

LIFELONG-LEARNING EMBEDDINGS: INCREMENTAL AND CONTINUAL REPRESENTATION LEARNING FOR DYNAMIC E-COMMERCE TRENDS

Anonymous authors

Paper under double-blind review

ABSTRACT

E-commerce is a highly dynamic domain where products and consumer behaviors evolve rapidly. Embedding-based representations are central to deep learning-based personalization systems, yet conventional embeddings are static and therefore, they cannot easily incorporate new tokens (e.g., new products) without retraining, which is costly and often infeasible due to privacy or data retention constraints. To address this, we propose Lifelong-Learning Embeddings, a framework that (1) incrementally extends embeddings to integrate new tokens, (2) adapts embedding dimensionality to balance expressiveness and efficiency, and (3) employs continual learning to mitigate catastrophic forgetting. Experiments on a real-world dataset and two benchmark datasets show that our approach consistently outperforms static embeddings in accuracy while incurring only modest training-time overhead, demonstrating its effectiveness and adaptability in dynamic e-commerce environments.

1 INTRODUCTION

Learning effective representations is a cornerstone of deep learning across domains such as computer vision (CV) (He et al., 2016; Chen et al., 2020) and natural language processing (NLP) (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020), where embeddings play a central role in mapping discrete inputs into continuous spaces that capture semantic and structural relationships. Embeddings enable models to generalize beyond raw symbols, underpinning advances in word representations (Mikolov et al., 2013; Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020), image features (He et al., 2016; Chen et al., 2020), as well as node embeddings (Abu-El-Haija et al., 2018; Menand & Seshadhri, 2024).

In e-commerce, embeddings serve a similarly fundamental function. Customer interactions, such as product views, clicks, or purchases, are tokenized into discrete identifiers (e.g., product IDs, URLs) that must be represented in dense vector spaces to power downstream tasks such as recommendation, purchase prediction, or personalization (Alves Gomes et al., 2022; Sheil et al., 2018b; Tercan et al., 2021; Srilakshmi et al., 2022; Vasile et al., 2016; Zeng et al., 2020; Huang et al., 2020; Chen et al., 2022; Zhou et al., 2018). Building on these dense representations, modern deep learning models leverage embeddings to capture evolving customer interests and item relationships, thereby driving state-of-the-art performance in personalization tasks (Zhou et al., 2019; Sun et al., 2019; Chen et al., 2022; Wang et al., 2025; Yang et al., 2024).

However, unlike in CV or NLP, e-commerce environments are inherently dynamic (Ferrera & Kessedjian, 2019). New products are continuously introduced, seasonal campaigns and emerging trends give rise to novel interaction patterns, and entire product segments may rapidly appear or disappear. This constant growth of the item vocabulary poses a major challenge for static embeddings, which typically require complete retraining to accommodate new tokens. Such retraining is computationally costly and memory-intensive, while also prone to catastrophic forgetting, where previously learned knowledge is overwritten by new information because retaining past data is often infeasible due to legal and economic constraints (European-Parliament, 2016; Tallon, 2010; Kiker, 2014; Demirel et al., 2024). At the same time, modern embedding tables in large-scale recommenders can reach terabyte scale, consuming vast resources and slowing training. Existing methods

often allocate uniform embedding sizes regardless of feature frequency, leading to inefficiencies and scalability limits (Luo et al., 2024;?).

This places three critical demands on any embedding mechanism used to support recommendation or prediction tasks in e-commerce:

- **Extendability:** seamlessly integrate new tokens without retraining from scratch.
- **Knowledge Retention:** preserve previously learned representations to mitigate catastrophic forgetting.
- **Compactness:** adapt the dimensionality of embeddings over time to balance expressiveness with computational efficiency.

To address these challenges, we propose Lifelong-Learning Embeddings (LLE), a dynamic embedding framework that integrates incremental vocabulary expansion, adaptive embedding dimensionality, and continual learning (CL) strategies. LLE is designed to incrementally accommodate new tokens while retaining past knowledge and maintaining a resource-efficient representation space. We evaluate LLE on large-scale industrial and benchmark e-commerce datasets, demonstrating consistent improvements over static baselines and providing detailed ablation studies to isolate the contribution of each component.

2 RELATED WORK AND ITS LIMITATIONS FOR E-COMMERCE

Foundational methods such as word2vec (Mikolov et al., 2013) and Transformers (Vaswani et al., 2017) have established dense semantic representations as the backbone of NLP. Out-of-vocabulary tokens are usually addressed with placeholder symbols (Sutskever et al., 2014) or subword techniques (Sennrich et al., 2015; Radford et al., 2019; Liu et al., 2019). However, these approaches are ill-suited to e-commerce, where product identifiers are arbitrary, non-compositional, and arrive at high frequency (cf. Figure 2). Unlike natural language, where vocabulary shifts gradually, e-commerce platforms encounter thousands of new items on a continuous basis, rendering static embedding strategies ineffective.

Research into dynamic embeddings has followed two main directions. The first optimizes *embedding dimensions* to reduce memory overhead. For instance, Luo et al. (2024) propose an approach that prunes or extends embedding chunks during training based on access frequency and gradient norms, achieving large memory savings in industrial recommender models. Dong et al. (2024) propose a Bayesian dimension selection framework for embeddings, using adaptive stochastic gradient and sparse regularization to prune irrelevant dimensions. Qu et al. (2024) tackles streaming recommendation by using reinforcement learning to allocate user and item embedding sizes under a strict memory budget, dynamically growing or shrinking embeddings as frequencies change. These works address scalability but not the integration of entirely new tokens.

