Impact of Style in Personalization of LLM

Balpreet Kaur
University of Massachusetts Amherst
Amherst, USA
bbalpreetkau@umass.edu

Aashnna Soni
University of Massachusetts Amherst
Amherst, USA
aashnnasoni@umass.edu

Prakriti Shetty
University of Massachusetts Amherst
Amherst, USA
psshetty@umass.edu

Abstract

As Large Language Models (LLMs) continue to evolve, personalization has become a vital focus, aiming to connect their universal capabilities with the rising demand for individualized interactions. Personalization methods often focus on either content-dependent or content-independent styles, overlooking the interplay between individual's writing style and content. To address this gap, we propose a novel approach that combines both content-independent and content-dependent style representations. By concatenating inputs from these two distinct approaches, we aim to capture a more comprehensive and nuanced profile of each user's unique linguistic style. This holistic method has the potential to significantly enhance the personalization capabilities of LLMs, enabling them to adapt more effectively to individual user preferences and communication patterns. Our research is motivated by the hypothesis that this combined approach will lead to more accurate and contextually appropriate personalized outputs, ultimately improving the quality and relevance of LLM-generated content across various applications and domains.

ACM Reference Format:
Balpreet Kaur, Aashnna Soni, and Prakriti Shetty. 2018. Impact of Style in Personalization of LLM. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Problem Statement

Recent approaches to LLM personalization include techniques such as authorship verification [8], which involves investigations into whether the nuances in an author's writing style can be strategically used to improve LLM adaptability to individual user preferences [6] for LLM personalization.

Our research aims to integrate content-dependent and content-independent style embeddings for a more comprehensive approach to LLM personalization.

To empirically assess this question, we propose the following:

Our approach starts with an innovative retrieval process using generative techniques to identify relevant documents from a user's historical activity, adapting to different user contexts for a comprehensive content gathering. To enhance retrieval, we generate multiple variations of the input text, broadening the search scope to capture diverse query expressions.

We then transform the original query and its variations into numerical embeddings, enabling a dense search across the user's profile documents to find semantically similar content. The top 'k' documents are selected to provide context for the language model, reflecting the user's linguistic preferences.

Next, we extract key passages from these documents for more focused inputs to the language model, improving title generation. Finally, we integrate content-independent style embeddings for author profiling using [8]'s model. By considering these style embeddings and ranking inputs, we combine both content-dependent and content-independent features, enriching the query expansion for personalized headline generation.

Mathematically, our approach is formalized as follows:
Given input $x \in X$ where X is defined to be of the following format:

$$X = \left\{ \begin{array}{l} \text{"id" : "—",} \\ \text{"input" : "—",} \\ \text{"profile" :} \left\{ \begin{array}{l} \text{"text" : "—",} \\ \text{"id" : "—",} \\ \ldots \end{array} \right\} \\ \ldots \end{array} \right\} \tag{1}$$

and ground-truth reference output $\tilde{y}_i \in Y$, where Y is defined to be of the following format :

$$Y = \left\{ \text{"golds" :} \left[\begin{array}{l} \text{"id" : "—",} \\ \text{"output" : "—"} \\ \ldots \end{array} \right] \right\} \tag{2}$$

Given a query passage q_p, we need to generate a headline y_i for the article and evaluate it against the corresponding reference output $\tilde{y}_i$, as indexed by id.

For a given sample (x_i, y_i) corresponding to user u, we create an input $\bar{x}_i$ by leveraging $\mathcal{R}$ to retrieve k items from the user's profile P_u, as defined by:

$$\bar{x}_i = \phi_p(x_i, \mathcal{R}(\phi_q(x_i), P_u, k)) \tag{3}$$

where ϕ_q is a query generation function, which converts the input x_i into a query q used to retrieve information from the user's profile, $\mathcal{R}(q, P_u, k)$ represents a retrieval model or technique, which takes the query q, the user's profile P_u, and retrieves k most relevant entries from the user's profile, and ϕ_p is a prompt construction function, which generates a personalized prompt for the user u by combining the input x_i with the retrieved entries [7].

