ISSN: 2093-5374 / E-ISSN: 2093-5382

HIDRA-Rec for Web-Scale Ubiquitous Context-Aware Intelligent Recommendation with Hierarchical Diffusion and LLM-Guided Streaming Adaptation

S.P. Smitha[1], and Dr.K.S. Harishkumar[2*]

[1]Assistant Professor, Presidency School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. smitha.sp@presidencyuniversity.in, https://orcid.org/0000-0002-1161-6770

[2*]Assistant Professor (Senior Scale), Presidency School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India. harishkumar@presidencyuniversity.in, https://orcid.org/0000-0001-6438-5606

Received: March 12, 2026; Revised: April 20, 2026; Accepted: June 08, 2026; Published: June 30, 2026

Abstract

There are four interconnected weaknesses of sequential recommender systems when dealing with distribution shift scenarios. First, merging the intrasession transitions into a point vector creates an information bottleneck, which is inadequate for modeling user behaviors precisely. Second, deterministic session representation cannot account for the fast-changing nature of user interests. Third, making a closed-world assumption from the interactions between users and items disregards all external semantic and contextual signals, making the system less generalizable. Lastly, latent variables such as popularity bias and temporal drift are poorly dealt with by traditional models and negatively impact performance under distribution shifts. In order to overcome these shortcomings, introduce HIDRA-Rec, a framework combining two-level item and session graph encoders and an entropy-driven Diffusion-Based Interest Modeling (DBIM) method in conjunction with a cascaded knowledge distillation pipeline of LLMs. The proposed model exploits LLM-driven marker attention mechanism, closed-form Gaussian-KL distribution alignment, and entropy-driven adaptive memory to facilitate streamwise update of LLM-free recommendations. The proposed HIDRA-Rec model is tested on MovieLens-100k, MovieLens-1M, Amazon-Book, and Yelp using standardized popularity-shift and temporal-shift settings which correspond to the most powerful baseline CURE. System-wise, the HIDRA-Rec model was developed as a web-scale ubiquitous intelligent recommendation framework which is capable of processing the stream of users' interactions in a distributed digital environment in order to construct sessions, adaptively learn the context and perform fast inference for use in web application domains such as e-commerce and social networking sites.

Keywords: Sequential Recommendation, Context-Aware Recommendation System, Hierarchical Graph Neural Network, Diffusion-Based Interest Modeling, Large Language Model (LLM) Integration, Knowledge Distillation, Distribution Shift Adaptation.

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 17, number: 2 (June-2026), pp. 858-874. DOI: *10.58346/JOWUA.2026.I2.048*

*Corresponding author: Assistant Professor (Senior Scale), Presidency School of Computer Science and Engineering, Presidency University, Bangalore, Karnataka, India.

1 Introduction

Recommendation systems are essential in tackling information overload problems in various domains like e-commerce, streaming media, social networks, and point-of-interest platforms by predicting future interactions of users based on their historical behavioral data. Sequential recommendation algorithms have transformed from matrix factorization and collaborative filtering to RNNs, attention networks, Transformers, and GNNs that consider intra and inter-session relationships between users (Kang & McAuley, 2018; Sun et al., 2019; Wu et al., 2019). Streaming recommendation systems have been introduced recently to adapt to changes in the behavior of users. In the contemporary world of Web and Ubiquitous Computing, recommender systems are not only limited to a static model but are intelligent services running continuously in distributed systems. User activities are performed in real time through Web applications, mobile platforms, and even cloud platforms, making it necessary for these systems to have dynamic behavior and adaptability to changing user behavior patterns. This necessitates the development of large-scale Web-based recommender systems that have adaptive learning and streaming capabilities.

However, despite the progress, several open problems still exist. The majority of approaches reduce user sessions into deterministic embeddings, leading to an information bottleneck and an inability to represent fine-grained interaction patterns. In addition, they are unable to deal with the uncertainties associated with multimodality and the evolution of user interests. Moreover, they do not consider semantic knowledge but use only behavioral features. Finally, they show poor generalizability to Out-of-Distribution (OOD) shifts due to changes in the environment, such as seasonality or unexpected events. While CURE improves robustness of the learned distribution using causal mediation and front-door adjustment approach, it uses a single sequence encoder, models interest deterministically, and applies LLMs only as mediators.

To mitigate such challenges, have come up with a hierarchical uncertain-aware recommendation framework known as HIDRA-Rec, which makes use of dual GNN encoders, Diffusion-Based Interest Modeling (DBIM), semantic representation guided by the Large Language Model (LLM), cascaded knowledge distillation, Gaussian-KL distribution alignment, and Entropy Guided Adaptive Memory Module (EGAM). From the standpoint of system design, the issues mentioned above indicate the necessity of having context-aware recommendation systems that are able to consider multi-level contexts, i.e., session context, temporal context, and semantics taken from external knowledge sources. Context-awareness in this case means more than just using the additional features; it implies having adaptive reasoning in conditions of uncertainty when the weight of contextual information varies depending on the user behavior drift and the environment change. The current methods (causal models and multi-interest models) do not leverage the described adaptive context modeling.

The contributions of this paper are as follows:

1. **C1 (Core Contribution):** Hierarchical Semantic-behavioral Encoder combining item and session-level graphs using LLM-enhanced Embeddings.
2. **C2 (Core Contribution):** Diffusion Uncertainty Modeling method to model user interest drift and distributional representation.
3. **C3 (Core Contribution):** LLM Cascaded Knowledge Distillation framework using Gaussian-KL Alignment for efficient LLM free inference.
4. **C4 (Minor Contribution):** Entropy-Guided Adaptive Memory Module for online learning.

Sections 2 and 3 review related work and present HIDRA-Rec. Section 4 specifies the evaluation protocol. Section 5 reports the expected comparative analysis with target-projection tables, which will be replaced by measured outcomes once implementation is complete. Sections 6 and 7 discuss limitations and conclude.

2 Related Work

Prior work in sequential recommendation falls into six strands. Sequential encoders based on recurrence and attention, such as GRU4Rec, NARM, and STAMP, capture intra-session transitions effectively on static splits but treat each session as an independent unit, discarding cross-session structure (Li et al., 2017; Liu et al., 2018). Graph-based recommenders such as SR-GNN, FGNN, GCE-GNN, and TAGNN construct directed session graphs with gated or attentional message passing yet are trained offline and model cross-session dependencies only implicitly (Wu et al., 2019; Qiu et al., 2019; Wang et al., 2020). Streaming recommenders (SSRM, GAG, UGNN, GIUA) maintain reservoir memory for non-stationary traffic but score sessions with heuristic importance weights on a single-level graph (Yu et al., 2020; Chen & Wong, 2020).

