Beyond Co-purchase Relation: Evolution of Complementary Recommendations at Allegro

Aleksandra Osowska-Kurczab*
aleksandra.kurczab@allegro.com
Allegro.com
Poland

Klaudia Nazarko*
klaudia.nazarko@allegro.com
Allegro.com
Poland

Eliška Kosturová*
eliska.kosturova@allegro.com
Allegro.com
Czech Republic

Lidia Wojciechowska
Allegro.com
Poland

Michał Bień†
NVIDIA
Poland

Abstract

When a customer adds a professional camera to their cart, should the system suggest a matching lens, a generic tripod, or another camera body? Complementary Product Recommendation is vital for comprehensive basket building, yet standard models often fail to distinguish between items that are merely bought together and those that truly work together. In this paper, we present AlleCompanion: a production-scale retrieval framework deployed at Allegro.com that transforms noisy behavioural signals into precise semantic compatibility. We mitigate the intrinsic noise in large-scale co-purchase traffic by combining data-level filtering heuristics with a category-constrained Two Tower architecture. Within this framework, the Category Adapter guides the model in the embedding space, constraining candidates within logically complementary boundaries. Since modelling authentic user behaviour at scale is inherently difficult, we introduce ComCat, a multi-source Complementary Categories Mapping. ComCat acts as a translational layer that distils meaningful patterns from noisy traffic into a maintainable and controllable solution, integrating expert rules, human-in-the-loop feedback, LLM-based reasoning, and statistical mining. Our experimental results demonstrate that combining explicit category-level constraints with neural architectures effectively filters out co-purchase noise to surface recommendations that satisfy real-world user needs. Serving over 20 million active users monthly, the framework delivers significant uplifts in attributed GMV for organic discovery and drives substantial revenue growth in sponsored placements.

CCS Concepts

• **Information systems → Recommender systems**.

*All authors contributed equally to this research.
†Work done at Allegro.com.

RecSys '26, Minneapolis, Minnesota, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keywords

Recommendation systems, Complementary recommendations, E-commerce, Content-based filtering

ACM Reference Format:
Aleksandra Osowska-Kurczab, Klaudia Nazarko, Eliška Kosturová, Lidia Wojciechowska, and Michał Bień. 2026. Beyond Co-purchase Relation: Evolution of Complementary Recommendations at Allegro. In *Proceedings of the Twentieth ACM Conference on Recommender Systems (RecSys '26), September 28–October 2, 2026, Minneapolis, Minnesota, USA*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

The objective of modern e-commerce recommendation systems has evolved from isolated item discovery towards modelling user's comprehensive purchase intent. While similarity-based models excel at selecting specific items (such as suggesting another smartphone), business value is often driven by complementarity. For example, recommending a protective case or a high-speed charger to a phone is what ultimately maximises shopping basket value and customer satisfaction. However, modelling complementarity presents a distinct challenge compared to similarity; it requires a fundamental shift from matching items with overlapping features (*substitutes*) to identifying items that work well together (*complements*).

Every month, over 20 million active buyers visit Allegro to explore a vast catalogue of products from more than 150 thousand sellers. Within this massive ecosystem, complementary recommendations stand out as one of the most profitable discovery mechanisms [14], driving engagement for both organic and sponsored content. We formulate these recommendations as a product-to-product relation where a target item functionally extends a query product. In practice, users frequently acquire these items together, a co-purchasing behaviour that is strongly encouraged by platform incentives to source multiple goods from the same seller. However, discovering genuine complementary pairs within historical transactions requires navigating complex, cross-category relationships bound by strict item-level compatibility — a fundamental shift from merely matching items with overlapping features.

Translating this goal into an accurate complementary retrieval system poses three distinct challenges. Firstly, recommendations must bridge the gap between broad category-level associations and precise item-level compatibility. This requires models to understand which product groups are complementary while strictly enforcing the compatibility on item-specific attributes like brand,

model or size. Secondly, the modelled relations must generalize to cold-start and long-tail items while respecting the inherent asymmetry of complementary pairs — for instance, while a phone charger complements a smartphone, a smartphone does not complement a charger. Finally, learning from historical co-purchase data is severely hindered by noise [11]. While transaction logs capture historical co-purchase behaviour, they frequently combine distinct intents. For example, a user purchasing multiple flavours of dog food, or both dog and cat food, represents a desire for substitute items or entirely separate needs rather than true functional complements. Overcoming this requires logical guidance beyond mere transaction frequency.

To address these challenges, we introduce AlleCompanion, an end-to-end framework for complementary product recommendation. Our work encompasses the complete lifecycle of the recommendation system, from architectural design, dataset construction and offline ablations to successful production deployment. The main contributions of this work are summarized as follows:

- **AlleCompanion**, Two Tower architecture featuring category-conditioned retrieval. It steers the latent space towards complementary categories while maintaining precise item-level alignment, bridging the gap between broad category-level associations and compatibility.
- **ComCat**, a multi-source mapping integrating expert rules, LLM insights, and statistical mining to guide the model's responses towards logically complementary relations.
- An empirical study on the **impact of various dataset definitions**, providing critical insights into the practical challenges of filtering out non-complementary purchase intents and isolating true item-level compatibility.
- **Lessons from offline ablations and online A/B testing**, demonstrating the model's practical value and performance gains for both organic and sponsored recommendations.

2 Related Work

E-commerce marketplaces drive basket growth through financial incentives (e.g. free delivery thresholds [17]), strategic UX placement and gamification mechanisms [1]. These elements are injected across the user journey — from "complete the look" suggestions to value-added services at checkout. To scale these strategies, platforms rely on complementary recommendation algorithms [8] that identify relevant items across diverse contexts.

However, mining these relationships is complex, as raw co-purchase logs often fail to distinguish between joint demand and mere alternatives [5, 16]. Recent research introduces distant supervision methods to filter this noise, such as excluding co-purchase pairs with high co-view overlap [5] or applying a substitute penalty [25]. While these heuristics improve signal quality, they often struggle with the nuances of complementarity or become too complex to maintain in production. We address this trade-off by refining behavioural filters and evaluating diverse data sources beyond simple co-purchase traffic.

