PDF Download
3701716.3735085.pdf
17 January 2026
Total Citations: 1
Total Downloads: 1179

Published: 08 May 2025

Citation in BibTeX format

WWW '25: The ACM Web Conference 2025
April 28 - May 2, 2025
Sydney NSW, Australia

Conference Sponsors:
SIGWEB

SHORT-PAPER

Enhancing News Recommendation with Hierarchical LLM Prompting

DANG KIEU, VinUniversity, Hanoi, Vietnam

DELVIN CE ZHANG, Pennsylvania State University, University Park, PA, United States

MINH DUC NGUYEN, VinUniversity, Hanoi, Vietnam

MIN XU, University of Technology Sydney, Sydney, NSW, Australia

QIANG WU, University of Technology Sydney, Sydney, NSW, Australia

DUNGDUY LE, VinUniversity, Hanoi, Vietnam

Open Access Support provided by:

Pennsylvania State University

University of Technology Sydney

VinUniversity

WWW '25: Companion Proceedings of the ACM on Web Conference 2025 (May 2025)
https://doi.org/10.1145/3701716.3735085
ISBN: 9798400713316

Enhancing News Recommendation with Hierarchical LLM Prompting

Hai-Dang Kieu
VinUniversity
HaNoi, VietNam
dang.kh@vinuni.edu.vn

Delvin Ce Zhang
The Pennsylvania State University
State College, United States
delvincezhang@gmail.com

Minh Duc Nguyen
VinUniversity
HaNoi, VietNam
duc.nm2@vinuni.edu.vn

Qiang Wu
University of Technology Sydney
Sydney, Australia
qiang.wu@uts.edu.au

Min Xu
University of Technology Sydney
Sydney, Australia
min.xu@uts.edu.au

Dung D. Le
VinUniversity
HaNoi, VietNam
dung.ld@vinuni.edu.vn

Abstract

Personalized news recommendation systems often struggle to effectively capture the complexity of user preferences, as they rely heavily on shallow representations, such as article titles and abstracts. To address this problem, we introduce a novel method, namely PNR-LLM, for **L**arge **L**anguage **M**odels for **P**ersonalized **N**ews **R**ecommendation. Specifically, PNR-LLM harnesses the generation capabilities of LLMs to enrich news titles and abstracts, and consequently improves recommendation quality. PNR-LLM contains a novel module, *News Enrichment via LLMs*, which generates deeper semantic information and relevant entities from articles, transforming shallow contents into richer representations. We further propose an attention mechanism to aggregate enriched semantic- and entity-level data, forming unified user and news embeddings that reveal a more accurate user-news match. Extensive experiments on MIND datasets show that PNR-LLM outperforms state-of-the-art baselines. Moreover, the proposed data enrichment module is model-agnostic, and we empirically show that applying our proposed module to multiple existing models can further improve their performance, verifying the advantage of our design.

CCS Concepts

• **Information systems** → **Personalization**; **Recommender systems**; **Language models**.

Keywords

News Recommendation, Personalization, Large Language Model.

ACM Reference Format:
Hai-Dang Kieu, Delvin Ce Zhang, Minh Duc Nguyen, Qiang Wu, Min Xu, and Dung D. Le. 2025. Enhancing News Recommendation with Hierarchical LLM Prompting. In *Companion Proceedings of the ACM Web Conference 2025 (WWW Companion '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3701716.3735085

WWW Companion '25, Sydney, NSW, Australia
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1331-6/2025/04
https://doi.org/10.1145/3701716.3735085

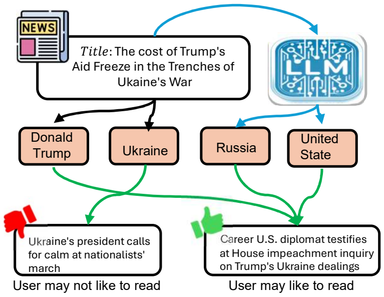

NEWS
Title: The cost of Trump's Aid Freeze in the Trenches of Ukaine's War
LLM
Donald Trump
Ukraine
Russia
United State
Ukraine's president calls for calm at nationalists' march
User may not like to read
Career U.S. diplomat testifies at House impeachment inquiry on Trump's Ukraine dealings
User may like to read

