Personalizing News Headlines with Retrieval-Augmented Generation

Jiajing Wan[1], **Samia Touileb**[1], **Lubos Steskal**[2], **and Lilja Øvrelid**[3]
[1]University of Bergen
[2]TV 2 Norway
[3]University of Oslo
{Jiajing.Wan, Samia.Touileb}@uib.no
Lubos.Steskal@tv2.no
liljao@ifi.uio.no

Abstract

We focus on personalized news headline generation, where we aim to improve headline generation by extending the generation context to incorporate the news reading history of users. In particular, we study a RAG-LLM-based system that customizes news headlines with user histories to improve news headline personalization. Our experiments show that our approach not only produces better headlines for specific users, but also makes the generated headlines closer to the original headlines. We experiment with different retrievers and analyze the generated outputs through systematic comparisons with both original and rewritten headlines. These analyses provide insights into the role of retrieval and personalization in headline generation, highlighting how the user history contributes to meaningful improvement while remaining aligned with original headlines[1].

1 Introduction

With the recent advances in large language models (LLMs), there has been a significant improvement in both the quality of generated content and the possibility to process various types of input (Kumar, 2024). LLMs can be used to generate personalized content tailored to individual user preferences, including tasks such as personalized news headline generation (Shi et al., 2025; Salemi et al., 2024; Ren et al., 2025). How news are presented might also be seen as a way to reduce news avoidance among intentional avoiders who feel overwhelmed by the large volume of available news (Skovsgaard and Andersen, 2020). Offering news in alternative framings can potentially help these users engage with news that they would otherwise overlook due to the phrasing of the headline.

One important contribution towards more personalized news headline generation is the introduction of the PENS (PErsonalized News headlineS) dataset (Ao et al., 2021). In addition to including news headlines and articles, Ao et al. (2021) also incorporate user browsing records as user behavior data to achieve personalization by customizing news headline from the user side. More recently, the LaMP dataset (Salemi et al., 2024) explored the use of Retrieval-Augmented Generation (RAG) for incorporating personalization into LLM generation , including personalized news headline generation. They proposed a more flexible and direct approach to concatenate and adjust input prompts for LLMs based on different candidate news. However, LaMP is limited and doesn't include full articles and metadata in both candidate news and user history, restricting its use for RAG-based personalization.

Customized personalization is approached through different mechanisms. Due to limited amounts of user-specific data, more stable customized personalization can be achieved by fine-tuning lightweight personalization modules (Song et al., 2023; Li et al., 2024), or adjusting the generation objective through offline goal-conditioned reinforcement learning to accommodate continuously evolving user interests (Tan et al., 2024). Nevertheless, these approaches mainly rely on encoding preferences as representations, which cannot be explicitly interpreted. In contrast, we aim to make use of the strengths of RAG to directly retrieve user records and incorporate them as contextual signals for LLMs.

In this paper, we propose a RAG-based system for customized news headline generation and apply it to the PENS dataset. In our experiments, we evaluate three different LLMs as generators and test various retrievers to investigate how user records retrieved under different retrieval setups influence the final generation outcome. We conduct a detailed analysis of headlines produced by different models, examining stylistic similarity, n-gram overlap, and entity-level statistics. Our experiments reveal that to generate headlines that best align with

[1]Our code will be made available on GitHub at https://github.com/lorylei/RAGpens.

Proceedings of the Second Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U), pages 55–67

user preferences, the retrieved user records that served as generation references should satisfy three conditions simultaneously: be of a certain quantity, have high relevance to the candidate news, and have adequate topical diversity. Furthermore, while RAG models demonstrates evidence of improved user-specific personalization, it also shows that the generated headlines are better aligned than general journalistic styles in the user's clicked history. Based on this finding, we conducted an in-depth study on how the model balances personalized signals with style consistency and proposed the necessity of a dual-reference evaluation.

2 Related work

Personalized news headline generation before LLMs Personalized news headline generation research has focused on generating a unified style without accounting for individual user preferences, e.g. by creating sensational news headlines (Xu et al., 2019). Ao et al. (2021) were the first to propose a framework for customized personalization of news headline generation, incorporating user browsing records as user representation in the generation process. Their baseline combines user representations with the input article and uses a Bi-LSTM pointer-generator network (See et al., 2017). User representations in personalized news headline generation tasks before LLMs were mostly incorporated during training in the form of embeddings (Ao et al., 2021; Zhang et al., 2022). Such representations can struggle to adapt context-dependent personalization signals to different types of news, thereby limiting personalization capabilities.

Personalized news headline generation using LLMs There is a growing attention to the personalization capabilities of LLMs, especially for adjusting their outputs based on user preferences with methods such as P-RLHF (Li et al., 2024) and difference-aware user modeling (Qiu et al., 2025). Salemi et al. (2024) introduced the LaMP benchmark for both personalized generation and classification tasks. However, in the personalized news headline generation dataset of the LaMP benchmark, neither the user records nor the candidate news contain complete news articles (the average article length is only 112 characters), which potentially limits the quality of the generated headlines.

Compared to previous works, and to fully exploit RAG and compare the effects of different retrievers, we apply RAG generation on the PENS dataset using complete news articles, and further analyze how the retrieved content from different retrievers influences performance.

3 Experimental setup

Dataset The publicly available PENS dataset (Ao et al., 2021) contains 113,762 news articles, with individual IDs, titles, body texts, and 15 news categories. For the test set, 103 native English speakers manually created 200 personalized headlines each, without seeing the original headlines. These rewritten headlines were then reviewed by editors. In addition, using each participant's browsing history, 50 news items were selected to create their personal user record. To support our RAG experiments, we reorganized the test set, by treating each manually created personalized headline and its associated article as a separate candidate instance. Each instance is linked to the user who created the personalized headline, and their full browsing history, resulting in 20,600 user-specific examples. In the experiments, we include the original news headlines as an additional point of comparison while treating the rewritten headlines as the gold standard for personalized reference , to better understand and compare the degree of personalization introduced by the generated and the rewritten headlines.

