Pseudo-Label NCF for Sparse OHC Recommendation: Dual-Representation Learning and the Separability–Accuracy Trade-off

PRONOB KUMAR BARMAN, University of Maryland, Baltimore County, USA
TERA L. REYNOLDS, University of Maryland, Baltimore County, USA
JAMES FOULDS, University of Maryland, Baltimore County, USA

Online Health Communities (OHCs) connect patients for peer support, but users face a discovery challenge when they have minimal prior interactions to guide personalization. We address recommendation under extreme interaction sparsity in a survey-driven setting where each user provides a 16-dimensional intake vector at registration and each support group has a structured feature profile. We augment three Neural Collaborative Filtering (NCF) architectures—Matrix Factorization (MF), Multi-Layer Perceptron (MLP), and NeuMF—with an auxiliary pseudo-label objective derived from survey–group feature alignment using cosine similarity mapped to $[0, 1]$. This yields Pseudo-Label NCF (PL-NCF), which learns dual embedding spaces: main embeddings optimized for ranking, and PL-specific embeddings intended to capture semantic alignment between users and support groups.

On a dataset of 165 users, 498 support groups, and 498 observed memberships (three per user), we evaluate ranking quality (HR@5, NDCG@5) primarily under a leave-one-out protocol—one validation and one test interaction per user, yielding approximately one training positive per user—which best reflects our target cold-start regime, with a train-validation-test (70/15/15) split for supplementary analysis. We analyze embedding structure in the leave-one-out setting using spherical k-means silhouette scores computed in the original high-dimensional embedding space, alongside 2D t-SNE visualizations for qualitative inspection. Under leave-one-out evaluation, all three PL variants improve ranking: MLP-PL improves HR@5 from 2.65 % to 5.30 %, NeuMF-PL from 4.46 % to 5.18 %, and MF-PL from 4.58 % to 5.42 %. Critically, PL-specific embedding spaces exhibit higher cosine silhouette scores than baseline main embeddings under per-seed optimal clustering with $k \in \{3, 4, 5, 6, 7, 8, 10\}$: MF-PL improves from 0.0394 to 0.0684; NeuMF-PL improves from 0.0263 to 0.0653. We also observe a separability–accuracy trade-off: main-embedding clusterability is negatively correlated with ranking accuracy (Spearman $\rho \approx -0.38$ on leave-one-out; $\rho \approx -0.59$ on the train-validation-test split), suggesting that embeddings optimized for ranking may sacrifice interpretability. These findings demonstrate that survey-derived pseudo-labels can regularize sparse-data training for improved ranking, while producing interpretable task-specialized embedding spaces via dedicated PL-specific representations.

CCS Concepts: • **Information systems** → **Recommender systems**; • **Applied computing** → *Health informatics*.

Additional Key Words and Phrases: recommender systems, neural collaborative filtering, pseudo-labeling, sparse-data recommendation, online health communities, representation learning, embedding visualization

Authors' Contact Information: Pronob Kumar Barman, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, Maryland, USA, pbarman1@umbc.edu; Tera L. Reynolds, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, Maryland, USA, reynoter@umbc.edu; James Foulds, Department of Information Systems, University of Maryland, Baltimore County, Baltimore, Maryland, USA, jfoulds@umbc.edu.

Manuscript submitted to ACM

ACM Reference Format:
Pronob Kumar Barman, Tera L. Reynolds, and James Foulds. 2026. Pseudo-Label NCF for Sparse OHC Recommendation: Dual-Representation Learning and the Separability–Accuracy Trade-off. In *Proceedings of ACM Conference on Recommender Systems (RecSys 2026).* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Online Health Communities (OHCs) support patients managing chronic conditions by enabling peer connection, experiential knowledge exchange, and coping strategies [3, 5, 19, 23]. Yet the moment when support can matter most is often the moment when recommendation is technically hardest: at onboarding, users need timely discovery of relevant support groups, but platforms have little to no behavioral history to personalize recommendations. As OHC catalogs expand, this cold-start regime creates a discovery bottleneck in which users must choose among dozens or hundreds of groups with only minimal interaction signals, increasing the risk of poor initial matches and early disengagement.

We study support-group recommendation when the primary available input is a structured intake survey. Our setting differs from (i) group recommender systems that aggregate preferences for established groups [1, 11, 20], which assume existing member preferences, and (ii) content-based methods that mine forum text [4], which require post-hoc community analysis. Instead, we adopt a feature-driven approach where survey responses yield a compact 16-dimensional representation of user needs, and each group has a corresponding aggregated feature profile. This structure also provides an interpretable alignment signal: the cosine similarity between user and group features, which we term *AlignFeatures*.

Neural Collaborative Filtering (NCF) models [14] capture non-linear user–item interactions beyond matrix factorization, but performance typically depends on interaction density. Graph-based recommenders such as LightGCN [13] and self-supervised methods [32] address sparsity via structural priors or contrastive objectives, yet healthcare recommendation demands conservative claims due to the gap between offline metrics and clinical outcomes [10]. To inject auxiliary supervision under extreme sparsity without requiring dense historical interactions, we introduce **Pseudo-Label Neural Collaborative Filtering (PL-NCF)**: we augment MF, MLP, and NeuMF with a soft pseudo-label objective derived from survey–group feature alignment. Crucially, PL-NCF maintains dual embedding spaces: main embeddings for ranking and PL-specific embeddings for semantic clustering, enabling specialization for distinct tasks.