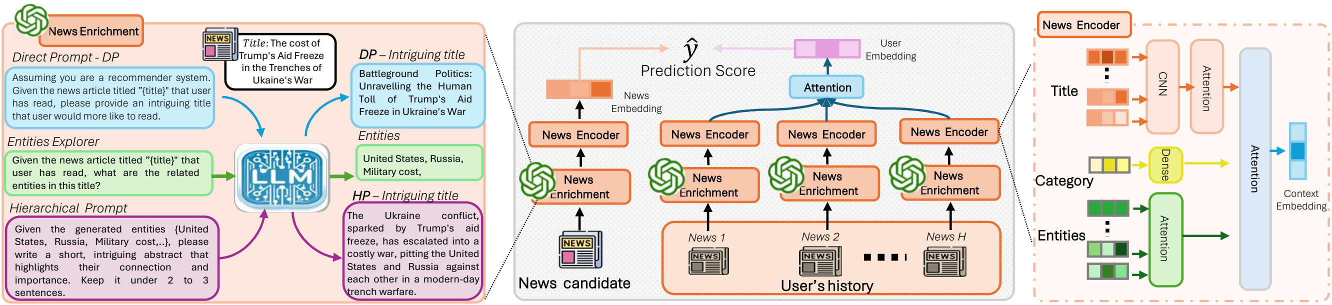
Figure 1: Illustration of a news connected to a relevant news through entities explored by LLM.

1 Introduction

With the rapid growth of online content, recommender systems play a crucial role in delivering personalized experiences to users. In the domain of news recommendation, the goal is to predict and recommend articles that align with individual users' interests. Given the fast-paced and dynamic nature of news, developing effective news recommender systems remains a challenging and important research problem [6, 13]. News recommendation systems face several challenges for example often contain sparse or ambiguous textual information, especially in short titles, which can limit accurate user preference modeling.

Traditional content-based methods model user preferences from article text using natural language processing and deep learning, employing CNNs, GRUs [8, 10, 12], and later enhancements like MINER [4] and LSTUR [1] with poly-attention and user embeddings for long-term interests. While effective, they overlook structured relationships between articles. Graph-based models such as DKN [9], DIGAT [7], and GLORY [14] enrich representations via entities and knowledge graphs, but often struggle with sparsity and incomplete data. Recently, PLMs and LLMs have improved text understanding and user modeling. Models like UNBERT [16] summarize reading history for better news encoding, while ONCE [5] leverage prompt-based LLMs for dynamic and semantically rich recommendations.

In this paper, we present our work on leveraging LLMs to enhance news recommendation by generating both deeper semantic information and relevant entities from articles, transforming shallow content into richer representations of user preferences.



Assuming you are a recommender system. Given the news article titled "{title}" that user has read, please provide an intriguing title that user would more like to read.

Given the news article titled "{title}" that user has read, what are the related entities in this title?

Given the generated entities {United States, Russia, Military cost,..}, please write a short, intriguing abstract that highlights their connection and importance. Keep it under 2 to 3 sentences.

Battleground Politics: Unravelling the Human Toll of Trump's Aid Freeze in Ukraine's War

United States, Russia, Military cost,

The Ukraine conflict, sparked by Trump's aid freeze, has escalated into a costly war, pitting the United States and Russia against each other in a modern-day trench warfare.

Figure 2: Illustration of our PNR-LLM.

