Diverse Perspectives: Revolutionizing Personalization in Language Models

Chirag Totla
University of Massachusetts Amherst
ctotla@umass.edu

Venkata Samyukta Malapaka
University of Massachusetts Amherst
vmalapaka@umass.edu

ABSTRACT

This paper investigates the impact of diversity on enhancing retrieval performance in the context of personalizing input prompts for large language models. The growing utilization of language models in various applications has led to increased diversity among users in terms of their backgrounds and preferences. Consequently, there is a heightened demand for generating user-specific outputs from these models. Building upon prior work in user-profile enrichment, particularly on the LaMP benchmark, we explore the effects of augmenting the diversity of retrieved profile items. Previous research has demonstrated that enhancing diversity contributes to improved performance and robustness in retrieval models. Our findings indicate that incorporating both similarity augmentation and diversity-enhanced profiles yields superior results compared to relying solely on similarity-based methods, showcasing the efficacy of incorporating a broader range of user information.

1 Introduction

In the era of widespread applications of Large Language Models (LLMs) in daily tasks, such as Bard for Google Search and ChatGPT for question answering, there is a growing emphasis on tailoring outputs to individual users. This focus on personalization aims to enhance user experience by accommodating diverse preferences. While existing research primarily concentrates on fine-tuning LLMs for specific tasks, there exists an untapped opportunity to delve into customizing LLMs for personalized user interactions (Bender et al., 2021).

Two principal approaches are prevalent in personalizing input for LLMs. The first involves training a dedicated LLM for each user, incorporating an exhaustive user profile. However, this method is often impractical due to the significant computational resources required for individual LLM training. The second approach, the focal point of our investigation, involves prompting the model by integrating elements from the user profile. This method aims to personalize the input to the model, resulting in unique and tailored outputs.

Diversity stands as a focal and extensively investigated subject within the realm of Information Retrieval, crucial for crafting robust models that can effectively cater to a diverse user base. The introduction of diversity in retrieved results has demonstrated its efficacy, especially when addressing a heterogeneous user population querying the same information. Users with distinct preferences might focus on varied aspects of a query, and incorporating diversity in the results aids in encompassing a broader spectrum of interests. Recognizing that an individual user's preferences extend beyond the topmost similar items to a query, the enhancement of diversity, in conjunction with similarity matching, becomes essential for a comprehensive understanding of user profiles.

The LaMP benchmark offers an extensive array of data points for each user profile. In an ideal scenario, feeding the complete profile to the model would yield optimal results. However, due to constraints in computing power and input size, the challenge lies in supplying the network with the most relevant data points to facilitate an optimal understanding of the user. The dilemma arises when selecting the top k-similar data points, as this might introduce redundant information into the user profile, potentially biasing the results toward specific aspects. Diversifying profile elements serves as a mechanism for providing negative feedback to the model, fortifying it against elements contradictory to the user's profile or those that represent mere noise. Additionally, this approach aids the model in generating optimal responses in instances where useful information from the user profile is absent.

In this paper, we introduce two distinct methodologies for integrating diversified profile elements into the input prompt. Our focus is on two classification tasks sourced from the LaMP benchmark: LaMP-1U, addressing Personalized Citation Identification, a binary classification task where the model must determine the paper a user is most likely to cite based on the paper topic and the user's other publications. The second task, LaMP-2U, revolves around Personalized News Categorization. In this task, the model is tasked with classifying articles written by a journalist, considering the article text and the author's previous articles.

KEYWORDS

LLM Personalization, Diversity Enhancement, Information Retrieval, Classification Task

2 Related Work

Fine-tuning is a method of additional training to make a model more aware of our dataset, so that we can leverage it to perform tasks based on our data, like answering specific questions about data that only we might care about or want to focus on. The LLM itself might not know anything internally about the set of data we are querying it about, but this information can be fed into the

LLM by the user himself. One such way is by combining user history behaviors with LLMs, personalized classifications can be done in recommender/classification/text-generation systems, improving the user experience.

2.1 Personalization Aware LLMs for Recommendation

The paper proposes a framework called PALR that combines user history behaviors with large language models (LLMs) for recommendation. User/item interactions are used for candidate retrieval, and a LLM-based ranking model is used to generate recommended items. The authors fine-tune a 7 billion parameters LLM for the ranking purpose, which takes retrieval candidates in natural language format as input and uses explicit instructions for result selection during inference. The experimental results demonstrate that the PALR framework outperforms state-of-the-art models in various sequential recommendation tasks. It takes retrieval candidates in natural language format as input and uses explicit instructions to select results during inference.

