Personalized News Categorization

Bavitha Reddy Modugu
University of Florida
bavitharedmodugu@ufl.edu

Abhinav Kapavarapu
University of Florida
a.kapavarapu@ufl.edu

Lohitha Maradani
University of Florida
lohitha.maradani@ufl.edu

Sri Harsha Talasila
University of Florida
talasilas@ufl.edu

Abhiram Veluru
University of Florida
ab.veluru@ufl.edu

Abstract

In the realm of computer systems equipped with Natural Language Processing (NLP), personalizing the user experience has become increasingly vital. Large Language Models (LLMs), which are sophisticated and expansive computer models, are pioneering new methods to customize virtual assistants and other language-based systems to individual preferences (Alireza Salemi, 2023). This project utilizes the potential of large language models to provide more personalized responses. This approach aims to surpass existing personalization techniques, particularly by minimizing information loss and more effectively addressing scenarios where limited data is available about a new user (Chris Richardson, 2023).

1 INTRODUCTION

In the era of digital abundance, the challenge of filtering relevant information has become paramount. This paper addresses the critical issue of personalized news categorization, a field where the precision and adaptability of Large Language Models (LLMs) can greatly enhance user experience. (Asai et al., 2022) Personalized news delivery systems using LLMs leverage vast datasets to learn and predict user preferences with remarkable accuracy, presenting solutions to the overwhelming problem of information overload.(Yang et al., 2023)

The significance of this technology cannot be overstated, particularly as users demand more tailored interactions. Traditional approaches often fail to meet these demands, providing generic results that lack personal relevance. For instance, consider a scenario where a journalist specializing in 'Healthcare Technology' writes about advancements in medical imaging. Traditional models might broadly categorize this under 'Technology' or 'Healthcare.' In contrast, a personalized LLM, aware of the journalist's focus, would accurately classify it under 'Healthcare Technology,' thus delivering more pertinent results to users interested in this niche.

However, implementing LLMs for such tailored experiences involves challenges such as maintaining user privacy, ensuring algorithmic fairness, and managing the subtleties of language interpretation. This paper delves into these challenges, presenting our methodology, the development of enhanced query generation functions, and innovative retrieval strategies. Through rigorous experiments and detailed analysis, we demonstrate how LLMs can be optimized to not only meet but exceed user expectations in news categorization, thereby pushing the boundaries of current technology.

2 BACKGROUND AND RELATED WORK

The idea of customizing natural language processing (NLP) systems is increasingly being recognized as crucial for improving user experience and tailoring responses to specific user preferences (Chiang et al., 2023) (Hyung Won Chung, 2022). Historically, research in this area has largely centered on search engines and recommender systems within the fields of information retrieval and human-computer interaction. However, the incorporation of personalization features into large language models (LLMs), particularly for tasks like text classification and generation, remains relatively unexplored.

The LaMP Benchmark: Filling a notable gap, the LaMP benchmark marks a substantial advancement in this area(Alireza Salemi, 2023). Departing from the conventional 'one-size-fits-all' approach of typical NLP benchmarks, LaMP introduces tasks tailored to individual preferences, offering a more robust evaluation framework(Alireza Salemi, 2023). This benchmark includes three classification tasks that are Personalized Citation Identification, Personalized News Categorization, and Personalized Product Rating and four text generation tasks like Personalized News Headline, Scholarly Title, Email Subject, and Tweet Paraphrasing. This variety enables a comprehensive evaluation of language models across various personalization contexts.

Retrieval-Augmented Personalization: A significant innovation within the LaMP benchmark is its retrieval augmentation method. This technique pulls personalized content from user profiles to craft customized prompts for large language models, overcoming the challenges associated with the input length restrictions of these models. Benchmark results indicate that language models that integrate user profile data perform better than those that do not incorporate such personalization features.

Zero-Shot and Fine-Tuned Models: The benchmark assesses both zero-shot and fine-tuned models, showing that fine-tuned models frequently outperform the zero-shot abilities of larger models. This finding is essential for advancing the development of more effective personalized language models. The launch of LaMP paves the way for multiple new research opportunities.

3 RESEARCH QUESTIONS AND OBJECTIVES

The core of our project is defined by a set of research questions and objectives aimed at pushing the boundaries of how LLMs can be optimized for news categorization tailored to individual preferences.(Xiang Ao, 2021) Our primary objective is to refine the query generation function, which is pivotal in fetching user-specific content. Additionally, we address the need to mitigate biases that may arise during the personalization process, ensuring fairness and equity in automated news delivery.