Unlike prior approaches that primarily focus on summarization, our method utilizes LLMs to extrapolate information beyond the title, enriching news articles with contextual details and entity relationships to address data sparsity. By integrating extrapolated entities, we believe that LLM-generated outputs maintain deeper semantic understanding while remaining concise and aligned with the original title. To our knowledge, this is the first study to employ LLMs to generate auxiliary information specifically for news recommendation rather than summarization. We conducted extensive experiments on two benchmark datasets, demonstrating that our approach outperforms recent state-of-the-art methods, effectively capturing richer semantic representations for improved recommendation performance.

2 METHODOLOGY

In this section, we will introduce our proposed approach, PNR-LLM as illustrated in Fig. 2. A core component of this framework is a news enrichment module comprising hierarchical prompting steps that leverage LLM to generate not only deeper semantic meaning but also concise information from articles, transforming the content into a richer representation.

2.1 Problem Formulation

The click history sequence of a user u can be denoted as $H_u = [d_1, d_2, ..., d_H]$, where H is the number of historical news articles. Each news article d_i has a title that contains a text sequence T_i consisting of T word tokens and an entity sequence denoted by E_i consisting of E entities. The objective is to predict the probability of interest of a given candidate news article d_c and user u.

2.2 LLM-based News Enrichment

As demonstrated in several studies, LLMs can enhance content-based recommendations; however, excessive elaboration on titles may distort their original meaning, reducing the effectiveness of recommendation systems. To address this, a structured prompting approach is crucial to guide LLMs in generating rich and meaningful titles while preserving their intent. Our framework leverages the strengths of LLMs by introducing **Hierarchical Prompting**, a multi-step reasoning process that systematically refines title generation. This method enables LLMs to process information more efficiently, ensuring the generated titles remain contextually rich, engaging, and aligned with the original content. The workflow follows three key steps: **(1) Direct Prompting** – Formulate targeted prompts to leverage the capabilities of an LLM in generating an engaging and contextually appropriate title. **(2) Exploration** – Utilize the LLM's vast knowledge base to identify and extract relevant entities associated with the article. We believe that if an article has been part of the LLM's training data, it can accurately determine the most relevant title. **(3) Hierarchical Prompting** – Integrate the initially generated title with the extracted relevant entities, prompting the LLM to refine and generate a more compelling and well-aligned final title. This step is particularly beneficial for news recommendation, as it ensures that the generated title balances informativeness and engagement, preventing misleading or overly generalized headlines while maintaining relevance to user interests. After finish this step, each news article d_i has an updated title that contains a a text sequence $T'_i = [w_1, w_2, ..., w_{T'}]$ consisting of T' word torkens and a new entity sequence which denoted by $E'_i = [e_1, e_2, ..., e_{E'}]$ consisting of E' entities.

2.3 News Encoder

The news encoder module is designed to learn representations of news articles by incorporating enriched news titles, entities, and topic categories. Similarly to NAML[10], we initialize word embeddings using GloVe and convert a news article's text from a sequence of words $\{w_1, w_2, \ldots, w_{T'}\}$ into a sequence of word embedding vectors: $\mathbf{x}_n = [\mathbf{x}_1^{\omega}, \mathbf{x}_2^{\omega}, \ldots, \mathbf{x}_{T'}^{\omega}]$. A convolutional neural network (CNN) layer is then applied to learn contextual word representations:

$$\mathbf{X}_n = \text{CNN}(\mathbf{x}_n) = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{T'}]. \tag{1}$$

Next, an attention layer is used to identify the most informative words in the titles. Let the attention weight of the i^{th} word in a news title be denoted as α_i^t, which is computed as follows:

$$a_i^t = \mathbf{q}_t^\top \tanh(\mathbf{W}_t \mathbf{x}_i^t + \mathbf{b}_t), \qquad \alpha_i^t = \frac{\exp(a_i^t)}{\sum_{j=1}^{T} \exp(a_j^t)}, \tag{2}$$

where $\mathbf{W}_t$ and $\mathbf{b}_t$ are learnable parameters, and q_t is the attention query vector. The final news title representation is obtained by computing a weighted sum of contextual word representations: $\mathbf{r}_n = \sum_{i=1}^{T'} \alpha_i^t \mathbf{x}_i$. Additionally, we incorporate entity representations to enhance the semantic richness of news embeddings. Fine-grained entity information helps establish connections between

Table 1: Dataset statistics.

