

# LaMP: When Large Language Models Meet Personalization

Presented by Choeun Kim

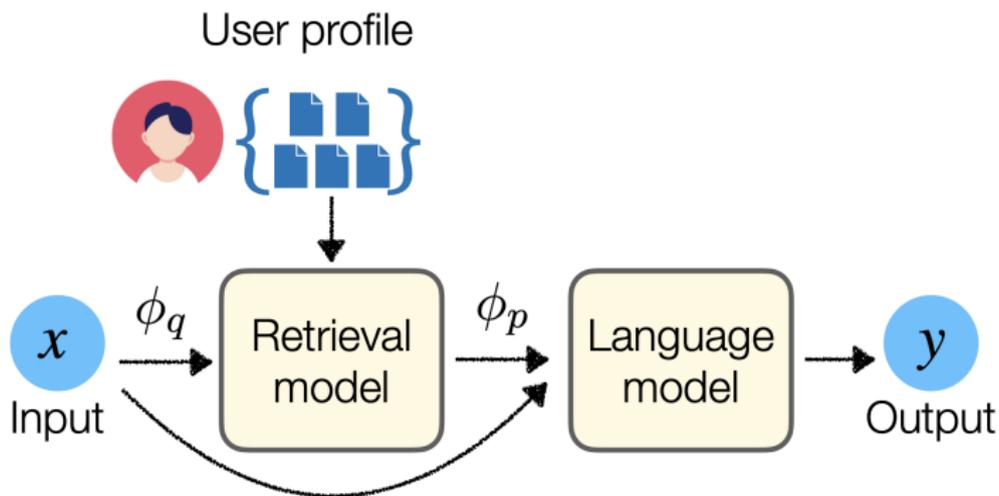
IDEA

October 26, 2025

# Motivation

- Existing NLP benchmarks (GLUE, SuperGLUE) follow a “one-size-fits-all” paradigm.
- Real-world users have distinct preferences and writing styles.
- Focus on developing strategies for training models personalized via user-specific inputs.
- **Goal:** Evaluate and enhance personalization in LLMs

# Overview



**Figure:** An overview of the retrieval-augmented method for personalizing LLMs.  $\phi_q$  and  $\phi_p$  represent query and prompt construction functions.

# Overview of LaMP Benchmark

**LaMP** = *Language Model Personalization*

7 tasks across classification and generation:

## Classification

- Citation Identification (binary)
- Movie Tagging (15 classes)
- Product Rating (5 classes)

## Generation

- News Headline
- Scholarly Title
- Email Subject
- Tweet Paraphrasing

# LaMP : classification

## Input

### LaMP-1: Personalized Citation Identification

For an author who has written the paper with the title "[TITLE]", which reference is related?  
Just answer with [1] or [2] without explanation.  
[1]: "[REF1]" [2]: "[REF2]"

### LaMP-2: Personalized Movie Tagging

Which tag does this movie relate to among the following tags?  
Just answer with the tag name without further explanation.  
tags: [sci-fi, based on a book, comedy, action, twist ending, dystopia, dark comedy, classic, ...] description: [MOVIE]

### LaMP-3: Personalized Product Rating

What is the score of the following review on a scale of 1 to 5?  
Just answer with 1, 2, 3, 4, or 5 without further explanation.  
review: [REVIEW]

## Output

[1]

comedy

4

## Profile

title: [TITLE]  
abstract: [ABSTRACT]

description: [MOVIE]  
tag: [TAG]

text: [REVIEW]  
score: [SCORE]

# LaMP : generation

## LaMP-4: Personalized News Headline Generation

Generate a headline for the following article: [ARTICLE]

The Best Cheap  
Wine: Two Buck  
Chuck vs Three  
Wishes

title: [TITLE]  
text: [ARTICLE]

## LaMP-5: Personalized Scholarly Title Generation

Generate a title for the following abstract of a paper: [ABSTRACT]

Attention is All  
You Need

title: [TITLE]  
abstract: [ABSTRACT]

## LaMP-6: Personalized Email Subject Generation

Generate a subject for the following email: [EMAIL]

A bug in the  
class HelloWorld

title: [TITLE]  
email: [EMAIL]