The second direction studies *dynamic or inductive embeddings*, where representations adapt to evolving interactions. Zhang & et al. (2021) approach models user-item trajectories with graph neural networks, supporting continuous updates and inductive handling of new nodes. Zheng et al. (2021) treats cold-start as a meta-learning task, using a matching network to rapidly form embeddings for new items given a few user interactions. Rossi et al. (2020) provide a general framework for dynamic graphs, where memory modules and message passing allow embeddings to update with each event and to inductively represent new nodes. These methods support new entities through on-the-fly computation or memory updates, but do not maintain persistent embedding tables that expand as in real-world e-commerce systems. Weimann & Conrad (2025) decouple embeddings from the model, using a dynamic cache updated in real time through self-supervised objectives, enabling new items to be added without retraining the whole model. However, the cache is periodically refreshed and filled with newly trained embedding representations, so previously learned representations are not preserved across iterations.

In e-commerce personalization systems, embeddings remain the backbone of click prediction (Zhou et al., 2019; Zhang et al., 2024), ranking (Cheng et al., 2016; Wang et al., 2017; Sun et al., 2019), and behavior modeling (Sheil et al., 2018b; Srilakshmi et al., 2022; Zhang et al., 2023). Yet, most models are retrained on static datasets or periodically refreshed, lacking mechanisms for incremental and continual representation learning. Existing dynamic embedding methods focus either on optimizing

dimension sizes or on inductively generating embeddings, but none provide a principled approach to incrementally expanding embedding representations with new tokens while also preserving past knowledge. This gap is particularly acute in e-commerce, where new products arrive continuously and embeddings must both grow dynamically and adapt continually without retraining from scratch. The LLE framework introduced in this work addresses this need.

3 LIFELONG-LEARNING EMBEDDINGS APPROACH

LLE enables dynamic adaptation to incoming data while preserving previously acquired knowledge. The framework consists of three core components as illustrated in figure 1: (1) incremental embedding update, which is able to represent changing token distributions over time, (2) adaptive embedding dimensionality to maintain an efficient and general embedding representation, and (3) training embedding with CL strategies to stabilize the embeddings and prevent drift when training on new distributions. Whereby adding new tokens (1) or extending dimensions (2) does not affect the outputs of models that rely on LLE.

LLE is designed as a drop-in replacement for existing embedding layers, requiring only minimal adjustment of downstream layers if the embedding dimension changes. Each component is self-contained and can be deployed independently. Furthermore, the framework is method-agnostic: initialization strategies, resizing criteria, or CL objectives can be integrated without architectural changes, enabling rapid adoption of future advances. In the following subsections, we describe each component and its integration.

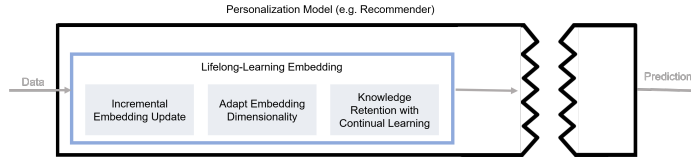


Figure 1: LLE framework: data feed an embedding module with three components: (1) incremental embedding update, (2) adaptive dimensionality, and (3) CL-based knowledge retention. LLE can replace any embedding layer in a prediction model, e.g. recommender.

3.1 INCREMENTAL EMBEDDING UPDATE

An embedding layer maps a token from a discrete vocabulary $\mathbb{V}$ to a d -dimensional latent space via a matrix $\mathbf{E} \in \mathbb{R}^{|\mathbb{V}| \times d}$, $d \in \mathbb{N}$, where each row encodes one token. The input is typically one-hot, so the layer functions as a lookup table of trainable parameters.

To handle evolving vocabularies, we introduce a dynamic update mechanism. Let $\mathbb{V}^{(t)}$ be the vocabulary at time t with embedding matrix $\mathbf{E}^{(t)} \in \mathbb{R}^{|\mathbb{V}^{(t)}| \times d}$, and $\mathbb{V}^{(t+1)}$ the updated vocabulary. The new embedding matrix $\mathbf{E}^{(t+1)}$ is defined as:

$$\mathbf{E}^{(t+1)}[i] = \begin{cases} \mathbf{E}^{(t)}[\pi(i)] & v_i \in \mathbb{V}^{(t)} \\ g(\mathbf{E}^{(t)}, v_i) & v_i \notin \mathbb{V}^{(t)}, \end{cases}$$

where $\pi(i)$ maps token indices and g is a function that initializes embeddings for unseen new tokens based on the weights from $\mathbf{E}^{(t)}$. For tokens that remain in both vocabularies, this definition guarantees that their embeddings are identical in $\mathbf{E}^{(t)}$ and $\mathbf{E}^{(t+1)}$, ensuring consistency across updates. The incremental embedding update supports both the growth and shrinkage of $\mathbb{V}$, enabling efficient and adaptive representations over time.

3.2 ADAPTIVE EMBEDDING DIMENSIONALITY

Beyond handling evolving vocabularies, embeddings must also adapt their latent dimensionality when the representational capacity of the current space is misaligned with the changing data. Given an embedding matrix $\mathbf{E}^{(t)} \in \mathbb{R}^{|\mathbb{V}^{(t)}| \times d}$ at time t , we allow both extension and reduction of the latent dimension of the embedding representation.

For an extension to $d' > d$, we allocate a new embedding matrix $\mathbf{E}^{(t+1)} \in \mathbb{R}^{|\mathbb{V}| \times d'}$. To preserve prior knowledge, the first d dimensions of each row are copied from $\mathbf{E}^{(t)}$, while the newly added $(d' - d)$ dimensions are initialized with zeros:

$$\mathbf{E}^{(t+1)}[i][d] = \mathbf{E}^{(t)}[i][d], \mathbf{E}^{(t+1)}[i][d : d'] = 0, \forall v_i \in \mathbb{V}.$$

This zero-initialization to handle new embedding dimensions ensures that the embedding outputs remain unchanged immediately after extension, so the model’s behavior is consistent while the new dimensions can be trained gradually. For a reduction to $d' < d$, a new embedding matrix $\mathbf{E}^{(t+1)} \in \mathbb{R}^{|\mathbb{V}| \times d'}$ is generated. Since direct truncation would discard information, CL is applied to transfer knowledge from the old higher-dimensional embedding $\mathbf{E}^{(t)}$ into the reduced representation.