In addition to behavioural logs, complementarity requires understanding compatibility between products through Knowledge Graphs [23] or Large Language Models (LLMs) [7]. LLMs, acting as repositories of world knowledge, can generate explanations [9], identify complementary concepts [6] or act as few-shot annotators [7, 20]. For platforms with sparse interaction logs, transferring such knowledge into a universal embedding space is crucial for identifying relations across the long tail items [15]. We extend these methodologies through ComCat, a multi-source category mapping that distils LLM-based reasoning and expert logic into a controllable translational layer for consistent cross-category retrieval.

The architectural landscape for complementary tasks has evolved from content-based similarity [11, 19] towards sequential modelling. GNNs have become the standard for capturing non-transitive dependencies [3, 10], while transformer-based architectures [12] address the temporal dynamics of basket building [24]. Current research focuses on hybrid frameworks that fuse GNN-derived structural knowledge with multi-modal content [18]. A prominent example is P-Companion [5], which utilises specialised embedding spaces to balance relevance with diversity [21]. Yet, while such hybrid systems achieve high precision, they often introduce significant computational overhead — particularly when involving real-time GNN inference — complicating real-time retrieval at scale. AlleCompanion builds upon the hybrid paradigm established by P-Companion, but prioritises architecture and maintenance simplicity.

3 Methods

3.1 Architecture Definition

AlleCompanion is a content-based model designed specifically for complementary product recommendation. Our proposed architecture features a Category Adapter, which projects the item embedding into a complementary category latent space. Furthermore, we utilise a category reconstruction loss [2] that enhances the model's ability to learn robust representations of these categories. The high-level architecture of AlleCompanion is illustrated in Figure 1.

3.1.1 Two Tower. The proposed approach utilises a Two Tower deep learning model, which maps query and target product features into a shared embedding space. The training objective is to maximise the inner product between the query and target vector representations. We frame this retrieval task as a classification problem, leveraging a sampled softmax loss function ($\mathcal{L}_{retrieval}$) [4] integrated with a mixed negative sampling strategy [22], and temperature scaling [13]. To improve convergence and ensure numerical stability, we implement parameter sharing between the query and target towers. This unified module is designated as the Product Encoder. Under the hood, the architecture consists of dedicated embedding tables that transform each individual item feature into a low-dimensional vector. Once concatenated, these vectors are processed through a multi-layer perceptron (FC) and L2-normalised to produce the final representation.

The model is trained in an item-to-item regime, learning the relationship between products co-purchased by a user within a short time window. To account for the volatility of a million-scale product catalogue, we represent each item via its content features, rather than ID features. In Vanilla Two Tower, by default, a product is represented using its title, price, and category features. This content-based approach allows for the seamless integration of additional textual and categorical features, such as product attributes and seller IDs, respectively. To capture broader structural relationships,

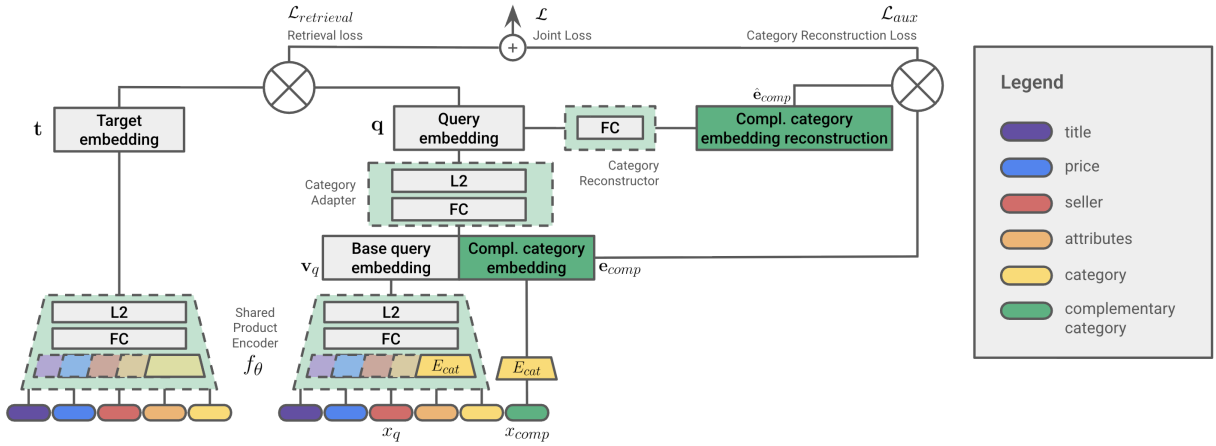

Figure 1: AlleCompanion architecture. Two Tower model extended with Category Adapter and Category Reconstruction Loss.

the category feature is derived from the hierarchical taxonomy of categories maintained in Allegro's product catalogue (e.g., *Allegro > Electronics > Smartphones)*.

3.1.2 Category Adapter. While a standard Two Tower architecture primarily focuses on identifying similar items in a shared embedding space, our goal is to explicitly direct the model towards retrieving complementary products. To achieve this, we introduce a Category Adapter module that constrains the retrieval process to a requested complementary category. Firstly, the query product features x_q are passed through the Shared Product Encoder f_θ to obtain the base query embedding $\mathbf{v}_q = f_\theta(x_q)$. In parallel, the target complementary category x_{comp} is projected into a dense vector $\mathbf{e}_{comp} = \mathbf{E}_{cat}(x_{comp})$, utilising the shared category embedding table $\mathbf{E}_{cat}$ from the Product Encoder. Lastly, these two vectors are concatenated and passed through a FC layer (weights $\mathbf{W}$ and bias $\mathbf{b}$) with normalisation to produce the final query representation:

$$\mathbf{q} = \text{Norm}(\mathbf{W}[\mathbf{v}_q \parallel \mathbf{e}_{comp}] + \mathbf{b})$$

Category Adapter is integrated into the query tower, where the product representation $\mathbf{v}_q$ is augmented with the target complementary category $\mathbf{e}_{comp}$. During training, this category is dynamically derived from the ground-truth target item to ensure that the model learns to associate specific products with their appropriate complementary classes.

