Improved Personalized Headline Generation via Denoising Fake Interests from Implicit Feedback

Kejin Liu[*‡]
Henan Institute of Advanced Technology, Zhengzhou University
Zhengzhou, China
liukejin@gs.zzu.edu.cn

Junhong Lian[*‡‖]
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
lianjunhong23s@ict.ac.cn

Xiang Ao[†‡§‖]
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
aoxiang@ict.ac.cn

Ningtao Wang
Independent Researcher
Hangzhou, China
ntwang25@gmail.com

Xing Fu
Independent Researcher
Hangzhou, China
fux008@gmail.com

Yu Cheng
Independent Researcher
Hangzhou, China
yu.cheng.info@gmail.com

Weiqiang Wang
Independent Researcher
Hangzhou, China
wang.weiqiang@gmail.com

Xinyu Liu[¶‖]
Institute of Computing Technology, Chinese Academy of Sciences
Beijing, China
liuxinyu@ict.ac.cn

Abstract

Accurate personalized headline generation hinges on precisely capturing user interests from historical behaviors. However, existing methods neglect personalized-irrelevant click noise in entire historical clickstreams, which may lead to hallucinated headlines that deviate from genuine user preferences. In this paper, we reveal the detrimental impact of click noise on personalized generation quality through rigorous analysis in both user and news dimensions. Based on these insights, we propose a novel **P**ersonalized **H**eadline **G**eneration framework via **D**enoising Fake Interests from **I**mplicit **F**eedback (**PHG-DIF**). PHG-DIF first employs dual-stage filtering to effectively remove clickstream noise, identified by short dwell times and abnormal click bursts, and then leverages multi-level temporal fusion to dynamically model users' evolving and multi-faceted interests for precise profiling. Moreover, we release **DT-PENS**, a new benchmark dataset comprising the click behavior of 1,000 carefully curated users and nearly 10,000 annotated personalized headlines with historical dwell time annotations. Extensive experiments demonstrate that PHG-DIF substantially mitigates the adverse effects of click noise and significantly improves headline quality, achieving state-of-the-art (SOTA) results on DT-PENS. Our framework implementation and dataset are available at https://github.com/liukejin-up/PHG-DIF.

CCS Concepts

• **Computing methodologies** → **Natural language generation**;
• **Information systems** → **Personalization**; **Data mining**.

Keywords

Personalized Headline Generation, User Preference Modeling, Implicit Feedback Analysis, Click Noise Denoising

ACM Reference Format:
Kejin Liu, Junhong Lian, Xiang Ao, Ningtao Wang, Xing Fu, Yu Cheng, Weiqiang Wang, and Xinyu Liu. 2025. Improved Personalized Headline Generation via Denoising Fake Interests from Implicit Feedback. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25), November 10–14, 2025, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746252.3761210

1 Introduction

Personalized headline generation has emerged as a pivotal strategy for enhancing user engagement on news platforms [2]. Prevailing methods typically condense user profiles from historical clickstreams into compact representations for personalization [1, 13, 34]. Despite the undeniable success of these personalized methods and continued progress, they predominantly rely on users' entire click history [26, 27, 32], overlooking a crucial characteristic of the clickstream: the inherent uncertainty of user click behaviors [8].

User click behaviors are influenced by a multitude of uncertain factors, extending beyond direct interest reflections [8, 29]. We regard clicks that are unrelated to personalization as "click noise" within historical clickstreams. Empirical analysis[1] shows that dwell

[*]Both authors contributed equally to this work.
[†]Corresponding author.
[‡]State Key Laboratory of AI Safety, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS).
[§]Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou, China.
[¶]High Performance Computer Research Center, ICT, CAS.
[‖]The authors are also with the University of Chinese Academy of Sciences, CAS.

CIKM '25, Seoul, Republic of Korea

ACM ISBN 979-8-4007-2040-6/2025/11
https://doi.org/10.1145/3746252.3761210

[1]Analysis based on PENS [2]. We randomly sampled users (e.g., U362229 in Figure 1) to calculate distributions of click history and corresponding news content length.

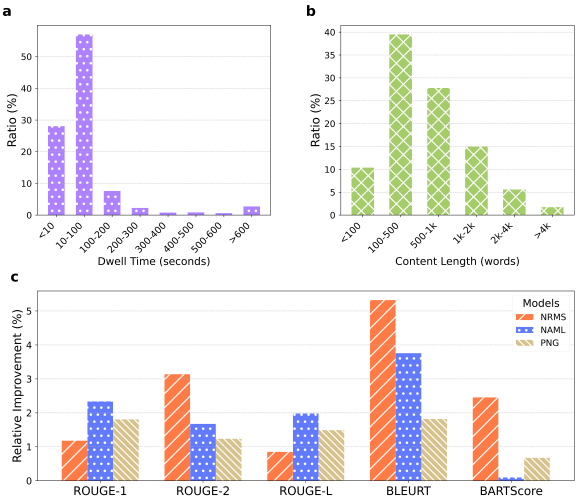


Figure 1: Fig. (a) and (b) present the ratio distribution of dwell times on the click history and news content length for a certain MSN user, respectively. Fig. (c) illustrates the relative improvement in evaluation after denoising.

time, the duration users spend reading news articles, effectively indicates click noise. A simple case study with a randomly selected MSN user reveals that 28.08% of clicked news exhibit dwell times less than 10 seconds, while merely 10.39% involve news content exceeding 100 words, as shown in Figure 1a-1b. Cognitive studies suggest even professional speed readers would struggle to fully comprehend 100-word content in under 10 seconds [21]. Moreover, transient news events (e.g., 2020 U.S. elections) induce temporal click surges containing incidental clicks from contextually uninterested users attracted by platform recommendations. By applying a simple, rule-based filtering method to exclude clicks with dwell times under 10 seconds and the top 0.1% of click-through rates during specific user impression log periods, we observed a significant improvement in personalization, as shown in Figure 1c.