Generation Models In our experiments, we use several instruction-tuned LLMs with comparable parameter scales (7–8B) and extended context windows of 128k tokens. This setup ensures a fair comparison across models while guaranteeing that the entire content of each news article can be fully covered within the input context. For generation, we set the maximum context length to 70k tokens to balance efficiency and coverage, and restrict the output length to 64 tokens. We use greedy decoding by setting the *temperature* to 0 and *best-of* to 1, ensuring deterministic outputs for headline generation. We use the following models:

- **Llama 3**: Model series designed to support multilingualism, coding, and reasoning (Grattafiori et al., 2024). We use *Llama-3.1-8B-Instruct* with 128k input context length.
- **Qwen**: LLMs designed to support a wide range of NLP and generation tasks (Team et al., 2024). We use *Qwen2.5-7B-Instruct*, an instruction-tuned variant with a size of 7B, and input context length of 128k.

- **Granite**: Family of models optimized for various tasks including long-context, instruction-following, and reasoning tasks (IBM, 2024). We use *Granite-3.3-8B-Instruct*, an 8B parameter model with a 128K token context window.

RAG-based Personalized Generation A RAG-based personalized generation model involves three main stages: retrieval, reranking, and generation. We compare the use of the BM25 algorithm (Robertson et al., 1995), a sparse lexical-matching method, with a dense retriever (Karpukhin et al., 2020), which encodes queries and documents into a shared semantic embedding space, to reorder user records during the retrieval phase. The BM25 retriever is implemented using the following formula:

$$\text{score}(Q, D) = \sum_{t \in Q} \text{IDF}(t) \cdot \frac{f(t,D) \cdot (k_1+1)}{f(t,D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)} \quad (1)$$

Each user record is represented as a document D, constructed by concatenating its title and article body. During retrieval, the candidate news article to be generated is used as the query Q. The BM25 algorithm then scores and ranks all documents D in the user profile compared to Q. The $f(t, D)$ is the term frequency of token t in document D; $|D|$ is the document length; and `avgdl` is the average length over the entire user profile. $\text{IDF}(t)$ represents each token's inverse document frequency[2].

For the dense retriever, we use the open-source model *bge-base-en-v1.5* (Xiao et al., 2023) as the pre-trained English text embedding model. In our implementation, the input query Q, corresponding to the candidate news article, is encoded with a "`query:`" prefix[3] into an embedding $\mathbf{q}$. Similarly, each news in the user records is treated as a candidate passage d_i, constructed by concatenating the article title and article body, and encoded with the prefix "`passage:`" to obtain embeddings $\mathbf{p}_i$. The similarity between `query` and `passage` embeddings is computed as the dot product: $s_i = \mathbf{p}_i^\top \mathbf{q}$. The passage with the highest score s_i is selected as the Top-1 retrieval result. After re-ranking for both retrievers, the model selects the top-k entries with the highest scores from the reordered records. These entries are then concatenated with the current input prompt to form a new prompt, which is subsequently passed to the generation model.

[2]We use the default hyper-parameters $k_1 = 1.5$ (controls TF saturation) and $b = 0.75$ (balances length normalization) in `BM25Okapi` from the `rank_bm25` Python package.

[3]This prefixing strategy is recommended for BGE models, which suggests using query instructions (BAAI, 2023).

Evaluation We use ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002), which measure n-gram overlap and structural similarity with references, as well as BERTScore (Zhang et al., 2019), which assesses semantic faithfulness. Following recent research on evaluation methods for style-personalized text generation (Jangra et al., 2025), we evaluate using `StyleDistance`, a content-independent style embedding model trained with contrastive learning on synthetic near-paraphrase data, which are generated by a LLM with controlled stylistic variations (Patel et al., 2025). The resulting embeddings allow us to compute stylistic similarity using cosine similarity.

We also implement an LLM-as-a-Judge evaluation setup using Mistral-7B-Instruct-v0.3 (AI, 2025) and GPT-4o-mini (OpenAI, 2024) to assess the personalization degree of the generated results, to examine whether evaluation results remain consistent across an open-source and a proprietary LLM, as shown in Table 11 in Appendix A. We use all user's past clicked headlines as references, and both LLM judges are tasked with selecting which headline among the generated headlines, the original headlines, and the rewritten headlines the user would be most interested in based on the provided references. We also compare our results to previous works (Ao et al., 2021; Yang et al., 2023; Lian et al., 2025).

4 Experiments

We present here our experiments, where we use RAG for personalized news headline generation[4].

LLM selection We evaluate the selected LLMs on the PENS dataset without incorporating user records. As shown in Table 1, the Llama 3 model achieves best results among all LLMs. We therefore focus primarily on this model in the remainder of our experiments. We can further observe that, compared with headlines generated by any existing model, the original news titles show higher overlap in terms of both n-gram similarity and semantic faithfulness with rewritten headlines. This suggests that users' personalized rewriting is not a complete redesign of the headline, but rather a subtle rewriting that retains the necessary content. We therefore, in our following experiments, compare the outputs generated by different RAG models with Llama 3 against both the original headlines and the rewritten

[4]All details about infrastructure and parameter settings of our experiments are available in Appendix A.1.

