*applied sciences*

MDPI

*Article*

# A NeuMF-Based Framework for Sparse E-Commerce Recommendation with Cold-Start Fallback

**Nermeen Kashef [1], Ola Ahmed Ali [1,*], El-Sayed M.E. Mostafa [1] and Mahmoud Gamal [2]**

[1] Department of Mathematics and Computer Science, Faculty of Science, Alexandria University, Alexandria 21526, Egypt; nermeen.kashief@alexu.edu.eg (N.K.)
[2] Department of Computing and Information Systems, Faculty of Computers and Data Science, Alexandria University, Alexandria 21526, Egypt; mahmoud.gamal@alexu.edu.eg
* Correspondence: ola.ali_pg@alexu.edu.eg

**Abstract**

Personalized recommender systems have become essential in modern e-commerce by improving product discovery and enhancing user experience. However, recommendation performance is often limited by sparse user–item interactions and cold-start scenarios, where insufficient historical data reduces personalization accuracy. This study evaluates a hybrid recommendation framework that uses a NeuMF-based model for users with historical interactions and a popularity-based fallback for completely unseen users. The NeuMF model combines Generalized Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) components, with Batch Normalization and Dropout examined through controlled ablation experiments. The evaluation separates three tasks: rating prediction, full-catalog Top-K ranking, and true user cold-start recommendation. On the Amazon Electronics dataset, the proposed NeuMF variant achieved a lower MAE than the baseline NeuMF model (0.8603 vs. 0.8837), while the baseline obtained a lower RMSE (1.2172 vs. 1.2362). The difference in NDCG@10 between the proposed and baseline models was not statistically significant. In full-catalog ranking, LightGCN achieved the strongest overall performance among the evaluated models. For completely unseen users, where no learned user embedding was available, a popularity-based fallback provided better ranking performance than the other tested non-personalized strategies. These results show that recommendation performance depends on the evaluation task and that rating prediction, Top-K ranking, and true cold-start recommendation should be assessed separately.

**Keywords:** neural collaborative filtering; cold-start recommendation; hybrid recommender systems; deep learning; collaborative filtering; sparse user–item interactions; e-commerce recommendation

Check for updates

Academic Editor: Christos Bouras

Received: 25 July 2026
Revised: 19 September 2026
Accepted: 21 September 2026
Published: 3 October 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Recommender systems have become an essential component of modern e-commerce platforms, enabling users to navigate increasingly large product collections and discover items that align with their interests and preferences. By analyzing historical interactions and behavioral patterns, recommendation algorithms support personalized decision-making, improve customer engagement, and enhance the overall efficiency of online shopping environments. Consequently, recommender systems have been widely adopted across digital marketplaces, media services, and content-delivery platforms to facilitate personalized user experiences [1–3].

Despite substantial advances in recommendation technologies, several challenges continue to limit their effectiveness in real-world applications. Among these challenges, data sparsity remains one of the most persistent problems. In large-scale e-commerce environments, users typically interact with only a small subset of available products, resulting in highly sparse user–item interaction matrices. This scarcity of observed interactions restricts the ability of recommendation models to learn accurate preference representations and often degrades recommendation quality [4,5]. Another major challenge is the cold-start problem, where recommendations must be generated for users or items with little or no historical information. In cold-start scenarios, conventional collaborative filtering techniques often fail to identify reliable user preference patterns because little or no historical interaction data are available. As a result, recommendation quality declines, reducing personalization and negatively affecting user satisfaction. Traditional recommendation methods, particularly Collaborative Filtering (CF) and Matrix Factorization (MF), have achieved considerable success by modeling latent relationships between users and items [6,7]. Nevertheless, these methods rely largely on linear interaction assumptions, which limit their ability to represent the complex and non-linear behavioral patterns commonly observed in modern recommendation environments.

Recent advances in deep learning have provided new opportunities to address these limitations. Neural Collaborative Filtering (NCF) has received considerable attention because it replaces conventional linear interaction functions with neural architectures that can capture more complex user–item relationships [8]. Several extensions have also been proposed, including graph-based, context-aware, and hybrid recommendation models. However, evaluating these models remains challenging in sparse recommendation settings. Rating prediction, Top-K ranking, and cold-start recommendation represent different tasks and should be evaluated separately. In addition, fair model comparison requires consistent data partitions, validation-based model selection, and evaluation protocols that reflect the intended recommendation task.

Based on these challenges, this study evaluates a hybrid recommendation framework for sparse e-commerce data. For users with interaction history, the framework uses a NeuMF-based model that combines Generalized Matrix Factorization (GMF) and a Multi-Layer Perceptron (MLP). Batch Normalization and Dropout are evaluated as additional components of this architecture. For completely unseen users, recommendations are generated using an item-popularity fallback without relying on a learned user embedding.

The framework is evaluated under three separate settings: rating prediction for users with interaction history, full-catalog Top-K ranking, and recommendation for completely unseen users. The evaluation also includes comparisons with graph-based recommenders, isolated ablation experiments, and an analysis of different training objectives.

The experiments are designed to evaluate each recommendation task separately. This avoids direct comparison between metrics that measure different aspects of recommendation performance and provides a clearer view of the strengths and limitations of each model.

The main contributions of this work are as follows:

A hybrid recommendation approach that separates personalized recommendation for warm users from item-based fallback recommendation for completely unseen users.

A reproducible evaluation protocol with fixed training, validation, warm-test, and cold-test partitions, allowing rating prediction, full-catalog Top-K ranking, and true cold-start recommendation to be evaluated separately.

A controlled comparison of baseline NeuMF, the proposed NeuMF variant, LightGCN, and NGCF, together with an isolated ablation analysis of the main NeuMF components.

An empirical analysis of different training objectives and statistical significance to distinguish improvements in rating prediction from improvements in recommendation ranking.

## 2. Related Work

Numerous recommendation approaches have been developed to improve personalization in e-commerce, ranging from traditional collaborative filtering and matrix factorization to neural and graph-based models. This section reviews the main approaches related to the present study, with particular attention to data sparsity, cold-start recommendation, and the evaluation of neural recommender systems. It also identifies the research gaps that motivate the experimental design used in this work.

### *2.1. Traditional Recommendation Approaches*

Early recommender systems primarily relied on Collaborative Filtering (CF) and Matrix Factorization (MF) techniques to model user preferences from historical interactions [2,5–7,9,10]. User-based and item-based CF methods estimate similarities among users or products [4,6,11], while MF techniques learn latent representations of users and items [7,9,10,12]. Although these methods have proven effective across many recommendation tasks [2,6], they generally rely on linear interaction patterns [8] and continue to face challenges related to data sparsity [4,13,14] and cold-start [14–19] scenarios in large-scale environments.

### *2.2. Deep Learning-Based Recommendation*

The increasing availability of large-scale interaction data has motivated the adoption of deep learning in recommender systems [13,20–26]. Neural Collaborative Filtering (NCF) represents one of the most influential developments in this area by replacing the linear interaction function of traditional MF with neural networks capable of learning complex non-linear relationships between users and items [7,8]. The NeuMF architecture further combines Generalized Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) components to improve representation learning and recommendation accuracy [8]. As a result, NCF-based models are now widely used as benchmark approaches in recommender-system research [8,13,27].

### *2.3. Advanced Recommendation Models*

Recent research has expanded neural recommendation frameworks by incorporating additional information sources and more sophisticated learning mechanisms [21,28–33]. Graph-based methods such as LightGCN and NGCF exploit higher-order user–item relationships through graph neural networks [34–37]. Although these approaches can improve recommendation performance, graph- and knowledge-based methods introduce additional modeling complexity compared with conventional collaborative-filtering approaches [29,30,38].

Hybrid recommendation frameworks have also been widely investigated [39,40]. By combining collaborative filtering with content-based methods, clustering techniques, or deep learning architectures, hybrid models aim to improve recommendation robustness and diversity [39,40].

As summarized in Table 1, recent recommender-system research has addressed cold-start recommendation and data sparsity using several approaches, including active learning, graph-based models, knowledge-aware methods, and hybrid frameworks. These methods differ in the information they require and in the way they handle users with limited or no interaction history. This diversity also highlights the importance of using clear evaluation protocols when comparing recommendation approaches, particularly when rating prediction, Top-K ranking, and true cold-start recommendation are considered.