Dataset	Version	# News	# Behaviors	# Users
MIND	SMALL	65,238	347,727	50,000
	LARGE	161,013	24,155,470	1,000,000

news articles when they share common entities, thereby improving the efficiency of the recommendation system. Each article's entities are transformed into a sequence of entity embedding vectors: $\mathbf{x}_e = [\mathbf{x}^e_1, \mathbf{x}^e_2, \dots, \mathbf{x}^e_{E'}]$. To learn entity representations $\mathbf{r}_e$ for each news, we employ a self-attention network, followed by an attention mechanism to aggregate multiple entities and derive a unified entity representation, similar to Eq. 2. Finally, we concatenate embedding of category $\mathbf{r}_c$, vector entity representation $\mathbf{r}_e$ and vector context representation $\mathbf{r}_n$, followed by an additional attention mechanism to obtain the final news representation.

2.4 User Representation

The **user encoder module** derives user representations from the embeddings of their browsed news, as illustrated in Fig 2. Since different news articles vary in their informativeness, we incorporate a **news attention network** to prioritize the most relevant items. Given the **news embeddings** for a user's reading history H_u, an **attention layer** refines and computes the final **user embedding**, following a process similar to Eq. 2.

3 Experiments

Datasets. Following [14], we evaluate our model on two versions of MIND dataset. MIND, sourced from Microsoft News, contains anonymized user behavior logs. The MIND-LARGE version includes one million users with at least five news clicks recorded between October 12 and November 22, 2019, while MIND-SMALL consists of 50,000 users. Key dataset statistics are summarized in Table 1.

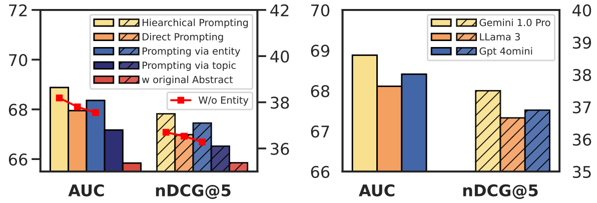
Experimental Setting and Implementation Details. We use negative sampling (4 negative samples per positive sample) with cross-entropy loss for model optimization. We adopt Adam optimizer [3] with a learning rate of 1×10^{-4}. We utilize the pre-trained TransE entity representations provided by MIND dataset. Enriched entity representations are initialized randomly. To ensure correctness and prevent redundancy, generated entities are verified using Wikipedia[1] (e.g., "United States" and "U.S." are treated equivalently). We use Gemini 1.0 Pro[2] for data enrichment. Following [14], models are evaluated using AUC, MRR, N@5 (nDCG@5), and N@10 (nDCG@10) over all impressions.

3.1 Empirical Evaluation

Comparison against Baselines. We evaluate the performance of various methods on MIND dataset. Our comparison includes conventional news recommendation models, including NAML [10], NPA [11], LSTUR [1], and NRMS [12], alongside recent models, including Glory [14], DIGAT [7], and GLoCIM [15]. As shown in Table 2, our model performs the best on MIND-SMALL and second best on MIND-LARGE. Incorporating enriched titles consistently improves the performance of our model. This improvement stems from the enhanced title representations, which better capture semantic connections between similar articles, benefiting text-based news encoders. Additionally, entity representations provide fine-grained information, further boosting recommendation accuracy.

Table 2: Performance on MIND. We reuse the results from existing works, and MINER and DIGAT do not have results on MIND-LARGE. † denotes the second best model.