Therefore, we attribute click noise in personalized user profiles to **the user dimension** and **the news dimension**. Users may rapidly exit from news due to misclicks or misleading headlines, resulting in clicks unrelated to their interests and forming click noise in the user dimension. In contrast, click noise in the news dimension typically arises from transient news events, which trigger a surge of clicks from non-specific users. These users are influenced by platform recommendations, not genuine interest. Both types of click noise hinder the accurate capture of user preferences, causing generated headlines to diverge from users' true interests. However, tackling historical click noise originating from both dimensions simultaneously remains challenging. A key challenge is the complexity of user click behaviors and the dynamic evolution of user interests, which complicates precise user profiling. Additionally, the absence of user historical dwell time data in existing personalized headline generation benchmark restricts further evaluation.

To remedy these challenges, we propose **PHG-DIF**, a novel **P**ersonalized **H**eadline **G**eneration framework via **D**enoising Fake Interests from **I**mplicit **F**eedback. PHG-DIF captures genuine user interests through a dual-filtering strategy, which filters out potential interference at both news-level and time-level. Concurrently, we capture users' multi-faceted preferences via dedicated modules for *Instantaneous Preference Learning (IPL)*, *Interest Evolution Analysis (IEA)*, and *Stable Interest Mining (SIM)*. These preferences are fused by multi-granular dynamic aggregation into a unified user representation that is subsequently injected at each decoding step of a breaking-news-aware pointer generator, thereby balancing factual accuracy with personalized headlines. Furthermore, we introduce **DT-PENS**, an extended benchmark derived from PENS [2]. DT-PENS includes complete dwell-time logs and nearly 10,000 human-validated personalized headlines for 1,000 carefully curated users, providing a comprehensive resource for mitigating historical click noise and personalized modeling.

The main contributions of this paper are summarized as follows:

- We propose a novel framework, PHG-DIF, which captures genuine user interests through dual-filtering and enhances personalization by multi-faceted user modeling.
- We introduce DT-PENS, an extended benchmark with nearly 10,000 personalized headlines for 1,000 carefully curated users annotated with historical dwell time, enabling more robust evaluation of personalized headline generation.
- Extensive experiments demonstrate PHG-DIF substantially mitigates the impact of click noise and significantly improves the quality of generated personalized headlines, achieving SOTA performance on DT-PENS benchmark.

2 Related Work

News platforms increasingly rely on automated headline generation to enhance user engagement and content distribution efficiency [2]. Early studies predominantly followed a content-compression paradigm, where headlines were either extracted or generated from news content to summarize the main idea, essentially forming a specialized instance of text summarization [12, 15, 16, 22]. Unified headlines often fail to accommodate individual user preferences, especially when readers focus on vastly different aspects of the same event, thus limiting user engagement [2]. Consequently, automated news headline generation is undergoing a paradigm shift from generic summarization towards personalization, a necessary transition driven by news platforms' pursuit of refined user operations and higher user retention.

PENS [2] was the first to formally define personalized news headline generation and introduced a large-scale public benchmark for offline evaluation. This spurred the development of various representative methods that encode user interests from click histories and integrate them into the headline generation process [1, 27, 34]. These approaches typically encode entire historical clickstreams into user-interest embedding generators, significantly outperforming generic models on metrics such as ROUGE and BLEU. GTP [26] further decomposed the generation process into a generic headline generation stage followed by personalized refinement. Concurrently, FPG [32] employed contrastive learning to constrain factual consistency, thus avoiding misleading attention-grabbing headlines.

The core of personalized generation lies in accurately modeling user interests. The field of recommendation systems has witnessed the evolution of several paradigms, from static vectors [9]

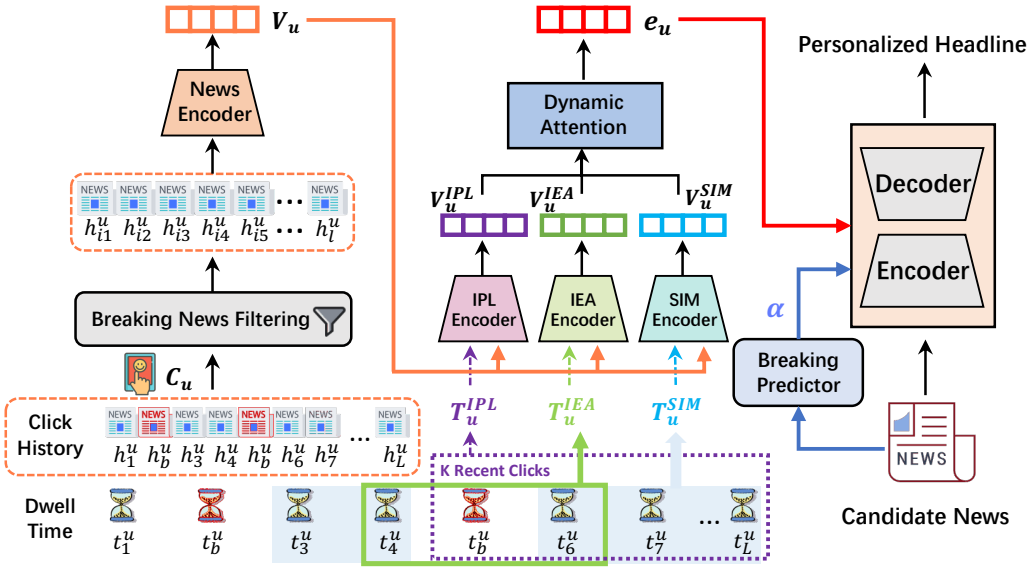


Figure 2: Overview of the proposed PHG-DIF framework.

to sequential attention mechanisms [35, 35], and subsequently to hierarchical interest modeling [18, 19]. Research on personalized headline generation has also begun exploring dimensions beyond content, including style, tone, and entity preferences. LaMP [23] introduced a personalization task leveraging authorial styles to achieve reader-side style matching, exploring personalization from a novel perspective. SCAPE [13] explicitly modeled user preferences across both content and stylistic dimensions, integrating short- and long-term interests to enhance fine-grained dynamic profiling.