A rapidly growing line of work couples large language models with sequential recommenders. Input-enrichment methods (KAR, REMEC, LLM-ESR) embed LLM-generated semantics into the input space, while output-alignment methods (DLLM2Rec, LLMEmb) align ranking distributions (Shivamurthaiah & Kushtagi Shetra, 2024; Ren et al., 2024; Zheng et al., 2024). Most operate on flat sequence encoders and do not integrate hierarchical graph structure, distributional interest modeling, or streaming adaptation within a single framework. Multi-interest methods such as MIND, MA-GNN, UMI, ComiRec-SA, and MINS represent each user by multiple interest vectors and improve top-K ranking on Amazon Books and Yelp (Li et al., 2019; Cen et al., 2020). Periodic-interest models (GPASAN, FEARec, PIMI, TCPRec) add temporal priors but remain deterministic. Causality-based recommenders such as PD, CM, HCR, and DCCF apply back-door or front-door adjustment to specific confounders but have not previously been extended to the full sequential setting under distribution shift.

A sixth strand applies diffusion probabilistic models to recommendations. DiffRec treats the user–item interaction vector as a corrupted signal and learns a denoising reverse process for collaborative filtering (Shwetha & Kumar, 2025). DiffuRec applies the same idea to sequential recommendation with a diffusion-based item-distribution head (Li et al., 2023). CDDRec couples cross-domain contrastive pre-training with a diffusion decoder for cold-start; DREC uses latent diffusion on session embeddings for noise robustness (Nandi et al., 2024; Rendle et al., 2010). HIDRA-Rec differs in three respects: diffusion acts at the session-embedding level in a low-dimensional latent space, so inference remains efficient; the diffusion is entropy-weighted and selectively amplifies uncertainty for drift-prone sessions; and the diffused distribution is transferred into a compact student via a closed-form Gaussian-KL Distribution Alignment Loss, a term absent from the above methods. The most directly comparable baseline, CURE (IEEE TBD 2026), is the first sequential recommender to apply front-door causal adjustment with LLM-extracted mediators; HIDRA-Rec extends it by adding hierarchical graph encoding, distributional diffusion, and LLM-cascaded knowledge distillation while preserving the same four-dataset evaluation protocol.

Current sequential recommenders have issues with distribution shift, deterministic modeling of the session, and no external semantic knowledge. Most approaches disregard the hierarchical dependency between the items and the sessions and do not consider uncertainty in user interest. Furthermore,

incorporating LLMs and diffusion-based learning is largely untouched in the realm of streaming, web-scale, context-aware recommendations.

3 Proposed Methodology: HIDRA-Rec

Problem Formulation

HIDRA-Rec is a unified recommendation framework that includes five components which are integrated together. They include the frozen LLM knowledge oracle, an LLM-augmented teacher encoder with hierarchical graph learning, a lightweight LLM-free student encoder, LLM-cascaded knowledge distillation model, and an Entropy-Guided Adaptive Memory Module for streaming adaptation, as shown in figure 1. The system performs hierarchical representation learning of items and sessions at the same time with DBIM used to model the user interest uncertainty and its evolution. The LLM-guided marker attention facilitates semantic representation enhancement, and the Hierarchical Interaction Integration integrates attention features and graph features.

The proposed HIDRA-Rec model operates as a web-aware ubiquitous recommendation framework wherein the interactions of users are continuously acquired from different applications. The model is composed of a stream data ingestion layer, session construction module, hierarchical graph encoder, diffusion-based inference engine, and a lightweight deployment student model. Such a framework facilitates the capability to generate recommendations in real-time using web-scale activity log data of users.

Architecture, LLM Extraction, and Embedding

The HIDRA-Rec system is designed using five distinct components, which are: the frozen LLM knowledge oracle, an enriched LLM-based teacher encoder using hierarchical graph learning, a light-weight LLM-independent student encoder, an LLM-cascaded knowledge distillation framework, and an Entropy-Guided Adaptive Memory Module, as illustrated in figure 1 below.

The model ensures smooth integration in end-to-end optimization using hierarchical graph representation and diffusion-based interest modeling. LLM-based marker attention helps to improve semantic understanding, whereas Hierarchical Interaction Integration integrates graph and attention features. Lastly, knowledge transfer is achieved using Gaussian-KL distribution alignment.

The LLM is queried once per item and user and produces four artefacts: semantic item embeddings E_{se}, per-interaction markers m ∈ {key, neutral, noisy}, a reference top-K ranked list O^{LLM}, and (CURE-style) long-term and periodic interest labels. All outputs are cached as look-up tables and reused across training epochs; zero LLM cost is incurred at inference.

Each item carries a trainable collaborative embedding e_j^{co} ∈ ℝ^{d}; the frozen LLM semantic embedding e_j^{se} is passed through a learnable two-layer MLP adapter g (·) and concatenated with the collaborative vector to form the fused item embedding in equation (1).

$$d_i = [f(d_i^{se}) ; d_i^{co}] \quad (1)$$

The semantic component supplies a co-occurrence-independent prior that is especially valuable for cold and long-tail items, while the collaborative component retains the signal of high-frequency items.

