Prompt-Based Generative News Recommendation (PGNR): Accuracy and Controllability

Xinyi Li[1(✉)], Yongfeng Zhang[2], and Edward C. Malthouse[1]

[1] Northwestern University, Evanston, IL, USA
XINYILI2024@u.northwestern.edu, ecm@northwestern.edu
[2] Rutgers University, Piscataway, NJ, USA
yongfeng.zhang@rutgers.edu

Abstract. Online news platforms often use personalized news recommendation methods to help users discover articles that align with their interests. These methods typically predict a matching score between a user and a candidate article to reflect the user's preference for the article. Given that articles contain rich textual information, current news recommendation systems (RS) leverage natural language processing (NLP) techniques, including the attention mechanism, to capture users' interests based on their historical behaviors and comprehend article content. However, these existing model architectures are usually task-specific and require redesign to adapt to additional features or new tasks. Motivated by the substantial progress in pre-trained large language models for semantic understanding and prompt learning, which involves guiding output generation using pre-trained language models, this paper proposes *Prompt-based Generative News Recommendation* (PGNR). This approach treats personalized news recommendation as a text-to-text generation task and designs personalized prompts to adapt to the pre-trained language model, taking the generative training and inference paradigm that directly generates the answer for recommendation. Experimental studies using the Microsoft News dataset show that PGNR is capable of making accurate recommendations by taking into account various lengths of past behaviors of different users. It can also easily integrate new features without changing the model architecture and the training loss function. Additionally, PGNR can make recommendations based on users' specific requirements, allowing more straightforward human-computer interaction for news recommendation.

Keywords: Large Language Model · Recommender Systems · Natural Language Processing · Information Retrieval

1 Introduction

The newspaper industry has experienced a steady and steep decline over the past decade. There have been widespread layoffs and closures, resulting in 'ghost

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14609, pp. 66–79, 2024.
https://doi.org/10.1007/978-3-031-56060-6_5

newspapers' and 'news deserts', where almost 200 out of 3,143 counties in the U.S have been left with no daily newspaper and 1,540 counties with only one weekly newspaper [1]. The demise of local newspapers is not only a commercial problem, but also a public and social problem. Communities without news organizations have seen an increase in government spending due to a lack of accountability [8]. Citizens who consume less news are unable to evaluate elected officials and are less likely to participate in voting. Reading news is one way for people to gain knowledge and to become more open-minded. Online platforms such as Google News and Microsoft News attract users to read news online [27]. However, in the current information-overloaded society, it is difficult for users to find news articles of interest from the massive set of articles published each day [14]. Therefore, it is important to design news RS to find articles of interest for users.

News RS typically involve three fundamental tasks: analyzing users' interests based on their past behaviors, comprehending news content by considering its contextual information, and predicting a user's preferences for candidate articles for personalized ranking [25]. News articles contain rich textual information, including their titles, bodies, and topics, making NLP techniques like Gated Recurrent Unit (GRU) [5], Long-short Term Memory (LSTM) [10], Convolutional Neural Network (CNN) [4], and attention mechanisms [21] popular choices for modeling users' interests and comprehending article contents [2,23,26].

There have recently been significant developments in pre-trained language models that can be used across various language tasks. These models can transfer knowledge from one task to another without extensive additional training, making them useful for fine-tuning for specific domains with little data compared to training a model from scratch. T5 [19], GPT-3 [3], BERT [7], and RoBERTa [15], are popular pre-trained language models that demonstrate impressive performance on NLP tasks. However, these models are usually large and complex, making it difficult to modify their structures or re-train them. To address this limitation and leverage the pre-trained language models, prompt learning [12], which provides specific prompts to guide the output generation, has been introduced. Prompt learning makes it possible to generate outputs that adapt to the input and has been an effective approach for various NLP tasks. Though some recent works attempted to explore pre-trained language models and prompt learning for news RS, they are mostly based on the slot filling paradigm using the BERT architecture or its variants [29], while the effectiveness of direct generative modeling based on large language models (LLMs) is still left unexplored.