However, limitations persist as most existing methods [1, 2] still rely on the "click equals interest" assumption for user modeling. Such an assumption often leads to the inadvertent incorporation of noise from users' historical clickstreams into their profiles. As a result, these profiles can deviate from genuine preferences. Indeed, studies show that noisy implicit feedback significantly undermines model robustness in news recommendation [8, 28, 30]. Furthermore, news content is highly time-sensitive, and user interests evolve rapidly. These factors make it challenging to accurately capture dynamic, multi-dimensional preferences. Filtering noise from user clicks and modeling their dynamic interests are key steps toward precise personalized headline generation. Nevertheless, this potential remains largely unexplored.

3 Problem Formulation

For any user u, the click history C_u is defined as an ordered sequence of N interactions: $C_u = [(h_1^u, t_1^u), (h_2^u, t_2^u), \ldots, (h_N^u, t_N^u)]$, where h_i^u denotes the headline of a news article clicked by user u, and t_i^u represents the corresponding dwell time. Our goal is to construct a valid click noise-irrelevant user click history $C_u^v = [(h_{k_1}^u, t_{k_1}^u), (h_{k_2}^u, t_{k_2}^u), \ldots, (h_{k_n}^u, t_{k_n}^u)] \subseteq C_u$ that captures u's genuine interests. Then, given a candidate news article with original headline h_x and body content b_x, we subsequently generate a personalized headline Y_u^v for user u based on C_u^v.

4 Our Framework

An overview of the PHG-DIF framework is shown in Figure 2.

4.1 User Modeling with Dual-Filtering

News-Level Filtering. We first construct breaking news candidates by selecting the top-M most clicked news headlines from the PENS dataset, forming set $B = \{h_j^b\}_{j=1}^M$, which serves as positive samples for training our breaking news predictor. For each user's click history $C_u = \{h_i^u\}_{i=1}^L$ where L denotes the total clicked news count, we mask dwell times ($t_i^u \leftarrow 0$) for $h_i^u \in B$, obtaining filtered sequence $H_u = (h_{i1}^u, \ldots, h_l^u)$.

Time-Level Filtering. Building upon the news-level filtered click history H_u, we further refine modeling of user's dwell time-based reading pattern. We align position indices between the filtered news sequence H_u and original dwell times sequence $T_u = (t_1^u, ..., t_L^u)$ via zero-padding. Each headline in H_u is first represented by its word embeddings and then encoded using a news encoder, which employs attention mechanism to generate the news representation V_u.

To effectively model the interaction between the user's historical dwell times and V_u, we introduce three specialized time-aware encoders. These encoders operate at multiple granularities to capture different aspects of user preference:

(1) Instant Preference Learning (IPL): We first capture the user's immediate interests by focusing on the most recent K click histories and their associated dwell times by IPL. Specifically, the dwell time T_u^{IPL} for the i-th item is defined as follows:

$$T_u^{\text{IPL}} = \begin{cases} t_i, & i \in [L-K+1, L], \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where t_i is the dwell time for the i-th item, L is total click histories. The user's instant preference representation v_u^{IPL} is then computed

as follows:

$$v_u^{\mathrm{IPL}} = w^{\mathrm{IPL}} \cdot V_u, \tag{2}$$

where w^{IPL} is learnable weight vector and V_u is the embedding matrix of the user's click histories.

By focusing on the most recent interactions, IPL effectively captures and prioritizes the user's current interests. For instance, even if a user historically favored the Golden State Warriors, their present focus on the Miami Heat is reflected in IPL, aligning with their current preferences.

(2) Interest Evolution Analysis (IEA): Then, we introduce IEA to model the dynamic changes in user interests based on their click history and dwell times within a certain time window n. By emphasizing temporal interaction patterns, IEA detects emerging interests, adapting to both abrupt and gradual changes in user behavior. The dwell time T_u^{IEA} for the i-th item and the user's evolving interest representation v_u^{IEA} are defined as:

$$T_u^{\mathrm{IEA}} = \begin{cases} t_i, & t_i \in [T-n, T], \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$v_u^{\mathrm{IEA}} = w^{\mathrm{IEA}} \cdot V_u, \tag{4}$$

where T is the current time, n is the length of the time window, and w^{IEA} is learnable weight vector.

(3) Stable Interest Mining (SIM): In SIM, we identify a user's long-term interests by analyzing news articles with consistently high dwell times across their click history. The threshold for stable interest is defined as the mean dwell time. For instance, if a user frequently engages with health and economics topics, these areas are considered their stable interests. The stable interest signals are formally defined as follows:

$$T_u^{\mathrm{SIM}} = \begin{cases} t_i, & t_i > mean(t), \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

$$v_u^{\mathrm{SIM}} = w^{\mathrm{SIM}} \cdot V_u, \tag{6}$$

where $mean(t)$ is the average dwell time across all interactions, and w^{SIM} is learnable weight vector.

4.2 Multi-Granular Dynamic Aggregation

We aggregate the embeddings from *IPL*, *IEA*, and *SIM* using a dynamic attention mechanism, enabling multi-level cross-time fusion to capture the user's multidimensional interest vector e_u. The matrices w^{i} ($i \in \{\mathrm{IPL}, \mathrm{IEA}, \mathrm{SIM}\}$) from the previous section are Min-Max scaled from T_u^{i}.

$$e_u = \mathrm{DynAttn}[v_u^{\mathrm{IPL}}, v_u^{\mathrm{IEA}}, v_u^{\mathrm{SIM}}]. \tag{7}$$

Here, *DynAttn* dynamically adjusts the weights of multi-granular interests, producing a unified representation e_u that reflects the user's true preferences.

4.3 Breaking-News-Aware Generator

The candidate news may be potential breaking news, which tends to be inherently more eye-catching. Users typically prefer objective headlines in such cases, requiring less personalization. A feasible approach is to train a BERT-based classifier B_ψ on the top–M% click-through headlines [3], as breaking news often revolves around similar topics, such as the lives and careers of public figures. During the headline generation phase, candidate news body b_x is processed by an encoder to obtain its dense representation v. Subsequently, a breaking news predictor utilizes v to yield a probability:

$$\alpha = B_\psi(v) \in [0, 1], \tag{8}$$

where a higher α indicates a greater likelihood that the article is breaking news. For breaking news, the headline should emphasize facts, while other news can be personalized to user preferences.