**Table 1.** Recent Recommender-System Studies Related to Cold-Start, Explainability, and Robust Recommendation (2025–2026).

| Study | Method | Dataset | Cold-Start Handling | Explainability | Main Limitation |
|---|---|---|---|---|---|
| Rajesh & Kumar (2025) [27] | Comparative study of CF, NCF, DeepFM, and LightGCN | MovieLens and benchmark datasets [41] | Limited | No | Focuses on performance comparison rather than proposing a new recommendation framework. |
| Tiwary et al. (2025) [28] | Knowledge Graph + Reinforcement Learning Recommendation | Knowledge graph datasets | Partial | Yes | Increased computational complexity and dependency on external knowledge graphs. |
| Zou et al. (2026) [29] | Graph Structures with Semantic Dependency Learning | Cold-start recommendation benchmarks | Yes | Limited | More computationally expensive than conventional NCF models. |
| De Pessemier et al. (2025) [42] | Active Learning-based Recommendation | MovieLens datasets | Yes | No | Requires user interaction during onboarding to collect preference signals. |
| Rodríguez-Baena et al. (2025) [43] | BinRec Biclustering-based Collaborative Filtering | Sparse recommendation datasets | Yes | No | Performance depends on bicluster quality and may suffer in highly dynamic environments. |
| Jun & Li (2025) [44] | Neighborhood-Aware Attention Collaborative Filtering | Benchmark recommendation datasets | Partial | Limited | Primarily addresses sparsity while offering limited cold-start support. |
| Sivakumar et al. (2026) [45] | CRAFT (Attention + Federated Learning) | Cold-start recommendation benchmarks | Yes | No | Federated training introduces additional communication overhead. |
| Liu & Li (2026) [46] | MAPTRec Cross-Domain Recommendation | Multi-domain recommendation datasets | Yes | No | Transfer learning effectiveness depends on source-domain quality. |
| Pasupuleti & Satapathy (2026) [47] | PRISM-Rec Knowledge Graph Recommendation | Knowledge-enhanced datasets | Partial | Limited | Requires rich semantic and relational information. |
| Proposed Framework | NeuMF-based recommendation + item-popularity fallback | Amazon Electronics | Yes—Fallback only (non-personalized) | No | Does not use user or item side information for personalized cold-start recommendation |

### 2.4. Research Gap

Despite substantial progress in recommender-system research [2,4,13,20], several evaluation challenges remain important in sparse recommendation settings. First, rating prediction and Top-K ranking are different tasks, but their results are not always clearly separated when model performance is reported. This can make it difficult to determine whether an improvement in prediction error also leads to better recommendation ranking.

Second, user cold-start evaluation requires special attention. A completely unseen user has no learned representation in a collaborative filtering model. Therefore, evaluating such users through untrained or unavailable user embeddings does not represent a realistic cold-start setting. A separate recommendation strategy is required until sufficient interaction history becomes available.

Third, fair comparison between recommendation models requires consistent data partitions, validation-based model selection, and common ranking protocols. Architectural modifications should also be evaluated separately to determine the contribution of each component rather than changing several components at the same time.

As summarized in Table 2, this study evaluates warm-user rating prediction, full-catalog Top-K ranking, and true user cold-start recommendation as separate tasks. The evaluation uses fixed data partitions and compares NeuMF-based models with graph-based recommenders under a common ranking protocol. In addition, isolated ablation experiments, different training objectives, and statistical tests are used to examine the effects of the evaluated model components and training choices.

**Table 2.** Comparison of Existing Recommendation Approaches and Identification of Research Gaps Addressed by the Proposed Framework.

| Feature | NCF | LightGCN | PRISM-Rec | CRAFT | Proposed Framework |
|---|---|---|---|---|---|
| Neural CF | ✓ | ✗ | ✗ | ✓ | ✓ |
| Graph Learning | ✗ | ✓ | ✓ | ✗ | ✗ |
| Knowledge Graph | ✗ | ✗ | ✓ | ✗ | ✗ |
| Cold-Start Handling | Limited | Limited | Partial | ✓ | Non-personalized fallback only |
| Explainability | ✗ | ✗ | Partial | ✗ | ✗ |
| Ablation Study | Limited | Limited | Limited | Limited | ✓ |
| Statistical Testing | Rare | Rare | Rare | Limited | ✓ |
| Computational Simplicity | High | Medium | Low | Low | Not Evaluated |

Note: ✓ indicates that the feature is supported/present, while ✗ indicates that the feature is not supported/absent.

## 3. Methodology

This section presents the methodology used to develop and evaluate the recommendation framework. It first describes the data source, preprocessing procedure, and construction of the training, validation, warm-test, and cold-test sets. It then explains the NeuMF-based recommendation model, the cold-start fallback strategy, and the training

configuration. Finally, the experimental design and evaluation metrics used for rating prediction, full-catalog Top-K ranking, and true cold-start recommendation are presented.

### 3.1. Overview of the Proposed Framework

The proposed framework addresses recommendation under two user conditions: users with historical interactions and completely unseen users. For users represented in the training data, recommendations are generated using a NeuMF-based collaborative filtering model. For completely unseen users, collaborative filtering cannot provide a learned user representation; therefore, an item-based fallback strategy is used until sufficient interaction history becomes available.

For warm users, the NeuMF-based model combines Generalized Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) components to model user–item interactions. The proposed variant additionally evaluates Batch Normalization and Dropout within the MLP branch. For true cold-start users, recommendations are generated without using a learned user embedding and rely only on item information derived from the training data.

The overall methodology consists of data preprocessing, fixed data partitioning, model training and validation, warm-user evaluation, and true cold-start evaluation. Figure 1 summarizes the complete experimental framework.

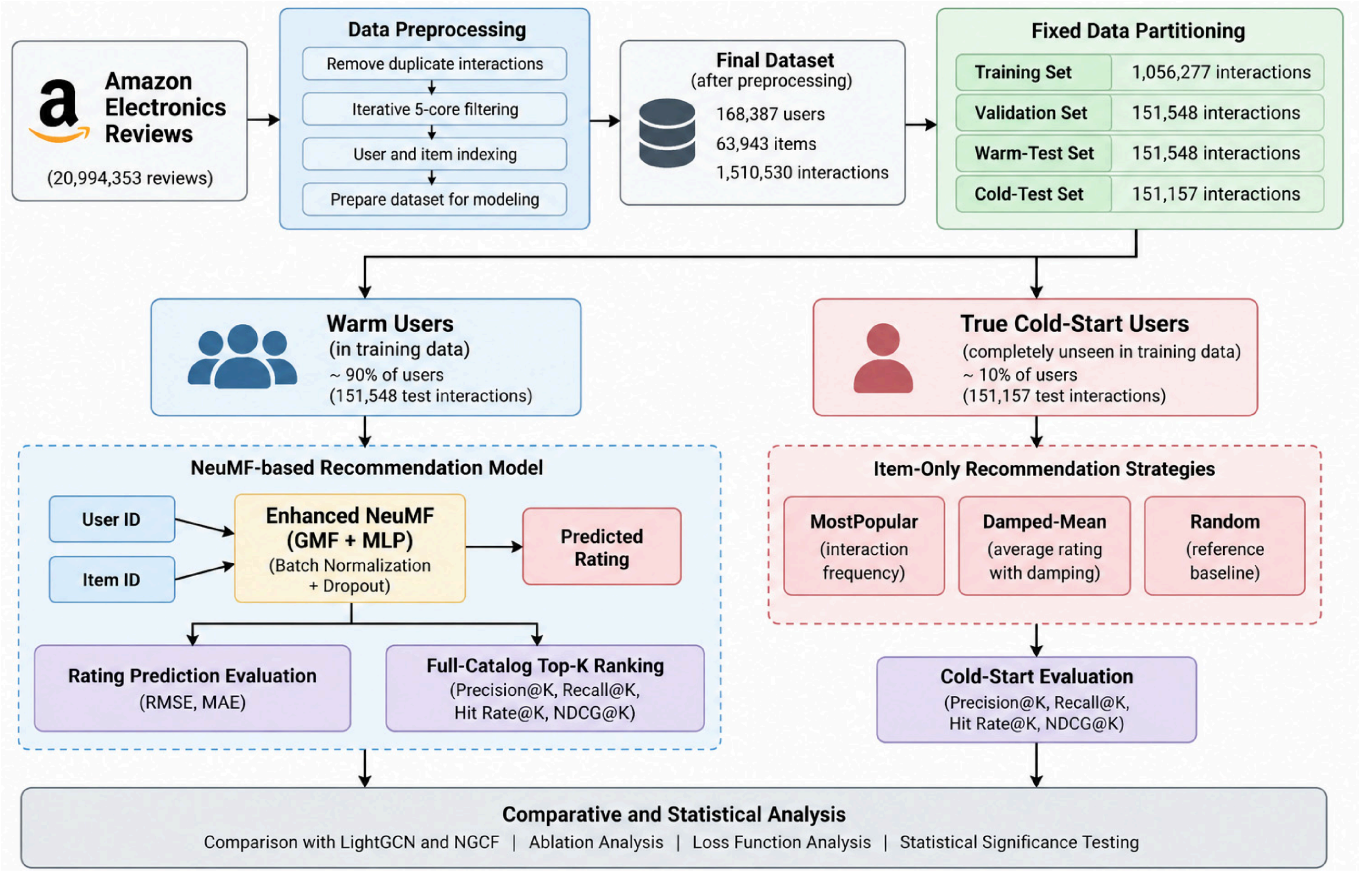



**Figure 1.** Overall Hybrid Framework.

### 3.2. Data Source and Description

The experiments were conducted using the Amazon Electronics review dataset as the primary data source [48]. The original dataset contained 20,994,353 review records. Duplicate user–item interactions were removed by retaining the most recent interaction for each user–item pair, resulting in 20,553,480 interactions.