	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore		
					F1	Precision	Recall
PENS-NAML†	26.69	10.01	23.02	/	/	/	/
FPG-GRU†	27.33	10.51	23.30	/	/	/	/
SCAPE†	34.26	14.97	28.36	/	/	/	/
Llama 3	31.18	12.63	26.00	34.48	87.08	86.99	87.20
Qwen	28.38	9.94	23.36	32.92	86.75	86.57	86.96
Granite	28.80	10.68	23.31	32.51	86.56	85.87	87.29
Original headline	**45.30**	**25.48**	**35.24**	**45.89**	**88.84**	**89.13**	**88.59**

Table 1: ROUGE and BLEU results of different LLMs compared to rewritten headlines. Results for Baseline PENS-NAML† are reproduced in our implementation following the settings of Ao et al. (2021). Results for FPG-GRU† and SCAPE† are reported from Yang et al. (2023) and Lian et al. (2025). The row with the original headline represents the comparison results between the original headlines and the rewritten headlines. Bold numbers indicate the best results in each column.

headlines. We also analyze the factors contributing to the observed differences.

4.1 RAG-based generation experiments

Retriever selection We use Llama 3 as the generation model and pair it with different retrievers. We also include a random retriever as a baseline. We also vary the number of retrieved user records k from 1 to 11, to explore the effect of adding more contextual information about the users.

The results of comparing the generated news headlines with the ROUGE and BLEU scores of the rewritten and the original news headlines are shown in Table 2 and Table 3 respectively. RAG-based models consistently outperform the non-personalized baseline across both original and rewritten headlines with BM25 at k>3, dense at k>4, and random at k>8 on all dimensions. The BM25 retriever achieves the best ROUGE and BLEU scores. Using the RAG framework simultaneously increases the n-gram similarity between the generation results, and both the original news headlines and the user-rewritten news headlines. This improvement, however, is more pronounced compared to the original news headlines.

BERTScore values remain relatively similar across models and the headlines we are comparing the generations against (original and rewritten). We also see that the generated news headlines are semantically closer to the original headlines (Table 7 and Table 8 in Appendix A). This suggests that all models preserve a similar level of semantics, and improvements from RAG mainly arise from lexical and stylistic changes rather than semantic ones. Nevertheless, the BERTScore between generated and original headlines shows a relatively clear increasing trend as the number of user records grows, with BM25-based RAG models performing slightly better. For the results compared to rewritten headlines, only Recall improves with more user records, while other dimensions show no clear pattern, and no retriever demonstrates an obvious advantage.

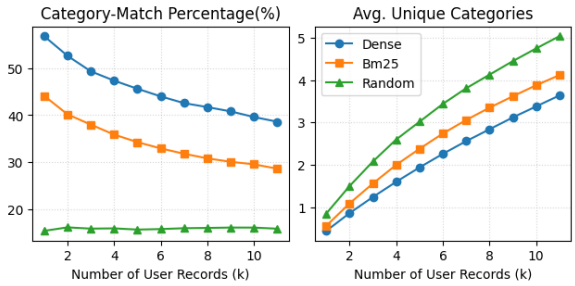

Figure 1: Category match and average number of unique categories across records using different retrievers.

Retrieved news category We also analyze the category of each news item and compute the (i) category-match percentage, which represents the proportion of retrieved user records sharing the same category, and (ii) the average number of unique categories, which is the average number of distinct categories among retrieved records. Shown in Figure 1, the dense retriever achieves the highest match percentage but the lowest category diversity. As the number of retrieved records k increases, category match consistently decreases for both dense and BM25 retrievers due to less relevant and low ranked items. All RAG models improve as the number of user records k increases. After k>7, the performance of both dense and BM25 retrievers are gradually stabilized, whereas the random retriever remains the weakest and most unstable, which confirms that ensuring sufficient category match relevance between the retrieved records and the candidate news is crucial. Overall, the BM25-based RAG achieves the best generation performance by maintaining a moderate trade-off between category relevance and category diversity.

	ROUGE-1			ROUGE-2			ROUGE-L			BLEU		
	BM25	dense	random	BM25	dense	random	BM25	dense	random	BM25	dense	random
w/o User records	40.03			20.24			35.09			41.64		
1 User records	39.03	39.63	39.23	19.60	19.74	19.33	34.09	34.55	34.19	40.86	41.35	40.95
2 User records	40.14	40.04	39.48	20.29	20.06	19.38	35.03	34.96	34.39	41.86	41.79	41.21
3 User records	40.48	40.03	39.52	20.60	20.13	19.59	35.32	34.99	34.51	42.18	41.74	41.22
4 User records	40.52	40.23	39.63	20.65	20.35	19.70	35.35	35.18	34.64	42.23	41.93	41.36
5 User records	40.73	40.36	39.75	20.85	20.44	19.85	35.54	35.25	34.74	42.43	42.06	41.48
6 User records	40.82	40.48	39.83	20.99	20.60	19.94	35.61	35.35	34.97	42.44	42.19	41.73
7 User records	40.90	40.49	39.90	21.08	20.64	19.94	35.69	35.35	34.80	42.54	42.21	41.62
8 User records	**40.95**	40.61	40.21	**21.14**	20.73	20.33	35.72	35.46	35.15	42.56	42.35	41.91
9 User records	40.95	40.68	40.05	21.13	20.88	20.21	**35.73**	35.54	35.02	**42.60**	42.40	41.92
10 User records	40.89	40.70	40.08	21.11	20.89	20.22	35.66	35.53	35.00	42.53	42.39	41.83
11 User records	40.93	40.69	40.13	21.12	20.88	20.35	35.73	35.53	35.11	42.57	42.39	41.84

Table 2: ROUGE and BLEU results compared to the original news headlines using different numbers of users' records. Bold numbers indicate best performance. All the best results are generated by Llama 3 combined with user records retrieved by BM25.