Inspired by Ao et al. [1, 2], we instantiate the personalized generator G_θ as a pointer-network decoder that conditions on both the encoded news representation v and the user interest vector e_u. At each decoding step t, the decoder state s_t attends over the encoder hidden states h_j, producing an attention distribution $a_{t,j}$ and a context vector $c_t = \sum_j a_{t,j} h_j$. Following See et al. [24], the final vocabulary distribution is a convex combination of the generator distribution $P_v(\cdot)$ and the copy distribution induced by a_t:

$$P(w) = \lambda_t \, P_v(w) + (1 - \lambda_t) \sum_{j: w_j = w} a_{t,j}, \tag{9}$$

$$\lambda_t = \sigma\big(W_\lambda [c_t; s_t; \alpha \, e_u] + b_\lambda\big), \tag{10}$$

where W_λ and b_λ are learnable parameters and $\sigma(\cdot)$ is the sigmoid function. The gating term λ_t is modulated by α: when $\alpha \to 1$ (breaking news), λ_t tends to favor factual copying from b_x; when $\alpha \to 0$, the generator leans toward user-tailored rewriting guided by e_u.

4.4 Model Optimization

PHG-DIF is optimized on the corpus $\mathcal{D} = (C_u, h_x, b_x, y)$ according to the procedure detailed in Algorithm 1. We first pre-train the encoder U_ξ on click-through labels, warm-up the decoder G_θ with *maximum likelihood estimation* (MLE), and fit the breaking-news classifier B_ψ by *binary cross-entropy*. The absence of personalized headline references makes it challenging to optimize personalized generation with purely supervised learning [26]. Following PNG [1], we therefore perform a policy-gradient fine-tuning step that maximizes the expected reward of sampled headlines. Specifically, the predicted breaking-news probability α is injected into the decoder gate to trade off personalization against factual fidelity, and an A2C search is adopted to estimate interim rewards. The optimization objective is defined as:

$$\mathcal{L}_{\mathrm{RL}} = -\mathbb{E}_{Y \sim G_\theta}\big[R(Y)\big], \tag{11}$$

where $R(Y)$ aggregates several headline-quality indicators.

5 Experimental Setup

5.1 DT-PENS Dataset

We introduce DT-PENS, a personalized news headline generation benchmark annotated with user dwell times, addressing the lack of user historical dwell time data for offline evaluation in PENS [2]. DT-PENS is a specialized dataset for personalized headline generation, featuring user dwell time annotations. DT-PENS is further developed in two phases from the anonymous user impressions in the training and validation sets of the original PENS.

5.1.1 The First Phase. We randomly sampled 1,000 users from the PENS validation data. For each user, we extracted their detailed click history, corresponding article dwell and exposure times (where

Algorithm 1: Model Optimization of PHG-DIF

Input: User Encoder U_ξ, Headline Generator G_θ, Breaking-News Predictor B_ψ, dataset $\mathcal{D}$
Output: Optimized parameters ξ, θ, ψ

Randomly initialize θ, ψ, ξ;
 # Phase 1: pre-train user encoder U_ξ;
while *not converged* **do**
  Sample (C_u) from $\mathcal{D}$;
  Update ξ via CTR prediction on C_u;
end
 # Phase 2: MLE warm-up of headline generator G_θ;
while *not converged* **do**
  Sample (C_u, b_x, h_x) from $\mathcal{D}$;
  Freeze ξ; compute $e_u = U_\xi(C_u)$, $v = \textsc{Enc}(b_x)$;
  Update θ by maximising $\log P_{G_\theta}(h_x \mid e_u, v)$;
end
 # Phase 3: train breaking-news predictor B_ψ;
for *each* (b_x, y) *in* $\mathcal{D}$ **do**
  $v = \textsc{Enc}(b_x)$; update ψ on (v, y) with BCE loss;
end
 # Phase 4: Policy-gradient fine-tuning (A2C);
while *not converged* **do**
  Sample (C_u, b_x) from $\mathcal{D}$; $e_u \leftarrow U_\xi(C_u)$; $v \leftarrow \textsc{Enc}(b_x)$;
  $\alpha \leftarrow B_\psi(v)$; $s_0 \leftarrow [\alpha e_u;\ v]$;
  Generate headline $Y \sim G_\theta$; compute reward $R(Y)$;
  Estimate advantage $\hat{A}_t$ and update θ, ξ via Eq. (11);
end
return ξ, θ, ψ

available), and news items that were exposed but not clicked. Subsequently, we leveraged Large Language Models (LLMs) to infer users' latent interests and generate preliminary personalized headlines based on these inferred interests and the candidate news. To ensure fairness, LLMs are not explicitly informed of the correlation between user click history and dwell time. Instead, we adopted a few-shot prompting strategy, which involved providing the models with partial anonymized historical interaction data and personalized headline samples derived from users in the original PENS test set, to guide the LLMs to learn and emulate the stylistic characteristics. We generated over 40K raw personalized headlines using multiple advanced LLMs, covering nearly 10K candidate news. By incorporating a rejection sampling mechanism, we then preliminarily filtered these to obtain a substantial corpus of candidate personalized headlines exhibiting high initial quality.

5.1.2 The Second Phase. We designed a rigorous and meticulous multi-level filtering pipeline to ensure the quality and suitability of the final reference headlines. Firstly, we removed overly long or short headlines, ensuring their length distribution is comparable to personalized headlines in the original PENS dataset [2]. Furthermore, we eliminated headlines found to be irrelevant to the news articles, containing factual inaccuracies, or exhibiting potential hallucinations. To implement this, we computed the semantic similarity between each generated headline and its corresponding news article body. Headlines falling below a predefined similarity threshold were flagged as potentially irrelevant or hallucinatory and subsequently discarded. Finally, all candidate headlines that passed the aforementioned automated filtering stages were submitted to human annotators for a final review. For each test instance, human annotators identified the headline that best reflected the user's historical preferences, designating it as the ground truth. The final DT-PENS dataset consists of 9,823 test instances from 1,000 unique readers. Detailed procedures for the dataset's construction are provided in Appendix B

5.2 Baselines

To comprehensively evaluate the performance of our proposed model, we select a diverse set of established baseline methods, encompassing both non-personalized and personalized methods.

