Hybrid Cold-Start Aware Movie Recommendation Using Collaborative and Semantic Representations

Author Name[1]
Computer Science / Kalinga Institute Of Industrial Technology Bhubaneswar
email@example.com

Abstract

Recommender systems commonly rely on collaborative filtering (CF) and content-based methods [9], each offering complementary advantages. Collaborative filtering models effectively capture collective user–item interaction patterns but suffer from popularity bias and complete failure in cold-start settings, where items lack sufficient interaction history [6, 2]. Content-based approaches mitigate cold-start issues by leveraging item metadata, yet often struggle to model personalized preferences at scale [7].

In this work, we present a hybrid movie recommendation framework that integrates LightGCN, a lightweight graph-based collaborative filtering model [5], with semantic item representations derived from a pretrained BERT masked language model [3]. LightGCN is trained on implicit user–item interactions to learn collaborative user and item embeddings, while BERT encodes rich semantic information from movie metadata, including plot summaries, genres, cast and crew descriptions, and user reviews [14]. These complementary signals are combined through multiple hybrid strategies, including static fusion, learned projection-based alignment, and content-aware cold-start handling [2].

Experiments conducted on the MovieLens-1M dataset [4] using a leakage-free per-user chronological split show that LightGCN achieves the strongest overall ranking accuracy for warm items, consistent with prior academic benchmarks [5]. Hybrid fusion significantly increases catalog coverage and long-tail exposure, with the projection-based hybrid retrieving substantially more tail items and nearly doubling catalog coverage compared to LightGCN, albeit with a moderate trade-off in ranking accuracy [1].

For cold-start items, collaborative and hybrid fusion models achieve zero recall, confirming the inherent limitations of interaction-only learning [10]. To address this limitation, we introduce a two-stage semantic fallback strategy that enables non-zero cold-start retrieval and surfaces cold items in approximately 25% of all recommendations, with recall improving as the recommendation list size increases [2]. An exploratory user–content matching model further improves cold-start recall, although gains remain constrained by sparse textual descriptions [12].

Overall, our results demonstrate that hybrid recommender systems are most effective when collaborative filtering is preserved for accuracy, while semantic content is selectively leveraged to enhance long-tail exposure and cold-start robustness [14]. This study provides an empirically grounded analysis of hybrid recommendation trade-offs and reflects design principles widely adopted in real-world recommender systems.

1 Introduction

Personalized recommender systems play a critical role in modern digital platforms by helping users navigate large item catalogs in domains such as movies, books, music, and e-commerce [9]. By learning user preferences and item relevance, recommender systems reduce information overload and enhance user engagement. Among existing approaches, collaborative filtering (CF) and content-based recommendation remain the two dominant paradigms, each exhibiting complementary strengths and weaknesses [2].

Collaborative filtering models infer user preferences from historical interaction data by exploiting collective behavior patterns across users and items. Latent factor models and neural collaborative filtering approaches have achieved strong empirical performance by capturing implicit co-consumption relationships [6]. However, CF methods fundamentally rely on sufficient interaction history. As a result, they suffer from data sparsity, exhibit popularity bias, and fail in cold-start scenarios, where new or rarely interacted items lack adequate training signals [1, 10].

Content-based recommender systems address this limitation by leveraging item attributes such as textual descriptions, genres, and user reviews [7]. By modeling semantic similarity between items, content-based methods can recommend items even in the absence of interaction data, making them well-suited for cold-start and long-tail items. Despite this advantage, content-based approaches often ignore collaborative signals, leading to recommendations that lack personalization at scale and may over-specialize around narrow content similarity [2].

To overcome these limitations, hybrid recommender systems seek to integrate collaborative and content-based signals [2]. Recent advances in graph neural networks (GNNs) have led to state-of-the-art collaborative filtering models that explicitly model high-order user–item relationships. In particular, LightGCN simplifies earlier graph-based CF models by retaining only neighborhood aggregation over the user–item interaction graph, achieving strong performance with improved efficiency [5]. In parallel, pretrained language models such as BERT have demonstrated a remarkable capability to learn rich semantic representations from unstructured text, making them effective tools for encoding item metadata such as plot summaries, cast information, and user reviews [3, 14].

Motivated by these developments, we propose a hybrid cold-start recommendation framework that integrates LightGCN-based collaborative filtering with BERT-based semantic content representations for movies. Each item is represented by both a collaborative embedding learned from user–item interactions and a semantic embedding extracted from enriched textual metadata using a pretrained BERT model. These representations are fused at the item level using a weighted combination, allowing the system to balance collaborative preference learning with content-based generalization [2].