3.1.3 Category Reconstruction Loss. To guide the Category Adapter towards more discriminative signals, we introduce an auxiliary reconstruction loss $\mathcal{L}_{aux}$. Specifically, we project the final query embedding $\mathbf{q}$ through a single FC layer to reconstruct the complementary category representation:

$$\hat{\mathbf{e}}_{comp} = \mathbf{W}_{aux}\mathbf{q} + \mathbf{b}_{aux}$$

We optimise this alignment using a sampled softmax loss, effectively minimising the distance between the projected query $\hat{\mathbf{e}}_{comp}$ and the true complementary category embedding $\mathbf{e}_{comp}$.

The final optimization objective $\mathcal{L}$ is formulated as a joint loss, combining the primary retrieval loss $\mathcal{L}_{retrieval}$ and the auxiliary reconstruction loss:

$$\mathcal{L} = \mathcal{L}_{retrieval} + \mathcal{L}_{aux}$$

3.2 Dataset Construction

Co-purchase signals serve as the primary basis for capturing latent complementary relations between products during the training of AlleCompanion. While they serve as a strong proxy for complementarity, raw transactional data is inherently noisy and often fails to distinguish true complementary items from substitutable ones. To address that, we refine the data through filtering and heuristics.

3.2.1 Raw dataset generation. We define a co-purchase session as a sequence of items $(i_1, i_2, \ldots, i_n)$ purchased by a single user within a specific time window. From these sessions, we derive ordered asymmetric pairs (i_j, i_k) where $j < k$, ensuring the sequence models the directionality of complementary needs.

The quality of these generated pairs depends heavily on the temporal window used to define co-occurrence. Empirical observations of Allegro traffic suggest that while same-cart purchases provide high frequency, they are often dominated by highly similar or identical items. Shifting focus towards longer session windows significantly increases item diversity and volume while preserving semantic relevance; by contrast, much longer windows introduce excessive noise from unrelated purchases. To address this trade-off, the exact session window duration was selected via hyperparameter tuning evaluated against a holdout validation set, balancing semantic connection and item diversity.

3.2.2 Behavioural filtering and heuristics. To ensure that the model learns from representative user behaviour rather than unrelated interactions or outliers, a multi-step filtering process is applied. First, "heavy buyers" whose transaction volume exceeds the 99th percentile are excluded to prevent high-frequency outliers from biasing the learned distribution. This is followed by heuristic rules, designed to distinguish true complementary pairs from substitutes or unrelated associations, focusing on two primary signals: pair count (the frequency of an item pair across sessions) and category

alignment (categorical overlap between the items' respective departments and categories).

Following established practices [5], we conducted an internal annotation task involving 400 item pairs sampled from the co-purchase dataset. Annotators assigned relationship labels—*complementary*, *substitutable*, or *unrelated*—requiring a consensus of two per pair. Analysis showed that enforcing a minimum pair count effectively filters niche, coincidental co-purchases, reducing the presence of unrelated items from 44% to 22%. However, excessively high thresholds predominantly isolate substitutes (increasing from 28% to 43%) rather than complements (which only rise from 29% to 36%).

To rebalance the dataset toward a target hierarchy of complementary > substitutes > unrelated, a category alignment heuristic was implemented. This heuristic requires product pairs to share the same department but belong to different categories (e.g., *Tripod* and *Camera Lens* within *Electronics*). Applying this constraint filtered the baseline pool —originally consisting of 36% complementary, 16% substitutes, and 48% unrelated pairs — into a rebalanced dataset comprising 61% complementary, 4% substitutes, and 35% unrelated pairs. The resulting structural bias toward complementary relationships aligns with established patterns showing that such pairings frequently span distinct product types [5].

All filters were applied to the testset to ensure consistency, with the exception of the minimum pair count. The exclusion of this specific threshold allows for the evaluation of model performance across a broader distribution of real-world traffic.

3.3 Online Deployment

Our online deployment architecture follows a standard embedding-based retrieval paradigm. The AlleCompanion model is periodically retrained on a single NVIDIA T4 (16GB) GPU, and the Approximate Nearest Neighbour index is refreshed daily using the Faiss library for efficient serving. This setup enables real-time recommendation retrieval with strict millisecond-level latency guarantees. In production, AlleCompanion serves as one of the retrievers for the recommendation carousel. It is deployed alongside collaborative filtering methods and is supplemented by heuristic fallbacks, such as category bestsellers, to ensure high coverage.

During online inference, explicit target items are naturally unavailable. To address this, we utilise a predefined complementary mapping to determine the valid set of target categories for a category of a given query. The model processes the query features and integrates the target category as a conditioning signal, dynamically projecting the query into a specific latent subspace [5]. Finally, the candidate groups retrieved from each respective target category are interleaved to diversify the output and enhance the visual presentation of the recommendation carousel. Decoupling the model from ComCat enables updates to complementarity rules without model retraining.

3.4 Complementary Categories Mapping

Determining the optimal target category for a given query item is a non-trivial challenge that requires balancing relevance with diversity to satisfy varied user needs. To address this, we use ComCat as a robust proxy for ground-truth complementarity. This approach ensures the model retrieves logical results even for cold-start products or items with low traffic coverage, where historical behavioural signals are often missing or unreliable.

3.4.1 Sources of complementary categories. To achieve both high precision and broad catalogue coverage, we utilise a heuristic ensemble that merges three distinct data sources, weighing expert domain knowledge against raw behavioural noise. Crucially, the mapping is modelled as a directed relationship, capturing the asymmetry of complementary categories (e.g., a primary purchase driving the need for an accessory).

