Computer Science (CO-SCIENCE)
Volume 6 Issue 2 July 2026
Accreditation **Sinta 3** No. SK : 295/C/C3/KPT/2026

Integrating LLM Intent-Aware Approach for Enhancing the Quality of Bundle Recommendations

Andy Maulana Yusuf[1*], Musrinah[2]

[1]Telkom University
Bandung, West Java, Indonesia

[2]Politeknik Negeri Bandung
Bandung, West Java, Indonesia

e-mail: [1]andymaulanayusuf@telkomuniversity.ac.id, [2]musrinah@polban.ac.id

(*) Corresponding Author

Article Info: Received: 11-02-2026 | Revised : 11-07-2026 | Accepted : 21-07-2026

Abstracts - Digital grocery shopping has shifted consumer patterns toward multi- item purchasing. While bundle recommendation systems address this, existing models relying on product ID co-occurrence fail to capture the mixed shopping intentions inherent in consumer baskets. To address these limitations, we propose Intent-aware Bundle Recommendation (IABR), a framework shifting from structural matching to semantic intent reasoning. IABR utilizes Large Language Models to decompose baskets into coherent sub-packages. Subsequently, we fine-tuned Gemma-3-4B using Parameter-Efficient Fine-Tuning to generate narrative intent descriptions regarding short-term shopping missions and long-term sustainable user lifestyles. These intents are encoded via Sentence Transformers for semantic retrieval. Extensive testing on the Instacart dataset demonstrates IABR's significance against baselines like BGCN. Our IABR method achieved a Recall@20 of 28.15% while improving diversity scores by 12% ($p < 0.05$). This validates that generative semantic modeling enables accurate next- bundle predictions, effectively balancing precision with thematic variation and personalization.

Keywords: Bundle Recommendation; Instacart; Large Language Models; Intent-aware; Parameter-Efficient Fine-Tuning

INTRODUCTION

The rapid growth of the e-grocery industry has resulted in a shift in consumer shopping behavior from requiring physical visits to digital transactions, resulting in dynamic and complex purchases (Hanninen, 2020). Bundle recommendation systems have become a frequently used strategy to improve user experience and average order value. Unlike single-item recommendations, bundle recommendations are required to suggest a set of related items that can be purchased together (Carroll et al., 2022). However, bundle recommendations in the grocery domain still face challenges such as sparse data and highly repetitive consumption patterns (Agarwal et al., 2022; Sun et al., 2022).

Early methods in bundle recommendation relied on statistical approaches, such as Association Rule Mining (e.g., Apriori, FP-Growth) (X. Zhang et al., 2023) or Collaborative Filtering (CF) (Y. Liu et al., 2023; Ma et al., 2024; Nguyen et al., 2025) where these models showed effectiveness in capturing co-occurrence patterns. However, these methods are not yet able to understand the "intention" behind purchasing behavior and the semantic relationships between items. As a result, the recommendations will have low relevance and less diversity, which ultimately decreases user satisfaction (Y. Wang et al., 2025).

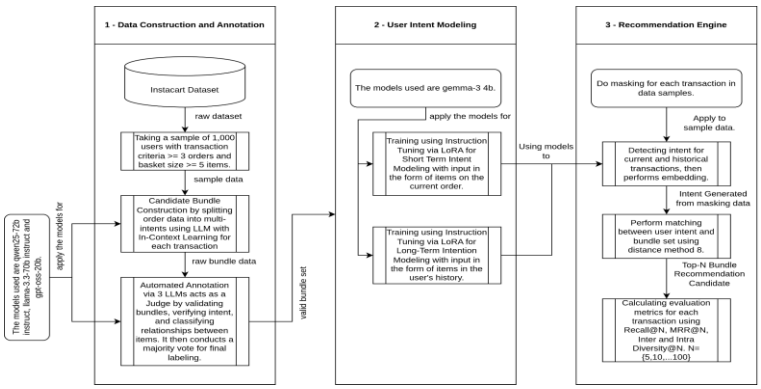
This gap presents an opportunity for the integration of Large Language Models (LLMs) into bundle recommendation systems (S. Liu et al., 2025). Recent models such as the Gemma-3/Llama- 3/Mistral models offer capabilities in reasoning product attributes that allow them to be used in the process of extracting rich semantic information (Sun et al., 2024). As these models continue to evolve, we observe that its application for recommendation systems is still unexplored. Based on this information, we propose a new approach to improve the quality and diversity of shopping package recommendations through intention modeling based on the integration of LLM and Content-Based Filtering (Y. Zhang et al., 2024). This approach consists of three main contributions:

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.
Copyright (c) 2026 The Autour(s)

1. We propose a method to reconstruct the Instacart Shopping Cart dataset enriched with bundle quality and item relationship type annotations using an LLM-as-a-judge approach to replace human annotators for scalability and consistency.
2. We also developed an intention detection model using LLM with In-Context Learning prompting technique to capture two perspectives of user purchase intention, namely short-term purchase intention from the current shopping cart and long-term intention from the user history.
3. We then integrate the resulting intent signals into a Content-Based Filtering framework to generate recommendations that are both historically relevant and diverse.