This dual-representation approach relates to multi-task learning [7, 26], where shared representations benefit from joint optimization of related tasks. However, rather than sharing all parameters, PL-NCF maintains separate embedding spaces for each objective, allowing task-specific specialization. Beyond top-K ranking metrics, we evaluate embedding geometry using spherical k-means, cosine silhouette scores [25], and 2D visualizations using t-SNE [18]. To avoid an unfair baseline in embedding analysis, we apply the same k-selection protocol to all models and representations: for each model, seed, entity, and representation, we evaluate $k \in \{3, 4, 5, 6, 7, 8, 10\}$ in the original embedding space and select the k that maximizes cosine silhouette with tie-break to the smallest k. Crucially, we observe a separability–accuracy trade-off: higher main-embedding clusterability negatively correlates with ranking performance (Spearman $\rho \approx -0.38$ on leave-one-out; $\rho \approx -0.59$ on the train-validation-test split), suggesting that embeddings optimized for ranking may sacrifice interpretability.

Research questions.

- **RQ1:** Do survey-derived soft pseudo-labels improve NCF ranking performance (HR@5, NDCG@5) under extreme interaction sparsity, and how do effects vary across MF, MLP, and NeuMF architectures?

- **RQ2:** Do PL models dedicated PL-specific embeddings exhibit superior clusterability compared to baseline main embeddings and PL models own main embeddings, supporting the dual-representation hypothesis?
- **RQ3:** Is the optimal number of embedding clusters k model- and representation-dependent, and does fixed $k = 5$ mischaracterize embedding structure?
- **RQ4:** What is the relationship between embedding clusterability and ranking accuracy in sparse-data regimes?

Contributions.

(1) **Dual-representation PL-NCF framework.** We extend MF, MLP, and NeuMF with an auxiliary feature-alignment objective that learns separate PL-specific embeddings alongside main ranking-optimized embeddings, enabling task-specialized representations.
(2) **Fair cluster-count selection for embedding analysis.** We show that the optimal number of clusters is model- and representation-dependent, and we apply the same k-selection procedure to all models to avoid biased comparisons.
(3) **Empirical analysis on sparse OHC data.** We show that PL variants improve leave-one-out ranking metrics and that PL-specific embeddings achieve higher cosine silhouette scores than baseline main embeddings under optimal-k spherical clustering.
(4) **Separability–accuracy trade-off.** We observe a negative correlation between main-embedding clusterability at fixed $k = 5$ and ranking performance, providing empirical evidence that interpretability and recommendation accuracy can compete under sparsity.

2 Related Work

2.1 OHC support and group formation

OHCs have been shown to support peer-to-peer sense-making and reduce isolation [19, 23], with evidence linking participation to improved self-management [5]. Prior work on healthcare group formation has leveraged the textual content of discussions, such as topic modeling [4]. Our focus differs: we assume onboarding-time access to structured survey features and study how to recommend support groups when interaction logs are sparse or minimal.

2.2 Group recommender systems and fairness

Group recommender systems typically aggregate individual preferences to recommend items to a group [1, 11, 20]. Fairness-aware group recommendation has been studied to avoid marginalizing less active members [15, 27]. In our setting, we recommend groups to individuals in sparse-data contexts; fairness remains important for healthcare deployment, but our work primarily addresses sparse supervision and representation learning.

2.3 Neural collaborative filtering and sparse-data recommendation

NCF [14] introduced GMF, MLP, and NeuMF to model non-linear interactions, building on matrix-factorization foundations [16, 24]. Graph-based recommenders such as NGCF [30] and LightGCN [13] further exploit connectivity structure, but performance often depends on interaction density. Sparse-data and cold-start recommendation can incorporate side information and self-supervised or contrastive objectives [31, 32]. In healthcare, conservative framing is critical because offline metrics do not necessarily translate to clinical outcomes [10]. We leverage structured surveys to provide a feature-alignment pseudo-label signal that augments sparse interaction supervision.

2.4 Pseudo-labeling and multi-task learning

Pseudo-labeling [17] is a simple semi-supervised learning approach; it is related to entropy minimization [12] and has been extended via curriculum strategies [8] and analyses of confirmation bias [2]. Our use of pseudo-labels differs from self-training: we derive fixed soft targets from survey–group alignment, avoiding feedback loops from a model labeling its own predictions.

Multi-task learning [7, 26] jointly optimizes related objectives to improve generalization. Multi-task recommender systems [33] have shown benefits from combining ranking with auxiliary signals. Our dual-representation approach maintains separate embedding spaces for each task, enabling stronger task-specific specialization than fully shared representations.

2.5 Healthcare recommender systems

Healthcare recommender systems have been applied to rehabilitation and mental-health interventions [9, 28]. Recent work has explored language-model-based methods for mental-health support and conversational recommendation with an emphasis on safety [6, 22]. Our work contributes a survey-driven sparse-data framework for recommending support groups in OHCs, with conservative claims rooted in available offline signals.

2.6 Embedding analysis and visualization

Silhouette analysis [25] provides a quantitative measure of cluster separation, while t-SNE [18] and UMAP [21] offer qualitative 2D visualizations of embedding geometry. We use 2D projections only for visualization, and we quantify cluster structure in the original embedding space to avoid distortion artifacts from non-linear dimensionality reduction.

3 Methodology

3.1 Problem formulation

Let $\mathcal{U} = \{u_1, \ldots, u_n\}$ be a set of n users and $\mathcal{G} = \{g_1, \ldots, g_m\}$ be a set of m support groups. Each user $u \in \mathcal{U}$ has a survey feature vector $\mathbf{x}_u \in \mathbb{R}^{16}$ derived from intake questionnaire responses, and each group $g \in \mathcal{G}$ has a feature profile $\mathbf{z}_g \in \mathbb{R}^{16}$ aggregated from member surveys (construction detailed in Section 4.1). Observed user–group memberships form an implicit-feedback matrix $\mathbf{Y} \in \{0, 1\}^{n \times m}$, where $y_{ug} = 1$ indicates observed membership.

