Understanding the Impact of RM3 Based Query Expansion in Personalizing Large Language Models

Aarthi Nunna*
anunna@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

Spurthi Tallam*
stallam@umass.edu
University of Massachusetts Amherst
Amherst, Massachusetts, USA

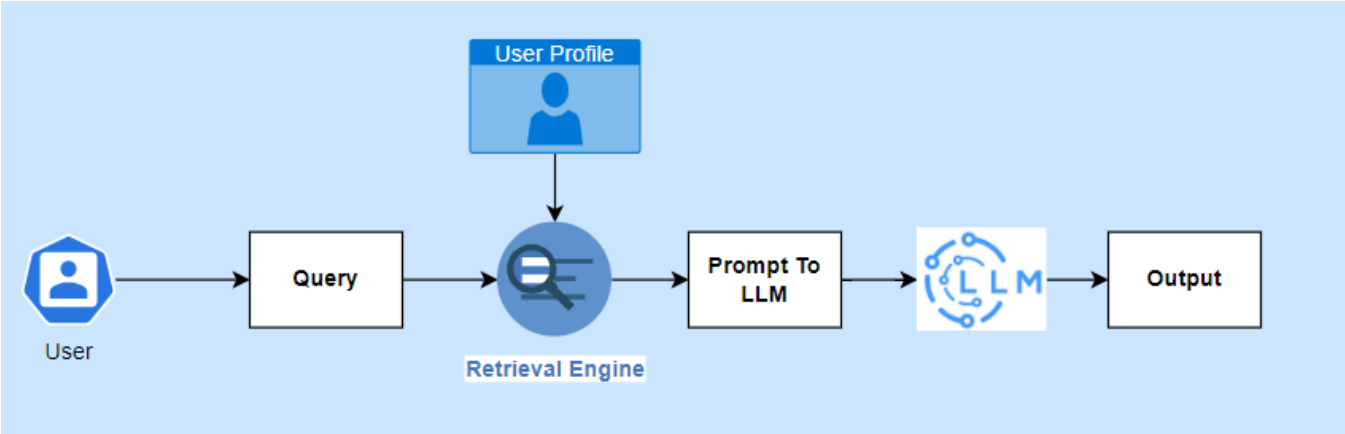

Figure 1. Personalization of Large Language Models

Abstract

Personalization of large language models is a growing field of research in the information retrieval community. With numerous applications across domains using retrieval to provide recommendations, filter out content, search engines, etc., there arises a need to personalize the output of LLMs to better cater to the users' wishes, aside from the ever-increasing need to improve the efficiency and accuracy of the retrieval process. State-of-the-art models like Okapi BM25, KL-Divergence, BERT, etc., view information in different lights, varying from simple tokens to complex word embeddings incorporating contextual information. Query expansion is one such technique to improve the retrieval process by expanding the query fed to a retrieval algorithm through iterations of identifying relevant terms that aid the model in better understanding a user query and narrowing down target concepts to search for. Our paper aims to observe and reason the impact of query expansion in the context of personalizing large language models.

CCS Concepts: • **Information systems** → **Query representation**; **Personalization**; **Language models**; **Clustering and classification**;

Keywords: Large Language Models, Personalization, Query Expansion, Relevance Models, BM25

ACM Reference Format:
Aarthi Nunna and Spurthi Tallam. 2023. Understanding the Impact of RM3 Based Query Expansion in Personalizing Large Language Models. In *Proceedings of (646 Final Project).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.

646 Final Project, Dec 12 , 2023, Amherst, MA

ACM ISBN 978-1-4503-2023-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The methodologies to retrieve information have transformed drastically over time. With a booming increase in computational resources and advancements in Natural Language Processing, Artificial Intelligence, Generative AI, etc., the research communities have modeled ideas ranging from sparse tokens and vectors to contextual and semantic embeddings. Large language models are fruitful products of deep learning techniques, particularly neural network architectures comprising millions of parameters and large datasets made available with a surplus of data that roams in the network today.