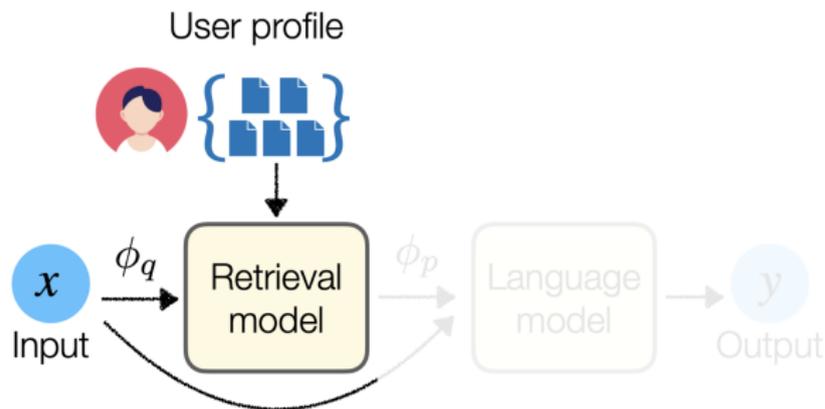
## LaMP-7: Personalized Tweet Paraphrasing

Paraphrase the following tweet without any explanation before or after it: [TWEET]

I hope so! what  
time do you get  
out? I get out at  
335

text: [TWEET]

# Retrieval Model

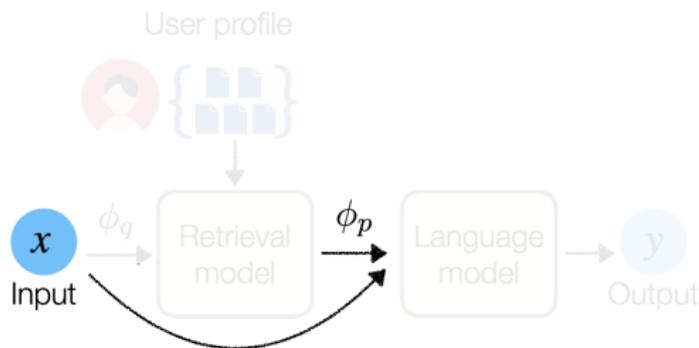


- context length constraint of LLMs
  - not all entries within a user profile are necessarily relevant to the specific input at hand
- Retrieval model  $\mathcal{R}$  selects top- $k$  relevant entries from the user profile  $P_u$  as input prompts ( $k$  is a hyperparameter)

# Retrieval Model

- BM25 : Term Matching
- Contriever : Dense semantic retrieval (SoTA)
- Recency : Most recent items
- Random : Baseline

# Retrieval Augmentations



Denote the selected top- $k$  entries as  $\mathbf{d} = (d_1, \dots, d_k)$ .

- IPA (In-Prompt Augmentation)

$$\phi_p(x, \mathbf{d}) = (x, (d_1, \dots, d_k))$$

- Fusion-in-Decoder (FiD) (p.13)

$$\phi_p(x, \mathbf{d}) = (x, \text{Dec}(\text{Enc}(x, d_1), \dots, \text{Enc}(x, d_k)))$$

# Two Evaluation Settings

- **User-based Split:** Personalization for new users.
  - train/val/test splits are made by partitioning across users, ensuring that no shared users appear across splits
- **Time-based Split:** Predicting future interactions for existing users.
  - train/val/test splits are made by partitioning user items ordered by time
  - the most recent user items are chosen to create the input-output pairs, with older items serving as user profiles