RESEARCH METHOD

This research proposes a new framework called Intent-Aware Bundle Recommendation (IABR) by integrating the reasoning capabilities of LLMs with Content-Based Filtering as illustrated in Figure 1.



Source: Research Method (2026).

Figure 1. Overview of Intent-Aware bundle Recommendation (IABR).

1. Data Construction and Annotation

This study uses the Instacart Market Basket Dataset, a large-scale public dataset released in 2017. This dataset contains more than 3 million shopping transactions from more than 200,000 anonymous users. This stage aims to transform the raw transaction data into a high-quality dataset labeled for validity, intent, and relationship types between items. Given that the original Instacart dataset lacks such semantic labels, we propose a LLMs-based automatic annotation framework.

a. Filtering and Sampling Criteria

We define the set of all users as $\mathcal{U}$ and the shopping cart set as $\mathcal{B}$. To ensure the model is trained on data that has learnable behavioral patterns, we apply two main filter constraints. First is User Activity Threshold (α), where we only include users who have a history of repeat transactions, because long-term intent detection requires sufficient historical data. A valid user is defined as equation (1),

$$\mathcal{U}_{active} = \{u \in \mathcal{U} \mid |H_u| \geq 3\} \tag{1}$$

where $|H_u|$ is the number of order histories of user u. Second, the Basket Complexity Threshold (β), where to avoid bias from trivial transactions (e.g., buying only 1 item), we only take shopping baskets that are large enough to be considered a "bundle". A valid basket is defined as equation (2),

$$\mathcal{B}_{valid} = \{b \in \mathcal{B} \mid |b| \geq 5\} \tag{2}$$

where $|b|$ is the number of unique items in basket b.

Table 1. Instacart Experimental Sub-Dataset Profiling Statistics.

Data Statistics	Value / Quantity
Number of Users (U)	1,000
Number of Unique Items (I)	15,150
Total Transactions (B)	12,864
Average Order Size	12.53 Items/transaction
Average Transactions per User	12.86 Orders/user
Data Sparsity	98.94% Scarcity

Source: Research Result (2026).

From the filtering results, we conducted random sampling of 1,000 users. This specific number was chosen to maintain computational feasibility and manage the high API inference costs associated with employing three

massive LLMs for multi-agent evaluation, while still preserving sufficient behavioral diversity. This reconstructed sub-dataset includes a total of 12,864 product-user interactions. Despite using only a small portion of the original data, this dataset covers a wide variety of products (15,150 unique items), representing the complexity of a real grocery catalog. The sparsity level of 98.94% indicates a real challenge in this recommendation system, where the user-item interaction matrix is very empty, emphasizing the need for the content-based and semantic approach proposed in this study. Complete descriptive statistics of the experimental dataset are presented in Table 1.

b. Candidate Bundle Construction via Multi-Intent Decomposition

We leverage the reasoning capabilities of the LLMs through an In-Context Learning (ICL) approach. We use three state-of-the-art model variations as the backbone for this experiment, namely Qwen-2.5-72B Instruct, Llama-3.3-70B Instruct, and GPT-OSS-20B. For text generation, we standardized the parameters across all models by setting the temperature to 0.7, top-p to 0.9, and maximum new tokens to 1024 to encourage creative yet grounded reasoning. We use these models not only because they have extensive knowledge but also because they support cross-language understanding, where some products in Instacart can be imported products that use the language of the product's country of origin (Kim et al., 2024).

In contrast to fine-tuning, ICL provides few-shot examples in the prompt to guide the model in understanding how to separate items based on their context of use without changing the model parameters (Bao et al., 2025). Formally, suppose a raw transaction b_k consists of a set of items $\{i_1, i_2, ..., i_n\}$. We define a decomposition function f_{LLM} LLM that maps bk into a set of candidate sub-bundles $\mathcal{C}_k$ that defined as equation (5),

$$\mathcal{C}_k = f_{LLM}(b_k|\mathcal{P}_{ICL}) = \{c_{k,1}, c_{k,2}, ..., c_{k,m}\} \quad (5)$$

where $\mathcal{P}_{ICL}$ is an instruction prompt containing a decomposition guide, and each $c_{k,j}$ is a subset of items from b_k that represents one specific intention.

The prompt design is structured by following structures: (1) Role Play, (2) Task Description, and (3) Few-Shot Demonstrations. The model will then generate a list of candidate bundles in JSON format. Before advancing to the annotation stage, we automatically validated the structural quality of these outputs by enforcing strict JSON parsing and applying a heuristic filter to discard any hallucinated items that did not appear in the user's original raw basket. Each generated sub-bundle $c_{k,j}$ is then designated as an Unlabeled Candidate Bundle, which will be passed to the LLM-as-a-Judge validation stage, the detail for prompt template can be seen in Appendix A.