Motivated by the power of LLMs and prompt learning, this paper presents a novel news recommendation model, PGNR (Prompt-based Generative News Recommendation), that treats the personalized news recommendation task as a text-to-text language generation task. In summary, the key contributions are:

- We introduce PGNR, a novel approach that predicts a user's preference for an article by applying personalized prompts that model the user's past behavior and article information. Unlike existing deep neural news RS, PGNR takes a generative recommendation paradigm and meanwhile allows various history lengths for different users throughout the training process.

- We incorporate language generation loss and ranking loss during model training to enhance the model's performance on the recommendation task.
- We demonstrate PGNR's flexibility in incorporating additional article features to improve recommendation performance without any need to modify its model architecture and training loss function.
- We investigate the potential for controlling news recommendations based on individual user requirements, which can enhance user experiences, improve human-computer interaction, and improve the interpretability of news RS with the help of personalized prompt learning. This is also the main advantage that distinguishes PGNR from existing news RS.

2 Related Works

Sequential News Recommendation. Since news articles are items with rich textual information, NLP techniques are often utilized to extract useful information from news contexts and understand users' interests [2,23,26]. Okura *et al.* [17] propose using a denoising autoencoder to study news representations and use a GRU network to model users' interests. An *et al.* [2] adopt CNN and the attention mechanism to learn a news representation from its title, topic and subtopic; learn a user's short-term representation using a GRU network; and learn a user's long-term representation using his/her ID embedding. The NRMS model proposed by Wu *et al.* [24] studies news representation from its title using a word-level, multi-head, self-attention and additive word-attention network, and studies a user's interest using a multi-head, self-attention network with the given historical clicked news sequence by a user. Wu *et al.* [22] also propose a neural news RS approach that studies news representation using an attentive multi-view network. Besides adopting various language models to better represent users and articles, An *et al.* [2] suggest focusing not only on users' short-term interests from their past behavior but also their long-term interests from their ID embedding, and Wu *et al.* [26] suggest being aware of temporal diversity when modeling the match between a user and an article. All of these deep news RS highly rely on the thriving of language techniques, but they are typically trained from scratch. In contrast, PGNR directly treats news RS as a text-to-text language generative task, utilizing prompt learning to generatively fine-tune the T5 language model [19] for news recommendation. Moreover, if further features are available, existing news RS models may require architectural modifications. However, PGNR can simply integrate them into its prompts without any need to modify its model architecture.

Pre-trained Language Models and RS. Motivated by the effectiveness of pre-trained language models and prompt learning techniques, RS researchers tend to formulate recommendation as a language task. Zhang *et al.* [28] convert the item-based recommendation to a text-based cloze task by modeling a user's historical interactions to a text inquiry. Li *et al.* [13] design personalized prompt learning for explainable recommendation by treating user and item IDs

as prompts. Cui *et al.* [6] propose M6-Rec, which converts a user's behavior to a text inquiry using general textual descriptions. Inspired by the T5 model [19] that studies a unified text-to-text generation model, Geng *et al.* [9] design a flexible and unified text-to-text paradigm called 'Pretrain, Personalized Prompt, and Predict Paradigm' (P5) for RS. Similar to P5 [9], our PGNR is also an encoder-decoder transformer that uses T5 as a backbone. However, different from P5, which relies on user IDs and item IDs [9] and may encounter challenges due to discrepancies between the semantic space of these IDs and that of the pre-trained language models, PGNR describes users' behaviors and news textually in the designed prompts. Zhang *et al.* [29] employ prompt learning for news recommendation by formulating it as a slot filling task for the [MASK] prediction. In contrast, PGNR formulates news recommendation as a direct generative recommendation task. Furthermore, to enhance the language model's performance on the recommendation task, PGNR innovatively integrates the ranking loss into the language generation loss throughout the training.

3 Methodology

Our objective is to estimate user u's preference $\hat{r}_{ui}$ for candidate article i by analyzing the user's interests based on previously read articles. This section describes PGNR and the loss function it employs to train the model parameters. The codes, dataset, and prompts are released at Github.[1]

3.1 Model Architecture

The PGNR employs the pre-trained T5 [19] as its backbone, utilizing transformer [21] blocks to build an encoder-decoder framework. As illustrated in Fig. 1, each user's sequential behavior is converted into a textual input sequence. The encoder processes this input sequence by summing the raw token embedding $X = \{x_1, x_2, \ldots, x_n\}$ with an additional position embedding to capture the token positional information.