3.3 EMBEDDING TRAINING WITH CONTINUAL LEARNING

CL provides a stability mechanism that restricts updates for shared tokens ($\mathbb{V}^{(t)} \cap \mathbb{V}^{(t+1)}$) to stay close to their historical representations, effectively controlling semantic drift. This allows the embedding space to adapt to new dimensionalities and new tokens without forgetting past knowledge. Thereby, applying CL to the proposed LLE approach is non-trivial. To make CL work in this setting, we need to solve three issues, since vocabularies may change and embedding dimensions may differ. Therefore, the following three design choices are needed:

1. Restrict knowledge perseveration to tokens present in both vocabularies.
2. Ensure that their indices are correctly matched.
3. Provide a mechanism to compare embeddings of possibly different dimension sizes.

First, we define the overlap of old and new vocabularies as

$$\tilde{\mathbb{V}} = \mathbb{V}^{(t)} \cap \mathbb{V}^{(t+1)}.$$

Only tokens in $\tilde{\mathbb{V}}$ are aligned, since new tokens in $\mathbb{V}^{(t+1)} \setminus \mathbb{V}^{(t)}$ have no prior representation.

Second, let $\pi^{(t)}$ and $\pi^{(t+1)}$ denote the index mappings into $\mathbf{E}^{(t)} \in \mathbb{R}^{|\mathbb{V}^{(t)}| \times d}$ and $\mathbf{E}^{(t+1)} \in \mathbb{R}^{|\mathbb{V}^{(t+1)}| \times d'}$, respectively. For each token $x \in \tilde{\mathbb{V}}$, its old and new embeddings are $\mathbf{E}^{(t)}[\pi^{(t)}(x), :]$ and $\mathbf{E}^{(t+1)}[\pi^{(t+1)}(x), :]$.

Third, to handle different embedding sizes ($d' \neq d$), we map both vectors into a common comparison space of dimension m_x :

$$\psi_x^{(t)} = P_x^{(t)} \mathbf{E}^{(t)}[\pi^{(t)}(x), :], \quad \psi_x^{(t+1)} = P_x^{(t+1)} \mathbf{E}^{(t+1)}[\pi^{(t+1)}(x), :],$$

where $P_x^{(t)} \in \mathbb{R}^{m_x \times d}$ and $P_x^{(t+1)} \in \mathbb{R}^{m_x \times d'}$ are comparison maps (e.g., identity or truncation if $d' = d$, padding for $d' > d$, or projection for $d' < d$).

Finally, the CL alignment loss is then defined as

$$\mathcal{L}_{\text{CL}} = \sum_{x \in \tilde{\mathbb{V}}} \ell_{\text{align}}(\psi_x^{(t+1)}, \psi_x^{(t)}),$$

where ℓ_{align} is a general alignment function (e.g., ℓ_2 -norm or cosine similarity).

Our approach is agnostic to the specific CL technique, and any method providing a regularization signal or teacher output from prior embeddings can be used. The total training loss is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{CL}} \mathcal{L}_{\text{CL}},$$

with $\mathcal{L}_{\text{task}}$ the supervised task loss (e.g., cross-entropy) and λ_{CL} a weighting factor. This enables embeddings to retain semantic structure across time while adapting to evolving token distributions. Minimizing $\mathcal{L}_{\text{CL}}$ bounds the drift of overlapping tokens, since for all $x \in \tilde{\mathbb{V}}$ the deviation $|\psi_x^{(t+1)} - \psi_x^{(t)}|$ is directly penalized. Thus, shared tokens remain anchored to their prior representations while the embedding space adapts to new dimensions and data.

In case of dimension upsizing ($d' > d$), the CL controls the gradients on the old dimensions with regularization so that semantics stay stable, while the new dimensions can gradually absorb additional information from the new distribution. When downsizing ($d' < d$), previous information is

lost and CL will align the new embedding $\mathbf{E}^{(t+1)}$ to the old space and reduce semantic loss. Even when the dimensionality stays the same ($d' = d$), CL provides stability by preventing overwriting of established semantics as token distributions shift. Finally, when increasing the token space, CL provides stability by anchoring shared tokens to their previous representations, ensuring minimal drift while the embedding space adapts.

4 DATASETS AND PREPROCESSING

We evaluate LLE on one proprietary industrial dataset and two commonly used public benchmarks. The industrial dataset spans 17 weeks and contains 53.8M events across 6.2M sessions. The YooChoose dataset¹, introduced for the RecSys Challenge 2015,² comprises 33M events from 9.2M sessions collected over 21 weeks, and has since become a widely used benchmark in the purchase prediction literature (Park et al., 2015; Romov & Sokolov, 2015; Lin et al., 2019; Esmeli et al., 2021; Alves Gomes et al., 2022). The RetailRocket dataset³ includes 2.75M events from 1.4M users spanning 18 weeks and is frequently adopted for user modeling tasks such as recommendation and churn prediction (Berger & Kompan, 2019; Fridrich & Dostál, 2022; Zeng et al., 2020; Sheil et al., 2018a).

Figure 2 illustrates the token dynamics over time. The datasets differ markedly in interaction volume, the rate of new token introduction, and the prevalence of unknown tokens, which allows for a comprehensive evaluation of LLE under diverse e-commerce patterns and evolving user behavior.

