Towards S^2-Challenges Underlying LLM-Based Augmentation for Personalized News Recommendation

Shicheng Wang[1,2], Hengzhu Tang[3], Li Gao[3*], Shu Guo[4*], Suqi Cheng[3], Junfeng Wang[3], Dawei Yin[3], Tingwen Liu[1,2], Lihong Wang[4]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Baidu Inc.
[4] National Computer Network Emergency Response Technical Team/Coordination Center
{wangshicheng, liutingwen}@iie.ac.cn , {hengzhutang, gaoli.sinh, chengsuqi}@gmail.com, guoshu@cert.org.cn, wangjunfeng@baidu.com, yindawei@acm.org, wlh@isc.org.cn

Abstract

Personalized news recommendation aims to recommend candidate news to the target user. Since the data and knowledge involved in traditional recommender systems are restricted, recent studies utilize large language models (LLMs) to generate news articles and augment the original dataset. However, despite the superiority of LLM-based augmentation in news recommendation, previous studies still suffer from two serious problems, i.e., **structure-level deficiency** and **semantic-level noise**. Since the LLM-based augmentation is mainly implemented at the semantic level, collaborative signals, the critical structure information in recommender systems, is neglected during the generation process. Thus, it is inappropriate to perform recommendation based on the augmented user-news bipartite, which manifests as multiple isolated cliques. Moreover, utilizing the open-world knowledge of LLMs to extend the closed systems will inevitably introduce noise information, leading to difficulties in mining users' real preferences. In this paper, we propose a novel **S**tructure-aware and **S**emantic-aware approach for **L**LM-**E**mpowered personalized **N**ews **R**ecommendation, named S^{2}LENR, to tackle the mentioned problems. Specifically, we propose a structure-aware refinement module to inject collaborative information in a parametric way, in order to construct a valid augmented bipartite. Besides, we devise a semantic-aware denoising module utilizing contrastive learning paradigm to overcome the negative effects of noise information. Finally, we calculate the relevance score between target user and candidate news representations. We conduct experiments on two real-world news recommendation datasets MIND-Large, MIND-Small and empirical results demonstrate the effectiveness of our approach from multiple perspectives.

Introduction

With the rapid development of the Internet and online news services, a large amount of news is published on numerous news platforms all the time, which makes users overwhelmed. Therefore, personalized news recommendation is necessary for news platforms to help users alleviate information overload and improve their reading experience. Briefly, news recommendation techniques aim to recommend candidate news that the target user may be interested in.

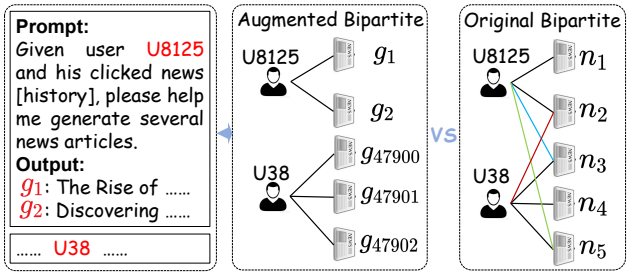

Figure 1: Illustration of structure-level deficiency: Each generated news only interacted with the specific user and thus generated bipartite lacks of collaborative information.

Since the data and knowledge involved are confined to specific application domains, traditional recommender systems are usually considered to be closed forms (Koren, Bell, and Volinsky 2009; Zhou et al. 2018). As a result, the characteristic of significant isolation from the external world tends to exert negative effects on exploring implicit relevant clues as well as mining users' preferences. Although recent studies have attempted to incorporate external knowledge graphs to break the restrictions of the closed forms (Liu et al. 2020; Wu et al. 2021), their performance is usually constrained due to the limited generalization ability of the structural data. Fortunately, the recent emergence of large language models (LLMs), such as GPT-4 (Achiam et al. 2023) or LLaMA (Touvron et al. 2023), has revolutionized the learning paradigm of various research fields, due to their multifarious exceptional abilities. To this end, with respect to the characteristics of the news recommendation task, a common solution is to generate synthetic news articles and augment the original dataset through fully exploiting the various capabilities of LLMs. For example, GERNE (Liu et al. 2023) constructs user-specific prompts and leverages pretrained semantic knowledge from LLMs to enrich news data.

However, despite the superiority of LLM-based augmentation in news recommendation, previous studies still suffer from two serious problems, i.e., **structure-level deficiency**

*Corresponding author.

and **semantic-level noise**. First, the LLM-based augmentation is mainly implemented at the semantic level, in view of the textual attributes of news articles. However, collaborative information is neglected during the generation process, as each generated news article is arbitrarily produced and merely interacted with the specific user. Obviously, there is a huge gap between the original data and the augmented data, since the latter bipartite manifests as multiple isolated cliques as shown in Figure 1. Thus, it is inappropriate to directly perform news recommendation based on the augmented user-news bipartite, where information is difficult to propagate effectively across distinct users. Second, though the open-world knowledge is another strength of LLMs, utilizing such universal knowledge to extend the closed recommender systems will inevitably introduce irrelevant information into users' behaviors, such as unrelated entities. Therefore, it is crucial to filter out such noise information, in order to obtain users' real preferences. Obviously, the above problems will impair user and news representations, which tends to degrade the recommendation performance ultimately.

