

# Personalized News Headline Generation via Multi-Agent User Simulation

Li Kang   Yuhan Zhao   Li Chen

Hong Kong Baptist University

{likang, csyhzha, lichen}@comp.hkbu.edu.hk

## Abstract

Personalized news headline generation has emerged as a promising avenue for optimizing user engagement and satisfaction in online news consumption. However, prevailing methods predominantly adopt an indirect modeling paradigm, inferring user preferences from related tasks rather than capturing their direct headline preferences. Direct modeling remains underexplored due to two primary challenges: (1) the scarcity of large-scale personalized headline datasets and (2) the risk of factual inaccuracies when models overly cater to subjective user preferences. To address these challenges, we propose **USG**, a novel multi-agent user simulator headline generation framework. USG introduces three specialized agents: the *User* provides subjective feedback to guide headline personalization; the *Editor* enforces editorial standards (i.e., factual consistency and no clickbait); and the *Writer* iteratively refines candidates via a Critique-and-Refine algorithm guided by feedback from the User and Editor agents. The multi-agent interactions yield optimal personalized headlines and reasoning chains distilled from the interaction trajectories, forming a high-quality dataset. This dataset directly supervises the fine-tuning of a standalone generator. Extensive experiments demonstrate that USG outperforms the strongest baseline by nearly 6% on personalization, achieving state-of-the-art performance on both personalization and factual consistency, while enabling efficient end-to-end inference. Our code is available at <https://github.com/Leah-HKBU/USG>.

## 1 Introduction

The headline is typically the first element a reader encounters when browsing news content online, playing a pivotal role in a user’s decision to engage with the news article (Tan et al., 2024). However, reliance on conventional, uniform headlines across a heterogeneous user base can result in missed op-

portunities for engagement, as users may overlook news items pertinent to their interests due to ambiguity or insufficient relevance in the headline formulation (Zhang et al., 2022). To this end, an increasing body of research has emerged around the personalization of news headlines, with the objective of generating tailored headlines for individual users to optimize both user engagement and satisfaction (Song et al., 2023; Cai et al., 2023).

A prevalent paradigm for personalized headline generation comprises two key steps (Ao et al., 2021; Yao, 2022; Cai et al., 2023): *first*, learn users’ preferences from related tasks; *second*, leverage these indirect preferences to guide headline generation. For instance, SP (Cai et al., 2023) constructs a specialized corpus based on signature phrases frequently encountered by users during news consumption, utilizing preferences for these phrases to inform subsequent headline generation. PNG (Ao et al., 2023) employs user preferences derived from a news recommender system to condition and personalize headline generation.

Despite their notable advancements, these approaches are rooted in an indirect modeling paradigm. By relying on user preferences inferred from related tasks, they personalize headlines without explicitly capturing users’ direct preferences for news headlines. This indirectness fails to fully reflect users’ actual interests in the context of headline personalization and may consequently lead to suboptimal outcomes. We rigorously validate the existence of this problem through mathematical analysis in Section 2. In our view, two key technical obstacles have led current methods to adopt indirect training approaches:

- **Data scarcity:** Directly learning users’ preferences for personalized headlines necessitates access to large-scale personalized headline datasets. In real-world settings, collecting such data requires extensive user studies, which are both

time-consuming and costly, often prohibitive in terms of time and resources (Song et al., 2023; Ouyang et al., 2022).

- **Factual accuracy:** Even with access to personalized headline data, directly trained models may be overly predisposed to cater to user preferences, potentially at the expense of factual accuracy (Xu et al., 2019; Tam et al., 2023). Thus, ensuring that personalized headlines remain factual and do not compromise the integrity of the underlying news remains another crucial concern.

To overcome these challenges, we propose **USG**, a multi-agent **u**ser **s**imulator **h**eadline **g**eneration framework. LLM-based user simulators aim to leverage the extensive background knowledge and advanced reasoning capabilities of large language models to effectively mimic user behavior (Achiam et al., 2023; Qin et al., 2024; Tang et al., 2024), serving as virtual avatars for a wide range of tasks (Wang et al., 2025; Liu et al., 2026; Corecco et al., 2024). Based on this technology, our framework consists of two primary stages (Figure 1): **Stage 1** employs interactive multi-agent simulation to synthesize a high-quality personalized headline dataset, and **Stage 2** fine-tunes a standalone generator on this dataset for efficient end-to-end inference. In particular, in Stage 1, inspired by recent role-playing agents (Tu et al., 2025; He et al., 2023), we introduce three carefully designed roles: the **User** ( $\mathcal{U}$ ) evaluates candidate headlines against individual preferences; the **Editor** ( $\mathcal{E}$ ) ensures factual correctness and editorial compliance via a contrastive data synthesis strategy covering factual, stylistic, and clickbait errors; and the **Writer** ( $\mathcal{W}$ ) iteratively refines headlines via a Critique-and-Refine algorithm guided by feedback from both  $\mathcal{U}$  and  $\mathcal{E}$ . To further improve the dataset quality, we propose two data processing strategies: (1) One-to-Many sampling, which pairs each article with diverse user contexts to capture robust user preference patterns and enrich the diversity of the dataset; and (2) rejection sampling to filter preference-misaligned samples. Leveraging the synthesized dataset, Stage 2 fine-tunes a standalone generator that enables efficient end-to-end inference without iterative multi-agent simulation, while providing interpretability by explicitly reasoning about why a specific headline is generated for a specific user.

Our main contributions are as follows:

- To the best of our knowledge, this is the first work to identify and formally characterize the

limitations of indirect training in the context of personalized news headline generation.

- We provide a theoretical analysis quantifying the deviation of indirect training from the realistic scenario and demonstrate how leveraging user simulation technology narrows this gap.
- We propose an LLM-driven multi-agent interaction framework (USG) comprising a User, an Editor, and a Writer agent that collaborate to provide direct supervision for personalized headline generation.
- Extensive experiments demonstrate that USG achieves state-of-the-art performance on both personalization and factual consistency. Furthermore, our stage-wise analysis shows that USG strikes an effective balance between headline quality and inference efficiency.

## 2 Preliminary

In this section, we compare the indirect personalized headline generation paradigm and the direct LLM simulator paradigm. Due to space constraints, we present the core theorems and comparative insights here. **Detailed and rigorous proofs are provided in Appendix A.**

**Problem Formulation** Let  $U$ ,  $C$ , and  $H$  denote the user, article body, and headline spaces, respectively. The true data distribution in the real world is denoted as  $\mathcal{P}_{\text{true}} = P_{\text{true}}(U, C, H)$ . The goal of personalized headline generation is to learn a conditional distribution  $P_{\theta}(H | U, C)$  that approximates the true underlying preference distribution:

$$P_{\text{ideal}}(H | U, C) = P_{\text{true}}(H | U, C). \quad (1)$$

**Limitations of Indirect Learning.** Indirect learning relies on a compressed latent proxy  $Z$  (e.g., tags, embeddings) inferred from auxiliary tasks, forming a Markov chain  $U \rightarrow Z \rightarrow H$ . Our analysis in Appendix A.2 demonstrates that the marginal log-likelihood of this paradigm suffers from two bounds:

- **Inference Gap:** Optimizing a surrogate objective causes a non-negative KL-divergence between the proxy encoder and the true posterior  $P(Z | H, U, C)$ .
- **Irreducible Latent Error** ( $\epsilon_{z\_loss}$ ): By the Data Processing Inequality, the low-dimensional proxy  $Z$  inevitably discards user nuances, resulting in an irreversible mutual information loss:  $\epsilon_{z\_loss} = I(H; U | C) - I(H; Z | C) > 0$ .

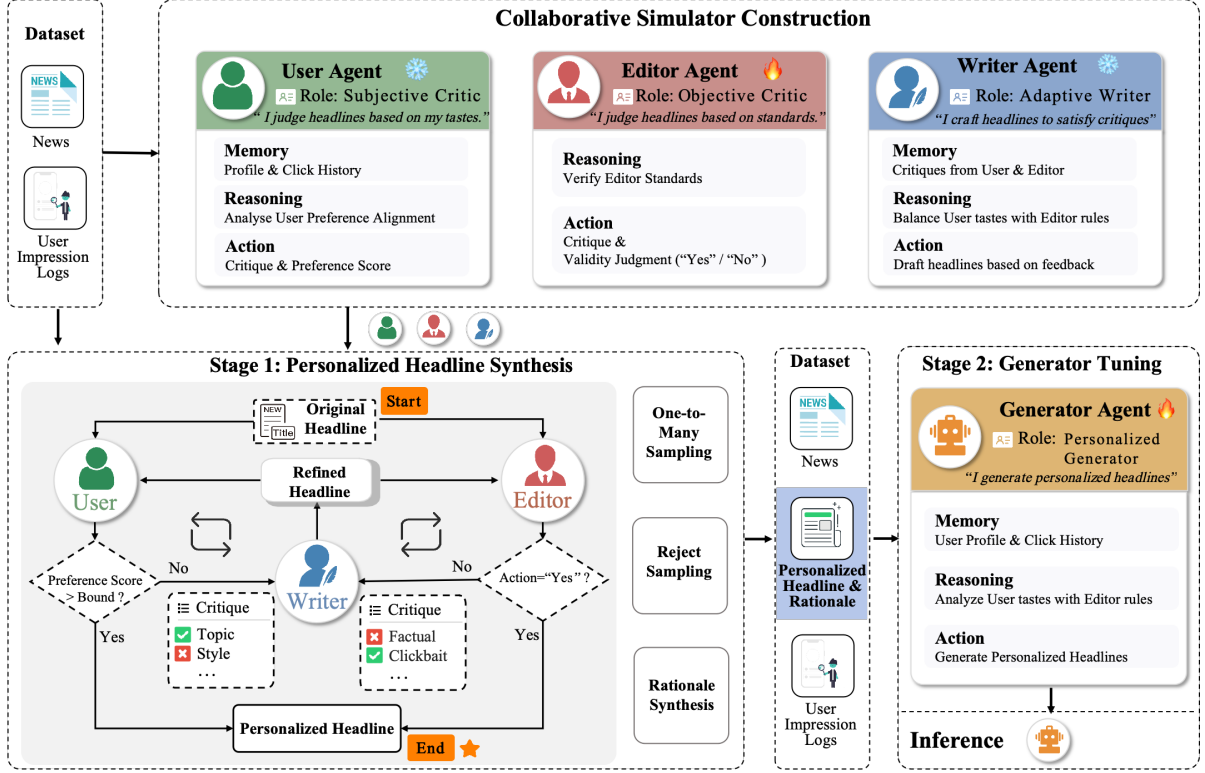


Figure 1: Overview of USG. Three specialized simulators, User ( $\mathcal{U}$ ), Writer ( $\mathcal{W}$ ), and Editor ( $\mathcal{E}$ ), collaborate in Stage 1 to synthesize a personalized headline dataset. In Stage 2, the Generator is fine-tuned on this dataset for efficient end-to-end inference.