	ROUGE-1			ROUGE-2			ROUGE-L			BLEU		
	bm25	dense	random	bm25	dense	random	bm25	dense	random	bm25	dense	random
w/o User records	31.18			12.63			26.00			34.48		
1 User records	30.15	30.74	30.66	11.90	12.14	12.12	25.39	25.84	25.88	33.55	32.27	34.04
2 User records	31.01	31.05	30.87	12.47	12.42	12.24	26.20	26.21	26.08	34.36	34.44	34.29
3 User records	31.29	31.13	30.99	12.68	12.53	12.36	26.43	26.32	26.19	34.58	34.52	34.32
4 User records	31.35	31.30	31.10	12.74	12.62	12.46	26.46	26.45	26.35	34.68	34.58	34.41
5 User records	31.38	31.41	31.11	12.74	12.73	12.46	26.48	26.54	26.33	34.72	34.73	34.50
6 User records	31.49	31.39	31.19	12.86	12.74	12.54	26.56	26.50	26.39	34.74	34.74	34.52
7 User records	31.50	31.46	31.23	12.90	12.77	12.56	26.52	26.56	26.42	34.80	34.76	34.57
8 User records	31.55	31.47	31.36	12.93	12.74	12.67	26.56	26.53	26.52	34.80	34.81	34.67
9 User records	**31.61**	31.49	31.32	12.99	12.81	12.71	**26.61**	26.57	26.47	**34.88**	34.81	34.69
10 User records	31.54	31.52	31.38	**13.00**	12.86	12.69	26.53	26.59	26.52	34.82	34.83	34.68
11 User records	31.45	31.50	31.46	12.91	12.84	12.81	26.47	26.56	26.59	34.78	34.80	34.75

Table 3: ROUGE and BLEU results compared to the rewritten news headlines using different numbers of users' records. Bold numbers indicate best performance. All the best results are generated by Llama 3 combined with user records retrieved by BM25.

Personalization evaluation Although ROUGE, BLEU, and BERTScore measure lexical overlap and semantic similarity between different headlines, they struggle to capture the personalization effects associated with abstract style in headline generation. To address this limitation, we use `StyleDistance` (Patel et al., 2025) to evaluate content-independent stylistic similarity between the original headline, the generated headline, and the rewritten headline that best represents a user's interests. In addition, we use the LLM-as-a-Judge framework described above to simulate user perspectives when choosing the headline they are most interested in among different headlines. The stylistic similarity results comparing the generated headlines with both the original and rewritten headlines change with the number of user records, and the choice of retrieval model as shown in Table 4. First, we observe that the original headlines are still the most stylistically similar to the rewritten headlines, and the generated headlines are stylistically closer to the original title. Overall, all RAG models improve the stylistic similarity of the generated headlines with both the original and rewritten headlines compared to the results without RAG. The BM25-based RAG model is also the best performing.

This result is consistent with the findings observed in the ROUGE and BLEU score tables. Increasing the number of user records enhances the stylistic similarity between the generated and original headlines. However, the highest stylistic similarity to the rewritten headlines occurs when the number of user records is five or fewer. As the number of user records increases over five, the stylistic similarity to the rewritten headlines begins to fluctuate and shows a slight decline.

In the LLM-as-a-judge experiment, we compare the framework with eight user records, since the performance of the RAG model stabilizes when k>7 on the ROUGE and BLEU evaluations with Mistral-7B-Instruct-v0.3 (AI, 2025) and GPT-4o-mini (OpenAI, 2024). The results, in Figure 2, show that when the headlines from the users' clicked histories are used as references, the judge models generally tend to prefer the original headlines over the generated headlines. Using RAG

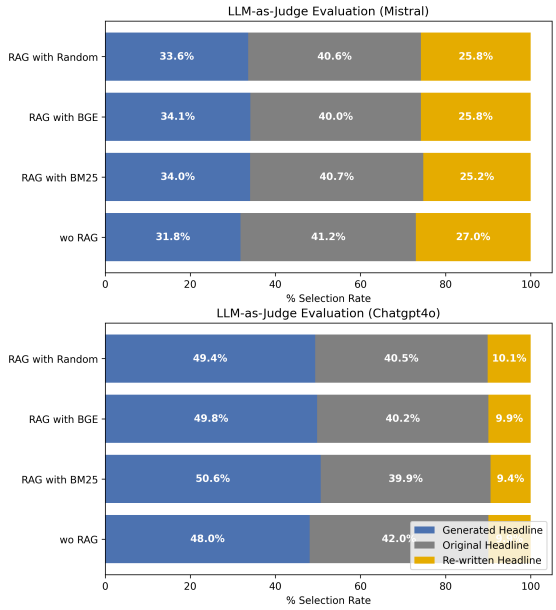

Figure 2: Comparison of selection rates for generated headlines versus original and rewritten news headlines in the LLM-as-a-judge evaluation. Overall, LLMs tend to prefer the generated headlines and original headlines.

	Rewritten headlines			Original headlines		
	BM25	dense	random	BM25	dense	random
Original headlines	76.62			/		
w/o User records	74.55			79.96		
1 User records	74.67	74.35	74.53	80.15	80.04	80.04
2 User records	74.73	74.56	74.58	80.32	80.15	80.15
3 User records	74.74	74.63	74.62	80.37	80.12	80.12
4 User records	74.76	74.65	**74.68**	80.33	80.21	80.21
5 User records	**74.76**	**74.69**	74.64	80.42	80.21	80.21
6 User records	74.73	74.6	74.61	80.44	80.21	80.21
7 User records	74.73	74.64	74.63	80.47	80.21	80.21
8 User records	74.74	74.57	74.58	80.48	80.23	80.23
9 User records	74.66	74.60	74.60	80.45	80.22	80.22
10 User records	74.68	74.53	74.59	80.49	80.21	80.21
11 User records	74.70	74.62	74.59	**80.52**	**80.27**	**80.27**

Table 4: The cosine similarity results, computed by `StyleDistance`. Bold numbers indicate the best performance in RAG. Both the results, which are stylistically closer to the original headline and the rewritten headline, were generated using Llama3-based RAG model with BM25.

models, especially those combined with effective retrievers, increases the probabilities that the generated headlines are selected. The rewritten headlines, which theoretically should best reflect user preferences, are the least favored by the models.