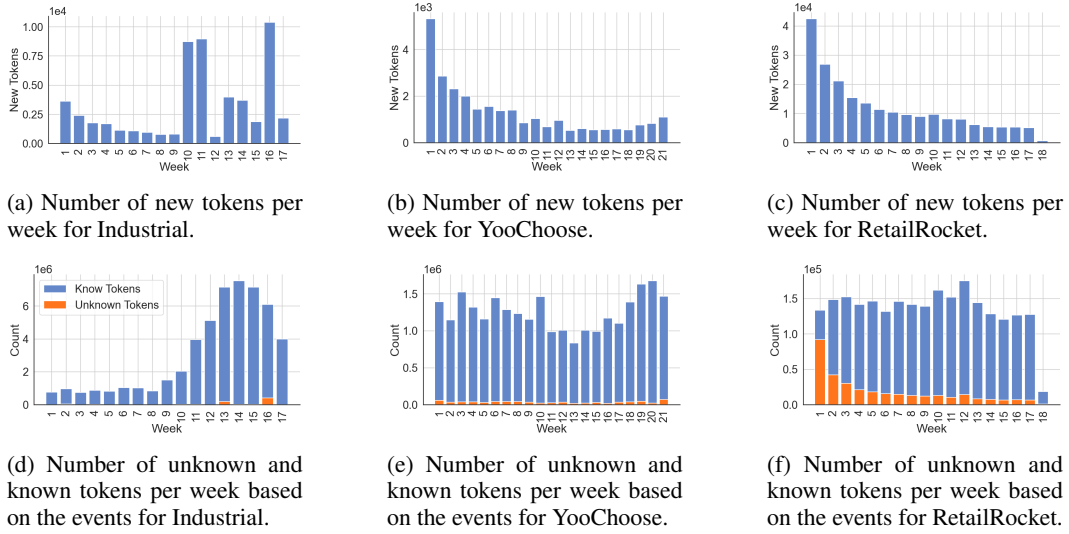


Figure 2: Token statistics over time across the three datasets.

We standardized all datasets into {session ID, interaction object, timestamp}. Sessions in RetailRocket were reconstructed by grouping consecutive user interactions with a gap of at most 30 minutes. Sessions shorter than three interactions or longer than the 99.5th percentile were removed. For the industrial dataset, interaction objects were extracted from URLs with query parameters removed to compactify the vocabulary. Customer IDs were used only to derive binary purchase labels (purchase vs. non-purchase) and then discarded.

As reported by Lo et al. (2016), consumer behavior exhibits weekly rhythms such as weekend peaks and payday effects, making this timeframe both behaviorally representative and practically relevant. Therefore, we sliced all datasets into *weekly increments* to simulate incremental learning. Weekly slicing provides a balance between granularity and stability: each batch contains sufficient inter-

¹Available at <https://www.kaggle.com/datasets/chadgostopp/recsys-challenge-2015>

²<https://recsys.acm.org/recsys15/challenge/>

³Available at <https://www.kaggle.com/datasets/RetailRocket/ecommerce-dataset>

action volume to train embeddings and task-models while avoiding the computational overhead of daily retraining.

5 EXPERIMENTAL SETUP

The objective of our experiments is to evaluate whether the proposed LLE method can effectively handle the dynamics of e-commerce characterized by continuously emerging tokens and shifting distributions while retaining past knowledge.

All experiments follow a weekly incremental learning protocol as highlighted in algorithm 1. At each step t , the embedding is updated with new data and then evaluated on the subsequent week $t + 1$ in a downstream prediction task. For the experiments, we follow a decoupled design: embeddings are first trained in a self-supervised fashion using context prediction with a cross-entropy loss (akin to Skip-Gram (Mikolov et al., 2013)), and subsequently fed into a Transformer classifier (two layers, adaptive attention heads) for session-level purchase prediction. This setup isolates the contribution of the embedding space itself, avoiding confounding effects of end-to-end fine-tuning. The Transformer model is retrained from scratch at each step to ensure that evaluation reflects only the quality of the updated embeddings, preventing any residual knowledge from earlier models from influencing performance.

Algorithm 1 Dynamic Embedding and Task Model Training

```

1: for each new data iteration  $D^{(t)}$  do
2:   Incremental embedding update
3:   Adapt embedding dimensionality
4:   Train embedding with CL
5:   Train task model with updated embedding
6:   Evaluate task performance on  $D^{(t+1)}$ 
7: end for

```

While the embedding and classifier are trained separately, the embedding dimensionality itself remains a design choice that strongly impacts generalization and efficiency. Previous work has explored different strategies for adapting embedding size, ranging from rule-based schedules to validation-driven criteria. In our experiments, we adopt a simple heuristic based on the gap between training and validation losses. Let $\mathcal{L}_{\text{train}}^{(e)}$ and $\mathcal{L}_{\text{val}}^{(e)}$ denote the training and validation losses at epoch e . We diagnose overfitting when

$$\frac{\mathcal{L}_{\text{val}}^{(e)}}{\mathcal{L}_{\text{train}}^{(e)}} > \gamma \quad \text{and} \quad \mathcal{L}_{\text{val}}^{(e)} > \mathcal{L}_{\text{val}}^{(e-1)},$$

with divergence threshold $\gamma > 1$. If this condition holds for multiple epochs, the embedding dimensionality d is iteratively reduced by step size $\Delta d \in \mathbb{N}$, $0 < \Delta d < d$ until validation loss improves:

$$d' = d - \Delta d.$$

Conversely, if validation loss continues to decrease without divergence, we increase dimensionality:

$$d' = d + \Delta d,$$

until $\mathcal{L}_{\text{val}}^{(e)}$ no longer improves.

