LaMP: When Large Language Models Meet Personalization

UMassAmherst
College of Information & Computer Sciences
Center for Intelligent Information Retrieval
Google Research

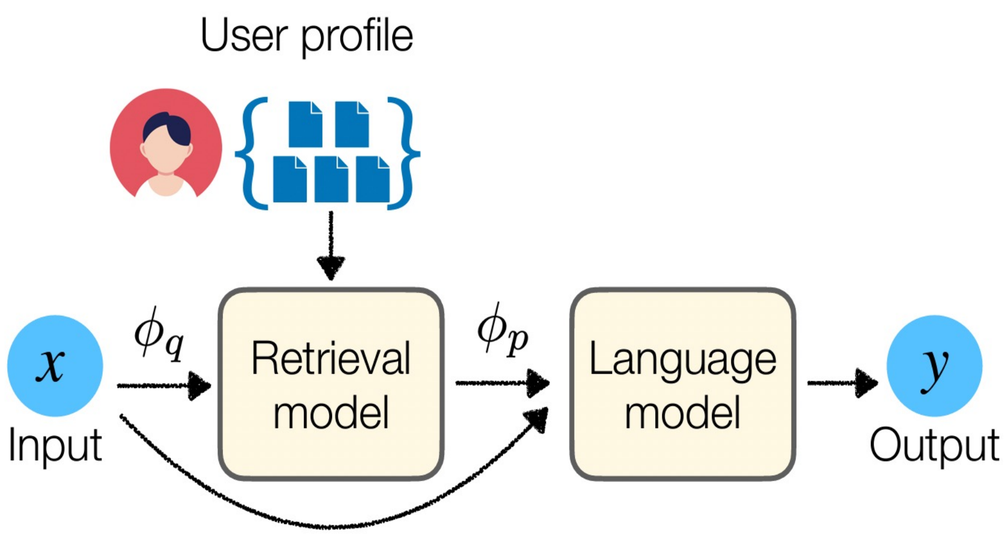
Alireza Salemi, Sheshera Mysore, Michael Bendersky, Hamed Zamani
{asalemi,smysore,zamani}@cs.umass.edu, bemike@google.com

Introduction

In recent years, the advent of large-scale language models (LLMs) has led to their widespread integration within numerous natural language processing (NLP) systems. As such, it is increasingly crucial to explore the potential for personalization of these models to cater to individual user requirements. In this research, we delve into this area of interest and present a benchmark for evaluating the personalization capabilities of LLMs.

Benchmark Website

Retrieval Augmentation for Personalizing LLMs



Research Enabled by LaMP

- Prompting LMs for Personalization.
- Evaluation of Personalized Text Generation
- Learning to Retrieve from User Profiles

The LaMP Benchmark

The LaMP benchmark consists of 7 diverse tasks, three text classification and four text generation. There are two settings:

Task	Task Type	Setting	#profile	#class
Citation Identification	Binary classification	user time	90.6 ± 53.8 84.1 ± 47.5	2
News Categorization	Categorical classification	user time	306.4 ± 286.6 191.0 ± 168.4	15
Product Rating	Ordinal classification	user time	188.1 ± 129.4 185.4 ± 129.3	5
News Headline Generation	Text generation	user time	287.1 ± 360.6 204.5 ± 250.7	-
Scholarly Title Generation	Text generation	user time	89.6 ± 53.8 87.8 ± 53.6	-
Email Subject Generation	Text generation	user time	80.7 ± 51.7 55.6 ± 36.3	-
Tweet Paraphrasing	Text generation	user time	17.7 ± 15.1 15.7 ± 14.8	-

Results

Task	Metric	FlanT5-base (fine-tuned)						Flant5-xxl(zero-shot)		GPT3.5 (zero-shot)	
		Non Personalized	Random	BM25	Contriever	Recency	Tuned Profile	Non Personalized	Tuned Profile	Non Personalized	Tuned Profile
Citation Identification	Accuracy	0.628	0.657	0.682	0.688	0.691	**0.714**	0.502	**0.636**	0.508	**0.634**
News Categorization	F1	0.574	0.634	0.614	0.656	0.645	**0.659**	0.520	**0.536**	0.519	**0.551**
Product Rating	MAE	0.280	0.279	0.278	0.281	0.279	**0.266**	0.333	**0.299**	0.677	**0.603**
News Headline Generation	Rouge-L	0.145	0.155	0.157	0.162	0.158	**0.162**	0.160	**0.172**	0.128	**0.140**
Scholarly Title Generation	Rouge-L	0.416	0.414	0.423	0.426	0.420	**0.431**	0.422	**0.433**	**0.355**	0.351
Email Subject Generation	Rouge-L	0.463	0.507	0.522	0.530	0.518	**0.533**	0.319	**0.387**	-	-
Tweet Paraphrasing	Rouge-L	0.416	0.456	0.457	0.455	0.453	**0.465**	**0.396**	0.389	**0.330**	0.318

- Personalizing language models yields better performance in all tasks in the fine-tuning setting and 6 out of 7 tasks in the zero-shot setting.
- The language model with personalized input on average achieves 15.6% and 12.5% improvement over non-personalized language model in fine-tuning and zero-shot settings, respectively.
- Fine-tuning smaller language models results in a better performance than zero-shot usage of LLMs.
- Our results show that the choice of retrieval method for selecting profile entries can significantly affect the performance of the language model.