Automated Co-Purchase Heuristics: This source identifies relationships by mining categorical co-occurrence patterns within purchase sessions. For each category, we generate directional permutations of items purchased within the same session—excluding high-volume outliers—to capture asymmetric category pairs (e.g., *Smartphone* → *Case*). The strength of these pairs is quantified using Jaccard similarity to ensure the signal is driven by specific co-purchase intent rather than global popularity. To prune accidental associations, we apply structural filters based on taxonomic tree distance, periodicity signals, and price ratio constraints.

Human-in-the-Loop Annotations: To address coverage gaps for cold-start and low-traffic categories, we integrated a human-in-the-loop component focused on less obvious pairings. Product pairs were sourced either from filtered co-purchase traffic or through an LLM-assisted workflow [6] that identifies complementary concepts. Expert annotators labelled these pairs as *complementary*, *substitutable*, or *unrelated*. Complementary signals were then aggregated to the category level to provide a reliable ground truth.

Rule-Based Expert Logic: We employ an expert-driven source to capture compatibility aspects that behavioural data might overlook. Business experts define high-precision rules based on fine-grained product parameters to ensure strict technical alignment between recommended items. Once aggregated to the category level, these rules provide a compatibility-enabled mapping entirely independent of purchase traffic.

3.4.2 Integration of the sources. The final mapping is constructed by merging these sources into a prioritised hierarchy based on expert assessment: Annotations, followed by Rule-Based logic, and finally Automated Heuristics as a broad-reach fallback. To adapt to specific business needs, this mapping can be extended with a same-category rule that maps a category to itself, such as headphones to headphones, thereby supporting the mining of alternative products. This configuration allows a single AlleCompanion instance to mix both complements and substitutes, without the overhead of maintaining multiple specialised models.

4 Results

We evaluate our framework through offline experiments and online A/B tests. Our offline analysis includes an ablation study of the AlleCompanion architecture, a summary of insights gained from complementary dataset construction, and an assessment of the ComCat mechanism. Finally, we report the performance gains observed during online A/B testing, focusing on key business metrics.

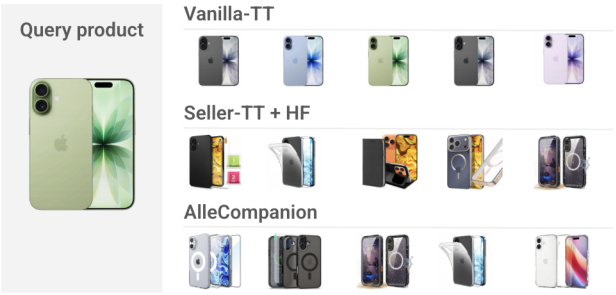

Figure 2: Recommendations generated by Vanilla-TT, Seller-TT + HF and AlleCompanion models.

4.1 Architecture Definition

The evaluation of AlleCompanion is strictly guided by our core design objectives. Specifically, the model must effectively learn from co-purchases to retrieve highly relevant items, explicitly generate candidates from targeted complementary categories, and promote same-seller recommendations to incentivise joint purchases. Furthermore, it must maintain a strong awareness of fine-grained product features essential for strict compatibility.

4.1.1 Metrics. To assess retrieval performance, we utilise standard Recall@k and MRR@k metrics. To provide a more qualitative analysis of the model's adherence to compatibility and platform incentives, we further evaluate the following consistency metrics:

- **Target Category Consistency** (Target Cat. Cons.): The average proportion of retrieved candidates that correctly belong to the intended complementary target category, evaluating the model's adherence to categorical mapping.
- **Seller Consistency** (Seller Cons.): The average proportion of recommended candidates offered by the exact same-seller as the query product, quantifying the model's effectiveness in encouraging single-parcel deliveries.
- **Attribute Consistency** (Attr. Cons.): The average overlap ratio of matching features (e.g., brand, model) between the query item and the retrieved candidates, serving as a proxy for strict item-level compatibility.

4.1.2 Model variants & evaluation protocol. To evaluate the effectiveness of our proposed architecture, we benchmark AlleCompanion against several baselines rooted in the standard Two Tower framework. This allows us to isolate the impact of our specific architectural contributions. We compare the following model configurations:

- **Vanilla Two Tower (Vanilla-TT)**: A standard dual-encoder trained on the transactional dataset described in Section 3.2, serving as a baseline for capturing general co-purchase relations.
- **Seller Two Tower (Seller-TT)**: An extension of the vanilla baseline that incorporates the seller ID feature, explicitly designed to encourage the model to generate same-seller recommendations.
- **Seller-TT with Hard Filtering (Seller-TT + HF)**: The Seller-TT model augmented with a post-processing heuristic. This baseline explicitly retrieves items from the complementary category by applying a same-seller hard filter to the candidates list after generation (300 candidates).
- **AlleCompanion (AlleCompanion w/o Attr)**: Our proposed architecture, trained with the seller ID feature. Unlike the hard-filtering baseline, this model utilises the Category Adapter to intrinsically guide candidate generation towards the target complementary categories directly within the latent space.
- **AlleCompanion**: The final version of our proposed model that incorporates fine-grained product attributes, extending its ability to enforce item-level compatibility constraints.

To ensure fair and consistent comparison, identical hyperparameters — determined through an independent tuning phase — were applied across all configurations. The models were trained using the AdamW optimizer, with the training process incorporating temperature scaling and mixed negative sampling. Training was conducted on a co-purchase dataset spanning 90 days of transaction data, refined through the filters and heuristics described in Section 3.2 (see Table 2 for detailed statistics). Performance was then evaluated on a subsequent 7-day holdout test set to ensure temporal separation. Two primary scenarios were employed:

- **Test**: A standard evaluation utilising ground-truth target categories from the holdout set
- **ComCat**: Evaluation using the generated category mapping (Section 3.4) to simulate online deployment where the ground-truth target category is unknown (Section 3.3).

4.1.3 Discussion. As observed in the results in Table 1, relying solely on the Vanilla-TT model to capture complementary relations leads to suboptimal retrieval performance, a degradation likely attributable to the high volume of noise inherent in raw co-purchase data. However, incorporating the seller feature (Seller-TT) proves highly beneficial, as it significantly strengthens the modelling of co-purchase dynamics. This improvement is largely driven by the high prevalence of same-merchant transactions in our data, with 59% of co-purchase pairs in the test set belonging to the same-seller. By explicitly capturing this signal, the model not only aligns with historical user behaviour but also substantially improves seller consistency in its recommendations.