In this work, we devise a novel **S**tructure-aware and **S**emantic-aware approach for **L**LM-**E**mpowered **N**ews **R**ecommendation, termed as $\mathrm{S^2LENR}$, which appears as a two-stage pipeline architecture. First of all, we design appropriate prompt templates to generate synthetic news articles, in order to augment the original dataset and build bridges to the external world. Second, we devise two innovative components to tackle the above-mentioned problems respectively. Specifically, we propose a structure-aware refinement module to construct a valid augmented user-prototype bipartite and then inject pseudo collaborative signals. In this way, we can mitigate the structure-level deficiency problem to a certain extent. Additionally, inspired by the contrastive learning ideology, we develop a semantic-aware denoising module to overcome the negative influence of noise information. Finally, we calculate the relevance score between target user and candidate news representations and decide whether to recommend or not. We conduct experiments on real-world datasets MIND-Large and MIND-Small to demonstrate the effectiveness of our proposed $\mathrm{S^2LENR}$.

The main contributions of this paper can be summarized as follows:

- To alleviate structure-level deficiency problem, we propose a structure-aware refinement module to inject collaborative information in a parametric way.
- We devise a semantic-aware denoising module based on contrastive learning, in order to attenuate the effects of semantic-level noise information.
- The empirical results demonstrate the effectiveness of our approach and verify the validity of introducing LLM-based augmentation to promote news recommendation.

Related Work

Personalized News Recommendation

News recommendation has recently attracted more and more attention with the growth of individual and social needs. Therefore, a variety of methods have been proposed, including ID-based methods and content-based methods. Most traditional ID-based methods achieved news recommendation based on collaborative filtering framework (Das et al. 2007; Wang and Blei 2011). They parameterized users and news in a latent space and aimed at reconstructing interactive behaviors. However, these methods always suffered from severe cold-start problem, to this end, content-based news recommendation methods have been proposed recently. In the early stage, researchers treat the task as a sequence modeling problem. For example, NAML (Wu et al. 2019c) leveraged a CNN network to learn news representations from news titles and categories. Then they learnt user representations through attentively aggregating click history. NRMS (Wu et al. 2019a) utilized homologous multi-head self-attention networks to learn news and user representations separately, capturing interactive information among word sequences and news sequences respectively. Next, with the development of graph neural network techniques, GNN-based methods are proposed to learn news and user representations. For example, GNewsRec (Hu et al. 2020) learnt users' short-term interests by applying attentive GRU neural network on click history, as well as users' long-term interests via graph neural networks. However, the above news recommendation methods are usually considered to be closed forms (Koren, Bell, and Volinsky 2009; Zhou et al. 2018) in which the data and knowledge involved are confined to specific news domains, inducing significant isolation from the external world. As we claimed before, the closed systems exert negative effects on exploring implicit relevant clues as well as mining users' preferences obviously.

Methodology

In this section, we first present the problem formulation of personalized news recommendation.

Task Definition. We denote U and N as the overall user and news sets. Given a candidate news $n_c \in N$ and target user $u \in U$ with his click history $[n_1, \cdots, n_L]$, we aim to learn their representations respectively. Afterwards we calculate the relevance score between their representations and decide whether to recommend n_c or not.

Then Figure 2 illustrates the overall architecture of our method $\mathrm{S^2LENR}$. On the basis of general encoders, we devise LLM-based augmentation module consisting of three steps to promote recommendation performance. We will elaborate our method in the subsequent sections in detail.

Basic Encoder

News Encoder. In this section, we introduce how to learn news semantic representations from news titles. Specifically, given title word sequence as $[w_1, \cdots, w_M]$, where w_i is denoted as the i-th word in title, we encode titles based on the fundamental Transformer framework.

First, at the bottom of news encoder module, it applies word embedding layer to convert each word w_i into corresponding vector e_i. Then, it applies a multi-head self-attention network to capture semantic relatedness among words $[e_1, \cdots, e_M]$. The representation of the word w_i learned by the s-th attention head h_i^s is calculated as fol-

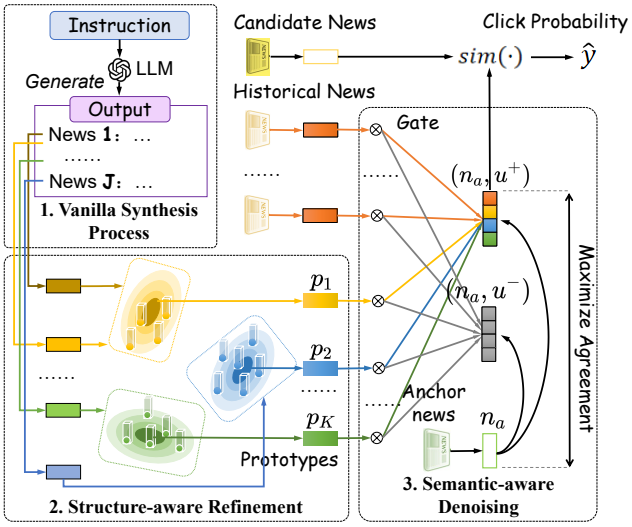

Figure 2: Overall framework of our S^{2}LENR method.