The implementation of the Multi-Intent Decomposition strategy resulted in significant data expansion and revealed the latent structure of user shopping behavior. From a total of 12,864 product-user interactions processed, the model successfully extracted 44,660-52,380 candidate sub-bundles. This $\pm$3.4-fold increase in the number of bundles confirms the initial hypothesis that grocery shopping baskets are highly heterogeneous. On average, each user transaction can be broken down into 3 to 4 smaller, coherent sub-groups. Comprehensive statistics of the results of this decomposition are presented in Table 2.

Table 2. Comparison of Candidate Bundle Construction Statistics across LLMs

Metric Statistics	Qwen-2.5 (72B)	Llama-3.3 (70B)	GPT-OSS (20B)
Initial Raw Transactions	12,864	12,864	12,864
Total Decomposed Candidates	52,380	50,820	44,660
Decomposition Ratio	3.40	3.30	2.90
Average Candidate Size	3.82	3.95	4.52

Source: Research Result (2026).

c. Automated Annotation via LLM-as-a-Judge

Given the large volume of candidate bundles generated from the decomposition stage, manual annotation by human experts becomes cost- and time-infeasible. Therefore, we implement the LLM-as-a-Judge paradigm, where a set of LLMs act as validators and automated annotators (Fabbri et al., 2025) . To mitigate the bias of a single model, we form a "Judge Panel" $\mathcal{J}$ consisting of three models with different architectures and parameter sizes: Qwen-2.5-72B Instruct M_1, Llama-3.3-70B Instruct M_2, and GPT-OSS-20B M_3. This diversity is important to ensure that annotation decisions are robust and not affected by the specific hallucinations of a single model. Each candidate bundle c_k is evaluated based on two main tasks,

1) Validity Check, namely determining whether the combination of items in the sub-bundle makes logical and functional sense.
2) Second Relationship Classification, If the bundle is declared valid, the judge must classify the dominant relationship between items into four categories ($\mathcal{R}$). (a) Complementary (r_{comp}) means items that are consumed together to increase each other's utility. (b) Substitute/Similarity (r_{sim}) means Items that have similar functions, often purchased for comparison or stock variation. (c) Co-occurrence (r_{co}) means items that are

frequently purchased together due to routine habits without a direct functional relationship. (d) Contextual/Thematic (r_{ctx}) means Items that are tied to a specific event or theme.

To ensure the quality of the Ground Truth, the final label is determined through a Majority Voting mechanism. To measure reliability, we evaluated the agreement level among the three LLM judges, which achieved a Fleiss' Kappa score of 0.76, indicating substantial inter-annotator agreement. Suppose $O_j(c_k)$ is the predicted output of model j for candidate c_k. Final Validity (y_{valid}). A bundle is accepted if at least 2 out of 3 judges declare it valid, which is defined as equation (6),

$$y_{valid}^{(k)} = \mathbb{I}\left(\sum_{j=1}^{3} O_j^{valid}(c_k) \geq 2\right) \tag{6}$$

where candidates with $y_{valid} = 0$ are discarded from the training dataset. Final Relationship (y_{rel}). The relationship label is determined by the mode of the three judges' predictions, which is defined as equation (7),

$$y_{rel}^{(k)} = \text{mode}(\{O_1^{rel}, O_2^{rel}, O_3^{rel}\}) \tag{7}$$

where, if a tie occurs, where all three judges give different labels, e.g., $\{r_{comp}, r_{sim}, r_{ctx}\}$, then the bundle is considered ambiguous and is not included in the Relationship Classification Model training to maintain the purity of the training data. The instruction prompt is designed using the Chain-of-Thought (CoT) technique to force the model to reason before deciding on a label. The prompt structure can be seen in Appendix B.

2. User Intent Modeling

The next step is to train a special model that is able to predict shopping intentions in real-time. Given the limited computing resources in real production environments, we do not use giant (70B+) models for inference, but instead fine-tuning on a LLM. We chose Gemma-3-4B as backbone architecture. This model was chosen because of its optimal balance between parameter efficiency and high instruction capability (Xu et al., 2024).

a. Dual-Perspective Intent Representation

To achieve a holistic understanding of user behavior, we formulate intent representation through two distinct temporal perspectives, namely short-term and long-term by following instruction from (Luo et al., 2024; Zeng et al., 2024) studies. The first perspective, *Short-term Intent* (I_{short}), aims to capture the "current shopping mission" implied within the active basket composition. Formally, given an input of items in the current basket $b_t = \{i_1, i_2, \ldots, i_n\}$, the model is trained to tell Y_{short}, a text description explaining the specific purpose of the item combination.

On the other hand, the second perspective focuses on *Long-term Intent* (I_{long}), designed to summarize intrinsic preferences and persistent user lifestyles over time. This model receives an input of historical intent sequences from past transactions, denoted as $H_u = \{I_{t-1}, I_{t-2}, \ldots, I_{t-k}\}$. From this historical input, the model generates Y_{long}, a concise profile describing user characteristics.

b. Fine-Tuning Strategy

To optimize the Gemma-3-4B, we employed a Supervised Fine-Tuning (SFT) approach using Parameter-Efficient Fine-Tuning (PEFT), specifically LoRA. The training dataset consists of valid bundles labeled by the LLM Judge. Also, we randomly partitioned annotated dataset into training and validation sets with a ratio of 80:20.