# Experiments : User-based

Dataset	Metric	FlanT5-base (fine-tuned)						
		Non-Personalized		Untuned profile, $k = 1$			Tuned profile	
		No-Retrieval	Random	Random	BM25	Contriever	IPA	FiD( $k = 16$ )
LaMP-1U: Personalized Citation Identification	Accuracy $\uparrow$	0.518	0.539	0.598	0.649	0.688	0.734	<b>0.754</b>
LaMP-2U: Personalized Movie Tagging	Accuracy $\uparrow$	0.468	0.442	0.497	0.524	0.536	0.556	<b>0.642</b>
	F1 $\uparrow$	0.435	0.403	0.459	0.480	0.506	0.519	<b>0.607</b>
LaMP-3U: Personalized Product Rating	MAE $\downarrow$	0.275	0.286	0.284	0.258	0.248	0.246	<b>0.236</b>
	RMSE $\downarrow$	0.581	0.607	0.602	0.573	0.563	0.565	<b>0.539</b>
LaMP-4U: Personalized News Headline Generation	ROUGE-1 $\uparrow$	0.153	0.159	0.162	0.167	0.173	<b>0.186</b>	0.180
	ROUGE-L $\uparrow$	0.140	0.147	0.148	0.153	0.159	<b>0.171</b>	0.166
LaMP-5U: Personalized Scholarly Title Generation	ROUGE-1 $\uparrow$	0.418	0.408	0.409	0.440	0.431	<b>0.450</b>	0.431
	ROUGE-L $\uparrow$	0.378	0.370	0.371	0.399	0.393	<b>0.409</b>	0.392
LaMP-6U: Personalized Email Subject Generation	ROUGE-1 $\uparrow$	0.379	0.473	0.486	0.586	0.572	<b>0.587</b>	0.567
	ROUGE-L $\uparrow$	0.358	0.457	0.470	0.570	0.558	<b>0.575</b>	0.555
LaMP-7U: Personalized Tweet Paraphrasing	ROUGE-1 $\uparrow$	0.509	0.510	0.514	0.521	0.524	<b>0.528</b>	0.517
	ROUGE-L $\uparrow$	0.455	0.457	0.460	0.468	0.471	<b>0.475</b>	0.464

# Experiments : Time-based

Dataset	Metric	FlanT5-base (fine-tuned)							
		Non-Personalized		Untuned profile, $k = 1$				Tuned profile	
		No-Retrieval	Random	Random	BM25	Contriever	Recency	IPA	FiD( $k = 16$ )
LaMP-1T: Personalized Citation Identification	Accuracy $\uparrow$	0.628	0.625	0.657	0.682	0.688	0.691	<b>0.714</b>	0.698
LaMP-2T: Personalized Movie Tagging	Accuracy $\uparrow$	0.506	0.513	0.518	0.539	0.533	0.549	0.564	<b>0.661</b>
	F1 $\uparrow$	0.443	0.449	0.456	0.472	0.475	0.492	0.519	<b>0.624</b>
LaMP-3T: Personalized Product Rating	MAE $\downarrow$	0.280	0.280	0.279	0.278	0.281	0.279	0.266	<b>0.250</b>
	RMSE $\downarrow$	0.615	0.616	0.612	0.614	0.606	0.608	<b>0.598</b>	<b>0.598</b>
LaMP-4T: Personalized News Headline Generation	ROUGE-1 $\uparrow$	0.159	0.160	0.169	0.171	0.176	0.173	<b>0.177</b>	0.170
	ROUGE-L $\uparrow$	0.145	0.147	0.155	0.157	0.162	0.158	<b>0.162</b>	0.157
LaMP-5T: Personalized Scholarly Title Generation	ROUGE-1 $\uparrow$	0.462	0.459	0.460	0.471	0.472	0.466	<b>0.479</b>	0.456
	ROUGE-L $\uparrow$	0.416	0.412	0.414	0.423	0.426	0.420	<b>0.431</b>	0.414
LaMP-6T: Personalized Email Subject Generation	ROUGE-1 $\uparrow$	0.479	0.500	0.525	0.537	0.545	0.532	<b>0.547</b>	0.540
	ROUGE-L $\uparrow$	0.463	0.452	0.507	0.522	0.530	0.518	<b>0.533</b>	0.525
LaMP-7T: Personalized Tweet Paraphrasing	ROUGE-1 $\uparrow$	0.462	0.474	0.505	0.508	0.505	0.503	<b>0.516</b>	0.502
	ROUGE-L $\uparrow$	0.416	0.457	0.456	0.457	0.455	0.453	<b>0.465</b>	0.450

# Appendix : Fusion in Decoder

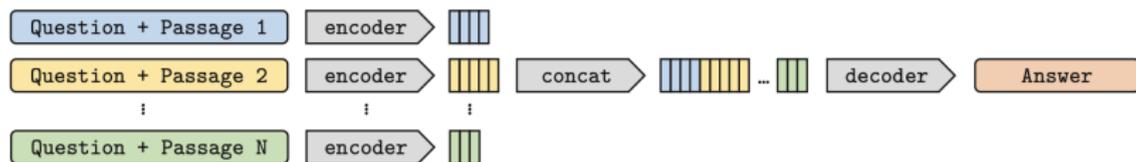


Figure 2: Architecture of the Fusion-in-Decoder method.

Figure: Architecture of the Fusion-in-Decoder method