2 Related Work

The growing demand for outputs that prioritize personal relevance and contextual awareness has driven significant research efforts to enhance large language models, aiming to deliver more personalized

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

and user-focused experiences. A.Salemi et al. (2023) emphasize the need for personalization in large language models in their study [7] and introduce the LaMP benchmark, a new framework for training and evaluating personalized LLMs. Through experiments involving zero-shot and fine-tuned models, the study highlights the significance of personalization across various language tasks and the effectiveness of these retrieval methods. Linguistic style plays a crucial role in language, and recent advances have focused on developing style representations using authorship verification tasks, which determine if two texts share the same author based on their style. However, AV-based representations may unintentionally capture content-specific information instead of pure style. Wegmann et al. (2022) [8] introduced a modified AV training approach that controls for content using conversation or domain labels. In [6], A. Neelakanteswara et al. (2024) used style embeddings to improve author profiling, aiming to enhance the personalization of large language models. Their results showed that style-based personalization performs slightly better than term or context matching. Creating effective zero-shot dense retrieval systems without relevance labels is challenging. L Gao et al. (2022) [2] worked on introducing Hypothetical Document Embeddings (HyDE). In HyDE approach, first, it uses an instruction-following language model to generate a hypothetical document based on a given query, capturing relevance patterns despite potentially containing inaccuracies. Next, an unsupervised contrastively learned encoder (Contriever) transforms this document into an embedding vector, which then guides the retrieval of similar real documents from the corpus. Their experiments demonstrated that HyDE outperforms the unsupervised dense retriever Contriever and performs comparably to fine-tuned models across tasks and languages.

As LLMs advance, there is a timely opportunity to address personalization challenges and explore innovative uses for LLMs in this space. In paper [1], J. Chen et al. (2024) review new LLM capabilities, potential applications, and current personalization challenges like lack of awareness of large models about private domain data and memory and comprehension in long conversations.

3 Approach

Given a query passage q_p, the goal is to generate a headline y_i and evaluate it against the reference output $\tilde{y}_i$.

The architecture consists of two parallel approaches to retrieve top k relevant articles from user profile. In the first approach, q_p is input into a large language model (LLM) to generate k query variants, which prompt the model to retrieve semantically similar articles. Using a dense retrieval model like Contriever, embeddings $e_{qi}, i \in [0, 1+k]$ are created for the query and its variants. These embeddings are indexed with FAISS to retrieve the top m documents, where m is a hyperparameter. This phase outputs a set of top m semantically similar documents.

In the second approach, we use the user profiles defined in the input dataset to extract style embeddings $s_i, i \in [0, |p|]$, where |p| indicates the number of text-id pairs defined as profiles, from [8]'s model. We then average out all the embeddings and build one mean style embedding s_p that represents style, and compute cosine similarity between all the user profiles and this mean style embedding. We then find the top-p most relevant documents in the user profile that matches the user style. We'll call these content-independent style embeddings.

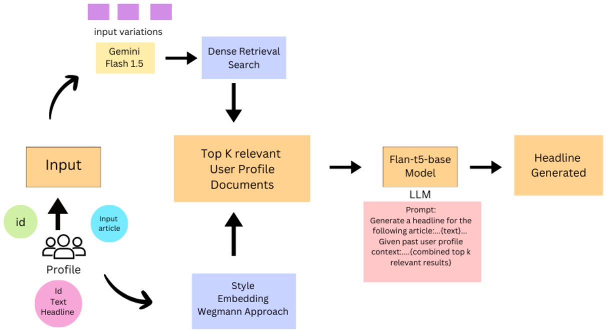
input variations
Gemini Flash 1.5
Dense Retrieval Search
Input
Top K relevant User Profile Documents
Flan-t5-base Model
LLM
Headline Generated
id
Input article
Profile
Id Text Headline
Style Embedding Wegmann Approach
Prompt: Generate a headline for the following article:....{text}... Given past user profile context:.....{combined top k relevant results}

Figure 1: Approach Architecture

To conclude, we need to integrate the content-independent and content-independent style embeddings. Recall that our outputs at this stage are top-x most relevant passages for both methods, where $x \in \{m, p\}$ for the content-independent phase and content-dependent phase respectively.