2.2 The Alkymi platform

If one wants to take advantage of generative AI in THEIR workflows without the time and resources required for fine-tuning, an alternative that verticalized platforms like Alkymi can offer is a process called Retrieval Augmented Generation (RAG). Instead of changing the underlying model with fine-tuning, RAG is a method where you provide the LLM with relevant information it can use to find the answer to a question, at the moment the user asks the question. Using semantic search and vector databases, the Alkymi platform can quickly identify this type of relevant information from a document or dataset and surface it to the model. Your questions are then augmented with your domain-specific information, without requiring fine-tuning the model itself.

2.3 LightOn's Paradigm

Given a sufficient amount of data and the adequate hardware, the LLM can be fine-tuned on a company's know-how. LLMs have the potential to revolutionize the way businesses interact with their customers, by creating personalized experiences for improved customer satisfaction and loyalty. While there are well-identified challenges associated with using LLMs, techniques such as few-shot learning and connecting LLMs with embeddings can help mitigate these challenges, enabling businesses to create more personalized and efficient customer experiences. By leveraging these techniques, businesses can improve customer satisfaction and loyalty, while also reducing the workload on customer support teams. All of the capabilities mentioned above are made possible by LightOn's Paradigm, an AI platform that elevates natural language processing and machine learning to the next level thanks to cutting-edge LLMs. With LightOn Mini, businesses can create personalized customer experiences by analyzing large amounts of customer data and generating insights that help tailor offerings to individual customers.

2.4 Information retrieval journal - The impact of result diversification on search behavior and performance by David Maxwell, Leif Azzopardi, Yashar Moshfeghi[4]

The primary **research question** of this study is: how does diversification affect the search performance and search behavior of people when performing ad-hoc topic and aspectual retrieval tasks?

Based on Information Foraging Theory (IFT), they inferred two hypotheses regarding search behaviors due to diversification, namely that (i) it will lead to searchers examining fewer documents per query, and (ii) it will also mean searchers will issue more queries overall. To this end, they performed a within-subjects user study using the TREC AQUAINT collection with 51 participants, examining the differences in search performance and behavior when using (i) a non-diversified system (BM25) versus (ii) a diversified system (BM25 + xQuAD) when the search task is either (a) ad-hoc or (b) aspectual. Our results show That when using the diversified system, participants were more successful in marking relevant documents, and obtained a greater awareness of the topics (i.e., identified relevant documents containing more novel aspects). These findings show that search behavior is influenced by diversification and task complexity. Furthermore, they also hypothesize that diversification will lead to a greater awareness of the topic, regardless of the task. Therefore, we expect searchers to encounter and find a greater variety of aspects when using the diversified system.

3 Methodology

Before diving deeper, we are formulating the problem mathematically, which is the similar to the one stated in the LaMP benchmark wherein for every sample (x_i, y_i) associated to a user u, we have three different components: a query generation function Φ_q for retrieving user u's profile. (2) a retrieval model that takes the query and retrieves k most relevant profile entries and (3) a prompt construction that assembles a prompt for user based on input x_i and the retrieved entries. We use the same functions for both the tasks except the retrieval one which is the focus of the

We have described two methods for diversity enhancement specific to each task as the profile is different for the tasks and needs to handle in a different manner as described further.

3.1 LaMP-1U: Personalized Citation Identification

In this task, the model is tasked with classifying the relevance of one paper to another based on their titles, with the user profile comprising of other papers by the same author. To diversify the data points integrated into the user profile for prompt generation, we propose a two-step filtering and augmentation function. Initially, we employ BM-25 to retrieve the top-k relevant results from the user profile based on the input titles. Subsequently, for each retrieved profile item, we employ BM-25 using their abstracts

as queries in the second step, excluding any documents already present in the fine-tuned profile.

The rationale behind this method is to incorporate a comprehensive range of information encompassed within research articles. Often, even if an article isn't directly related to the current topic, it may appear as a citation in the past work section or offer an alternative method to address the same issue, or even present a direct contradiction. To comprehend these varied facets, it's imperative to expose the model to a broader array of articles beyond the most similar ones. This approach enhances the model's resilience to negative feedback, ultimately contributing to improved decision-making capabilities. We insert similar documents to the ones retrieved in the first step rather then incorporating the same number of random elements as to not stray away too much from the original query in the need of diversification.