Our goal is to learn a scoring function $s_\theta(u, g) \in [0, 1]$ that ranks groups for each user under extreme sparsity (three observed memberships per user). The survey features $\mathbf{x}_u$ and $\mathbf{z}_g$ are always available, enabling feature-driven prediction even with minimal interaction history.

3.2 Survey representations and AlignFeatures pseudo-labels

User features. Each user 16-dimensional vector $\mathbf{x}_u$ concatenates two survey components: (i) six Q33 weights capturing support-preference dimensions and (ii) ten Q26 weights reflecting demographic and health-condition factors. Weights within each component are normalized to sum to 1, yielding a unit-simplex representation per component.

Group features. Each group profile $\mathbf{z}_g \in \mathbb{R}^{16}$ is constructed via k-nearest-neighbor aggregation with $k = 6$ of member survey features, with three repetitions per user during dataset generation. This yields 498 groups for 165 users (details in Section 4.1).

AlignFeatures pseudo-label. For each observed user–group pair (u, g) with $y_{ug} = 1$, the dataset provides a feature-alignment score:

$$\text{Align}(u, g) = \frac{\cos(\mathbf{x}_u, \mathbf{z}_g) + 1}{2} \in [0, 1], \tag{1}$$

mapping cosine similarity to the unit interval. We treat $\tilde{y}_{ug} = \text{Align}(u, g)$ as a fixed soft target during training, motivated by a homophily assumption common in peer-support settings: similarity in structured needs and context, as captured by intake surveys, is a reasonable proxy for short-list suitability when behavioral evidence is absent.

3.3 PL-NCF architectures

We compare three NCF architectures [14]—Matrix Factorization (MF), Multi-Layer Perceptron (MLP), and NeuMF—each with a baseline variant (binary supervision only) and a pseudo-label (PL) variant.

Baseline models.

- **MF**: $\hat{y}_{ug} = \sigma(\mathbf{p}_u^\top \mathbf{q}_g)$, where $\mathbf{p}_u, \mathbf{q}_g \in \mathbb{R}^{d_{\text{MF}}}$ and σ is the sigmoid.
- **MLP**: $\hat{y}_{ug} = \text{MLP}([\mathbf{p}_u; \mathbf{q}_g])$, with $\mathbf{p}_u, \mathbf{q}_g \in \mathbb{R}^{d_{\text{MLP}}}$.
- **NeuMF**: $\hat{y}_{ug} = \sigma(\mathbf{w}^\top [\mathbf{p}_u^{\text{GMF}} \odot \mathbf{q}_g^{\text{GMF}}; \text{MLP}([\mathbf{p}_u^{\text{MLP}}; \mathbf{q}_g^{\text{MLP}}])])$.

PL models with dual embedding spaces. PL variants introduce a second set of embeddings:

- **Main embeddings** $\mathbf{p}_u, \mathbf{q}_g$ optimized for interaction prediction.
- **PL-specific embeddings** $\mathbf{p}_u^{\text{PL}}, \mathbf{q}_g^{\text{PL}} \in \mathbb{R}^{d_{\text{PL}}}$ dedicated to learning feature-alignment structure.

The PL branch computes:

$$a_{ug}^{\text{PL}} = \cos(\mathbf{p}_u^{\text{PL}}, \mathbf{q}_g^{\text{PL}}) = \frac{\mathbf{p}_u^{\text{PL}\top} \mathbf{q}_g^{\text{PL}}}{\|\mathbf{p}_u^{\text{PL}}\|_2 \|\mathbf{q}_g^{\text{PL}}\|_2}, \tag{2}$$

which is fused with the main prediction pathway (architecture-dependent fusion). For MF-PL and NeuMF-PL, we additionally project $\mathbf{x}_u$ to dimension d_{PL} and compute cosine similarity against $\mathbf{q}_g^{\text{PL}}$ as an auxiliary signal.

Architectural details. We use standard NCF-style architectures with model-specific embedding sizes. MF baselines use $d_{\text{MF}} = 64$; MF-PL uses $d_{\text{MF}} = 96$ for main MF embeddings and $d_{\text{PL}} = 32$ for the PL branch. MLP uses $d_{\text{MLP}} = 32$ and a single hidden layer of size 32. NeuMF uses $d_{\text{GMF}} = 32$, $d_{\text{MLP}} = 64$, and MLP layers $\{128, 64, 32\}$; PL variants use $d_{\text{PL}} = 32$. Figure 1 illustrates the full PL-NCF dual-representation architecture.

3.4 Training objective

We train with on-the-fly uniform negative sampling: for each observed positive (u, g) with $y_{ug} = 1$, we sample one negative $g' \sim \mathcal{G} \setminus \{g' \mid y_{ug'} = 1\}$. The baseline loss is binary cross-entropy:

$$\mathcal{L}_{\text{BCE}} = -\sum_{(u,g) \in \mathcal{D}} \big[y_{ug} \log \hat{y}_{ug} + (1 - y_{ug}) \log(1 - \hat{y}_{ug}) \big], \tag{3}$$

where $\mathcal{D}$ is the training set including sampled negatives.