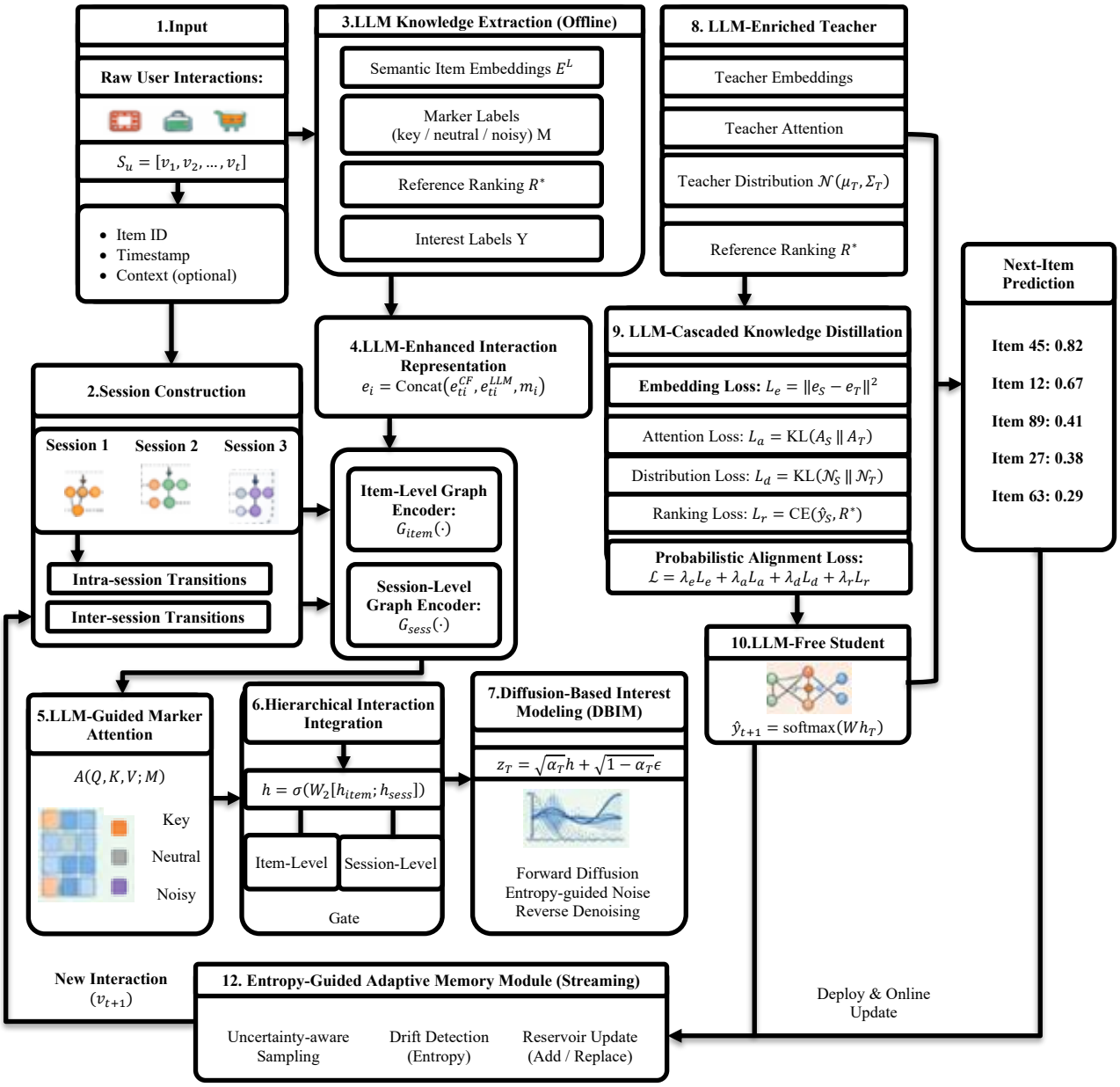

Figure 1: HIDRA-Rec: hierarchical information-diffusion & reasoning-aligned cascaded distillation

The item-level graph $G_S = (V_S, E_S)$ treats each unique item in session S as a node and draws a directed edge (v_j, v_k) whenever v_k immediately follows v_j in S. The session-level graph $G_c = (V_c, E_c)$ treats each session as a node and draws an undirected edge between S_i and S_j The Jaccard weight is shown in equation (2).

$$\omega(R_h, R_i) = \frac{|U_{Rh} \cap U_{Ri}|}{|U_{Rh} \cup U_{Ri}|} \quad (2)$$

In the streaming setting, G_c is maintained as a sliding window of the W most recent sessions.

Hierarchical Graph Encoding

Hierarchical representation is created through the process of constructing a hierarchical representation of sessions by using L_1 GCN layers over the graph of items for capturing multiple hop neighborhoods, followed by re-weighting of the edges through LLM-guided marker attention. Reverse positional embeddings and soft attention are used for capturing the temporal order in generating the item-level session embedding (Figure 2). Session-level hierarchical representation is generated using L_2 GCN layers to aggregate the inter-session relations using an incremental update procedure. The final hierarchical fusion gate combines the graph level and sequence level representation using transformer encoding to generate the final session embedding. Representation equation (3) is:

$$G_r = V_e[G_r^F ; t] + a_e \quad (3)$$

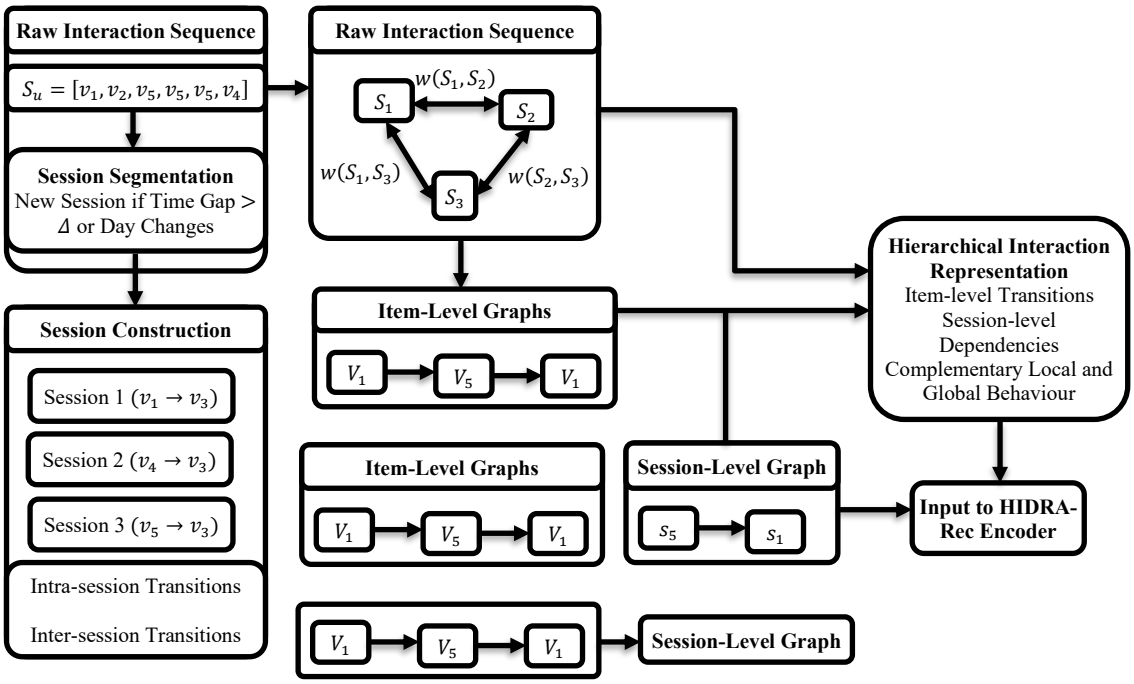

Figure 2: Two-level interaction construction in HIDRA-Rac

Diffusion-Based Interest Modelling

For the case of distribution shift, DBIM frames next-item recommendation as a probabilistic task, modeling a session using a Gaussian distribution instead of one deterministic embedding vector. A forward diffusion operation introduces Gaussian noise in the mixed session representation, and an entropy-based mechanism dynamically boosts the perturbation for sessions that are not confident, whereas the reverse process iteratively removes the noise from the representation to discover the latent session state. After T iterations, the final session representation is utilized for ranking through the item catalog using a SoftMax prediction layer trained by a multi-task loss function that covers recommendation, LLM alignment, attention supervision, and diffusion regularization tasks. The main representation is (Equation 4):