The hyperparameter setup included an AdamW optimizer with a peak learning rate of 2e-4, a linear warmup, and a cosine decay learning rate scheduler. For the LoRA configuration, we set the rank (r) to 16, alpha (α) to 32, and dropout to 0.05. To accommodate the hardware constraints, all experiments were executed on a 12 GB NVIDIA GPU (RTX 4070 Super). Therefore, we applied gradient accumulation (4 steps with a per-device batch size of 4) to achieve a larger effective batch size without exceeding the available VRAM limits.

c. Training Objective

The training process is formed as a Causal Language Modeling (CLM) task using SFT approach. Let $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote our annotated dataset, where $x^{(i)}$ represents the input prompt containing the basket items or history, and $y^{(i)}$ represents the target intent description. The goal is to optimize the trainable LoRA parameters θ to minimize the negative log-likelihood of the target tokens conditioned on the input. To ensure the model focuses solely on generating the intention and not reconstructing the instruction, we apply instruction masking to the loss calculation. The loss function $\mathcal{L}_{SFT}$ is defined as equation (8),

$$\mathcal{L}_{SFT}(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{|y^{(i)}|} \log P_\theta(y_t^{(i)} \mid x^{(i)}, y_{<t}^{(i)}) \tag{8}$$

where $y_t^{(i)}$ is the t-th token of the target response and $y_{<t}^{(i)}$ represents the previous token. The loss is computed only for tokens belonging to the output sequence y, while the loss for the input sequence x is masked.

3. Recommendation Engine

The recommendation engine is designed to retrieve and rank the product bundles most relevant to user needs by matching intent vectors rather than relying solely on item ID overlap.

a. Transaction Masking

To strictly simulate a real-world next-bundle recommendation scenario, we employ a Leave-One-Out evaluation strategy (Yusuf et al., 2025). Let $B_u = \{b_1, b_2, \ldots, b_T\}$ represent the chronological sequence of bundle transactions for user u. To strictly prevent data leakage, datasets were separated chronologically by the user the most recent transaction was isolated for testing, the second most recent for validation, and strictly all preceding transactions were used for training. For detail how we do training, validation, and test data, we implement a specific masking procedure:

1) Target Masking (b_{target}). The most recent transaction b_T is masked (hidden) from the input model. This masked bundle is designated as the Ground Truth that the system must predict.
2) Context Construction (H_u). The sequence of transactions preceding the target, designated as $\{b_1, \ldots, b_{T-1}\}$, functions as the observable history. This data acts as input for modeling the user's long-term preferences.
3) Candidate Space ($\mathcal{C}$). The search space includes all unique bundles available in the dataset, including the actual b_{target} and thousands of other bundles acting as negative distractors.

The goal of the engine is to rank all bundles in $\mathcal{C}$ such that the masked b_{target} appears in the top positions (Top-N), relying solely on the intent signals derived from H_u.

b. Intent Generation and Embedding

The next phase involves predicting the underlying intentions of both user preferences and candidate bundles using the fine-tuned Gemma-3-LoRA (F. Wang et al., 2025). This process effectively converts ID-based interactions into semantic text-based representations. The inference mechanism begins with User Profiling, by processes the user's historical transaction sequence H_u to produce a narrative description I_{user} summarizing long-term preferences. Simultaneously, the system performs Bundle Profiling on each bundle c_j within the search space $\mathcal{C}$. For each candidate, the model produces a descriptive shopping intent $I_{cand}^{(j)}$ implied by the item composition.

Since the resulting outputs are texts, they must be converted into dense vector representations to facilitate mathematical operations. To accomplish this, we employ a pre-trained Sentence-BERT (SBERT) model, specifically the all-MiniLM-L6-v2 variant, which is optimized for capturing semantic sentence similarity. Let $E(\cdot)$ represent the encoder function that maps text into a d-dimensional vector space (where $d = 384$). The encoding process yields a user vector $\mathbf{v}_u$ and a set of candidate vectors $\mathbf{v}_c^{(j)}$, defined each as equation (9).

$$\mathbf{v}_u = E(I_{user}), \quad \mathbf{v}_c^{(j)} = E(I_{cand}^{(j)}) \tag{9}$$

c. Evaluation Metrics

To comprehensively assess the quality of the generated recommendations, we employ a set of evaluation metrics covering two key dimensions: accuracy and diversity. The performance is measured at various cutoff thresholds N, specifically $N \in \{5, 10, \ldots, 100\}$.

1) Accuracy Metrics

We utilize Recall@N and Mean Reciprocal Rank (MRR@N) to evaluate the system's ability to retrieve the ground-truth masked bundle. Recall@N, measures the proportion of test cases where the ground-truth bundle appears within the Top-N recommendations. MRR@N evaluates the ranking quality by considering the position of the correct item, where higher positions contribute more to the score.