An iterative 5-core filtering procedure was then applied to remove users and items with insufficient interaction history. After filtering, a seeded user subsample was constructed and the 5-core filtering procedure was applied again. The final experimental dataset contained 168,387 users, 63,943 items, and 1,510,530 interactions.

This processed dataset was fixed before the official model evaluation and was used consistently across the experiments reported in this study. The final dataset composition and the experimental data partitions are summarized in Table 3.

**Table 3.** Summarizes the final dataset and the experimental partitions used throughout the study.

| Data Component | Users/Cohort | Interactions | Purpose |
|---|---|---|---|
| Final Amazon Electronics Dataset | 168,387 users | 1,510,530 | Complete processed dataset |
| Training Set | Warm users | 1,056,277 | Model training |
| Validation Set | Warm users | 151,548 | Model selection and early stopping |
| Warm-Test Set | Warm users | 151,548 | Rating prediction and warm-user evaluation |
| Cold-Test Set | 16,839 cold users | 151,157 | True cold-start evaluation |
| Top-K Ranking Cohort | 10,000 warm users | — | Full-catalog ranking evaluation |

### 3.3. Data Preprocessing

After duplicate removal and iterative 5-core filtering, user and item identifiers were mapped to numerical indices suitable for embedding-based learning. Original rating values were retained for explicit rating-prediction experiments.

The processed data were divided into separate training, validation, warm-test, and cold-test partitions. Ten percent of the users were reserved as true cold-start users and were completely excluded from model training. The remaining users formed the warm-user population.

For warm users, interactions were ordered chronologically. Earlier interactions were used for training, while later interactions were reserved for validation and testing. The final partitions contained 1,056,277 training interactions, 151,548 validation interactions, 151,548 warm-test interactions, and 151,157 cold-test interactions. The validation set was used for model selection and early stopping, while the test sets remained separate until final evaluation.

For full-catalog Top-K evaluation, a fixed cohort of 10,000 warm users was selected before model testing. Items with ratings of 4 or 5 were considered relevant. Items already observed in the training and validation sets were removed from the candidate recommendations before ranking. All dataset partitions and evaluation cohorts were fixed before the official experiments.

### 3.4. Proposed Recommendation Model

This subsection describes the NeuMF-based recommendation model evaluated in this study. The architecture combines Generalized Matrix Factorization (GMF) and Multi-Layer Perceptron (MLP) branches to learn complementary user–item interaction representations. A baseline NeuMF model is used as the reference architecture, while the proposed variant incorporates Batch Normalization and Dropout within the MLP branch. Both models use

the same embedding dimensions, fusion mechanism, and bounded prediction layer to ensure a controlled comparison.

#### 3.4.1. Embedding Layer

Each user and item is represented by a dense latent embedding learned from the training interactions. Let uu and ii denote a user and an item, respectively. Separate user and item embeddings are used in the GMF and MLP branches, with an embedding dimension of k = 32. These representations are optimized jointly during model training.

These embeddings provide compact representations of user preferences and item characteristics and serve as inputs to the interaction-learning components of the model.

#### 3.4.2. GMF Branch

The first interaction-learning component employs Generalized Matrix Factorization (GMF) to model linear relationships between users and items. Given user embedding (p_u) and item embedding (q_i), the GMF representation is computed through element-wise multiplication:

$$e_{ui}^{GMF} = p_u \odot q_i \tag{1}$$

where ($\odot$) denotes element-wise multiplication.

This branch captures direct user–item interactions through element-wise multiplication of the corresponding latent representations.

#### 3.4.3. MLP Branch

To capture non-linear user–item interaction patterns, a Multi-Layer Perceptron (MLP) is employed alongside the GMF branch. The user and item embeddings of the MLP branch are concatenated and passed through three fully connected layers containing 64, 32, and 16 neurons, respectively.

$$z_{MLP} = f_{MLP}([p_u; q_i]) \tag{2}$$

where ($f_{MLP}$) represents the non-linear transformation learned by the network.

Rectified Linear Unit (ReLU) activation is applied after each hidden layer to enhance representational capacity and facilitate optimization.

#### 3.4.4. Regularization Mechanisms

Batch Normalization and Dropout are evaluated within the MLP branch of the proposed NeuMF variant. Batch Normalization is applied to stabilize the intermediate activations, while Dropout with a rate of $p = 0.2$ is used to reduce overfitting during training. These components are not included in the baseline NeuMF model, allowing their effects to be evaluated under the same experimental configuration.

#### 3.4.5. Fusion and Prediction Layer

The outputs of the GMF and MLP branches are combined through a fusion layer to exploit both linear and non-linear interaction information.

$$z_{fuse} = \left[ e_{ui}^{GMF}; z_{MLP} \right] \tag{3}$$

The 32-dimensional $^{\mathrm{GMF}}$ representation and the 16-dimensional MLP representation are concatenated and passed directly to the final prediction layer.

The predicted rating is computed as

$$\hat{y}ui = 4\sigma\left( w^T zfuse + b \right) + 1 \tag{4}$$

where σ (·) denotes the sigmoid activation function, w is the learned weight vector, and b is the bias term. The transformation constrains the predicted rating to the original rating range of 1–5. The same prediction layer is used by both the baseline and proposed NeuMF models.

### 3.5. Cold-Start Recommendation Strategy

A user is considered a true cold-start user when no interaction from that user is included in the training set. Consequently, no learned user embedding is available for these users, and the collaborative filtering model is not used to generate their recommendations.

Three item-only recommendation strategies were evaluated for true cold-start users: MostPopular, damped-mean popularity, and random recommendation. All item statistics were calculated using the training set only. MostPopular ranks items according to their interaction frequency in the training data. The damped-mean strategy combines each item's mean rating with the global mean using a fixed damping parameter of λ = 10, while random recommendation was included as a reference baseline.

The cold-start evaluation therefore measures the performance of recommendation strategies that do not depend on a learned user representation. Based on the comparative evaluation, the strongest item-only strategy is used as the fallback until sufficient user interaction history becomes available for personalized recommendation. The architecture of the proposed NeuMF variant, including the GMF and MLP branches, fusion layer, and prediction layer, is illustrated in Figure 2.

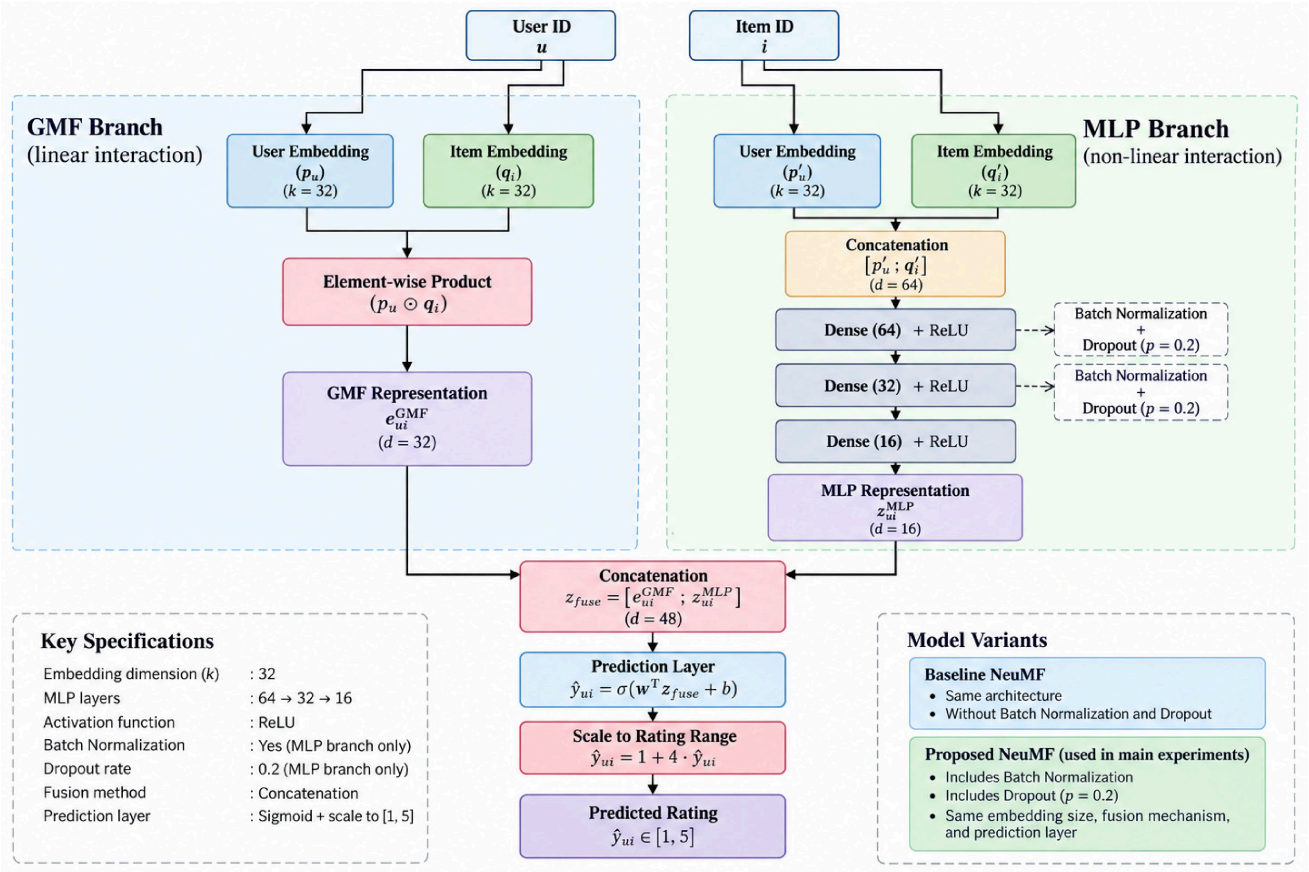



**Figure 2.** Architecture of the proposed NeuMF variant with Batch Normalization and Dropout.

### 3.6. Ensemble Variant for Ablation Analysis

An ensemble variant was included as part of the ablation analysis to examine whether averaging independently trained NeuMF models provides additional benefit. The ensemble does not represent an additional component of the main proposed architecture.