This gap in indirect learning arises from a fundamental structural limitation imposed by disjoint training objectives. We thus advocate for direct modeling via LLM simulators. Despite intrinsic bottlenecks such as simulation accuracy and hallucinations, the rapid evolution of scaling laws and alignment methodologies has significantly mitigated these concerns (Zhang et al., 2024; Wang et al., 2025). We argue that this paradigm shift represents a promising direction for more faithful personalized headline generation.

### 3 Methodology

We propose **USG**, a novel multi-agent user simulator headline generation framework (see Figure 1). USG operates in two stages: (1) *interactive multi-agent simulation* to synthesize preference-aligned headlines with rationales, and (2) *generator fine-tuning* to train an end-to-end generator  $\mathcal{G}_\theta$ .

In the first stage, given article content  $c$  and user context  $u$ , we coordinate a Writer ( $\mathcal{W}$ ), an Editor ( $\mathcal{E}$ ), and a User ( $\mathcal{U}$ ) simulator to search for the optimal headline  $h^*$ :

$$h^* = \arg \max_{h \in \mathcal{H}_{\mathcal{W}}(c)} \mathcal{R}_{\mathcal{U}}(h) \quad \text{s.t.} \quad \mathcal{C}_{\mathcal{E}}(h) = 1 \quad (2)$$

where  $\mathcal{H}_{\mathcal{W}}(c)$  is the candidate set from  $\mathcal{W}$ ,  $\mathcal{R}_{\mathcal{U}}(\cdot)$  is the preference score from  $\mathcal{U}$ , and  $\mathcal{C}_{\mathcal{E}}(\cdot) \in \{0, 1\}$  is the editorial constraint. Crucially, we extract a grounding rationale  $r^*$  from the multi-turn interaction trajectory, yielding a high-quality synthetic corpus  $\mathcal{D}_{\text{syn}} = \{(c, u, r^*, h^*)\}$ .

In the second stage, we fine-tune a standalone generator  $\mathcal{G}_\theta$  on  $\mathcal{D}_{\text{syn}}$ . At inference time,  $\mathcal{G}_\theta$  directly generates both  $r^*$  and  $h^*$  end-to-end, eliminating the costly multi-agent loop to achieve both efficiency and interpretability.

#### 3.1 Collaborative Simulator Architectures

This section details the roles and mechanisms of the collaborative simulators in our framework. As shown in Figure 2, the Editor audits headlines against objective editorial standards, the User agent evaluates personalization alignment, and the Writer generates headlines based on their feedback. Full prompt templates for all agents are provided in Appendix D.

**User ( $\mathcal{U}$ )** The  $\mathcal{U}$  acts as a proxy for real users, offering feedback to guide the Writer agent toward headline generation that aligns with user preferences (see Figure 2). Adopting the prevalent archi-



Figure 2: Prompt schemas and interaction overview of the three collaborative agents. The Writer generates a candidate headline and receives critiques from both the Editor and User agent to iteratively refine its output.

texture of autonomous agents (Wang et al., 2024; Sumers et al., 2023; Chen et al., 2026), it comprises three modules: *Memory*, *Reasoning*, and *Action*, where the Memory Module encodes long-term profiles ( $\mathbf{u}_{\text{prof}}$ ) and short-term interactions ( $\mathbf{u}_{\text{hist}}$ ). As explicit user profiles are unavailable in real-world logs,  $\mathbf{u}_{\text{prof}}$  is synthesized from raw click history via an LLM (Kang et al., 2025).

To achieve a holistic preference judgment,  $\mathcal{U}$  integrates both semantic and collaborative signals. Specifically, while the LLM excels at capturing textual style preferences from the memory  $M_u = \{\mathbf{u}_{\text{prof}}, \mathbf{u}_{\text{hist}}\}$ , it typically struggles to capture behavioral collaborative signals. We bridge this gap by incorporating NRMS (Wu et al., 2019b), a dedicated news recommendation model trained on such signals, to provide a complementary behavioral relevance score  $S$ . Through chain-of-thought (CoT) reasoning (Wei et al., 2022),  $\mathcal{U}$  jointly reasons over both dimensions to yield the final evaluation:

$$(r_u, \mathcal{R}_{\mathcal{U}}(h)) \leftarrow \mathcal{U}(M_u \parallel h \parallel S_{\text{NRMS}} \parallel \rho_{\mathcal{U}}) \quad (3)$$

where  $\mathcal{R}_{\mathcal{U}}(h)$  is the preference score for candidate  $h$ ,  $r_u$  is a natural language critique, and  $\rho_{\mathcal{U}}$  denotes the prompt template.

**Editor ( $\mathcal{E}$ )** The  $\mathcal{E}$  serves as an objective editor, providing critical feedback in producing headlines that strictly adhere to editorial standards (see Figure 2). It comprises *Reasoning* and *Action* modules.

To equip  $\mathcal{E}$  with robust error-detection capabilities, training on both positive and negative samples is required. However, standard news corpora typically contain only valid headlines (Wu et al., 2020; Ao et al., 2021). To address the lack of negative samples, we adopt a *Contrastive Data Synthesis* strategy (Kim et al., 2023). Specifically, we instruct GPT-4.1 (Achiam et al., 2023) to generate an

invalid headline  $h^-$  and a corresponding critique rationale  $r^-$  based on three error categories (Maynez et al., 2020; Xu et al., 2019): (1) *Factual Inconsistency*, conveying false information that contradicts the source article; (2) *Misleading/Clickbait*, using exaggerated or sensationalist language to distort the content; and (3) *Stylistic Inappropriateness*, failing to present in a concise and compelling headline style. The final dataset is constructed as  $\mathcal{D}_{\text{edit}} = \mathcal{D}_{\text{pos}} \cup \{(c, h^-, r^-, \text{"Incorrect"})\}$ , where original instances in  $\mathcal{D}_{\text{pos}}$  are labeled as “Correct” and augmented with rationales justifying their compliance. We fine-tune  $\mathcal{E}_{\theta}$  via LoRA (Hu et al., 2022) on  $\mathcal{D}_{\text{edit}}$ .

During inference, given article  $c$  and candidate headline  $h$ ,  $\mathcal{E}$  performs CoT reasoning (Wei et al., 2022) to assess editorial compliance:

$$(r_e, \mathcal{C}_{\mathcal{E}}(h)) \leftarrow \mathcal{E}_{\theta}(c \parallel h \parallel \rho_{\mathcal{E}}) \quad (4)$$

where  $\mathcal{C}_{\mathcal{E}}(h) = \mathbb{I}(y_{\text{pred}} = \text{"Correct"}) \in \{0, 1\}$  indicates headline validity,  $r_e$  is a natural language critique, and  $\rho_{\mathcal{E}}$  denotes the prompt template.

**Writer ( $\mathcal{W}$ )**  $\mathcal{W}$  serves as the core generative component, responsible for drafting and iterating personalized headlines (see Figure 2). It comprises three modules: *Memory*, *Reasoning*, and *Action*, where the Memory Module dynamically logs interaction history, archiving iterative critique feedback from both  $\mathcal{U}$  and  $\mathcal{E}$  (Shinn et al., 2023). Formally, given the source article  $c$ , user memory  $\mathbf{u}$ , and the accumulated interaction history  $\mathcal{M}$ ,  $\mathcal{W}$  synthesizes a reasoning chain  $r_w$  followed by the candidate headline  $h$ , formalized as:

$$(r_w, h) \leftarrow \mathcal{W}(c \parallel \mathbf{u} \parallel \mathcal{M} \parallel \rho_{\mathcal{W}}) \quad (5)$$

where  $\rho_{\mathcal{W}}$  denotes the prompt template.

### 3.2 Stage 1: Personalized Headline Synthesis

We propose an iterative **Critique-and-Refine** algorithm (Madaan et al., 2023; Saunders et al., 2022) to synthesize a high-quality personalized headlines dataset  $\mathcal{D}_{\text{syn}} = \{(c_i, u_i, r_i^*, h_i^*)\}_{i=1}^N$ . For each instance, we first evaluate the original headline  $h_{\text{org}}$ . If  $\mathcal{R}_{\mathcal{U}}(h_{\text{org}} | u) > \tau_{\text{high}}$  (where  $\tau_{\text{high}}$  is the high-confidence bound),  $h_{\text{org}}$  is directly accepted as the optimal target  $h^*$ . Otherwise, we initialize the  $\mathcal{W}$ 's memory buffer as  $\mathcal{M}_0 = \{h_{\text{org}}, \mathcal{I}_{\text{org}}\}$  using the initial simulator critique and trigger a multi-agent refinement cycle (Li et al., 2023; Cai et al., 2025): (1) **Candidate Generation:** At iteration  $k$ ,  $\mathcal{W}$  drafts a new candidate  $h_k \sim P_{\mathcal{W}}(h | c, \mathbf{u}, \mathcal{M}_{k-1})$ . (2) **Objective Constraint Check:**  $\mathcal{E}$  evaluates  $h_k$ . If  $\mathcal{C}_{\mathcal{E}}(h_k, c) = 0$ , corrective feedback  $\mathcal{I}_{\mathcal{E}}$  is generated; we update  $\mathcal{M}_k \leftarrow \mathcal{M}_{k-1} \cup \{(h_k, \mathcal{I}_{\mathcal{E}})\}$  and revert to step (1). (3) **Subjective Preference Optimization:** If  $h_k$  is valid ( $\mathcal{C}_{\mathcal{E}} = 1$ ),  $\mathcal{U}$  assigns a preference score  $s_k = \mathcal{R}_{\mathcal{U}}(h_k | u)$ . If  $s_k < \tau_{\text{high}}$ , a preference-misalignment critique  $\mathcal{I}_{\mathcal{U}}$  is issued. We then update the memory as  $\mathcal{M}_k \leftarrow \mathcal{M}_{k-1} \cup \{(h_k, \mathcal{I}_{\mathcal{U}}, \mathcal{I}_{\mathcal{E}})\}$  and proceed to the next iteration.

The refinement process terminates when  $s_k \geq \tau_{\text{high}}$  or the maximum iteration  $T$  is reached, yielding the final accepted headline  $h^*$ . However, training the model solely on target headlines often leads to suboptimal generalization, as it lacks explicit supervision on how to effectively balance user preferences against editorial constraints (Lightman et al., 2023; Mitra et al., 2023). To alleviate this, we propose a *Rationale Synthesis* strategy. By leveraging the comprehensive multi-agent interaction trajectory  $\mathcal{T} = \{(h_k, \mathcal{I}_{\mathcal{U}_k}, \mathcal{I}_{\mathcal{E}_k})\}_{k=0}^{k^*}$  as grounding evidence, we prompt an LLM to synthesize a coherent rationale  $r^*$  that explicitly captures the reasoning process behind  $h^*$ :

$$r^* \leftarrow \text{LLM}(\mathcal{T} \parallel c \parallel u \parallel h^* \parallel \rho_{\text{rat}}) \quad (6)$$

where  $\rho_{\text{rat}}$  is the prompt template, thereby establishing the dataset  $\mathcal{D}_{\text{syn}}$ .

