

# Suggest, complement, inspire: story of Two Tower recommendations at Allegro.com

Aleksandra Osowska-Kurczab\*

aleksandra.kurczab@allegro.com

Allegro.com

Poland

Klaudia Nazarko\*

klaudia.nazarko@allegro.com

Allegro.com

Poland

Mateusz Marzec\*

mateusz.marzec@allegro.com

Allegro.com

Poland

Lidia Wojciechowska

Allegro.com

Poland

Eliška Kremeňová

Allegro.com

Czech Republic

## Abstract

Building large-scale e-commerce recommendation systems requires addressing three key technical challenges: (1) designing a universal recommendation architecture across dozens of placements, (2) decreasing excessive maintenance costs, and (3) managing a highly dynamic product catalogue. This paper presents a unified content-based recommendation system deployed at Allegro.com, the largest e-commerce platform of European origin. The system is built on a prevalent Two Tower retrieval framework, representing products using textual and structured attributes, which enables efficient retrieval via Approximate Nearest Neighbour search. We demonstrate how the same model architecture can be adapted to serve three distinct recommendation tasks: similarity search, complementary product suggestions, and inspirational content discovery, by modifying only a handful of components in either the model or the serving logic. Extensive A/B testing over two years confirms significant gains in engagement and profit-based metrics across desktop and mobile app channels. Our results show that a flexible, scalable architecture can serve diverse user intents with minimal maintenance overhead.

## CCS Concepts

• **Information systems** → **Recommender systems**.

## Keywords

Recommendation systems, Large Scale Retrieval, Complementary recommendations, Diverse recommendations, E-commerce

## ACM Reference Format:

Aleksandra 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 (RecSys '25)*, September 22–26, 2025, Prague, Czech Republic. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3705328.3748135>

\*All authors contributed equally to this research.

## 1 Introduction

Allegro is a major Central European e-commerce marketplace where over 20 million active buyers connect with more than 150 thousand sellers monthly to discover and purchase products. Recommendations are pivotal in driving organic and sponsored content discovery, with a significant share of attributed Gross Merchandise Value (GMV) and advertising revenue. Managing dozens of recommendation placements across the user journey introduces challenges in maintaining a universal and coherent system. Additionally, the platform's dynamic catalogue – with its vast array of products, sellers, and users – introduces complexities such as cold-start and long-tail distributions, highlighting the need for scalable and adaptable recommendation strategies.

A common approach to solving a recommendation problem is content-based filtering [13], which suggests items to users based on the intrinsic attributes of those items and historical user preferences. It matches item features to a user's profile, effectively framing recommendation as an information retrieval task focused on content similarity [2]. Alongside collaborative filtering [8], this method serves as a fundamental candidate generator in large-scale recommendation systems [2].

The Two Tower (TT) model [5, 10] is prominent in industrial recommendation systems due to its balance between predictive effectiveness and serving efficiency. It encodes inputs – typically a *query* representing user context and a *target* representing an item from the catalogue – into a shared embedding space using a deep learning model (DLRM), where relevance is inferred via the dot product of embeddings. Given the scale of large e-commerce platforms [14], exact similarity search becomes computationally infeasible; consequently, the combination of a DLRM encoder and Approximate Nearest Neighbour (ANN) indexing [3, 6] is crucial for effective TT-based content recommendations.

With a wide range of placements to populate, making dedicated solutions for each scenario is impractical and expensive [12]. To address this, the industry usually turns to one of two paths: either large foundation models [4, 17], which require sophisticated infrastructure for serving, or many domain-specific models [7, 11], which then result in high maintenance costs and complexity. We propose an architectural design to alleviate the issue:

- we present a unified platform architecture tailored to content-based recommendations and demonstrate its effectiveness across three seemingly distinct tasks: similarity, complementary and inspirational recommendations (Fig. 1).

- we generalise these tasks by purposefully redefining the complementary and inspirational recommendation problems as variants of similarity search.
- we summarise insights from two years of continuous A/B testing on the Allegro platform to highlight the business impact of the proposed solution.

## 2 Methods

### 2.1 Similarity search with the Two Tower

Two Tower model (Similarity-TT) is a canonical application of DLRM-based retrieval [2, 5, 16] deployed at Allegro for similarity search tasks (e.g. retrieving substitute or similar products). It is trained to maximise the similarity between vector representations of query and target items from the product catalogue (item-to-item regime [5]). Due to the massive amount of products (in a scale of hundreds of millions), the task is formulated as a classification problem with sampled softmax loss function [2] and mixed negative sampling strategy [15].