Predictions from the independently trained proposed NeuMF models were combined using arithmetic averaging:

$$\hat{y}ensemble = \frac{1}{M}\sum m = 1^{M}\widehat{y_m} \quad (5)$$

where M denotes the number of independently trained models and $\hat{y}m$ is the prediction produced by model m. No additional model training or test-based selection was performed for the ensemble.

### 3.7. Training Configuration

The NeuMF models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 2048. The maximum number of training epochs was set to 20. For the rating-prediction experiments, Mean Squared Error (MSE) was used as the primary optimization objective.

Early stopping was based exclusively on validation performance, with a patience of three epochs and a minimum improvement of $10^{-4}$. The model checkpoint with the best validation performance was restored before final test evaluation. The main NeuMF experiments were repeated using three random seeds: 42, 123, and 456.

Graph-based models were trained using their native pairwise ranking objective. Hyperparameter selection was performed using validation NDCG@10, and the test set was evaluated only after model selection was completed. This separation ensured that test results were not used for hyperparameter selection.

### 3.8. Experimental Design

The experimental evaluation was organized into six complementary parts:

True Cold-Start Evaluation: Evaluates item-only recommendation strategies for users completely excluded from the training set.

Warm-User Evaluation: Evaluates rating prediction and full-catalog Top-K ranking for users represented in the training data.

Graph-Based Benchmark Comparison: Compares the NeuMF models with LightGCN and NGCF under the same full-catalog ranking protocol.

Isolated Ablation Analysis: Examines the individual effects of the GMF and MLP branches, Batch Normalization, Dropout, and the ensemble variant.

Loss-Function and Class-Imbalance Analysis: Compares MSE, binary cross-entropy (BCE), and Bayesian Personalized Ranking (BPR) objectives and examines their effects under the observed rating distribution.

Statistical Analysis: Evaluates selected model differences using statistical significance tests, confidence intervals, and effect sizes.

### 3.9. Evaluation Metrics

Evaluation metrics were selected according to the recommendation task. Explicit rating prediction was evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). Full-catalog recommendation ranking was evaluated using Precision@K, Recall@K, Hit Rate@K, and Normalized Discounted Cumulative Gain (NDCG@K) at K = 5, 10, and 20.

For the pointwise binary-classification experiment using BCE, ratings of 4 or 5 were treated as positive interactions. Precision, Recall, F1-score, ROC-AUC, and PR-AUC were reported for this experiment separately from the Top-K ranking metrics.

Because rating prediction, pointwise binary classification, and full-catalog ranking represent different evaluation protocols, their metric values are interpreted only within their corresponding tasks and are not directly compared across protocols.

### 3.10. Metric Harmonization Note

The experiments in this study address distinct recommendation tasks and therefore use task-specific evaluation protocols. Rating prediction is evaluated using RMSE and MAE, full-catalog Top-K ranking is evaluated using ranking metrics at predefined values of K, and pointwise binary classification is evaluated separately using classification metrics. True cold-start recommendation is also evaluated independently using item-only strategies for users completely excluded from the training set. Consequently, absolute metric values should be interpreted within their corresponding evaluation protocol and should not be directly compared across different tasks.

## 4. Results and Discussion

This section presents the results of the revised experimental evaluation. The experiments examine three main recommendation settings: rating prediction for warm users, full-catalog Top-K ranking, and true cold-start recommendation for completely unseen users. Additional analyses compare the NeuMF-based models with graph-based recommenders, isolate the effects of individual architectural components, examine different training objectives, and evaluate the statistical significance and effect size of selected model differences.

### 4.1. Experimental Setup

All experiments were implemented in Python 3.12.5. The NeuMF-based models were implemented using TensorFlow/Keras, while LightGCN and NGCF were implemented using PyTorch 2.12.0 + cpu (CPU- only).

All experiments used the fixed Amazon Electronics dataset and the predefined data partitions described in Section 3. NeuMF models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 2048. Training was performed for a maximum of 20 epochs, with early stopping based only on validation performance. The best validation checkpoint was restored before final test evaluation.

The main NeuMF experiments were repeated using three random seeds (42, 123, and 456). Rating prediction and full-catalog ranking were evaluated separately. RMSE and MAE were used for rating prediction, while Precision@K, Recall@K, Hit Rate@K, and NDCG@K were used for ranking evaluation at K = 5, 10, and 20. Graph-based models were selected using validation NDCG@10 before evaluation on the fixed test cohort. The experimental settings and configurations used throughout the evaluation are summarized in Table 4.

**Table 4.** Summarizes the experimental settings used throughout the evaluation.

| Setting | Configuration |
|---|---|
| Dataset | Amazon Electronics |
| Final interactions | 1,510,530 |
| Training interactions | 1,056,277 |
| Validation interactions | 151,548 |
| Warm-test interactions | 151,548 |

**Table 4.** *Cont.*

| Setting | Configuration |
|---|---|
| Cold-test interactions | 151,157 |
| NeuMF embedding dimension | 32 |
| NeuMF MLP layers | $64 \rightarrow 32 \rightarrow 16$ |
| Proposed-model Dropout | 0.2 |
| Optimizer | Adam |
| Learning rate | 0.001 |
| Batch size | 2048 |
| Maximum epochs | 20 |
| Early-stopping patience | 3 |
| Minimum improvement | $10^{-4}$ |
| Random seeds | 42, 123, 456 |
| Ranking cutoffs | K = 5, 10, 20 |

### *4.2. Warm-User Recommendation Performance*

To evaluate recommendation performance for users represented in the training data, the baseline NeuMF model was compared with the proposed NeuMF variant. Rating prediction and full-catalog ranking were evaluated separately because they measure different aspects of recommendation performance. The warm-user performance results for the baseline and proposed NeuMF models are summarized in Table 5.

**Table 5.** Warm-User Performance Comparison Between Baseline and Proposed NeuMF Models.

| Metric | Baseline NeuMF | Proposed NeuMF |
|---|---|---|
| RMSE ↓ | 1.21721 ± 0.00062 | 1.23617 ± 0.00151 |
| MAE ↓ | 0.88374 ± 0.00146 | 0.86030 ± 0.00351 |
| Precision@5 ↑ | 0.00013 ± 0.00001 | 0.00028 ± 0.00009 |
| Recall@5 ↑ | 0.00067 ± 0.00007 | 0.00138 ± 0.00043 |
| NDCG@5 ↑ | 0.00039 ± 0.00005 | 0.00098 ± 0.00031 |
| Precision@10 ↑ | 0.00017 ± 0.00002 | 0.00017 ± 0.00004 |
| Recall@10 ↑ | 0.00167 ± 0.00019 | 0.00167 ± 0.00040 |
| NDCG@10 ↑ | 0.00072 ± 0.00003 | 0.00107 ± 0.00029 |
| Precision@20 ↑ | 0.00015 ± 0.00001 | 0.00009 ± 0.00003 |
| Recall@20 ↑ | 0.00297 ± 0.00026 | 0.00180 ± 0.00051 |
| NDCG@20 ↑ | 0.00105 ± 0.00007 | 0.00110 ± 0.00032 |

Note: ↓ indicates that lower values represent better performance, whereas ↑ indicates that higher values represent better performance.

The results show a different pattern for rating prediction and recommendation ranking. For rating prediction, the baseline NeuMF model achieved the lower RMSE (1.21721), while the proposed model achieved the lower MAE (0.86030 compared with 0.88374 for the baseline). This indicates that the proposed model reduced the average absolute prediction error, but it did not reduce larger squared errors measured by RMSE.