Both the encoder and decoder consist of a stack of H identical layers. At the ℓ-th layer of the encoder, the output $X^{\ell-1}$ from the previous layer undergoes a multi-head self-attention mechanism, which generates a set of attention weights for each token based on its interactions with all other tokens in the sequence, resulting in:

$$MH(X^{\ell-1}) = [head_1, \ldots, head_h]W^O$$

where

$$head_i = Attention(X^{\ell-1}W_i^Q, X^{\ell-1}W_i^K, X^{\ell-1}W_i^V).$$

The attention applies the scaled dot product attention

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V,$$

[1] https://github.com/imrecommender/PGNR.

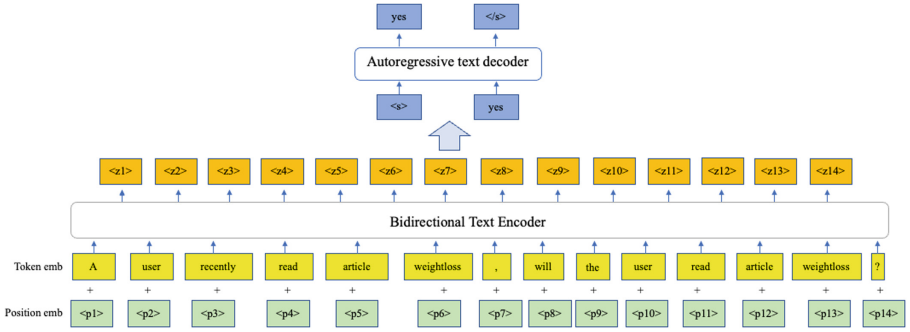


Fig. 1. PGNR utilizes an encoder-decoder framework, where a user's historical behavior is converted into a text inquiry and each news article is described textually, and then PGNR generates the answer to indicate a user's preference to a candidate article through an auto-regressive decoder.

where Q, V and K represent query, value and key of dimension d respectively. The output then undergoes a residual connection and layer normalization, resulting in O^ℓ, which is further passed through a position-wise feed-forward neural network to get

$$ReLU(O^\ell W_1 + b_1)W_2 + b_2.$$

The output of this feed-forward network is added to the original input tokens using a residual connection, and the resulting sequence is normalized using layer normalization to get X^ℓ. In the aforementioned formulas, $\theta = \{W^O, W_i^Q, W_i^K, W_i^V, W_1, W_2, b_1, b_2\}$ are model parameters.

The output of the encoder is a sequence of continuous representations denoted as $Z = \{z_1, z_2, \ldots, z_n\}$. Subsequently, given Z, the decoder engages in auto-regressive generation to produce an output sequence $Y = \{y_1, y_2, \ldots, y_m\}$. The decoder employs a linear transformation and a softmax layer to obtain a probability distribution over all tokens during the generation.

3.2 Training Loss Function

PGNR treats the news RS as a text-to-text language generation task, thus the language generation loss function (i.e., negative log-likelihood (NLL)) is applied to estimate the model parameters θ for auto-regressive model

$$L_{NLL} = -\sum_t \log P_\theta(y_t | y_{<t}, X),$$

where L_{NLL} measures how well the language model can generate the observed output sequence. However, RS often care about how well a model ranks items for a given user, so pair-wise or list-wise training are often applied to maximize the margin between a user u's preference for a clicked positive sample $\hat{r}_{u,pos}$ and that for an unclicked negative sample $\hat{r}_{u,neg}$. To improve the language model's

performance in news recommendation task, we incorporate L_{NLL} and Bayesian Personalized Ranking (BPR) loss L_{BPR} [18]

$$L = (1 - \lambda)L_{NLL} + \lambda L_{BPR},$$

where λ is a positive hyper-parameter, and L_{BPR} is defined as

$$L_{BPR} = - \sum_{(u,pos,neg)} \log(\sigma((\hat{r}_{u,pos} - \hat{r}_{u,neg}))),$$

considering all pair of clicked positive and unclicked negative items for users in training. Here, $\sigma(\cdot)$ denotes the sigmoid function.

