RAGs to Style: Personalizing LLMs with Style Embeddings

Abhiman Neelakanteswara **Shreyas Chaudhari** **Hamed Zamani**
University of Massachusetts Amherst
aneelakantes@umass.edu

Abstract

This paper studies the use of style embeddings to enhance author profiling for the goal of personalization of Large Language Models (LLMs). Using a style-based Retrieval-Augmented Generation (RAG) approach, we meticulously study the efficacy of style embeddings in capturing distinctive authorial nuances. The proposed method leverages this acquired knowledge to enhance the personalization capabilities of LLMs. In the assessment of this approach, we have employed the LaMP benchmark, specifically tailored for evaluating language models across diverse dimensions of personalization. The empirical observations from our investigation reveal that, in comparison to term matching or context matching, style proves to be marginally superior in the development of personalized LLMs.

1 Introduction

In the dynamic realm of Large Language Models (LLMs), achieving personalization has evolved from a mere aspiration to a vital goal (Flekova, 2020; Dudy et al., 2021). The continuous growth of LLMs emphasizes the need to customize their outputs based on individual user preferences. However, the challenge lies in bridging the gap between the inherent universal capabilities of these models and the increasing demand for personalized interactions. A relevant example from science fiction illustrates this phenomenon: in Marvel comics, JARVIS, the personal AI system in Iron Man, is crucial to Tony Stark but requires adaptation to Peter Parker's unique style for optimal service.

Addressing this challenge requires a deep understanding of individual writing styles. Consequently, our research employs style embeddings for author profiling, driven by the question: Can style embeddings capture the author profile for user-personalized retrieval, enhancing the overall personalization of LLMs? This inquiry explores whether the nuances in an author's writing style can be strategically used to improve LLM adaptability to individual user preferences.

To empirically assess this question, our study adopts the LaMP (Large Language Models Meet Personalization) benchmark (Salemi et al., 2023). This benchmark rigorously compares the effectiveness of style embeddings with traditional BM25 retrieval methods. The style embeddings used come from the model introduced by Wegmann et al. (2022), complemented by Google's Flan-T5 (Chung et al., 2022) small models for generation. Additionally, the Flan-T5 base model is used for evaluation, benchmarked against the performance outlined in the original LaMP paper.

The quantitative insights from this study aim to contribute significantly to the personalized language models landscape. By highlighting the relative importance of a user's stylistic nuances in identifying user patterns compared to the content itself, our research provides a nuanced understanding useful for refining user-specific interactions with LLMs. The systematic exploration of style embeddings and their impact on author profiling in user-personalized retrieval not only offers empirical clarity but also sets the stage for future advancements in optimizing the personalization capabilities of advanced LLMs. In an era where personalization is trending, our findings aim to inform and influence the ongoing development of language models, fostering a more sophisticated and effective era of human-computer interactions.

2 Related Work

Significant strides have been taken in the realms of authorship verification and style embeddings within the existing body of literature. In the pioneering work by Wegmann et al. (2022), a distinctive approach to authorship verification was undertaken. Their study focused on a sentence em-

Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024), pages 119–123

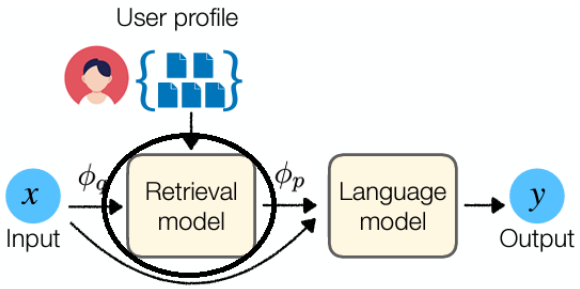

Figure 1: Architecture used to evaluate the original LaMP benchmark and the component in focus

bedding model designed explicitly to encapsulate linguistic style, distinguishing itself from conventional sentence transformer models (Reimers and Gurevych, 2019) that primarily captured textual content and semantics.

Further exploration in the domain of authorship attribution (AA) and authorship verification (AV) was conducted by Tyo et al. (2022). Their noteworthy findings revealed a compelling outcome: a conventional n-gram-based model demonstrated superior performance in five out of seven AA tasks, while BERT-based models excelled in the remaining two AA tasks and all AV tasks.

In the pursuit of effective methodologies, Coates and Bollegala (2018) demonstrated the simplicity and efficacy of employing the average of word embeddings as a "meta" embedding. Additionally, the utilization of average word embeddings in author profiling was investigated by Bayot and Gonçalves (2016), concluding that this approach outperformed tf-idf(Jones, 2021) in the context of author profiling tasks.

Synthesizing these advancements, we introduce a novel approach that employs the average style embedding as the author's identity. By identifying documents closely aligned with the average embedding, our methodology aims to intricately capture the essence of the author, thereby contributing a distinct perspective to the evolving discourse in this domain.

3 Dataset

We evaluate our methodology using the LaMP-7U, LaMP-7T, and LaMP-4U datasets[1], each carefully selected to scrutinize distinct facets of our approach. LaMP-7 encompasses a tweet paraphrasing dataset, while LaMP-4 focuses on a dataset for generating news article titles. The designations "U" and "T" signify whether the profiles are segregated among different users in the train, validation, and test sets or distributed across distinct timestamps for the same user.