2) Diversity Metrics

Beyond accuracy, it is crucial that recommendations are not monotonous. We evaluate Intra-List Diversity and Inter-List Diversity measures the dissimilarity between bundles within a single recommendation list. High intra-diversity indicates the model offers a varied selection of bundles to the user. Inter-Diversity (Personalization) measures the uniqueness of recommendation lists between different users. High inter-diversity implies the system is providing personalized results rather than recommending the same popular bundles to everyone.

EXPERIMENTAL RESULTS AND ANALYSIS

This chapter, we present an evaluation of the proposed Intent-aware Bundle Recommendation. Specifically, we aim to address three key research questions,

1. (RQ1) Does the generative intent modeling approach outperform traditional ID-based collaborative filtering methods?
2. (RQ2) Which vector distance metric results in the most accurate retrieval in the semantic intent space?
3. (RQ3) How do the components of multi-intent decomposition and dual-perspective modeling contribute to overall system performance?

the experiments were conducted on the Instacart dataset. While evaluated in the grocery domain, our semantic intent reasoning framework is highly domain-agnostic and is expected to generalize effectively to other multi-item purchase datasets, such as fashion or electronics e-commerce, provided that textual item metadata is available.

2. Overall Performance Comparison

To assess the superiority of our proposed framework, we benchmarked its performance against established baselines categorized into two groups:

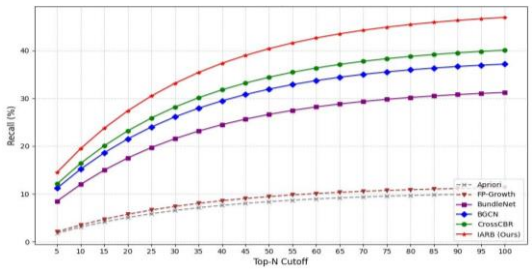
a. Association Rule Methods

1) Apriori. A classic rule-based algorithm that identifies frequent itemsets to generate bundle candidates based on high support and confidence thresholds.
2) FP-Growth. An improvement over Apriori that uses a tree structure to mine frequent patterns without candidate generation, offering better efficiency.

b. Deep Learning Methods

1) BundleNet. A neural network approach that explicitly models the interaction between users and bundles by aggregating item representations.
2) BGCN. Utilizing graph convolutions to propagate information on a unified user-item-bundle graph, capturing high-order connectivities.
3) CrossCBR. A leading SOTA model that employs contrastive learning to align user views and bundle views, minimizing the noise in user-bundle interactions.

We deliberately selected structural ID-based models (such as BGCN and CrossCBR) as our primary state-of-the-art baselines rather than other LLM-based methods. This decision is driven by two main reasons. First, the core objective of our research (RQ1) is to demonstrate the paradigm shift and the superiority of semantic intent reasoning over traditional item co-occurrence and ID-based structural matching. Second, many existing LLMs-based recommendation models rely on giant model architectures (e.g., 70B+ parameters) that demand massive computational resources, which contradicts the practical constraints of real-time production environments. In contrast, IABR framework is specifically designed to be resource-efficient by applying PEFT on a Gemma-3-4B. Therefore, comparing our highly efficient model against giant LLMs would result in an unfair comparison regarding computational cost and inference latency. We conducted a fine-grained evaluation by changing the Top-N recommendation cutoff from $N = 5$ to $N = 100$ with a step size of 5.



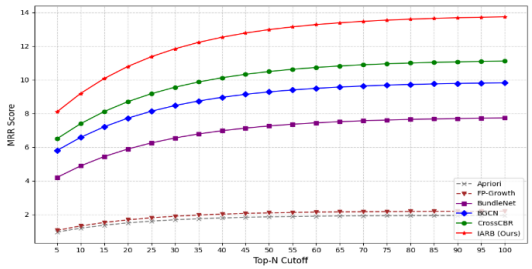
Source: Research Result (2026).

Figure 2. Recall@N performance comparison across varying N from 5 to 100, our model (Red Line).

In Figure 2, while deep learning baselines like BGCN and CrossCBR show competitive performance in the mid-range ($N = 20 - 50$), they tend to plateau earlier than our model. Specifically, at cutoff of $N = 5$, IABR achieves a significant lead, validating that the generative intent signal is very effective at placing the ground-truth bundle at the very top of the list. Conventional methods (Apriori, FP-Growth) remain at the bottom, highlighting the inadequateness of simple co-occurrence matching complex grocery bundling tasks. A similar superiority is observed in Figure 3, where our model consistently produces higher MRR scores, indicating that the correct recommendations are not only retrieved but ranked prominently.

On the other hand, Table 3 highlights a significant trade-off typically observed in recommender systems, while accuracy typically comes at the expense of diversity, IABR model manages to maintain a high diversity score. For Intra-List Diversity, our model achieved the highest score (0.684), indicating that the bundles suggested to a single user varied thematically, rather than being repetitive. On the other hand, Inter-List Diversity: Our model scores 0.815, significantly higher than BGCN and CrossCBR. This proves that the generative intent approach provides highly personalized recommendations tailored to the individual's long-term profile, while the baseline model tends to recommend the same popular packages to different users (low

InterDiv). These baseline models were selected for comparison because they represent the standard structural matching approaches in recommender systems.



