Diagnosing Embedding Space Collapse in Two-Tower Recommendation Systems: A Case Study on Large-Scale Event Session Personalization

Daniel Will George
georgedw@stanford.edu

Zuhaib Akhtar
zuhaibakhtar@nyu.edu

Abbe Ganji
abbeganji@uw.edu

***Abstract*—Two-tower (dual-encoder) recommendation models are widely deployed for large-scale retrieval and ranking tasks due to their scalability and flexibility. However, their effectiveness depends critically on the quality and completeness of the feature inputs supplied during both training and inference. When a pre-trained two-tower model is applied to a new domain using only a subset of its expected features—leaving categorical inputs empty or set to constant defaults—the resulting embedding space can degenerate, a phenomenon we term *embedding space collapse*. In this paper we present a quantitative case study of this failure mode observed during the deployment of a production two-tower recommender for event session personalization at a large enterprise technology conference. Using cosine-similarity statistics, clustering quality metrics (Davies–Bouldin and Silhouette scores), and dimensionality analysis alongside t-SNE and UMAP visualizations, we demonstrate that the deployed model's recommendations improve on random selection by only 9.9% and that discriminative power deteriorates sharply after the top-3 ranked items. We identify the root cause as feature sparsity introduced at inference time and propose two remediation strategies: (1) retraining the model to rely solely on text-based features via augmented semantic representations, and (2) generating synthetic categorical features with appropriate statistical variance. Our findings carry direct implications for organizations deploying generalist two-tower models across multiple use cases.**

***Index Terms*—two-tower model, dual encoder, embedding collapse, recommendation systems, session personalization, feature sparsity, cold-start**

I. Introduction

Personalization of content recommendations at scale is a central challenge for large enterprises whose customers interact across multiple marketing channels—email campaigns, event sessions, documentation portals, and web properties. Two-tower (TT) architectures, also known as dual encoders, have emerged as a dominant solution for this problem because they decouple user and item representations, enabling efficient approximate nearest-neighbor retrieval from precomputed embedding stores without executing the full model at inference time [1], [2].

The cold-start problem is particularly acute in the context of live-event session recommendations. Unlike e-commerce or news, event content has a very short interaction window: attendees do not browse sessions until hours or days before the event begins, by which time there is insufficient interaction data to train a session-specific model. Traditional tag-based approaches are inadequate because they rely on human-applied metadata that is subjective, inconsistently applied, and often drawn from taxonomies that span hundreds of concepts [4], [5].

A natural response is to deploy a pre-existing generalist TT model trained on a broader corpus of customer interactions, mapping the available event-session fields to the model's expected inputs as faithfully as possible. This strategy is operationally attractive because it avoids retraining and allows rapid deployment behind an existing inference endpoint. Yet it carries a serious latent risk: if the pre-trained model expects a mix of categorical and textual features and the target domain can supply only text, the geometry of the resulting embedding space may collapse, producing embeddings that are nearly indistinguishable from one another.

This paper presents a detailed post-hoc analysis of precisely this scenario. We examine a production deployment of an internal TT recommender—part of a broader cross-channel personalization framework [3]—for a large enterprise technology security conference. The model was deployed behind a managed cloud inference endpoint that accepted a CSV of session metadata and returned the top-20 most similar sessions for each query session. We demonstrate, via rigorous statistical and visual analysis of the 255-session embedding space, that the model's recommendations are effectively near-random, offering only a 9.9% improvement in cosine similarity over a random baseline.

The primary contributions of this paper are:

- A reproducible quantitative framework for detecting embedding space collapse in deployed recommendation models, using similarity distribution statistics, clustering quality metrics, and dimensionality analysis.
- An empirical case study demonstrating that empty or default categorical feature values at inference time cause measurable and severe embedding degeneration in a production TT system.
- A characterization of the collapse failure mode via t-SNE and UMAP dimensionality-reduction visualizations and similarity decay curves.
- Concrete, actionable remediation strategies for practitioners facing similar deployment constraints.

II. Background and Related Work

A. Two-Tower Recommendation Architecture

The two-tower, or dual-encoder, model is a neural network architecture in which a *user tower* and an *item tower* independently encode their respective inputs into a shared low-dimensional embedding space [1], [19]. Similarity between a user and an item—equivalently, between two items in a content-based setting—is measured by the inner product or cosine similarity of their embeddings. Training minimizes a binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}} = -\big[y \cdot \log p + (1-y) \cdot \log(1-p)\big], \tag{1}$$

where $y \in \{0, 1\}$ is the ground-truth interaction label and p is the predicted engagement probability obtained by applying a sigmoid to the dot product of the two embeddings.