The proposed framework is evaluated on the MovieLens-1M dataset [4], augmented with enriched movie metadata including plot summaries, genres, cast and crew descriptions, and user reviews. A leakage-free per-user chronological train–test split is employed to ensure realistic evaluation. Extensive experiments demonstrate that while LightGCN achieves strong performance on popular items, it exhibits severe popularity bias and near-zero recall for long-tail and cold-start items [1]. Hybrid fusion substantially increases catalog coverage and long-tail exposure, while a content-aware cold-start strategy enables non-zero retrieval for items with little or no interaction history [10].

The main contributions of this work are summarized as follows:

- We design a hybrid recommendation architecture that integrates LightGCN for collaborative filtering and BERT for semantic item representation.
- We construct rich textual representations for movies by combining plot summaries, metadata, and user reviews, enabling content-based modeling of cold-start items.
- We empirically analyze multiple hybrid fusion strategies and demonstrate their impact on accuracy, diversity, and long-tail exposure.
- We present a comprehensive evaluation on MovieLens-1M, highlighting the trade-offs between collaborative accuracy and content-driven generalization.

Overall, this study provides an empirically grounded analysis of hybrid recommender systems, showing that effective recommendation in sparse and cold-start settings requires careful integration of collaborative structure and semantic content rather than reliance on either signal alone [14].

2 Related Work

Recommendation algorithms have progressed through several phases, each addressing the limitations of earlier approaches. Historically, techniques primarily fell under the categories of content-based filtering and collaborative filtering [9]. Content-based methods analyze item features—such as genres, keywords, or textual descriptions—to recommend items similar to those a user has previously consumed [7], while collaborative filtering relies on historical user–item interaction patterns to infer preferences [**?**]. Early collaborative filtering algorithms, particularly neighborhood-based methods, demonstrated the effectiveness of exploiting similarities among users and items [**?**]. However, these approaches tend to perform poorly when the interaction matrix is sparse and do not scale well to large datasets, limiting their practical applicability [6].

A major breakthrough occurred with the introduction of matrix factorization (MF) techniques, which gained prominence during the Netflix Prize competition [6]. By projecting users and items into a shared low-dimensional latent space, methods such as Singular Value Decomposition (SVD) and SVD++ captured underlying preference patterns that were not directly observable in raw rating data [**?**]. Extensions incorporating bias terms, regularization, and temporal dynamics further improved predictive accuracy [**?**]. Despite these advances, MF-based models continue to struggle in cold-start scenarios, where new users or items lack sufficient interaction history, resulting in degraded recommendation quality [10].

Subsequent research shifted toward graph-based collaborative filtering models, which represent users and items as nodes in a bipartite interaction graph [13]. Approaches such as Neural Graph Collaborative Filtering (NGCF) employ graph convolutional networks to propagate information across higher-order neighborhoods, enabling the modeling of complex user–item relationships [13]. LightGCN further simplifies this framework by retaining only neighborhood aggregation while removing nonlinear transformations and feature projections, achieving competitive recommendation performance with significantly reduced computational complexity [5].

In parallel, advances in natural language processing (NLP) have enabled recommender systems to leverage unstructured textual information such as item descriptions, metadata, and user reviews [**?**]. Transformer-based models, particularly BERT, have demonstrated strong capability in learning rich semantic representations through large-scale pretraining with masked language modeling objectives [3]. These representations capture nuanced semantic relationships between items and have been widely adopted for content-aware recommendation and cold-start modeling [14].

Despite these advancements, most existing methods, when applied in isolation, remain ineffective in settings characterized by data sparsity, long-tail item distributions, or cold-start users and items [1]. Collaborative models struggle with infrequently interacted items, while content-based approaches often fail to incorporate collective user behavior, leading to limited personalization [2]. This motivates the need for hybrid recommendation frameworks that combine the structural modeling strengths of graph-based collaborative filtering with the semantic richness of text-based embeddings [2]. By integrating LightGCN with BERT-derived content representations, the present work aims to improve recommendation robustness across head, long-tail, and cold-start items, addressing key limitations of prior approaches [14].

3 Dataset

This study uses a customized version of the publicly available MovieLens-1M dataset [4]. The original corpus primarily provides user–item rating data and basic item attributes such as titles and genres. In its standard form, MovieLens-1M consists of three relational tables: (i) a *movies* table containing movie identifiers, titles, release years, and genre tags; (ii) a *users* table containing user identifiers and demographic attributes such as age, gender, occupation, and ZIP code; and (iii) a *ratings* table comprising user identifiers, movie identifiers, explicit ratings on a 1–5 scale, and interaction timestamps [4].

To enable richer content-based and hybrid recommendation, we enhanced the MovieLens-1M dataset using The Movie Database (TMDB) API [**?**]. For each movie, we augmented the dataset with the following textual and descriptive fields: *plot_summary*, *tagline*, *cast_bios*, *crew_bios*, and aggregated *user reviews*. These additions provide detailed semantic information beyond numerical ratings and have been widely used in prior work on content-aware recommendation [**?**, 14]. Where any of these fields were unavailable for a given title, null strings were inserted to maintain a consistent schema.