The transformer architecture proposed by the authors in [18] has revolutionized the Natural Language Processing industry by achieving state-of-the-art results in machine translation, summarization, query answering, and other such niche tasks by demonstrating superior parallelization capabilities and improved performance on long-range dependencies. The advent of LLMs has seen the rise of powerful models such as Generative Pre-Trained Transformers, popularly referred to as GPT, offering a spectrum of use cases ranging from query-answering to data analysis [15]. These models throw light on the growing need for personalization as the end-users often seek output tailored to their needs.

LLMs have transformed traditional information retrieval through passive filtering to that based on active user engagement, allowing for the proactive exploration of user data and requests that fine-tune the outputs of LLMs to deliver genuine, interactive, and explainable information. They have significantly broadened the breadth of personalization from naively gathering personalized information to building personalized complex services [5].While the advantages and indispensable need for personalization are realized, there is a lack of study in evaluation methodologies for these models that are essential for analysis and production. Thus, LaMP provides an all-encompassing framework for assessment that integrates a variety of language tasks necessitating personalization [17].

LaMP focuses on multiple classification and text generation tasks comprising user-based and time-based data to understand user interactions and predict future answers in a manner personalized to suit the user. LaMP proposes to append numerous user profiles to the user query to generate a retrieval-augmented query that assists the LLM in the generation of personalized outputs. However, there is the underlying problem of identifying relevant user profile data from all the available data points of a user's profile as they would not all contribute positively to helping the LLM learn the user needs. Thus, the research problem is divided into two parts, the first concentrating on filtering relevant data points from a user profile and augmented prompt generation, and the second focusing on fine-tuning the LLM.

Our paper dives deeper into the first part of the research problem tackled by LaMP, the retrieval-augmented generation. We primarily devote this paper to understanding the influence of query expansion and relevance models in RAG (retrieval-augmented generation) and its corresponding impact on the LLM's output. Relevance models have a foundation built on language models and the probability of word occurrences [9]. The probability distributions computed through relevance models are plugged into query expansion formulae to perform query reformulation. Thus, our experiments play around with the parameters pertaining to query expansion and relevance models, while establishing a baseline through the use of state-of-the-art models such as BM25 to understand the impact of the same.

2 Related Work

Query expansion can be performed using many models and algorithms whose choice depends on the query itself, the output conditions, the corpus, computation and resource requirements, etc. For instance, [7] proposes to utilize the generative capabilities of LLMs to expand the query. Their methodology offers greater weight to the initial query by mentioning the query multiple times in their prompt and concatenating this with the prompts obtained from the LLM. The concatenated string is later fed to foundational models such as BM25 to obtain a rank list. Metrics such as NDCG, Recall, etc., draw comparisons between the performance of various LLMs that expand the query. On similar lines, [19] employ LLMs to generate pseudo-documents obtained through few-shot prompting the LLM with necessary input data.

An approach taken by [14] is to perform query expansion by selecting relevant terms used to expand a query in a hybrid manner. The hybrid approach is a mixture of distribution-based term selection such as Kullback-Leibler Divergence (KLD) [1], and association-based term selection such as Local Context Analysis (LCA) [20] and Relevance-based Language Model (RM3) [8]. The idea originates through initially acquiring candidate expansion through a distribution-based approach following which they undergo refinement based on the degree of association of terms with the original query terms.

Authors of [11] have conducted research specific to personalizing LLMs for improving academic prose and idea generation. They have implemented AUTOGEN ("AI Unique Tailored Output GENerator") models that are LLMs fine-tuned on the previously published scholarly work of the authors of the paper. The model is observed to have outperformed the base GPT-3 model, emphasizing the importance of LLM personalization. [4] take on another approach to query expansion by using contextual embeddings of words combined with a cluster-based model that selects terms that capture the user interests better and introduces diversity based on these words.

3 Theoretical Background

3.1 BM25

Classical probabilistic models establish their foundation on the probability ranking principle stated by Stephen E. Robertson. One such widely used model to rank documents based on their relevance to a query is BM25. The algorithm is a document-generative model that accounts for parameters such as term frequency [16], document length, and the inverse document frequency of query terms. BM25 is an improved version of the earlier TF-IDF model as it introduces term saturation to prevent overly influential terms [2]. Despite its limitations and the upcoming varying representations of text that better represent information, BM25 has its

roots settled deeply into the IR field and is still profound and is thus still used as an initial retriever for multiple models.