While the above procedure details the synthesis for a single user-article pair, a naive one-to-one sampling (i.e., pairing each article with only one user) introduces a severe risk of *shortcut learning*. Specifically, the downstream generator could easily bypass the user preferences and directly map the article content to a generic headline, failing to capture true personalization. To force the model to genuinely perceive user preferences and prevent such conditional collapse, we propose a *One-to-Many*

synthesis strategy (Chen et al., 2024). For each article  $c$ , we anchor the article content and sample a diverse set of users  $\mathcal{U}_c = \{u^{(1)}, \dots, u^{(M)}\}$ , synthesizing a distinct personalized headline  $h^{*(m)}$  for each  $u^{(m)} \in \mathcal{U}_c$ . This multi-target framing presents the generator with contrasting user-headline pairs for the exact same article context. Consequently, the model is compelled to model the joint conditional distribution  $P(h | c, u)$ , learning how identical core content must be distinctively tailored to appeal to heterogeneous user preferences, rather than collapsing into user-agnostic outputs.

To guarantee data quality, we subsequently employ a rigorous rejection sampling mechanism to eliminate unreliable synthesized instances (Wei et al., 2026) based on two criteria: (1) **Behavioral Alignment Filter:** We validate the user simulator's initial preference score  $s_{\text{org}}$  for the original headline against the ground-truth interaction label  $y \in \{0, 1\}$  from real logs. Instances are discarded if they exhibit severe behavioral mismatches:  $(y = 1 \wedge s_{\text{org}} < \tau_{\text{low}}) \vee (y = 0 \wedge s_{\text{org}} > \tau_{\text{high}})$ , where  $\tau_{\text{low}}$  and  $\tau_{\text{high}}$  represent the lower and upper confidence bounds, respectively. (2) **Infinite Optimization Loop Pruning:** If the refined headline  $h^*$  after  $T$  iterations still fails to surpass  $\tau_{\text{low}}$ , it indicates an intrinsic irrelevance between the article topic and the user. Such samples are discarded to avoid training on low-confidence noise.

### 3.3 Stage 2: Generator Tuning

In this stage, we utilize the synthesized dataset  $\mathcal{D}_{\text{syn}}$  to fine-tune a specialized Generator  $\mathcal{G}_{\theta}$  capable of producing personalized headlines end-to-end. The Generator comprises *Memory*, *Reasoning*, and *Action* modules (Wang et al., 2024; Sumers et al., 2023), where Memory encodes the long-term profile  $\mathbf{u}_{\text{prof}}$  and short-term click history  $\mathbf{u}_{\text{hist}}$ .

We fine-tune  $\mathcal{G}_{\theta}$  via LoRA (Hu et al., 2022), optimizing the cross-entropy loss over target sequence  $Y = (r^*, h^*)$  conditioned on  $X = (c, \mathbf{u}_{\text{prof}}, \mathbf{u}_{\text{hist}})$ :

$$\mathcal{L}_{\mathcal{G}}(\theta) = -\mathbb{E}_{(X,Y) \sim \mathcal{D}_{\text{syn}}} [\log P_{\theta}(Y | X)] \quad (7)$$

During inference,  $\mathcal{G}_{\theta}$  autoregressively produces the reasoning chain  $r_g$  and personalized headline  $h$  via:

$$(r_g, h) \leftarrow \mathcal{G}_{\theta}(c \parallel \mathbf{u}_{\text{prof}} \parallel \mathbf{u}_{\text{hist}} \parallel \rho_{\mathcal{G}}) \quad (8)$$

where  $\rho_{\mathcal{G}}$  denotes the prompt template.

## 4 Experiment

In this section, we conduct extensive experiments to evaluate the effectiveness of USG. Specifically,

we aim to answer the following research questions. Additionally, case study and human evaluation results are provided in Appendix B.4 and Appendix B.5, respectively. To further assess real users’ preference for generated headlines, we also conduct a user study, detailed in Appendix B.6.

- **RQ1:** *How does the performance of USG compare against state-of-the-art baselines?*
- **RQ2:** *How do the multi-agent pipeline and generator individually contribute to the effectiveness and efficiency of USG?*
- **RQ3:** *How do different components of USG contribute to the overall performance?*
- **RQ4:** *How do key hyperparameters influence the model performance?*

## 4.1 Experimental Setup

**Datasets** We employ the PENS (Ao et al., 2021) dataset, a widely adopted benchmark for personalized news headline generation<sup>1</sup>. It comprises three components: (1) a news corpus of 113,762 articles, (2) a training set of user impression logs from the Microsoft News platform, and (3) a test set of personalized headlines manually authored by 103 native English speakers and reviewed by professional editors. Dataset statistics are summarized in Table 1. Notably, the fundamental challenge of this task lies in the supervision gap: training data contains only implicit behavioral signals without explicit headline annotations, yet the model must generate personalized headlines.

Table 1: Statistics and data fields of the PENS dataset.

| Dataset Split | Statistics                           | Key Data Fields                                                |
|---------------|--------------------------------------|----------------------------------------------------------------|
| News Corpus   | 113,762 Articles                     | News ID, Headline, Body, Category, Entities                    |
| Training Set  | 445,000 Users<br>500,000 Impressions | User ID, Click History, Impressions Log (Clicked / Unclicked)  |
| Test Set      | 103 Annotators<br>20,600 Headlines   | User ID, Click History, News ID, Manual Personalized Headlines |

**Implementation Details** We employ Qwen3-8B (Yang et al., 2025) as the backbone for all agents ( $\mathcal{U}$ ,  $\mathcal{E}$ ,  $\mathcal{W}$ ) and the generator  $\mathcal{G}$ . The maximum number of iterations is set to  $T = 3$ , and the One-to-Many strategy samples  $M = 3$  distinct users per article. More details are provided in Appendix B.1.

<sup>1</sup>Currently, there are no publicly available datasets that provide the specific user interaction data and preferred news headlines for news content required in this study.

**Evaluation Protocols** Following prior work (Ao et al., 2021, 2023; Song et al., 2023), we evaluate from two perspectives: for personalization, we report ROUGE-1/2/L (Lin, 2004) to assess lexical overlap and BLEURT (Sellam et al., 2020) to gauge semantic alignment. For factual consistency, we adopt FactCC (Kryściński et al., 2020) to verify adherence to the source article. More details are provided in Appendix B.2.

**Baselines** To validate the effectiveness of our approach, we conduct a comprehensive evaluation against two categories of baselines: (1) general LLMs, including Qwen3-8B, Qwen3-14B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and GPT-4.1 (Achiam et al., 2023) under both zero-shot and few-shot settings; and (2) a range of state-of-the-art (SOTA) personalized headline generation models: PENS-NRMS, PENS-NAML (Ao et al., 2021), NAML-HAN (Yao, 2022), PNG (Ao et al., 2023) and GTP (Song et al., 2023). Baseline details are in Appendix B.3.

## 4.2 RQ1: Overall Performance Comparison

Table 2 presents the overall experimental results. We observe the following:

**USG achieves state-of-the-art performance.** USG consistently outperforms all baselines across all metrics, achieving over 20% relative improvement in ROUGE-1 and 14% in BLEURT over the strongest personalized baseline, while surpassing GPT-4.1 in factual consistency. We attribute this to our direct training paradigm, which aligns the generator with high-quality multi-agent synthesized data to simultaneously satisfy user preferences and editorial constraints.

**General LLMs exhibit a trade-off between personalization and factuality.** Few-shot prompting improves lexical and semantic alignment over zero-shot but reduces factuality. In zero-shot, lacking explicit guidance on user-preferred headlines, the model prioritizes source fidelity. In the few-shot setting, examples enable personalization but often over-optimize preference alignment at the expense of factual accuracy. Notably, USG consistently outperforms both settings, underscoring the efficacy of our approach.

**Traditional personalized models lack factual grounding.** Despite strong personalization, traditional models consistently yield suboptimal FactCC scores, reflecting their limited world knowledge when trained solely on news corpora and user click

Table 2: Overall performance comparison. “NA” denotes not available and “IM” indicates injection methods (Ao et al., 2021). R-1/2/L represent ROUGE-1/2/L scores, respectively. *Improv.* indicates the relative improvement of USG over the strongest baseline. All improvements are statistically significant ( $p < 0.05$ ).

| Types                          | Methods           | Metrics |              |              |              |              |              |
|--------------------------------|-------------------|---------|--------------|--------------|--------------|--------------|--------------|
|                                |                   | IM      | R-1(%) ↑     | R-2(%) ↑     | R-L(%) ↑     | BLEURT(%) ↑  | FactCC(%) ↑  |
| <b>LLMs<br/>(Zero-shot)</b>    | Qwen3-8B          | NA      | 23.73        | 7.63         | 20.64        | 37.53        | 67.02        |
|                                | Qwen3-14B         | NA      | 24.06        | 8.74         | 22.04        | 38.50        | 70.27        |
|                                | Llama-3.1-8B      | NA      | 22.29        | 7.20         | 20.01        | 36.72        | 65.71        |
|                                | GPT-4.1           | NA      | 25.21        | 9.06         | 22.88        | 38.64        | <u>73.33</u> |
| <b>LLMs<br/>(Few-shot)</b>     | Qwen3-8B          | NA      | 24.39        | 7.96         | 21.38        | 37.55        | 66.66        |
|                                | Qwen3-14B         | NA      | 25.38        | 9.04         | 22.53        | 39.10        | 67.98        |
|                                | Llama-3.1-8B      | NA      | 23.74        | 7.59         | 21.11        | 37.41        | 60.60        |
|                                | GPT-4.1           | NA      | 26.44        | 9.76         | 23.08        | <u>39.93</u> | 72.81        |
| <b>Personalized<br/>Models</b> | PENS-NRMS         | ①       | 24.25        | 8.65         | 21.42        | 32.68        | 48.39        |
|                                | PENS-NRMS         | ②       | 24.57        | 8.88         | 21.66        | 33.12        | 48.57        |
|                                | PENS-NRMS         | ③       | 23.69        | 8.37         | 20.90        | 32.06        | 47.98        |
|                                | PENS-NAML         | ①       | 24.16        | 8.61         | 21.35        | 32.49        | 48.73        |
|                                | PENS-NAML         | ②       | 24.20        | 8.65         | 21.39        | 32.58        | 48.91        |
|                                | PENS-NAML         | ③       | 24.17        | 8.63         | 21.37        | 32.33        | 48.62        |
|                                | NAML-HAN          | NA      | 24.39        | 8.64         | 21.48        | 32.30        | 53.21        |
|                                | PNG               | NA      | 23.44        | 7.70         | 20.97        | 28.16        | 62.26        |
|                                | GTP               | NA      | 28.06        | 10.98        | 23.67        | 39.38        | 56.12        |
|                                | <b>USG (Ours)</b> | NA      | <b>33.88</b> | <b>14.17</b> | <b>26.74</b> | <b>45.56</b> | <b>74.80</b> |
|                                | <i>Improv.</i>    | —       | +20.74%      | +29.05%      | +12.97%      | +14.10%      | +2.00%       |

logs. USG successfully bridges this gap, synthesizing the high factuality of foundation LLMs with the personalization capabilities of specialized models.