For full-catalog ranking, the proposed model achieved higher mean NDCG values at K = 5, 10, and 20. At K = 10, NDCG increased from 0.00072 for the baseline to 0.00107 for

the proposed model. However, Precision@10 and Recall@10 were approximately the same for both models. At K = 20, the baseline achieved higher Precision and Recall.

The absolute ranking values are low because the evaluation was performed over the full item catalog rather than a small sampled candidate set. Therefore, these results should be interpreted within the full-catalog ranking protocol. Statistical analysis presented later in this section also shows that the difference in NDCG@10 between the baseline and proposed models was not statistically significant.

As shown in Figure 3, the proposed NeuMF model achieved higher mean NDCG values at K = 5, 10, and 20. However, the difference at NDCG@10 was not statistically significant, as discussed in the statistical analysis.

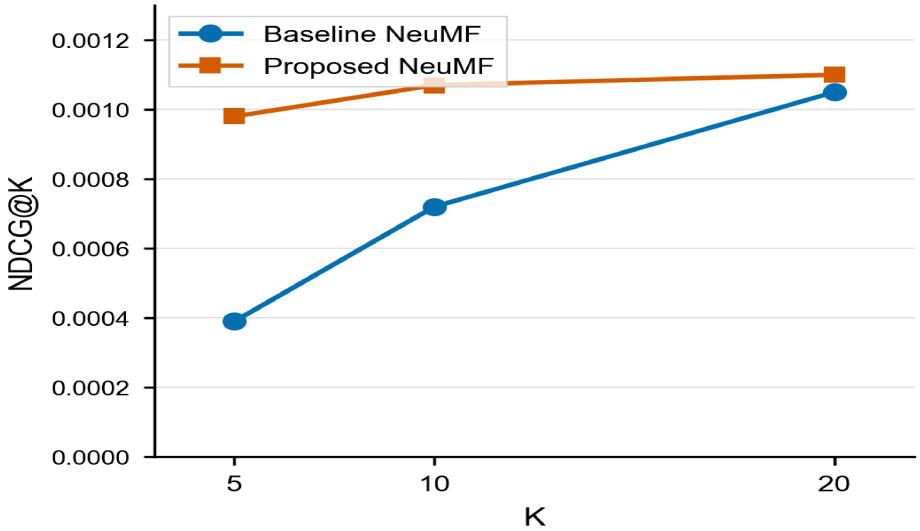



**Figure 3.** NDCG@K comparison between the baseline and proposed NeuMF models under full-catalog warm-user ranking.

### *4.3. Isolated Ablation Study*

To examine the effect of each architectural component separately, an isolated ablation study was conducted using the same data partitions and evaluation protocol. The evaluated variants included an MLP-only model, a GMF-only model, the baseline NeuMF model, NeuMF with Dropout only, NeuMF with Batch Normalization only, the full proposed NeuMF model with Batch Normalization and Dropout, and an ensemble variant based on the proposed model. The performance of all evaluated ablation variants is summarized in Table 6.

The ablation results show that the effect of each architectural component depends on the evaluation objective. For rating prediction, the MLP-only model achieved the lowest RMSE among the individually trained variants, while the proposed NeuMF model achieved the lowest MAE. The ensemble produced a further reduction in MAE, but this improvement did not extend to ranking performance.

Ranking results showed a different pattern. The GMF-only variant achieved the highest NDCG@10 (0.00671), clearly exceeding both the baseline NeuMF model (0.00072) and the proposed model (0.00107). In contrast, adding Batch Normalization alone reduced NDCG@10 to 0.00017. Dropout produced a small numerical increase over the baseline, but this difference was not statistically supported in the later statistical analysis.

As shown in Figure 4, the ensemble variant did not achieve better ranking performance. Figure 5 shows a different pattern for rating prediction. Among the individually trained variants, the MLP-only model achieved the lowest RMSE (1.21695), while the proposed NeuMF model achieved the lowest MAE (0.86030). The ensemble variant further reduced MAE to 0.85296.

**Table 6.** Results of the Isolated Ablation Study.

| Variant | Configuration | RMSE ↓ | MAE ↓ | NDCG@10 ↑ |
|---|---|---|---|---|
| A0 | MLP-only | 1.21695 ± 0.00072 | 0.88351 ± 0.00151 | 0.00061 ± 0.00003 |
| A1 | GMF-only | 1.30412 ± 0.00118 | 1.10103 ± 0.00186 | 0.00671 ± 0.00040 |
| A2 | Baseline NeuMF | 1.21721 ± 0.00062 | 0.88374 ± 0.00146 | 0.00072 ± 0.00003 |
| A3 | NeuMF + Dropout | 1.22275 ± 0.00112 | 0.86952 ± 0.00265 | 0.00080 ± 0.00013 |
| A4 | NeuMF + BatchNorm | 1.26991 ± 0.00409 | 0.88466 ± 0.01486 | 0.00017 ± 0.00008 |
| A5 | Proposed NeuMF | 1.23617 ± 0.00151 | 0.86030 ± 0.00351 | 0.00107 ± 0.00029 |
| A6 | Ensemble Variant | 1.21756 | 0.85296 | 0.00058 |

Values for A0–A5 are reported as mean ± standard deviation across three random seeds. A6 is a predetermined arithmetic ensemble and is therefore reported as a single result. ↓ indicates that lower values represent better performance, whereas ↑ indicates that higher values represent better performance.

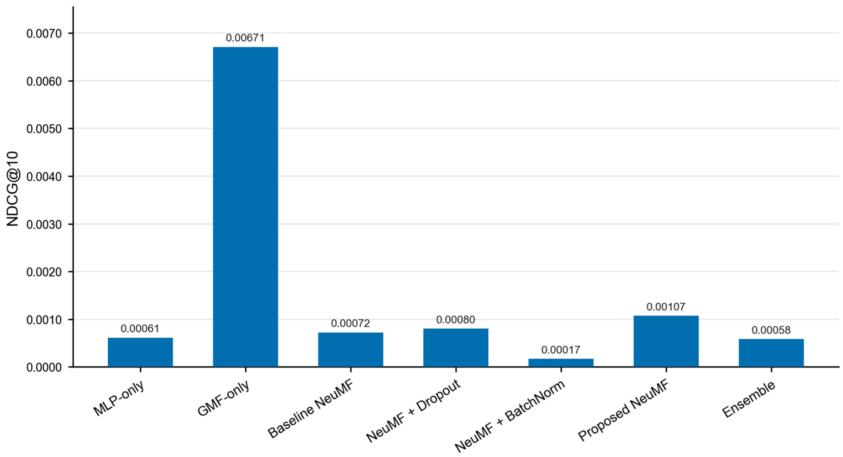



**Figure 4.** NDCG@10 performance of the model variants in the isolated ablation study.

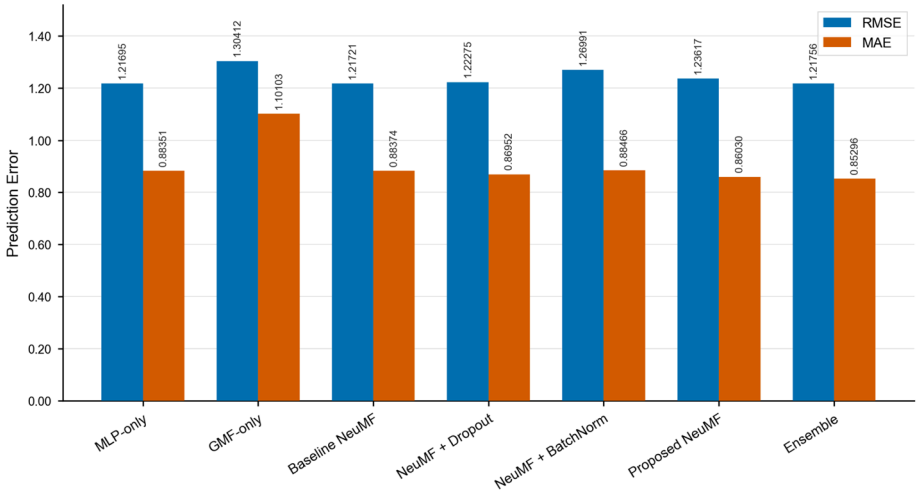



**Figure 5.** RMSE and MAE comparison of the model variants in the isolated ablation study.

The full proposed model therefore should not be interpreted as the best configuration across all evaluation tasks. Its main advantage was lower MAE, while the GMF-only model provided stronger full-catalog ranking performance. These results show that architectural changes that improve rating prediction do not necessarily improve Top-K recommendation quality.

### 4.4. Statistical Significance Analysis

Statistical analysis was conducted to determine whether the observed differences between the evaluated models were statistically meaningful. Paired Wilcoxon signed-rank tests were applied to per-user results, together with bootstrap confidence intervals and Holm correction for multiple comparisons. Effect sizes were also examined to distinguish statistical significance from practical importance. The results of the statistical significance analysis for the key model comparisons are summarized in Table 7.