The **non-personalized methods** include BART [10], a bidirectional and autoregressive transformer model, and T5-small [20], a smaller variant of the T5 model designed for efficient text generation. These methods generate headlines without considering user preferences, serving as a general performance benchmark.

For **personalized methods**, we consider three personalized news headline generation methods that integrate user-specific preferences, namely PENS-EBNR, PENS-NRMS, and PENS-NAML, as mentioned by Ao et al. [2]. Other personalized methods include PNG [1], which tailors generated news headlines to individual users based on multi-perspective interests, and GTP [26], which improves personalized headlines through pre-training and achieved SOTA performance on the original PENS benchmark.

5.3 Evaluation Metrics

To ensure a fair comparison with previous studies on personalized news headline generation Ao et al. [2], Song et al. [26], we evaluated the quality of generated personalized news headlines using several evaluation metrics. For lexical similarity between the generated and reference headlines, we employ ROUGE-n [14], which measures the overlap of n-grams and is widely used in text summarization evaluation[2]. To evaluate the semantic quality of the generated headlines, we utilize two model-based evaluation methods: BLEURT [25] and BARTScore [33]. BLEURT[3] captures semantic similarity and provides robust quality judgments. BARTScore[4] assesses fluency, grammar, and alignment with the input text by leveraging BART's language understanding and generation probabilities.

5.4 Implementation Details

To construct DT-PENS, we generated raw personalized headlines using multiple advanced LLMs, including o1-mini [7], GPT-4o [6], GLM-4-plus [5], and the Qwen-2.5 series [31]. All models were accessed via API endpoints with prompt engineering and a sampling temperature of 0.7. The collected raw headlines were then scored automatically by Qwen-2.5-72B , which served as an LLM judge proxy. Raw headlines receiving extremely low scores were filtered out and resampled. The prompt templates and further details are provided in Appendix B. In the dual-filtering, we first selected the top 0.1% of news articles in PENS [2] based on click-through

[2]We use the rouge package provided by https://github.com/pltrdy/rouge for evaluation.
[3]https://huggingface.co/spaces/evaluate-metric/bleurt.
[4]https://huggingface.co/ZoneTwelve/BARTScore.

Table 1: Our main experimental results. "*-w/o*" indicates component ablation (relative % change in parentheses).

Methods	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	BARTScore
• Non-personalized methods					
BART [10]	18.05	6.70	17.19	26.36	58.03
T5-small [20]	18.90	6.84	17.46	29.60	59.71
• Personalized methods					
PENS-EBNR [2]	23.15	7.19	21.33	44.79	62.28
PENS-NRMS [2]	22.89	7.01	21.15	43.40	60.27
PENS-NAML [2]	23.11	7.17	21.21	44.18	62.20
PNG [1]	23.24	7.28	21.47	45.63	62.29
GTP [26]	24.01	7.58	22.22	48.09	63.80
PHG-DIF$_{\text{ours}}$	**24.33***	**7.99***	**22.47***	**48.50***	**65.64***
-w/o IPL	24.08(-1.03%)	7.62(-4.63%)	22.23(-1.07%)	48.19(-0.64%)	63.86(-2.71%)
-w/o IEA	24.16(-0.70%)	7.64(-4.38%)	22.29(-0.80%)	48.21(-0.60%)	63.95(-2.58%)
-w/o SIM	23.65(-2.80%)	7.32(-8.39%)	21.76(-3.16%)	46.43(-4.27%)	62.78(-4.36%)

*The symbol * denotes the significance level with $p \leq 0.05$. **Bold** font indicates the best-performing method.*

rate (CTR), which formed the breaking news set for the news-level filtering. For time-level filtering, we set $k = 30$ for the *IPL*, applied a one-week sliding window for the *IEA*, and determined the *SIM* threshold based on the mean dwell time after excluding outliers (dwell time > 3000 s). The encoder was implemented with 8-headed attention, while the decoder uses beam search with a beam width of 5. The user model was pre-trained on a CTR prediction task, using a peak learning rate of $1e-5$. During PHG-DIF framework training, we applied a peak learning rate of $1e-7$ and executed an Advantage Actor-Critic (A2C) search with 16 sampled sequences. We use the NVIDIA A800 80GB GPU for our experiments.

6 Results and Analysis

In this section, we analyze our experiments to address the following research questions:

- **RQ1:** How does PHG-DIF perform compared to competitive non-personalized and personalized headline generation baselines?
- **RQ2:** What factors affect the performance of PHG-DIF?
- **RQ3:** How do users perceive the personalized news headlines generated by PHG-DIF?
- **RQ4:** Why does PHG-DIF achieve these improvements?

6.1 Overall Performance (RQ1)

To answer RQ1, we perform a comprehensive experimental evaluation of PHG-DIF. Table 1 reports the main results. Our proposed PHG-DIF framework outperforms all baseline methods across evaluation metrics, demonstrating that denoising users' historical click-streams markedly improve personalized headline generation.

Compared to non-personalized methods, all personalized methods, including PHG-DIF, exhibit substantial gains as we expected. This finding underscores that personalization enhances headline quality by aligning with user preferences, highlighting the value of user-oriented strategies in news headline generation.

In fine-grained comparisons among personalized methods, we observe that the GTP achieves significantly higher ROUGE-L and BARTScore than earlier pointer-network-based approaches due to its strong pre-training backbone. This suggests that generic headline-generation pre-training yields beneficial effects for personalization, consistent with the observations of Yang et al. [32]. Nevertheless, our pointer-generator-based PHG-DIF still surpasses the GTP. We attribute this advantage to the dual-filtering mechanism in our user modeling, which is designed to effectively remove noise from users' click histories. By applying this refined filtering, PHG-DIF distills a purer and more representative user interest profile from noisy interactions. These results indicate that in highly personalized contexts with complex user data, specialized noise filtering and dynamic interest modeling have the potential to surpass the generalization of pre-trained models.