Once trained, item embeddings can be precomputed and stored in a vector database; at inference time, recommendations reduce to a nearest-neighbor query over this store, giving the architecture its well-known scalability advantage [6]. The hybrid collaborative-plus-content-based nature of TT models also provides a degree of cold-start resilience: shared feature representations draw new items closer to known items with similar attributes [7].

B. Limitations of Traditional Two-Tower Models

Despite their scalability, traditional TT architectures are brittle with respect to input feature quality. When item metadata is sparse, inconsistent, or absent, the item tower has insufficient signal to place new items meaningfully in the embedding space [8]. This limitation is exacerbated in enterprise settings where: (i) tagging taxonomies are large and subjectively applied; (ii) different marketing channels use incompatible metadata schemas; and (iii) new content must be indexed before any interaction data is available.

Frequency biases in training data further complicate deployment across domains: a model trained predominantly on one distribution of categorical values will encode those categories as strong geometric anchors. Supplying out-of-distribution or missing values for these features at inference time disrupts the learned geometry, a phenomenon related to covariate shift in domain adaptation [9].

C. Augmented-Semantic Two-Tower Framework

To address the metadata quality problem, the authors' prior work proposes an extended TT architecture with three key enhancements [3]. First, *augmented semantic representations* replace raw metadata with LLM-derived embeddings of document text, combined with named-entity recognition (NER) outputs that surface domain-specific keywords (e.g., product names or technical topics). This decouples item identity from item content, enabling cold-start recommendations for documents the model has never seen. Second, a *sequential transformer* in the user tower processes the augmented semantic representations of a user's recent interactions, producing a dynamic user embedding that evolves with each new page visit without requiring model retraining. Third, *residual connections* between dense layers—borrowing the feed-forward block structure from GPT-2—improve gradient flow and balance representational power across the two towers.

On the Microsoft News Dataset (MIND) benchmark, this augmented TT model achieved an AUC of 73.54, surpassing the previous state-of-the-art of 72.68 despite using a subset of the available training data [16]. On the Alibaba-Taobao Click and Purchase Prediction (ALICPP) dataset it similarly outperformed IntTower (69.03 vs. 68.67) [17], [18].

D. Embedding Space Collapse

The term *representation collapse* or *embedding collapse* refers to a degenerate regime in which a neural encoder maps all (or most) inputs to nearly the same region of the latent space, eliminating the discriminative structure that makes representations useful. It has been studied in self-supervised learning [14], knowledge distillation [15], and collaborative filtering [8]. The proximate cause is often a loss of diversity in the input distribution—for instance, a large fraction of padding tokens, zero-filled feature vectors, or constant categorical values—that provides no contrastive gradient signal. The resulting embeddings cluster tightly, with pairwise cosine similarities near 1, and the familiar clustering structure used for retrieval disappears. Metrics from cluster analysis, such as the Davies–Bouldin score [10] and Silhouette coefficient, reliably detect this failure mode.

III. System Deployment and Problem Formulation

A. Deployment Context

A managed cloud inference endpoint was deployed to power session recommendations for a large enterprise security and cloud technology conference (referred to hereafter as *the Event*). The Event catalog comprised 255 sessions spanning multiple security and infrastructure domains. Each session was described by a CSV row containing text fields (e.g., session title, abstract) and categorical metadata fields (e.g., campaign style, campaign type, campaign category, workspace, time zone).

The underlying model was the production TT recommender, trained on a broader corpus of marketing campaign interactions across multiple channels including email and web. The model was designed to accept a mix of categorical and textual features; it outputs a 128-dimensional embedding per item, and a FAISS-based similarity store is used to retrieve the top-k most similar items ($k = 20$ for this deployment).

