452-461.
- [11] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in Proc. Int. Conf. Learn. Representations, 2016.
- [12] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in Proc. 26th Int. Conf. World Wide Web, 2017, pp. 173-182.
- [13] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, and C. Pal, "Towards deep conversational recommendations," in Proc. 32nd Conf. Neural Information Processing Systems, 2018, pp. 9748-9758.
- [14] Y. Zhang and X. Chen, "Explainable recommendation: A survey and new perspectives," *Found. Trends Inf. Retr.*, vol. 14, no. 1, pp. 1-101, 2020.
- [15] W. Lei, X. He, Y. Miao, Q. Wu, R. Hong, M.-Y. Kan, and T.-S. Chua, "Estimation-Action-Reflection: Towards deep interaction between conversational and recommender systems," in Proc. 13th ACM Int. Conf. Web Search and Data Mining, 2020, pp. 304-312.

- [16] D. Jannach, A. Manzoor, W. Cai, and L. Chen, "A survey on conversational recommender systems," *ACM Comput. Surv.*, vol. 54, no. 5, Art. no. 105, pp. 1-36, 2021.
- [17] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, and E. Chen, "A survey on large language models for recommendation," *World Wide Web*, vol. 27, no. 5, Art. no. 60, 2024.
- [18] J. Lin, X. Dai, Y. Xi, W. Liu, B. Chen, H. Zhang, Y. Liu, C. Wu, X. Li, C. Zhu, H. Guo, Y. Yu, R. Tang, and W. Zhang, "How can recommender systems benefit from large language models: A survey," *ACM Trans. Inf. Syst.*, 2024.
- [19] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, and Q. Li, "Recommender systems in the era of large language models (LLMs)," *arXiv preprint arXiv:2307.02046*, 2023.
- [20] L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," in *Proc. Joint Int. Conf. Computational Linguistics, Language Resources and Evaluation*, 2024, pp. 10146-10159.
- [21] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-REC: Towards interactive and explainable LLMs-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.
- [22] D. Chen, "Online Retail II," *UCI Machine Learning Repository*, 2012, doi: 10.24432/C5CG6D.
- [23] Bitext, "Customer-support-llm-chatbot-training-dataset," *GitHub repository*, 2024.
- [24] M. Li, S. Jullien, M. Ariannezhad, and M. de Rijke, "A next basket recommendation reality check," *ACM Trans. Inf. Syst.*, vol. 41, no. 4, Art. no. 116, pp. 1-29, 2023, doi: 10.1145/3587153.
- [25] M. Ariannezhad, S. Jullien, M. Li, M. Fang, S. Schelter, and M. de Rijke, "ReCANet: A repeat consumption-aware neural network for next basket recommendation in grocery shopping," in *Proc. 45th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2022, pp. 1240-1250, doi: 10.1145/3477495.3531708.
- [26] Z. Shao, S. Wang, Q. Zhang, W. Lu, Z. Li, and X. Peng, "An empirical study of next-basket recommendations," *arXiv preprint arXiv:2312.02550*, 2023.
- [27] D. Jannach, "Evaluating conversational recommender systems," *Artif. Intell. Rev.*, vol. 56, pp. 2365-2400, 2023, doi: 10.1007/s10462-022-10229-x.

BIOGRAPHIES OF AUTHORS

Qi Xin received a master's degree from the University of Pittsburgh, USA. His research interests include applied artificial intelligence, machine learning, predictive modeling, intelligent classification, and NLP/LLM-driven decision-support systems for data-intensive applications.