For evaluation, we report AUC scores for each week, training time, and analyze the performance–efficiency trade-off. To measure the impact of the framework, we compare LLE against two reference setups:

- **Baseline 1:** Embedding trained from scratch at each week using only that week’s data.
- **Baseline 2 (Upper bound):** Embedding retrained from scratch at each week but using data from all weeks up to the current point (not feasible in practice but provides an oracle-like reference).

To disentangle the effect of each component in LLE, we conduct a systematic ablation study that varies the following factors:

- **New Token Initialization:** none (retraining), random, unknown-token, or average-based initialization.
- **Dimensionality Adaptation:** with or without adaptive dimensionality adjustment.
- **Continual Learning:** no CL, single regularization-based methods (EWC (Kirkpatrick et al., 2017), MAS (Aljundi et al., 2018), LwF (Li & Hoiem, 2016)), and their combinations.

We focus on regularization-based CL since memory-based replay approaches, while effective (Shin et al., 2017), are limited by scalability, privacy (e.g., GDPR (European-Parliament, 2016)), and storage costs in practice.

All experiments are implemented in PyTorch using AdamW (learning rate 1×10^{-3}) with a batch size of 128. Each variant is executed independently ten times, and we report averaged results for robust comparison. Unless stated otherwise, the standard LLE configuration employs average-based initialization for new tokens and MAS for CL. Experiments are conducted on a server with 96 Xeon Platinum 8168 CPU cores (2.7GHz) and a single Nvidia Tesla V100 GPU. Although we evaluate LLE in a decoupled setting for clarity, the framework is directly compatible with end-to-end architectures and can be integrated as a drop-in replacement for standard embedding layers.

6 RESULTS AND DISCUSSION

Figure 3 shows that LLE consistently outperforms the lower-bound baseline and approaches the upper bound across all datasets. These results indicate that LLE can retain knowledge without access to historical data, which is often limited in practice due to storage costs or privacy constraints. In this evaluation, LLE is applied with incremental updates, average-based initialization for new tokens, and MAS for continual learning, a configuration that yielded a stable and adaptable performance across datasets.

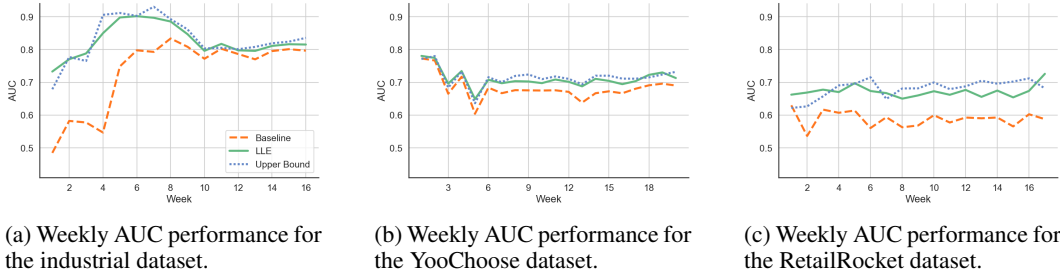


Figure 3: AUC results over time for LLE compared to Baseline (lower bound) and Upper Bound across all three datasets.

On the Industrial dataset, LLE clearly outperforms the lower baseline in the first ten weeks and nearly matches the upper bound. After week 10, however, the performance gap narrows as interactions increase from about 1M to over 6M by week 14 (cf. Figure 2), rendering recent data dominant and past information less relevant.

For both YooChoose and RetailRocket, LLE consistently improves over the lower baseline and remains close to the upper bound, despite very different data characteristics. This shows that even under contrasting conditions, LLE effectively leverages past knowledge and adapts to new data.

Figure 4 compares training time. The baseline is the fastest but also the least effective. The upper bound achieves higher accuracy but is the slowest, as it requires full retraining each week with training time scaling linearly in the added data. LLE introduces only modest overhead compared to the baseline while avoiding the escalating costs of the upper bound.

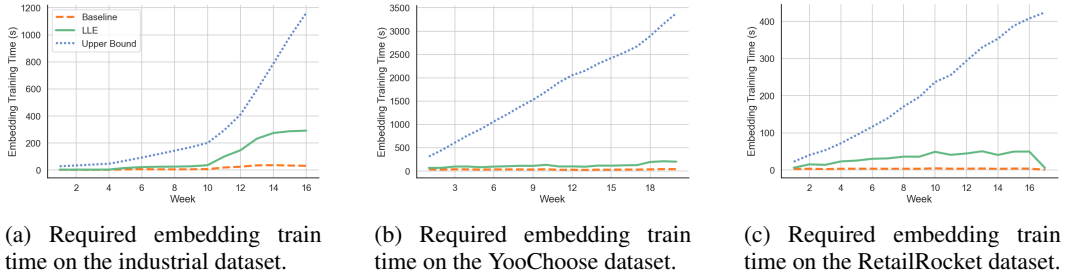


Figure 4: Training time per week for LLE compared to Baseline and Upper Bound across all datasets.

The findings demonstrate that LLE achieves a favorable balance between adaptability and efficiency. By incrementally integrating new tokens and stabilizing representations through CL regularization, LLE adapts to evolving data distributions while mitigating forgetting, without the need to retrain on the full history. This design enables LLE to approach the accuracy of the upper bound while maintaining computational costs closer to the lower bound.

7 ABLATION STUDY: COMPONENT CONTRIBUTIONS OF LLE

We evaluate the contribution of initialization strategies and CL. The results indicate that the overall design of LLE provides the primary performance gains, while the specific choice of initialization scheme (e.g., unknown-token vs. average) or CL variant (e.g., EWC, MAS, LwF, etc.) acts more as a hyperparameter that needs to be tuned to the needs of a particular use case.