4 Experiments

We aim to investigate several research questions about the performance of PGNR:

- **RQ1**: How does PGNR perform compared to other baselines in the task of sequential news recommendation?
- **RQ2**: Does PGNR possess the adaptability to incorporate additional article features to enhance recommendation performance?
- **RQ3**: Is it possible for PGNR to produce personalized recommendations according to the specific needs of users?
- **RQ4**: How does the ranking loss L_{BPR} impact the performance of PGNR?
- **RQ5**: How does the definition of additional article features, such as its diversity, have on the performance of PGNR?

4.1 Dataset

We conduct experiments over the Microsoft News dataset (MIND) [27], which is the only well-established benchmark for researchers in the field of news RS [29]. Following common settings [27], the impressions collected from November 9, 2019 to November 14, 2019 are used for model training, while the impressions on November 15, 2019 are used for validation and testing. A summary of the statistical details of the used dataset is provided in Table 1.

4.2 Baseline Methods

We compare the performance of PGNR with some representative baselines:

- **MostPop** and **RecentPop** [11]: recommend the top-K popular articles based on the number of real-time news clicks each article receives, with **RecentPop** considers the clicks in the past 24 h.
- **LSTUR** [2]: captures users' interests by modeling both their long- and short-term preferences.

Table 1. Statistics of MIND used for model evaluations.

#users	#news	#impressions	avg. history length	avg. click rate (%)	avg. title length	#category
141,935	71,671	297,715	23.56	0.10	10.77	18

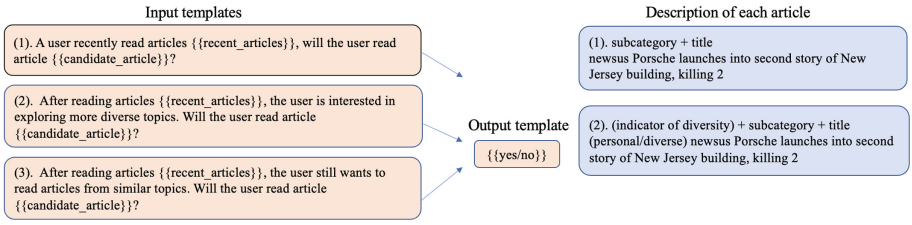


Fig. 2. The personalized prompts are created by designing input-target templates, wherein the relevant fields in the prompts are replaced with corresponding information from the raw data. In this study, the model denoted by PGNR $(i-j)$ employs input template (i) and article description (j). The phrasing of these prompts are common phrasing for recommendations [9,29].

- **TANR** [23]: employs a topic-aware news encoder, and utilizes an attention network to select essential words from the news title and important news from the user's past behavior.
- **NRMS** [24]: models users' and articles' representations via a multi-head self-attention network.
- **NAML** [26]: models users' and articles' representations via multi-view self-attention.
- **Prompt4NR** [29]: approaches news recommendation by formulating it as a slot filling task. For a fair comparison, we opt for its discrete action prompt because it closely corresponds to our designed prompts.

4.3 Implementation Details

The personalized prompts used are presented in Fig. 2. The personalization is from the depiction of a user's past behaviors and the detailed description of each article in input templates. For creating personalized recommendations, input template (1) is utilized. Meanwhile, input templates (2) and (3) are employed to evaluate the controllability of PGNR based on specific user requests, such as exploring more topics or reading articles from similar topics next. To maintain clarity and uniformity, a standardized target template of {yes/no} is adopted.

Different from existing deep neural news recommendations, where a user's preference is calculated through the dot-product of the news embedding and the user embedding, the user's preference for an article $\hat{r}_{ui}$ used for personalized ranking is estimated as the probability that the output from the auto-regressive decoder is 'yes'. During the inference stage, constrained text generation is used, given the prior knowledge that the target output is limited to either 'yes' or 'no'.