lows:

$$h_i^s = V_s^w \sum_{j=1}^{M} \alpha_{i,j}^s \cdot e_j, \quad \alpha_{i,j}^s = \frac{\exp(e_i^T Q_s^w e_j)}{\sum_{t=1}^{M} \exp(e_i^T Q_s^w e_t)}, \tag{1}$$

where $V_s^w \in \mathbb{R}^{d/S \times d}$ and $Q_s^w \in \mathbb{R}^{d \times d}$ are the projection parameters in the s-th self-attention head, and $\alpha_{i,j}$ indicates the interaction score between the word w_i and w_j. Then the multi-head representation of word w_i is concatenated as $h_i \in \mathbb{R}^d$, i.e., $h_i = [h_i^1; h_i^2; \cdots; h_i^S]$, where S denotes the number of separate self-attention heads. Finally, it applies an additive attention network to aggregate contextual word representations into a news representation n, formulated as:

$$n = \sum_{i=1}^{M} \alpha_i^w \cdot h_i, \quad \alpha_i^w = \frac{\exp(q_\alpha^T \cdot h_i)}{\sum_{j=1}^{M} \exp(q_\alpha^T \cdot h_j)}, \tag{2}$$

where $q_\alpha \in \mathbb{R}^d$ is a projection parameter vector. In this way, we are able to obtain representations for real news, denoted as $n_i, i \in (1, |N|)$.

User Encoder. Since there is semantic relatedness between news articles clicked by the same user, we apply a multi-head self-attention mechanism among the entire click history $[n_1, \cdots, n_L]$. Similarly, the enhanced news representations by the s-th attention head n_i^s are formulated as:

$$n_i^s = V_s^n \sum_{j=1}^{N'} \beta_{i,j}^s \cdot n_j, \quad \beta_{i,j}^s = \frac{\exp(n_i^T Q_s^n n_j)}{\sum_{t=1}^{N'} \exp(n_i^T Q_s^n n_t)}, \tag{3}$$

where $Q_s^n \in \mathbb{R}^{d \times d}$ and $V_s^n \in \mathbb{R}^{d/S \times d}$ are the projection parameters in the s-th self-attention head, and $\beta_{i,j}$ indicates the interaction score between the news n_i and n_j. Then the multi-head representation of news n_i is concatenated as $n_i \in \mathbb{R}^d$, i.e., $n_i = [n_i^1; n_i^2; \cdots; n_i^S]$. Next it applies an additive attention network to aggregate contextual news representations into a user representation u, formulated as:

$$u = \sum_{i=1}^{n_{N'}} \beta_i \cdot n_i, \quad \beta_i = \frac{\exp(q_\beta^T \cdot n_i)}{\sum_{j=1}^{n_{N'}} \exp(q_\beta^T \cdot n_j)}, \tag{4}$$

where $q_\beta \in \mathbb{R}^d$ is a projection vector, and β_i denotes the contribution weight of news n_i to user representation.

LLM-based Augmentation

Vanilla Synthesis Process. As recent studies (Zhao et al. 2023) claim that LLMs possess exceptional language understanding and logical reasoning abilities, we demand the LLMs to comprehend users' historical interactions and then generate synthetic news articles, in order to provide abundant semantic information and break the limitations of the closed systems.

Specifically, given a prompt π and template $\mathtt{F}(\cdot)$, we can get the instruction $\mathtt{F}(history, \pi)$ and query the LLM to generate the news articles:

$$G \leftarrow \text{LLM}(\mathtt{F}(history, \pi)), \tag{5}$$

where $\mathtt{F}(\cdot)$ aims to fill the clicked news history into the slots of prompt π and G represents the entire generated news corpus. The prompt π is defined as follows:

> Assuming that you are a news article generator carrying divergent thinking. Given a specific user and his few clicked news history in list format **history**, please first summarize keywords of interest to the user and infer his potential reading preference. Then help me generate at most **#NUM** news articles step by step. The generated news articles must be reasonable, diverse, fluent and have low similarities to historical contents.
> Please output with the following FORMAT:
> ⟨start output⟩
> Title : ⟨text⟩; Abstract : ⟨text⟩; Topic : ⟨words⟩;
> Keywords : ⟨words⟩; Evidence : ⟨text⟩
> ⟨end output⟩

For each user, by adding the clicked history that simply utilizes the delimiter [SEP] to concatenate the textual context of each news article into prompt π, this process can generate multiple synthetic news articles by querying the LLM. Notably, since LLMs perhaps can not perform well when dealing with temporal information, we ignore the historical relevant information and merely focus on understanding and exploring users' reading preferences from semantic perspective. To maintain consistency with the original news format, we provide clear definitions in prompt π for instructing LLM to generate news in the pre-defined format.