Qualitative analysis To gain a better understanding of the effect of the RAG-based generation using user histories we perform a manual analysis. We select 100 examples in which the RAG-based model outperforms the w/o RAG baseline in both ROUGE and BLEU scores, when evaluated on both the rewritten and original headlines, and additionally check 20 examples where the RAG-based model performs worse than the w/o RAG baseline under the same evaluation settings, and present two representative examples from both sides in Table 5.

Generally, we observe that the model with BM25 is better at capturing phrase-level stylistic patterns and incorporating them into headline generation (e.g. the formulation of the "Top news: . . . " headline prefix). However, when the original news headlines or the rewritten news headlines employ more abstract and descriptive language, extracting phrases from the browsing history and reproducing key information from the news content in the generated headline becomes less effective. This is illustrated by the second example in Table 5, where none of the generated headlines match the more abstract description of "Going Dark and Fake Voyages". To further investigate how the model extracts stylistic patterns from user histories and applies them to headline generation, we analyze another 100 examples of the model that achieve best score on rewritten headlines in terms of `StyleDistance` (RAG with BM25, $K = 5$) and the model that shows more stable performance on ROUGE scores (RAG with BM25, $K = 8$). From these results, we discuss two representative examples. In Table 6, we can see that as the number of user records increases, the model tends to favor a concise, entity-centered style inherited from the user's profile, which may lead it to overlook certain style-level repetitions (e.g. the use of quotation marks in "different"). To further validate this observation, we conduct a named entity recognition analysis on all generated headlines from different RAG models using DistilBERT (Sanh et al., 2019). As shown in Figure 3, the average number of entities contained in each generated headline first slightly decreases and then increases as more user records are incorporated. Moreover, the generated headlines consistently contain a higher number of entities than both the original and the rewritten headlines, and the headline produced by RAG with BM25 have the highest average number of entities. The remaining examples can be found in Table 10 and Table 9 in Appendix A.

Input news	Here's what you need to know about what's happening in Phoenix. APS worker dead after fire in underground electrical vault in downtown Phoenix According to Phoenix FD...	
Original news headline	Top Phoenix news: APS worker dead after fire in electrical vault; buildings still without power.	
Rewritten news headline	APS worker dead after fire in underground electrical vault in downtown Phoenix; high-rise power cut down possibly lasts for a few days	
Model	Generation	User profile
RAG(BM25,K=8)	Top Phoenix news: APS worker dead after fire, Phoenix buildings without power, and more.	1: Top Berkeley news... ...
RAG(Dense,K=8)	APS worker dead in downtown Phoenix fire, more	1: 1 dead,several injured... ... 7: Top Berkeley news... ...
RAG(Random,K=8)	Phoenix News Briefs: Worker Dies in Underground Fire, Harassment Allegations, and More	1: Trump raises 2020 stakes... ...
w/o RAG	Phoenix News Briefs: Fatal Fire, Harassment Allegations, and NBA Updates	/
Input news	As the U.S. looks to enforce sanctions aimed at halting trade with states such as Iran and Venezuela, there's increased scrutiny on the tricks...	
Original headline	Going Dark and Fake Voyages: The Tricks Used to Dodge Trade Sanctions	
Rewritten headline	Trade sanctions: how sanctioned countries still trade	
Model	Generation	User profile
RAG(BM25,K=8)	Sanctions Busters: How Ships Are Evading Trade Restrictions	1: America... trade deficit... ...
RAG(Dense,K=8)	Sanctions Busters: How Iran and North Korea Evade Trade Restrictions	1: America... trade deficit... 2: ... Iran strikes... ...
RAG(Random,K=8)	Sanctions Busters: How Iran and North Korea Evade Trade Restrictions	1: Trump walks back statements ... ...
w/o RAG	Sanctions-Busting Tactics Used to Hide Trade with Blacklisted Countries	/

Table 5: Examples of personalized news headline generation based on different RAG frameworks. Bold text indicates the model with the best performance in each example. Blue text refers to phrase-level repetitions.

Titles in browsing history from user profile		- Trump Slams California On Homelessness, Threatens To 'Intercede' - Bombing Range or Nature Preserve? A Battle for Control of the Nevada Desert - Chris Watts Says He Found God After Choking the Life Out of His Wife and Daughters - Trump vows to deport'millions' of migrants, but it's unclear if there is a plan for mass arrests and removals - Iowa weather: New record for wettest 12 months in state history - Louisville considering handing over youth detention center to state - UFC 239: Ben Askren believes he's one of the best in the world and plans on beating up Jorge Masvidal to prove it - Throwback: the secret Jaguar concept car they never made
Generated news headline	RAG(BM25,K=8)	**North Korean Leader Kim Jong-un Believes** US is Seeking Regime Change, But Thinks Trump is Different
	RAG(BM25,K=5)	US Intelligence Says North Korea's **Kim Jong-un Believes** Trump is "Different"
Original news headline		**Kim Jong-un believes** Trump is "different," State Dept's intel arm assesses
Rewritten news headline		**North Korean Leader Kim Believes** Trump to be "Different" According to State Department Sources
Titles in browsing history from user profile		- Mauldin PD sex scandal: SLED declines investigation request - Trump Approves Strikes on Iran, but Then Abruptly Pulls Back - Underprivileged black youth at NASA became pioneer for racial equality - At G-20, Donald Trump to talk to Saudis about Iran - not Khashoggi - Heavy Rains, Growing Sinkhole Causes Problems For Rostraver Towing Company - Cloud of Cancer-Causing Chemical Hangs Over the Houston Channel - Iowa State student charged with sexual abuse pleads to lesser charge - Automakers Send Letter to Trump about Plan to Lower Fuel-Economy Rules
Generated news headline	RAG(BM25,K=8)	North Carolina **Animal Sanctuary Cited for Safety Violations** After Lion **Mauled Intern**
	RAG(BM25,K=5)	**Animal Sanctuary Cited** for Safety Violations After Lion Kills Intern
Original news headline		**Animal sanctuary cited** for safety violations months lion **mauls intern**
Rewritten news headline		Safety Violation Cited When Lion Fatally **Mauled Intern**

Table 6: Examples of personalized news headline generation based on different numbers of records. Bold text indicates similar phrases. Blue text indicates similar style-level repetitions. In these examples, the model using 8 records achieves higher ROUGE scores in its generated outputs, while the model with 5 records produces results with better stylistic similarity compared to rewritten headlines.