Our final step will be to choose a concatenated set of the passages retrieved from both the methods, and then append this nuanced context of updated documents and titles to the original query. This expanded query is then sent as an input to the LLM, which then generates our final output y_i, the title of the query passage.

For evaluation purposes, this y_i is benchmarked against $\tilde{y}_i$ from the reference outputs of the dataset.

4 Experiments

4.1 Dataset

We use the **LaMP-4 dataset** from the LaMP benchmark [7] for personalized news headline generation. It was chosen for its complexity, showcasing how journalists balance content fidelity with unique styles. Unlike simpler tasks, generative text tasks are more challenging, making LaMP-4 ideal for evaluating personalized language models. The dataset includes Huffington Post articles [5], where each entry has an input x (article body), a headline y, and a user profile P_u. The profile contains historical article-headline pairs (x_{u_i}, y_{u_i}), enabling models to learn author-specific styles.

4.2 Evaluation

In our study, we aim to conduct a comparative evaluation to assess the effectiveness of personalization in large language models (LLMs). We will use rouge-1 and rouge-L as the evaluation metrics. Rouge scores evaluate the quality of generated answers by comparing them to human-written reference responses [4]. ROUGE-L measures the longest common subsequence-based statistics and ROUGE-1 evaluates the overlap of unigrams. Higher rouge scores indicate a closer match to the reference, suggesting that the generated text aligns well with the expected output.

4.3 Baseline Comparison

For baseline comparison in our research, we draw inspiration from the study "RAGs to Style: Personalizing LLMs with Style Embeddings"* by Neelakanteswara et al. (2023) [6], which explores the effectiveness of style embeddings in enhancing author profiling for personalized Large Language Models (LLMs). Their approach utilizes a style-based Retrieval-Augmented Generation (RAG) method on the LaMP benchmark [7], comparing it against traditional BM25 and semantic similarity-based methods like Contriever [3] for personalized retrieval.

We have implemented two distinct RAG pipelines, each employing a unique methodology to retrieve the top-k articles from the user profile.

4.3.1 Pipeline 1: BM25-Based Retrieval. In the first approach, we use the BM25 model to evaluate the similarity between the input article, treated as the query, and the user's profile articles. BM25 scores are computed based on term frequency, inverse document frequency, and document length normalization, providing a robust measure of textual relevance. The top-k articles with the highest BM25 scores are retrieved and used as context alongside the input article in the prompt for the LLM. This pipeline leverages the efficiency and interpretability of BM25 for document ranking.

$$\text{BM25}(q, d) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, d) \cdot (k_1 + 1)}{f(q_i, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \tag{4}$$

4.3.2 Pipeline 2: Style Embedding-Based Retrieval. The second pipeline employs the model proposed by Wegmann et al. (2022) [8] to extract style embeddings from the input article. These embeddings represent distinct stylistic dimensions of the author's writing. We compute the average of these embeddings across the user profile articles to encapsulate the overall stylistic tendencies of the author. This averaged embedding serves as a robust representation of the user's linguistic style, as the individual dimensions inherently signify distinct stylistic attributes.