Structure-aware Refinement. As described before, for each user, we have already obtained synthetic news articles to enhance users' preference. Nevertheless, compared to the original user-news bipartite, the augmented bipartite manifests as multiple isolated cliques and lacks collaborative information, which is not suitable for recommendation.

Intuitively, a feasible solution is to introduce auxiliary components and establish connections between the cliques. Inspired by the concept of prototype learning, we integrate

multiple generated news with high correlations to form a corresponding prototype, even if they belong to different users. Furthermore, taking into account the connections between users and generated news, we regard the latter as temporary hubs and then construct a bipartite where users interact with prototypes directly. As a result, we introduce pseudo collaborative information in the user-prototype bipartite, as the mentioned isolated cliques disappeared and information can be propagated across different users. Specifically, the main procedure consists of prototype representation learning and user-prototype structure construction.

We first input the generated news articles into basic news encoder and the corresponding representations are denoted as $g_j, j \in (1, |G|)$. Since we are unaware of the practical content of the prototypes, we assume that there are K parametric prototypes. Next, we adopt attention-based information aggregation mechanism to obtain robust prototype representation $p_i, i \in (1, K)$ separately as follows:

$$\beta_{i,j} = \frac{\exp W_1(p_i||g_j)}{\sum_{k=1}^{|G|} \exp W_1(p_i||g_k)}, \tag{6}$$

$$p_i = ReLU(p_i + W_2 \sum_{j=1}^{|G|} \beta_{i,j} \cdot g_j), \tag{7}$$

where $W_1 \in \mathbb{R}^{1\times 2d}, W_2 \in \mathbb{R}^{d\times d}$ are learnable parameter matrices. Moreover, it is worth noting that the prototypical features p tend to be more robust to noise than initial features g of generated news.

Afterwards, we employ straightforward matrix multiplication to calculate correlation matrix W containing affiliation weights between the generated news articles and prototypes. Then we choose the index t of the maximum weight in the j-th row of W to indicate the prototype that the generated news g_j belonging to:

$$t = \underset{i}{\operatorname{argmax}} \frac{\exp(W_{j,i})}{\sum_{k=1}^{K} \exp(W_{j,k})}, i \in (1, K) \tag{8}$$

Currently, we can construct interactions between users and prototypes directly. Then, prototypes can be spliced after the historical clicked behaviors. In this way, we are capable of inferring users' potential preferences from the original click history and interactions with inductive prototypes.

Semantic-aware Denoising. Although prototypical features are resistant to noise to some extent, there is still unpleasant misleading information, due to the open-world knowledge of LLMs. In this section, we design to filter out noise information while preserving effective information.

For simplicity, both real news and additional prototypes are denoted as n uniformly. At first, we devise an element-wise gated mechanism to filter news-level noise as follows:

$$gate = Sigmoid(W_f \times LN(n)), \tag{9}$$

where $W_f \in \mathbb{R}^{d\times d}$ denotes shared parameter matrix and $LN(\cdot)$ indicates layer norm operation. Therefore, the opposite pair of representations for each news, containing effective information and noise information respectively, are calculated through element-wise product operation:

$$n^+ = gate \odot n, \quad n^- = (1 - gate) \odot n. \tag{10}$$

Afterwards, we deliver the opposite pairs n^+ and n^- into the basic user encoder, in order to acquire corresponding user representations u^+ and u^- separately, where u^+ is intended to reflect users' actual preference. Intuitively, compared to u^-, u^+ should be closer to the candidate news n_a that target user will truly click on. Therefore, inspired by the local-local contrast methods recently (Chen et al. 2020; Tian, Krishnan, and Isola 2020), we propose to impose constraints on the resulting pairs. According to the universal process of contrastive learning, we treat (u^+, n_a) as positive pair, and (u^-, n_a) as corresponding negative pairs above all. Then we construct the contrastive learning objective as follows:

$$\mathcal{L}_{ssl} = -\sum_{n_a \in TS} \log \frac{\exp(f(u^+, n_a))}{\exp(f(u^-, n_a))}, \tag{11}$$

where TS denotes the training set and $f(\cdot)$ is a scoring function for sample pairs, i.e., cosine similarity. Through optimizing this objective function, valid information is retained in u^+ while noise information is forced to remain in u^-.

Click Predictor

Simple but effective, the click probability score y is computed by the inner product between the target user representation u^+ and the candidate news representation n_c, i.e., $\hat{y} = u^+ \cdot n_c$.