For PL models, when AlignFeatures $\tilde{y}_{ug}$ is available for pair (u, g), we add a soft-label cross-entropy term:

$$\mathcal{L}_{\text{PL}} = -\sum_{(u,g) \in \mathcal{D}_{\text{PL}}} \big[\tilde{y}_{ug} \log \hat{y}_{ug} + (1 - \tilde{y}_{ug}) \log(1 - \hat{y}_{ug}) \big], \tag{4}$$

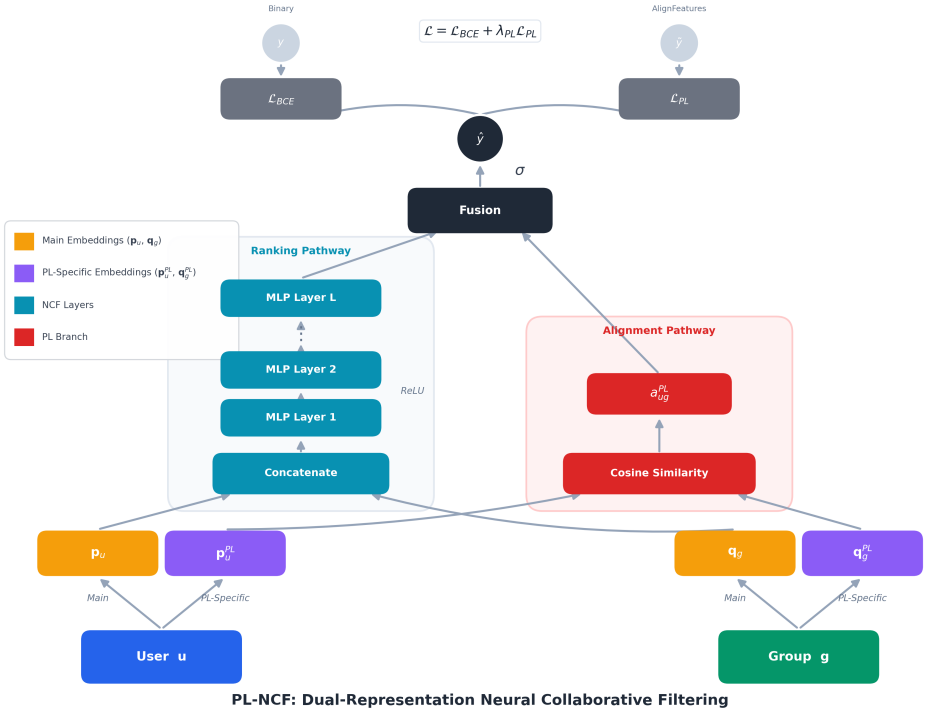


Fig. 1. Overview of the PL-NCF dual-representation architecture. Each user u and group g maintain separate main embeddings ($\mathbf{p}_u$, $\mathbf{q}_g$) for the ranking pathway and PL-specific embeddings ($\mathbf{p}_u^{\text{PL}}$, $\mathbf{q}_g^{\text{PL}}$) for the alignment pathway. The ranking pathway processes main embeddings through NCF layers, while the alignment pathway computes cosine similarity a_{ug}^{PL} between PL-specific embeddings. Both pathways are fused and supervised jointly by binary cross-entropy $\mathcal{L}_{\text{BCE}}$ and pseudo-label loss $\mathcal{L}_{\text{PL}}$.

where $\mathcal{D}_{\text{PL}} \subseteq \mathcal{D}$ contains pairs with AlignFeatures scores. The combined objective is:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda_{\text{PL}}\mathcal{L}_{\text{PL}}, \tag{5}$$

where λ_{PL} controls pseudo-label supervision strength.

3.5 Embedding clustering and visualization

After training, we extract user and group embedding matrices. For PL models, we extract both main embeddings and PL-specific embeddings.

Spherical k-means clustering in high-dimensional space. We apply k-means clustering to ℓ_2-normalized embeddings (spherical k-means) [25]. For each embedding set, we evaluate $k \in \{3, 4, 5, 6, 7, 8, 10\}$ and compute cosine silhouette scores using cosine distance on normalized embeddings. The silhouette computation is performed in the original embedding space, not on any 2D projection. For each model, seed, entity, and representation, we select the optimal k (the k maximizing cosine silhouette; tie-break: smallest k), then report mean and standard deviation across seeds. For the separability–accuracy analysis in Section 5.3, we additionally report main-embedding silhouette at fixed $k = 5$ to avoid post-hoc tuning effects.

2D visualization using t-SNE.. We project high-dimensional embeddings to 2D using t-SNE [18] for qualitative inspection. We apply ℓ_2 normalization before projection and use cosine distance with perplexity 15. Crucially, spherical k-means cluster labels and silhouette metrics are computed in the original high-dimensional embedding space; the resulting high-dimensional cluster labels are then overlaid on the 2D t-SNE coordinates.

4 Experimental Setup

4.1 Dataset

We use a survey-driven support-group dataset constructed from OHC intake questionnaires. The dataset contains:

- **Users**: $n = 165$, each with a 16-dimensional survey vector $\mathbf{x}_u$.
- **Support groups**: $m = 498$, each with a 16-dimensional aggregated feature profile $\mathbf{z}_g$ generated via k-nearest-neighbor aggregation with $k = 6$ of member surveys with three repetitions per user.
- **Observed memberships**: 498 total (three per user), representing synthetic group assignments based on feature similarity.

This synthetic construction reflects a bootstrapping scenario where groups are created from user needs rather than organic community formation.

4.2 Evaluation protocols

We evaluate under two data-split regimes:

(1) **Leave-one-out (primary)**: For each user, one interaction is held out for validation and one for test; remaining interactions (approximately one per user) form the training set. This protocol best reflects the extreme cold-start regime central to our work.
(2) **Train-validation-test split (70/15/15)**: A random split over all 498 observed memberships, providing supplementary analysis under slightly less extreme sparsity.