4 METHODOLOGY

4.1 Dataset

LaMP, or Language Model Personalization, is a thorough benchmark consisting of datasets specifically designed for the development and assessment of methods that personalize language models. This benchmark includes seven unique datasets for various classification and text generation tasks. In this experiment, the 'LaMP 2: Personalized News Categorization' dataset is utilized. It features an input prompt for a language model, enhanced by a detailed profile section. This section contains numerous articles, each defined by its text, title, and assigned category.


```
    "profile": [
      {
        "text": "The three make a trip of atypical opera themes, but no new opera brought the Met as much controversy as Klinghoffer.",
        "title": "Klinghoffer",
        "category": "culture & arts",
        "id": "1000"
      },
      {
        "text": "Though I might not subscribe to every prayer in the Siddur, I always use the synagogue time for my own prayer of thanks f
or being alive and the multitudes of blessings I enjoy.  I want to let the Lord know I haven't forgotten them.  Then, leaving the hall, th
e yarmulke still in place on my head, I head home feeling a little purer.",
        "title": "Alone in the Synagogue",
        "category": "religion",
        "id": "1001"
      },
      {
        "text": "I expected him to do well and show me some of his work. The money could be renewed for the second semester and the next y
ear of college. Though I imagined that Mel would want to maintain periodic contact -- at least it was what I hoped -- I was wrong.",
        "title": "Help Is on the Way",
        "category": "religion",
        "id": "1002"
      },
```


Figure 1: User Profile Example

4.2 Experimental Setup

In this study, the BM25 retrieval algorithm was chosen due to its impressive track record in information retrieval tasks. For the experiments, the Flan-T5-base language model was selected, which supports up to 512 tokens. Personalization involves using a retriever function to fetch the top-k articles from a profile section. This retriever is supplied with the output from the Query Generation Function ($\phi(q)$) to collect the top-k documents. Following this, the top-k documents are used to construct the input prompt for the LLM via the Prompt Generation Function ($\phi(p)$), and the Flan-T5-base LLM then processes this prompt.

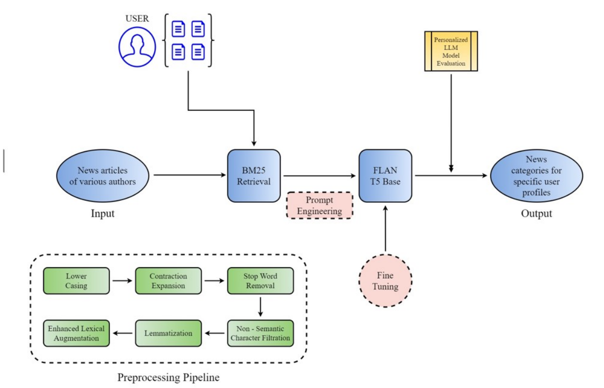

Figure 2: Experiment Procedure

4.3 Query Generation Function ($\phi(q)$)

In the process of tailoring Large Language Models (LLMs) for improved query generation, an essential phase is the preprocessing of textual data. This pipeline is structured to modify and tune the input articles (queries) to enhance the efficiency of the BM25 retrieval algorithm. The subsequent sections outline each stage of the preprocessing routine:

Lowercasing - Each character in the article was converted to lowercase to avoid any discrepancies due to case sensitivity during the retrieval process.

Contraction Expansion - Common contractions such as "don't" were expanded to "do not" to make the text clearer and more interpretable. This expansion helped enhance the semantic understanding of the text.

Stop Word Removal - Words like "the", "is", and "in", which occur often and provide minimal informational value, were eliminated from the text. This resulted in a more concise and pertinent set of terms for the retrieval process.

Non-Semantic Character Filtration - Characters lacking semantic importance, such as special symbols and extra spaces, were removed.

Lemmatization - Words were lemmatized based on their part of speech, turning forms like "running" into "run".

Enhanced Lexical Augmentation - Synonym Enrichment Technique - In this advanced step, the preprocessed text is subjected to a synonym enrichment process. Every token in the article is matched with an appropriate synonym, thereby expanding the diversity and scope of the query terms. This approach is designed to cover a broader spectrum of relevant documents from the user profile, embracing a wide array of lexical variations. This strategic expansion aims to enrich the context and breadth of the query, enhancing the retrieval process to be more comprehensive and effective.