**Table 7.** Statistical Analysis of Key Model Comparisons.

| Comparison | Metric | Mean Difference | 95% CI | Adjusted $p$-Value | Interpretation |
|---|---|---|---|---|---|
| Proposed vs. Baseline NeuMF | MAE | −0.02344 | [−0.02442, −0.02244] | <0.0001 | Significant, small effect |
| Proposed vs. Baseline NeuMF | NDCG@10 | +0.00035 | [−0.00017, 0.00085] | 0.3032 | Not significant |
| GMF-only vs. Baseline NeuMF | NDCG@10 | +0.00599 | — | <0.0001 | Significant |
| Dropout vs. Baseline NeuMF | NDCG@10 | +0.00008 | — | 0.8540 | Not significant |
| BatchNorm vs. Baseline NeuMF | NDCG@10 | −0.00056 | — | 0.0019 | Significant decrease |
| BPR vs. MSE | NDCG@10 | +0.00165 | — | <0.0001 | Significant |
| LightGCN vs. Proposed NeuMF | NDCG@10 | +0.00752 | [0.00568, 0.00942] | <0.0001 | Significant |

The proposed NeuMF model achieved a lower MAE than the baseline model, with a mean difference of −0.02344. This difference was statistically significant ($p < 0.0001$), although the effect size was small. This result supports a modest improvement in rating prediction rather than a large overall performance gain.

In contrast, the difference in NDCG@10 between the proposed and baseline NeuMF models was not statistically significant after Holm correction (adjusted $p = 0.3032$). Therefore, the higher mean NDCG@10 of the proposed model should be treated as a numerical difference rather than evidence of a statistically reliable ranking improvement.

The ablation results were also supported by the statistical analysis. The GMF-only model showed a significant improvement in NDCG@10 over the baseline NeuMF model, while the small improvement produced by Dropout alone was not statistically significant. Batch Normalization alone resulted in a small but statistically significant decrease in ranking performance.

The benchmark and loss-function comparisons showed larger ranking differences. BPR achieved significantly higher NDCG@10 than MSE, while LightGCN significantly outperformed the proposed NeuMF model in full-catalog ranking. These results further show that statistical conclusions depend on the recommendation task and evaluation metric.

### 4.5. Benchmark Comparison with Graph-Based Models

Graph-based recommendation models were included to compare the NeuMF-based models with ranking-oriented approaches. LightGCN and NGCF were evaluated using the same warm-user ranking cohort, candidate catalog, relevance definition, and masking procedure used for the NeuMF models. Their hyperparameters were selected using validation NDCG@10 before the final test evaluation.

Table 8 presents the full-catalog ranking results at K = 5, 10, and 20. Rating-prediction metrics such as RMSE and MAE are not reported for LightGCN and NGCF because these models were trained using their native pairwise ranking objective.

**Table 8.** Full-Catalog Ranking Comparison with Graph-Based Models.

| Model | Precision@5 | Recall@5 | NDCG@5 | Precision@10 | Recall@10 | NDCG@10 | Precision@20 | Recall@20 | NDCG@20 |
|---|---|---|---|---|---|---|---|---|---|
| Baseline NeuMF | 0.00013 | 0.00067 | 0.00039 | 0.00017 | 0.00167 | 0.00072 | 0.00015 | 0.00297 | 0.00105 |
| Proposed NeuMF | 0.00028 | 0.00138 | 0.00098 | 0.00017 | 0.00167 | 0.00107 | 0.00009 | 0.00180 | 0.00110 |
| LightGCN | 0.00183 | 0.00915 | 0.00764 | 0.00112 | 0.01115 | 0.00827 | 0.00085 | 0.01704 | 0.00975 |
| NGCF | 0.00118 | 0.00589 | 0.00463 | 0.00074 | 0.00739 | 0.00510 | 0.00055 | 0.01090 | 0.00598 |

LightGCN achieved the strongest full-catalog ranking performance across all reported values of K. At K = 10, LightGCN achieved an NDCG@10 of 0.00827, compared with 0.00510 for NGCF, 0.00107 for the proposed NeuMF model, and 0.00072 for the baseline NeuMF model.

The proposed NeuMF model achieved a higher mean NDCG than the baseline at K = 5, 10, and 20. However, the difference between the proposed and baseline models at NDCG@10 was not statistically significant, as reported in Section 4.4. In contrast, LightGCN significantly outperformed the proposed NeuMF model in full-catalog ranking.

These results show that the proposed NeuMF model is not the strongest model for Top-K ranking. Its main advantage is observed in rating prediction, particularly in MAE, whereas LightGCN provides stronger ranking performance. This difference further supports the need to evaluate rating prediction and recommendation ranking as separate tasks.

### *4.6. Cold-Start Evaluation*

The cold-start experiment evaluates recommendation performance for users who were completely excluded from the training set. These users have no learned user embeddings and are therefore not evaluated using the NeuMF model. Instead, three item-only recommendation strategies were compared: MostPopular, damped-mean popularity, and random recommendation. All item statistics used by these strategies were calculated from the training data only.

Of the 16,839 cold-start users, 16,747 were included in the ranking evaluation. The remaining users had no relevant in-catalog items under the evaluation protocol. The candidate catalog contained 63,629 items observed in the training data. The performance of the evaluated item-only recommendation strategies for true cold-start users is summarized in Table 9.

MostPopular achieved the best performance across all reported cold-start ranking metrics. At K = 10, it achieved an NDCG@10 of 0.01244 and a Hit Rate@10 of 0.07434, compared with 0.00130 and 0.00979, respectively, for the damped-mean strategy. Random recommendation produced the lowest performance.

Importantly, no learned user representation was used in this experiment. The results therefore represent a true user cold-start setting rather than collaborative filtering applied to held-out user identifiers. Based on these results, MostPopular was selected as the fallback strategy for completely unseen users until sufficient interaction history becomes available for personalized recommendation.

These findings do not indicate that the cold-start problem is fully resolved. Instead, they show that a simple item-popularity strategy can provide recommendations when personalized collaborative filtering cannot yet be applied.

**Table 9.** True Cold-Start Recommendation Results at K = 10.

| Strategy | Precision@10 | Recall@10 | Hit Rate@10 | NDCG@10 |
|---|---|---|---|---|
| MostPopular | 0.00908 | 0.01222 | 0.07434 | 0.01244 |
| Damped Mean ($\lambda = 10$) | 0.00116 | 0.00138 | 0.00979 | 0.00130 |
| Random | 0.00011 | 0.00014 | 0.00109 | 0.00013 |

### *4.7. Limitations*

Although the revised experiments provide a more controlled evaluation, several limitations should be considered. First, the official evaluation was conducted on the Amazon Electronics dataset only. Therefore, the results should not be interpreted as evidence of cross-dataset or cross-domain generalization. Evaluation on additional datasets using the same controlled protocol remains necessary.

Second, full-catalog ranking was performed over a large item catalog, resulting in low absolute ranking metrics. These values should be interpreted within this evaluation setting rather than compared directly with results obtained using sampled candidate sets.

Third, the cold-start strategy relies on item popularity and does not provide personalized recommendations for completely unseen users. It provides a temporary fallback until sufficient user interaction history becomes available. Future work may incorporate item content, contextual information, or other side information to improve personalization for new users.

From an engineering perspective, the experiments were conducted in an offline research environment and did not measure production-level factors such as inference latency, memory consumption, scalability under concurrent requests, or online user engagement. These aspects should be evaluated before deploying the framework in a large-scale e-commerce system.

Finally, the experiments show that model performance depends strongly on the target task and training objective. The proposed NeuMF variant improved MAE but did not provide the strongest Top-K ranking performance, while LightGCN achieved better ranking results. Further work is therefore needed to investigate architectures and ranking-oriented objectives that better balance rating prediction and recommendation ranking.

### *4.8. Loss-Function and Class-Imbalance Analysis*

To examine the effect of the training objective on recommendation performance, additional experiments were conducted using MSE, binary cross-entropy (BCE), and Bayesian Personalized Ranking (BPR). The analysis also considered the imbalance in the observed rating distribution.

In the training set, 869,527 interactions (82.32%) had ratings of 4 or 5 and were treated as positive interactions, while 186,750 interactions (17.68%) were below this threshold. This corresponds to a positive-to-negative ratio of approximately 4.66:1. The effect of the different training objectives on full-catalog ranking performance is summarized in Table 10.

**Table 10.** Effect of Training Objective on Full-Catalog Ranking.

| Training Objective | NDCG@10 |
|---|---|
| MSE | $0.00072 \pm 0.00003$ |
| BCE | $0.00034 \pm 0.00009$ |
| BPR | $0.00237 \pm 0.00025$ |

BPR achieved the highest NDCG@10 among the three evaluated training objectives, reaching 0.00237 compared with 0.00072 for MSE and 0.00034 for BCE. The improvement of BPR over MSE was statistically significant. This result indicates that a ranking-oriented objective is better aligned with the full-catalog ranking task than the rating-prediction objective used by the baseline NeuMF model.