4.3 Metrics

Ranking metrics. We evaluate top-K ranking with $K = 5$ using sampled evaluation: for each held-out positive, we sample 99 negatives uniformly from groups not observed in training for that user, forming a candidate set of 100. We report HR@5 and NDCG@5. Metrics are aggregated over test users and reported as percentages.

Clustering metrics. We quantify embedding clusterability via cosine silhouette score computed on ℓ_2-normalized embeddings with cosine distance. For each model, seed, entity, and representation, we select $k \in \{3, 4, 5, 6, 7, 8, 10\}$ maximizing cosine silhouette, then report mean and standard deviation across seeds (Table 3). For the separability–accuracy analysis, we use main-embedding silhouette at fixed $k = 5$.

4.4 Implementation details

Models are implemented in PyTorch and trained for 20 epochs with AdamW. We report mean and standard deviation over five random seeds (42, 52, 62, 122, 232). Hyperparameters follow the architecture-specific settings in Section 3.2 through Section 4.1, with model- and protocol-specific λ_{PL} values (MF-PL: 0.03 leave-one-out and 0.40 train-validation-test split; MLP-PL: 0.25 and 0.20; NeuMF-PL: 0.35 and 0.50). For 2D visualizations, we use t-SNE with ℓ_2-normalized

Table 1. Recommendation performance under leave-one-out evaluation (primary), mean ± std over 5 seeds.

Model	HR@5 (%)	NDCG@5 (%)
MF	4.58 ± 1.93	2.70 ± 1.33
MF-PL	**5.42 ± 2.25**	**3.32 ± 1.51**
MLP	2.65 ± 1.63	1.41 ± 0.97
MLP-PL	**5.30 ± 1.88**	**2.97 ± 1.12**
NeuMF	4.46 ± 0.91	2.50 ± 0.58
NeuMF-PL	**5.18 ± 1.25**	**3.02 ± 0.63**

Table 2. Supplementary result: Recommendation performance on train-validation-test (70/15/15) split, mean ± std over 5 seeds.

Model	HR@5 (%)	NDCG@5 (%)
MF	4.57 ± 1.86	2.90 ± 1.47
MF-PL	2.57 ± 1.86	2.15 ± 1.48
MLP	0.57 ± 0.78	0.23 ± 0.32
MLP-PL	**1.43 ± 1.75**	**0.72 ± 1.04**
NeuMF	3.43 ± 0.78	2.43 ± 0.78
NeuMF-PL	**6.29 ± 0.78**	**3.90 ± 0.35**

Table 3. Cosine silhouette scores under leave-one-out evaluation (mean ± std over 5 seeds). Each entry uses per-seed optimal $k \in \{3, 4, 5, 6, 7, 8, 10\}$ computed in the original embedding space.

Model	User Main	User PL	Group Main	Group PL
MF	0.0394 ± 0.0018	–	0.0318 ± 0.0007	–
MF-PL	0.0265 ± 0.0020	**0.0684 ± 0.0050**	0.0223 ± 0.0013	**0.0572 ± 0.0011**
MLP	0.0687 ± 0.0013	–	0.0577 ± 0.0015	–
MLP-PL	0.0680 ± 0.0026	**0.0716 ± 0.0028**	0.0569 ± 0.0017	**0.0567 ± 0.0015**
NeuMF	0.0263 ± 0.0018	–	0.0222 ± 0.0008	–
NeuMF-PL	0.0256 ± 0.0011	**0.0653 ± 0.0022**	0.0220 ± 0.0004	**0.0571 ± 0.0015**

embeddings, cosine distance, perplexity 15, and seed-specific random state. All clustering decisions and silhouette metrics are computed in the original embedding space and only visualized in 2D via overlay.

5 Results

5.1 Ranking performance

Table 1 reports our primary leave-one-out results, and Table 2 provides supplementary results on the train-validation-test split.

5.2 Embedding clustering quality

Table 3 reports cosine silhouette scores for user and group embeddings under leave-one-out. PL-specific embeddings achieve higher clusterability than both baseline embeddings and PL models own main embeddings.

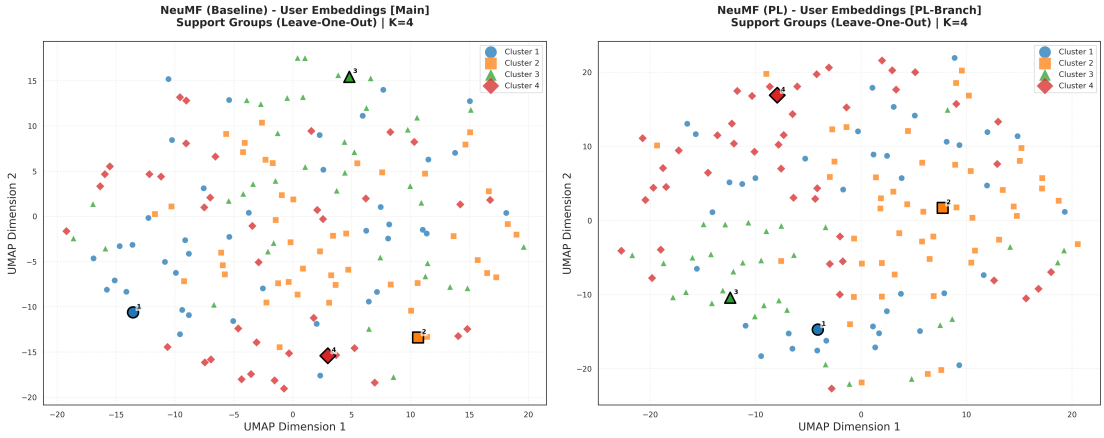


Fig. 2. t-SNE visualization of user embeddings under leave-one-out evaluation. Left: NeuMF baseline main embeddings. Right: NeuMF-PL PL-specific embeddings. Cluster labels are computed via spherical k-means in the original embedding space and overlaid on the 2D coordinates for visualization only.