While applying a hard post-filter (Seller-TT + HF) enforces target category alignment and boosts retrieval metrics, the approach is fundamentally unscalable. Even when oversampling an initial candidate pool 15 times larger than the target list size, the median number of successfully retrieved candidates decreased to 7, highlighting the severe inefficiency of post-hoc filtering.

In contrast, our proposed AlleCompanion w/o Attr architecture demonstrates substantial improvements in both Recall@20 and MRR@20. Inherently, by guiding candidate generation, it ensures high target category and seller consistency without the scalability bottlenecks of post-filtering. Finally, augmenting this architecture with fine-grained product attributes (AlleCompanion) yields the strongest overall performance, boosting both relevance metrics and attribute consistency. This enhances the model's capacity to recommend not only complements but also compatible items.

We observe that attribute consistency is highest for the baseline Two-Tower models; this is expected, as their primary objective is

Table 1: Offline results of complementary recommendations with AlleCompanion benchmarked against other content-based models. Best results are bolded.

	Evaluation: Test					Evaluation: ComCat	
Model	Recall@20 ↑	MRR@20 ↑	Target Cat. Cons. ↑	Seller Cons. ↑	Attr. Cons. ↑	Recall@20 ↑	MRR@20 ↑
Vanilla-TT	0.0447	0.0103	0.0654	0.2237	**0.4225**	-	-
Seller-TT	0.1219	0.0286	0.0967	**0.9309**	0.3309	-	-
Seller-TT + HF	0.3782	0.1789	**1.0000**	0.8334	0.2320	0.0451	0.0177
AlleCompanion w/o Attr	0.4537	0.1947	0.8283	0.6605	0.2155	0.0913	0.0244
AlleCompanion	**0.4567**	**0.2000**	0.8267	0.6619	0.2181	**0.0952**	**0.0258**

to retrieve similar items with overlapping features. Conversely, AlleCompanion shifts its focus towards cross-category compatibility, yielding an attribute consistency of 21.8%. Importantly, this aligns closely with the ground-truth distribution of the holdout test set, where organic co-purchases exhibit an empirical baseline of 21% attribute consistency.

Figure 2 provides a qualitative comparison of the candidates generated by the Vanilla-TT, Seller-TT + HF, and AlleCompanion models. In this example, the query item is an *Apple iPhone 17*, and the target complementary category is *Phone Cases and Covers*. The Vanilla-TT model focuses exclusively on the similarity aspect, primarily retrieving other iPhones that differ only in colour or built-in memory configuration. While the Seller-TT + HF model's candidates are successfully restricted to the target category, it fails to maintain strict compatibility, proposing cases for the *iPhone 17 Pro* (featuring three lenses instead of two). In contrast, the AlleCompanion model provides a diverse selection of *iPhone 17* cases, demonstrating its ability to simultaneously respect category constraints and fine-grained compatibility requirements.

4.2 Dataset Construction

While heuristics (Section 3.2) increase share of complementary pairs, they don't narrow down to precise item-level compatibility across specific attributes like brand, model or size. We hypothesized that integrating expert domain knowledge can shift the focus from broad category logic to strict technical alignment, capturing the underlying reasons *why* products truly complement one another.

4.2.1 Dataset variants & evaluation protocol. To evaluate the impact of this expert knowledge on model performance, the Transactions dataset (Section 3.1) is compared against three variants derived from rule-based expert logic (Section 3.4). These variants are designed to explore the trade-offs between data volume and precision (Table 2):

- **Transactions**: The primary dataset derived from co-purchase traffic, utilising the behavioural filters and heuristics previously described.
- **Filtered Transactions**: A subset of the Transactions dataset restricted to pairs that strictly match expert compatibility rules. While this produces a high-precision set of complementary pairs, it results in a 77% reduction in total data volume.
- **Expert Rules**: To maintain volume while minimising noise, synthetic pairs were generated using category and attribute constraints derived from rule-based expert logic. To ensure the relevance of the resulting pairs, the dataset was restricted to active products with high user engagement and constrained by price-proximity rules.
- **Expert Rules + Same-Seller**: A variant of the Expert Rules dataset with an additional same-seller filter applied to align the training distribution with business requirements.

Experiments employ the AlleCompanion architecture defined in Section 3.1, utilising the model parametrization and evaluation protocol detailed in Section 4.1.

4.2.2 Discussion. As illustrated in Table 3, the base Transactions dataset remains superior in terms of Recall@20 and MRR@20 across both evaluation scenarios, which is explained by its close alignment with the underlying test distribution. However, a notable trade-off is observed with the Filtered Transactions variant. While this configuration drops in relevance metrics, it provides the highest absolute boost to attribute consistency. This suggests that although expert filters restrict the model's discovery range, they successfully enforce stricter compatibility standards.

Conversely, training models exclusively on synthetic Expert Rules yields poor performance, emphasising that exposure to at least a portion of the historical co-purchase logs remains essential for effective recommendation. A compelling middle ground is offered by utilising synthetic datasets for pretraining followed by transaction-based finetuning, which maintains high relevance metrics while marginally increasing attribute consistency.

4.3 Complementary Categories Mapping

Fundamentally, the ComCat mapping addresses two main challenges: the lack of explicit target categories in online environments, and the need to mitigate noise in historical co-purchase datasets (Section 3.2). To evaluate its effectiveness, an offline ablation study was conducted across four configurations of the mapping framework (Section 3.4).

While the platform spans over 20,000 categories, results show that all configurations consistently cover approximately 42% of source query categories. As shown in Table 4, while this source coverage remains stable, target category coverage increases significantly from 27.55% in the baseline to 43.06% in the final configuration. This confirms that adding more sources enriches the mapping with a deeper variety of high-quality target pairs without requiring a massive expansion of the source category set. The impact of this enrichment is most evident in the target distribution: while the median (p50) count of target categories remains stable at 3 or 4, the

Table 2: Statistics for training configurations corresponding to 90 days of purchase activity.