The BCE experiment also illustrates the effect of class imbalance on pointwise classification metrics. At a decision threshold of 0.5, BCE achieved a Precision of 0.80187, Recall of 0.98359, and F1-score of 0.88348. The predicted-positive rate was 0.96972, showing that the model classified most interactions as positive. At a lower threshold of 0.3, all evaluated interactions were classified as positive, resulting in a Recall of 1.0. This demonstrates why very high or perfect Recall values can occur under an imbalanced pointwise-classification protocol and should not be interpreted as evidence of strong Top-K recommendation performance.

Overall, the results show that the choice of loss function should reflect the target recommendation task. MSE remains suitable for explicit rating prediction, whereas BPR produced stronger ranking performance in this experiment. BCE classification results were strongly influenced by the high proportion of positive ratings and by the selected decision threshold. The NDCG@10 performance across the different training objectives is illustrated in Figure 6.

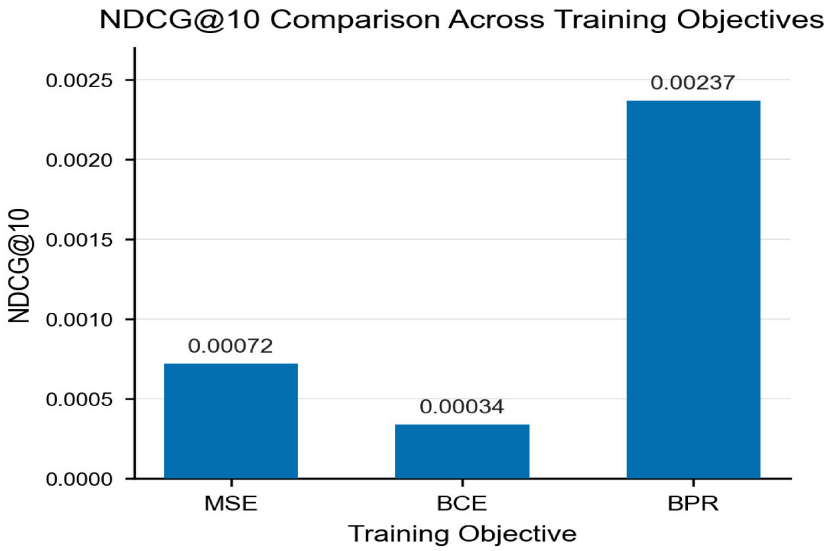



**Figure 6.** NDCG@10 comparison across MSE, BCE, and BPR training objectives under the full-catalog ranking protocol.

### *4.9. Discussion*

The revised experiments show that recommendation performance depends strongly on the task being evaluated. The proposed NeuMF model achieved a lower MAE than the baseline NeuMF model, indicating an improvement in average rating-prediction error. However, the baseline achieved a lower RMSE, showing that the proposed model did not improve all rating-prediction measures.

A similar distinction was observed in the ranking experiments. Although the proposed NeuMF model achieved a higher mean NDCG@10 than the baseline, this difference was not statistically significant. The ablation study also showed that the GMF-only variant achieved stronger full-catalog ranking performance than the full proposed model. Therefore, combining additional architectural components did not consistently improve ranking quality.

The ablation results also clarify the different behavior of the GMF and MLP components. In the evaluated sparse setting, the GMF-only variant achieved stronger full-catalog ranking performance than the more complex NeuMF configurations. One possible explanation is that GMF directly models multiplicative user–item latent interactions using a relatively simple representation, whereas the non-linear MLP branch introduces additional parameters and modeling flexibility. This additional complexity may support pointwise rating-prediction objectives without necessarily improving full-catalog Top-K ranking. Similarly, Batch Normalization alone significantly reduced NDCG@10 (adjusted $p = 0.0019$), while the numerical improvement produced by Dropout alone was not statistically significant (adjusted $p = 0.8540$). Therefore, these components should not be interpreted as general ranking improvements; their effects are task- and metric-dependent under the evaluated protocol.

The graph-based comparison provided further evidence of this task-dependent behavior. LightGCN achieved the strongest Top-K ranking performance among the evaluated models and significantly outperformed the proposed NeuMF model in NDCG@10. In addition, the loss-function analysis showed that BPR produced better ranking performance than MSE and BCE. Together, these results indicate that ranking-oriented models and objectives are better suited to the full-catalog recommendation task under the evaluated setting.

The true cold-start experiment addressed a different recommendation condition. Completely unseen users were not assigned learned user embeddings and were evaluated using item-only strategies. MostPopular achieved the best performance among the tested cold-start strategies and was therefore selected as the fallback method. This approach does not provide personalized recommendations for new users, but it allows recommendations to be generated until sufficient interaction history becomes available.

Overall, the findings do not support a single model as the best choice for every recommendation task. The proposed NeuMF variant provides an advantage in MAE, LightGCN performs better for full-catalog Top-K ranking, and MostPopular provides a simple fallback for completely unseen users. These results highlight the importance of matching the model, training objective, and evaluation protocol to the intended recommendation task.

## 5. Conclusions and Future Work

This section summarizes the main findings of the study and outlines directions for future research.

### *5.1. Conclusions*

This study evaluated a hybrid recommendation framework for sparse e-commerce environments under three separate recommendation settings: rating prediction for warm users, full-catalog Top-K ranking, and recommendation for completely unseen users. The framework uses a NeuMF-based model for users with historical interactions and an item-popularity fallback for true cold-start users.

For rating prediction, the proposed NeuMF variant achieved a lower MAE than the baseline NeuMF model (0.86030 vs. 0.88374), while the baseline achieved a lower RMSE (1.21721 vs. 1.23617). Statistical analysis showed that the improvement in MAE was significant, although its effect size was small. In full-catalog ranking, the proposed model achieved a higher mean NDCG@10 than the baseline, but this difference was not statistically significant.

The ablation analysis showed that the effects of the architectural components depend on the evaluation task. The GMF-only variant achieved the strongest NDCG@10 among the evaluated NeuMF variants, while the proposed model provided its main advantage in MAE.

The loss-function analysis also showed that BPR produced stronger ranking performance than MSE and BCE under the full-catalog ranking protocol.

The graph-based benchmark comparison showed that LightGCN achieved the strongest overall Top-K ranking performance among the evaluated models. This result indicates that improvements in rating prediction do not necessarily lead to improvements in recommendation ranking.

For completely unseen users, no learned user embedding was used. Instead, item-only recommendation strategies were evaluated, and MostPopular achieved the strongest performance among the tested cold-start methods. This strategy provides a simple fallback until sufficient user interaction history becomes available for personalized recommendation.

Overall, the results show that no single model or training objective performs best across all recommendation tasks. Rating prediction, full-catalog ranking, and true cold-start recommendation should therefore be evaluated separately using methods and metrics that match the intended recommendation setting.

### *5.2. Future Work*

Future research can extend this work in several directions. First, true cold-start recommendation can be improved by incorporating item content, product metadata, contextual information, or other side information rather than relying only on popularity-based fallback recommendations.

Second, future work can investigate ranking-oriented models and training objectives. The results of this study showed that LightGCN achieved stronger full-catalog ranking performance than the NeuMF-based models, while BPR outperformed MSE and BCE for NDCG@10. Further research can therefore explore architectures and objectives designed specifically for Top-K recommendation.

Third, the framework should be evaluated on additional datasets using the same controlled preprocessing, data-partitioning, model-selection, and evaluation procedures. This would allow future studies to examine whether the findings generalize across different recommendation domains and interaction densities.

Finally, future work can investigate richer recommendation evaluation in real-world environments, including online user behavior, click-through rate, user satisfaction, and other user-centered measures. Such evaluation would complement the offline metrics used in this study and provide a better understanding of the practical value of the recommendation strategies.

**Author Contributions:** Software, O.A.A.; Supervision, N.K., E.-S.M.E.M. and M.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article will be made available by the authors on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Li, X.; Karahanna, E. Online Recommendation Systems in a B2C E-Commerce Context: A Review and Future Directions. *J. Assoc. Inf. Syst.* **2015**, *16*, 1–42. [CrossRef]
2. Ricci, F.; Rokach, L.; Shapira, B. *Recommender Systems Handbook*, 3rd ed; Springer: Cham, Switzerland, 2022.