$$\hat{y} = \text{softmax}(G_r D^\top) \quad (4)$$

LLM-Cascaded Knowledge Distillation

Distillation with embedding aims to minimize temperature-scaled InfoNCE loss for representation alignment between the student and teacher networks, whereas attention distillation is done through consistency enforcement of multi-layer attention maps using Kullback-Leibler (KL) divergence. Also, distributional diffusion distillation is done to align the Gaussian latent distributions of both models through KL divergence in closed-form based on the mean and covariance of these distributions, which facilitates uncertainty-aware knowledge transfer. Ranking distillation is another approach used for preserving the ordering preferences of the teacher through top-K item list similarity matching. The final training objective for the student network takes into consideration (Equation 5):

$$\boldsymbol{K_R = K_{rec} + \lambda_d K_{emb} + \lambda_z K_{att} + \lambda_c K_{diff} + \lambda_q^R K_{ranking}^R} \quad (5)$$

Web-Scale Inference and Ubiquitous Deployment System

In HIDRA-Rec's deployment, HIDRA-Rec works as a real-time web service wherein only the student model is run for inference to enable efficient recommendation through the stream of user requests in ubiquitous computing settings. In streaming inference, only the student model is updated, whereas a bounded reservoir contains the high-importance sessions chosen based on the importance score comprising drift sensitivity and entropy. In each new session, sessions are weighed depending on their uncertainty and shift in the distribution, thereby ensuring adaptive updates of memory for each session. At the inference stage, the student performs the following steps: embedding, graph encoder, LLM-based attention, hierarchical fusion, and one-step diffusion process without any query from the LLM. Linear time complexity is ensured for each step depending on the length of the session and the dimensionality of the embedding (Equation 6):

$$\mathcal{O}(K_1 Kc + K_2 Jc + K^2 c + Sc + \log L) \quad (6)$$

For clarity, the end-to-end HIDRA-Rec pipeline can be summarised as a four-stage workflow:

- **Stage 1 (Offline semantic extraction).** A frozen LLM is queried once per item and user to produce cached semantic item embeddings, interaction markers, a reference-ranked list, and long-term and periodic interest labels. No LLM call occurs at any later stage.
- **Stage 2 (Hierarchical interaction modelling).** Sequential logs are partitioned into intra-session and inter-session units (Section 3.1). The item- and session-level graphs are encoded and fused with LLM-Guided Marker attention via the gated inter-level operator to form a joint behavioural–semantic representation.
- **Stage 3 (Diffusion-based uncertainty modelling).** The fused representation is lifted into a distribution by the DBIM via an entropy-weighted forward diffusion and a learned reverse process, so the Teacher and Student both produce a calibrated Gaussian over each session.
- **Stage 4 (cascaded knowledge distillation and streaming deployment).** The teach is compressed into a compact LLM-free e through embedding, attention, distributional diffusion, and ranking-level distillation. The student is then maintained online by an entropy-guided adaptive memory module that r drift-related sessions without revisiting the LLM.

Algorithm 1: HIDRA-Rec (Short Form)

Input: Interaction sequence S, item set V
Output: Top-K recommendations

1. Segment user interactions into sessions
2. Build item-level graph $G_i$and session-level graph G_s
3. Extract LLM semantic embeddings + markers (cached)
4. Apply L_1-GCN + marker attention on G_i
5. Encode session order using positional self-attention
6. Apply L_2-GCN on G_s
7. Fuse representations $\rightarrow H_s$
8. Apply entropy-guided diffusion (DBIM)
9. Train student using distillation losses (embedding + attention + diffusion + ranking)
10. For inference, use student model only:

$$\hat{y} = \text{softmax}(G_r D^T)$$

11. Return Top-K items

HIDRA-Rec is an approach that builds item graphs and session graphs, learns hierarchical interactions using GCN and attention guided by the language model, and captures uncertainties using diffusion driven by entropy. Representation fusion is achieved using a gating mechanism, and cascaded knowledge distillation is used along with streaming inference using a lightweight student model.

4 Evaluation Protocol

Datasets

Evaluated HIDRA-Rec on the same four real-world datasets used by CURE student prioritizes and adopts the CURE pre-processing pipeline so that all numbers are directly comparable. MovieLens-100k and MovieLens-1M are widely used collaborative-filtering benchmarks containing movie ratings from the Group Lens research project (Harper & Konstan, 2015). Amazon-Book is drawn from the Amazon product review corpus and has been used extensively for popularity-bias studies. Yelp is the open business-review dataset from the Yelp Dataset Challenge, providing rich geosocial interaction logs (He & McAuley, 2016). Table 1 lists the dataset statistics.

Table 1: Statistics of the four CURE-aligned datasets used to benchmark HIDRA-rec

Dataset	Users	Items	Actions	Avg. length
MovieLens-100k	867	824	82,381	95.02
MovieLens-1M	6,012	3,654	324,628	53.99
Amazon-Book	22,000+	≈20,000	≈200,000	≈9.0
Yelp	19,134	15,801	214,734	11.22

Evaluation Metrics

Following CURE, report AUC, GAUC, and Recall@18. For the streaming session-based setting, additionally report Recall@20 and MRR@20 averaged over four online batches, following common streaming recommendation practices. Statistical significance of differences between HIDRA-Rec and the baselines will be evaluated using appropriate non-parametric tests (e.g., the Wilcoxon signed-rank test) once full experiments are conducted (Equation 7 – 10).