4.4 Retrieval and Prompt Generation Function ($\phi(p)$)

The BM25 retriever was utilized across all 13 settings to retrieve the top 'k' documents from the 'profile' section for each news article, with experiments conducted using various values of 'k'- specifically 1, 2, and 3. Enhanced Lexical Augmentation was applied to the articles, leading to a noticeable improvement in BM25 scores for certain profile articles.

The method for generating prompts in this experiment aligns with the procedures established in the LaMP benchmark. The format for the prompt in the LaMP-2: Personalized News Categorization Task is structured as follows:

Sample Article 1 -

Scores without Enhanced Lexical Augmentation -

```
[array([0.        , 0.        , 4.78241772, 0.        , 2.39120886,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 3.35023668,
        2.65139221, 0.        , 0.        , 0.        , 0.        ,
        1.6318087 , 0.        , 0.        , 1.5891885 , 0.        ,
        0.        , 0.        , 4.57092271, 5.97501645, 7.799573  ,
        2.39145184, 0.        , 0.        ])]
```

Scores with Enhanced Lexical Augmentation -

```
[array([0.        , 0.        , 4.78241772, 0.        , 5.78841147,
        0.        , 0.        , 0.        , 0.        , 0.        ,
        0.        , 0.        , 0.        , 0.        , 3.35023668,
        2.65139221, 0.        , 0.        , 0.        , 0.        ,
        1.6318087 , 0.        , 0.        , 1.5891885 , 0.        ,
        0.        , 0.        , 4.57092271, 5.97501645, 7.799573  ,
        4.78290368, 0.        , 0.        ])]
```

Sample Article 2 -

Scores without Enhanced Lexical Augmentation -

```
[array([2.39518005, 4.81358388, 2.7630574 , 5.008226  , 2.7630574 ,
        2.05052647, 0.        , 0.        , 2.504113  , 0.        ,
        2.43042248, 0.        , 0.        , 0.        , 2.32767498,
        2.62342662, 0.        , 0.        , 0.        , 0.        ,
        0.        , 2.78409674, 5.59638553, 1.72031541, 1.41091647,
        1.87445871, 1.52678144, 0.        , 1.44336514, 0.        ,
        2.34757186, 1.21907243, 0.        ])]
```

Scores with Enhanced Lexical Augmentation -

```
[array([4.7903601 , 4.81358388, 2.7630574 , 7.512339  , 6.11726411,
        2.05052647, 0.        , 0.        , 2.504113  , 0.        ,
        2.43042248, 0.        , 0.        , 0.        , 2.32767498,
        2.62342662, 0.        , 0.        , 0.        , 0.        ,
        0.        , 2.78409674, 7.23241828, 1.72031541, 1.41091647,
        1.87445871, 1.52678144, 0.        , 1.44336514, 0.        ,
        2.34757186, 1.21907243, 0.        ])]
```

Table 1: Performance improvement in BM25 retrieval scores when Enhanced Lexical Augmentation is used.

Per Profile Entry Prompt (PPEP) - The category for the article is indicated by: "Pi[text]" is "Pi[category]".

Aggregated Input Prompt (AIP) - This is created by concatenating the prompts: concat([PPEP(P1), ..., PPEP(Pn)], ", and "). [INPUT], where 'concat' is a function used to merge strings from its first parameter by inserting the string from its second parameter between them.

The PPEP function generates a prompt for each entry retrieved from the profile, and [INPUT] represents the task's input.

4.5 Language Model

In this research, we employed the Flan-T5-base language model in two distinct configurations: one version was fine-tuned, while the other was used as-is, without any modifications. The fine-tuning involved adjusting the model with a learning rate of 3e-4 over three training epochs, which significantly improved its performance. The FlanT5-base model was chosen for its efficient runtime. This efficiency was crucial, particularly as it supported the strategy of maximizing user data inclusion within the 512-token input length limit. Additionally, the selection of the FlanT5-base was influenced by its demonstrated superiority in the LaMP benchmark experiments, where it outperformed larger models such as FlanT5-XXL and ChatGPT in zero-shot scenarios. The enhanced performance of the FlanT5-base model, especially in its fine-tuned state, was a key factor in its selection for the experiment.

5 EVALUATION AND RESULTS

Our evaluation of the LLM's performance in personalized news categorization revealed promising results. Using a set of metrics such as accuracy, precision, recall, and F1 score, we assessed the model across various configurations and parameter settings. Initial results indicate a marked improvement in the accuracy and personalization of news categorization, with the model demonstrating an enhanced ability to align news content with individual user profiles effectively. These results provide a quantitative foundation supporting the effectiveness of our query generation and retrieval strategies.