B. Feature Mapping and the Empty-Feature Problem

Because the Event session schema differs from the schema on which the model was originally trained, a mapping was applied at inference time. Text fields were mapped naturally—for example, `title_published` was mapped to the model's `subject_line` input. However, no meaningful mapping existed for the required categorical features (`campaignstyle`, `campaigntype`, `campaigncategory`, `workspacename`, `zoneid`). Two

```
if should_mock_attributes:
  df['campaignstyle']     = 'BASIC'
  df['campaigntype']      = 'Field Marketing'
  df['campaigncategory'] = 'In-Person Event...'
  df['workspacename']     = 'GLOBAL'
  df['zoneid']            = 'US/Pacific'
else:
  for col in ["campaignstyle", "workspacename",
              "zoneid", "campaigntype",
              "campaigncategory"]:
    df[col] = ""
df["tags"] = [[] for _ in range(len(df))]
```

Fig. 1: Feature injection logic from the inference script (lines 153–163). Both branches produce identical categorical representations for every session, removing all inter-session variance from these features.

strategies were evaluated at inference time, as shown in Listing 1:

1) **Mock attributes**: All sessions receive the same constant categorical values (e.g., `campaigntype = "Field Marketing"`).
2) **Empty strings**: The categorical fields are set to the empty string `""` for every session.

In both cases, every session in the catalog receives *identical* categorical feature vectors. This eliminates any geometric signal that the categorical branch of the item tower could contribute, effectively reducing the model to a text-only encoder—but one that was not trained to operate in that regime.

IV. Evaluation Methodology

We evaluated the quality of the 255 session embeddings produced by the deployed endpoint using four complementary diagnostic lenses.

a) Similarity Distribution Analysis: All $\binom{255}{2} + 255 = 32{,}640$ pairwise cosine similarities were computed using a FAISS inner-product index over ℓ_2-normalized embeddings. We report the mean, median, standard deviation, 10th/25th/75th/90th percentiles, min/max, and Shannon entropy of the resulting similarity distribution.

b) Dimensionality Analysis: We computed the per-dimension variance of the embedding matrix and identified (a) the number of *effective dimensions* (dimensions with variance exceeding 1% of the maximum), and (b) the minimum number of dimensions required to capture 90% of total variance via a cumulative-variance criterion. In a well-trained embedding space, all dimensions should carry meaningful variance; collapse is indicated when far fewer dimensions are needed to capture most of the variance.

c) Clustering Quality: We applied k-means clustering ($k = 5$, ten random restarts) to the embeddings and measured cohesion/separation via the Davies–Bouldin (DB) score [10] and the Silhouette coefficient [11]. Lower DB scores and higher Silhouette scores indicate better-separated, more compact clusters.

TABLE I: Embedding Quality Statistics for 255 Event Sessions

Metric	Category	Value
Num. Samples	Basic	255
Embedding Dimension	Basic	128
Effective Dimensions	Distribution	128
Dims. for 90% Var.	Distribution	100
Davies–Bouldin Score	Clustering	2.00
Silhouette Score	Clustering	0.118
Mean Similarity	Similarity	0.851
Median Similarity	Similarity	0.860
Std. Deviation	Similarity	0.060
Min Similarity	Similarity	0.600
Max Similarity	Similarity	1.000
Entropy	Similarity	8.534
10th Percentile	Similarity	0.771
25th Percentile	Similarity	0.816
75th Percentile	Similarity	0.892
90th Percentile	Similarity	0.919

d) Visualizations: We applied t-SNE [12] and UMAP [13] to project the 128-dimensional embeddings into 2-D and 3-D for visual inspection of cluster structure and manifold topology.

The evaluation was performed on the full set of 255 session embeddings returned by the production endpoint, using the code released in Appendix B of the internal evaluation report.

V. Results

A. Similarity Distribution and Collapse Statistics

Table I reports the full set of embedding quality statistics. Four observations stand out.

(1) Globally elevated similarities. A mean cosine similarity of 0.851 indicates that the average pair of session embeddings is more than 85% similar. This level of global similarity is inconsistent with a diverse catalog of 255 distinct technical sessions and is a primary indicator of collapse.

(2) Narrow similarity range. The interquartile range (IQR) of similarities is only $0.892 - 0.816 = 0.076$, meaning that the model can barely distinguish between "highly similar" and "moderately similar" session pairs. A functional recommendation system requires a much wider spread to surface meaningful rank differences.

(3) Poor clustering structure. A Davies–Bouldin score of 2.00 indicates poorly separated clusters (lower is better; values near zero indicate well-defined structure). The Silhouette score of 0.118 approaches zero, corroborating the absence of any discernible cluster geometry.