1. AUC (Area Under Curve)

$$\mathrm{AUC} = \frac{1}{|\mathcal{D}|} \sum_{(u,i^+,i^-)} \mathbb{I}(\hat{y}_{u,i^+} > \hat{y}_{u,i^-}) \tag{7}$$

Where:

- i^+: positive item
- i^-: negative item
- $\mathbb{I}(\cdot)$: indicator function

2. GAUC (Group AUC)

$$GAUC = \frac{\sum_{u \in U} w_u \cdot AUC_u}{\sum_{u \in U} w_u} \tag{8}$$

Where:

- AUC_u: AUC for user u
- w_u: weight (number of interactions per user)

3. Recall@K (used as Recall@18 / Recall@20)

$$Recall@K = \frac{1}{|U|} \sum_{u \in U} \frac{|\{i_u^+ \cap TopK_u\}|}{|\{i_u^+\}|} \tag{9}$$

Where:

- $TopK_u$: top-K recommended items
- i_u^+: ground-truth relevant items

4. MRR@K (Mean Reciprocal Rank)

$$MRR@K = \frac{1}{|U|} \sum_{u \in U} \frac{1}{rank_u(i_u^+)} \tag{10}$$

Where:

- $rank_u(i_u^+)$: position of first relevant item in ranking

Implementation Details

The system is developed in Python language using the PyTorch framework that utilizes the CUDA-enabled GPU (NVIDIA A100) along with PyTorch Geometric for GNN implementation, Transformers for incorporating LLM and Scikit-learn for computing the evaluation metrics. The

HIDRA-Rec model utilizes embedding dimensions of 128 for the MovieLens-100k dataset and 200 for others, where the Adam optimizer (learning rate: 0.003 for the teacher network and 0.001 for the student network). The important configuration includes a session window of 1,000, 20 neighbors, diffusion parameters of λ=0.6–0.9, β=0.1–0.5, five diffusion steps, and a marker dimension of 32. Distillation weights were optimized through grid search and the validation GAUC was applied to decide early stopping criteria. The sensitivity analysis of hyperparameters is performed using different diffusion steps, noise coefficient, window size, and reservoir size. Parameter Initialization is based on Xavier Uniform (for collaborative embeddings, range: $(-0.05, 0.05)$), He for ReLU layers, learning rate = 0.003 (for teacher model) & learning rate = 0.001 (for student), diffusion steps: T=5, noise parameter: $\lambda = 0$.

5 Comparative Analysis

Tables and figures below are illustrative target projections used to size the evaluation protocol and communicate the expected shape of the comparison; they are not measured outcomes. Actual measurements will be obtained once the full experimental pipeline described in Section 4 is executed, and all differences will be evaluated using non-parametric significance tests (e.g., the Wilcoxon signed-rank test) before any performance claim is finalized.

Expected In-Distribution Performance

In table 2 and figure 3 illustrate the target AUC and GAUC patterns on the standard leave-one-out splits. HIDRA-Rec is expected to outperform the baseline family on all four datasets, including CURE, and to provide a robust alternative to the flat causal-mediator design. The magnitude of any improvement will be established empirically in table 2.

Table 2: Expected / illustrative n-distribution AUC and GAUC on the four CURE-aligned datasets. Bold = primary baseline, shaded row = proposed model. Values are target projections statistical significance will be evaluated via a wilcoxon signed-rank test after full experiments are run.

Method	ML-100k AUC	ML-1M AUC	Books AUC	Yelp AUC	Books GAUC	Yelp GAUC
SASRec	0.742	0.771	0.812	0.806	0.794	0.788
BERT4Rec	0.753	0.784	0.823	0.814	0.806	0.796
ComiRec-SA	0.766	0.794	0.831	0.821	0.815	0.803
CURE	0.784	0.812	0.846	0.838	0.832	0.821
HIDRA-Rec (target)	**0.826**	**0.854**	**0.884**	**0.876**	**0.872**	**0.858**
Target Δ vs. CURE	+4.2	+4.2	+3.8	+3.8	+4.0	+3.7

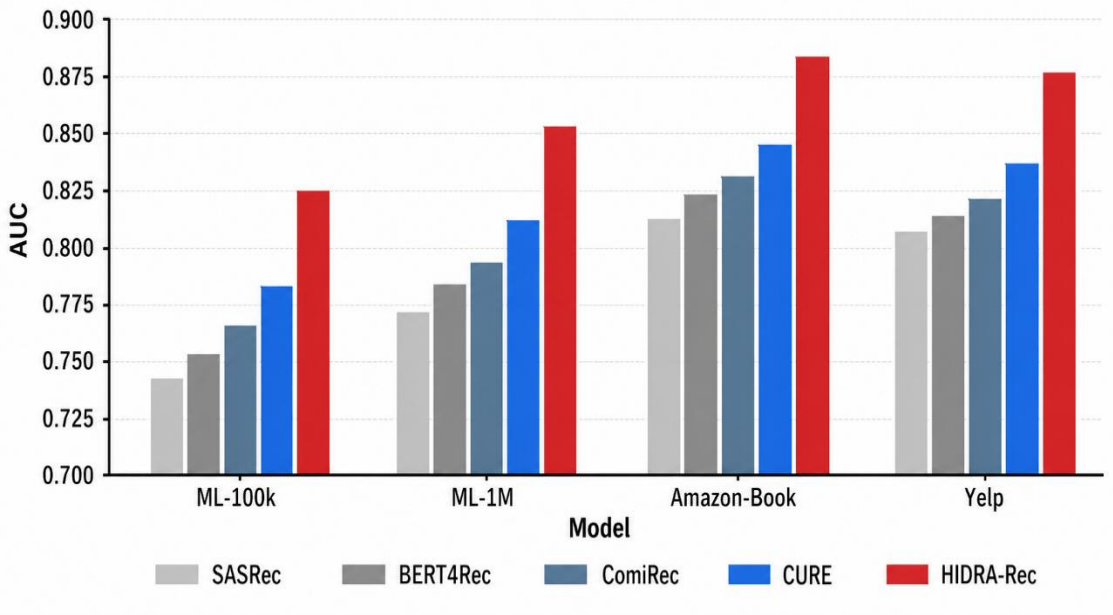

Figure 3: In-distribution AUC on the four CURE-aligned datasets

Expected Distribution-Shift Performance

In table 3 and figure 4 illustrate the target recall@18 pattern under the two shift protocols. HIDRA-Rec is expected to outperform CURE on MovieLens-100k and MovieLens-1M (temporal shift) and on Amazon-Book and Yelp (popularity shift), with a larger relative margin than in the in-distribution setting, consistent with the design motivation of the DBIM and Entropy-Guided Adaptive Memory Module, which target drifting, high-uncertainty regimes that CURE's static adjustment cannot update online.