3.2 Relevance Feedback

The process of obtaining feedback from users regarding the relevance of search results and using the feedback to produce better ranking lists and thus a better search engine result page is called relevance feedback. The system may explicitly obtain feedback through user interaction, such as surveys, or implicitly by monitoring user interactions, such as click-data. A major use-case of relevance feedback is performing query expansion based on the feedback resulting in significant improvement of search results.

3.3 Query Expansion

Vocabulary mismatch is a fundamental problem in information retrieval as languages often can express the same or similar ideas in multiple ways. One of the ways to address this issue is query expansion. Through expanding the query, we include more terms that convey the same or similar concepts that aid the retrieval engine in retrieving documents that match the query topics but were initially unable to retrieve. Numerous techniques for query expansion such as using a global view of the corpus, external resources like WordNet, and relevance information are present. Our paper focuses on using relevance information in specific pseudo-relevance feedback to perform query expansion.

3.4 Pseudo-Relevance Feedback

Pseudo-relevance feedback [10], also known as blind feedback or local query expansion centers around the assumption that documents retrieved by a first-stage retriever are relevant to the query and can thus be used to identify relevant query terms using different formulae [3]. Pseudo-relevance feedback (PRF) is an efficient technique despite the probability of the initial assumption failing, as it prevents the need to acquire feedback through external means that are often tough or expensive to collect. Relevance models employ PRF, and our experiments use one such model.

3.5 Relevance Models

Relevance models work towards estimating a better query language model from which the query terms and candidate query terms are sampled. Our paper uses a relevance model called RM3, a linear interpolation of the maximum likelihood estimate of a term belonging to a query language, and the relevance model RM1 which works on the independent and identically distributed (IID) assumption.

$$P(W|\theta_Q) = \sum_{D \in F} P(W|\theta_D) \prod_{i=1}^{|q|} P(q_i|\theta_d) \tag{1}$$

Equation 1 refers to the final equation generated by the IID assumption of RM1.

$$P(W|\theta_Q) = \alpha P_{\text{MLE}}(w|\theta_Q) + (1-\alpha) P_{\text{RM1}}(W|\theta_Q) \tag{2}$$

The retrieval model then performs query expansion using equation 2 to determine potential candidates or candidates with high scores, implying a high probability of their generation from the newly computed query language distribution.

4 Experiments

4.1 Dataset

We have used two of the classification-based datasets provided by LaMP [17], LaMP2 which consists of data for personalized new article categorization, and the LaMP3 dataset for personalized product rating.

4.1.1 LaMP2: Personalized News Categorization. The LaMP2 dataset is a modified version of the dataset provided by the authors of [12]. The dataset predicts the user's news article category from a set of pre-established categories by understanding the previously written news article categories that are present in the dataset as the user profile. As the nature of the task is classification, the evaluation metrics considered are accuracy and macro-averaged F1 score. To conduct our experiments and draw conclusions, we have used all 5914 samples present in the training set of the dataset. Each sample pertains to a single unique user and their user profile of varying lengths.

4.1.2 LaMP3: Personalized Product Rating. The LaMP3 dataset is an alteration of the dataset provided by [13]. The dataset predicts the rating a user would reward a product based on a user review, by extracting relevant information from the user's past reviews and ratings. As the nature of the task is an ordinal multi-class classification problem, the evaluation metrics considered are RMSE and MAE. To conduct our experiments and draw conclusions, we shuffle the training set of the entire dataset with a fraction of 1 and a random state of 42 and consider the first 1500 samples. We consider only a subset of the samples to carry out the research as the required resources exceeded the capacity of the experimental environment setup as explained in the following sections.