Model Training

Recent news recommendation researches (Wu et al. 2019a; Wang et al. 2023) focus on retrieving relevant news articles from the candidate pool, rather than ranking them. Therefore, the objective is formalized as classification task and negative sampling techniques is used for model training. Denoting clicked candidate news in the training set as positive sample n_i, i.e., $y_{u,i} = 1$, then we randomly choose P non-clicked candidate news as negative samples $[n_{i,1}, \cdots, n_{i,P}]$ from the same impression displayed to the target user. The corresponding click probability scores of the positive and P negative samples are calculated as $\hat{y}_i^+$ and $[\hat{y}_{i,1}^-, \hat{y}_{i,2}^-, \cdots, \hat{y}_{i,P}^-]$ respectively. Finally, the supervised classification loss function is formulated as:

$$\mathcal{L}_{ce} = -\sum_{i \in TS} \log \frac{\exp(\hat{y}_i^+)}{\exp(\hat{y}_i^+) + \sum_{j=1}^{P} \exp(\hat{y}_{i,j}^-)}. \tag{12}$$

Since we have already obtained the main classification loss function and auxiliary contrastive loss function, we define the final loss function in a joint paradigm as:

$$\mathcal{L} = \mathcal{L}_{ce} + \alpha \cdot \mathcal{L}_{ssl}, \tag{13}$$

where α is a hyper-parameter that makes a trade-off between classification loss and contrastive loss.

During the inference stage, candidate news with top scores y are selected to be recommended to the target user.

Experiment

This section conducts experiments to evaluate the performance of our model, S^2LENR.

# News	161,013	# Users	1,000,000
# News category	20	# Impression	15,777,377
# Entity	3,299,687	# Click behavior	24,155,470
Avg. title len.	11.52	Avg. abstract len.	43.00
Avg. body len.	585.05		

Table 1: Statistic information of MIND-Large dataset.

Dataset and Evaluation Metrics

Following previous work (Wu et al. 2019a; Qi et al. 2021), we conduct extensive experiments on two large-scale real-world datasets, MIND-Large [1] and MIND-Small [2], to evaluate the effectiveness of our method. MIND-Large dataset collected from Microsoft News platform contains two record documents. One document describes text content of news, including titles and abstracts. The other document describes click behaviors between users and news. These total interactions are gathered from October 12 to November 22, 2019 (six weeks). The click behaviors in the first four weeks are regarded as user reading history, the behaviors in the penultimate week is applied for training, and the data in last week is used for performance evaluation. Detailed statistic information about MIND-Large dataset is summarized in Table 1. Besides, we call GPT-4 API [3] to generate news articles. After filtering texts with wrong forms, we retain approximate 230,000 and 160,000 generated news articles for each dataset respectively. We independently repeated each experiment 10 times and reported average results in terms of AUC, MRR(Voorhees et al. 1999), nDCG@5(Järvelin and Kekäläinen 2002) and nDCG@10. In classification scenario, AUC is the most important one among them.

Baselines

To demonstrate the effectiveness of the proposed news recommendation method, we compare our $\mathrm{S}^2\mathrm{LENR}$ with the following typical content-based baselines: (1) **EBNR** (Okura et al. 2017) employs a GRU network to learn user representations from clicked news history. (2) **DKN** (Wang et al. 2018) utilizes an adaptive attention network to learn user representations considering relatedness between candidate news and historical news. (3) **NPA** (Wu et al. 2019b) employs personalized attention networks to learn individual representation for each user. (4) **NAML** (Wu et al. 2019c) leverages CNN networks to model news semantic representations and learns user representations through attentively aggregating clicked news. (5) **LSTUR** (An et al. 2019) models short-term user interests via a GRU network and long-term user interests via user ID embeddings respectively from two perspectives. (6) **NRMS** (Wu et al. 2019a) learns news representations and user representations through multi-head self-attention networks respectively. (7) **GNewsRec** (Hu et al. 2020) models users' short-term interests by applying attentive GRU neural network and users' long-term interests via graph neural networks based on user-news-topic graph. (8) **GERL** (Ge et al. 2020) utilizes the neighborhood of news and users on the user-news bipartite network to enhance their representations. (9) **User-as-Graph** (Wu et al. 2021) proposes to represent each user as a personalized heterogeneous graph built from their behaviors, including extra entity information. (10) **HDNR** (Wang et al. 2023) learns user and news representations in hyperbolic space with exponential-growth volume and design a re-weighting aggregation module to alleviate conformity bias. (11) **LENR** imitates the architecture of **GERNE** (Liu et al. 2023), which performs a two-stage pipeline framework combing vanilla LLM-based augmentation with previous NRMS model. [4]

Implementation Details

For a fair comparison, following (Wu et al. 2019a), we introduce experimental and hyper-parameters settings of our method. For news content modeling, we utilize the first 30 words of news titles to learn corresponding news representations. In addition, a special character [PAD] is used for filling when the length of word sequence does not meet the condition. Besides, we adopt pre-trained Glove embeddings (Pennington, Socher, and Manning 2014) for word initialization. For user interest modeling, we treat the recent 50 clicked news as users' reading history. Moreover, news and user representations, as well as latent embeddings are both 400-dimensional vectors, i.e., $d = 400$. For hyper-parameters, the number of generated news $\#\mathrm{NUM}$ in prompt π is 5, the number of prototypes K is set to 1000, the joint learning weight α is set to 0.1, and the negative sampling ratio is 4. In addition, we utilize dropout technique and Adam optimizer (Kingma and Ba 2014) for training. The dropout rate and learning rate are 0.1 and 0.001 respectively.