3.2 LaMP-2T: Personalized News Categorization

The proposed method for tackling the LaMP-2T: Personalized News Categorization dataset involves employing the BM25 algorithm to select the most relevant article for each news category within a given profile. In cases where there are multiple articles per category, this approach aims to identify and include the most pertinent one. The selected articles are then incorporated into the input prompt when presenting a new article for classification. The rationale behind this method lies in the effectiveness of BM25 in information retrieval scenarios, where it can accurately assess the relevance of documents based on their content. Leveraging BM25 for personalized news categorization introduces a dynamic and context-aware element to the classification process, potentially leading to more accurate and contextually relevant categorization results. The novelty of this method lies in its integration of BM25 within a personalized news categorization context, contributing to a more tailored and precise classification approach for diverse news articles.

4 Experimental Setup

In evaluating our methodologies, we adhere to the same configuration employed in the LaMP benchmark. Specifically, we utilize the Flan-T5-base model with the AdamW optimizer, employing a learning rate of 5×10^{-5}. To initiate the training process, we allocate 5% of the total learning steps for warm-up, employing a linear schedule, and set the weight decay at 10^{-4}. The training duration spans 10 epochs. For the LaMP-1U task, we cap the maximum output tokens at 10, while for the LaMP-2T task, this limit is set to 512. Both tasks share an input token setting of 512 tokens.

Consistent with the LaMP benchmark methodology, we employ a similar approach to integrate profile elements into the input prompt. In the case of the LaMP-1U task, we append the title of each data point in the fine-tuned profile to the paper titles in the input prompt. Parameters akin to the original benchmark are employed to constrain the length of each profile entry in the input prompt. Conversely, for the LaMP-2T task, we introduce fine-tuned profile elements at the beginning of the input query, including their text and corresponding category.

Table 1: Results for LaMP-2T Task employing BM25 with k=1.5 and b=0.75 as hyper parameters.

Metric	Metric Value
Accuracy	80.038
F1 score	79.038

Table 1: Results for LaMP-1U Task. k denotes the number of retrieved results in the first step and k' denotes the number of retrieved results in the second step.

Metric	Value of k	Value of k'	Metric Value
Accuracy	5	2	59.684
Accuracy	4	2	61.834
Accuracy	3	3	62.529
Accuracy	3	2	62.520
Accuracy	2	2	61.930

5 Results

The examination of Table-2 reveals a marginal enhancement in performance compared to the untuned BM-25 from the LaMP benchmark on Task-1, albeit trailing behind the outcomes achieved by Contriver. Notably, in alignment with the original paper's findings, our results affirm that increasing the value of k positively influences performance up to a certain threshold, beyond which detrimental effects are observed. Our findings indicate that, for an equivalent number of elements in the user profile, introducing diversity contributes to the model's resilience to additional information. While our results hint at the potential for improved performance by substituting Contriver for BM-25 in our methodology, further investigation is warranted to definitively establish whether diversification yields performance gains. Nevertheless, our study underscores the value of diversification in incorporating additional information without adversely impacting performance, suggesting the potential for combining different retrieval techniques to achieve enhanced results.

Table 1 reports the performance metrics for the LaMP-2T task using the proposed BM25-based method with hyper-parameters k=1.5 and b=0.75. The achieved accuracy is 80.038%, reflecting the proportion of correctly categorized news articles. The F1 score, balancing precision and recall, stands at 79.038%. These results suggest that the BM25 algorithm, when applied to personalized news categorization, demonstrates a solid level of accuracy and effectiveness. The high accuracy indicates a strong capability to correctly assign news articles to their respective categories, while the F1 score underscores a balanced performance in terms of both precision and recall. These findings highlight the potential of the BM25 approach for context-aware news categorization, offering a promising solution for accurately classifying diverse news articles within the LaMP-2T dataset.

REFERENCES

[1] PALR: Personalization Aware LLMs for Recommendation. By Zheng Chen, Ziyan Jiang DOI: https://typeset.io/papers/palr-personalization-aware-llms-for-recommendation-9l5t5vzv

[2] The Role of Large Language Models in creating Personalized Customer Experiences. DOI:https://www.lighton.ai/blog/lighton-s-blog-4/the-role-of-large-language-models-in-creating-personalized-customer-experiences-28

[3] Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

[4] Maxwell, D., Azzopardi, L. & Moshfeghi, Y. The impact of result diversification on search behaviour and performance. *Inf Retrieval J* **22**, 422–446 (2019). https://doi.org/10.1007/s10791-019-09353-0