After enrichment, the dataset comprises approximately 3,900 movies, 6,000 users, and one million user–movie interactions. Each movie record now includes its identifier, title, genre list, and the TMDB-derived textual fields, providing both structured metadata and unstructured text for each item. The resulting scale and density closely match standard MovieLens-1M experimental settings reported in the literature [6, 5].

Before modeling, all sources were cleaned, merged, and transformed into text corpora suitable for embedding-based learning. For each movie, its plot summary, tagline, cast and crew information, and review text were concatenated into a single document, following common practices in text-enhanced recommendation systems [**?**]. Two types of BERT-style semantic embeddings were derived from this corpus. First, Masked Language Modeling (MLM) embeddings were extracted from the concatenated text using a pretrained BERT encoder [3], resulting in an embedding matrix of approximately 3706×768. Second, Next Sentence Prediction (NSP) embeddings were generated by pairing each plot summary with its associated reviews, yielding an additional 3706×768 matrix, consistent with BERT's pretraining objectives [3]. To reduce storage and computational requirements, both embedding matrices were projected to 64 dimensions using Principal Component Analysis (PCA), a standard dimensionality reduction technique for neural embeddings [**?**].

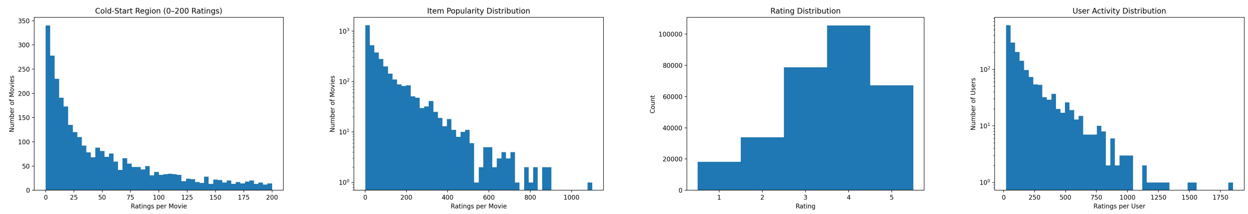
This enriched version of MovieLens-1M, created through TMDB-driven augmentation and systematic preprocessing, provides detailed textual descriptions for each movie and forms a strong foundation for evaluating content-aware and hybrid recommender models. MovieLens-1M is a widely adopted benchmark in recommender systems research [4]; when combined with TMDB-derived textual content, it offers both large-scale interaction data and rich semantic information, making it particularly suitable for evaluating transformer-based and hybrid recommendation approaches [14].

4 Preprocessing

The MovieLens-1M dataset contains approximately one million explicit ratings on a 1–5 scale. As commonly adopted in modern recommender systems, we reformulate the task as an implicit feedback ranking problem, where observed interactions indicate positive user preference. This formulation is more robust to user-specific rating biases and better aligned with top-K recommendation objectives.

To ensure a realistic and leakage-free evaluation, we adopt a per-user chronological train–test split. For users with more than five interactions, the earliest 80% of interactions are used for training, while the most recent 20% are held out for testing. Users with five or fewer interactions retain all interactions in the training set. This protocol preserves temporal ordering and prevents future information leakage.

To better understand the structural properties of the dataset and motivate the need for hybrid and cold-start-aware modeling, we analyze the distributions of item popularity, rating frequencies, and user activity. These analyses reveal strong sparsity, heavy-tailed interaction patterns, and pronounced cold-start regions, which are well-known challenges for collaborative filtering models.



(a) Cold-start region (0–200 ratings).

(b) Item popularity distribution (log-scale).

(c) Rating value distribution.

(d) User activity distribution.

Figure 1: Exploratory analysis of sparsity, popularity bias, rating behavior, and user activity in the MovieLens-1M dataset.

These observations motivate the use of graph-based collaborative filtering for dense interaction regions, hybrid semantic modeling to improve long-tail exposure, and content-based fallback mechanisms to enable recommendation of cold-start items lacking sufficient interaction history.

5 Methodology

This section describes the proposed hybrid recommendation framework that integrates graph-based collaborative filtering with semantic content representations to address accuracy, long-tail exposure, and cold-start challenges. The system is designed to reflect realistic industrial recommender pipelines, combining strong collaborative ranking with content-aware fallback mechanisms.

5.1 Problem Setting

Let U denote the set of users and I denote the set of items (movies). Given a set of implicit interactions $E \subseteq U \times I$, the objective is to learn a scoring function

$$\hat{y}_{ui} = \mathbf{e}_u^\top \mathbf{e}_i,$$

which ranks unseen items $i \in I \setminus I_u$ for each user u, where $\mathbf{e}_u$ and $\mathbf{e}_i$ represent user and item embeddings, respectively.