Model	MIND-SMALL			MIND-LARGE		
	AUC	MRR	N@5 (N@10)	AUC	MRR	N@5 (N@10)
NPA	64.65	30.01	33.14 (39.47)	65.92	32.07	34.72 (40.37)
NRMS	65.63	30.96	34.13 (40.52)	67.66	33.25	36.28 (41.98)
NAML	66.12	31.53	34.88 (41.09)	66.46	32.75	35.66 (41.40)
LSTUR	65.87	30.78	35.15 (40.15)	67.08	32.36	35.15 (40.93)
MINER	68.07	32.93	N.A (42.62)	N.A.	N.A.	N.A. (N.A.)
DKN	62.90	28.37	30.99 (37.41)	64.07	30.42	32.92 (38.66)
DIGAT	67.82	32.65	36.25 (42.49)	N.A.	N.A.	N.A. (N.A.)
GERL	65.27	30.10	32.93 (39.48)	68.10	33.41	36.34 (42.03)
GLORY	67.68	32.45	35.78 (42.10)	69.04	33.83	37.53 (43.69)
GLoCIM	68.21†	33.02†	36.69† (42.78†)	**69.52**	**34.34**	**37.89 (44.08)**
PNR-LLM (ours)	**68.88**	**33.58**	**37.50 (43.49)**	69.32†	34.26†	37.81† (43.79†)



(a) Effect of different promptings on model performance.

(b) Effect of different LLMs for data enrichment.

Figure 3: Effect of different (a) promptings and (b) LLMs.

Analysis of Different Promptings. Following this, we compare our hierarchical prompting with different prompting strategies. To highlight the effectiveness of our method, we adopt the approach from GNR [2], which utilizes topics for news representation. Additionally, we evaluate the abstracts provided in the dataset. Furthermore, we compare title generation without hierarchical steps, including direct prompting and entity-based prompting. As shown in Fig. 3a, hierarchical prompting demonstrates the best performance, followed by direct prompting and entity-based prompting, indicating that structured, stepwise processing enhances recommendation accuracy. Topic-based prompting and using the original abstract show relatively lower performance, suggesting that entity enrichment contributes significantly to model effectiveness.

Analysis of Different LLMs for Data Enrichment. The effectiveness of different LLMs is compared in Fig. 3b. Gemini 1.0 Pro achieves the highest scores for both AUC and nDCG@5, outperforming LLaMA 3[3] and GPT-4o mini[4] . This suggests that most of LLMs could capture semantic information by following prompting techniques, leading to improved recommendation accuracy.

3.2 Ablation Study

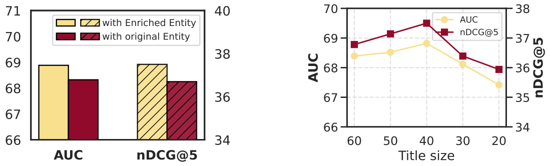
Effect of Pre-trained Language Models for Text Encoding: We replace our CNN-based text encoder with DistilBERT model.

[1] https://www.wikidata.org/w/api.php

[2] https://huggingface.co/papers/2312.11805

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B

[4] https://platform.openai.com/docs/models#gpt-4o-mini



(a) Effect of entity enrichment. **(b) Effect of title token size.**

Figure 4: Performance comparison of PNR-LLM with entities provided by MIND-SMALL and our method.

Table 3: Performance of different methods on MIND-SMALL using PLM as a new encoder.

Metrics	NPA	NRMS	LSTUR	CAUM	DIGAT	MINER	PNR-LLM w/ DistillBert
AUC	67.78	68.60	68.60	70.04	68.77	69.61	**70.54**
MRR	33.24	32.97	32.97	34.71	33.46	33.97	**35.26**
N@5	36.19	36.55	36.55	37.89	37.14	37.62	**38.07**
N@10	41.95	42.78	42.78	43.57	43.39	43.90	**44.23**

Table 4: Performance of existing approaches using our enriched title on MIND-SMALL, denoted with "++".