Let $t_i(1), t_i(2), \ldots, t_i(N)$ be articles from user i. Let

$$\vec{s}_i(t_i(1)), \vec{s}_i(t_i(2)), \ldots, \vec{s}_i(t_i(N)) \tag{5}$$

be the vector embeddings obtained from the style embedding model for each of the articles. Then, we can represent the style of the user as $\vec{S}_i$, where:

$$\vec{S}_i := \frac{1}{N}\left[\vec{s}_i(t_i(1)) + \vec{s}_i(t_i(2)) + \ldots + \vec{s}_i(t_i(N))\right] \tag{6}$$

Alternatively, this can be expressed as:

$$\vec{S}_i = \frac{1}{N}\sum_{j=1}^{N} \vec{s}_i(t_i(j)) \tag{7}$$

Next, the cosine similarity between the input's style embedding and the averaged user style embedding is calculated. Articles in the user profile are ranked in descending order of their similarity scores, ensuring that the top-k retrieved documents are those most closely aligned with the author's overall stylistic tendencies. These documents are then provided as context, along with the input article, for the LLM. By prioritizing stylistic alignment, this pipeline aims to capture the nuanced essence of the author's writing style.

In both pipelines, the top-k articles are provided as context to the LLM for headline generation, with results evaluated using ROUGE scores. The first pipeline focuses on term matching retrieval, while the second emphasizes stylistic alignment, both enhancing personalized generation in the Retrieval-Augmented Generation process. The results from both these pipelines serve as baseline benchmarks, against which we compare the outcomes of our experimental approach, that we perform next.

4.3.3 Pipeline 3: Dense Search-Based Retrieval. This pipeline is designed to enhance the retrieval process by integrating semantic understanding and efficient indexing techniques, enabling high precision in retrieving contextually relevant documents.

Generation of Hypothetical Variations: To facilitate personalized and contextually relevant retrieval, we leveraged Gemini Flash 1.5, a state-of-the-art large language model (LLM) developed by Google. This model excels in generating text that aligns semantically with the input while maintaining stylistic fidelity. Given a new article intended for title generation, Gemini Flash 1.5 was prompted to produce semantically related articles or abstracts. These generated documents act as hypothetical variations of the input, preserving the original essence while diversifying the contextual landscape. By broadening the pool of comparable documents, this approach enhanced the efficiency of retrieving top-K relevant documents, thereby enriching downstream tasks.

Contextual Representation Encoding: For encoding, we utilized *Contriever*, a dense retrieval model designed to extract highly relevant embeddings for large-scale datasets. By encoding the generated documents, Contriever captured fine-grained contextual nuances, enabling precise alignment with user-specific requirements. Its focus on contextual relevance improved the granularity and reliability of the information fed into our personalized retrieval framework.

Indexing and Dense Search Retrieval: The retrieval process was powered by FAISS, an optimized library for similarity search and vector indexing. Using Contriever-encoded embeddings, we built a tailored FAISS index for each user profile, mapping the embeddings to their respective document IDs. This index provided a robust foundation for rapid and accurate similarity-based searches.

To enhance retrieval, the embeddings of the hypothesis documents were queried against the FAISS index, ensuring quick access to documents with the highest contextual relevance. This seamless integration of FAISS and Contriever allowed for efficient dense searches that prioritized user-specific data, ultimately augmenting the personalization capabilities of our language model framework. By efficiently narrowing down relevant documents, our approach significantly improved the quality of outputs tailored to individual user profiles.

Hybrid Approach Finally, we integrate the top-k relevant articles retrieved from Pipeline 2, which utilizes style embeddings to represent the user's profile articles, with the top-k results obtained from Pipeline 3. This combined approach allows us to incorporate both content-dependent and content-independent embeddings to construct a comprehensive representation of the author's user profile. In our experimental framework, these top-k articles are subsequently provided to the large language model (LLM) as context, alongside the input article, to generate the final headline.

Sample Prompt

Generate a headline for the following article:
Your loans were made at the height of the housing bubble, and looked like a great deal at the time. By using a HELOC as a "piggyback" second mortgage, you were not required to make a down payment or to purchase mortgage insurance.
Given past user profile context:
Title: The Rise of the Housing Market
Text: During the housing bubble, many buyers relied on risky loans, leading to long-term financial issues.
Title: Mortgage Trends
Text: HELOCs gained popularity for their flexibility but posed risks when used as second mortgages.