5 Discussion

Our results demonstrate that LLMs achieve strong performance compared with previous models. The evaluation results for both ROUGE and style distance show that news headlines generated by the RAG models tend to be closer to the rewritten headlines compared to non-personalized generation. However, the RAG models taking into account the user history seems to maintain the similarity to the original headline more effectively. Notably, our BM25-based RAG consistently delivers

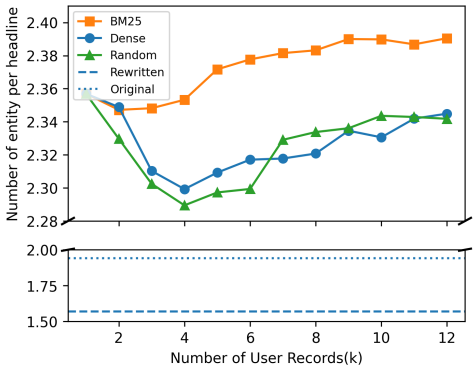

Figure 3: The average number of entities per headline across user records k.

stronger performance than dense retrievers, despite its lower retrieval accuracy. Based on our qualitative analysis, we observe that BM25's sensitivity to features in lexical overlap beyond content relevance, can sometimes help generation by retrieving stylistically similar records. However, we still lack more precise evidence to clarify which specific properties of BM25, and under what conditions, contribute to this behavior. During our manual inspection and qualitative analysis, we found that most generated headlines follow a mainstream journalistic style: concise, and event-centered. The RAG framework tends to confuse the mainstream journalistic style present in users' browsing histories with the users' personal reading preferences, which may limit the adaptability of the generation model to users favoring alternative styles.

A similar discrepancy occurs in evaluation, where recurring headline styles from user history are confused with user preferences. Although the results from both the LLM-as-a-judge and `StyleDistance` evaluations indicate that RAG can improve stylistic similarity, this improvement remains limited. Across different evaluations, the headlines generated by RAG are often closer to the original news headlines. In the LLM-as-a-judge evaluation, we observe that when the clicked headlines are used as references, the model generally prefers either the original headlines, which share the general media styles with the user's history clicks, or the generated headlines that also follow this style. This creates a disconnect from using rewritten headlines as the gold reference for personalization, raising an important question: when defining a personalization reference, should we rely on the user's observed behavior or on their preferred linguistic expression?

Based on the current results, when user history is used as retrieval context, RAG-based models tend to reproduce stylistic patterns that are common in news media, making it difficult to disentangle general journalistic style from user-specific preferences. Existing work on personalized generation has not explicitly addressed this distinction. Moreover, this ambiguity is not only a modeling issue but also reflected in the data itself. There is no strict boundaries in the users' personalized style and the standard news style they are frequently exposed to, which is supported by our observations that rewritten headlines remain consistently closer to original headlines in terms of both n-gram overlap and stylistic similarity than any generated outputs. Therefore, we contend that both types of headlines should jointly serve as gold references for personalization: one as a reference for the general news style alignment, and the other as a reference for the user's style expression alignment. This dual alignment allows the model to preserve general journalistic conventions while also capturing subtle variations that reflect user preferences.

6 Conclusion and future work

The RAG-based approaches for personalized news headline generation demonstrate that using the users' browsing history can effectively improve personalized news headline generation. We further find that in RAG systems, where retrieved records serve as user representation, quantity, diversity, and relevance of user records together play important roles for the generation. Results from our multiple experiments further reveal that, both in generation and evaluation, LLMs exhibit a strong reliance on general journalistic writing paradigms. In this context, RAG's performance stems primarily from the model's improved ability to learn and reproduce the styles of the news media that are embedded in users' records, rather than from effectively capturing users' personal reading preferences. This finding highlights a fundamental challenge in personalized text generation: distinguishing between stylistic conformity and true personalization.

Moving forward, while preserving the model's capacity to learn news writing styles, we plan to explore how records capturing users' reading styles can be identified and more effectively integrated into headline generation , while disentangling it from global media styles and designing evaluation protocols that better reflect personalization.

Limitations

Due to computational limitations, we were restricted during our experiments to only evaluate three open-source LLMs with comparable parameter scales. Using other, bigger or even more advanced, models could potentially further improve the baseline performance of headline generation.

Additionally, again due to computational restrictions, we did not use all 50 user records available for each news item in the dataset. It is therefore possible that the performance trends observed across different RAG experiments may change when k>11. Also, we have not fully exploited the retrieved results. During our qualitative analysis, both the dense and BM25 retrievers occasionally include the same user record in the prompt. However, since all records are ranked in descending order by retrieval score, the BM25-based model produced better generation results. This suggests that the ranking order of user records may play an important role in shaping the generation outcome.

We also lack methods for detecting the factuality and consistency of the generated headlines. None of the existing evaluation methods in this research, whether n-gram–based metrics, style similarity measures, or BERTScore, are capable of addressing this issue. Based on the `StyleDistance` score and our qualitative analysis, we found that when using the RAG model to extract style from user records, it is difficult to separate the user's personal reading style from the general media stylistic choice. Moreover, incorporating more user records will further overshadow the user's preferred reading style.