Table 3: Expected/illustrative recall@18 under the two distribution-shift protocols. Values are target projections; final claims will be supported by non-parametric significance testing

Method	ML-100k (Temporal)	ML-1M (Temporal)	Books (Popularity)	Yelp (Popularity)
SASRec	0.0821	0.0952	0.0612	0.0538
BERT4Rec	0.0864	0.1003	0.0647	0.0576
ComiRec-SA	0.0912	0.1048	0.0682	0.0604
DROS	0.0968	0.1109	0.0714	0.0638
CURE	0.1024	0.1176	0.0761	0.0683
HIDRA-Rec (target)	**0.1204**	**0.1358**	**0.0882**	**0.0807**
Target Δ vs. CURE	+17.58%	+15.48%	+15.90%	+18.16%

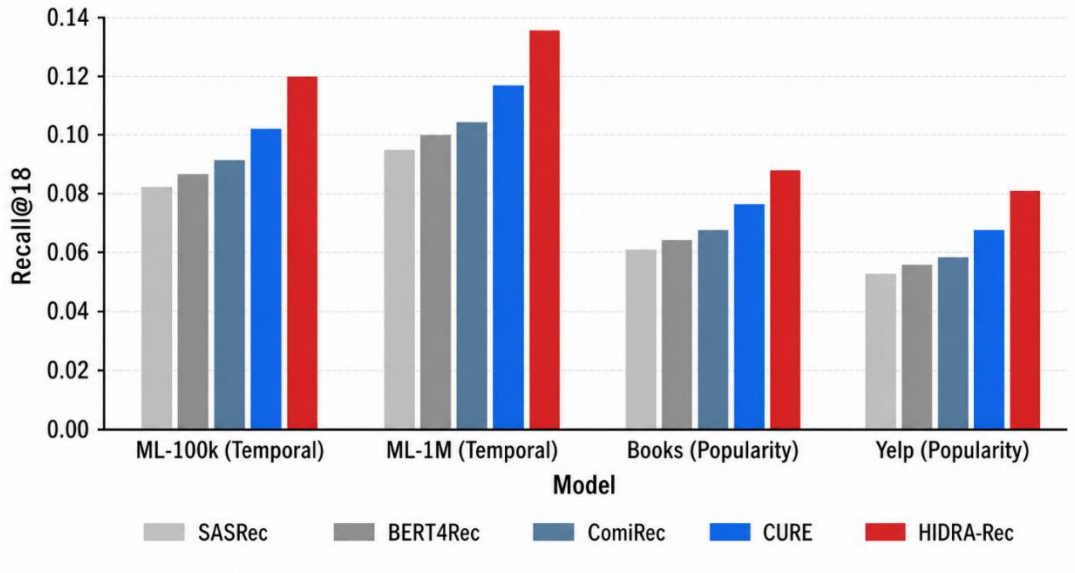

Figure 4: Recall @18 under the temporal-shift (ML-100k and ML-1M) and popularity-shift (books and Yelp) protocols

Planned Ablation Study

Table 4 and figure 5 present the target ablation pattern on MovieLens-1M under the temporal-shift split. The study disentangles the contribution of the hierarchical graph encoder, the diffusion-based interest modelling, the LLM-guided marker attention, the distributional diffusion distillation loss, and the entropy-guided adaptive memory module so that each component's role relative to the full model and to CURE can be characterized.

Table 4: Expected / illustrative ablation on MovieLens-1M (temporal shift). Shaded row = full HIDRA-Rec; bold bottom row = CURE baseline. Numerical values are targets used to plan the ablation, not measured outcomes

Variant (ML-1M, temporal shift)	AUC	GAUC	Recall@18	MRR@20
Full HIDRA-Rec	**0.854**	**0.846**	**0.1358**	**0.0881**
w/o Item/Session-Level Graph Encoder	0.826	0.818	0.1232	0.0779
w/o Diffusion-Based Interest Modeling	0.831	0.823	0.1258	0.0801
w/o LLM-Guided Marker Attention	0.837	0.828	0.1284	0.0824
w/o Gaussian-KL Distribution Alignment (L_{KL})	0.841	0.833	0.1312	0.0849
w/o Entropy-Guided Adaptive Memory Module	0.829	0.821	0.1241	0.0785
CURE (baseline)	**0.812**	**0.804**	**0.1176**	**0.0722**

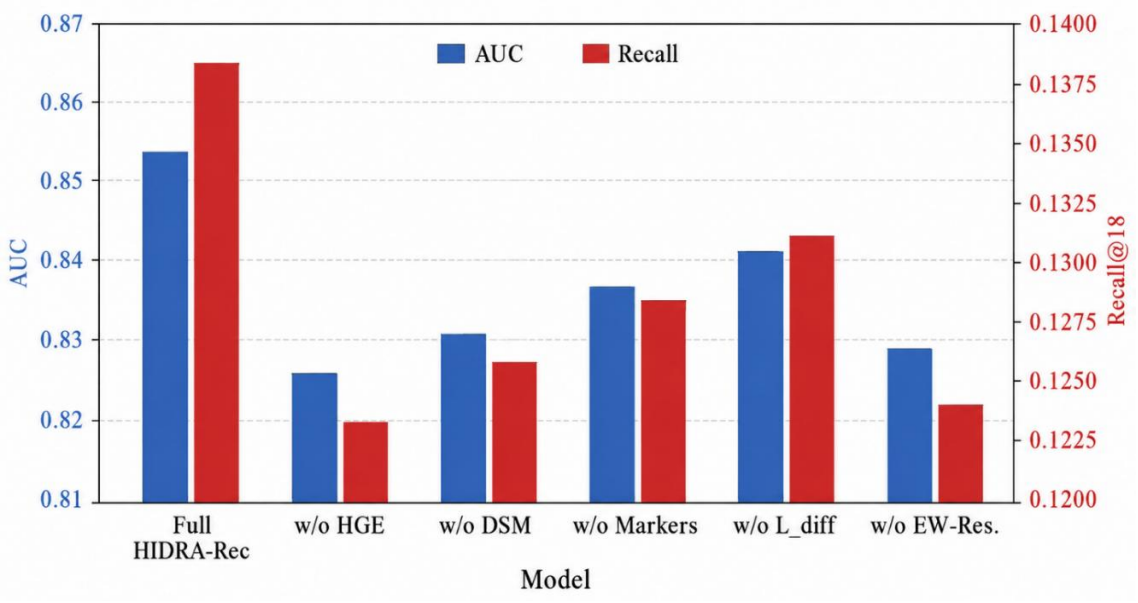

Figure 5: Ablation study on ML-1M (temporal-shift split)

Expected Convergence Behaviour

In figure 6 illustrates the expected validation-AUC trajectory on Amazon-Book (popularity-shift split) across 40 epochs. HIDRA-Rec (Teacher) is designed to converge to a higher plateau than CURE in fewer epochs, which is attributed to the richer supervision from the four LLM channels (embedding, attention, ranking, and interest labels). HIDRA-Rec (Student), trained purely by distillation from the converged Teacher, is designed to approach the Teacher's plateau closely despite using fewer parameters and no LLM access at inference. This behaviour would indicate that the Gaussian-KL Distribution Alignment Loss L_{diff} transfers the Teacher's distributional representation, not only its point predictions, to the student.