### 4.3 RQ2: Stage-wise Analysis

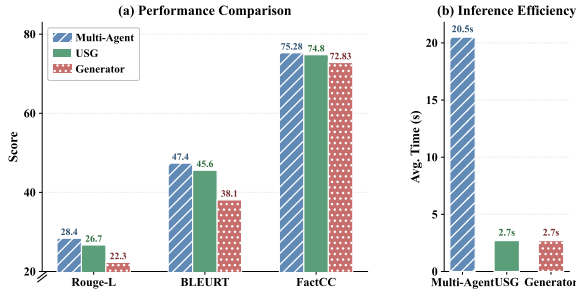


Figure 3: Multi-agent and Generator Analysis.

To quantify the contributions of the multi-agent pipeline (Stage 1) and the generator (Stage 2), we design two comparisons in terms of both headline quality and inference efficiency: (1) replacing the multi-agent synthesis with headlines generated by GPT-4.1; and (2) evaluating multi-agent outputs directly without generator fine-tuning.

Figure 3 shows the results, from which we observe: (1) **Multi-agent pipeline drives performance gains.** Training on multi-agent synthesized headlines consistently outperforms training on GPT-4.1-generated headlines across both personalization and factuality, confirming that data

quality stems from multi-agent collaboration where  $\mathcal{U}$  and  $\mathcal{E}$  jointly contribute to preference alignment and editorial quality. (2) **Generator enables efficient deployment.** While directly deploying multi-agent outputs achieves the highest quality, it requires 20.5s per headline compared to 2.7s for the generator ( $7.6\times$  speedup).

Overall, **USG achieves an effective balance between quality and efficiency:** the multi-agent pipeline provides high-quality supervision for the generator, which enables efficient end-to-end inference at the cost of a marginal performance drop.

### 4.4 RQ3: Ablation Study

To validate key components of USG, we analyze six variants: *w/o MA* replaces the multi-agent synthesis process with a same-budget single-agent baseline; *w/o User* and *w/o Editor* exclude the User and Editor simulators to remove subjective and objective constraints, respectively; *w/o Warm-Start* disables the original headline as a refining anchor; *w/o O2M* and *w/o Rejection* remove One-to-Many sampling and post-synthesis rejection sampling, respectively.

Table 3 reveals three insights: (1) **Multi-Agent Collaboration is Effective:** *w/o MA* drops all five metrics simultaneously, demonstrating that multi-agent collaboration yields higher-quality synthetic data than a single agent. (2) **User and Editor Agents are Complementary:** *w/o User* maximizes

Table 3: Comprehensive ablation study results.

| Variants       | R-1          | R-2          | R-L          | BLEURT       | FactCC       |
|----------------|--------------|--------------|--------------|--------------|--------------|
| <b>USG</b>     | 33.88        | 14.17        | 26.74        | 45.56        | 74.80        |
| w/o MA         | 30.32        | 11.43        | 24.47        | 42.92        | 71.23        |
| w/o User       | 28.53        | 9.13         | 23.32        | 42.33        | <b>77.71</b> |
| w/o Editor     | <b>35.90</b> | <b>16.13</b> | <b>28.11</b> | <b>47.50</b> | 69.33        |
| w/o Warm-Start | 30.41        | 11.93        | 24.59        | 43.59        | 75.14        |
| w/o O2M        | 31.61        | 12.51        | 25.27        | 43.69        | 75.12        |
| w/o Rejection  | 32.69        | 12.83        | 25.66        | 44.89        | 74.92        |

FactCC but severely drops BLEURT and ROUGE, while *w/o Editor* maximizes personalization at the cost of a sharp FactCC drop (74.80→69.33). This confirms that the User agent drives personalization while the Editor agent enforces factual consistency, and both are indispensable. (3) **Efficacy of Auxiliary Modules:** All three modules consistently benefit personalization. Specifically, *w/o Warm-Start* degrades personalization while slightly increasing FactCC; we speculate that skipping initialization forces from-scratch generation, requiring more alignment negotiation with the Editor and thus restricting available interaction turns for the User agent. *w/o O2M* drops personalization scores, proving that diverse user simulation is conducive to capturing robust preference patterns. *w/o Rejection* compromises ROUGE/BLEURT while maintaining stable FactCC scores, demonstrating that the rejection criteria effectively filter out misaligned samples to ensure data purity.

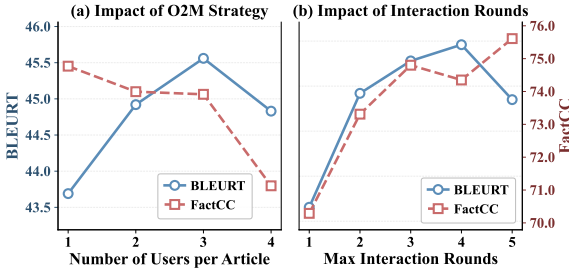


Figure 4: Parameter Sensitivity Analysis.

#### 4.5 RQ4: Parameter Sensitivity Analysis

We investigate the framework sensitivity to two key hyperparameters in Figure 4: the number of users sampled per article ( $M \in \{1, 2, 3, 4\}$ ) for the One-to-Many (O2M) strategy, and the maximum number of interaction rounds ( $T \in \{1, 2, 3, 4, 5\}$ ) between agents. We observe two core trends: (1) **Impact of O2M Strategy ( $M$ ):** Scaling  $M$  from 1 to 3 monotonically boosts BLEURT (from 43.69 to 45.56), confirming that exposing the generator

to diverse user perspectives enhances preference learning. However, further scaling to  $M = 4$  degrades performance, while FactCC exhibits a continuous downward trend. This indicates that excessive user diversity introduces conflicting optimization signals that overwhelm the generator and compromise factual grounding. (2) **Impact of Interaction Rounds ( $T$ ):** Both metrics improve steadily from  $T = 1$  to  $T = 3$ , confirming that sufficient rounds are essential for iterative refinement. Beyond  $T = 3$ , the two metrics diverge, reflecting the trade-off between personalization and factuality: additional rounds increasingly favor user preference over factual grounding, until the Editor reasserts editorial constraints at  $T = 5$ .

## 5 Related Work

**Personalized Headline Generation** Personalized headline generation aims to produce tailored headlines for individual users based on news content. However, collecting large-scale, high-quality personalized headlines for training is rarely feasible in practice (Ao et al., 2021). As a result, the mainstream paradigm first infers users’ indirect preferences from related tasks and generates personalized headlines conditioned on them (Yao, 2022; Song et al., 2023; Yang et al., 2023). For example, EUI-PENS (Zhang et al., 2022) leverages existing user representations from news recommender systems as external knowledge to guide headline generation. SP (Cai et al., 2023) extracts signature phrases from users’ reading histories to customize headlines. PNG (Ao et al., 2023) introduces a knowledge-aware user encoder to model user interests from reading behavior, incorporated into a pointer-generator network. Tan et al. (2024) incorporates click history into prompts as indirect preferences via an LLM-based approach. Appendix C reviews further related work on LLM-based user simulators for data synthesis.

## 6 Conclusion

In this work, we identify the fundamental limitations of existing indirect paradigms for personalized news headline generation, which arise as practical compromises due to data scarcity and factual accuracy constraints. To address these issues, we propose USG, an LLM-based multi-agent framework that integrates subjective user modeling with objective editorial oversight through collaborative simulation. By synthesizing high-quality,

preference-aligned headlines for direct supervision, USG enables headline generators that achieve superior personalization without sacrificing factual consistency. Our theoretical analysis quantifies the expressiveness gap in indirect training and demonstrates how USG bridges this divide. Extensive experiments on the PENS dataset validate our framework, achieving state-of-the-art performance on both personalization and factuality metrics, a notable breakthrough given the historical tradeoff between these objectives.

## 7 Limitations

While USG demonstrates strong performance, several limitations remain. Our approach relies on LLMs, whose usage entails non-trivial monetary costs and engineering constraints (e.g., latency and reliability). In addition, despite substantial related work on personalized headline generation, publicly available datasets remain scarce, as collecting user click histories paired with human-written headlines is costly and time-consuming; to the best of our knowledge, only one publicly available dataset is directly applicable to our setting. Finally, since we adopt a multi-agent framework that relies on LLM-generated intermediate outputs to synthesize training data, this process may be susceptible to risks of synthetic-data bias and error propagation.

## 8 Ethics Statement

This study used publicly available datasets and complied with their original licenses and intended uses. We also conducted a user study to evaluate the proposed method. All participants were adults and signed informed consent forms before participation. Our study did not pose risks to privacy, fairness, or other ethical concerns arising from the proposed method or experimental protocol. No personally identifiable information or other sensitive attributes were collected in this study.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- 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. *TKDE*, 18(3):1–20.
- 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 *IJCNLP*, pages 82–92.
- Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.
- Pengshan Cai, Kaiqiang Song, Sangwoo Cho, Hongwei Wang, Xiaoyang Wang, Hong Yu, Fei Liu, and Dong Yu. 2023. Generating user-engaging news headlines. In *ACL*, pages 3265–3280.
- Shihao Cai, Jizhi Zhang, Keqin Bao, Chongming Gao, Qifan Wang, Fuli Feng, and Xiangnan He. 2025. Agentic feedback loop modeling improves recommendation and user simulation. In *Proceedings of the 48th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 2235–2244.
- Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *WWW*, 27(4):42.
- Weixin Chen, Yuhao Zhao, Jingyuan Huang, Zhihe Ye, Mingxuan Ju, Tong Zhao, Neil Shah, Li Chen, and Yongfeng Zhang. 2026. Memrec: Collaborative memory-augmented agentic recommender system. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44515–44544.
- Nathan Corecco, Giorgio Piatti, Luca A Lanzendörfer, Flint Xiaofeng Fan, and Roger Wattenhofer. 2024. Suber: An rl environment with simulated human behavior for recommender systems. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain—Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pages 2210–2217.
- Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. Lego: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Feiran Huang, Yuanchen Bei, Zhenghang Yang, Junyi Jiang, Hao Chen, Qijie Shen, Senzhang Wang, Fakhri Karray, and Philip S Yu. 2025. Large language model simulator for cold-start recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 261–270.
- Li Kang, Yuhan Zhao, and Li Chen. 2025. Exploring the potential of llms for serendipity evaluation in recommender systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 746–754.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.
- Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9332–9346.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.
- Lei Li, Yongfeng Zhang, Dugang Liu, and Li Chen. 2024. Large language models for generative recommendation: A survey and visionary discussions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10146–10159.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Fei Liu, Xinyu Lin, Hanchao Yu, Mingyuan Wu, Jianyu Wang, Qiang Zhang, Zhuokai Zhao, Yinglong Xia, Yao Zhang, Weiwei Li, et al. 2026. Recoworld: Building simulated environments for agentic recommender systems. In *Companion Proceedings of the ACM Web Conference 2026*, pages 650–659.
- 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*, pages 1872–1881.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.
- Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Coda, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.
- 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*, 35:27730–27744.
- Choongwon Park. 2025. Llm as user simulator: Towards training news recommender without real user interactions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '25*, page 3080–3084, New York, NY, USA. Association for Computing Machinery.
- Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13025–13048.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.
- 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*, pages 7881–7892.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

- Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. 2023. General then personal: Decoupling and pre-training for personalized headline generation. *TACL*, 11:1588–1607.
- Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.
- Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.
- Xiaoyu Tan, Leijun Cheng, Xihe Qiu, Shaojie Shi, Yuan Cheng, Wei Chu, Yinghui Xu, and Yuan Qi. 2024. Enhancing personalized headline generation via of-fine goal-conditioned reinforcement learning with large language models. In *KDD*, pages 5762–5772.
- Yihong Tang, Zhaokai Wang, Ao Qu, Yihao Yan, Zhaofeng Wu, Dingyi Zhuang, Jushi Kai, Kebing Hou, Xiaotong Guo, Jinhua Zhao, et al. 2024. Itinera: Integrating spatial optimization with large language models for open-domain urban itinerary planning. In *EMNLP*, pages 1413–1432.
- Geng Tu, Bingbing Wang, Erik Cambria, Wenjie Li, and Ruifeng Xu. 2025. Supportplay: A multi-agent role-playing system for personalized and sustained multimodal emotional support conversation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2915–2918.
- Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.
- Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. 2025. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):1–37.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Tianjun Wei, Huizhong Guo, Yingpeng Du, Zhu Sun, Huang Chen, Dongxia Wang, and Jie Zhang. 2026. Mirroring users: Towards building preference-aligned user simulator with user feedback in recommendation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6046–6071.
- Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576*.
- Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019b. Neural news recommendation with multi-head self-attention. In *EMNLP-IJCNLP*, pages 6389–6394.
- Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, et al. 2020. Mind: A large-scale dataset for news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3597–3606.
- Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning. 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)*, pages 3065–3075.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Zhao Yang, Junhong Lian, and Xiang Ao. 2023. Fact-preserved personalized news headline generation. In *ICDM*, pages 1493–1498. IEEE.
- Jiaohong Yao. 2022. Personalized news headline generation system with fine-grained user modeling. In *MSN*, pages 581–585. IEEE.
- An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. 2024. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pages 1807–1817.
- Kui Zhang, Guangquan Lu, Guixian Zhang, Zhi Lei, and Lijuan Wu. 2022. Personalized headline generation with enhanced user interest perception. In *ICANN*, pages 797–809. Springer.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyang Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, pages 400–410.

## A Theoretical Analysis

In this section, we formulate a theoretical framework to rigorously analyze the indirect personalized headline generation paradigm and the direct LLM simulator paradigm.

### A.1 Problem Formulation

Let  $U$ ,  $C$ , and  $H$  denote the user, article body, and headline spaces, respectively. The true data distribution in the real world is denoted as  $P_{\text{true}} = P_{\text{true}}(U, C, H)$ . The goal of personalized headline generation is to learn a conditional distribution  $P_{\theta}(H | U, C)$  that approximates the true underlying preference distribution:

$$P_{\text{ideal}}(H | U, C) = P_{\text{true}}(H | U, C). \quad (9)$$

### A.2 Analysis of Indirect Learning

Modeling Eq. 9 requires a dataset  $\mathcal{D} = \{(u_i, c_i, h_i)\}$  sampled from  $P_{\text{true}}$ . However, collecting such data at scale necessitates extensive user participation, which is time-consuming, costly, and ultimately unscalable. As a practical workaround, the indirect learning paradigm has been widely adopted. Indirect learning introduces a latent variable  $Z \in \mathcal{Z}$  (e.g., user tags, embeddings) inferred from auxiliary tasks. The modeling process consists of two components:

1. A proxy inference model  $Q_{\phi}(Z | U)$  (the encoder).
2. A conditional generator  $P_{\psi}(H | Z, C)$  (the decoder).

The distribution induced by this paradigm is the marginal distribution over the latent variable  $Z$ :

$$\begin{aligned} P_{\text{indirect}}(H | U, C) \\ = \int_{\mathcal{Z}} P_{\psi}(H | Z, C) Q_{\phi}(Z | U) dZ. \end{aligned} \quad (10)$$

#### A.2.1 The Optimization Gap

Because the marginal log-likelihood involves an intractable integral, indirect methods optimize surrogate objectives, leading to a structural inference gap.

**Theorem A.2.1** (Decomposition of Indirect Objective). For any sample  $(u, c, h) \sim P_{\text{true}}$ , the log-likelihood of the indirect model decomposes into a reconstruction objective and a strictly non-negative

inference gap:

$$\begin{aligned} \log P_{\text{indirect}}(H | U, C) \\ = \mathbb{E}_{Z \sim Q_{\phi}} [\log P_{\psi}(H | Z, C)] \\ + D_{\text{KL}}(Q_{\phi}(Z | U) \parallel P_{\text{post}}(Z | H, U, C)) \end{aligned} \quad (11)$$

where  $P_{\text{post}}(Z | H, U, C) = \frac{P_{\psi}(H | Z, C) Q_{\phi}(Z | U)}{P_{\text{indirect}}(H | U, C)}$  is the true posterior.

*Proof.* We derive this decomposition by analyzing the KL-divergence between the proxy encoder  $Q_{\phi}$  and the true posterior  $P_{\text{post}}$ . By definition:

$$\begin{aligned} D_{\text{KL}}(Q_{\phi}(Z | U) \parallel P_{\text{post}}(Z | H, U, C)) \\ = \mathbb{E}_{Z \sim Q_{\phi}} [\log Q_{\phi}(Z | U) \\ - \log P_{\text{post}}(Z | H, U, C)] \\ = \mathbb{E}_{Z \sim Q_{\phi}} \left[ \log Q_{\phi}(Z | U) \right. \\ \left. - \log \frac{P_{\psi}(H | Z, C) Q_{\phi}(Z | U)}{P_{\text{indirect}}(H | U, C)} \right] \\ = -\mathbb{E}_{Z \sim Q_{\phi}} [\log P_{\psi}(H | Z, C)] \\ + \log P_{\text{indirect}}(H | U, C). \end{aligned} \quad (12)$$

Rearranging the terms yields the stated equality:

$$\begin{aligned} \log P_{\text{indirect}}(H | U, C) \\ = \underbrace{\mathbb{E}_{Z \sim Q_{\phi}} [\log P_{\psi}(H | Z, C)]}_{\text{Reconstruction Objective}} \\ + \underbrace{D_{\text{KL}}(Q_{\phi} \parallel P_{\text{post}})}_{\text{Inference Gap}}. \end{aligned} \quad (13)$$

The KL term is strictly non-negative due to Gibbs' inequality.  $\square$

**Implication:** Indirect training typically maximizes only the first term (Reconstruction Objective). However, since  $Q_{\phi}(Z | U)$  is often pre-trained on auxiliary tasks (e.g., tag prediction) and fixed, it cannot dynamically adapt to approximate the *true posterior*  $P_{\text{post}}(Z | H, U, C)$  required for a specific headline. This results in a permanent **Inference Gap**.

#### A.2.2 Irreducible Bottleneck

Beyond optimization, indirect methods suffer from a fundamental representation limit. Since  $\dim(Z) \ll \dim(U)$ , the Data Processing Inequality (DPI) implies unavoidable information loss.

**Proposition A.2.1** (Bottleneck). Let the generation process be a Markov chain  $U \rightarrow Z \rightarrow H$ . The mutual information satisfies:

$$I(H; U | C) \geq I(H; Z | C). \quad (14)$$

The inequality is strict,  $I(H; U | C) > I(H; Z | C)$ , unless  $Z$  is a sufficient statistic of  $U$  for  $H$ . Given the high complexity of user preference  $U$  and the low dimensionality of proxy  $Z$ ,  $Z$  cannot capture all variations in  $U$  relevant to  $H$ . This defines the **Irreducible Error** of the indirect paradigm:

$$\epsilon_{z\_loss} = I(H; U | C) - I(H; Z | C) > 0. \quad (15)$$

### A.3 Analysis of LLM Simulator Training

The simulator paradigm attempts to bypass the latent bottleneck  $Z$  by directly modeling the joint distribution. However, realistically, an LLM cannot ingest the abstract, infinite-dimensional true user preference  $U$ . Instead, it relies on a function  $\tau(\cdot)$  to simulate the user (e.g., prompts constructed from historical reading summaries), and samples from its own approximation of the world. The realistic simulator distribution is:

$$\begin{aligned} \mathcal{P}_{\text{sim}}(U, C, H) \\ = P_{\text{sample}}(U, C) \cdot P_{\text{LLM}}(H | \tau(U), C). \end{aligned} \quad (16)$$

**Theorem A.3.1** (Simulator Divergence Decomposition). The KL-divergence between the true world and the practical simulator exactly decomposes into sampling bias and a fidelity gap:

$$\begin{aligned} D_{\text{KL}}(\mathcal{P}_{\text{true}} \parallel \mathcal{P}_{\text{sim}}) \\ = \underbrace{D_{\text{KL}}(P_{\text{true}}(U, C) \parallel P_{\text{sample}}(U, C))}_{\text{Sampling Bias}} \\ + \mathbb{E}_{(U, C) \sim \mathcal{P}_{\text{true}}} [\text{Fidelity Gap}]. \end{aligned} \quad (17)$$

where Fidelity Gap =  $D_{\text{KL}}(P_{\text{true}}(H | U, C) \parallel P_{\text{LLM}}(H | \tau(U), C))$ .