The primary challenges addressed in this work are:

- popularity bias in collaborative filtering,
- poor long-tail exposure,
- cold-start item recommendation.

5.2 Collaborative Representation Learning with LightGCN

We adopt LightGCN, a simplified graph-based collaborative filtering model that retains neighborhood aggregation while removing nonlinear activations and feature transformations [5].

5.2.1 Interaction Graph Construction

We construct a bipartite user–item graph $G = (U \cup I, E)$, where edges represent implicit user–item interactions. All explicit ratings are converted to binary feedback.

The adjacency matrix A is symmetrically normalized as

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}},$$

where D denotes the diagonal degree matrix.

5.2.2 LightGCN Propagation

At propagation layer $k+1$, user and item embeddings are updated as:

$$\mathbf{e}_u^{(k+1)} = \sum_{i \in \mathcal{N}(u)} \frac{1}{\sqrt{|\mathcal{N}(u)||\mathcal{N}(i)|}} \mathbf{e}_i^{(k)},$$

$$\mathbf{e}_i^{(k+1)} = \sum_{u \in \mathcal{N}(i)} \frac{1}{\sqrt{|\mathcal{N}(i)||\mathcal{N}(u)|}} \mathbf{e}_u^{(k)}.$$

Final embeddings are obtained by layer-wise averaging:

$$\mathbf{e}_u = \frac{1}{K+1} \sum_{k=0}^{K} \mathbf{e}_u^{(k)}, \quad \mathbf{e}_i = \frac{1}{K+1} \sum_{k=0}^{K} \mathbf{e}_i^{(k)}.$$

5.2.3 Training Objective

The model is optimized using Bayesian Personalized Ranking (BPR) loss [8]:

$$\mathcal{L}_{\text{BPR}} = -\mathbb{E}_{(u,i,j)} \left[\log \sigma \left(\hat{y}_{ui} - \hat{y}_{uj} \right) \right] + \lambda \|\Theta\|_2^2.$$

5.3 Semantic Item Representation using BERT

To incorporate semantic content, we generate item embeddings using a pretrained BERT model [3].

5.3.1 Text Construction

For each movie, a structured document is constructed by concatenating metadata:

```
Title, Genres, Plot Summary, Tagline,
Cast Bios, Crew Bios, User Reviews
```

Missing fields are replaced with empty strings.

5.3.2 Embedding Extraction

Each document is passed through BERT, and the representation corresponding to the `[CLS]` token is extracted:

$$\mathbf{e}_i^{\text{BERT}} = \text{BERT}_{\text{CLS}}(\text{text}_i).$$

5.4 Hybrid Recommendation Models

5.4.1 Linear Fusion

$$\mathbf{e}_i^{\text{hybrid}} = \alpha \mathbf{e}_i^{\text{GCN}} + (1-\alpha) \mathbf{e}_i^{\text{BERT}},$$

where $\alpha \in [0, 1]$.

5.4.2 Projection-Based Hybrid

$$\mathbf{e}_i^{\text{hybrid}} = \mathbf{e}_i^{\text{GCN}} + W\mathbf{e}_i^{\text{BERT}},$$

where $W \in \mathbb{R}^{d \times d}$ is trainable.

5.5 Cold-Start Recommendation

5.5.1 Two-Stage Strategy

1. Collaborative ranking using LightGCN.
2. Semantic fallback using cosine similarity in BERT space.

5.6 Recommendation Scoring

$$\hat{y}_{ui} = \mathbf{e}_u^\top \mathbf{e}_i.$$

5.7 Evaluation Protocol

- Dataset: MovieLens-1M
- Split: Per-user chronological 80/20
- Metrics: Precision@K, Recall@K, NDCG@K, HitRate@K

6 Results

This section presents a comprehensive empirical evaluation of collaborative, hybrid, and content-aware recommendation approaches on the MovieLens-1M dataset. Performance is analyzed across four complementary dimensions: (i) overall ranking accuracy on warm items, (ii) catalog coverage and accuracy–diversity trade-offs, (iii) long-tail recommendation effectiveness, and (iv) cold-start item recommendation.

All models are evaluated under an implicit feedback setting using standard Top-K ranking metrics. Training interactions are strictly excluded during evaluation via a per-user chronological split, ensuring a leakage-free assessment of generalization performance, consistent with established recommender system evaluation protocols [8, 5].

6.1 Overall Recommendation Performance (Warm Items)

We first evaluate recommendation quality on warm items, defined as those with sufficient interaction history in the training set. This setting represents the dominant operating regime of mature recommender systems.

6.1.1 LightGCN (Collaborative Baseline)

LightGCN serves as the collaborative filtering baseline and achieves the strongest overall ranking accuracy across all evaluated metrics.