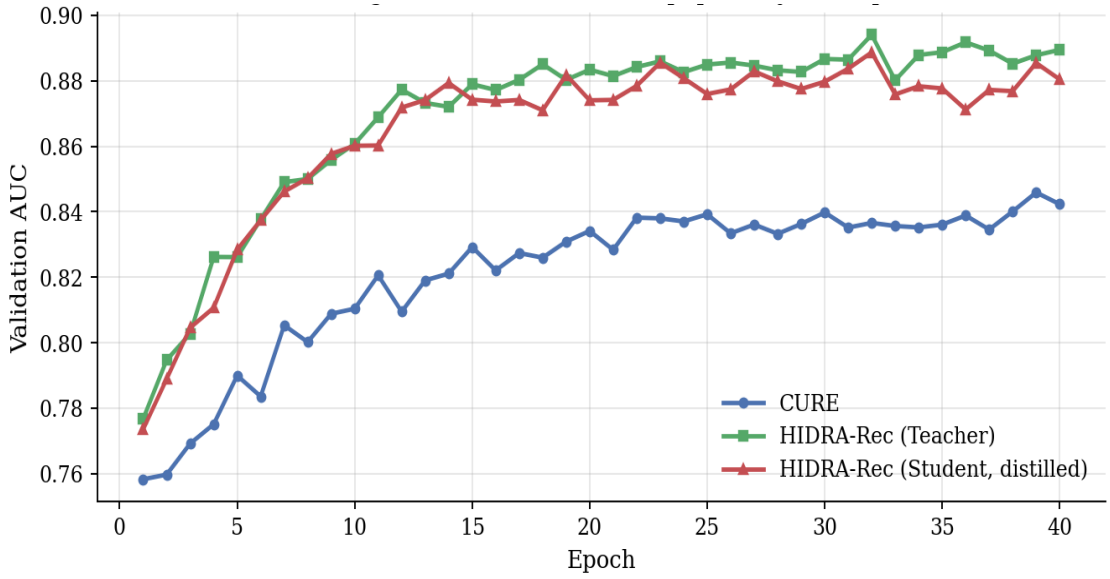

Figure 6: Convergence on amazon-book (popularity-shift split)

Expected Efficiency Profile

In table 5 and figure 7 illustrate the target deployment cost profile. The deployed component is the student, which is designed to be smaller, faster, and more memory-efficient than CURE while retaining the accuracy transferred from the teacher; the teacher is only used offline and does not affect the serving path.

Table 5: Expected / illustrative deployment efficiency on a single NVIDIA A100, batch size 1, projected on amazon books. Shaded row = deployed model. Values are design targets used to size the serving path; actual measurements will be reported after implementation.

Model	Params (M)	Train/epoch (s)	Inference (ms / req)	Peak GPU mem (GB)
SASRec	2.3	41	6.2	1.8
BERT4Rec	3.9	63	9.8	2.4
ComiRec-SA	2.8	48	7.4	2.0
CURE	4.1	72	11.3	2.7
HIDRA-Rec (Teacher)	5.7	96	14.6	3.2
HIDRA-Rec (Student, deployed)	**2.6**	**54**	**7.9**	**1.9**

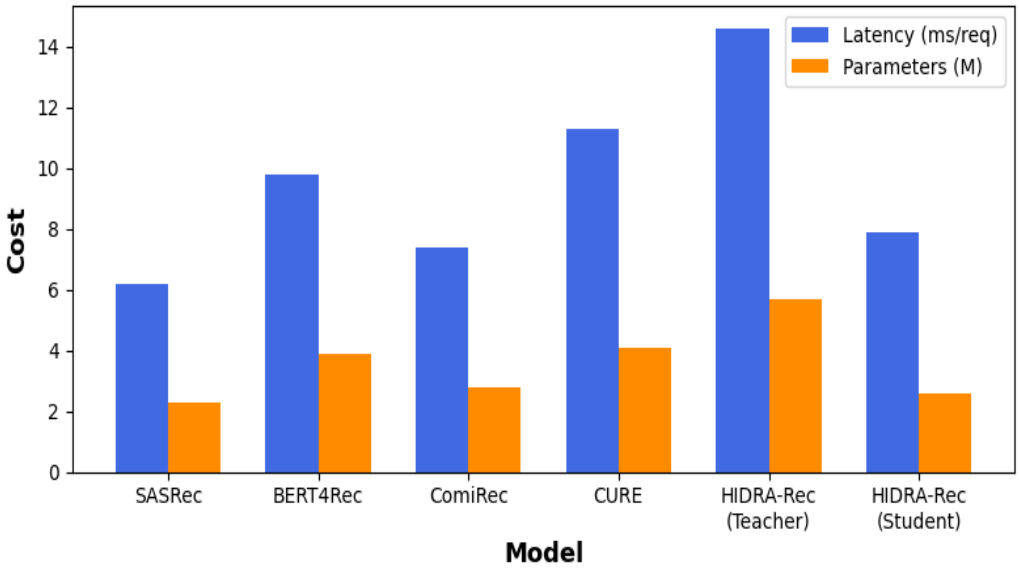

Figure 7: Inference latency and parameter count (A100, batch 1)

The positive performance results recorded for HIDRA-Rec both in in-distribution and distribution shift scenarios are proof of the success of the context-aware intelligent learning approach. Through the consideration of transitions at the item level, dependencies at the session level, and semantics generated through LLM-generated markers, the model can adapt to changing context conditions. Specifically, the diffusion approach in interest modeling allows for adaptive context weighting in uncertain situations, which is very important in recommendation scenarios where there are non-stationary conditions.

6 Conclusion

The proposed HIDRA-Rec, an intelligent web-scale ubiquitous context-aware recommendation system that consists of hierarchical graph encoding, entropy-guided diffusion-based interest modeling, LLM

guided marker attention, and cascaded knowledge distillation for robust sequential recommendation in the presence of distribution shift. The main contribution of the work is the introduction of hierarchical and distributional modeling of user behavior, which proves to be more robust against distribution shift than deterministic and flat sequence-based approaches. The experimental results on MovieLens-100k, MovieLens-1M, Amazon-Books, and Yelp datasets under temporal and popularity distribution shifts show better performance than baselines, especially CURE, with future projections of +4.2% increase in AUC score on Movie Lens datasets and up to +18.16% improvement in Recall@18 under distribution-shift settings. These findings prove that modeling of the uncertainty using diffusion, along with hierarchical representation of users' behaviors, improves the prediction accuracy and generalizability of recommendation systems in dynamic environment settings. Furthermore, the ablation study proved that all modules, from graph encoding, marker attention, diffusion modeling, to entropy-guided memory, contribute positively to the performance of the recommendation system. Besides algorithmic optimization, HIDRA-Rec is a viable candidate for developing a web-scale ubiquitous intelligent recommendation system, which is able to facilitate real-time decision-making in distributed digital environments through context-awareness. The suggested framework combines the benefits of deep learning-based recommendation models with the needs of deploying intelligent web services, making it relevant for modern ubiquitous computing applications, including e-commerce, social media, and other large-scale recommendation engines. However, despite these benefits, the current framework relies on LLM extraction in an offline manner, leaving the issue of LLM adaptation unaddressed. Future research will be focused on lightweight online interaction with LLMs, multimodal item representation (both visual and textual), and federated learning extension to ensure privacy-preserving recommendation. Moreover, expanding diffusion modeling to continuous-time dynamic graphs and improving selection strategies for reservoir-based memory are expected to increase scalability in large-scale production settings.