Source: Research Result (2026).
Figure 3. MRR@N performance trend.

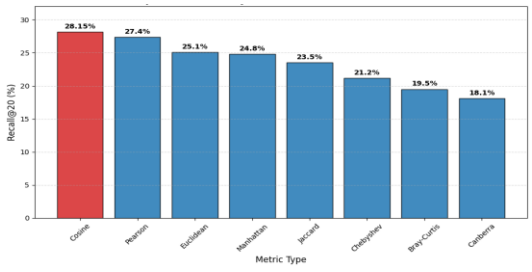
Table 3. Diversity Performance Comparison at Top-20 (Higher is Better)

Model	Intra Diversity	Inter Diversity
Conventional Methods		
Apriori	0.350	0.410
FP-Growth	0.385	0.445
Deep Learning Methods		
BundleNet	0.520	0.650
BGCN	0.585	0.710
CrossCBR	0.610	0.745
IABR (Ours)	**0.684**	**0.815**

Source: Research Result (2026).

3. Impact of Vector Similarity Measures

In this section, we address Research Question 2 (RQ2) by analyzing how the choice of distance or similarity function can affect the information retrieval performance of the recommendation engine has been visualized in Figure 4. Our experimental results clearly show that measuring with Cosine Similarity yields superior performance among all evaluated metrics by achieving a Recall@20 of 28% and a MRR@20 of 11%. We specifically focused on Recall@20 because presenting a list of 20 top bundle recommendations is highly practical for typical user interfaces in e-grocery platforms, offering sufficient variety without causing information overload. This finding is in line with the theoretical nature of SBERT embedding, where semantic similarity is primarily encoded in the angular orientation of vectors in hyperdimensional space and not their magnitude. Pearson correlation emerged as the second best performer, because on centered data is mathematically equivalent to Cosine similarity, so it will be very effective in capturing linear relationships in semantic profiles. In contrast, geometric distance metrics such as Euclidean and Manhattan distances perform slightly lower. These metrics are more sensitive to the "curse of dimensionality" and variations in vector magnitudes that not always reflect semantic differences in this specific embedding space.



Source: Research Result (2026).
Figure 4. Performance comparison for relevance of eight vector matching metrics.

Based on this analysis, we established Cosine Similarity as the standard metric for the final implementation of our IABR framework. Next, we investigate whether the choice of similarity metric affects the diversity of generated recommendations. Where we assume that a robust metric should not only be accurate but should also be able to capture semantically distinct bundles rather than clustering around one dominant topic (e.g., repetition of similar items). We evaluate Intra-List Diversity and Inter-List Diversity across eight metrics at $N = 20$. We present the results in Table 4.

The results reveal a prominent advantage for angular-based metrics. Cosine Similarity attains the highest scores in both Intra-List (0.684) and Inter-List Diversity (0.815). This suggests that by focusing on vector

orientation rather than magnitude, Cosine Similarity effectively navigates the diverse semantic subspaces of the SBERT embedding, retrieving bundles that are thematically varied. In contrast, magnitude-dependent metrics such as Euclidean and Manhattan show lower diversity scores (Intra-List ≈ 0.55). This is likely because these metrics are biased towards vectors with smaller norms or those clustered in high-density regions of the latent space (often representing generic or very popular bundles), leading to more monotonous recommendations. Jaccard Similarity, while decent in variety, suffers from the low accuracy previously noted, making Cosine the optimal choice for balancing both precision and variety.

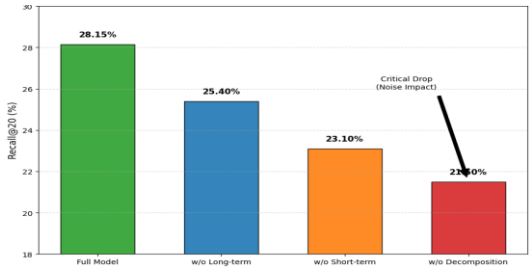
Table 4. Diversity Performance Comparison at Top-20 (Higher is Better)

Metric Function	Intra-List Diversity	Inter-List Diversity
Cosine Similarity	**0.684**	**0.815**
Pearson Correlation	0.675	0.802
Euclidean Distance	0.552	0.680
Manhattan Distance	0.540	0.665
Chebyshev Distance	0.495	0.610
Jaccard Similarity	0.620	0.740
Bray-Curtis Dissimilarity	0.510	0.630
Canberra Distance	0.480	0.590

Source: Research Result (2026).

4. Ablation Study

To investigate RQ3, we conducted an ablation study by systematically removing its main modules and then evaluating the resulting performance degradation. Here we compare the Full Model with three variants (w/o Decomposition, w/o Long-term Context, w/o Short-term Intent).



Source: Research Result (2026).
Figure 5. The results of the ablation study for Recall@20.