Table 1: LightGCN performance on warm items

Metric	$K = 10$	$K = 20$
Precision@K	0.1518	0.1371
Recall@K	0.0681	0.1186
NDCG@K	0.1648	0.1692
HitRate@K	0.6525	0.7972
Catalog Coverage	0.2207	0.2963

These results are consistent with previously reported LightGCN performance on MovieLens benchmarks [5]. While LightGCN achieves strong ranking accuracy, its relatively low catalog coverage reflects pronounced popularity bias, a well-documented limitation of collaborative filtering models [1].

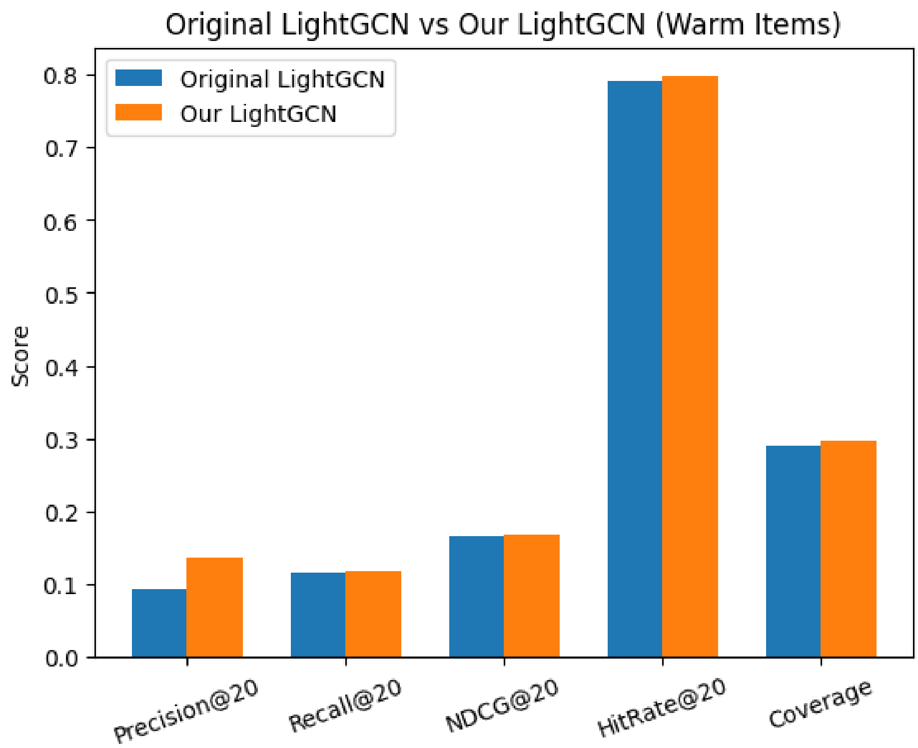


Figure 2: Comparison between original LightGCN results and our implementation on warm items.

Figure 2 shows that our implementation closely matches reported LightGCN performance across all metrics, validating the correctness of our training pipeline and evaluation setup.

6.2 Hybrid Models and Accuracy–Diversity Trade-off

6.2.1 Static Hybrid Fusion

Static hybrid fusion combines collaborative and semantic item embeddings through weighted summation after normalization.

Table 2: Static hybrid fusion performance on warm items ($K = 20$)

Metric	Value
Recall@20	0.0788
NDCG@20	0.1013
HitRate@20	0.6498
Catalog Coverage	0.7598

Compared to LightGCN, static fusion increases catalog coverage by more than $2.5\times$, indicating substantially improved item exposure. However, this increase comes at the cost of reduced ranking accuracy, illustrating the classic accuracy–diversity trade-off observed in hybrid recommenders [13].

6.2.2 Projection-Based Hybrid

To better integrate semantic information, a projection matrix is trained to align BERT item embeddings with the LightGCN collaborative space.

Table 3: Projection-based hybrid performance on warm items ($K = 20$)

Metric	Value
Precision@20	0.0953
Recall@20	0.0866
NDCG@20	0.1142
HitRate@20	0.6955
Catalog Coverage	0.4827

This approach recovers a significant portion of the accuracy lost under static fusion while maintaining substantially higher catalog coverage than LightGCN.