4.2 Model Architecture

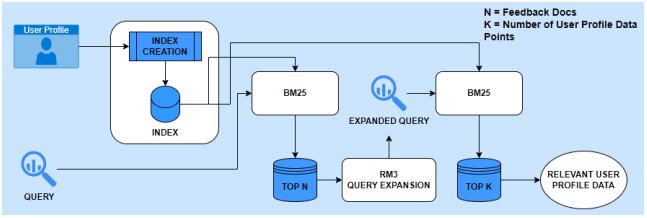

Figure 2. Stage 1

Figure 2 describes Stage 1 of personalizing LLMs. This stage focuses on retrieval augmented generation, i.e., creating prompts that are a concatenation of relevant user-profiles and the initial user query. Each sample of the dataset undergoes stage 1. The first step is the index creation of the user profiles, following which the retrieval process begins. BM25 is initially used to score the previous data of the user based on its relevance to the current query. Subsequently, the RM3 model uses N-relevant articles or reviews to expand the user query with an alpha score of 0.6. Finally, BM25 uses the expanded query to score the data again to obtain the top K-relevant user profile data points.

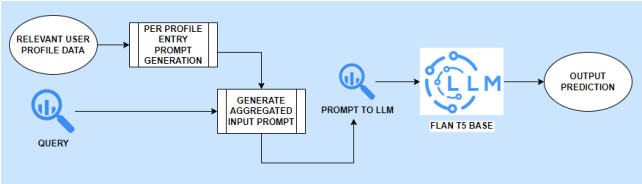

Figure 3. Stage 2

Figure 3 describes Stage 2 of personalizing LLMs. This stage focuses on obtaining the final prediction from the LLM by submitting a newly augmented prompt. Each sample of the dataset undergoes stage 2. The per-entry prompt generation function iterates through top K-relevant user profile data points which are concatenated together and the initial user query. On acquiring the newly created LLM prompt, the prompt is fed to the Flan-T5-base LLM, and its prediction is generated.

4.3 Experimental Setup

Our experimental setup comprised Google Colab with a T4 GPU (Tesla T4) runtime to conduct our research. We created multiple notebooks using Python to index the data, perform query expansion, and use the Flan-T5-base LLM. The resources provided by Colab are as follows:

- **System RAM** - 12.7 GB
- **GPU RAM** - 15.0 GB
- **Disk** - 78.2 GB

4.3.1 Libraries and Frameworks.

- **Python Terrier** - PyTerrier is a declarative platform used to perform information retrieval experiments in Python. Its internal support for indexing and retrieval operations relies on the Java-based Terrier information retrieval platform. For conducting our experiments, we use PyTerrier for its modules to index user profiles, query expansion using RM3, and pipelining the retrieval model.
- **Dask** - Dask is a library provided by Python to support parallel computing by maximizing the usage of the already present resources such as CPU. We use the Dask data frame to create the index in parallel to speed up the index creation.
- **Transformers** - The Transformers library is often used to work with natural language processing (NLP) and deep learning models, particularly for transformer-based models like BERT, GPT, and others. We also use the PyTorch dependency to optimize the utilization of the resources.
- **scikit-learn** - The scikit-learn library consists of many modules oriented towards machine learning and deep learning. We use the sklearn.metrics module that provides multiple functions for evaluating the performance of machine learning models. In particular, we use metrics like accuracy, F1 score, RMSE, and MAE.

4.3.2 Large Language Model.

- The FLAN-T5 LLM [6] is an enhanced version of the T5 LLM fine-tuned to suit a mixture of tasks. The FLAN-T5-base is a variation of the FLAN-T5 released by Google. It has a default input token size of 512 and has 250M parameters.

5 Results and Analysis

Table 1 tabularizes the data from the multiple experiments run on the LaMP2 dataset with varying parameters. The parameters varied are K - the number of relevant user profile data points added to the query, and Fb Docs - the number of feedback documents considered by RM3. The evaluation metrics as mentioned earlier are accuracy and macro-averaged F1 score.

BM25	RM3	K	Fb Docs	Fb Terms	F1	Accuracy
Yes	-	1	-	-	**0.162**	**0.462**
Yes	Yes	1	3	10	0.153	0.448
Yes	Yes	1	2	10	0.149	0.447
Yes	Yes	1	10	10	0.179	0.467
Yes	-	2	-	-	0.152	0.430
Yes	Yes	2	3	10	0.153	0.428
Yes	Yes	2	2	10	0.151	0.429
Yes	Yes	2	10	10	0.158	0.434
Yes	-	4	-	-	0.120	0.404
Yes	Yes	4	3	10	0.111	0.402
Yes	Yes	4	2	10	0.116	0.400
Yes	Yes	4	10	10	**0.321**	**0.500**

Table 1. Results for the LAMP2 Dataset for Personalized News Categorization

Our analysis is split into two parts. One based on the number of relevant user profile data points (K) and the other based on the number of feedback documents used in the query expansion process by RM3 (N).