Metrics	NRMS	NRMS++	NAML	NAML++	LSTUR	LSTUR++
AUC	65.63	**67.86**	66.12	**68.01**	65.87	**66.98**
MRR	30.96	**32.61**	31.53	**32.75**	30.78	**32.16**
N@5	34.13	**36.02**	34.88	**36.29**	35.15	**35.82**
N@10	40.52	**42.26**	41.09	**42.46**	40.15	**41.41**

The results in Table 3 demonstrate that our approach with a PLM-based encoder consistently outperforms other methods, further validating its effectiveness. This improvement can be attributed to the fact that the original news titles are often sparse and lack sufficient contextual information, making it difficult for PLMs to fully capture user preferences. However, with the enriched and extrapolated titles generated by our method, the input becomes more concise, meaningful, and semantically informative, allowing the PLM to better understand and encode user interests, ultimately leading to enhanced recommendation performance.

Performance of Baseline Models Using Our Enriched Title. The title generated from our hierarchical prompting further improves conventional baseline models. In this ablation study, we replace original titles with our generated titles and evaluate baseline models. The results in Table 4 confirm the effectiveness of enriched titles in enhancing news recommendations.

Other ablation study. We also compare the enriched entities generated by our approach with the original entities provided in the MIND dataset. As shown in Fig. 4, incorporating these enriched entities further enhances recommendation performance, demonstrating the value of entity extrapolation. In addition, since the generated titles tend to be longer than the original ones, we conduct an analysis to determine the optimal token length for the new titles. Our results indicate that setting the token length to 40 achieves the best performance.

4 Conclusion

In this work, we enhance news recommendation by incorporating enriched titles by utilizing a hierarchical prompting strategy. Experiments demonstrate that these enrichments consistently improve performance across various baseline models. Additionally, PLM-based encoders further boost effectiveness. The results highlight the potential of leveraging LLMs for enriching textual and entity information. One future work is to connect the generated entities with a knowledge graph to improve recommendation accuracy.

References

[1] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 336–345.

[2] Shen Gao, Jiabao Fang, Quan Tu, Zhitao Yao, Zhumin Chen, Pengjie Ren, and Zhaochun Ren. 2024. Generative News Recommendation. In *Proceedings of the ACM on Web Conference 2024*. 3444–3453.

[3] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[4] Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. MINER: Multi-Interest Matching Network for News Recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 343–352.

[5] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. 452–461.

[6] Qijiong Liu, Jieming Zhu, Quanyu Dai, and Xiao-Ming Wu. 2024. Benchmarking News Recommendation in the Era of Green AI. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 971–974.

[7] Zhiming Mao, Jian Li, Hongru Wang, Xingshan Zeng, and Kam-Fai Wong. 2022. DIGAT: Modeling News Recommendation with Dual-Graph Interaction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6595–6607.

[8] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained Interest Matching for Neural News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 836–845.

[9] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-Aware Network for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference* (France) *(WWW '18)*. International World Wide Web Conferences, Republic and Canton of Geneva, 1835–1844.

[10] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, 3863–3869.

[11] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2576–2584.

[12] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Multi-Head Self-Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing(EMNLP)*. Association for Computational Linguistics, Hong Kong, China, 6389–6394.

[13] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Trans. Inf. Syst.* 41, 1, Article 24 (Jan. 2023), 50 pages.

[14] Boming Yang, Dairui Liu, Toyotaro Suzumura, Ruihai Dong, and Irene Li. 2023. Going Beyond Local: Global Graph-Enhanced Personalized News Recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore) *(RecSys '23)*. Association for Computing Machinery, USA, 24–34.

[15] Zhen Yang, Wenhui Wang, Tao Qi, Peng Zhang, Tianyun Zhang, Ru Zhang, Jianyi Liu, and Yongfeng Huang. 2025. GLoCIM: Global-view Long Chain Interest Modeling for news recommendation. In *Proceedings of the 31st International Conference on Computational Linguistics*. 6855–6865.

[16] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. 2021. UNBERT: User-News Matching BERT for News Recommendation.. In *IJCAI*, Vol. 21. 3356–3362.