Dataset	Trainset	Unique items	Unique categories
Transactions	3,611,848	779,199	7,555
Filtered Transactions	829,841	326,238	3,004
Expert Rules	3,015,027	1,446,727	5,268
Expert Rules + Same-Seller	2,553,622	1,029,244	5,131

Table 3: Performance of the AlleCompanion model across various training configurations. Best and second-best results are bolded and italicized, respectively.

		Evaluation: Test			Evaluation: ComCat		
Pretrain	**Finetune**	**Recall@20 ↑**	**MRR@20 ↑**	**Attr. Cons. ↑**	**Recall@20 ↑**	**MRR@20 ↑**	**Attr. Cons. ↑**
–	Transactions	**0.4567**	**0.2000**	0.2181	**0.0957**	**0.0260**	0.2411
–	Filtered Transactions	0.3222	0.1221	**0.2433**	0.0801	0.0213	**0.2541**
–	Expert Rules	0.0818	0.0243	0.1815	0.0127	0.0035	0.2070
–	Expert Rules + Same-Seller	0.2717	0.0960	0.2017	0.0471	0.0124	0.2288
Expert Rules	Transactions	0.4458	0.1932	0.2183	0.0934	0.0253	0.2404
Expert Rules + Same-Seller	Transactions	0.4429	0.1918	0.2191	0.0926	0.0252	0.2420

95th percentile (p95) expands substantially from 3 to 10 categories as annotations, expert rules, and query categories are integrated.

To evaluate how these structural changes perform under realistic production conditions, an evaluation dataset was generated based on historical user traffic with known query items and engaged co-purchased or co-clicked categories. This setup enables the measurement of online-like CTR and CVR metrics in an offline experiment. Within this traffic-based dataset, the 42% source coverage translates to 99.8% of live traffic, confirming that the framework successfully provides at least one complementary target for virtually every relevant user interaction. Consequently, these mappings prove highly effective in practice by covering the specific categories that drive the majority of user interactions. These quality gains are directly reflected in the final metrics: relative to the baseline (1), the full ensemble in (4) yields a +450% increase in CTR and a +398% boost in CVR. The substantial performance leap observed in (4) when including the same-category relation (Section 3.4.2) suggests that, within the context of the analysed placement, users have a strong expectation for alternative product suggestions alongside complementary ones.

4.4 Online A/B testing

To validate the real-world efficacy of the AlleCompanion model, a series of online A/B tests was conducted across key platform placements. All organic carousels were limited to same-seller items, aligning with business logic intended to help users reach the free delivery threshold (MOV). Performance was measured primarily through Visit Conversion (v-CVR), representing the ratio of visits resulting in a purchase, Carousel Conversion (c-CVR), representing the ratio of carousel clicks resulting in a purchase, and the Gross Merchandise Value (GMV) directly linked to carousel interactions. A summary of these results is provided in Table 5.

All A/B tests were performed on the web-facing version of Allegro (available on desktop and mobile web, referenced as Web) and the Allegro mobile app (referenced as App). Each experiment was conducted over a two-week period and routed 100% of the platform traffic. AlleCompanion was evaluated as an additional retrieval source alongside existing production models.

4.4.1 Product Page. Testing on the Product Page focused on two carousels — Sponsored and Organic — to evaluate how different recommendation strategies align with shifting user needs. In both placements, AlleCompanion was benchmarked against the existing production baseline, which combined item-to-item collaborative filtering on purchases, seller bestsellers, and seller new arrivals.

In the Sponsored placement (1), titled *Suggestions for you*, AlleCompanion was evaluated utilising the ComCat mechanism, focusing exclusively on complements. This configuration yielded a +0.53%∗ and +0.13% increase in v-CVR on Web and App, respectively. Although non-significant fluctuations were observed in GMV during the test, the carousel alone saw an approximate 50% boost in ad revenue across both platforms, attributed to uplifts in CTR.

Conversely, the Organic carousel (2), titled *Order in one shipment*, serves as a broader discovery tool. Initial evaluations suggested that purely complementary recommendations were overly restrictive for this placement, as users in this context often seek alternative products to enrich their selection alongside complements. By expanding the mapping with a same-category relation to support a "*complements + substitutes*" combination (Section 3.4.2), the system effectively broadened its discovery range. This was reflected by a boost in GMV of +9.35%∗ on App and +8.05%∗ on the Web, while v-CVR showed neutral fluctuations.

4.4.2 Cart Placements. Cart placements appear during the final stages of the user journey, including the "pre-cart" pop-up window

Table 4: Ablation Study of Category Mapping Sources based on offline traffic data from organic carousel on product page. Best results are bolded.

Sources of category mapping					Taxonomy coverage ↑		Target counts ↑		Traffic metrics ↑	
No.	Heuristics	Annotations	Rules	Same-Category	Query	Targets	p50	p95	CTR	CVR
(1)	✓				41.77%	27.55%	3	3	-	-
(2)	✓	✓			41.89%	28.69%	3	5	+25%	+36%
(3)	✓	✓	✓		42.32%	32.46%	3	9	+125%	+136%
(4)	✓	✓	✓	✓	**42.32%**	**43.06%**	**4**	**10**	**+450%**	**+398%**

Table 5: Summary of Online A/B Test Results. Metrics represent relative change against the production baseline. (∗ denotes statistical significance at $p < 0.005$).