Using query expansion models such as RM3 with BM25's implicit feedback set improved the pre-trained LLM's capability of predicting the right category of personalized news articles, and this is especially evident for higher values of K. The values below are a comparison of BM25 with tuned RM3 model (best configuration).

1. K = 1:
   - F1 score increased from 0.162 to 0.179
   - Accuracy increased from 0.462 to 0.467
2. K = 2:
   - F1 score increased from 0.152 to 0.158
   - Accuracy increased from 0.430 to 0.434
3. **K = 4**:
   - F1 score increased from 0.120 to 0.321 **(~2.68x)**
   - Accuracy increased from 0.404 to 0.5 **(~1.24x)**

From the RM3 perspective, increasing the number of feedback documents from 2 to 10 for query expansion improved the performance of the LLM model.

1. N = 2:
   - F1 score increased from 0.1495 to 0.1790
   - Accuracy increased from 0.4472 to 0.4665
2. **N = 10**:
   - F1 score increased from 0.1156 to 0.3208 **(~2.77x)**
   - Accuracy increased from 0.3999 to 0.5 **(~1.25x)**

Thus, we observe that when the RM3 model is properly tuned with the appropriate number of feedback documents and terms for query expansion offers potential benefits in retrieving an expanded set of data points from a user's profile. This, in turn, aids in generating a personalized user query prompt for downstream LLM tasks.

Table 2 tabularizes the data from the multiple experiments run on the LaMP3 dataset by varying the same set of parameters as before, K and Fb Docs. However, the evaluation metrics as mentioned earlier are now MAE, mean absolute error, and RMSE, root mean square error.

BM25	RM3	K	Fb Docs	Fb Terms	MAE	RMSE
Yes	-	1	-	-	0.473	0.787
Yes	Yes	1	3	10	0.499	0.821
Yes	Yes	1	10	10	0.472	0.799
Yes	-	4	-	-	0.422	0.741
Yes	Yes	4	3	10	0.423	0.741
Yes	Yes	4	10	10	0.418	0.742

Table 2. Results for the LAMP3 Dataset for Personalized Product Rating

As per 2, we observe a decrease in the Mean Absolute Error and the Root Mean Square Error when more number of user profile data points (K increases from 1 to 4) are appended to prompt sent to the LLM.

1. Only BM25
   - MAE decreased from 0.473 to 0.422
   - RMSE decreased from 0.787 to 0.741
2. Fb Docs = 3, Fb Terms = 10
   - MAE decreased from 0.499 to 0.423
   - RMSE decreased from 0.821 to 0.741
3. Fb Docs = 10, Fb Terms = 10
   - MAE decreased from 0.472 to 0.418
   - RMSE decreased from 0.799 to 0.742

Query expansion using RM3 exhibited slight improvement for the LaMP3 dataset as observed in 2. We find a marginal decrease in Mean Average Error (MAE) when comparing the tuned RM3 model to BM25, that is for k = 4, MAE values were 0.418 and 0.422. However, we faced the obstacle imposed by the Flan-T5-base model's fixed number of input token limitation, preventing us from conducting experiments with values of K greater than 4.

6 Conclusion and Future Work

In a nutshell, our paper focuses on studying the impact of query expansion, particularly by leveraging the concept of pseudo-relevance feedback and relevance model to perform the query expansion. Despite constraints imposed by our resource, we notice a significant improvement in metrics to further consider query expansion in personalizing large language models. To overcome the block imposed by the input-token limit, we propose a solution of summarizing the reviews as a precursor to constructing the user-personalized query prompt. Furthermore, indexing the summarized text for retrieval along with employing appropriate dense retrieval models for ranking before leveraging the query expansion capabilities of the RM3 model has the potential to enhance the performance of Language Models (LLMs) for downstream personalization tasks. Further research could be directed in the direction of experimenting the impact of query expansion on LLM personalization with larger large language models (eg. BERT-QE [21]) that are fine-tuned on the dataset as fine-tuning has often shown notable results.