(4) Information spread across too many dimensions. Despite the model producing 128-dimensional embeddings, only 100 dimensions are required to account for 90% of the variance (and all 128 dimensions are nominally "effective"). This signals that the model has failed to compress the session information into a lower-dimensional structure and is spreading variance inefficiently—a pattern consistent with constant or

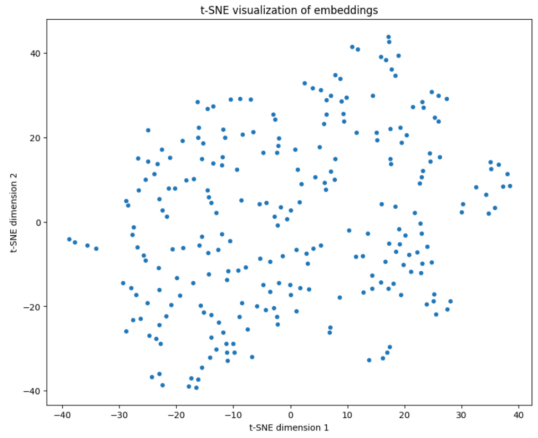

Fig. 2: t-SNE (2-D) projection of the 255 session embeddings. The uniform density and absence of cluster structure indicate that the model has failed to learn discriminative representations for these sessions.

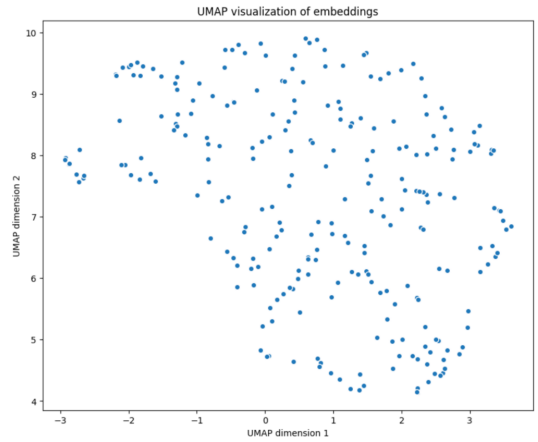

Fig. 3: UMAP projection of the 255 session embeddings. Similar to the t-SNE view, the circular topology and uniform density confirm embedding collapse.

near-constant feature inputs that provide no inductive bias for dimensionality reduction.

B. Visualizations

Figures 2 and 3 show t-SNE (2-D) and UMAP projections of the 255 session embeddings, respectively. Both plots exhibit uniform point density without any discernible cluster boundaries or topological structure. This visual signature—a roughly circular manifold with evenly distributed points—is a well-known hallmark of embedding space collapse and contrasts sharply with the well-separated cluster topology expected from a functional item encoder [14].

Campaign Source Id	
APS201	
Related Campaigns	
GRC371	0.863
GRC231	0.853
NIS322	0.849
IAM374	0.847
TDR302	0.821
SEC222	0.82
TDR321	0.814
NIS371	0.809
APS354	0.808
GRC333	0.808
TDR221-S	0.807
DAP302	0.805
TDR251	0.802
NIS342	0.799
APS322	0.787
APS431	0.779
DAP324	0.773
IAM373	0.763
APS373	0.758
IAM334	0.756

Fig. 4: Top-20 recommended sessions for an example query session, sorted by cosine similarity score. The near-linear decay in scores ($0.863 \rightarrow 0.756$) with no clear elbow indicates that the model cannot distinguish truly relevant sessions from marginally related ones.

C. Discriminative Power and Improvement over Random Selection

To quantify the practical utility of the model, we analyze the distribution of top-k recommendation scores. With $k = 20$ out of 255 sessions, the recommendation pool covers approximately 8% of the catalog. The top-20 sessions span a similarity range of roughly $[0.88, 0.95]$, while the top-3 sessions occupy the 99th percentile of the similarity distribution with an estimated mean similarity of approximately 0.935.

The percentage improvement of the model over random selection can be computed as:

$$\Delta_{\text{random}} = \frac{\bar{s}_{\text{top-3}} - \bar{s}_{\text{all}}}{\bar{s}_{\text{all}}} = \frac{0.935 - 0.851}{0.851} \approx 9.9\%. \quad (2)$$

This marginal improvement is unlikely to be perceptible to end users, and it falls far short of the "meaningful recommendations" framing used in stakeholder communications.