				Platform	
No.	**Placement**	**Type**	**Metric**	**Web ↑**	**App ↑**
(1)	Product Page	Sponsored	v-CVR	+0.53%∗	+0.13%
			GMV	+0.19%	-0.01%
(2)	Product Page	Organic	v-CVR	+0.57%	-0.07%
			GMV	+8.05%∗	+9.35%∗
(3)	Pre-Cart	Organic (filtered)	v-CVR	-0.28%∗	-0.04%
			c-CVR	-0.51%	-0.42%
			GMV	+0.17%	-0.27%
(4)	Pre-Cart	Organic (finetuned)	v-CVR	-0.13%	-0.17%
			c-CVR	+0.69%	+0.35%
			GMV	-0.06%	-0.28%
(5)	In-Cart	Organic	c-CVR	+4.98%∗	+1.21%
			GMV	+21.25%∗	+15.73%∗

and the "in-cart" view during checkout. In these placements, AlleCompanion was evaluated against a production baseline consisting of item-to-item collaborative filtering on purchases, seller bestsellers, and recurring purchases. These placements often feature carousels with slogans such as *Buy more from this seller to get free delivery*, explicitly prompting users towards the MOV threshold. Since these placements specifically target order value optimization near checkout, GMV serves as the primary metric, and our analysis focuses exclusively on this outcome.

Two different strategies were tested in the Pre-cart layer. The first (3) evaluated variants from the dataset construction experiments (Section 4.2) to assess whether functional compatibility is preferred over relevance measured in offline. Specifically, the (3) Filtered Transactions variant (selected for its high attribute consistency) and the (4) Pretrained + Finetuned Same-Seller variant (selected as a balanced middle ground) were evaluated. The second strategy utilised a base transactional model with a same-category mapping (Section 3.4.2) to introduce alternative product suggestions. While most core business metrics across both strategies showed non-significant fluctuations in GMV and c-CVR across platforms, the (3) Filtered Transactions variant led to a statistically significant -0.28%∗ drop in v-CVR on Web. Overall, these findings suggest that neither the inclusion of alternatives nor the use of expert-filtered rules fully satisfied user needs for this specific intent.

Interestingly, the In-cart placement (5) revealed a different dynamic. The same-category mapping that yielded neutral results in the pre-cart layer performed exceptionally well during the final checkout stage. In carousels encouraging users to consolidate their shipments, the model achieved increases of +15.73%∗ and +21.25%∗ in GMV on App and Web, respectively.

Based on A/B tests (1), (2) and (5), the model has been deployed to all three placements. Each deployment was justified by a statistically significant improvement in at least one primary metric (CVR or GMV) on at least one platform, without negatively impacting the remaining metrics.

5 Conclusions

We present AlleCompanion, a universal production-scale retrieval framework for complementary item recommendations. To guide basket building effectively, our design separates item-level fit from category-level intent, combining explicit item-level constraints in the input feature space with category guidance imposed via category adapter. By decoupling the ComCat category mapping from the retrieval model, this architecture provides a flexible, production-friendly paradigm where category policies can be updated dynamically without expensive model retraining.

Our online experiments reveal a crucial insight: while baseline behavioural filtering is necessary to clear out transaction outliers, the model benefits most from exposure to the full remaining distribution of co-purchase traffic, whereas enforcing strict, domain-specific restrictions degrades overall retrieval performance. Consequently, we decouple data precision from model training by shifting expert logic, LLM-based reasoning, and human feedback entirely into our ComCat mapping layer. Importantly, while designed for complementary recommendations, our production deployment revealed that the strongest business gains emerged when expanding the system to support a mixed related-product strategy (blending complements with same-category alternatives). This highlights that real-world checkout intents often favor broader discovery over strictly complementary options. Deployed at scale to serve over 20 million active users at product page and cart, AlleCompanion successfully bridges the gap between co-purchase and compatibility, delivering a +8-9% GMV increase in organic discovery on the product page, +15-21% GMV uplift in the cart, and a 50% revenue boost in sponsored placements.

While AlleCompanion successfully addresses large-scale complementary retrieval, certain design choices introduce opportunities for further development. Firstly, because our framework relies on

the ComCat mapping as the primary driver of complementary constraints, its precision is naturally bounded by the granularity of the underlying taxonomy, which can occasionally obscure item-level nuances or encounter coverage gaps in extreme cold-start categories. Although this trend is not currently visible in the traffic logs, as the framework successfully covers 99.8% of active customer interactions, the architecture could be generalized to learn directly from raw product features. Secondly, the current deployment prioritizes broad item-to-item retrieval and interleaving to boost carousel diversity, leaving room to incorporate user personalization directly into candidate generation or via a dedicated downstream ranking layer. Finally, evaluating multi-item complementary intent in an offline manner remains difficult due to the inherent feedback loop from the production system. Consequently, online A/B testing remains the gold standard for reliably measuring true multi-item purchase dynamics.

Acknowledgments

This work is the result of a collaborative effort within the recommendation systems team at Allegro. We would like to extend our gratitude to fellow researchers Paweł Młyniec, Mateusz Marzec, and Bartłomiej Szołkowski for their invaluable support. We also thank the engineering team — Jakub Demianowski and Mateusz Lamecki — as well as former members Krzysztof Szczepański, Maciej Arciuch, Elwira Hołowko and Marcin Cylke for their foundational contributions to ML-based complementary recommendations at Allegro. Special thanks also go to the Paloma team behind the expert rule-based systems.

References

[1] Talha Bayır and Gökhan Akel. 2024. Gamification in mobile shopping applications: A review in terms of technology acceptance model. *Multimedia Tools and Applications* 83, 16 (May 2024), 47247–47268. doi:10.1007/s11042-023-16823-7

[2] Koby Bibas, Oren Sar Shalom, and Dietmar Jannach. 2023. Semi-supervised Adversarial Learning for Complementary Item Recommendation. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. ACM, New York, NY, USA, 1804–1812. doi:10.1145/3543507.3583462

[3] Huajie Chen, Jiyuan He, Weisheng Xu, Tao Feng, Ming Liu, Tianyu Song, Runfeng Yao, and Yuanyuan Qiao. 2023. Enhanced Multi-Relationships Integration Graph Convolutional Network for Inferring Substitutable and Complementary Items. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (June 2023), 4157–4165. doi:10.1609/aaai.v37i4.25532

[4] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, 191–198. doi:10.1145/2959100.2959190

[5] Junheng Hao, Tong Zhao, Jin Li, Xin Luna Dong, Christos Faloutsos, Yizhou Sun, and Wei Wang. 2020. P-Companion: A Principled Framework for Diversified Complementary Product Recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. ACM, Virtual Event Ireland, 2517–2524. doi:10.1145/3340531.3412732

[6] Jie Huang, Yifan Gao, Zheng Li, Jingfeng Yang, Yangqiu Song, Chao Zhang, Zining Zhu, Haoming Jiang, Kevin Chen-Chuan Chang, and Bing Yin. 2023. CCGen: Explainable Complementary Concept Generation in E-Commerce. doi:10.48550/arXiv.2305.11480 arXiv:2305.11480.