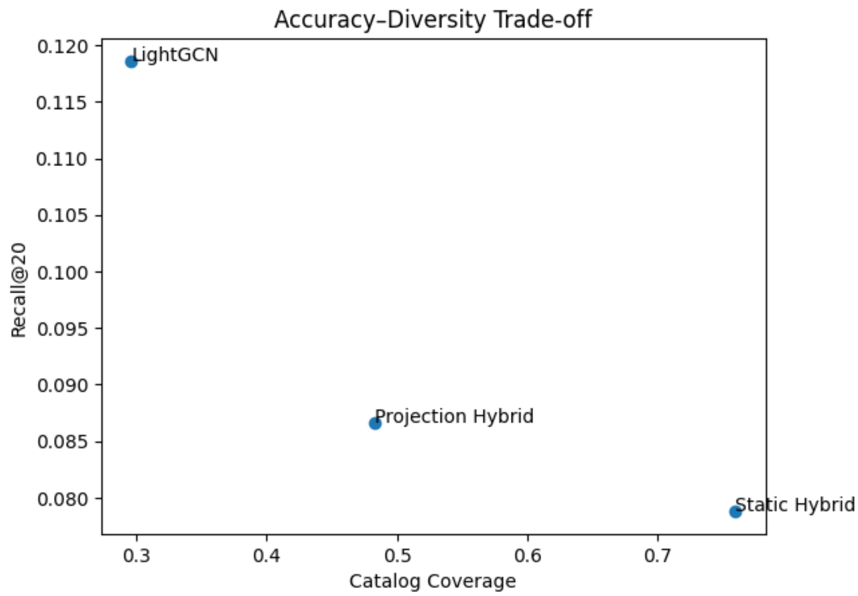


Figure 3: Accuracy–diversity trade-off across collaborative and hybrid models.

Figure 3 visualizes the trade-off between Recall@20 and catalog coverage. LightGCN achieves the highest recall but lowest coverage, static fusion maximizes coverage at the expense of accuracy, and the projection-based hybrid occupies a favorable middle ground.

6.3 Long-Tail Recommendation Performance

We evaluate long-tail effectiveness by restricting evaluation to tail items, defined as the bottom 80% of items ranked by interaction frequency.

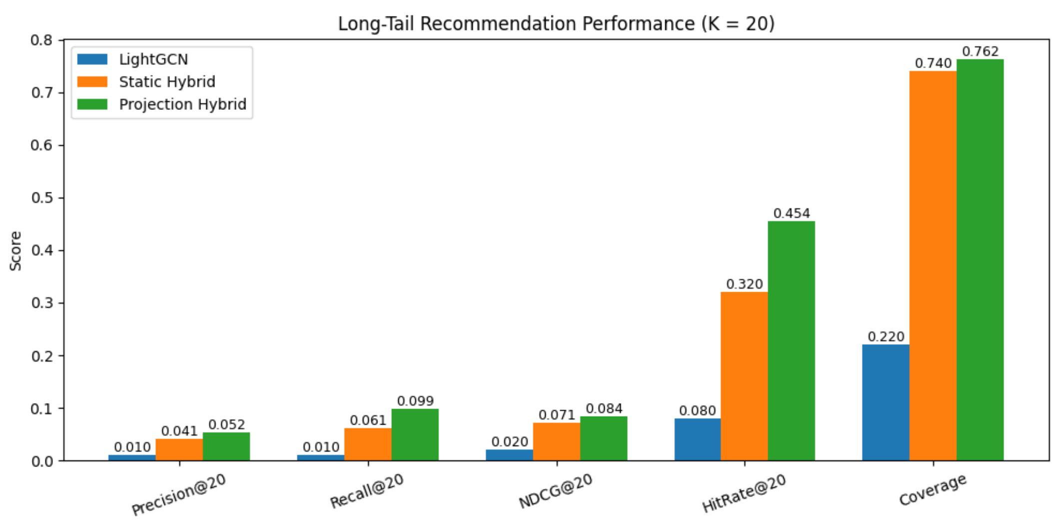


Figure 4: Long-tail recommendation performance comparison at $K = 20$.

Table 4: Projection-based hybrid performance on tail items ($K = 20$)

Metric	Value
Precision@20	0.0525
Recall@20	0.0985
NDCG@20	0.0842
HitRate@20	0.4541
Catalog Coverage	0.7623

As shown in Figure 4, the projection-based hybrid significantly outperforms LightGCN in long-tail recall and hit rate. Nearly 45% of users receive at least one relevant tail item in their Top-20 recommendations, confirming that semantic alignment effectively mitigates popularity bias.

6.4 Cold-Start Item Recommendation

Cold-start items are defined as those with five or fewer interactions in the training set.

6.4.1 Collaborative and Hybrid Fusion Models

LightGCN and both hybrid fusion models achieve zero recall for cold-start items:

$$\text{Recall@}K = 0, \quad \text{HitRate@}K = 0.$$

This confirms the fundamental limitation of interaction-based learning for items without historical feedback [11].

6.4.2 Two-Stage Semantic Fallback Strategy

To overcome this limitation, we employ a two-stage recommendation strategy: (i) collaborative ranking for warm items, followed by (ii) semantic similarity-based retrieval for cold items using BERT embeddings.