6.2 Ablation Study (RQ2)

6.2.1 Impact of the Three Time-aware Encoders. PHG-DIF embeds three time-aware encoders, detailed in Section 4.1. One important question that arises is how each of these time-aware encoder modules contributes to the overall performance of PHG-DIF. To address this question, we conduct ablation studies on three variants, each omitting one encoder. The results are shown in Table 1, where "*-w/o*" denotes the removal of the corresponding component. The results yield three observations: 1) All three encoders are essential. Removing IPL, IEA, or SIM individually causes noticeable drops on every evaluation metric, demonstrating that modeling instantaneous, evolving, and stable interests is critical. 2) Stable interest mining (SIM) has the greatest impact. Its removal causes the biggest degradation, highlighting the importance of long-term preference modeling. 3) Instantaneous preference learning (IPL) and interest evolution analysis (IEA) are complementary. While SIM is most influential, IPL and IEA are indispensable. Removing either produces moderate yet non-negligible losses, confirming the need to capture real-time and evolving interests. Overall, the ablation results indicate that omitting any time-aware encoder degrades performance, thereby validating the effectiveness of the full design.

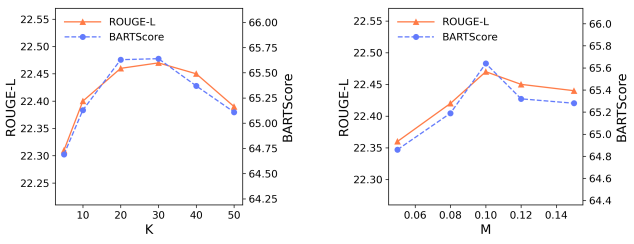
6.2.2 Ablation on Breaking News Handling. Breaking news is inherently compelling enough that readers will engage with it even

Table 2: Ablation study results for breaking news handling.

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEURT	BARTScore
PHG-DIF (full)	**24.33**	**7.99**	**22.47**	**48.50**	**65.64**
-w/o BF	24.12	7.63	22.25	48.20	63.91
-w/o BP	24.29	7.82	22.40	48.41	65.16

without personalized headlines. Our approach therefore prioritizes factual accuracy for such news. We conducted an ablation study with two variants to examine the impact of removing either of the two breaking-news components in PHG-DIF. Table 2 presents the results for two variants, where "*-w/o BF*" denotes removing the training-time *Breaking News Filtering* (BF), "*-w/o BP*" denotes removing the inference-time *Breaking Predictor* (BP). When BF is ablated, performance drops markedly on every metric, confirming that news-level filtering is indispensable for eliminating collaborative popularity noise and for preserving genuine user signals in interest modeling. Conversely, omitting the BP component, which compels the generator to personalize all headlines, results in some factual breaking news headlines being replaced by less precise rewrites, consequently lowering semantic and n-gram scores. This finding indicates that BP accurately identifies breaking news at inference, allowing the model to adapt its generation strategy, thereby preserving factual accuracy for these items instead of invariably prioritizing personalization. The importance of safeguarding headline factuality for user experience, as evidenced by our results, aligns with the findings of Yang et al. [32].

6.2.3 Influence of History Length K in IPL. In this section, we investigate the impact of the historical click window size K within the IPL module on overall model performance. As illustrated in Figure 3a, our experimental results show that varying K across the set $\{5, 10, 20, 30, 40, 50\}$ leads to a rapid improvement in model performance with increasing K, which peaks at $K = 30$. We posit that an excessively small K may prevent the model from capturing a sufficiently broad recent user history, resulting in inadequate modeling of instant preferences. Conversely, further increasing K beyond this optimum might lead to functional overlap with modules designed for mid-to-long-term interests (e.g., IEA, SIM). This overlap could diminish IPL's distinct role in capturing short-term, immediate interests and, furthermore, potentially degrade performance due to information redundancy or introduced noise.



(a) (b)

Figure 3: Impact of IPL history length K and breaking news threshold M on model performance.

Table 3: Results of the user study with rankings.

Methods	Fluency ↓	Consistency ↓	Attractiveness ↓
T5-small	2.95	3.02	3.25
PNG	2.11	3.68	3.09
GTP	2.84	1.68	1.95
PHG-DIF$_{\text{ours}}$	**2.10**	**1.62**	**1.77**

6.2.4 Sensitivity to Threshold M. Following GTP [26], we define breaking news as items ranking in *the top-M%* by click-through rate (CTR). This definition aims to distinguish user clicks primarily driven by trending events or platform recommendations from those reflecting pure personal interest. A similar idea is also reflected in FPG [32], which restricts training to news items clicked by a limited number of users. To ascertain the optimal value for M, we performed a sensitivity analysis on breaking news click-through rate threshold. We evaluated M across a range from 0.05 to 0.15, affecting the label of 5,606,741 news articles in the dataset. The results revealed optimal model performance at $M = 0.10$ (corresponding to 0.10% CTR threshold). We infer that an excessively low M (e.g., < 0.10) might be overly stringent, potentially leading to the omission of some breaking news that has garnered significant public attention. Conversely, an excessively high M (e.g., > 0.10) could cause a substantial volume of regular news to be misclassified as breaking news, which would then unnecessarily bypass the personalized rewriting process. This misclassification not only dilutes personalized user interest signals but also compromises the overall effectiveness of personalization. Hence, a moderate M best balances popularity bias suppression, factual headlines for breaking news, and robust personalization elsewhere.

6.3 User Study (RQ3)

To gain deeper insights into RQ3 and further evaluate the practical effectiveness of personalized news headlines generated by PHG-DIF, we conducted a user study. We recruited 5 native English-speaking graduate students and compensated them according to our approved participant guidelines. Participants were asked to select 100 news articles from a preselected list, thereby constructing their user preference profiles. This process aimed to simulate their personalized historical clickstreams. Subsequently, four different personalized models generated headlines for 20 unseen news articles based on each participant's historical click data. Participants assessed the generated headlines across three dimensions: **fluency**, **consistency**, and **attractiveness**, and ranked the headlines produced by each model (with *1 being the best* and *4 the worst*). Notably, participants were unaware of the source of each headline and were allowed to assign the same ranking score to different headlines for the same news article. Finally, we calculated the average ranking of headlines generated by each model to obtain a ranking score, as shown in Table 3. As observed, PHG-DIF achieved the highest scores in fluency, consistency, and attractiveness, suggesting that its personalized headlines are more aligned with users' true interests and have a greater potential to engage readers.