Overall Performance

The main purpose of this section is to verify the effectiveness of our method. We conduct experiments to compare our proposed $\mathrm{S}^2\mathrm{LENR}$ with several baselines on MIND-Large and then apply them to MIND-Small dataset for supplement. The overall performance results are displayed in Table 2 and Table 3 respectively, where the best results are in bold and the second best are underlined.

Obviously, we have several observations according to the tables: (1) First, our proposed method $\mathrm{S}^2\mathrm{LENR}$ outperforms all baselines in terms of AUC on both two real-world datasets, achieving 1.64% and 1.4% improvement comparing to state-of-the-art baselines respectively. Besides, in terms of other evaluation metrics, our method performs excellent as well, which demonstrates the effectiveness and adaptability to data scale of our method. (2) Apparently, the performance of the LENR is better than the corresponding baseline NRMS, which demonstrates the effectiveness of the augmentation module. Nevertheless, the combination is too straightforward to explore the potential of LLM-based augmentation. Taking into account the structure-level and semantic-level problems, our proposed method achieves a

[1] https://msnews.github.io/

[2] A small version of the MIND-Large dataset by randomly sampling 50,000 users and their behavior logs.

[3] https://openai.com/

[4] We do not compare with GERNE, since they include extra images beyond the original dataset.

Method	AUC	MRR	nDCG@5	nDCG@10
EBNR	65.42	31.24	33.76	39.47
DKN	64.60	31.32	33.84	39.48
NPA	66.69	32.24	34.98	40.68
NAML	66.86	32.49	35.24	40.91
LSTUR	67.73	32.77	35.59	41.34
NRMS	67.76	33.05	35.94	41.63
GNewsRec	67.53	32.68	35.46	41.17
GERL	68.24	33.46	36.38	42.11
User-as-Graph	69.23	34.14	37.21	43.04
HDNR	69.98	34.10	37.97	44.20
LENR	69.47	33.87	37.52	43.48
S^2LENR (Ours)	**71.13**	**35.32**	**39.38**	**45.51**

Table 2: Performance of different methods on MIND-Large Dataset (%).

Method	AUC	MRR	nDCG@5	nDCG@10
EBNR	61.62	28.07	30.55	37.07
DKN	63.99	28.95	31.73	37.07
NPA	64.28	29.64	32.28	38.93
NAML	64.30	29.81	32.64	39.11
LSTUR	65.68	30.44	33.49	39.95
NRMS	65.43	30.74	33.13	39.66
GNewsRec	65.91	30.50	33.56	40.13
GERL	66.22	30.89	34.28	40.50
User-as-Graph	66.71	31.13	34.51	40.95
HDNR	68.23	32.61	36.10	42.29
LENR	67.04	31.71	34.98	41.24
S^2LENR (Ours)	**69.17**	**33.23**	**36.93**	**43.18**

Table 3: Performance of different methods on MIND-Small Dataset (%).

significant improvement. In addition, the subsequent ablation study further proves the importance of solving the problems. (3) In general, the methods that introduce external knowledge usually perform better, i.e., User-as-Graph and ours. Therefore, the observation certificates that breaking the limitations of the closed systems contributes to achieve a considerable improvement. Meanwhile, our S^2LENR outperforms User-as-Graph, due to the exceptional generalization ability and extensive open-world knowledge of LLMs.

Performance on Distinct Specific Scenarios

It is well-known that the cold-start users problem is one of the most serious challenges in the news recommendation task. The scenario where users have limited browsing history makes it difficult to understand users' preferences and perform recommendation accurately. To further illustrate the effectiveness of our method, we conduct an in-depth performance comparison oriented to cold-start users as well as warm-start users respectively. To be specific, we define users with less than $\delta \in \{5, 10\}$ click records as cold-start users empirically and then extract the corresponding partial dataset. For simplicity, we compare with three typical baselines, NRMS, User-as-Graph and HDNR. The Performance results are evaluated on MIND-Small dataset and recorded in Table 4.

As shown in the tables, we have several observations: (1) Regardless of the scenarios, our proposed S^2LENR consistently outperforms the typical baselines in terms of all evaluation metrics and obtains around 1.4% improvement in terms of AUC, which confirm the effectiveness and the generalization ability of our method. (2) From an overall perspective, the recommendation performance specific to cold-start users shows significant declines compared to warm-start users. However, we discover that the performance improvement in cold-start scenarios outweighs that in warm-start scenarios, demonstrating the effectiveness of our method in dealing with such a challenging problem. (3) Then we conduct comparative experimental analysis aiming at cold-start users. Though User-as-Graph incorporates extra knowledge graph to promote performance on cold-start users, the relatively lower results further demonstrate the superiority of integrating LLMs. Besides, although HDNR takes into account the power-law distribution property and thus incorporates hyperbolic space to alleviate the cold-start users problem, our method still performs better and we attribute the performance improvement to the introduction of external open-world knowledge.