4.4 Experimental Setup

We implemented baselines - pipeline 1, which involves a BM25 retrieval for the most semantically similar passages to our query passage, and then expanding our query with these passages before sending it to the LLM for generating the output title. The second baseline is replacing the BM25 retrieval with [8]'s style embedding model(pipeline 2) to extract context based on the past profile of the user, and using that to append to the input of the LLM.

Our novel approach combines Pipeline 2: Style Embedding-Based Retrieval and Pipeline 3: Dense Search-Based Retrieval to generate top-k relevant user profile documents for headline generation. In Pipeline 2, the style embeddings approach uses a maximum sequence length of 512, ensuring that input sentences longer than this are truncated. The sentence embeddings are generated using a pooling layer with a word embedding dimension of 768. Mean pooling of all token embeddings is used for sentence-level representation. These settings ensure the embeddings effectively capture stylistic characteristics. In Pipeline 3, we use the Gemini-1.5-flash API to generate text variants, with the number of variants set to 3, determined by the profile size. Increasing number of variants can enhance recall by capturing diverse semantic meanings but may increase computational cost and increase noise. The variants are encoded using contriever, and retrieves the top-3 documents via FAISS. Results from both pipelines are merged and fed into a large language model - Google Flan T5 Base to produce contextually rich headlines. This hybrid approach effectively integrates stylistic relevance in generating more accurate news headline. We chose a sample size of 100 passages due to computational constraints. The computational cost of generating embeddings and performing dense searches, especially at scale, is a challenge. Furthermore, the author style evaluation relies partially on human judgment, making this a significant limitation.

The implementation code for the project is available in the GitHub repository :- https://github.com/Balpreetkaur291/Impact-of-Style-in-Personalization-of-LLM

4.5 Results

In this section, we present the evaluation results for the different retrieval pipelines. We focus on two metrics: ROUGE-1 and ROUGE-L, which are standard measures for assessing the quality of the generated text in terms of its overlap with reference text.

Table 1 summarizes the performance of four different retrieval approaches: BM25, Style Embedding, Dense Search, and Hybrid. The results show varying degrees of effectiveness across the different pipelines, with the Hybrid approach achieving the highest performance in both metrics.

For ROUGE-1, which measures the recall of unigrams, the Hybrid model outperforms the other models, yielding a score of 0.0985. It is closely followed by the BM25 model with a score of 0.0989. The Dense Search and Style Embedding models have lower scores, indicating that they are less effective at capturing the relevant content for headline generation.

Similarly, for ROUGE-L, which takes into account the longest common subsequence (LCS) between the generated and reference headlines, the Hybrid model again demonstrates the best performance with a score of 0.0894. The BM25 model, with a score of 0.0892, also performs quite well, suggesting that it is particularly effective in maintaining the structural integrity of the headline. In contrast, the Dense Search and Style Embedding models, with scores of 0.0736 and 0.0787 respectively, exhibit relatively weaker performance.

Metric	BM25	Style Embedding	Dense Search	Hybrid
ROUGE-1	0.0989	0.0826	0.0787	0.0985
ROUGE-L	0.0892	0.0787	0.0736	0.0894

Table 1: Performance Comparison of Different Approaches on the LaMP-4 Dataset

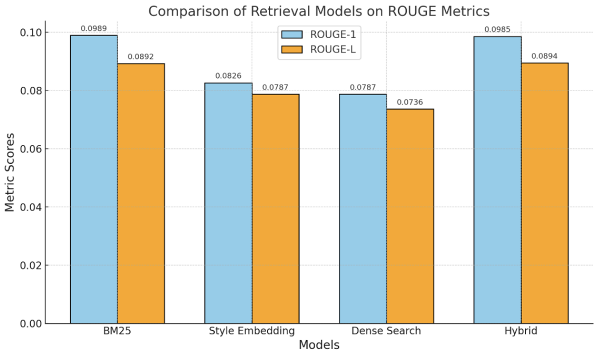

Figure 2: Comparison of retrieval models across ROUGE-1 and ROUGE-L metrics.