5.3 Separability–accuracy trade-off

We analyze the relationship between main-embedding silhouette at fixed $k = 5$ and ranking performance across model–seed runs. Spearman rank correlation [29] reveals a negative relationship:

- **Leave-one-out (primary)**: $\rho_{\text{sil, HR@5}} \approx -0.38$
- **Train-validation-test split (supplementary)**: $\rho_{\text{sil, HR@5}} \approx -0.59$

This indicates that models with more clusterable main embeddings tend to achieve lower ranking accuracy in this sparse-data setting.

5.4 2D visualization using t-SNE

Figure 2 shows an example t-SNE projection of user embeddings under leave-one-out comparing NeuMF baseline main embeddings to NeuMF-PL PL-specific embeddings. Points are colored by spherical k-means cluster labels computed in the original embedding space, with k selected by the same high-dimensional silhouette procedure.

6 Discussion

6.1 When does pseudo-labeling improve ranking

Pseudo-labeling exhibits protocol- and architecture-dependent effects. Under leave-one-out—our primary evaluation reflecting extreme sparsity with approximately one training positive per user—all three PL variants consistently improve both HR@5 and NDCG@5, demonstrating that pseudo-label regularization is most beneficial when collaborative signals are minimal. This consistency across all architectures is a key finding. Under the train-validation-test split, gains are architecture-dependent: NeuMF-PL achieves large improvements, while MF-PL degrades. This divergence suggests that when more training data is available, hybrid architectures can leverage auxiliary signals constructively, whereas bilinear models may experience interference from pseudo-label objectives that conflict with interaction-structure learning.

6.2 Why do PL-specific embeddings cluster better

Three mechanisms explain PL-specific embedding improvements:

(1) **Objective decoupling**: main embeddings optimize ranking loss; PL-specific embeddings are shaped by feature-alignment supervision. Separate embedding spaces enable task-specific specialization.
(2) **Cosine-consistent geometry**: AlignFeatures derives from cosine similarity, and we evaluate clustering using spherical k-means with cosine silhouette, aligning supervision and evaluation geometry.
(3) **Feature-grounded semantics**: PL-specific embeddings receive direct supervision from survey-derived similarities, encoding interpretable feature-based groupings independent of interaction structure.

6.3 Implications for healthcare recommendation

Survey-driven recommendation is pragmatic in healthcare: intake questionnaires are standard practice, and structured features avoid privacy-sensitive text mining. However, AlignFeatures is not ground-truth preference; offline metric improvements do not guarantee engagement or clinical outcomes without real-world validation. Our dataset is small and synthetically constructed via neighbor aggregation, limiting generalizability. We position this work as a proof-of-concept for dual-representation learning under extreme sparsity, not a deployable clinical system.

6.4 Broader implications for representation learning

The negative correlation between main-embedding clusterability and ranking accuracy suggests caution when interpreting intrinsic embedding metrics or visually appealing 2D projections as proxies for downstream performance. Dual-representation architectures provide a practical mechanism to obtain task-specialized embeddings when interpretability and ranking compete.

7 Limitations

Feature-similarity proxy for preference. AlignFeatures pseudo-labels are derived from survey–group feature similarity, not observed user engagement or satisfaction. Improvements in offline HR@5 under sampled evaluation do not directly imply improved real-world engagement or health outcomes.

Small-scale synthetic dataset. The dataset contains 165 users and 498 support groups with three memberships per user, constructed via nearest-neighbor aggregation of survey features rather than organic community formation. The small scale limits statistical power and increases variance across seeds.

2D projections are qualitative only. t-SNE provides intuitive 2D visualizations of embedding geometry but involves non-linear dimensionality reduction with sensitivity to hyperparameters and random seeds. We therefore compute clustering and silhouette scores in the original embedding space and use 2D projections only for qualitative visualization via label overlays.

Limited baseline breadth. We compare PL-augmented versus baseline NCF variants but do not include non-NCF baselines such as purely feature-based ranking or recent self-supervised approaches. These omissions limit claims about broader state-of-the-art performance.

8 Conclusion

We presented PL-NCF, a survey-driven pseudo-labeling framework for Neural Collaborative Filtering in sparse-data support-group recommendation. This dual-representation approach maintains separate main embeddings optimized for ranking and PL-specific embeddings shaped by feature-alignment pseudo-labels, enabling task specialization. Under leave-one-out evaluation—our primary protocol reflecting extreme cold-start sparsity—pseudo-labeling consistently improves ranking across all three architectures, while train-validation-test split results are architecture-dependent. In the leave-one-out setting, PL-specific embedding spaces exhibit higher clusterability than baseline main embeddings under optimal-k spherical clustering, and we observe a separability–accuracy trade-off: embedding clusterability can be negatively correlated with ranking quality. Future work must validate survey-derived alignment signals against real OHC engagement and outcomes and explore richer survey encoders under conservative and fairness-aware evaluation.