For instance, in YooChoose, the top-performing variants differ only in initialization or CL choice but all achieve mean AUCs around 0.7 over all weeks, which is about +0.07 over the baseline. These findings indicate that the benefit comes from the LLE framework itself rather than from a specific initialization or CL variant.

To further explore this, Figure 5 shows the weekly AUC difference from the baseline across datasets. Thereby, shaded regions indicate variance across runs. We group the approaches into three categories: (1) *Both* incremental initialization and CL (orange), (2) *Only incremental* initialization (blue), and (3) *Only CL* (green).

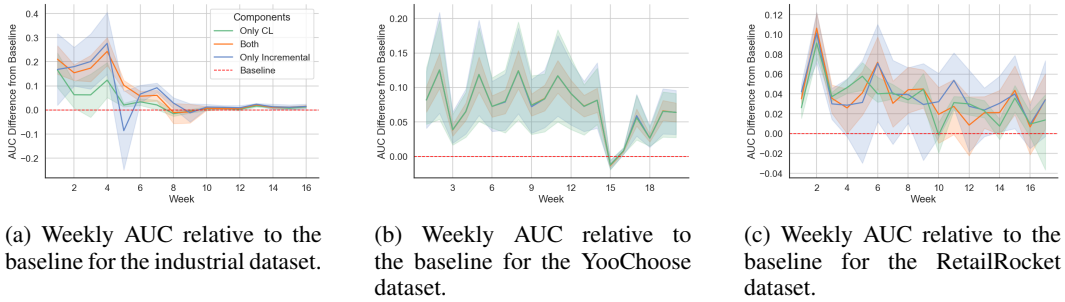


Figure 5: Ablation results of using different LLE components on the three datasets.

In the industrial dataset, incremental strategies adapt quickly but show higher variance, CL stabilizes performance but reacts more slowly, and combining both provides the most consistent improvements. The variance across methods decreases after week 10, coinciding with the sharp rise in interaction volumes (cf. Figure 2). This aligns with the earlier performance results (Figure 3a), where the gap between LLE and the baselines narrows once abundant new data dominates. The ablation thus reinforces that when fresh data overwhelms prior knowledge, different strategies converge in performance.

In YooChoose, combining incremental strategies and CL produces smoother performance than either alone, reflecting the benefit of balancing rapid adaptation with knowledge retention. Both mechanisms help, but their combination yields the most stable AUC improvements week by week.

In RetailRocket, the embedding dimensionality is the only one that undergoes shrinkage (cf. Figure 6), which forces the model to discard past information. This explains the high variance of the incremental-only configuration, which at times even falls below the baseline, and the overall limited performance gains. By contrast, CL-only shows the lowest variance and most stable results. This highlights the role of CL in aligning new embeddings with the information of the old space, thereby mitigating the loss induced by dimensionality reduction and stabilizing performance despite the challenging dynamics of this dataset.

As aforementioned, embeddings constitute massive tables, and compact representations are essential for memory efficiency. Adaptive dimensionality plays a complementary role: while larger embeddings may contain all the information needed to solve a task, they are often redundant and costly. Figure 6 illustrates how LLE dynamically regulates capacity across datasets. In RetailRocket, dimensions shrink from ~ 28 to ~ 13 , indicating that too much capacity was initially allocated, likely due to the influx of new tokens requiring richer representations, which later became redundant and was compressed to prevent overfitting. In the industrial dataset, the opposite pattern emerges. Dimensions start small, but as the event volume grows after week 10, capacity is gradually expanded to avoid under-parameterization. YooChoose remains consistently high (32–35) to accommodate its large and diverse vocabulary. These dynamics show that adaptive resizing does not directly improve AUC but ensures efficiency by allocating capacity where needed and compressing when redundancy dominates.

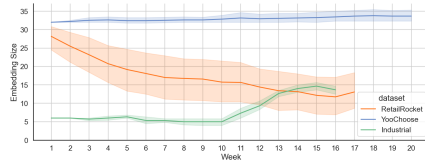


Figure 6: Average embedding size over time.

8 CONCLUSION

We introduced Lifelong-Learning Embeddings, a drop-in framework that incrementally integrates new tokens, adapts embedding dimensionality to balance capacity and compactness, and uses CL to retain prior semantics. Across an industrial dataset and two public benchmarks, LLE consistently outperforms training from scratch and approaches the upper bound without requiring access to the full history. The ablation study shows that the specific choice of initialization or CL (EWC/MAS/LwF/combo) has a minor impact on accuracy. As long as some LLE mechanism is applied, gains are robust. Whereby, dimensionality adaptation contributes primarily to efficiency.

9 RESEARCH LIMITATIONS AND FUTURE DIRECTIONS

We focused on a weekly, decoupled protocol to isolate representation quality. The behavior under fully end-to-end training or alternative update cadences (e.g., daily or event-driven) remains unexplored in this research. Furthermore, new-token initialization relied on lightweight schemes (random, unknown, average). Richer metadata- or structure-aware approaches could further improve cold-start handling. Our ablation study suggests that the main benefit comes from applying incremental or CL mechanism at all. This highlights LLE’s strength in flexibility, but also indicates that gains may plateau without more specialized strategies.

Future work should therefore extend LLE to richer initialization strategies, investigate task-aware or streaming update schedules, and examine integration into end-to-end training regimes.

REPRODUCIBILITY STATEMENT

We provide the full implementation of LLE and all experiments in the supplementary material, along with a lightweight Jupyter notebook that highlights the core components of LLE. The notebook can be run on a standard local machine without requiring high-end computational resources.