Ablation Study

In this section, we conduct elaborate ablation experiments on MIND-Small dataset, in order to further evaluate the effectiveness of all innovative components in our method. The results of comparative experiments are recorded in Table 5. In detail, SaR and SaD are the abbreviations for structure-aware refinement and semantic-aware denoising respectively. Notably, the variant "w/o SaR & SaD" is equivalent to the baseline LENR. Besides, we replace the LLM-based augmentation with traditional data augmentation methods, including representation-level Mixup (Zhang et al. 2017) and token-level Concat (Summers and Dinneen 2019) techniques. Notably, the augmentation variants are unable to incorporate with the structure-aware refinement module, since they merely utilize the historical real news articles where the structure-deficiency problem no longer exists. Then we discuss how each component affects the recommendation performance, according to corresponding variant.

By analyzing the results of Table 5, we have the following observations: (1) As shown in the top of the table, the performance decreases correspondingly in terms of all evaluation metrics when we remove different module gradually. Apparently, the phenomenon demonstrates the effectiveness of each proposed innovative components as well as the necessity to address the structure-level deficiency and semantic-level noise problems. (2) In the meantime, we observe that removing the structure-aware refinement module results in more significant performance degradation than removing the denoising module from multiple perspectives, which proves that incorporating collaborative information contributes more to the overall model in recommendation scenarios. (3) In addition, keeping the same settings for the remaining modules, we compare the variants "w/o SaR" and "repl. Mixup" (and "repl. Concat") and discover that the

Threshold	Method	Cold-Start Users ($L < \delta$)				Warm-Start Users ($L \geq \delta$)			
		AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
$\delta = 5$	NRMS	61.46	31.66	34.22	40.37	65.82	29.52	32.58	39.06
	User-as-Graph	62.51	30.30	33.81	40.29	67.53	32.29	35.49	41.39
	HDNR	65.06	31.85	35.58	42.01	68.88	34.55	37.69	43.63
	S^{2}LENR	65.97	32.42	36.31	42.73	69.75	35.06	38.32	44.17
	$\Delta_{\mathcal{H}}$	+1.40	+1.79	+2.05	+1.71	+1.26	+1.48	+1.67	+1.24
$\delta = 10$	NRMS	62.83	31.52	34.30	40.38	66.31	29.07	32.16	38.75
	User-as-Graph	63.92	30.71	34.18	40.55	68.04	32.39	35.64	41.55
	HDNR	66.06	32.29	35.94	42.17	69.30	34.90	37.85	43.80
	S^{2}LENR	66.97	32.85	36.65	42.97	70.21	35.37	38.53	44.33
	$\Delta_{\mathcal{H}}$	+1.38	+1.73	+1.98	+1.90	+1.31	+1.35	+1.80	+1.21

Table 4: Performance results towards cold-start/warm-start users on MIND-Small Dataset.

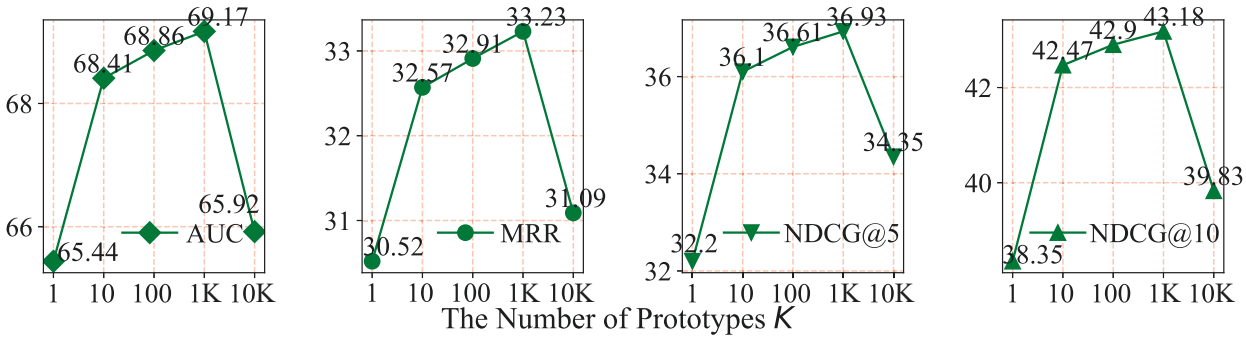

Figure 3: Performance with different values of the number of prototypes.

Method	AUC	MRR	nDCG@5	nDCG@10
S^{2}LENR	**69.17**	**33.23**	**36.93**	**43.18**
w/o SaR	68.20	32.18	35.74	42.02
w/o SaD	68.49	32.45	36.02	42.28
w/o SaR & SaD	67.04	31.71	34.98	41.24
repl. Mixup	67.23	32.17	35.65	41.78
repl. Concat	66.50	31.03	34.45	40.67

Table 5: Effect of each module performs in our model on MIND-Small Dataset.