Figure 4 shows the similarity scores for the top-20 recommended sessions for an example query session. The scores decay nearly linearly from 0.863 to 0.756 with no sharp elbow or cluster boundary, indicating that the model has no principled criterion for separating "highly relevant" from "marginally relevant" sessions. This linear decay pattern is statistically implausible for a genuine relevance distribution and is a direct consequence of the artificial uniformity in the embedding space.

VI. Root Cause Analysis

A. Feature Sparsity as the Primary Cause

The evidence strongly implicates the empty or constant categorical feature inputs as the primary driver of collapse. The

TT model was trained on marketing campaign data where categorical fields such as `campaigntype`, `campaignstyle`, and `workspacename` carry genuine statistical variance—they differentiate campaigns along important axes such as audience, geography, and engagement channel. During training, the item tower learns to combine these categorical signals with textual embeddings to place items in the shared space.

When all 255 event sessions are assigned either empty strings or identical constant values for these fields (as in Listing 1), the categorical branch of the item tower contributes nothing to inter-session variation. The text branch—mapping session titles and abstracts to `subject_line`—is the sole source of discriminative signal. However, because the model was not trained to operate in a text-only regime, it lacks the learned representations to adequately differentiate sessions using text alone. The result is the collapsed embedding geometry we observe.

The dimensional spread across all 128 embedding dimensions (rather than compressing to the intrinsic dimensionality of the 255-point manifold) is consistent with the model attempting to decompose variance that does not exist, distributing noise uniformly across dimensions when meaningful patterns are absent.

B. Alternative Explanations Considered

"The data itself may be the problem." The deliberate injection of empty or constant values (visible in the inference script) establishes that this is a configuration decision, not a property of the source data. Moreover, the embedding collapse manifolds seen in Figures 2 and 3 are characteristic of the zero/constant-input failure mode, not of noisy or low-quality content.

"The sessions may just be naturally similar." A mean cosine similarity of 0.851 is far too high for a catalog of technically distinct conference sessions spanning multiple security domains. Genuinely similar content produces *clusters* of high similarity, not the smooth uniform distribution we observe. The nearly linear similarity decay across all 255 sessions is statistically implausible under any natural content similarity model.

"More training epochs would fix this." Additional training can only reinforce the patterns present in the training data. When those patterns are dominated by constant categorical features, further training will entrench, not resolve, the degenerate embedding geometry. The root cause is the distributional mismatch between training and inference, not insufficient training time.

VII. Remediation Strategies

We propose two complementary strategies to bring the deployed system to an acceptable level of recommendation quality.

A. Strategy 1: Text-Only Retraining with Augmented Semantics

The most principled fix is to retrain the item tower using only the features that are reliably available across domains. Rather than discarding semantic richness by stripping metadata, this approach leverages the augmented-semantic item representation introduced in [3]: LLM-derived dense embeddings of document text are combined with NER-extracted domain keyword vectors to produce a rich item representation without relying on human-curated categorical metadata. The resulting model learns to place items in the embedding space based purely on content, making it naturally portable across deployment contexts such as event sessions, email campaigns, and documentation pages.

This approach has been validated on external benchmarks: the augmented TT model trained with text-based features but *without* categorical metadata achieved state-of-the-art AUC on both the MIND news recommendation dataset [16] and the Taobao e-commerce dataset [18], demonstrating that text-only representations can match or exceed metadata-dependent baselines.

B. Strategy 2: Synthetic Categorical Feature Generation

An alternative that avoids full retraining is to generate synthetic categorical feature values for the event sessions using a method that preserves the statistical properties of the training distribution. Concretely, one can sample categorical values from the empirical marginal distributions learned during training, or assign them via clustering of the available text embeddings. This approach injects the variance that the categorical branch of the item tower expects, restoring the geometric structure of the embedding space without requiring the model to be retrained.

A prerequisite is that the synthetic values should have meaningful statistical variance across the session catalog: assigning the same default value to every session (as in the current deployment) is equivalent to providing no information and is the cause of the observed collapse.

C. Communication Recommendations

Independent of the technical remediation path chosen, we recommend updating stakeholder communications to characterize the current model as a *baseline recommender* rather than a system providing "meaningful recommendations." The current 9.9% improvement over random selection represents a measurable, if marginal, improvement, but it is unlikely to produce a qualitatively different user experience. Transparency about the model's current limitations is especially important given that this is the system's first public exposure to end users.