*Proof.* By the definition of the KL-divergence for joint distributions:

$$\begin{aligned} D_{\text{KL}}(\mathcal{P}_{\text{true}} \parallel \mathcal{P}_{\text{sim}}) \\ = \mathbb{E}_{(u, c, h) \sim \mathcal{P}_{\text{true}}} \left[ \log \frac{P_{\text{true}}(u, c, h)}{P_{\text{sim}}(u, c, h)} \right]. \end{aligned} \quad (18)$$

Factorizing both numerator and denominator using the chain rule of probability:

$$P_{\text{true}}(u, c, h) = P_{\text{true}}(u, c) \cdot P_{\text{true}}(h | u, c). \quad (19)$$

$$\begin{aligned} P_{\text{sim}}(u, c, h) \\ = P_{\text{user}}(u) P_{\text{content}}(c) \cdot P_{\text{LLM}}(h | \tau(u), c). \end{aligned} \quad (20)$$

Substitute these back into the log term:

$$\begin{aligned} \log \frac{P_{\text{true}}(u, c, h)}{P_{\text{sim}}(u, c, h)} &= \log \frac{P_{\text{true}}(u, c)}{P_{\text{user}}(u) P_{\text{content}}(c)} \\ &\quad + \log \frac{P_{\text{true}}(h | u, c)}{P_{\text{LLM}}(h | \tau(u), c)}. \end{aligned} \quad (21)$$

Taking the expectation  $\mathbb{E}_{(u, c, h) \sim \mathcal{P}_{\text{true}}}$  over the sum, the first term depends only on  $(U, C)$ , and the second term conditions on them:

$$\begin{aligned} \text{Term 1} &= \mathbb{E}_{(u, c) \sim \mathcal{P}_{\text{true}}} \left[ \log \frac{P_{\text{true}}(u, c)}{P_{\text{user}}(u) P_{\text{content}}(c)} \right] \\ &= D_{\text{KL}}(P_{\text{true}}(U, C) \parallel P_{\text{user}}(U) P_{\text{content}}(C)) \end{aligned} \quad (22)$$

$$\begin{aligned} \text{Term 2} &= \mathbb{E}_{(u, c) \sim \mathcal{P}_{\text{true}}} \left[ \mathbb{E}_{h \sim P_{\text{true}}(\cdot | u, c)} \left( \log \frac{P_{\text{true}}(h | u, c)}{P_{\text{LLM}}(h | \tau(u), c)} \right) \right] \\ &= \mathbb{E}_{(U, C) \sim \mathcal{P}_{\text{true}}} \left[ D_{\text{KL}}(P_{\text{true}}(H | U, C) \parallel P_{\text{LLM}}(H | \tau(U), C)) \right]. \end{aligned} \quad (23)$$

Summing Term 1 and Term 2 completes the proof.  $\square$

Crucially, **the fidelity gap of the simulator is non-zero**. Building upon 17, the gap arises from both the learned distribution  $P_{\text{LLM}}(\cdot)$  and the mapping function  $\tau(\cdot)$ , indicating that the LLM-based simulator incurs its own irreducible practical errors. We can further decompose this fidelity gap into two components:  $\epsilon_{\tau\_loss}$ , which captures the model's capacity to accurately generate headlines, and  $\epsilon_{\text{hallu}}$ , which accounts for the intrinsic error of the LLM, such as hallucinations.

### A.4 Theoretical Comparison and Conclusion

By analyzing both paradigms through a unified theoretical lens, we identify a fundamental trade-off in their respective error compositions. Neither approach is theoretically infallible; rather, they are characterized by distinct failure modes. The generation risk of the indirect method is dominated by Inference Gap +  $\epsilon_{z\_loss}$ , while the simulator is dominated by Sampling Bias +  $\epsilon_{\tau\_loss}$  +  $\epsilon_{\text{hallu}}$ . The core justification for this paradigm shift lies in the distinction between structural and scalable

errors. The Inference Gap in indirect learning represents a rigid structural bottleneck—an irreducible discrepancy stemming from disjoint training objectives. Conversely, the simulator’s errors, such as *Sampling Bias* and *LLM Intrinsic Error* ( $\epsilon_{\text{hallu}}$  and  $\epsilon_{\tau_{\text{loss}}}$ ), are scalable limitations. While these are not entirely eliminated, they are systematically ameliorated by the advancement of scaling laws, sophisticated sampling strategies and state-of-the-art alignment protocols (Zhang et al., 2024; Wang et al., 2025; Argyle et al., 2023). In summary, our theoretical analysis highlights two key insights:

- **Rationale for paradigm shift:** The shift to LLM-based simulator training is theoretically well-founded, as it eliminates the irreducible error of indirect methods and benefits from continuous advances in LLM technology.
- **The Foundational Role of Indirect Learning:** While we advocate for a shift toward direct modeling, indirect learning methods remain fundamentally significant. They facilitated critical advancements in personalized headline synthesis during the era when simulation fidelity was bottlenecked by substantial intrinsic errors (high  $\epsilon_{\text{hallu}}$  and  $\epsilon_{\tau_{\text{loss}}}$ ). Furthermore, these approaches established the theoretical and empirical foundations for modeling user preferences within data-constrained regimes, offering robustness where early-stage simulators lacked the requisite capacity.

## B Experiment

### B.1 Implementation Details

To ensure reproducibility, our framework is implemented using the LLaMA-Factory library (Zheng et al., 2024). The specific configurations for agent architecture, training, and synthesis strategies are detailed below.

**Agent Configuration and User Context.** We select Qwen3-8B (Yang et al., 2025) as the backbone for all agents ( $\mathcal{U}$ ,  $\mathcal{E}$ ,  $\mathcal{W}$ ) and the generator  $\mathcal{G}$ . Regarding user modeling, the long-term user profile ( $\mathbf{u}_{\text{prof}}$ ) is derived from the user’s most recent 200 click interactions, while the short-term memory ( $\mathbf{u}_{\text{hist}}$ ) stores the latest 20 clicks.

**Training Setup.** The Editor agent and the downstream generator are fine-tuned using LoRA with a rank of  $r = 8$  and an alpha of  $\alpha = 16$ , applying adapters across all linear layers. The learning rate is set to  $1 \times 10^{-4}$ . The Editor is trained for 2

epochs and the generator for 3 epochs. The Editor’s training set uses a 1 : 3 positive-to-negative sample ratio to improve error discrimination.

**Headline Synthesis Strategy.** In the *Critique-and-Refine* process (Stage 1), we empirically set the lower acceptance bound  $\tau_{\text{low}} = 0.3$  and the high-confidence bound  $\tau_{\text{high}} = 0.85$ . The maximum iteration count is limited to  $T = 3$ . Implementing the *One-to-Many* strategy, we sample 3 distinct users for each article body. This process yields a final synthetic dataset  $\mathcal{D}_{\text{syn}}$  comprising 27,036 personalized headlines.

We conduct experiments on a workstation equipped with an Intel Xeon Gold 6248R 3.00 GHz CPU, eight NVIDIA A100 80 GB GPUs, and 1.5 TB of RAM.

### B.2 Evaluation Protocols

To ensure a rigorous and fair evaluation, we adopt protocols consistent with prior studies in personalized headline generation (Ao et al., 2021, 2023; Yao, 2022; Yang et al., 2023; Song et al., 2023). Specifically, we assess model performance from three complementary perspectives: lexical overlap, semantic alignment, and factual consistency. We employ the standard ROUGE-1, ROUGE-2, and ROUGE-L metrics (Lin, 2004)<sup>2</sup> to measure n-gram overlap and sentence-level structural similarity between the generated and manually written headlines. Beyond surface-level matching, we utilize BLEURT (Sellam et al., 2020)<sup>3</sup> to assess the semantic similarity between the generated outputs and the reference headlines. To strictly evaluate factual consistency with respect to the source content, we adopt FactCC (Kryściński et al., 2020)<sup>4</sup>, a claim verification model that assesses whether the generated headline is factually supported by the source article.

### B.3 Baselines

To validate the effectiveness of our approach, we conduct a comprehensive evaluation against two categories of baselines: (1) general LLMs, and (2) specialized personalized headline generation models. For the former, we assess representative open-source models (Qwen3-8B (Yang et al., 2025), Qwen3-14B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024)) as well as the closed-source GPT-4.1 (Achiam et al., 2023), under

<sup>2</sup><https://pypi.org/project/rouge/1.0.1/>

<sup>3</sup><https://github.com/google-research/bleurt>

<sup>4</sup><https://huggingface.co/manueldeprada/FactCC>

both zero-shot and few-shot settings. For the latter, we compare against a range of state-of-the-art (SOTA) models on this task:

- **PENS-NRMS** (Ao et al., 2021): Utilizes the NRMS (Wu et al., 2019b) click-through rate prediction model to capture user preferences, injecting these latent representations into the headline generator.
- **PENS-NAML** (Ao et al., 2021): Building on PENS, this variant employs the NAML (Wu et al., 2019a) preference model to enhance user modeling and is recognized as the best-performing baseline in the original PENS study.
- **NAML-HAN** (Yao, 2022): Incorporates sentence-level informativeness into the recommendation model, using a hierarchical attention network to guide the generation of high-quality headlines.
- **PNG** (Ao et al., 2023): Designs a knowledge-aware user encoder and utilizes a user-perturbed pointer-generator network to explicitly condition headline generation on user interests.
- **GTP** (Song et al., 2023): Adopts a decoupled two-stage paradigm that first generates generic candidates and subsequently customizes them via information self-boosting mechanisms to achieve personalization.

#### B.4 Case Study

We present a case study from the experimental data in Table 4, comparing the personalized headlines generated by our model USG against the state-of-the-art baselines from two categories: (1) a general LLM (GPT-4.1, zero-shot), and (2) a specialized personalized headline generation model (GTP). In Table 4, **Fact.** denotes factual consistency with the article, and **Pers.** denotes personalization quality with respect to the reference headline, which is the manually written headline by the user in the test set. To further illustrate USG’s data synthesis process, Table 5 details the multi-agent interaction underlying this example. As shown in Table 4, USG achieves the best overall performance, producing a headline that is both factually consistent and well-personalized. GTP produces a headline containing

Table 4: Case study. ✓: correct; ✗: incorrect; △: partial.

|                                 |                                                                                                                                                                                                                                                                                                                            |              |              |
|---------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|--------------|
| <b>User Profile</b>             | <b>Topic Interests:</b> The user shows interest in politics, particularly U.S. politics and news, [...], lifestyle, and occasionally engages with health. [...]<br><b>Style Preferences:</b> The user prefers detailed news articles with a mix of formal and sometimes more casual or entertainment-oriented tones. [...] |              |              |
| <b>Short-term Click History</b> | <ul style="list-style-type: none"> <li>• Bernie Sanders details plan to cancel \$1.6 trillion in student loan debt</li> <li>• 31 Easy No-Cook Dinners to Make in July</li> <li>• California residents wondered if today’s earthquake was “the big one”</li> </ul>                                                          |              |              |
| <b>Article</b>                  | Learn what pain doctors are really thinking when you visit them, and how you can better manage your symptoms. [...] Chances are, you’re <b>never</b> going to be 100 percent pain-free—especially if you’re experiencing chronic back or knee pain. [...]                                                                  |              |              |
| <b>Reference</b>                | Understand Your Pain Better                                                                                                                                                                                                                                                                                                |              |              |
| <b>Method</b>                   | <b>Generated Headline</b>                                                                                                                                                                                                                                                                                                  | <b>Fact.</b> | <b>Pers.</b> |
| GTP                             | Pain Doctors’ Tips for 100% Pain-Free                                                                                                                                                                                                                                                                                      | ✗            | △            |
| GPT-4.1                         | What Pain Doctors Really Want You to Know: Managing Your Symptoms Effectively                                                                                                                                                                                                                                              | ✓            | ✗            |
| <b>USG (Ours)</b>               | Expert Pain Tips: Understand Your Pain Better                                                                                                                                                                                                                                                                              | ✓            | ✓            |

the phrase “100% Pain-Free,” which directly contradicts the article’s key claim that one is *never* going to be 100 percent pain-free. While GTP demonstrates partial personalization alignment, it fails on factual consistency. GPT-4.1, on the other hand, generates a factually correct headline but lacks personalization, producing a generic and overly formal headline that does not align with the user’s preference for a casual, engaging tone.