Table 4: A case on personalized headline generation affected by click noise. Red text highlights content related to click noise, and blue text represents the user's true interests.

Click History		Dwell Time
British Ambassador to the U.S., Kim Darroch, resigns after Trump criticism		366s
There's a democratic civil war brewing over decriminalizing migration		115s
A wooden sculpture of Melania Trump was unveiled on the banks of the Sava River		3s
The true cost of high deductible health care plans		7s
Report: Durant, Irving planned to team up before 2018-19 season began		249s
What were the Warriors thinking on Stephen Curry's final shot?		5s
The Knicks failed to sign Kevin Durant and Kyrie Irving		486s
• ***Case 1***		
Original Headline:	Here's when social security benefits could be cut	
PNG:	Possible impact of Melania Trump's social security reductions	✗
GTP:	The benefits of Trump's proposed social security reductions	✓
Ours:	Exploring implications of Trump's proposed social security cuts	✓
• ***Case 2***		
Original Headline:	Calling BS on Stephen A. Smith's explosive claims on First Take	
PNG:	Stephen Curry calls Stephen A.Smith's claims about the Warriors	✗
GTP:	Kevin Durant addresses Stephen A. Smith's claims on his Warriors Era	✗
Ours:	Kevin Durant denies Stephen A. Smith's claims about his departure	✓

6.4 Case Study (RQ4)

To investigate why our PHG-DIF leads to improvements, we conducted a case study comparing the performance of the baseline PNG [1] and GTP [26]. Table 4 outlines a user's click history, replete with dwell times that differentiate genuine interests from incidental clicks (i.e., click noise), alongside headlines generated by our proposed PHG-DIF and the baseline methods. The click history reveals a typical pattern where short-duration clicks represent click noise, contrasting with longer engagements that signify true user interests. Our analysis focuses on how effectively each model discerns these nuances. We found that the PNG erroneously interprets mistakenly clicked news as genuine user interests during the user modeling. This results in headlines that are both misaligned with user preferences and factually inaccurate. While the GTP produces headlines that are consistent with the news content, it still incorporates fake user interests. For instance, the generated headlines include information about the Warriors, despite the user's actual interest being in Kevin Durant and the Knicks. In contrast, our PHG-DIF method effectively filters out click noise, accurately capturing the user's true interests and generating headlines that better align with user preferences. The empirical case Study results underscore PHG-DIF's significant potential for enhancing user experience in real-world news recommendation systems, particularly in addressing the persistent challenge of click noise.

7 Conclusion

In this paper, we present PHG-DIF, a novel framework to tackle the challenges of click noise in user historical clickstreams. PHG-DIF employs a robust dual-filtering strategy that removes click noise at both news-level and time-level, isolating genuine user interests. Three specialized time-aware encoders then capture instantaneous, evolving, and stable preferences, yielding a precise user representation from noisy interaction data. By denoising fake interests from implicit feedback, PHG-DIF effectively improves the precision of user profiles, leading to more relevant and accurate personalized headline generation. We further introduce DT-PENS, a new personalized headline generation benchmark with dwell time annotations for better evaluation. Extensive experiments on DT-PENS demonstrate that PHG-DIF significantly enhances headline quality and outperforms multiple competitive baseline methods.

Acknowledgments

The research work is supported by National Key R&D Plan No. 2022YFC3303303, the National Natural Science Foundation of China under Grant (No. U2436209, 62476263), the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDB0680201, Beijing Nova Program 20230484430, the Innovation Funding of ICT, CAS under Grant No. E461060.

A Limitations and Discussion

Despite the significant improvements in headline quality and personalization achieved by our PHG-DIF framework for personalized news headline generation, we acknowledge several limitations that warrant further exploration in future research. Firstly, our PHG-DIF framework depends on historical clicks and dwell times. When interactions are sparse or noisy, the dual filtering module may not

Raw Personalized Headline Generation Prompt

You are to generate a personalized news headline tailored to a user's interests. The headline should attract user attention and remain factually accurate.
Please follow the instructions below:

Inputs:

- User history: {`clicked headlines`, `dwell times`, `exposure times`, `uninterested headlines`}
- Candidate news: {`news body`}
- Reference examples: {`original news body`, `original headline`, `human-written headline`}

Steps:

1. Identify user interests from the provided history.
2. Extract relevant facts from the candidate news.
3. Match user interests to news content.
4. Refer to example headlines for style and tone.
5. Write a concise, personalized headline reflecting both the user's interests and news facts.

Output (JSON format):

```
{
  "matched_interest_points": ["..."],
  "generated_headline": "...",
  "description": "A brief rationale explaining how
       the headline achieves factual accuracy and
       personalization."
}
```

(a)

Automated Headline Evaluation Prompt

You are an impartial judge. Evaluate the quality of the provided candidate headline based solely on the candidate news content and the scoring criteria below. Assign a single score that best reflects its overall quality and relevance.

Inputs:

- Candidate news: {`news body`}
- Candidate headline: {`generated headline`}

Scoring Criteria:

0: Irrelevant, hallucinated, or invalid headline; not related to news facts.
1: Weakly related; overly generic or misleading; lacks informativeness.
2: Partially relevant; covers some key facts but misses main points or is ambiguous.
3: Good coverage; accurate, relevant, and clear; partially personalized or engaging.
4: Very good; highly relevant, factually correct, concise, and well-aligned with user interests.
5: Excellent; fully relevant, factually perfect, maximally personalized, and editorially optimal.

Output (JSON format):

```
{
  "score": int,
}
```

(b)

Figure 4: Prompt templates for instructing LLMs in the DT-PENS dataset construction.

recover true interests, which can lower headline quality. The temporal fusion models gradual preference change yet may lag under abrupt shifts, such as breaking news. Our evaluation uses DT-PENS from a single platform with a limited cohort. Transfer to other platforms and user groups is uncertain. Cold start and missing click cases are underrepresented in current benchmarks. We recognize that the characteristics of this dataset may not be fully applicable to certain platforms or user groups. Specifically, for cases involving sparse or missing click behavior, such as cold-start scenarios, these issues extend beyond the scope of current PENS and DT-PENS benchmarks and are left for future work.