[7] Byung Eun Jeon, Ryan Bae, and Xiao Bai. 2025. Leveraging Large Language Models for Complementary Product Ads Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*. ACM, Seoul Republic of Korea, 4837–4841. doi:10.1145/3746252.3760940

[8] Linyue Li and Zhijuan Du. 2024. Complementary Recommendation in E-commerce: Definition, Approaches, and Future Directions. doi:10.48550/arXiv.2403.16135 arXiv:2403.16135.

[9] Zelong Li, Yan Liang, Ming Wang, Sungro Yoon, Jiaying Shi, Xin Shen, Xiang He, Chenwei Zhang, Wenyi Wu, Hanbo Wang, Jin Li, Jim Chan, and Yongfeng Zhang. 2024. Explainable and Coherent Complement Recommendation Based on Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. ACM, Boise ID USA, 4678–4685. doi:10.1145/3627673.3680028

[10] Haitong Luo, Xuying Meng, Suhang Wang, Hanyun Cao, Weiyao Zhang, Yequan Wang, and Yujun Zhang. 2024. Spectral-Based Graph Neural Networks for Complementary Item Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (March 2024), 8868–8876. doi:10.1609/aaai.v38i8.28734

[11] Mansi Ranjit Mane, Stephen Guo, and Kannan Achan. 2019. Complementary-Similarity Learning using Quadruplet Network. doi:10.48550/arXiv.1908.09928 arXiv:1908.09928.

[12] Soroush Mokhtari, Muhammad Tayyab Asif, and Sergiy Zubatiy. 2026. T-REX: Transformer-Based Category Sequence Generation for Grocery Basket Recommendation. doi:10.48550/arXiv.2603.06631 arXiv:2603.06631.

[13] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. https://arxiv.org/abs/1807.03748v2

[14] Aleksandra Maria Osowska-Kurczab, Klaudia Nazarko, Mateusz Marzec, Lidia Wojciechowska, and Eliška Kremeňová. 2025. Suggest, Complement, Inspire: Story of Two-Tower Recommendations at Allegro.com. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. ACM, Prague Czech Republic, 1095–1098. doi:10.1145/3705328.3748135

[15] Rastislav Papso. 2023. Complementary Product Recommendation for Long-tail Products. In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM, Singapore Singapore, 1305–1311. doi:10.1145/3604915.3608864

[16] Kai Sugahara, Chihiro Yamasaki, and Kazushi Okamoto. 2024. Is It Really Complementary? Revisiting Behavior-based Labels for Complementary Recommendation. In *18th ACM Conference on Recommender Systems*. ACM, Bari Italy, 1091–1095. doi:10.1145/3640457.3691705

[17] Zhu Sun, Jie Yang, Kaidong Feng, Hui Fang, Xinghua Qu, and Yew Soon Ong. 2022. Revisiting Bundle Recommendation: Datasets, Tasks, Challenges and Opportunities for Intent-aware Product Bundling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Madrid Spain, 2900–2911. doi:10.1145/3477495.3531904

[18] Junting Wang, Chenghuan Guo, Jiao Yang, Yanhui Guo, Yan Gao, and Hari Sundaram. 2025. Multi-modal Relational Item Representation Learning for Inferring Substitutable and Complementary Items. doi:10.48550/arXiv.2507.22268 arXiv:2507.22268.

[19] Da Xu, Chuanwei Ruan, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2020. Knowledge-aware Complementary Product Representation Learning. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. ACM, Houston TX USA, 681–689. doi:10.1145/3336191.3371854

[20] Chihiro Yamasaki, Kai Sugahara, and Kazushi Okamoto. 2025. Knowledge-Augmented Relation Learning for Complementary Recommendation with Large Language Models. In *2nd Workshop on Generative AI for E-Commerce (RecSys '25)*. ACM. doi:10.48550/arXiv.2509.05564

[21] An Yan, Chaosheng Dong, Yan Gao, Jinmiao Fu, Tong Zhao, Yi Sun, and Julian Mcauley. 2022. Personalized Complementary Product Recommendation. In *Companion Proceedings of the Web Conference 2022*. ACM, Virtual Event, Lyon France, 146–151. doi:10.1145/3487553.3524222

[22] Ji Yang, Xinyang Yi, Derek Zhiyuan Cheng, Lichan Hong, Yang Li, Simon Xiaoming Wang, Taibai Xu, and Ed H. Chi. 2020. Mixed Negative Sampling for Learning Two-tower Neural Networks in Recommendations. In *Companion Proceedings of the Web Conference 2020 (WWW '20)*. ACM, New York, NY, USA, 441–447. doi:10.1145/3366424.3386195

[23] Nasser Zalmout, Chenwei Zhang, Xian Li, Yan Liang, and Xin Luna Dong. 2021. All You Need to Know to Build a Product Knowledge Graph. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event Singapore, 4090–4091. doi:10.1145/3447548.3470825

[24] Wei Zhang, Zeyuan Chen, Hongyuan Zha, and Jianyong Wang. 2022. Learning from Substitutable and Complementary Relations for Graph-based Sequential Product Recommendation. *ACM Transactions on Information Systems* 40, 2 (April 2022), 1–28. doi:10.1145/3464302

[25] Huasha Zhao, Luo Si, Xiaogang Li, and Qiong Zhang. 2017. Recommending Complementary Products in E-Commerce Push Notifications with a Mixture Model Approach. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Shinjuku Tokyo Japan, 909–912. doi:10.1145/3077136.3080676