3. Bektemyssova, G.; Shyntore, G.; Akhmer, Y. Application of recommender systems for e-commerce. *Procedia Comput. Sci.* **2024**, *231*, 329–334. [CrossRef]
4. Aljunid, M.F.; Manjaiah, D.H.; Hooshmand, M.K.; Ali, W.A.; Shetty, A.M.; Alzoubah, S.Q. A collaborative filtering recommender systems survey. *Neurocomputing* **2024**, *594*, 128489. [CrossRef]
5. Adomavicius, G.; Tuzhilin, A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 734–749. [CrossRef]
6. Bobadilla, J.; Ortega, F.; Hernando, A.; Gutiérrez, A. Recommender systems survey. *Knowl.-Based Syst.* **2013**, *46*, 109–132. [CrossRef]
7. Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30–37. [CrossRef]
8. He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; Chua, T.-S. Neural Collaborative Filtering. In Proceedings of the International Conference on World Wide Web, Perth, Australia, 3–7 April 2017; pp. 173–182. [CrossRef]
9. Aggarwal, C.C. *Recommender Systems: The Textbook*; Springer: Cham, Switzerland, 2016.
10. Rendle, S. Factorization Machines. In Proceedings of the IEEE International Conference on Data Mining (ICDM), Sydney, Australia, 13–17 December 2010; pp. 995–1000. [CrossRef]
11. Isinkaye, F.O.; Folajimi, Y.O.; Ojokoh, B.A. Recommendation systems: Principles, methods and evaluation. *Egypt. Inform. J.* **2015**, *16*, 261–273. [CrossRef]
12. Park, H.; Jung, J.; Kang, U. A comparative study of matrix factorization and random walk with restart in recommender systems. In Proceedings of the IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; pp. 1024–1032. [CrossRef]
13. Gheewala, S.; Xu, S.; Yeom, S. In-depth survey: Deep learning in recommender systems—Exploring prediction and ranking models, datasets, feature analysis, and emerging trends. *Neural Comput. Appl.* **2025**, *37*, 10875–10947. [CrossRef]
14. Guan, J.; Chen, B.; Yu, S. A hybrid similarity model for mitigating the cold-start problem of collaborative filtering in sparse data. *Expert Syst. Appl.* **2024**, *249*, 123700. [CrossRef]
15. Schein, A.I.; Popescul, A.; Ungar, L.H.; Pennock, D.M. Methods and metrics for cold-start recommendations. In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Tampere, Finland, 11–15 August 2002; pp. 253–260. [CrossRef]
16. Kannout, E.; Grzegorowski, M.; Grodzki, M.; Nguyen, H.S. Clustering-based frequent pattern mining framework for solving cold-start problem in recommender systems. *IEEE Access* **2024**, *12*, 13678–13698. [CrossRef]
17. Liu, J.; Zhang, Z.; Cheung, R.C. MoToRec: Sparse-regularized multimodal tokenization for cold-start recommender. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 15324–15332. [CrossRef]
18. Rajapakse, D.; Leith, D. User cold-start learning in recommender systems using Monte Carlo tree search. *ACM Trans. Recomm. Syst.* **2025**, *3*, 1–23. [CrossRef]
19. Mishra, K.N.; Mishra, A.; Barwal, P.N.; Lal, R.K. Natural Language Processing and Machine Learning-Based Solution of Cold Start Problem Using Collaborative Filtering Approach. *Electronics* **2024**, *13*, 4331. [CrossRef]
20. Zhang, S.; Yao, L.; Sun, A.; Tay, Y. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* **2019**, *52*, 5. [CrossRef]
21. Guo, H.; Tang, R.; Ye, Y.; Li, Z.; He, X. DeepFM: A factorization-machine based neural network for CTR prediction. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Melbourne, Australia, 19–25 August 2017; pp. 1725–1731. [CrossRef] [PubMed]
22. Sedhain, S.; Menon, A.K.; Sanner, S.; Xie, L. AutoRec: Autoencoders meet collaborative filtering. In Proceedings of the International Conference on World Wide Web, Florence, Italy, 18–22 May 2015; pp. 111–112. [CrossRef]
23. Strub, F.; Gaudel, R.; Mary, J. Hybrid recommender system based on autoencoders. In Proceedings of the Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016; pp. 11–16. [CrossRef]
24. Xia, H.; Wang, Y. Enhancing Neural Collaborative Filtering for Product Recommendation by Integrating Sales Data and User Satisfaction. *Electronics* **2025**, *14*, 3165. [CrossRef]
25. Peng, S.; Siet, S.; Ilkhomjon, S.; Kim, D.-Y.; Park, D.-S. Integration of Deep Reinforcement Learning with Collaborative Filtering for Movie Recommendation Systems. *Appl. Sci.* **2024**, *14*, 1155. [CrossRef]
26. Padhy, N.; Suman, S.; Priyadarshini, T.S.; Mallick, S. A Recommendation System for E-Commerce Products Using Collaborative Filtering Approaches. *Eng. Proc.* **2024**, *67*, 50. [CrossRef]
27. Rajesh, D.B.; Kumar, A. Collaborative filtering models an experimental and detailed comparative study. *Sci. Rep.* **2025**, *15*, 31667. [CrossRef] [PubMed]
28. Tiwary, N.; Mohd Noah, S.A.; Fauzi, F. Prioritising explainable AI-driven recommendations with knowledge graphs and reinforcement learning. *J. King Saud Univ. Comput. Inf. Sci.* **2025**, *37*, 156. [CrossRef]
29. Zou, L.; Chen, Q.; Deng, Z.; Li, G. Enhancing user cold-start recommendation with graph structures and semantic dependencies. *Knowl.-Based Syst.* **2026**, *336*, 115280. [CrossRef]

30. Pasupuleti, M.; Satapathy, S.M. KGERA: Knowledge graph enhanced reasoning architecture for recommendation systems. *Sci. Rep.* **2026**, *16*, 13994. [CrossRef] [PubMed]
31. Mateos, P.; Bellogín, A. A systematic literature review of recent advances on context-aware recommender systems. *Artif. Intell. Rev.* **2024**, *58*, 20. [CrossRef]
32. Wu, L.; Zheng, Z.; Qiu, Z.; Wang, H.; Gu, H.; Shen, T.; Liu, Q.; Chen, E. A survey on large language models for recommendation. *World Wide Web* **2024**, *27*, 60. [CrossRef]
33. Ali, W.; Ammad-Ud-Din, M.; Zhou, X.; Zhang, Y.; Shao, J. Communication-efficient federated neural collaborative filtering with multi-armed bandits. *ACM Trans. Recomm. Syst.* **2025**, *4*, 1–28. [CrossRef]
34. Wang, X.; He, X.; Wang, M.; Feng, F.; Chua, T.-S. Neural graph collaborative filtering. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 165–174. [CrossRef]
35. He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; Wang, M. LightGCN: Simplifying and powering graph convolution network for recommendation. In Proceedings of the ACM SIGIR Conf. Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 639–648. [CrossRef]
36. Sharma, S.; Lee, Y.-C.; Nambi, S.; Salian, A.; Shah, S.; Kim, S.-W.; Kumar, S. A survey of graph neural networks for social recommender systems. *ACM Comput. Surv.* **2024**, *56*, 1–34. [CrossRef]
37. Lu, H.; Chen, Z. SocialJGCF: Social Recommendation with Jacobi Polynomial-Based Graph Collaborative Filtering. *Appl. Sci.* **2024**, *14*, 12070. [CrossRef]
38. Zangari, L.; Raso, S.; Tagarelli, A. Transformers for graph-based recommender systems: A survey. *Mach. Intell. Res.* **2026**, *23*, 34–69. [CrossRef]
39. Çano, E.; Morisio, M. Hybrid recommender systems: A systematic literature review. *Intell. Data Anal.* **2019**, *21*, 1487–1524.
40. Biswas, P.K.; Liu, S. A hybrid recommender system for recommending smartphones to prospective customers. *Expert Syst. Appl.* **2022**, *208*, 118058. [CrossRef]
41. Harper, F.M.; Konstan, J.A. The MovieLens datasets: History and context. *ACM Trans. Interact. Intell. Syst.* **2015**, *5*, 1–19. [CrossRef] [PubMed]
42. Pessemier, D.T.; Willems, B.; Martens, L. Active learning algorithm for alleviating the user cold start problem of recommender systems. *Sci. Rep.* **2025**, *15*, 24493. [CrossRef] [PubMed]
43. Rodríguez-Baena, D.; Gómez-Vela, F.; Lopez-Fernandez, A.; García-Torres, M.; Divina, F. BinRec: Addressing data sparsity and cold-start challenges in recommender systems with biclustering. *Appl. Intell.* **2025**, *55*, 830. [CrossRef]
44. Jun, J.; Li, Y. Collaborative filtering recommendation algorithm based on fine-grained mining and neighborhood awareness attention. *Discov. Artif. Intell.* **2025**, *5*, 156. [CrossRef]
45. Sivakumar, N.; John, R.S.; Bijo, A.; Suganeshwari, G.; Anbalagan, S.; Thandapani, S. CRAFT: Cold-start recommender with attention and federated training. *Sci. Rep.* **2026**, *16*, 17150. [CrossRef] [PubMed]
46. Liu, Q.; Li, Q. MAPTRec: A multi-view attention framework for heterogeneous cross-domain recommendation. *Soc. Netw. Anal. Min.* **2026**, *16*, 11. [CrossRef]
47. Pasupuleti, M.; Satapathy, S.M. PRISM-Rec: Progressive relational interest and semantic modeling for knowledge graph-enhanced recommendation. *J. King Saud Univ. Comput. Inf. Sci.* **2026**, *38*, 208. [CrossRef]
48. McAuley, J.; Targett, C.; Shi, Q.; van den Hengel, A. Image-based recommendations on styles and substitutes. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, 9–13 August 2015; pp. 43–52. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.