B DT-PENS Dataset Details

This section further details the construction pipeline of the DT-PENS dataset. To acquire raw personalized headlines from multiple advanced LLMs, we employ a *few-shot prompting* strategy [17]. We design a prompt template, as shown in Figure 4a, to guide LLMs in generating raw personalized headlines. The LLMs are provided with detailed historical click data for each user, including headlines of clicked news, dwell and exposure times, and unclicked headlines, along with few-shot examples from the original PENS test set. This setup enables LLMs to discern user interests and generate personalized headlines from the user's inferred perspective, aligning with their genuine interests.

Before forwarding over 40K raw personalized headlines for final human vetting, we adopt the *LLM-as-Judge* paradigm [4, 11] to perform automatic scoring. We first devise a headline-quality rubric and instantiate it in a prompt template, as illustrated in Figure 4b. The judge LLM is then instructed to assign each headline an integer score from 0 to 5, where higher values indicate better headline quality. Headlines scoring below 2 are discarded and re-sampled until the target corpus size is reached, substantially reducing the downstream manual workload.

We post-process the outputs from the two-stage LLMs using a JSON parser, and any output that fails to parse is immediately discarded. All retained candidates are subsequently examined by human annotators. To guarantee objectivity, at least three annotators independently review each candidate. A candidate personalized headline is accepted only if at least two-thirds of the annotators concur that it meets the quality criteria.

GenAI Usage Disclosures

We acknowledge the use of Generative AI (GenAI) tools in preparing this work. Specifically, we used multiple advanced LLMs to generate raw samples for the DT-PENS dataset (see Section 5.4), with all GenAI-generated content manually reviewed and validated. GenAI tools were also used for grammar checking and language refinement. We are fully responsible for all content and confirm this disclosure complies with ACM policy on GenAI use.

References

[1] Xiang Ao, Ling Luo, Xiting Wang, Zhao Yang, Jiun-Hung Chen, Ying Qiao, Qing He, and Xing Xie. 2023. Put Your Voice on Stage: Personalized Headline Generation for News Articles. *ACM Transactions on Knowledge Discovery from Data* 18, 3 (2023), 1–20.

[2] Xiang Ao, Xiting Wang, Ling Luo, Ying Qiao, Qing He, and Xing Xie. 2021. PENS: A dataset and generic framework for personalized news headline generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 82–92.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[4] Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. Can LLM be a Personalized Judge? *arXiv preprint arXiv:2406.11657* (2024).

[5] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793* (2024).

[6] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[7] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720* (2024).

[8] Hao Jiang, Chuanzhen Li, and Mingxiao An. 2024. Time Matters: Enhancing Pre-trained News Recommendation Models with Robust User Dwell Time Injection. *arXiv preprint arXiv:2405.12486* (2024).

[9] Hyoung R Kim and Philip K Chan. 2003. Learning implicit user interest hierarchy for context in personalization. In *Proceedings of the 8th international conference on Intelligent user interfaces*. 101–108.

[10] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. doi:10.18653/v1/2020.acl-main.703

[11] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579* (2024).

[12] Zhengpeng Li, Jiansheng Wu, Jiawei Miao, and Xinmiao Yu. 2022. News headline generation based on improved decoder from transformer. *Scientific Reports* 12, 1 (2022), 11648.

[13] Junhong Lian, Xiang Ao, Xinyu Liu, Yang Liu, and Qing He. 2025. Panoramic Interests: Stylistic-Content Aware Personalized Headline Generation. In *Companion Proceedings of the ACM on Web Conference 2025*. 1109–1112.

[14] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[15] Ling Luo, Xiang Ao, Yan Song, Feiyang Pan, Min Yang, and Qing He. 2019. Reading like HER: Human reading inspired extractive summarization. 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)*. 3033–3043.

[16] Ramesh Nallapati, Bowen Zhou, Cicero Nogueira Dos Santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. 280–290.

[17] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[18] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5446–5456.

[19] Mingjie Qian, Yongsen Zheng, Jinghui Qin, and Liang Lin. 2023. HutCRS: Hierarchical user-interest tracking for conversational recommender system. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10281–10290.

[20] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[21] Keith Rayner, Elizabeth R Schotter, Michael EJ Masson, Mary C Potter, and Rebecca Treiman. 2016. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest* 17, 1 (2016), 4–34.

[22] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 379–389.

[23] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When Large Language Models Meet Personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7370–7392.

[24] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. doi:10.18653/v1/p17-1099

[25] Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7881–7892.

[26] Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. 2023. General then Personal: Decoupling and Pre-training for Personalized Headline Generation. *Transactions of the Association for Computational Linguistics* 11 (2023), 1588–1607.

[27] Xiaoyu Tan, Leijun Cheng, Xihe Qiu, Shaojie Shi, Yuan Cheng, Wei Chu, Yinghui Xu, and Yuan Qi. 2024. Enhancing Personalized Headline Generation via Offline Goal-conditioned Reinforcement Learning with Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5762–5772.

[28] Wenjie Wang, Fuli Feng, Xiangnan He, Liqiang Nie, and Tat-Seng Chua. 2021. Denoising implicit feedback for recommendation. In *Proceedings of the 14th ACM international conference on web search and data mining*. 373–381.

[29] Ruobing Xie, Cheng Ling, Yalong Wang, Rui Wang, Feng Xia, and Leyu Lin. 2021. Deep feedback network for recommendation. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*. 2519–2525.

[30] Ruobing Xie, Lin Ma, Shaoliang Zhang, Feng Xia, and Leyu Lin. 2023. Reweighting Clicks with Dwell Time in Recommendation. In *Companion Proceedings of the ACM Web Conference 2023*. 341–345.

[31] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[32] Zhao Yang, Junhong Lian, and Xiang Ao. 2023. Fact-Preserved Personalized News Headline Generation. In *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1493–1498.

[33] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems* 34 (2021), 27263–27277.

[34] Kui Zhang, Guangquan Lu, Guixian Zhang, Zhi Lei, and Lijuan Wu. 2022. Personalized headline generation with enhanced user interest perception. In *International Conference on Artificial Neural Networks*. Springer, 797–809.

[35] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.