The high amount of volatile products makes training a distinct and learnable embedding for each product ID challenging and prone to overfitting [9]. To address this, each product is represented by its content features, such as title, price, and category (with a hierarchical taxonomy). Each feature is transformed into a low-dimensional vector via a dedicated embedding table. All feature vectors are concatenated, passed through a multi-layer perceptron (FC), and L2-normalised. We apply weight tying between the query and target towers to enhance training speed and stability, creating a shared component collectively called the Product Encoder.

The architecture of the service has a distinct separation into offline and online components. Offline processing consists of data preparation, model training, evaluation, and index building. Training data comprises co-viewed product pairs, filtered to retain only pairs meeting a minimum co-occurrence threshold. Given the model's lightweight architecture, training can be completed efficiently on a machine with a single GPU (NVIDIA T4 16GB). Our models are retrained regularly, and ANN indexes are refreshed daily. The Faiss library [3] handles index creation and online serving. Recommendations are served in real time with millisecond-level latency. The online service processes an incoming request by encoding the associated product using the Product Encoder and retrieving similar candidates from the indexed product catalogue.

Representing products solely based on their content features provides several advantages within Allegro's dynamic environment. The content-based recommendations complement existing collaborative filtering models and effectively mitigate the cold-start problem by generating representations for new products directly from their features. The lightweight content-based model eliminates the need for extensive product ID embedding tables, improving computational efficiency for offline training and online serving.

### 2.2 Complementary Two Tower

Complementary Two Tower (Complementary-TT) aims to provide supplementary product recommendations (e.g. tennis balls to tennis rackets). It is achieved by modifying the Product Encoder while keeping the same serving architecture as in similarity-TT.



Figure 1: Recommendations generated by Similarity-TT, Complementary-TT and Inspirational-TT models.

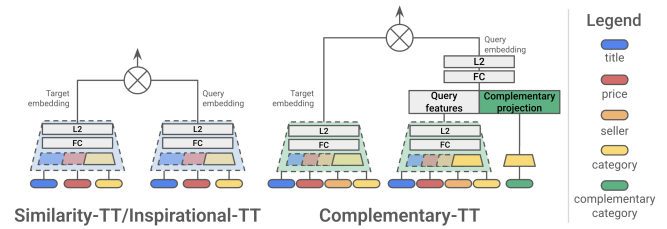


Figure 2: Comparison of two proposed TT architectures.

The modifications are implemented in the query tower, with the target tower remaining unchanged (Fig. 2). The query tower takes an additional input, a complementary categories mapping (one-to-many), derived from the statistical models fit on co-purchase data, external annotations and domain knowledge. Firstly, the target category is taken from the mapping and embedded with the same embedding tables as category inputs in Product Encoder. Then, the query product embedding is concatenated with the target category embedding and used as the final representation of the query [7]. Inspired by [1], the loss function is enhanced with the target category reconstruction error to enforce the correctness of the target category embedding. Besides the default Similarity-TT input features, the seller feature is used (due to co-purchase incentives, users tend to buy products from the same seller in a single transaction). The model is trained on co-purchase pairs, filtered to examples that follow a complementarity relation heuristic.

The online serving infrastructure remains the same as in the Similarity-TT model, with the only modification being the querying of the model with both the query product and the target complementary category. When the complementary categories mapping points to multiple target categories, the resulting groups of candidates are interleaved to enhance visual appearance of the carousel.

### 2.3 Inspirational Two Tower

The goal of the Inspirational Two Tower (Inspirational-TT) algorithm is to encourage exploration by suggesting personalised products that are diverse and relevant to the user's browsing history. Such recommendations are obtained using the Similarity-TT Product Encoder with hierarchical ANN indexes to enable controllable diversification (similarly to [11]).

A hierarchical ANN index is constructed as follows: products are embedded with the Product Encoder, and then the k-means clustering algorithm is applied to product representations to generate

$k$  distinct clusters (second-level indexes). Each cluster is represented by the centroid and added to the top-level index. The user is represented by the last 100 product views from the past 7 days, aggregated into categories. The most recently viewed product from a given category serves as a category representative. During the inference, the Product Encoder is queried with those representatives. Next, the top-level index is queried with the obtained representations, and  $n$  closest clusters per category are selected: the more clusters, the more diverse the results. To avoid results that are too similar to the query,  $l$  closest clusters may be skipped. In the third step, the most similar products to each category are fetched from second-level indexes and interleaved as the list of candidates.

While the Product Encoder architecture remains untouched, the online infrastructure must be adjusted by implementing hierarchical ANN indexes and introducing the aggregation of users' recently viewed products.