The deliberate choice of LaMP-7 (Go et al., 2009) stems from its unique composition, housing a tweet paraphrasing dataset that inherently facilitates the effective capture of user style. Tweets, characterized by less filtered content compared to articles or abstracts, serve as an ideal substrate for discerning nuanced stylistic elements. Extending this rationale, LaMP-4 (Misra, 2022) also captures user style through the titles generated from news articles specific to each user. This meticulous selection of datasets not only ensures a comprehensive evaluation but also underscores the versatility of our approach in accommodating diverse textual genres and user-specific linguistic nuances.

4 Retrieval Model

This study specifically directs its attention to one constituent of the LaMP architecture—the Retrieval model, as depicted in Figure 1. Our investigation answers a fundamental query: Can style embeddings effectively capture the author profile to facilitate user-personalized retrieval?

To accomplish this, we employ the model proposed by Wegmann et al. (2022) to extract the style embeddings from the input. Subsequently, by computing the average of these embeddings to encapsulate the overall stylistic tendencies of the author, we arrange the inputs in descending order based on their cosine similarity to the average embedding. The top k results are then retrieved and employed as input for our LLM. Our architecture can be seen in Figure 2.

The use of the average embedding as a representation of user style proves to be a reasonable choice, as the dimensions within the style vector inherently signify distinct aspects of the user's linguistic style. As studied in Coates and Bollegala (2018), the simple linear average of the sentence embeddings captures the "meta" information, which in our case, captures the author's writing profile. For instance, considering our focus on LaMP-7, a user's tweeting style may encompass frequent usage of abbreviations such as "ur" or "brb." A few dimensions within the style vector quantifies the extent to which the user incorporates abbreviations, thereby contributing to an effective representation of the user's writing style. Formally,

[1]https://lamp-benchmark.github.io/download

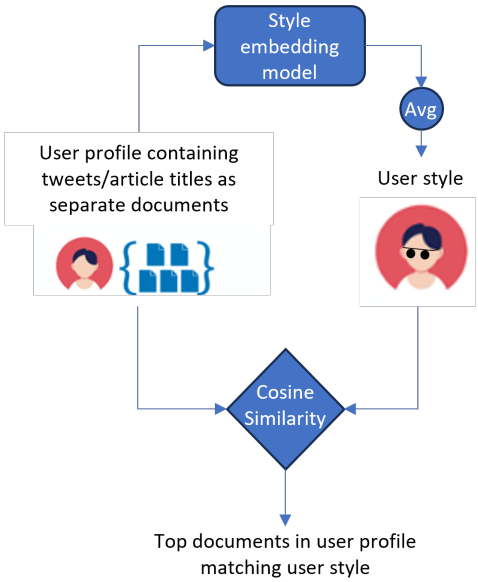

Figure 2: Style-based retrieval workflow

let $t_i(1), t_i(2), ..., t_i(N)$ be tweets from user i. Let $\vec{s_i}(t_i(1)), \vec{s_i}(t_i(2)), ..., \vec{s_i}(t_i(N))$ be the the vector embeddings obtained from the style embedding model for each of the tweets. Then, we can represent the style of the user as $\vec{S_i}$, where:

$$\vec{S_i} := \frac{1}{N}[\vec{s_i}(t_i(1)) + \vec{s_i}(t_i(1)) + ... + \vec{s_i}(t_i(1))]$$

$$\vec{S_i} := \frac{1}{N}\sum_{j=1}^{N} \vec{s_i}(t_i(j))$$

5 Empirical Analysis

5.1 Experimental Setup

We evaluate most of our results using the Flan-T5-small model by Google. The small version was chosen over the base version due to GPU resource constraints and costs. The T5 architecture (Raffel et al., 2019) has been found to be the state-of-the-art when it comes to instruction-based Text-to-Text generation tasks (Chung et al., 2022). The primary reason for choosing this family of LLMs is to compare with the metrics shown in the original LaMP paper (Salemi et al., 2023).

The models are finetuned using the same hyperparameters set in the LaMP paper. We used the AdamW (Loshchilov and Hutter, 2017) optimizer with a learning rate of 5×10^{-5}. We set 5% of the total training steps as warmup steps (Kim et al., 2021) using a linear warmup scheduler. We also set a weight decay of 10^{-4}. We set the maximum input and output lengths to 512 tokens. We have set the truncation strategy to be from left to prevent the main input, apart from the profile, being truncated. As it is a generative model, we train it for 20 epochs (Salemi et al., 2023). We also employ beam search (Freitag and Al-Onaizan, 2017) as the decoding algorithm with a beam size of 4 in all experiments. Beam search generates a sentence by evaluating the best word amongst the different beams at every step instead of choosing only one word at a time. This improves the model's ability to generate high-quality predictions The validation set is evaluated on the ROUGE metrics (Lin, 2004). k represents the number of documents retrieved for fine-tuning the generative model.

We have also used the Flan-T5-base model for two of our experiments, which is shown in the results. When compared with the results from LaMP paper, this establishes that style embeddings do indeed improve the results.