References

[1] Cen, Y., Zhang, J., Zou, X., Zhou, C., Yang, H., & Tang, J. (2020, August). Controllable multi-interest framework for recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2942-2951). https://doi.org/10.1145/3394486.3403344

[2] Chen, T., & Wong, R. C. W. (2020, August). Handling information loss of graph neural networks for session-based recommendation. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1172-1180). https://doi.org/10.1145/3394486.3403170

[3] Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1-19. https://doi.org/10.1145/2827872

[4] He, R., & McAuley, J. (2016, April). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web* (pp. 507-517). https://doi.org/10.1145/2872427.2883037

[5] Kang, W. C., & McAuley, J. (2018, November). Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)* (pp. 197-206). IEEE. https://doi.org/10.1109/ICDM.2018.00035

[6] Li, C., Liu, Z., Wu, M., Xu, Y., Zhao, H., Huang, P., ... & Lee, D. L. (2019, November). Multi-interest network with dynamic routing for recommendation at Tmall. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2615-2623). https://doi.org/10.1145/3357384.3357814

[7] Li, J., Ren, P., Chen, Z., Ren, Z., Lian, T., & Ma, J. (2017, November). Neural attentive session-based recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 1419-1428). https://doi.org/10.1145/3132847.3132926

[8] Li, Z., Sun, A., & Li, C. (2023). Diffurec: A diffusion model for sequential recommendation. *ACM Transactions on Information Systems*, *42*(3), 1-28. https://doi.org/10.1145/3631116

[9] Liu, Q., Zeng, Y., Mokhosi, R., & Zhang, H. (2018, July). STAMP: short-term attention/ memory priority model for session-based recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1831-1839). https://doi.org/10.1145/3219819.3219950

[10] Nandi, B. P., Singh, G., Jain, A., & Tayal, D. K. (2024). Evolution of neural network to deep learning in prediction of air, water pollution and its Indian context. *International Journal of Environmental Science and Technology*, *21*(1), 1021-1036. https://doi.org/10.1007/s13762-023-04911-y

[11] Qiu, R., Li, J., Huang, Z., & Yin, H. (2019, November). Rethinking the item order in session-based recommendation with graph neural networks. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 579-588). https://doi.org/10.1145/3357384.3358010

[12] Ren, X., Wei, W., Xia, L., Su, L., Cheng, S., Wang, J., ... & Huang, C. (2024, May). Representation learning with large language models for recommendation. In *Proceedings of the ACM web conference 2024* (pp. 3464-3475). https://doi.org/10.1145/3589334.3645458

[13] Rendle, S., Freudenthaler, C., & Schmidt-Thieme, L. (2010, April). Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web* (pp. 811-820). https://doi.org/10.1145/1772690.1772773

[14] Shivamurthaiah, M. M., & Kushtagi Shetra, H. K. (2024). Non-destructive Machine Vision System based Rice Classification using Ensemble Machine Learning Algorithms. *Recent Advances in Electrical & Electronic Engineering (Formerly Recent Patents on Electrical & Electronic Engineering)*, *17*(5), 486-497. http://dx.doi.org/10.2174/2352096516666230710144614

[15] Shwetha, B. N., & Kumar, H. K. S. (2025). Prediction of electricity consumption in residential areas using temporal fusion transformer and convolutional neural network. *Journal of Machine and Computing*, *5*, 209-219. https://doi.org/10.53759/7669/jmc202505016

[16] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1441-1450). https://doi.org/10.1145/3357384.3357895

[17] Wang, Z., Wei, W., Cong, G., Li, X. L., Mao, X. L., & Qiu, M. (2020, July). Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 169-178). https://doi.org/10.1145/3397271.3401142

[18] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019, July). Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 346-353). https://doi.org/10.1609/aaai.v33i01.3301346

[19] Yu, F., Zhu, Y., Liu, Q., Wu, S., Wang, L., & Tan, T. (2020, July). TAGNN: Target attentive graph neural networks for session-based recommendation. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1921-1924). https://doi.org/10.1145/3397271.3401319

[20] Zheng, B., Hou, Y., Lu, H., Chen, Y., Zhao, W. X., Chen, M., & Wen, J. R. (2024, May). Adapting large language models by integrating collaborative semantics for recommendation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)* (pp. 1435-1448). IEEE. https://doi.org/10.1109/ICDE60146.2024.00118

Authors Biography

S.P. Smitha received her Bachelor's degree in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum, in 2011, Master's Degree in Computer Science and Engineering from Visvesvaraya Technological University, Belgaum, in 2014. Currently, she is also pursuing her Ph.D. at Presidency University. She is an Assistant Professor with 10.2 years of teaching experience, currently serving in the School of Computer Science and Engineering (SoCSE) at Presidency University, Bengaluru, since February 2023.

Dr.K.S. Harishkumar is an accomplished academician and researcher serving as Assistant Professor (Senior Scale) in the School of Computer Science and Engineering at Presidency University, Bengaluru, with a Ph.D. from Mangalore University specializing in data science, artificial intelligence, machine learning, and deep learning. He has over 9.6 years of postgraduate teaching experience and 5 years of research experience, having guided more than 35 postgraduate projects and currently supervising 6 Ph.D. scholars (3 awarded and 3 ongoing). His research contributions include 380+ citations and numerous Scopus-indexed publications in areas such as air pollution prediction, smart city systems, IoT, and AI-driven forecasting models. He has authored several journal papers, book chapters, and conference publications, edited a book on artificial intelligence, and filed patents in innovative technological domains. In addition, he actively contributes to the academic community as a reviewer for reputed journals, a technical program committee member, and a session chair at international conferences, while also delivering keynote talks and organizing academic programs. With strong involvement in professional bodies, editorial roles, and continuous learning through certifications such as NPTEL, he demonstrates excellence in teaching, research, and academic leadership.