References

Mistral AI. 2025. Mistral-7b-instruct-v0.3. https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. Hugging Face model card.

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

BAAI. 2023. Baai/bge-base-en-v1.5. https://huggingface.co/BAAI/bge-base-en-v1.5. Hugging Face model card.

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*.

IBM. 2024. Granite 3.3-8b-instruct. https://huggingface.co/ibm-granite/granite-3.3-8b-instruct.

Anubhav Jangra, Bahareh Sarrafzadeh, Silviu Cucerzan, Adrian de Wynter, and Sujay Kumar Jauhar. 2025. Evaluating style-personalized text generation: Challenges and directions. *arXiv preprint arXiv:2508.06374*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Pranjal Kumar. 2024. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10):260.

Xinyu Li, Ruiyang Zhou, Zachary C Lipton, and Liu Leqi. 2024. Personalized language modeling from personalized human feedback. *arXiv preprint arXiv:2402.05133*.

Junhong Lian, Xiang Ao, Xinyu Liu, Yang Liu, and Qing He. 2025. Panoramic interests: Stylistic-content aware personalized headline generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 1109–1112.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o-mini model documentation. https://platform.openai.com/docs/models#gpt-4o-mini. Accessed: 2025-02-13.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2025. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8662–8685.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. 2025. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. *arXiv preprint arXiv:2503.02450*.

Han Ren, Xiaona Chang, and Xia Li. 2025. Neural headline generation: A comprehensive survey. *Neurocomputing*, page 129633.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. *Okapi at TREC-3*. British Library Research and Development Department.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. LaMP: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, Bangkok, Thailand. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Teng Shi, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Yang Song, and Han Li. 2025. Retrieval augmented generation with collaborative filtering for personalized text generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1294–1304.

Morten Skovsgaard and Kim Andersen. 2020. Conceptualizing news avoidance: Towards a shared understanding of different causes and potential solutions. *Journalism studies*, 21(4):459–476.

Yun-Zhu Song, Yi-Syuan Chen, Lu Wang, and Hong-Han Shuai. 2023. General then personal: Decoupling and pre-training for personalized headline generation. *Transactions of the Association for Computational Linguistics*, 11:1588–1607.

Xiaoyu Tan, Leijun Cheng, Xihe Qiu, Shaojie Shi, Yuan Cheng, Wei Chu, Yinghui Xu, and Yuan Qi. 2024. Enhancing personalized headline generation via offline goal-conditioned reinforcement learning with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5762–5772.

Qwen Team et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2:3.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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.

Zhao Yang, Junhong Lian, and Xiang Ao. 2023. Fact-preserved personalized news headline generation. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 1493–1498. IEEE.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Appendix

A.1 Infrastructure details

All the experiments were executed on an NVIDIA A100-SXM4 GPU with 40GB memory, using the vLLM framework, with GPU memory utilization capped at 70%, enabling efficient batched inference of up to 32 sequences in parallel.

A.2 BERTScore of different RAG frameworks

In Table 7, although the difference between models is not large, we can still see that as the number of user records increases, results of all RAG models improve. Already when $k = 2$, RAG with BM25 outperforms the generation without any user records, reaching the best performance in *F1* compared to original news headlines.

A.3 The examples of Personalized generation with RAG

From Table 9 and Table 10, we observe that headlines generated by RAG-based models are more closely aligned with the original news headlines, and the model using fewer user records tend to be relatively less similar. In Table 10, RAG with BM25 generally exhibits a stronger tendency to generate headlines with more detailed information.

	BERTScore								
	F1			PRECISION			RECALL		
	BM25	dense	random	BM25	dense	random	BM25	dense	random
w/o User records	89.21			89.46			89.00		
1 User records	89.02	89.16	89.07	89.16	89.30	89.26	88.92	89.05	88.91
2 User records	89.24	89.27	89.15	89.37	89.44	89.36	89.15	89.14	88.99
3 User records	89.30	89.28	89.17	89.43	89.47	89.40	89.21	89.13	88.97
4 User records	89.31	89.31	89.19	89.43	89.49	89.42	*89.29*	89.16	89.00
5 User records	89.35	89.32	89.21	89.46	89.49	89.43	89.29	89.18	89.02
6 User records	89.37	89.33	89.22	*89.47*	89.50	89.42	*89.30*	89.21	89.04
7 User records	89.37	89.33	89.21	89.47	89.49	89.41	89.32	89.22	89.04
8 User records	**89.38**	89.35	89.26	89.46	**89.50**	89.44	*89.33*	89.25	89.10
9 User records	89.38	89.35	89.24	89.45	89.49	89.43	**89.35**	89.25	89.10
10 User records	89.37	89.36	89.25	89.44	89.50	89.42	89.34	89.26	89.11
11 User records	89.38	89.36	89.26	89.45	89.50	89.44	89.35	89.26	89.12

Table 7: BERTScore of different numbers of RAG frameworks compared to original news headlines. Bold numbers indicate the best performance. The best results in F1 and Recall are generated by Llama 3 combined with user records retrieved by the BM25 retriever. The best results in Precision are generated combined with user records retrieved by the dense retriever.

A.4 The construction of LLM-as-a-Judge

The prompt used in both Mistral-7B-Instruct-v0.3 (AI, 2025) and GPT-4o-mini (OpenAI, 2024) to construct the LLM-as-a-Judge is shown in the Table 11. During the evaluation, we use headlines from the past 50 records of the current user as reference texts. For each evaluation instance, the rewritten headline, original headline, and generated headline are randomly shuffled and assigned to labels A, B, and C before being presented to the judge model, preventing the LLM from exhibiting order dependencies during evaluation.