### 3 Results

Recommendation systems at Allegro.com operate at production scale, powering online services with a throughput of 20k requests per second and 40ms p99 CPU latency. Models serve to address a variety of user intents, from exploration to purchase decisions. One recommendation system can be used to produce outputs for different recommendation scenarios (Fig. 1). With the query product being a bike, the regular TT model retrieves very similar products, mainly bicycles of the same brand and model, varying only in colour. The Complementary-TT model returns recommendations of supplementary products, such as helmets and knee pads. Results obtained from the Inspirational-TT model are characterised by more diversity: they are visually appealing and loosely connected accessories, such as bicycle bells, lamps, and decorations.

#### 3.1 Product related recommendations

Product page remains one of the most prominent placements for recommendations at Allegro.com. It showcases products and enables navigation via recommendation carousels. Both Similarity-TT and Complementary-TT models were evaluated here as fallback candidate generators to collaborative-filtering (co-viewed and co-purchased model respectively). A/B tests were run with division to desktop and mobile app traffic.

Similarity-TT model was evaluated in the carousel "Others also viewed", which serves to present substitute products. Complementary-TT was tested in the carousel "Order in one parcel," which helps users find additional products from the same seller to complement their order and reach a minimal order value for free delivery. Due to business priorities, the latter test was run in the Sports, Travel and Fashion departments. The two primary metrics for online evaluation were Click-Through Rate (CTR) and GMV per visit. The first metric evaluates user engagement, while the second measures financial profits delivered by the tested solution.

Both TT and Complementary-TT positively influence CTR, suggesting that content-based models can capture hidden product relations and recommend items well suited to user needs (Table 1). Increases in GMV per visit are mainly visible in desktop traffic, which can be explained by the differences in usage patterns between desktop and mobile devices.

**Table 1: A/B test results of Similarity-TT and Complementary-TT on product page. Values indicate % change from the baseline, \* denotes statistical significance at the 0.01 level.**

model	mobile app		desktop	
	CTR ↑	GMV ↑	CTR ↑	GMV ↑
Similarity-TT	+2.11% *	+0.13%	+2.37% *	+0.29%
Complementary-TT	+1.62% *	+0.09%	+1.06% *	+0.31%

**Table 2: A/B test results of introducing inspirational recommendations on product page (desktop traffic). Values indicate % change from the baseline, \* denotes statistical significance at the 0.01 level, and the best result is bolded.**

view	CTA ↑	CVR ↑	bounce rate ↓	exit rate ↓
carousel	+3.12% *	+1.38% *	-4.09% *	-1.82% *
infinite feed	<b>+4.15% *</b>	<b>+2.22% *</b>	<b>-5.74% *</b>	-1.66% *

#### 3.2 Inspirational recommendations

Highly relevant product page recommendations perfectly fit customers with a purchase intent. However, the product page should also provide incentives to explore more and engage users looking for inspiration. Enhancing the product page with an Inspirational-TT recommender in the "How about..." section was the subject of an A/B test on desktop traffic. Two test variants differed in how inspirational recommendations were displayed: as an additional carousel vs an infinite feed. In both variants, models were queried with a single, currently viewed product (without the extended context of previously viewed products).

The focus of the A/B test was to improve engagement metrics, such as the Call To Action (CTA) and Conversion Rate (CVR), which calculate the ratio of users who engaged with the carousel. In addition, two auxiliary metrics were evaluated: exit rate and bounce rate, which measure the ratio of users who abandoned the page.

Presenting inspirational content on the product page led to substantial improvements in user engagement (Table 2). The CTA increased by 3.12% in the carousel view and by 4.15% in the infinite feed layout. It indicates that inspirational recommendations capture user attention and encourage them to explore more.

### 4 Conclusions

This work demonstrates that Two-Tower system architecture for similarity search can be easily adjusted for serving complementary or inspirational content. Extensive A/B tests show improvements in both engagement and profit-based metrics while maintaining minimal maintenance costs. This proves that the proposed solution can scale and adapt to numerous recommendation placements, fulfilling diverse user intents across their journey on the platform. Nonetheless, performance depends on content features quality, and deployment success requires coherence between encoder and indexing mechanism. Future work involves integrating user context into the model and evaluating its production impact.

## Presenters' bio

**Aleksandra Osowska-Kurczab, PhD** is a Machine Learning Manager leading the research team developing retrieval and ranking models deployed in e-commerce applications at Allegro.com. Her research interests include representation learning and robustness. **Klaudia Nazarko** is a Machine Learning Research Engineer at Allegro.com, where she works on e-commerce recommender systems. Her work is focused on personalised retrieval and ranking. **Mateusz Marzec** is a Machine Learning Research Engineer at Allegro.com, specialising in machine learning for e-commerce recommendations. Daily, he works on improving personalisation in retrieval systems.