LLM-based augmentation outperforms traditional data augmentation methods. This fair result verifies the validity of introducing external universal knowledge to break the limitations of the closed recommender systems.

Hyper-Parameter Analysis

In this section, to further evaluate the sensitivity of our proposed method to hyper-parameters, we conduct detailed experiments on the MIND-Small dataset with different values of the number of prototypes K, which affects the performance of structure-aware refinement module. Specifically, Figure 3 shows the trend variation of our method with various number of parametric prototypes. At the beginning, taking into account the information aggregation within the adaptive prototype learning procedure, the lower K indicates that different users are more likely to share similar semantic information. Apparently, the indistinguishable information not only has little effect on exploring users' preference, but also introduces a mass of noise inevitably. Then it can be seen that with the increases of K, the performance of our S^{2}LENR will also gradually improve accordingly in terms of all metrics, which demonstrates the effectiveness of alleviating the structure-level deficiency problem through constructing user-prototype relations. However, when the value of K becomes larger, the performance of our model decreases. This is because the constructed user-prototype collaborative information tends to disappear, leaving the structure-level deficiency problem unsolved.

Conclusion

In this paper, we propose a novel structure-aware and semantic-aware approach for LLM-empowered personalized news recommendation, namely S^{2}LENR. Specifically, we first utilize the superiority of LLMs to enhance the original dataset and break the limitations of the closed recommender systems. Then, we devise a structure-aware refinement module to construct user-prototype bipartite and inject pseudo collaborative signals. Besides, we design a semantic-aware denoising module inspired by the contrastive learning ideology, in order to retain effective semantic information as well as filter out noise. Extensive experiments on real-world datasets validate the effectiveness of our approach.

Acknowledgments

We sincerely thank all the anonymous reviewers for their comments and suggestions. This work is supported by the National Natural Science Foundation of China (No.62106059), the National Natural Science Foundation of China (No.62406319), and the Youth Innovation Promotion Association of CAS (Grant No. 2021153).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, M.; Wu, F.; Wu, C.; Zhang, K.; Liu, Z.; and Xie, X. 2019. Neural news recommendation with long-and short-term user representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 336–345.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Das, A. S.; Datar, M.; Garg, A.; and Rajaram, S. 2007. Google news personalization: scalable online collaborative filtering. In *Proceedings of the 16th international conference on World Wide Web*, 271–280.

Ge, S.; Wu, C.; Wu, F.; Qi, T.; and Huang, Y. 2020. Graph enhanced representation learning for news recommendation. In *Proceedings of The Web Conference 2020*, 2863–2869.

Hu, L.; Li, C.; Shi, C.; Yang, C.; and Shao, C. 2020. Graph neural news recommendation with long-term and short-term interest modeling. *Information Processing & Management*, 57(2): 102142.

Järvelin, K.; and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4): 422–446.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8): 30–37.

Liu, D.; Lian, J.; Wang, S.; Qiao, Y.; Chen, J.-H.; Sun, G.; and Xie, X. 2020. KRED: Knowledge-aware document representation for news recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, 200–209.

Liu, Q.; Chen, N.; Sakai, T.; and Wu, X.-M. 2023. A First Look at LLM-Powered Generative News Recommendation. *arXiv preprint arXiv:2305.06566*.

Okura, S.; Tagami, Y.; Ono, S.; and Tajima, A. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1933–1942.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Qi, T.; Wu, F.; Wu, C.; Yang, P.; Yu, Y.; Xie, X.; and Huang, Y. 2021. HieRec: Hierarchical user interest modeling for personalized news recommendation. *arXiv preprint arXiv:2106.04408*.

Summers, C.; and Dinneen, M. J. 2019. Improved mixed-example data augmentation. In *2019 IEEE winter conference on applications of computer vision (WACV)*, 1262–1270. IEEE.

Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Voorhees, E. M.; et al. 1999. The trec-8 question answering track report. In *Trec*, volume 99, 77–82.

Wang, C.; and Blei, D. M. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 448–456.

Wang, H.; Zhang, F.; Xie, X.; and Guo, M. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*, 1835–1844.

Wang, S.; Guo, S.; Wang, L.; Liu, T.; and Xu, H. 2023. HDNR: A Hyperbolic-Based Debiased Approach for Personalized News Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 259–268.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019a. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*.

Wu, C.; Wu, F.; An, M.; Huang, J.; Huang, Y.; and Xie, X. 2019b. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2576–2584.

Wu, C.; Wu, F.; Ge, S.; Qi, T.; Huang, Y.; and Xie, X. 2019c. Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, 6389–6394.

Wu, C.; Wu, F.; Huang, Y.; and Xie, X. 2021. User-as-Graph: User Modeling with Heterogeneous Graph Pooling for News Recommendation. In *IJCAI*, 1624–1630.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou, G.; Zhu, X.; Song, C.; Fan, Y.; Zhu, H.; Ma, X.; Yan, Y.; Jin, J.; Li, H.; and Gai, K. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 1059–1068.