9 Reproducibility

We release code, configuration, and derived artifacts needed to reproduce the reported results, including trained model checkpoints across five seeds, extracted embedding matrices, per-k silhouette grids, and 2D visualization figures. We emphasize that all clustering decisions and cluster quality metrics are computed in the original embedding space; 2D projections are used only to visualize those high-dimensional cluster labels.

References

[1] Sihem Amer-Yahia, Senjuti Basu Roy, Ashish Chawlat, Gautam Das, and Cong Yu. 2009. Group recommendation: Semantics and efficiency. *Proceedings of the VLDB Endowment* 2, 1 (2009), 754–765. https://www.vldb.org/pvldb/vol2/vldb09-858.pdf Accessed 2026-01-01.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*. IEEE, 1–8. https://arxiv.org/abs/1908.02983 arXiv preprint, accessed 2026-01-01.

[3] Pronob Kumar Barman, James R. Foulds, and Tera L. Reynolds. 2026. Understanding User Perceptions of Human-Centered AI-Enhanced Support Group Formation in Online Healthcare Communities. arXiv:2603.11237 https://arxiv.org/abs/2603.11237 arXiv preprint.

[4] Pronob Kumar Barman, Tera L Reynolds, and James Foulds. 2025. Facilitating Online Healthcare Support Group Formation Using Topic Modeling. *Studies in health technology and informatics* 329 (2025), 1049–1053.

[5] R. Bernardi and P. Wu. 2017. The impact of online health communities on patients' health self-management. In *Proceedings of the International Conference on Information Systems (ICIS)*. Seoul, South Korea.

[6] Andreas Bucher, Sarah Egger, Inna Vashkite, Wenyuan Wu, and Gerhard Schwabe. 2025. "It's Not Only Attention We Need": Systematic Review of Large Language Models in Mental Health Care. *JMIR Mental Health* 12, 1 (2025), e78410. https://pubmed.ncbi.nlm.nih.gov/41186978/ Accessed 2026-01-01.

[7] Rich Caruana. 1997. Multitask learning. *Machine Learning* 28 (1997), 41–75.

[8] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2021. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 6912–6920. https://cdn.aaai.org/ojs/16852/16852-13-20346-1-2-20210518.pdf Accessed 2026-01-01.

[9] Yoon-Seop Chang, Mikyong Han, Boosun Jeon, Jae-Chul Kim, and NohSam Park. 2023. An Neural Collaborative Filtering (NCF) based Recommender System for Personalized Rehabilitation Exercises. In *2023 14th International Conference on Information and Communication Technology Convergence (ICTC)*. IEEE, 1292–1297. https://ieeexplore.ieee.org/document/10393615 Accessed 2026-01-01.

[10] Zeshan Fayyaz, Mahsa Ebrahimian, Dina Nawara, Ahmed Ibrahim, and Rasha Kashef. 2020. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *applied sciences* 10, 21 (2020), 7748. https://www.mdpi.com/2076-3417/10/21/7748 Accessed 2026-01-01.

[11] Alexander Felfernig, Ludovico Boratto, Martin Stettinger, and Marko Tkalcic. 2018. Group recommender systems. *Springer* 10 (2018), 3284588.

[12] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised Learning by Entropy Minimization. In *Advances in Neural Information Processing Systems*. 529–536. https://proceedings.neurips.cc/paper_files/paper/2004/file/96f2b50b5d3613adf9c27049b2a888c7-Paper.pdf Accessed 2026-01-01.

[13] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 639–648. https://dl.acm.org/doi/10.1145/3397271.3401063 Accessed 2026-01-01.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the International World Wide Web Conference (WWW)*. 173–182. https://arxiv.org/abs/1708.05031 arXiv preprint, accessed 2026-01-01.

[15] Junjie Jia, Fen Wang, Huijuan Wang, and Shilong Liu. 2024. A Fairness Group Recommendation Algorithm Based On User Activity. *International Journal of Computational Intelligence Systems* 17, 1 (2024), 204. https://link.springer.com/article/10.1007/s44196-024-00602-9 Accessed 2026-01-01.

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

[17] Dong-Hyun Lee. 2013. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *ICML Workshop on Challenges in Representation Learning*. https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks Accessed 2026-01-01.

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[19] Paul Marshall, Millissa Booth, Matthew Coole, Lauren Fothergill, Zoe Glossop, Jade Haines, Andrew Harding, Rose Johnston, Steven Jones, Christopher Lodge, et al. 2024. Understanding the impacts of online mental health peer support forums: realist synthesis. *JMIR Mental Health* 11 (2024), e55750. https://mental.jmir.org/2024/1/e55750 Accessed 2026-01-01.

[20] Judith Masthoff. 2011. Group recommender systems: Combining individual models. In *Recommender Systems Handbook*. Springer, 677–702.

[21] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018). arXiv:1802.03426 https://arxiv.org/abs/1802.03426 Accessed 2026-01-01.

[22] Abdulaziz Mohammed, Mingwei Zhang, Gehad Abdullah Amran, Husam M Alawadh, Ruizhe Wang, Amerah Alabrah, and Ali A Al-Bakhrani. 2025. A social information sensitive model for conversational recommender systems. *PeerJ Computer Science* 11 (2025), e3067. https://pubmed.ncbi.nlm.nih.gov/40989476/ Accessed 2026-01-01.

[23] John A Naslund, Kelly A Aschbrenner, Lisa A Marsch, and Stephen J Bartels. 2016. The future of mental health care: peer-to-peer support and social media. *Epidemiology and psychiatric sciences* 25, 2 (2016), 113–122. https://pmc.ncbi.nlm.nih.gov/articles/PMC4830464/ Accessed 2026-01-01.