	BERTscore								
	F1			RECISION			RECALL		
	bm25	dense	random	bm25	dense	random	bm25	dense	random
w/o User records	87.08			86.99			87.20		
1 User records	86.90	86.97	87.01	86.75	86.79	86.89	87.09	87.20	87.16
2 User records	87.08	87.11	87.08	86.92	86.97	86.99	87.28	87.28	87.22
3 User records	87.12	87.14	87.11	86.96	87.01	**87.03**	87.32	87.30	87.23
4 User records	87.12	87.15	87.11	86.94	87.02	87.02	87.33	87.32	87.24
5 User records	87.13	87.16	87.12	86.94	87.02	87.03	87.36	87.33	87.25
6 User records	87.14	87.15	87.12	86.95	87.01	87.02	87.38	87.34	87.27
7 User records	87.14	87.16	87.11	86.95	87.01	87.00	87.38	87.35	87.26
8 User records	87.15	87.16	87.13	86.95	87.01	87.00	**87.39**	87.36	87.29
9 User records	87.15	87.16	87.12	86.94	87.00	86.99	87.39	87.36	89.30
10 User records	87.14	**87.17**	87.13	86.93	87.01	86.99	87.39	87.37	89.30
11 User records	87.14	87.16	87.14	86.94	87.00	87.01	87.39	87.37	87.32

Table 8: BERTScore of different numbers of RAG frameworks compared to rewritten news headlines. Bold numbers indicate the best performance in each standard. The best result in F1 is generated by Llama 3 combined with user records retrieved by the dense retriever, the best result in Precision is generated combined with random retrieved user records, and the best result in Recall are generated combined with user records retrieved by BM25.

News index: 416	Generated news headline	RAG(BM25,K=8)	**More Than 50 Tips to Help You Lose Weight and Keep It Off**
		RAG(BM25,K=5)	56 Weight Loss Tips **to Help You Lose Weight and Keep It Off**
	Original news headline		**More Than 50 Tips to Help You Lose Weight and Keep It Off**
	Rewritten news headline		50 Detailed Tips To Help You Melt Those Stubborn Pounds
News index: 444	Generated news headline	RAG(BM25,K=8)	Foods to **Cut Out on the Keto Diet**
		RAG(BM25,K=5)	Foods to Avoid **on the Keto Diet**
	Original news headline		These Are the Foods You Absolutely Must **Cut Out** While **on the Keto Diet**
	Rewritten news headline		The Foods **The Keto Diet** Absolutly Does Not Allow
News index: 658	Generated news headline	RAG(BM25,K=8)	**Toronto Maple Leafs Sign Alexander Kerfoot and Cody Ceci** to New Contracts
		RAG(BM25,K=5)	Maple Leafs **Sign Alexander Kerfoot and Cody Ceci** to New Contracts
	Original news headline		**Toronto Maple Leafs to sign Alexander Kerfoot, Cody Ceci**
	Rewritten news headline		Maple Leafs Announces New Deals

Table 9: Additional examples of personalized news headline generation based on different numbers of records. Bold text indicates similar phrases between the generated headlines, original and rewritten headlines.

News index: 79	Original news headline	Kevin Durant, Kyrie Irving to Nets **'pretty much done'**?"
	Rewritten news headline	Kevin Durant and Kyrie Irving are to the Brooklyn Nets?
	RAG(BM25,K=8)	NBA Rumors: Kevin Durant and Kyrie Irving to Brooklyn Nets **"Pretty Much Done"**
	RAG(Dense,K=8)	NBA Rumors: Kevin Durant and Kyrie Irving to Join Brooklyn Nets
	RAG(Random,K=8)	NBA Rumors: Kevin Durant and Kyrie Irving to Join Brooklyn Nets
	w/o RAG	NBA Rumors: Durant and Irving Linked to Brooklyn Nets
News index: 213	Original headline	Kawhi Leonard had Lakers change meeting place so he could meet Paul George after?
	Rewritten headline	NBA Deals: Kawhi Leonard
	RAG(BM25,K=8)	Kawhi Leonard's Westlake Village Meeting with Lakers May Have Been a Ploy to Recruit Paul George
	RAG(Dense,K=8)	Kawhi Leonard May Have Outmaneuvered Lakers in Meeting Location Request
	RAG(Random,K=8)	Kawhi Leonard's Westlake Village Meeting with Lakers May Have Been a Power Play
	w/o RAG	Kawhi Leonard's Mysterious Meeting with Lakers Raises Questions
News index: 777	Original headline	**The Best New Makeup** Products to Try in 2019
	Rewritten headline	Makeup Products That You Should Definitely **Add To Your Collection**
	RAG(BM25,K=8)	**The Best New Makeup** Launches of April You Need to Know About
	RAG(Dense,K=8)	**The Best New Makeup** Launches This Month
	RAG(Random,K=8)	Beauty Team's Top Picks for April's Best New Makeup Launches
	w/o RAG	**The Best New Makeup** Launches to **Add to Your Collection** This Month

Table 10: Additional examples of personalized news headline generation based on different RAG frameworks. Bold text indicates similar phrases between the generated headlines, original, and rewritten headlines. In the last case, the model without RAG demonstrates better performance.

System	You are a helpful style judge that provides concise answers to style related questions based on the provided text. You output exactly one of A, B or C.
User	Choose exactly ONE among A, B, or C that BEST MATCHES the users' past preference. Focus on tone, register, vocabulary, punctuation, and phrasing. Ignore content/topic. Output EXACTLY one of: A, B, or C. No extra text. [USER STYLE EXAMPLES]: {reference_text} [Candidates]: A: {A} B: {B} C: {C} Your answer(A/B/C):

Table 11: The prompt used in LLM-as-a-Judge.