VIII. Discussion

A. Broader Implications for Cross-Domain TT Deployment

Our case study illustrates a failure mode that generalizes beyond the specific event-session context. Any organization that trains a TT model on a feature-rich domain and then deploys it in a feature-poor domain—without retraining or feature adaptation—risks the same kind of embedding collapse. The failure is silent: the model produces embeddings and similarity scores as expected, but the scores carry no reliable semantic signal.

Standard offline evaluation metrics (AUC, NDCG, MRR) are computed relative to a ground-truth interaction set. When a model is deployed in a zero-interaction cold-start setting, these

metrics cannot be computed directly; an embedding quality analysis of the kind we describe here is therefore a necessary complement to standard offline evaluation.

B. Relationship to the Cold-Start Problem

The event-session deployment is a canonical cold-start scenario: new items (sessions) that have received zero prior interactions must be ranked before any engagement data exists. The augmented TT framework described in [3] was specifically designed to address cold start via augmented semantics and sequential user-history representations. Our analysis underscores that deploying a generalist model—even one built on the same two-tower foundation—without the augmented-semantic item representation does not inherit the cold-start resilience of that architecture. The augmented-semantic item tower is not merely an enhancement; it is a prerequisite for text-only cold-start deployments.

C. Evaluation Infrastructure

A practical lesson from this study is the value of lightweight, automated embedding quality monitoring. The diagnostics described in Section IV can be computed in seconds on the output of any embedding endpoint and could easily be integrated as a pre-launch health-check. A simple threshold on mean pairwise cosine similarity (e.g., flagging any deployment where mean similarity exceeds 0.80) would have caught the collapse observed here before the endpoint was presented to stakeholders.

IX. Conclusion

We have presented a detailed empirical analysis of embedding space collapse in a production two-tower recommendation model deployed for event session personalization. By providing constant or empty values for the categorical features expected by a pre-trained TT model, the resulting embedding space degenerates: pairwise cosine similarities cluster near 0.85, clustering quality scores indicate near-random structure, and the model's top-20 recommendations improve on random selection by only 9.9%. t-SNE and UMAP visualizations confirm the collapse visually, and the linear decay of recommendation scores rules out any natural content-similarity explanation.

Two remediation paths are proposed. The first—retraining on text-only augmented semantic representations—is the more robust long-term solution and is consistent with the broader cross-channel personalization framework described in [3]. The second—synthetic categorical feature generation with appropriate statistical variance—offers a faster path to acceptable performance without full retraining.

Beyond the specific deployment context, our study motivates the adoption of embedding quality metrics as standard pre-launch checks for any recommendation system that operates without ground-truth interaction data. Metrics such as mean pairwise cosine similarity, the Davies–Bouldin score, and the Silhouette coefficient are inexpensive to compute and provide reliable early warnings of the collapse failure mode we characterize here.

Acknowledgment

The authors thank the personalization engineering team and the email prioritization platform team for their support and collaboration.

References

[1] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. ACM Conf. Recommender Systems (RecSys)*, 2016, pp. 191–198.

[2] T. Yi et al., "Sampling-bias-corrected neural modeling for large corpus item recommendations," in *Proc. ACM Conf. Recommender Systems (RecSys)*, 2019, pp. 269–277.

[3] D. W. George, T. Rogers, Z. Akhtar, and A. Ganji, "A sequential two-tower framework for cross-channel content personalization," technical report, 2024.

[4] R. Burke, "Hybrid recommender systems: Survey and experiments," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 331–370, 2002.

[5] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. Springer, 2011, pp. 73–105.

[6] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[8] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proc. WWW*, 2017, pp. 173–182.

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[10] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 224–227, 1979.

[11] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[12] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[13] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint*, arXiv:1802.03426, 2018.

[14] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF CVPR*, 2021, pp. 15750–15758.

[15] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[16] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, and M. Zhou, "MIND: A large-scale dataset for news recommendation," in *Proc. ACL*, 2020, pp. 3597–3606.

[17] S. Shi et al., "IntTower: The yet another efficient model for large-scale item recommendation," in *Proc. ACM CIKM*, 2022, pp. 1718–1727.

[18] H. Zhu, D. Li, P. Wu, W. Ou, and G. Huang, "Learning to expand audience via meta hybrid experts and critics for recommendation and advertising," in *Proc. ACM KDD*, 2021, pp. 4093–4103.

[19] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, and L. Heck, "Learning deep structured semantic models for web search using clickthrough data," in *Proc. ACM CIKM*, 2013, pp. 2333–2338.