[24] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. 452–461. https://arxiv.org/abs/1205.2618 arXiv preprint, accessed 2026-01-01.

[25] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65. https://www.sciencedirect.com/science/article/pii/0377042787901257 Accessed 2026-01-01.

[26] Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098* (2017). https://arxiv.org/abs/1706.05098 Accessed 2026-01-01.

[27] Dimitris Sacharidis. 2019. Top-N Group Recommendations with Fairness. In *Proceedings of the ACM Symposium on Applied Computing*. 1663–1670. https://dsachar.net/publication/2019-sac-s/2019-sac-s.pdf Accessed 2026-01-01.

[28] Emily Slade, Stefan Rennick-Egglestone, Fiona Ng, Yasuhiro Kotera, Joy Llewellyn-Beardsley, Chris Newby, Tony Glover, Jeroen Keppens, Mike Slade, et al. 2024. The implementation of recommender systems for mental health recovery narratives: evaluation of use and performance. *JMIR Mental Health* 11, 1 (2024), e45754. https://mental.jmir.org/2024/1/e45754/ Accessed 2026-01-01.

[29] Charles Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1 (1904), 72–101.

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

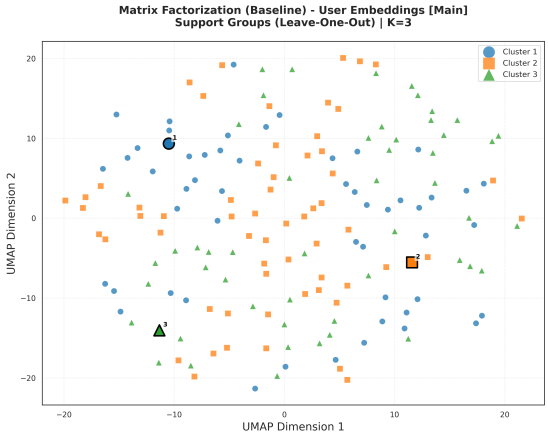
[31] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive learning for cold-start recommendation. In *Proceedings of the 29th ACM international conference on multimedia*. 5382–5390. https://arxiv.org/abs/2107.05315 arXiv preprint, accessed 2026-01-01.

[32] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. 2023. Self-supervised learning for recommender systems: A survey. *IEEE Transactions on Knowledge and Data Engineering* 36, 1 (2023), 335–355. https://arxiv.org/abs/2203.15876 arXiv preprint, accessed 2026-01-01.

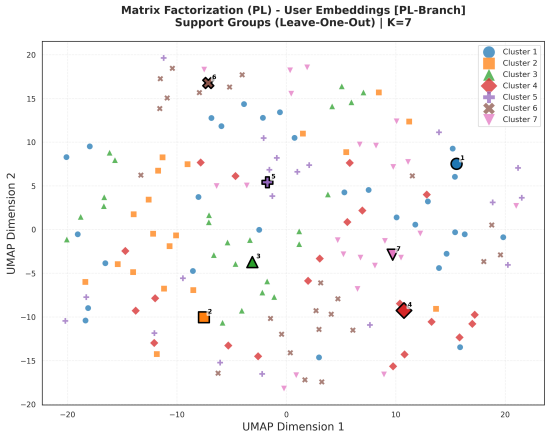
[33] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM conference on recommender systems*. 43–51.

A Additional Embedding Visualizations

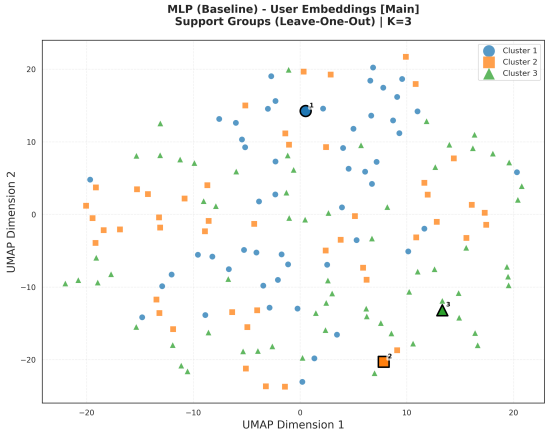
Figure 3 provides additional t-SNE visualizations under leave-one-out comparing baseline main embeddings to PL-specific embeddings for MF and MLP. These figures are included for qualitative inspection and are not used as inputs to clustering or evaluation.



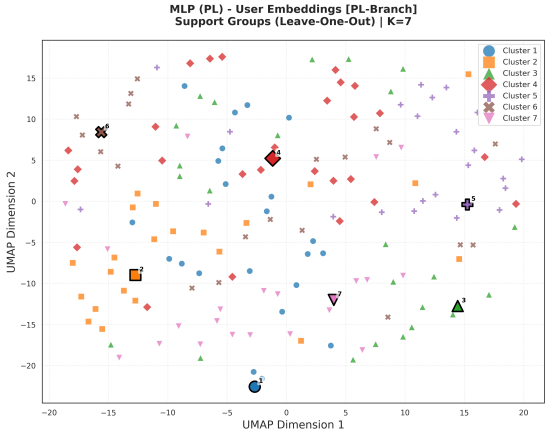
(a) MF baseline main embeddings



(b) MF-PL PL-specific embeddings



(c) MLP baseline main embeddings



(d) MLP-PL PL-specific embeddings

Fig. 3. Additional t-SNE visualizations under leave-one-out comparing baseline main embeddings (left) to PL-specific embeddings (right) for MF and MLP. Cluster labels are computed via spherical k-means in the original embedding space and overlaid on 2D projections for visualization only.