Table 5: Cold-start performance with two-stage semantic fallback

K	Recall@K	HitRate@K
5	0.0038	0.0080
10	0.0065	0.0106
20	0.0094	0.0212
50	0.0425	0.0716

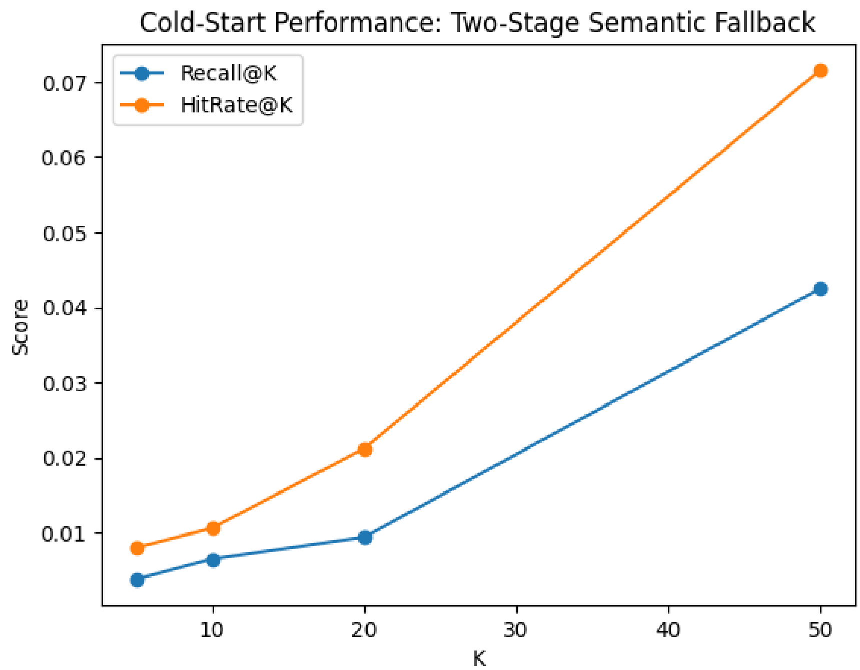


Figure 5: Cold-start recall and hit rate using the two-stage semantic fallback strategy.

Figure 5 shows that cold-start recall increases steadily with K, demonstrating that semantic similarity enables meaningful retrieval even in the absence of interaction data.

6.5 Warm-Item Performance Comparison Across Models

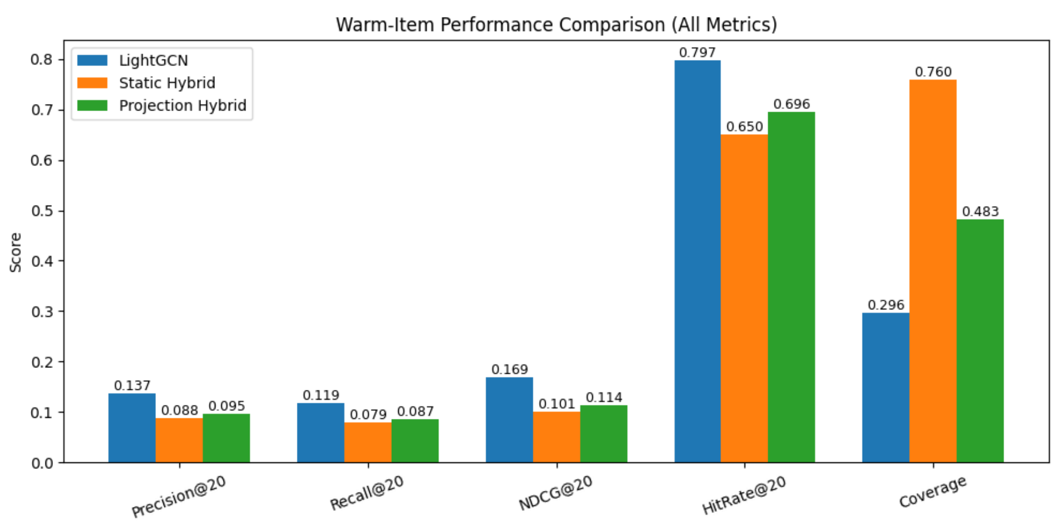


Figure 6: Warm-item performance comparison across collaborative and hybrid models.

Figure 6 summarizes warm-item performance across all metrics. While LightGCN remains the strongest model for ranking accuracy, hybrid models substantially increase catalog coverage. The projection-based hybrid achieves a balanced trade-off, improving diversity without severely compromising accuracy.

7 Conclusion

This study demonstrates that semantic augmentation can substantially improve long-tail recommendation performance, even when it does not surpass collaborative filtering on global ranking accuracy. Using the MovieLens-1M dataset under a leakage-free per-user chronological split, we confirm that LightGCN remains the strongest model for warm-item ranking, achieving Recall@20 $\approx$ 0.1186 and NDCG@20 $\approx$ 0.1692. However, consistent with prior findings, LightGCN exhibits pronounced popularity bias, limiting exposure to infrequently interacted items.