The multi-agent interaction trace in Table 5 reveals several key observations. First, the Editor agent serves as an objective evaluator: given the original clickbait headline “24 Secrets Pain Doctors Won’t Tell You,” it correctly flags both the fabricated numerical claim (“24 secrets” is unsupported by the article) and the sensational framing, grounding its critique in factual consistency, clickbait detection, and stylistic appropriateness standards. In Iteration 1, it further identifies entity inaccuracy and misleading implication in the Writer’s output. Second, the User agent provides subjective,

Table 5: Multi-agent interaction process among Writer, Editor, and User for personalized news headline synthesis.

|                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|-------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Original News Headline: “24 Secrets Pain Doctors Won’t Tell You”</i> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Editor Agent</b>                                                     | <p><b>Reasoning:</b> Factual Consistency: The article does not mention “24 secrets” or that pain doctors are withholding information; this is a fabricated number and implication. No Misleading/Clickbait: The headline is misleading and sensationalizes the content. Stylistic Appropriateness: The phrase “Secrets Pain Doctors Won’t Tell You” is clickbait and sensational, not appropriate for a factual article about pain management.</p> <p><b>Critique:</b> This headline violates (1) Factual Consistency: “24 secrets” is not supported by the article; and (2) No Misleading/Clickbait &amp; Stylistic Appropriateness: “Secrets Pain Doctors Won’t Tell You” is clickbait framing, unsuitable for a factual article.</p> <p><b>Decision:</b> ✗ Incorrect</p> |
| <b>Iteration 1</b>                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Writer Agent</b>                                                     | <p><b>Reasoning:</b> The headline is concise, factual, and addresses the key points from the article. It avoids sensationalism and aligns with the user’s interest in health and detailed, informative content.</p> <p><b>Output:</b> “Expert Pain Management Tips from Board-Certified Doctors”</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <b>Editor Agent</b>                                                     | <p><b>Reasoning:</b> No Misleading/Clickbait: The headline could imply a more prescriptive or structured guide than provided in the article. Entity Accuracy: “Board-certified doctors” is not the main subject of the article.</p> <p><b>Critique:</b> This headline violates (1) No Misleading/Clickbait: it implies a more structured guide than the article provides; and (2) Entity Accuracy: “board-certified doctors” is not the main subject of the article.</p> <p><b>Decision:</b> ✗ Incorrect</p>                                                                                                                                                                                                                                                                |
| <b>Iteration 2</b>                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Writer Agent</b>                                                     | <p><b>Reasoning:</b> Key facts include understanding pain causes and proper use of heat and ice. User preferences are addressed by focusing on the content and maintaining a casual tone while avoiding sensationalism.</p> <p><b>Output:</b> “Pain Management Tips from Experts: Understanding Your Pain”</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Editor Agent</b>                                                     | <p><b>Reasoning:</b> Factual Consistency: The headline accurately reflects the article’s focus on pain management, expert insights, and advice about understanding pain causes. No misleading or exaggerated claims. Entity Accuracy: “Experts” is appropriate as the article features pain doctors and specialists. Stylistic Appropriateness: The headline is concise, clear, and not generic or narrative.</p> <p><b>Critique:</b> This headline meets all standards.</p> <p><b>Decision:</b> ✓ Correct</p>                                                                                                                                                                                                                                                              |
| <b>User Agent</b>                                                       | <p><b>Reasoning:</b> The topic of pain management aligns well with the user’s interest in health, indicated by the strong NRMS score of 0.615. However, the headline is formal and does not align with the user’s preference for a mix of formal and casual tones. The recent click history does not include health-related content, but the NRMS score suggests potential interest.</p> <p><b>Critique:</b> The headline’s formal tone doesn’t align with the user’s preference for a mix of formal and casual.</p> <p><b>Score:</b> 0.75</p>                                                                                                                                                                                                                              |
| <b>Iteration 3</b>                                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Writer Agent</b>                                                     | <p><b>Reasoning:</b> The article discusses pain management insights from experts and the importance of understanding pain causes. The user has a strong interest in finance, politics, and entertainment, with a preference for a mix of formal and casual tones. The new headline integrates these elements by focusing on expert advice in a more casual, relatable manner.</p> <p><b>Output:</b> “Expert Pain Tips: Understand Your Pain Better”</p>                                                                                                                                                                                                                                                                                                                     |
| <b>Editor Agent</b>                                                     | <p><b>Reasoning:</b> Factual Consistency: The headline accurately reflects the article, which provides advice from pain doctors and discusses ways to understand and manage pain better. No Misleading/Clickbait: The language is neutral and not sensational or exaggerated. Stylistic Appropriateness: The headline is concise and direct, not a generic narrative sentence. Entity Accuracy: The main subject is “pain doctors,” which matches the article’s focus.</p> <p><b>Critique:</b> This headline meets all standards.</p> <p><b>Decision:</b> ✓ Correct</p>                                                                                                                                                                                                     |
| <b>User Agent</b>                                                       | <p><b>Reasoning:</b> The candidate headline aligns well with the user’s interest in health, as indicated by the NRMS score and recent click history. The headline is clear and direct but could be more engaging by incorporating a casual or humorous tone to better match the user’s style preferences.</p> <p><b>Critique:</b> Good topic match. The headline is clear but could be more engaging, consider a slightly more casual tone to better match the user’s preferences.</p> <p><b>Score:</b> 0.85</p>                                                                                                                                                                                                                                                            |

personalization-oriented feedback: in Iteration 2, although the Editor approves the headline, the User agent assigns a score of 0.75, noting that the formal tone does not match the user’s preference for a mix of formal and casual styles. Third, the Writer progressively refines its output in response to both the Editor’s and User agent’s feedback, ultimately producing a headline in Iteration 3 that satisfies both the Editor (factually correct) and the User agent (score: 0.85), demonstrating the effectiveness of the collaborative multi-agent framework.

We further note an interesting limitation: the original clickbait headline “24 Secrets Pain Doctors Won’t Tell You” contains an implicit numerical claim. Upon careful inspection of the article, there are indeed 24 tips discussed in the content; however, this number is never explicitly stated in the text. Detecting such implicit numerical information requires non-trivial reasoning over the full article, which current LLMs are not yet reliably capable of. This suggests that while our framework effectively handles explicit factual inconsistencies, implicit reasoning over latent article structure remains an open challenge for future work.

## B.5 Human Evaluation

To complement automatic metrics, we conduct a human evaluation on 1000 randomly sampled test instances. Five annotators compare headlines generated by USG against the state-of-the-art baselines from two categories: (1) a general LLM (GPT-4.1, zero-shot), and (2) a specialized personalized headline generation model (GTP). For each instance, annotators are presented with the article body and the user’s 50 most recent clicked headlines, along with two headlines in random order without knowledge of which system generated them. Following [Liu et al. \(2025\)](#), annotators judge which headline is better across five dimensions: **Fluency**, **Factuality**, **Attractiveness**, **Personalization** (abbreviated as “Personal.” in Figures 5 and 6), and **Overall** quality. Specifically, annotators respond to the following questions for each dimension:

- *Fluency: Which headline is more fluent and grammatically correct?*
- *Factuality: Which headline more faithfully reflects the facts presented in the article?*
- *Attractiveness: Which headline is more engaging?*

- *Personalization: Which headline better matches this user’s preference based on their click history?*
- *Overall: Taking all aspects into account, which headline is of higher overall quality?*

Each comparison yields a *Win* (USG is preferred), *Tie* (comparable quality), or *Lose* (baseline is preferred) judgment.

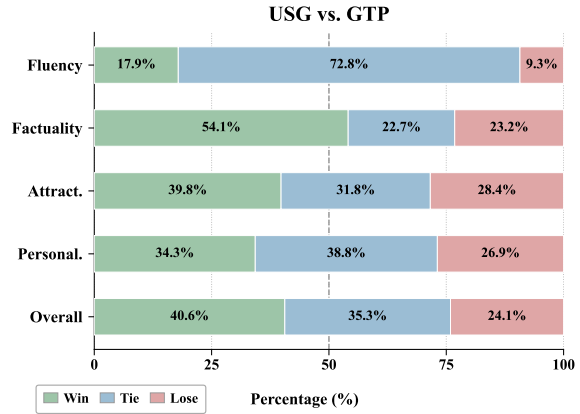


Figure 5: Human evaluation results (USG vs. GTP).

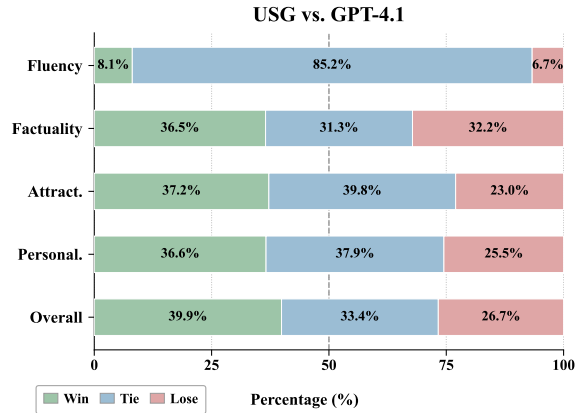


Figure 6: Human evaluation results (USG vs. GPT-4.1).