5.2 Results

Use of the average style embedding for retrieval following which, using the retrieved documents for generation shows a marked improvement over the benchmark metrics shown in the LaMP paper. Amongst the metrics used, ROUGE-1 refers to the overlap of unigrams and ROUGE-L evaluates the longest common subsequence based statistics.

5.2.1 Style-based retrieval outperforms non-personalized and BM25 retrieval

The variation of the score across the type of retrieval can be seen in Table 1. Here, we can see the average style embedding retrieval clearly outperforms the non-personalized retrieval and BM25 retrieval. This shows that the style embeddings are effective, even when working with a very small model like FlanT5-small.

The results from FlanT5-base as well as the experiment on LaMP-4U are reported in Table 2. The LaMP-4U experiments show that style embeddings performs better than non-personalized performance presented in the LaMP paper and is almost as good as random retrieval. It has to be noted that this is with a much smaller model compared to FlanT5-base.

The more interesting result is the FlanT5 results on the LaMP-7U and 7T datasets. We see a marked improvement, just using one document for retrieval with style embeddings. This shows that usage of style embeddings leads to significantly improved personalization.

Dataset	Metric	FlanT5-small (fine-tuned)		
		Non-Personalized	BM25	Style Embedding
LaMP-7U	ROUGE-1	0.488	0.504	**0.507**
	ROUGE-L	0.435	0.451	**0.454**
LaMP-7T	ROUGE-1	0.481	0.499	**0.500**
	ROUGE-L	0.427	**0.447**	**0.447**

Table 1: Impact of type of retrieval on performance $k = 1$. Style-based retrieval clearly outperforms both non-personalized retrieval and BM25 retrieval.

Model	Dataset	Metric	Best LaMP model	Style Embedding
FlanT5-base	LaMP-7U	ROUGE-1	0.526	**0.534**
		ROUGE-L	0.471	**0.475**
	LaMP-7T	ROUGE-1	0.518	**0.531**
		ROUGE-L	0.467	**0.478**
FlanT5-small	LaMP-4U	ROUGE-1	-	0.163
		ROUGE-L	-	0.149

Table 2: Results from FlanT5-base and LaMP-4U. Style-based retrieval is significantly better performing than the finetuned models used in the LaMP benchmark.

5.2.2 An increase in *k* values doesn't increase performance

We should expect that increasing k, the number of retrieved documents, should increase the performance of the models. Counter-intuitively, this was shown not to be the case in Table 3. This could be an issue of the smaller model and results may vary with larger models like FlanT5-base. The best possible explanation for this peculiar behaviour is, since we are choosing the document that most represents the author's style, any other document could confuse the model if it deviates too much from the style of the author.

Metric	$k = 1$	$k = 3$
ROUGE-1	**0.507**	0.498
ROUGE-L	**0.454**	0.446

Table 3: Impact of k on performance for dataset LaMP-7U. An increase in k does not correlate with an increase in performance.

6 Conclusion

This study shows the effectiveness of style embeddings in user personalized retrieval and personalization of LLMs. There is a significant performance increase when using average style embeddings to capture the identity of an author over both term matching retrieval like BM25 and semantic similarity retrieval like Contriever (Izacard et al., 2021). The author's style, therefore, can be more important than just the words the author uses. This is a very interesting finding. Further research has to be done on this to better represent style as a vector along with using appropriate non-linear functions for combining the vectors and utilize these embeddings for author profiling.

There is a concern for privacy of author data when it comes to the task of author profiling. But using the average embedding from a pretrained style embedding model does not require the author to share any data with the owner of the LLM. The retrieval and RAG can be performed on the user's machine without the need for powerful machinery. Hence, the scope for data leaks is very limited.

In conclusion, the integration of style embeddings for author profiling within a personalized retrieval framework, as demonstrated through our LaMP benchmark evaluations, not only showcases promising advancements in tailoring language models but also underscores the significance of considering individual writing styles for the future development of personalized, context-aware linguistic technologies.

References

Roy Bayot and Teresa Gonçalves. 2016. Multilingual author profiling using word embedding averages and svms. In *2016 10th International Conference on Software, Knowledge, Information Management & Applications (SKIMA)*, pages 382–386.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph,

Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Joshua Coates and Danushka Bollegala. 2018. Frustratingly easy meta-embedding – computing meta-embeddings by averaging source word embeddings.

Shiran Dudy, Steven Bedrick, and Bonnie Lynn Webber. 2021. Refocusing on relevance: Personalization in nlg. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2021:5190–5202.

Lucie Flekova. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Annual Meeting of the Association for Computational Linguistics*.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. In *NMT@ACL*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022.

Karen Spärck Jones. 2021. A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, 60:493–502.

Chiheon Kim, Saehoon Kim, Jongmin Kim, Donghoon Lee, and Sungwoong Kim. 2021. Automated learning rate scheduler for large-batch training. *ArXiv*, abs/2107.05855.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Rishabh Misra. 2022. News category dataset. *ArXiv*, abs/2209.11429.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, Dublin, Ireland. Association for Computational Linguistics.