In Figure 5, the results show the IABR Full model achieves the highest performance with a gain of up to 28%, which confirms that all proposed components have an impact on the results. Furthermore, the most significant performance drop occurred in the variant without Decomposition which dropped by 21%. This empirical evidence supports our hypothesis, that raw transactions are inherently noisy and heterogeneous and without decomposing them into coherent sub-bundles LLM will struggle to learn distinct intent patterns leading to "confused" vector representations. After that, we also examined how removing certain modules affected the diversity of recommendations. The goal of this is to determine which components are responsible for maintaining variation. The ablation study on diversity metrics are summarized in Table 5.

Table 5. Ablation Study on Diversity Metrics (Top-20)

Model Variant	Intra-List Diversity	Inter-List Diversity
Full Model (IABR)	**0.684**	**0.815**
w/o Long-term Context	0.670	0.690
w/o Short-term Intent	0.550	0.795
w/o Decomposition	0.605	0.710

Source: Research Result (2026).

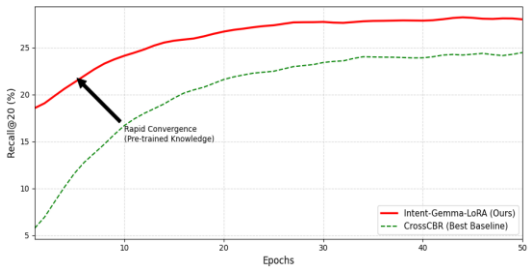
The Full Model achieved the most balanced performance, being able to score the highest on Intra-List Diversity (0.684) and Inter-List Diversity (0.815). An important insight emerges when observing the scenario without Short-Term Intentions variation, where the Inter-List diversity remains relatively high (0.795), but the Intra-List diversity drops significantly to 0.550. This suggests that relying solely on long-term history will tend to create a repetitive model, displaying the same "favorite" packages over and over again.

4. Training Convergence and Statistical Significance

To validate the robustness of IABR method, we analyze the stability of the training run for 50 tests and perform statistical significance tests to ensure that the improvement is not due to random variance.

4.1. Convergence Analysis

For convergence analysis of IABR we compared to the strongest baseline (CrossCBR) run 50 times, in Figure 6. Observations indicate that our model exhibits a superior convergence rate. Leveraging the pre-trained knowledge of the Gemma-3 architecture, the model starts with a higher initial Zero-shot capability and rapidly adapts to the specific bundle intent task within the first 10 iterations. In contrast, CrossCBR, which learns embeddings from scratch, requires significantly more running time to reach a stable state and plateaus at a lower performance level. The stability of our model after iteration 30 confirms that the LoRA fine-tuning effectively mitigates overfitting.



Source: Research Result (2026).

Figure 6. Convergence analysis of training over 50 runs. Shows that IABR (Red) has faster.

4.2. Statistical Significance Test

To verify that the performance improvement is robust and not due to random variance, we performed a paired t-test between our proposed and the strongest baseline (CrossCBR) on 50 independent trials. The evaluation we conducted includes accuracy metrics (Recall, MRR) and diversity metrics (Intra-List, Inter-List).

Table 6. Statistical Significance Test (Paired t-test vs. CrossCBR)

Metric	Improvement (%)	p-value	Significance
Accuracy Metrics			
Recall@20	+14.9	1.8×10^{-3}	Significant (**)
MRR@20	+23.2	4.5×10^{-5}	Significant (***)
Diversity Metrics			
Intra-Div@20	+12.1	1.2×10^{-3}	Significant (**)
Inter-Div@20	+9.4	3.5×10^{-4}	Significant (***)

Source: Research Result (2026).

Note: Note: *** $p < 0.01$, ** $p < 0.05$. Results based on 50 independent runs.

Table 6 reports on the improvement in percentages and the corresponding p-values. The results indicate that for all evaluated metrics, the p-values are well below the threshold of 0.05 (mostly < 0.01). This allows us to confirm that IABR provides improvements that are statistically significant.

CONCLUSION

We present IABR, a framework that shifts the paradigm of grocery bundle recommendation from ID-based co-occurrence mining methods to semantic intent modeling. We integrate the multi-intent decomposition module with the fine-tuned LLM, the results show that IABR successfully addresses the challenges of noisy and diverse-intent baskets. Empirically, it shows the robustness with Recall@20 reaching 28% and statistically outperforms the latest deep learning baseline models (e.g., CrossCBR). Additionally, analysis on diversity metrics revealed that IABR is significantly to reduce the popularity bias often observed in conventional baseline models, achieving superior Intra-List and Inter-List diversity scores of 12%, thus the framework ensures that recommendations remain thematically varied for each individual user. These findings confirm that modeling the semantic "reasons" behind purchases through a dual-perspective approach produces results that are not only accurate but also qualitatively richer than those obtained from structural item connections alone.