The T5 pre-trained checkpoint [19] serves as PGNR's backbone, consisting of 6 layers in both the encoder and decoder components, with a dimension size of 512 and an 8-headed attention mechanism. The SentencePiece [20] tokenizer is used with a vocabulary size of 32,128 sub-word units. A batch size of 16 is employed during training. To incorporate the ranking loss L_{BPR} for each user, a pair of positive and negative sample is generated every time, resulting in the generation of 32 input-target templates for each batch. The peak learning rate is set to 10^{-3}, and the maximum length of input tokens is restricted to 512. The warmup strategy is applied with the warmup stage set to the first 5% of all iterations to adjust the learning rate during training. PGNR is trained for 10 epochs with AdamW optimization on four NVIDIA RTX A6000 GPUs. For training with negative sampling, we used a positive-to-negative sample ratio of approximately 1:4.

5 Performance Evaluations

5.1 Sequential News Recommendations (RQ1)

We assess the effectiveness of PGNR in sequential news recommendations and ensure a fair comparison with other baseline methods by incorporating information on the subcategory and title of the news articles for all methods. For users with a history length shorter than the setting of the history length, existing deep neural news RS add vector embeddings to fill in the remaining articles to a fixed length while PGNR simply adds padding tokens at the end of the input template. As a result, PGNR is capable of considering different lengths of history throughout the training compared to deep neural baselines LSTUR, TANR, NRMS, and NAML.

The experimental results shown in Table 2 provide several insights about personalized news RS. Firstly, we see the methods based on popularity turn out to very competitive baselines for news recommendation. This could be attributed to users' tendency to favor popular articles. Furthermore, the measurement of news popularity could also be influenced by impression bias. Secondly, the performance of our approach without tuning the hyper-parameter λ in the training objective, PGNR (1–1), is comparable to other deep neural baselines. This is because our approach also considers users' interests from their historical behaviors, attempts to understand the articles read by a user before, and employs the attention mechanism. Lastly, adjusting the hyper-parameter λ leads to superior performance of PGNR (1–1)* over Prompt4NR, which adopts a slot filling paradigm for news recommendation. This observation emphasizes the effectiveness of treating news recommendation with a direct generative language model and underscores the potency of our methodology in introducing ranking loss and utilizing paired data to enhance the language model's performance in the news recommendation task. Overall, the performance indicates that treating personalized news RS as a language generation task and utilizing constrained text generation with the assistance of prompt learning is an effective approach for sequential news RS.

Table 2. Experimental results on Mean Reciprocal Rank (MRR), Normalized Discounted Cumulative Gain (NDCG@k), and Hit Ratio (HR@k). * indicates that the hyper-parameter λ in the training objective is adjusted; otherwise, $\lambda = 0$. Bold numbers represent best performance utilizing solely subcategory and title information of articles. ↑ indicates the improved performance when additional article feature is included. The statistical significance was assessed using the Student's t-test, with a significance level of $p < 0.1$.

Methods	MRR	HR@5	NDCG@5	NDCG@10
MostPop	0.2699	0.4899	0.2906	0.3510
RecentPop	0.2704	0.4939	0.2924	0.3519
LSTUR	0.2522	0.4715	0.2712	0.3352
TANR	0.2918	0.5519	0.3241	0.3876
NRMS	0.2847	0.5253	0.3101	0.3763
NAML	0.2943	0.5426	0.3235	0.3870
Prompt4NR	0.2997	0.5487	0.3249	0.3880
PGNR (1–1)	0.2924	0.5450	0.3218	0.3862
PGNR (1–1)*	**0.3084**	**0.5574**	**0.3387**	**0.4012**
PGNR (1–2)*	0.3168 ↑	0.5688 ↑	0.3454↑	0.4068↑

5.2 Incorporate Additional Article Features (RQ2)

Following a comprehensive examination of the dataset, it was discovered that 54% of the 297,715 impressions are associated with articles covering distinct topics compared to those read by users before. Interestingly, users still click on articles within these new topics. This discovery motivated us to improve recommendation accuracy by enhancing article descriptions through the introduction of a diversity signal.