Acknowledgments

We would like to express our gratitude to the authors of the LaMP paper for their valuable contributions to the IR field by introducing quality benchmarks and datasets to work with. We would also like to thank our fellow authors for their partnership and commitment throughout the research process. Additionally, we acknowledge the resources and knowledge provided by the University of Massachusetts Amherst that made this work possible.

References

[1] Sara Alnofaie, Mohammed Dahab, and Mahmoud Kamal. 2016. A Novel Information Retrieval Approach using Query Expansion and Spectral-based. *International Journal of Advanced Computer Science and Applications* 7, 9 (2016). https://doi.org/10.14569/IJACSA.2016.070950

[2] Giambattista Amati. 2009. *BM25*. Springer US, Boston, MA, 257–260. https://doi.org/10.1007/978-0-387-39940-9_921

[3] R. Attar and A. S. Fraenkel. 1977. Local Feedback in Full-Text Retrieval Systems. *J. ACM* 24, 3 (jul 1977), 397–417. https://doi.org/10.1145/322017.322021

[4] Elias Bassani, Nicola Tonellotto, and Gabriella Pasi. 2023. Personalized Query Expansion with Contextual Word Embeddings. *ACM Trans. Inf. Syst.* 42, 2, Article 61 (dec 2023), 35 pages. https://doi.org/10.1145/3624988

[5] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Defu Lian, and Enhong Chen. 2023. When Large Language Models Meet Personalization: Perspectives of Challenges and Opportunities. arXiv:2307.16376 [cs.IR]

[6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG]

[7] Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query Expansion by Prompting Large Language Models. arXiv:2305.03653 [cs.IR]

[8] Nasreen Abdul Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004 (NIST Special Publication, Vol. 500-261)*, Ellen M. Voorhees and Lori P. Buckland (Eds.). National Institute of Standards and Technology (NIST). http://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf

[9] Victor Lavrenko and W. Bruce Croft. 2001. Relevance Based Language Models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (New Orleans, Louisiana, USA) *(SIGIR '01)*. Association for Computing Machinery, New York, NY, USA, 120–127. https://doi.org/10.1145/383952.383972

[10] Hang Li, Ahmed Mourad, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. 2023. Pseudo Relevance Feedback with Deep Language Models and Dense Retrievers: Successes and Pitfalls. *ACM Trans. Inf. Syst.* 41, 3, Article 62 (apr 2023), 40 pages. https://doi.org/10.1145/3570724

[11] Sebastian Porsdam Mann, Brian D. Earp, Nikolaj Møller, Suren Vynn, and Julian Savulescu. 2023. Autogen: A Personalized Large Language Model for Academic Enhancement–Ethics and Proof of Principle. *American Journal of Bioethics* 23, 10 (2023), 28–41. https://doi.org/10.1080/15265161.2023.2233356

[12] Rishabh Misra. 2022. News Category Dataset. arXiv:2209.11429 [cs.CL]

[13] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. 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)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 188–197. https://doi.org/10.18653/v1/D19-1018

[14] Dipasree Pal, Mandar Mitra, and Kalyankumar Datta. 2013. Query Expansion Using Term Distribution and Term Association. arXiv:1303.0667 [cs.IR]

[15] Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

[16] S. E. Robertson and K. Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science* 27, 3 (May 1976), 129–146. https://doi.org/10.1002/asi.4630270302

[17] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406 [cs.CL]

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query Expansion with Large Language Models. arXiv:2303.07678 [cs.IR]

[20] Jinxi Xu and W. Bruce Croft. 2000. Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Trans. Inf. Syst.* 18, 1 (jan 2000), 79–112. https://doi.org/10.1145/333135.333138

[21] Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized Query Expansion for Document Reranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4718–4728. https://doi.org/10.18653/v1/2020.findings-emnlp.424

Received 12 December 2023