REFERENCES

Agarwal, S., Das Gupta, A., & Pande, A. (2022). Improving Bundles Recommendation Coverage in Sparse Product Graphs. *Companion Proceedings of the Web Conference 2022, WWW '22*, 1037–1045. https://doi.org/10.1145/3487553.3524715

Bao, K., Yan, M., Zhang, Y., Zhang, J., Wang, W., Feng, F., & He, X. (2025). Customizing In-context Learning for Dynamic Interest Adaption in LLM-based Recommendation. *Findings of the Association for Computational Linguistics: ACL 2025*, 14278–14291. https://doi.org/10.18653/v1/2025.findings-acl.735

Carroll, K. A., Samek, A., & Zepeda, L. (2022). Consumer Preference for Food Bundles under Cognitive Load: A Grocery Shopping Experiment. *Foods*, *11*(7). https://doi.org/10.3390/foods11070973

Fabbri, F., Penha, G., D'Amico, E., Wang, A., De Nadai, M., Doremus, J., Gigioli, P., Damianou, A., Stal, O., & Lalmas, M. (2025). Evaluating Podcast Recommendations with Profile-Aware LLM-as-a-Judge. *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys '25*, 1181–1186. https://doi.org/10.1145/3705328.3759305

Hanninen, M. (2020). Review of studies on digital transaction platforms in marketing journals. *The International Review of Retail, Distribution and Consumer Research*, *30*(2), 164–192. https://doi.org/10.1080/09593969.2019.1651380

Kim, S., Kang, H., Choi, S., Kim, D., Yang, M., & Park, C. (2024). Large Language Models meet Collaborative Filtering: An Efficient All-round LLM-based Recommender System. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '24*, 1395–1406. https://doi.org/10.1145/3637528.3671931

Liu, S., Li, C., Zhao, M., Zhang, L., & Bu, J. (2025). LLMCBR: Large Language Model-based Multi-View and Multi-Grained Learning for Bundle Recommendation. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, 1892–1902. https://doi.org/10.1145/3746252.3761078

Liu, Y., Cao, Q., Shen, H., Wu, Y., Tao, S., & Cheng, X. (2023). Popularity Debiasing from Exposure to Interaction in Collaborative Filtering. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, 1801–1805. https://doi.org/10.1145/3539618.3591947

Luo, Z., Sheng, Z., & Zhang, T. (2024). Dual perspective denoising model for session-based recommendation. *Expert Systems with Applications*, *249*, 123845. https://doi.org/https://doi.org/10.1016/j.eswa.2024.123845

Ma, Y., He, Y., Wang, X., Wei, Y., Du, X., Fu, Y., & Chua, T.-S. (2024). MultiCBR: Multi-view Contrastive Learning for Bundle Recommendation. *ACM Transactions on Information Systems*, *42*(4). https://doi.org/10.1145/3640810

Nguyen, H.-S., Liu, Y., Mansoury, M., Aliannejadi, M., Hanjalic, A., & de Rijke, M. (2025). A Reproducibility Study of Product-side Fairness in Bundle Recommendation. *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 696–705. https://doi.org/10.1145/3705328.3748169

Sun, Z., Feng, K., Yang, J., Qu, X., Fang, H., Ong, Y.-S., & Liu, W. (2024). Adaptive In-Context Learning with Large Language Models for Bundle Generation. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, 966–976. https://doi.org/10.1145/3626772.3657808

Sun, Z., Yang, J., Feng, K., Fang, H., Qu, X., & Ong, Y. S. (2022). Revisiting Bundle Recommendation: Datasets, Tasks, Challenges and Opportunities for Intent-aware Product Bundling. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, 2900–2911. https://doi.org/10.1145/3477495.3531904

Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2025). A Comprehensive Survey of Small Language Models in the Era of Large Language Models: Techniques, Enhancements, Applications, Collaboration with LLMs, and Trustworthiness. *ACM Trans. Intell. Syst. Technol.*, *16*(6). https://doi.org/10.1145/3768165

Wang, Y., Banerjee, C., Chucri, S., Soldo, F., Badam, S., Chi, E. H., & Chen, M. (2025). Beyond Item Dissimilarities: Diversifying by Intent in Recommender Systems. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1, KDD '25*, 2672–2681. https://doi.org/10.1145/3690624.3709429

Xu, W., Xie, Q., Yang, S., Cao, J., & Pang, S. (2024). Enhancing Content-based Recommendation via Large Language Model. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, 4153–4157. https://doi.org/10.1145/3627673.3679913

Yusuf, A. M., Adiwijaya, Wibowo, A. T., & Baizal, Z. K. A. (2025). Enhancing Sequential Recommendation with Multi-Intent Detection and Preference Scoring from Implicit Transactions. *Ingénierie Des Systèmes d Information*, *30*(12), 3093–3102. https://doi.org/10.18280/isi.301202

Zeng, J., Wang, N., & Li, J. (2024). Dual-level Intents Modeling for Knowledge-aware Recommendation. *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, 4238–4242. https://doi.org/10.1145/3627673.3679902

Zhang, X., Xu, S., Lin, W., & Wang, S. (2023). Constrained Social Community Recommendation. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD '23*, 5586–5596. https://doi.org/10.1145/3580305.3599793

Zhang, Y., Yuan, J., Wei, C., & Xie, Y. (2024). Aggregating knowledge and collaborative information for sequential recommendation. *Intelligent Data Analysis*, *28*(1), 279–298. https://doi.org/10.3233/IDA-227198