For each user, an article within the impression is categorized as 'diverse' if its topic differs from those of the T most recently read articles by the user. Otherwise, it is classified as 'personal'. With PGNR, all that is required to include the diversity signal is to add this signal to the description using description (2) from Fig. 2. The results of PGNR (1–2)* from Table 2 indicate that PGNR can readily accommodate diversity signals, leading to enhanced recommendation performance without modifying the model architecture or training loss function.

5.3 Controllability of PGNR (RQ3)

We define a RS as controllable when it can tailor recommendations to individual users' preferences. This control is crucial because users might want to explore different topics after reading some articles, and their reading habits vary. Therefore, there should be a mechanism for users to express their preferences to the RS, enabling it to generate personalized recommendations aligned with their specific interests. Current news RS may not recognize this preference and keep

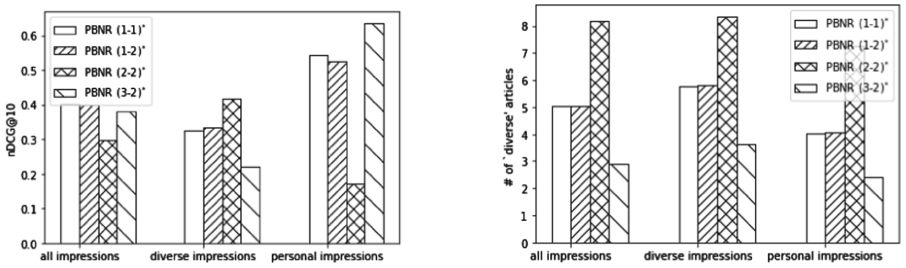
suggesting articles with similar content. We now test how our model can consider users' requests to enhance the news RS accordingly. In particular, if a user expresses a preference for exploring articles from a new topic, we expect PGNR to recommend such articles to the user. Conversely, if the user wants to read content similar to what they have previously engaged with, we anticipate the model to provide such recommendations.

To achieve this, we employ input template (2) and article description (2) as shown in Fig. 2 to assess the controllability of PGNR in recommending articles that are tagged as 'diverse' (i.e., from a new topic for the user). Similarly, we use input template (3) and article description (2) to evaluate whether PGNR can provide personal recommendations when necessary. To demonstrate PGNR's effectiveness in considering users' preferences in generating recommendations, we test its performance on three groups of testing impressions: (1) all impressions in the test dataset, (2) diverse impressions in the test set where all clicked articles are labeled as 'diverse', and (3) personal impressions in the test set where all clicked articles are labeled as 'personal'.

Figure 3 compares the performances among different prompts for news recommendation. Subfigure (a) demonstrates that PGNR (1–1)* and PGNR (1–2)* perform similarly in providing sequential news recommendations, while PGNR (2–2)* and PGNR (3–2)*, which aim to recommend articles based on users' preferences for topic diversity, perform worse than PGNR (1–1)* and PGNR (1–2)*. However, PGNR (2–2)*, which targets articles labeled as 'diverse', performs better than all other models in terms of diverse impressions, while PGNR (3–2)*, which targets articles labeled as 'personal', performs better than all other models in terms of personal impressions. Subfigure (b) also presents the number of recommended articles labeled as 'diverse' among the top-10 recommendations. As expected, PGNR (2–2)* suggests a more varied range of topics, while PGNR (3–2)* recommends a limited range. These findings confirm the controllability of PGNR in terms of enabling readers to tailor it to provide either personal or diverse recommendations based on readers' preferences, which is beyond the capability of most currently existing news RS.

5.4 Ablation Study on Ranking Loss (RQ4)

This section describes an ablation study of the training objective function to assess the impact of jointly training the ranking loss L_{BPR} and the language generation loss L_{NLL} in the training process. The results are presented in Fig. 4. We find that not considering L_{BPR} results in sub-optimal recommendations. If λ is too small, the model fails to utilize the benefits of adopting L_{BPR}. Conversely, if λ is too large, the performance of the language model in generating responses may be overlooked, leading to a decline in overall performance. This observation highlights the significance of jointly considering and carefully balancing the ranking loss and the language generation loss during training to enhance the language model's performance on recommendation task.



a Model performances on three groups of testing impressions.

b The number of 'diverse' articles within the top-10 recommendations.