## Acknowledgments

This work is the result of a collaborative effort in the area of recommendation systems at Allegro.com for the past 2 years. We thank the alumni researchers: Michał Bień, Elwira Hołowko, Piotr Januszewski, Marcin Cylke for their foundational contributions to ML projects, and the engineering team: Maciej Arciuch, Mateusz Lamecki, Jakub Demianowski, Krzysztof Szczepański, Uładzislau Dziuba for their implementation support.

## References

- [1] 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 (WWW '23)*. ACM, 1804–1812. doi:10.1145/3543507.3583462
- [2] 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
- [3] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. doi:10.48550/arXiv.2401.08281 arXiv:2401.08281.
- [4] Hamed Firooz, Maziar Sanjabi, Adrian Englhardt, Aman Gupta, Ben Levine, Dre Olgiati, Gungor Polatkan, Iuliia Melnychuk, Karthik Ramgopal, Kirill Talanin, Kutta Srinivasan, Luke Simon, Natesh Sivasubramoniapillai, Necip Ayan, Qingquan Song, Samira Sriram, Souvik Ghosh, Tao Song, Vignesh Kothapalli, Xiaoling Zhai, Ya Xu, Yu Wang, and Yun Dai. 2025. 360Brew: A Decoder-only Foundation Model for Personalized Ranking and Recommendation. doi:10.48550/arXiv.2501.16450 arXiv:2501.16450 version: 1.
- [5] Daniel A. Galron, Yuri M. Brovman, Jin Chung, Michal Wieja, and Paul Wang. 2018. Deep Item-based Collaborative Filtering for Sparse Implicit Feedback. doi:10.48550/arXiv.1812.10546 arXiv:1812.10546.
- [6] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. 2020. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In *Proceedings of the 37th International Conference on Machine Learning (ICML '20)*. JMLR.org, Article 364, 10 pages.
- [7] 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 (CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2517–2524. doi:10.1145/3340531.3412732
- [8] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 173–182. doi:10.1145/3038912.3052569
- [9] Yi-Ping Hsu, Po-Wei Wang, Chantat Eksombatchai, and Jiajing Xu. 2024. Taming the One-Epoch Phenomenon in Online Recommendation System by Two-stage Contrastive ID Pre-training. In *18th ACM Conference on Recommender Systems*. ACM, Bari Italy, 838–840. doi:10.1145/3640457.3688053
- [10] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole Wu, Alisson Azzolini, Dmytro Dzhulgakov, Andrey Malleevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. doi:10.48550/arXiv.1906.00091 arXiv:1906.00091.
- [11] Aditya Pal, Chantat Eksombatchai, Yitong Zhou, Bo Zhao, Charles Rosenberg, and Jure Leskovec. 2020. PinnerSage: Multi-Modal User Embedding Framework for Recommendations at Pinterest. In *Proceedings of 26th ACM International Conference on Knowledge Discovery (KDD '20)*. ACM. doi:10.1145/3394486.3403280
- [12] Iulia Paun. 2020. Efficiency-Effectiveness Trade-offs in Recommendation Systems. In *Fourteenth ACM Conference on Recommender Systems*. ACM, Virtual Event Brazil, 770–775. doi:10.1145/3383313.3411452
- [13] Jieun Son and Seoung Bum Kim. 2017. Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications* 89 (Dec. 2017), 404–412. doi:10.1016/j.eswa.2017.08.008
- [14] Tian Wang, Yuri M. Brovman, and Sriganesh Madhvanath. 2021. Personalized Embedding-based e-Commerce Recommendations at eBay. doi:10.48550/arXiv.2102.06156 arXiv:2102.06156.
- [15] 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)*. Association for Computing Machinery, New York, NY, USA, 441–447. doi:10.1145/3366424.3386195
- [16] Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM, Copenhagen Denmark, 269–277. doi:10.1145/3298689.3346996
- [17] Jiaqi Zhai, Lucy Liao, Xing Liu, Yueming Wang, Rui Li, Xuan Cao, Leon Gao, Zhaojie Gong, Fangda Gu, Jiayuan He, Yinghai Lu, and Yu Shi. 2024. Actions speak louder than words: trillion-parameter sequential transducers for generative recommendations. In *Proceedings of the 41st International Conference on Machine Learning (ICML '24)*. JMLR.org, Article 2414, 26 pages.