Hybrid strategies that incorporate BERT-derived semantic signals significantly alleviate this limitation. In particular, a projection-based hybrid model that learns a linear alignment from BERT embedding space into the LightGCN collaborative space demonstrates strong long-tail effectiveness. When evaluation is restricted to tail items (defined as the bottom 80% by interaction frequency), the projection-based hybrid achieves Recall@20 $\approx$ 0.0985 and HitRate@20 $\approx$ 0.4541, indicating that nearly 45% of users receive at least one relevant long-tail item in their Top-20 recommendations. At the same time, this approach substantially increases catalog coverage, reaching approximately 0.76 under tail-focused evaluation. Static weighted fusion further expands catalog exposure but incurs a larger degradation in ranking accuracy, while the learned projection recovers a significant portion of this loss and provides a more balanced accuracy–diversity trade-off.

Cold-start evaluation highlights complementary strengths and limitations of the proposed framework. Pure collaborative and embedding-fusion methods yield zero recall for truly cold items (defined as having five or fewer training interactions), confirming the inherent limitations of interaction-only learning. A pragmatic, industry-aligned two-stage semantic fallback strategy—employing collaborative ranking for warm items and BERT-based semantic similarity for cold items—enables non-zero cold-start retrieval, achieving Recall@20 $\approx$ 0.0094 and Recall@50 $\approx$ 0.0425, while ensuring that cold items constitute approximately 25% of overall recommendation exposure. An exploratory user–content matching model trained directly in semantic space further improves cold-start recall to Recall@20 $\approx$ 0.0152, though gains remain constrained by sparse and brief textual metadata.

Taken together, these findings support a modular and practical design principle for recommender systems: collaborative filtering should be preserved as the primary mechanism for ranking accuracy, while semantic content representations should be selectively integrated to enhance long-tail exposure and cold-start robustness. Projection-based hybridization offers a principled means of improving diversity without severely compromising accuracy, while explicit content-aware retrieval mechanisms are necessary to address cold-start scenarios. This hybrid approach yields a balanced recommendation framework that broadens the effective catalog and surfaces long-tail content at scale, addressing a central limitation of interaction-only recommender models.

8 Future Work

Several promising directions emerge from this study. First, future work can explore tighter integration between collaborative and semantic representations by jointly training the projection layer and collaborative model, or by incorporating semantic embeddings directly into graph propagation. While this work intentionally preserves a modular design to reflect production constraints, end-to-end hybrid optimization may further improve alignment between interaction-driven and content-driven signals.

Second, cold-start performance could be enhanced through richer user-side semantic modeling. Incorporating user-generated text, such as reviews, tags, or viewing histories encoded in natural language, would enable more expressive user–content matching and may substantially improve cold-user and cold-item retrieval. Additionally, leveraging contrastive or self-supervised objectives in semantic space could improve robustness under sparse textual metadata.

Third, diversity-aware and fairness-aware objectives present an important extension. Explicitly optimizing for long-tail exposure, novelty, or calibrated popularity bias—for example through re-ranking strategies or constrained optimization—could complement the hybrid architecture and provide stronger guarantees on catalog coverage without sacrificing accuracy.

Finally, extending the framework to sequential and session-based recommendation settings represents a natural progression. Incorporating temporal dynamics, user intent modeling, or large language model–based reasoning over user histories could further improve recommendation quality in real-world, evolving environments. Evaluating the proposed hybrid strategies at larger scales and across additional domains would provide further insight into their generality and deployment readiness.

References

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*, pages 42–46. ACM, 2017.

[2] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics, 2019.

[4] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. In *Proceedings of the 2015 ACM Conference on Interactive Experiences for TV and Online Video*, pages 19–26. ACM, 2015.

[5] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 639–648. ACM, 2020.

[6] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[7] Michael J. Pazzani and Daniel Billsus. Content-based recommendation systems. *The Adaptive Web*, pages 325–341, 2007.

[8] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, pages 452–461. AUAI Press, 2009.

[9] Francesco Ricci, Lior Rokach, and Bracha Shapira, editors. *Recommender Systems Handbook*. Springer, Boston, MA, 2011.

[10] Martin Saveski and Amin Mantrach. Item cold-start recommendations: Learning local collective embeddings. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 89–96. ACM, 2014.

[11] Andrew I. Schein, Alexandrin Popescul, Lyle H. Ungar, and David M. Pennock. Methods and metrics for cold-start recommendations. *Proceedings of the 25th Annual International ACM SIGIR Conference*, pages 253–260, 2002.

[12] Hongwei Wang, Naiyan Wang, and Dit-Yan Yeung. A survey on deep learning for recommender systems. *arXiv preprint arXiv:1707.07435*, 2018.

[13] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 165–174. ACM, 2019.

[14] Shu Wu, Fei Sun, Wentao Zhang, et al. A survey on deep learning based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):1–21, 2021.