Publicly available datasets (YooChoose and RetailRocket) can be used to reproduce our results (cf. Section Datasets and Preprocessing). The industrial dataset cannot be released due to confidentiality, but it was included in the experiments to better reflect real-world dynamics.

Upon acceptance, we will release the complete implementation on GitHub to support further use, improvement, and extension by the community.

GENAI USAGE DISCLOSURE

In the preparation of this manuscript, the authors used generative AI tools, including ChatGPT (OpenAI), DeepL Write, and Grammarly, exclusively for language polishing and grammar correction. These tools were not used for generating content, data analysis, experimental design, or drawing conclusions.

The authors affirm that all substantive ideas, algorithms, experimental setups, and analyses presented in this work are original and authored by the listed contributors. We take full responsibility for the integrity, accuracy, and originality of the published content.

REFERENCES

- Sami Abu-El-Haija, Bryan Perozzi, Rami Al-Rfou, and Alex Alemi. Watch your step: learning node embeddings via graph attention. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, pp. 9198–9208, Red Hook, NY, USA, 2018. Curran Associates Inc.
- Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part III*, pp. 144–161, Berlin, Heidelberg, 2018. Springer-Verlag. ISBN 978-3-030-01218-2. doi: 10.1007/978-3-030-01219-9_9. URL https://doi.org/10.1007/978-3-030-01219-9_9.
- Miguel Alves Gomes, Richard Meyes, Philipp Meisen, and Tobias Meisen. Will this online shopping session succeed? predicting customer’s purchase intention using embeddings. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management, CIKM ’22*, pp. 3707–3716, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9236-5/22/10. doi: 10.1145/3511808.3557127. URL <https://doi.org/10.1145/3511808.3557127>.
- Patrik Berger and Michal Kompan. User modeling for churn prediction in e-commerce. *IEEE Intelligent Systems*, 34(2):44–52, 2019. doi: 10.1109/MIS.2019.2895788.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- Chi Chen, Hui Chen, Kangzhi Zhao, Junsheng Zhou, Li He, Hongbo Deng, Jian Xu, Bo Zheng, Yong Zhang, and Chunxiao Xing. Extr: Click-through rate prediction with externalities in e-commerce sponsored search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2732–2740, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3534678.3539053.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML’20*. JMLR.org, 2020.

- Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. Wide & deep learning for recommender systems. In *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, DLRS 2016, pp. 7–10, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450347952. URL <https://doi.org/10.1145/2988450.2988454>.
- Mert Demirer, Diego J Jiménez Hernández, Dean Li, and Sida Peng. Data, privacy laws and firm production: Evidence from the gdpr. Working Paper 32146, National Bureau of Economic Research, February 2024. URL <http://www.nber.org/papers/w32146>.
- 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: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.
- Tianning Dong, Yan Sun, and Faming Liang. Deep network embedding with dimension selection. *Neural Networks*, 179:106512, 2024. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2024.106512>. URL <https://www.sciencedirect.com/science/article/pii/S0893608024004362>.
- Ramazan Esmeli, Mohamed Bader-El-Den, and Hassana Abdullahi. Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets*, 31(3):697–715, 2021.
- European-Parliament. Regulation (eu) 2016/679 of the european parliament and of the council, 2016. URL <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679>.
- Cécile Ferrera and Eowyn Kessedjian. Evolution of e-commerce and global marketing. *International Journal of Technology for Business (IJTB)*, 1(1):33–38, 2019.
- Martin Fridrich and Petr Dostál. User churn model in e-commerce retail. *Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration*, 30(1), 2022.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.
- Guojing Huang, Qingliang Chen, and Congjian Deng. A new click-through rates prediction model based on deep&cross network. *Algorithms*, 13(12):342, 2020. doi: 10.3390/a13120342.
- Dennis R Kiker. Defensible data deletion: A practical approach to reducing cost and managing risk associated with expanding enterprise data. *Richmond Journal of Law & Technology*, 20(2):6, 2014.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL <https://www.pnas.org/doi/abs/10.1073/pnas.1611835114>.
- Zhizhong Li and Derek Hoiem. Learning without forgetting. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 614–629, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.
- Weiqliang Lin, Natasa Milic-Frayling, Ke Zhou, and Eugene Ch’ng. Predicting outcomes of active sessions using multi-action motifs. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 9–17, New York, NY, USA, 2019. Association for Computing Machinery. doi: 10.1145/3350546.3352495.

- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Caroline Lo, Dan Frankowski, and Jure Leskovec. Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16*, pp. 531–540, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939729. URL <https://doi.org/10.1145/2939672.2939729>.
- Qinyi Luo, Penghan Wang, Wei Zhang, Fan Lai, Jiachen Mao, Xiaohan Wei, Jun Song, Wei-Yu Tsai, Yuxi Hu, and Xuehai Qian. FIITED: Fine-grained embedding dimension optimization during training for recommender systems, 2024. URL <https://openreview.net/forum?id=gDDW5zMKFe>.
- Nicolas Menand and C. Seshadhri. Link prediction using low-dimensional node embeddings: The measurement problem. *Proceedings of the National Academy of Sciences*, 121(8):e2312527121, 2024. doi: 10.1073/pnas.2312527121. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2312527121>.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.
- Chanyoung Park, Donghyun Kim, Jinoh Oh, and Hwanjo Yu. Predicting user purchase in e-commerce by comprehensive feature engineering and decision boundary focused under-sampling. In *Proceedings of the 2015 International ACM Recommender Systems Challenge, RecSys '15 Challenge*, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336659. doi: 10.1145/2813448.2813517. URL <https://doi.org/10.1145/2813448.2813517>.
- Yunke Qu, Liang Qu, Tong Chen, Xiangyu Zhao, Quoc Viet Hung Nguyen, and Hongzhi Yin. Scalable dynamic embedding size search for streaming recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM '24*, pp. 1941–1950, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679638. URL <https://doi.org/10.1145/3627673.3679638>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Peter Romov and Evgeny Sokolov. Recsys challenge 2015: Ensemble learning with categorical features. In *Proceedings of the 2015 International ACM Recommender Systems Challenge, RecSys '15 Challenge*, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336659. doi: 10.1145/2813448.2813510. URL <https://doi.org/10.1145/2813448.2813510>.
- E. Rossi, B. Chamberlain, F. Frasca, D. Eynard, F. Monti, and M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*, 2020. URL <https://arxiv.org/abs/2006.10637>.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.
- Humphrey Sheil, Omer Rana, and Ronan Reilly. Predicting purchasing intent: automatic feature learning using recurrent neural networks, 2018a. URL <http://arxiv.org/abs/1807.08207>.
- Humphrey Sheil, Omer Rana, and Ronan Reilly. Predicting purchasing intent: automatic feature learning using recurrent neural networks. *arXiv preprint arXiv:1807.08207*, 2018b.

- Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pp. 2994–3003, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Madiraju Srilakshmi, Gourab Chowdhury, and Sudeshna Sarkar. Two-stage system using item features for next-item recommendation. *Intelligent Systems With Applications*, 14:200070, 2022. ISSN 2667-3053. doi: <https://doi.org/10.1016/j.iswa.2022.200070>. URL <https://www.sciencedirect.com/science/article/pii/S2667305322000114>.
- Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19*, pp. 1441–1450, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357895. URL <https://doi.org/10.1145/3357384.3357895>.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, NIPS'14*, pp. 3104–3112, Cambridge, MA, USA, 2014. MIT Press.
- Paul P. Tallon. Understanding the dynamics of information management costs. *Commun. ACM*, 53(5):121–125, May 2010. ISSN 0001-0782. doi: 10.1145/1735223.1735253. URL <https://doi.org/10.1145/1735223.1735253>.
- Hasan Tercan, Christian Bitter, Todd Bodnar, Philipp Meisen, and Tobias Meisen. Evaluating a session-based recommender system using prod2vec in a commercial application. In *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 1: ICEIS*, pp. 610–617. INSTICC, SciTePress, 2021. ISBN 978-989-758-509-8. doi: 10.5220/0010400706100617.
- Flavian Vasile, Elena Smirnova, and Alexis Conneau. Meta-prod2vec: Product embeddings using side-information for recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 225–232, New York, NY, USA, 2016. Association for Computing Machinery.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Kefan Wang, Hao Wang, Wei Guo, Yong Liu, Jianghao Lin, Defu Lian, and Enhong Chen. Dlf: Enhancing explicit-implicit interaction via dynamic low-order-aware fusion for ctr prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, pp. 2213–2223, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3729956. URL <https://doi.org/10.1145/3726302.3729956>.
- Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. Deep & cross network for ad click predictions. *ADKDD'17*, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351942. doi: 10.1145/3124749.3124754. URL <https://doi.org/10.1145/3124749.3124754>.
- K. Weimann and T.O.F. Conrad. Felrec: Efficient handling of item cold-start with dynamic representation learning in recommender systems. *International Journal of Data Science and Analytics*, 20: 2937–2950, 2025. doi: 10.1007/s41060-024-00635-5. URL <https://doi.org/10.1007/s41060-024-00635-5>.
- Zhiming Yang, Haining Gao, Dehong Gao, Luwei Yang, Libin Yang, Xiaoyan Cai, Wei Ning, and Guannan Zhang. Mlora: Multi-domain low-rank adaptive network for ctr prediction. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24*, pp. 287–297, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400705052. doi: 10.1145/3640457.3688134. URL <https://doi.org/10.1145/3640457.3688134>.

- Jiangwei Zeng, Yan Chen, Haiping Zhu, Feng Tian, Kaiyao Miao, Yu Liu, and Qinghua Zheng. User sequential behavior classification for click-through rate prediction. In *Database Systems for Advanced Applications. DASFAA 2020 International Workshops*, pp. 267–280. Springer, Berlin/Heidelberg, Germany, 2020. doi: 10.1007/978-3-030-59413-8_22.
- Hengyu Zhang, Junwei Pan, Dapeng Liu, Jie Jiang, and Xiu Li. Deep pattern network for click-through rate prediction. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’24*, pp. 1189–1199, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657777. URL <https://doi.org/10.1145/3626772.3657777>.
- Xue Zhang, Fusen Guo, Tao Chen, Lei Pan, Gleb Beliakov, and Jianzhang Wu. A brief survey of machine learning and deep learning techniques for e-commerce research. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4):2188–2216, 2023. ISSN 0718-1876. doi: 10.3390/jtaer18040110. URL <https://www.mdpi.com/0718-1876/18/4/110>.
- Y. Zhang and et al. Cope: Modeling continuous propagation and evolution on interaction graph. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*. Association for Computing Machinery, 2021. doi: 10.1145/3459637.3482443.
- Yujia Zheng, Siyi Liu, Zekun Li, and Shu Wu. Cold-start sequential recommendation via meta learner. 35:4706–4713, May 2021. doi: 10.1609/aaai.v35i5.16601. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16601>.
- Chang Zhou, Jinze Bai, Junshuai Song, Xiaofei Liu, Zhengchao Zhao, Xiushi Chen, and Jun Gao. Atrank: An attention-based user behavior modeling framework for recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 4 2018. doi: 10.1609/aaai.v32i1.11618. URL <https://ojs.aaai.org/index.php/AAAI/article/view/11618>.
- Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5941–5948, 2019. doi: 10.1609/aaai.v33i01.33015941.