4.6 Analysis and Conclusion

The results demonstrate that the Hybrid retrieval pipeline, which combines the benefits of style embeddings with a dense retrieval model, outperforms the other approaches in both ROUGE-1 and

ROUGE-L. This combination allows the model to not only capture semantic relationships through the style embeddings but also maintain relevant content through the dense retrieval model. The effectiveness of the Hybrid model suggests that integrating multiple retrieval techniques can enhance headline generation tasks by providing more nuanced and contextually appropriate results.

While the Dense Search and Style Embedding-based pipelines provide useful contributions, their lower performance, especially in ROUGE-L, suggests that relying solely on either content-dependent or content-independent embeddings may limit the effectiveness of the retrieval process. The Dense Search model, despite its ability to perform semantic matching, shows lower performance in generating structurally relevant headlines, while the Style Embedding model struggles with context-specific retrieval.

The BM25-based retrieval pipeline, on the other hand, remains a strong performer, particularly for ROUGE-1, demonstrating that traditional information retrieval models are still effective in tasks involving content matching. However, the Hybrid approach, by combining the strengths of both the Dense Search and Style Embedding models, emerges as the most promising for generating high-quality, personalized headlines.

In conclusion, the Hybrid retrieval approach significantly enhances headline generation by combining content-dependent and content-independent embeddings, leading to better semantic and structural relevance. Future work could further explore the optimization of the Hybrid model, experiment with alternative large language models, and refine retrieval techniques to improve performance in similar text generation tasks.

Contributions

The contributions of the team members to this research project are outlined as follows:

- **BM25 Model-Based Retrieval Pipeline:** Prakriti Shetty
- **Style Embedding-Based Retrieval Pipeline:** Aashnna Soni
- **Dense Search-Based Retrieval Pipeline:** Balpreet Kaur
- **RAG Implementation Using a Hybrid Approach:** Aashnna Soni, Balpreet Kaur
- **Result Analysis and Parameter Tuning:** Aashnna Soni, Balpreet Kaur
- **Report Writing:** Aashnna Soni, Balpreet Kaur

5 Acknowledgments

We extend our gratitude to Professor Negin Rahimi for her support in helping us understand the foundational concepts of Information Retrieval. A special mention to our TA Harshith, who helped us with the topic selection for our project. Additionally, the assignments provided by the professor and TA were instrumental in deepening our understanding of key IR concepts, which were crucial for the successful implementation of our project.

References

[1] 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. *World Wide Web* 27, 4 (2024), 42.

[2] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2022. Precise zero-shot dense retrieval without relevance labels. *arXiv preprint arXiv:2212.10496* (2022).

[3] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised Dense Information Retrieval with Contrastive Learning. *Transactions on Machine Learning Research* (2022).

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

[5] Rishabh Misra. 2022. News category dataset. *arXiv preprint arXiv:2209.11429* (2022).

[6] Abhiman Neelakanteswara, Shreyas Chaudhari, and Hamed Zamani. 2024. RAGs to Style: Personalizing LLMs with Style Embeddings. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*. 119–123. https://aclanthology.org/2024.personalize-1.11/

[7] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization. *arXiv preprint arXiv:2304.11406* (2023).

[8] Anna Wegmann, Dong Nguyen, Dong Nguyen, Dominik Schlechtweg, and Hinrich Schütze. 2022. Same Author or Just Same Topic? Towards Content-Independent Style Representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, Spandana Gella et al. (Eds.). Association for Computational Linguistics, 249–268. https://doi.org/10.18653/v1/2022.repl4nlp-1.26