Inter-annotator agreement is measured using Fleiss’  $\kappa$  ([Fleiss, 1971](#)), with all dimensions yielding  $\kappa > 0.6$ , indicating substantial agreement. As illustrated in Figures 5 and 6, USG achieves more wins than losses across all dimensions against both baselines, validating the effectiveness of our framework. Against GTP, USG demonstrates a decisive advantage in Factuality (54.1% Win vs. 23.2% Lose) and outperforms in Personalization (34.3% Win vs. 26.9% Lose), confirming that the Editor agent effectively enforces factual correctness, while our direct training paradigm better captures

user preferences than indirect approaches. Against GPT-4.1 (zero-shot), USG shows a clear lead in Personalization (36.6% Win vs. 25.5% Lose) and Overall (39.9% Win vs. 26.7% Lose), demonstrating that our multi-agent framework captures individual user preferences more effectively than zero-shot prompting. Notably, USG achieves competitive Factuality (36.5% Win vs. 32.2% Lose), consistent with our automatic FactCC results. The large Tie proportions in Fluency indicate comparable linguistic quality across systems.

## B.6 User Study

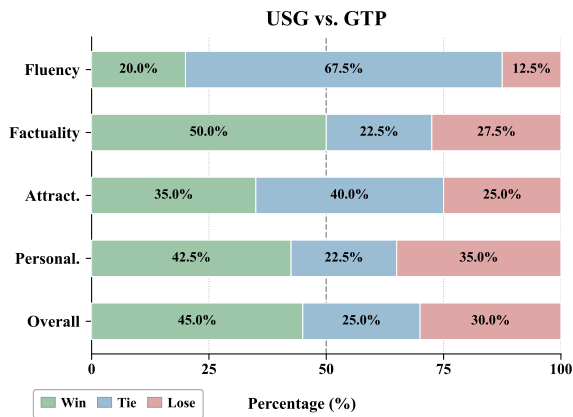


Figure 7: User study results (USG vs. GTP).

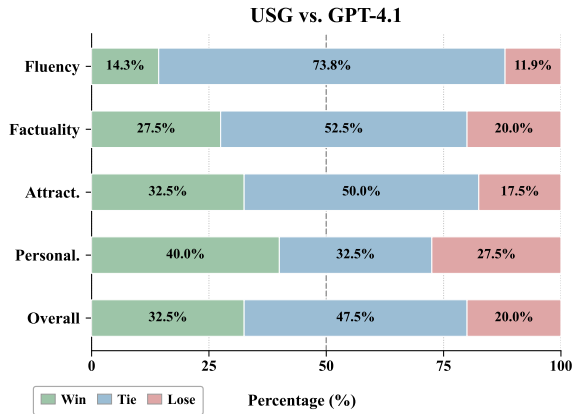


Figure 8: User study results (USG vs. GPT-4.1).

To assess real users’ preferences for generated headlines, we conduct a small-scale user study. Following Liu et al. (2025); Ao et al. (2023), we recruited five English-speaking graduate students as participants. Each participant was first shown a preselected pool of news headlines and asked to select 50 that matched their personal interests; these selections served as a proxy for the participant’s historical click behavior. Based on this data, we

generated personalized headlines for 40 previously unseen news articles. For each article, participants were shown the article body along with two candidate personalized headlines presented in random order, without knowledge of which system produced each one, and asked to indicate their preference along the same five dimensions as in Section B.5: Fluency, Factuality, Attractiveness, Personalization, and Overall quality.

As shown in Figures 7 and 8, USG achieves more wins than losses across all dimensions against both baselines, most notably in Factuality and Personalization. These results are consistent with the trends observed in the human evaluation, suggesting that USG’s advantages generalize beyond simulated preferences to genuine user preferences.

## C Related Work

### LLM-Based User Simulator for Data Synthesis

Personalization tasks typically rely on extensive user-centric data (e.g., interaction logs) to effectively model user preferences (Chen et al., 2024). However, such data is often scarce or inaccessible. Aside from the natural sparsity inherent in cold-start scenarios (Huang et al., 2025), optimizing for subjective objectives (e.g., user satisfaction) necessitates costly and time-consuming user studies (Liu et al., 2026). Furthermore, some collections may raise privacy and compliance concerns (Li et al., 2024). Recently, LLM-based user simulators have emerged as a potential solution for synthesize data at scale, alleviating sparsity while avoiding privacy risks (Zhang et al., 2024; Wang et al., 2025). For example, LAUS (Park, 2025) proposes replacing real user logs with synthetic interactions, leveraging an LLM to predict click probabilities based on generated user histories to construct recommendation datasets from scratch. Similarly, SUBER (Corecco et al., 2024) constructs comprehensive synthetic agents with detailed profiles and interaction histories. These agents simulate responses (e.g., ratings or clicks) that are strictly consistent with their assigned personas, yielding high-quality training signals. RecoWorld (Liu et al., 2026) establishes a multi-turn interaction framework between LLM-simulated user agents and the recommender system to synthesize high-quality user satisfaction data. ColdLLM (Huang et al., 2025) employs an LLM to simulate potential user interactions based on item attributes. This process generates synthetic records that transform cold-start items into "warm" items,



Figure 9: Prompt template for the User agent  $\mathcal{U}$ , which evaluates candidate headlines and provides subjective personalization feedback.

thereby optimizing item embeddings via data augmentation. USERMIRRORER (Wei et al., 2026) integrates the Engel–Kollat–Blackwell (EKB) cognitive model to simulate user decision-making processes. Crucially, it implements a rigorous filtering pipeline to discard low-quality simulations, ensuring the synthesized data aligns closely with real-world user behavior.

### D Prompt Templates

This section provides the complete prompt templates used in our framework. Figures 9, 10, and 11 show the prompts for the User ( $\mathcal{U}$ ), Editor ( $\mathcal{E}$ ), and Writer ( $\mathcal{W}$ ) agents used in Stage 1, and Figure 12 shows the prompt for the Generator ( $\mathcal{G}$ ) in Stage 2.

### E AI Use Disclosure Statement

We employed Gemini for grammatical corrections and spell-checking. All core ideas, methodological design, theoretical derivations, experiments, and results are entirely our own work.

| Editor Agent Prompt Template                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are a Senior News Editor responsible for auditing headline quality and ensuring adherence to editorial standards.</p> <p><b>Task:</b><br/> Validate the headline against the Quality Standards below and provide constructive critique to guide the Writer's refinement.</p> <hr/> <p><b>Quality Standards:</b></p> <ol style="list-style-type: none"> <li>1. Factual Consistency: the headline must faithfully convey core article information without hallucination</li> <li>2. No Misleading/Clickbait: avoid exaggerated or sensationalist language that distorts content</li> <li>3. Stylistic Appropriateness: concise (under 20 words) and engaging, not a generic narrative sentence</li> <li>4. Entity Accuracy: main subject/person must be correctly identified</li> </ol> <hr/> <p><b>Article Body:</b> {article_body}<br/> <b>Headline:</b> {headline}</p> <hr/> <p><b>Output JSON:</b><br/> { "analysis": "step-by-step verification against all standards",<br/> "critique": "If meets ALL standards: None. If violates ANY standard: state Violates [Standard Name] and explain (e.g., Violates Factual Consistency: contains unsubstantiated claim)",<br/> "decision": "[correct] or [incorrect]" }</p> |

Figure 10: Prompt template for the Editor agent  $\mathcal{E}$ , which audits headline quality against editorial standards and provides objective feedback.

| Writer Agent Prompt Template                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are a Personalized Headline Writer balancing editorial standards and user preferences.</p> <p><b>Task:</b><br/> Generate a headline that simultaneously satisfies two objectives:</p> <ol style="list-style-type: none"> <li>1. Editorial Quality: factually accurate, no misleading/clickbait, stylistically appropriate, entity-accurate</li> <li>2. User Preference: aligned with the user's topic interests and reading style</li> </ol> <p><b>Generation Constraints:</b></p> <ol style="list-style-type: none"> <li>1. Extractive: use vocabulary from the article body only (no hallucinations)</li> <li>2. Concise: under 20 words</li> <li>3. User-written style: avoid journalistic jargon.</li> <li>4. Address ALL critique instructions from the Memory Module</li> </ol> |
| <hr/> <p><b>Memory Module</b> (Accumulated Interaction History): {memory}</p> <hr/>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <p><b>User Context:</b><br/> [Long-term Profile]: {u_prof}<br/> [Short-term Click History]: {u_hist}</p> <hr/>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
| <p><b>Article Body:</b> {article_body}</p> <hr/>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <p><b>Now reason step by step:</b><br/> Step 1 – Fact Extraction: identify the most important event, outcome, and entity from the article body<br/> Step 2 – Preference Inference: what topics and styles does this user prefer based on profile and history?<br/> Step 3 – Critique Integration: for each entry in Memory, ensure the new headline addresses editor and user feedback<br/> Step 4 – Generation: produce a headline that maximizes editorial validity and user preference</p> <hr/>                                                                                                                                                                                                                                                                                          |
| <p><b>Output JSON:</b><br/> { "reasoning": "(1) key facts extracted, (2) inferred user preferences, (3) how Memory critiques are addressed",<br/> "headline": "the generated personalized headline (max 20 words)" }</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |

Figure 11: Prompt template for the Writer agent  $\mathcal{W}$ , which generates and refines personalized headlines based on feedback from the Editor and User agents.

| Generator Prompt Template                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>You are a Personalized Headline Writer balancing editorial standards and user preferences.</p> <p><b>Task:</b><br/> Generate a headline that simultaneously satisfies two objectives:</p> <ol style="list-style-type: none"> <li>1. Editorial Quality: adherence to the Editorial Standards below</li> <li>2. User Preference: aligned with the user's topic interests and reading style</li> </ol> <p><b>Generation Constraints:</b></p> <ol style="list-style-type: none"> <li>1. Extractive: use vocabulary from the article body only (no hallucinations)</li> <li>2. Concise: under 20 words</li> <li>3. User-written style: avoid journalistic jargon</li> </ol> <p><b>Editorial Standards:</b></p> <ol style="list-style-type: none"> <li>1. Factual Consistency: faithfully convey core article information without hallucination</li> <li>2. No Misleading/Clickbait: avoid exaggerated or sensationalist language that distorts content</li> <li>3. Stylistic Appropriateness: concise and engaging, not a generic narrative sentence</li> <li>4. Entity Accuracy: main subject/person must be correctly identified</li> </ol> <hr/> <p><b>Memory Module:</b><br/> [User Long-term Profile]: {u_prof}<br/> [User Short-term Click History]: {u_hist}</p> <hr/> <p><b>Article Body:</b> {article_body}</p> <hr/> <p><b>Now reason step by step:</b><br/> Step 1 – Fact Extraction: identify the most important event, outcome, and entity from the article body<br/> Step 2 – Preference Inference: what topics and styles does this user prefer based on profile and history?<br/> Step 3 – Generation: produce a headline that maximizes editorial validity and user preference</p> <hr/> <p><b>Output JSON:</b><br/> { "reasoning": "(1) key facts extracted, (2) inferred user preferences, (3) brief explanation of the generation process",<br/> "personalized_headline": "the generated personalized headline (max 20 words)" }</p> |

Figure 12: Prompt template for the Generator  $\mathcal{G}$ , which directly generates personalized headlines.