Fig. 3. Evaluation of PGNR's controllability to make recommendations based on individual user requirements. PGNR (2–2)* and PGNR (3–2)* are omitted from Table 2 as they focus on evaluating the model's controllability through resampled training data based on requirements.

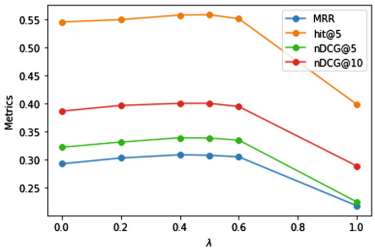
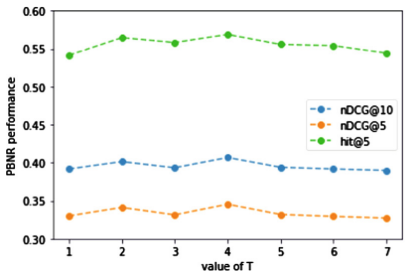


Fig. 4. PGNR performance on sequential recommendation with different λ values – weight on ranking loss.

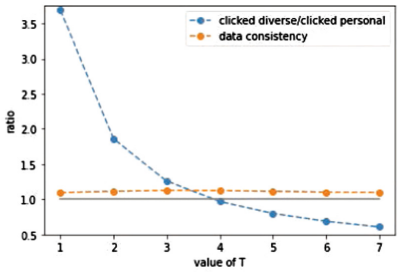
5.5 Influence of Definition of Diversity (RQ5)

We have shown that the PGNR can effectively incorporate extra article feature, such as whether an article is 'diverse' or 'personal' for the user, to improve its performance. In this section, we present our experimental analysis of how the definition of the diversity of articles affects the model's recommendations.

Figure 5a illustrates the performance of PGNR's recommendations with different threshold values, T, used to classify articles as either 'diverse' or 'personalized'. It shows a general trend that the performance decreases as T either increases or decreases, and we observe that $T = 4$ is an appropriate choice for defining articles' diversity to achieve the best recommendation performance. One underlying reason for the findings is the memorization ability of the large language model [16]. To assess the influence of an article's diversity, we analyzed the proportion of articles labeled as 'diverse' versus 'personal' that were clicked on, denoted as clicked diverse/clicked personal. Based on Fig. 5b, our results demonstrate that when T equals 4, the proportion of clicked articles labeled as 'diverse' is approximately equal to those labeled as 'personal', indicating no dominant label during the training process. This observation implies that the



a Variation in recommendation performance as T values change in defining articles' diversity.



b Visualization of training data samples and the consistency between the training data and the testing data with different T values.

Fig. 5. Evaluation of PGNR performance using different threshold value T in defining articles' diversity. For each user, an article within the impression is categorized as 'diverse' if its topic differs from those of the T most recently read articles by the user.

language model may memorize the 'diverse' signal when generating the output sequence for the testing data. Optimal performance of PGNR is achieved when the memorization capability of the large language model is reduced. Since language models have the capability of memorization, it is crucial to carefully define these additional features when incorporating additional features to enhance the model's performance.

6 Conclusion

This work introduces a novel generative news recommendation approach called PGNR that capitalizes on the strengths of pre-trained language models and prompt learning. Rather than considering news recommendation as a conventional task, we treat it as a text-to-text language generation task. To enhance the language model's performance in the recommendation domain, we incorporate both ranking and language generation losses during model training. Our experimental findings show that PGNR outperforms existing baselines in recommendation accuracy and does not require a fixed length of history for all users throughout the training process. This improvement can be attributed to the enhanced language understanding capabilities of pre-trained language models. Unlike other baselines that may necessitate a change in the model's architecture to integrate additional article features, PGNR remains unchanged in structure and training loss function, allowing easy integration of extra features through prompt design. PGNR also stands out from the existing news RS methods in its ability to produce personalized recommendations to meet users' specific needs, improving the human-computer interaction in the domain of news RS through the memorization capabilities of LLMs.