K' - Documents Retrieved	Enhanced Lexical Augmentation	Accuracy	F1 Score
1	No	0.422	0.521
	Yes	0.425	0.526
2	No	0.381	0.476
	Yes	0.375	0.471
3	No	0.351	0.448
	Yes	0.353	0.449

Table 2: Performance Metrics without LLM Fine-Tuning

6 ANALYSIS AND CONCLUSION

6.1 Limitations

The study highlights a few limitations affecting the effectiveness of personalized language models (LLMs). Firstly, there is a potential bias towards certain user profiles in the training data, which could hinder the model's ability to personalize effectively for a diverse range of users. This bias

K' - Documents Retrieved	Enhanced Lexical Augmentation	Accuracy	F1 Score
1	No	0.866	0.867
	Yes	0.868	0.869
2	No	0.870	0.871
	Yes	0.872	0.873
3	No	0.875	0.876
	Yes	0.872	0.873

Table 3: Performance Metrics with LLM Fine-Tuning

is exacerbated by the dependence on the availability and diversity of user data; when data is limited or lacks variety, the model's capacity for effective personalization is significantly weakened. Furthermore, the resource-intensive nature of fine-tuning LLMs and implementing lexical augmentation poses challenges for scalability and practicality, particularly in environments with limited computational resources. These advanced techniques, while enhancing personalization, may not be feasible in resource-constrained settings. Additionally, the fine-tuning process itself risks overfitting, where the model becomes overly specialized to the training data, reducing its adaptability and accuracy for user profiles or scenarios not covered during training. Such limitations suggest that while personalized LLMs have potential, their deployment must be carefully managed to ensure they are effective and equitable across all user demographics.

6.2 Conclusion

This project improved the query generation function $\phi(q)$ in Large Language Models (LLMs) for personalized news categorization, making the queries more effective and aligning news content more closely with user preferences for a customized news experience. This enhancement also mitigated bias in AI-driven news sorting, advancing fairness and ethical AI practices. The study demonstrated the capacity of LLMs to deliver precise, relevant, and unbiased digital experiences. Nonetheless, several challenges were recognized, including limited diversity in user profiles, reliance on extensive user data, scalability issues, and privacy concerns. These issues highlight the necessity for the continuous development of responsible and efficient AI systems. Future initiatives will concentrate on advancing AI personalization while ensuring fairness, privacy, and a focus on user needs.

7 FUTURE SCOPE

The future scope of this project on personalized news categorization using Large Language Models (LLMs) involves several promising directions. To address the limitations identified, such as bias and dependency on extensive user data, efforts will concentrate on diversifying the training datasets to ensure broader representation across various demographics. This will help mitigate bias and enhance the model's effectiveness across a wider array of user preferences. Additionally, advancements in algorithm efficiency and model compression techniques are essential to make LLMs more scalable and less resource-intensive, enabling their deployment even in resource-constrained environments. Privacy-enhancing technologies will also be a focus, ensuring that personalization does not compromise user confidentiality. Furthermore, ongoing research will explore the integration of ethical AI practices to prevent unfair biases and promote transparency in AI-driven systems. These efforts aim to refine AI personalization technologies, making them more equitable, efficient, and respectful of user privacy and diversity.

References

Michael Bendersky Hamed Zamani Alireza Salemi, Sheshera Mysore. 2023. *LaMP: When Large Language Models Meet Personalization*. arXivLabs.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen tau Yih. 2022. Task-aware retrieval with instructions.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. LMSYS ORG.

Kellen Gillespie Sudipta Kar Arshdeep Singh Zeynab Raeesy Omar Zia Khan Abhinav Sethy Chris Richardson, Yao Zhang. 2023. *Integrating Summarization and Retrieval for Enhanced Personalization via Large Language Models*. arXivLabs.

Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter sentiment classification using distant supervision*. researchgate.

Shayne Longpre Barret Zoph Yi Tay William Fedus Yunxuan Li Xuezhi Wang Mostafa Dehghani Siddhartha Brahma et al Hyung Won Chung, Le Hou. 2022. *Scaling Instruction-Finetuned Language Models*. arXivLabs.

Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. *GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation*. arXivLabs.

Ling Luo Ying Qiao Qing He Xing Xie Xiang Ao, Xiting Wang. 2021. *"PENS: A Dataset and Generic Framework for Personalized News Headline Generation"*. Association for Computational Linguistics.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. 2023. Palr: Personalization aware llms for recommendation.