In the future, we will consider incorporating more and multimodal news information in prompts for news recommendation. Additionally, while our study

employs manually designed personalized prompts, future research could explore automated approaches to prompt design, which would allow the system to design prompts more efficiently and independently. The current token limit of 512 may pose challenges in handling long user news interactions, suggesting a potential avenue for future investigation. Furthermore, we aim to extend our exploration to leverage LLMs for enhancing other recommendation metrics, such as the recommendation diversity, in contrast to the emphasis on recommendation accuracy in the current work.

References

1. Abernathy, P.: The expanding news desert, center for innovation and sustainability in local media (2018)
2. An, M., Wu, F., Wu, C., Zhang, K., Liu, Z., Xie, X.: Neural news recommendation with long-and short-term user representations. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 336–345 (2019)
3. Brown, T., et al.: Language models are few-shot learners. Adv. Neural. Inf. Process. Syst. **33**, 1877–1901 (2020)
4. Chen, Y.: Convolutional neural network for sentence classification. Master's thesis, University of Waterloo (2015)
5. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
6. Cui, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: M6-Rec: generative pretrained language models are open-ended recommender systems. arXiv preprint arXiv:2205.08084 (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Gao, P., Lee, C., Murphy, D.: Municipal borrowing costs and state policies for distressed municipalities. J. Financ. Econ. **132**(2), 404–426 (2019)
9. Geng, S., Liu, S., Fu, Z., Ge, Y., Zhang, Y.: Recommendation as language processing (RLP): a unified pretrain, personalized prompt & predict paradigm (P5). arXiv preprint arXiv:2203.13366 (2022)
10. Graves, A., Graves, A.: Supervised Sequence Labelling with Recurrent Neural Networks, Long Short-term Memory, pp. 37–45. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24797-2
11. Ji, Y., Sun, A., Zhang, J., Li, C.: A re-visit of the popularity baseline in recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1749–1752 (2020)
12. Jin, W., Cheng, Y., Shen, Y., Chen, W., Ren, X.: A good prompt is worth millions of parameters? Low-resource prompt-based learning for vision-language models. arXiv preprint arXiv:2110.08484 (2021)
13. Li, L., Zhang, Y., Chen, L.: Personalized prompt learning for explainable recommendation. arXiv preprint arXiv:2202.07371 (2022)
14. Lian, J., Zhang, F., Xie, X., Sun, G.: Towards better representation learning for personalized news recommendation: a multi-channel deep fusion approach. In: IJCAI, pp. 3805–3811 (2018)
15. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

16. Mireshghallah, F., Uniyal, A., Wang, T., Evans, D.K., Berg-Kirkpatrick, T.: An empirical analysis of memorization in fine-tuned autoregressive language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1816–1826 (2022)
17. Okura, S., Tagami, Y., Ono, S., Tajima, A.: Embedding-based news recommendation for millions of users. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1933–1942 (2017)
18. Qi, T., Wu, F., Wu, C., Huang, Y.: PP-Rec: news recommendation with personalized user interest and time-aware news popularity. arXiv preprint arXiv:2106.01300 (2021)
19. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(1), 5485–5551 (2020)
20. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Wu, C., Wu, F., An, M., Huang, J., Huang, Y., Xie, X.: Neural news recommendation with attentive multi-view learning. arXiv preprint arXiv:1907.05576 (2019)
23. Wu, C., Wu, F., An, M., Huang, Y., Xie, X.: Neural news recommendation with topic-aware news representation. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 1154–1159 (2019)
24. Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., Xie, X.: 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), pp. 6389–6394 (2019)
25. Wu, C., Wu, F., Huang, Y., Xie, X.: Personalized news recommendation: methods and challenges. ACM Trans. Inform. Syst. **41**(1), 1–50 (2023)
26. Wu, C., Wu, F., Qi, T., Li, C., Huang, Y.: Is news recommendation a sequential recommendation task?. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2382–2386 (2022)
27. Wu, F., et al.: MIND: a large-scale dataset for news recommendation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 3597–3606 (2020)
28. Zhang, Y., et al.: Language models as recommender systems: evaluations and limitations. In: I (Still) Can't Believe It's Not Better! NeurIPS 2021 Workshop (2021)
29. Zhang, Z., Wang, B.: Prompt learning for news recommendation. arXiv preprint arXiv:2304.05263 (2023)