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

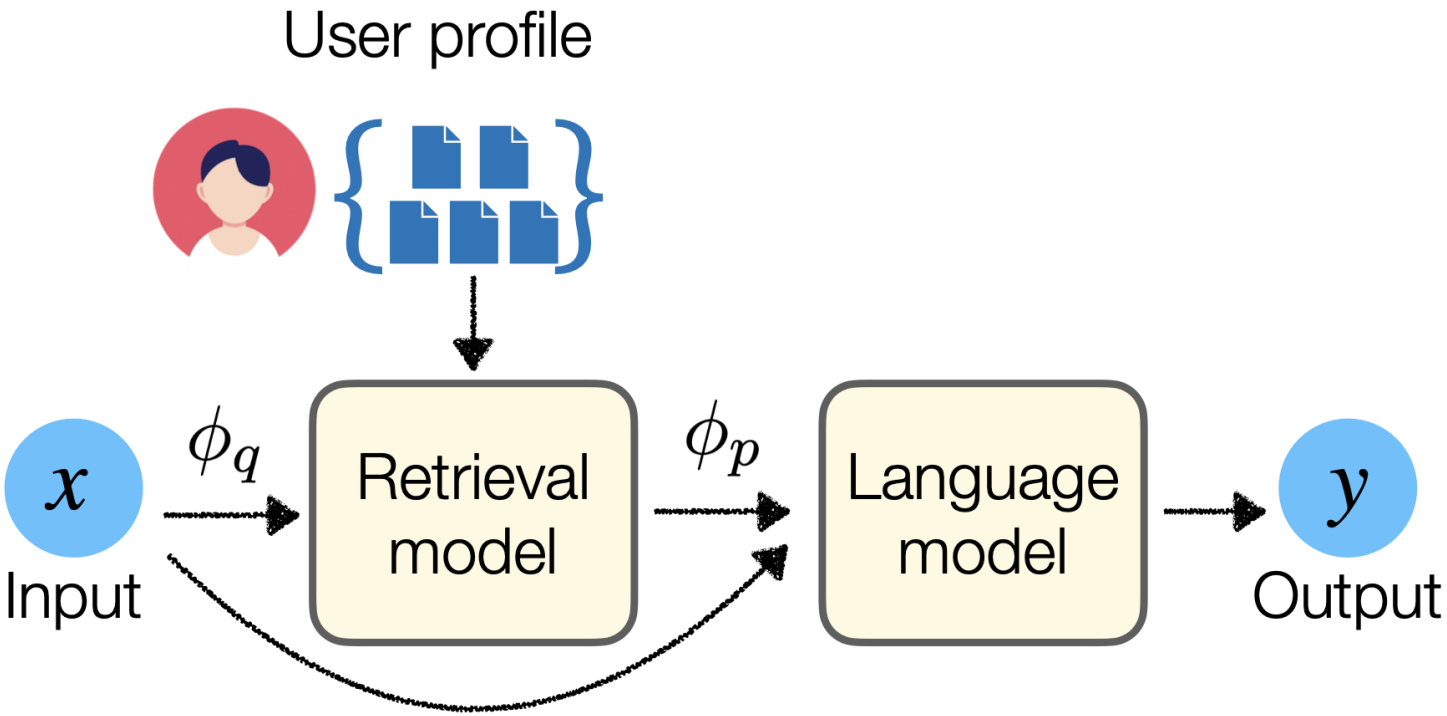


Figure: An overview of the retrieval-augmented method for personalizing LLMs. ϕ_q and ϕ_p represent query and prompt construction functions.

Overview of LaMP Benchmark

LaMP = *Language Model Personalization*
7 tasks across classification and generation:

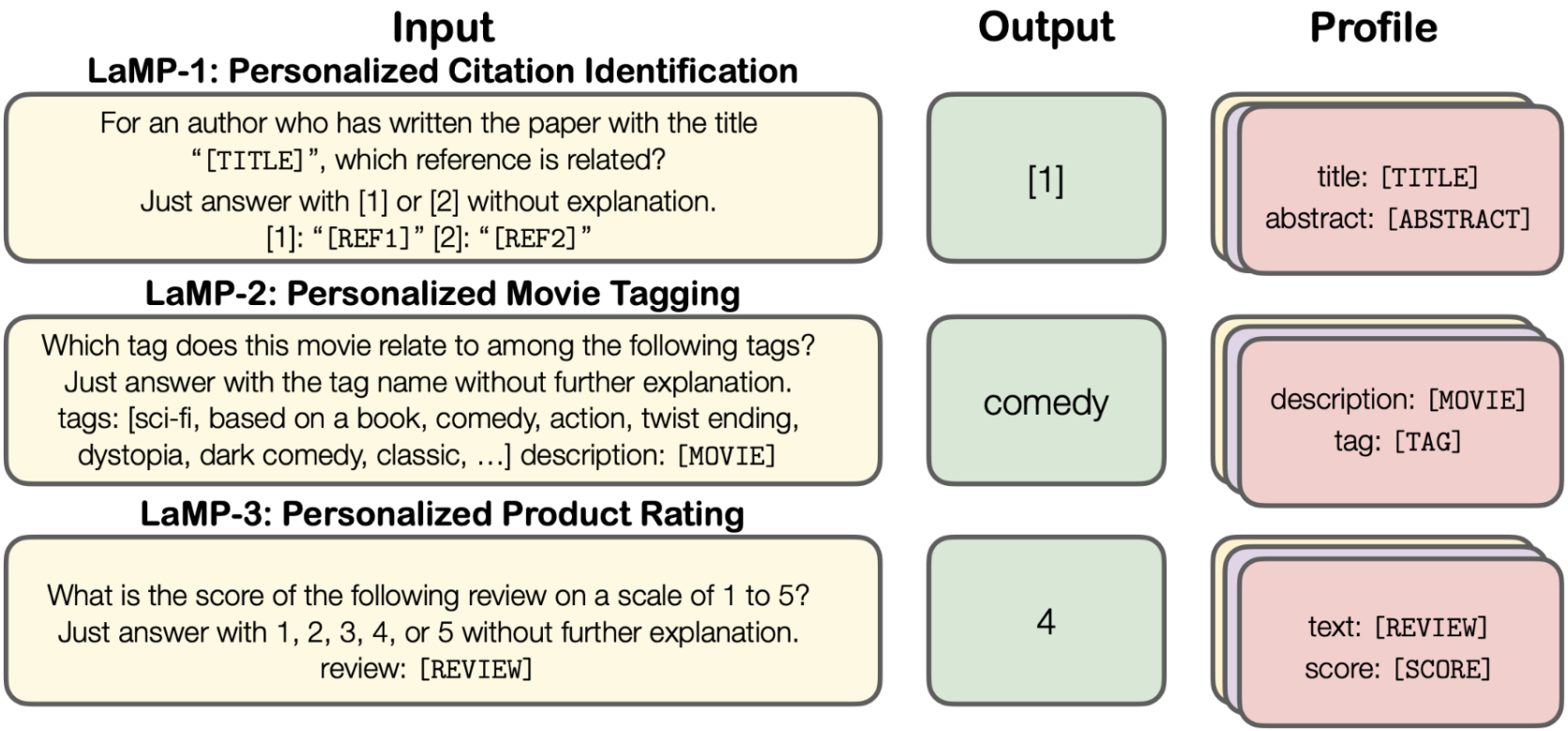
Classification

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

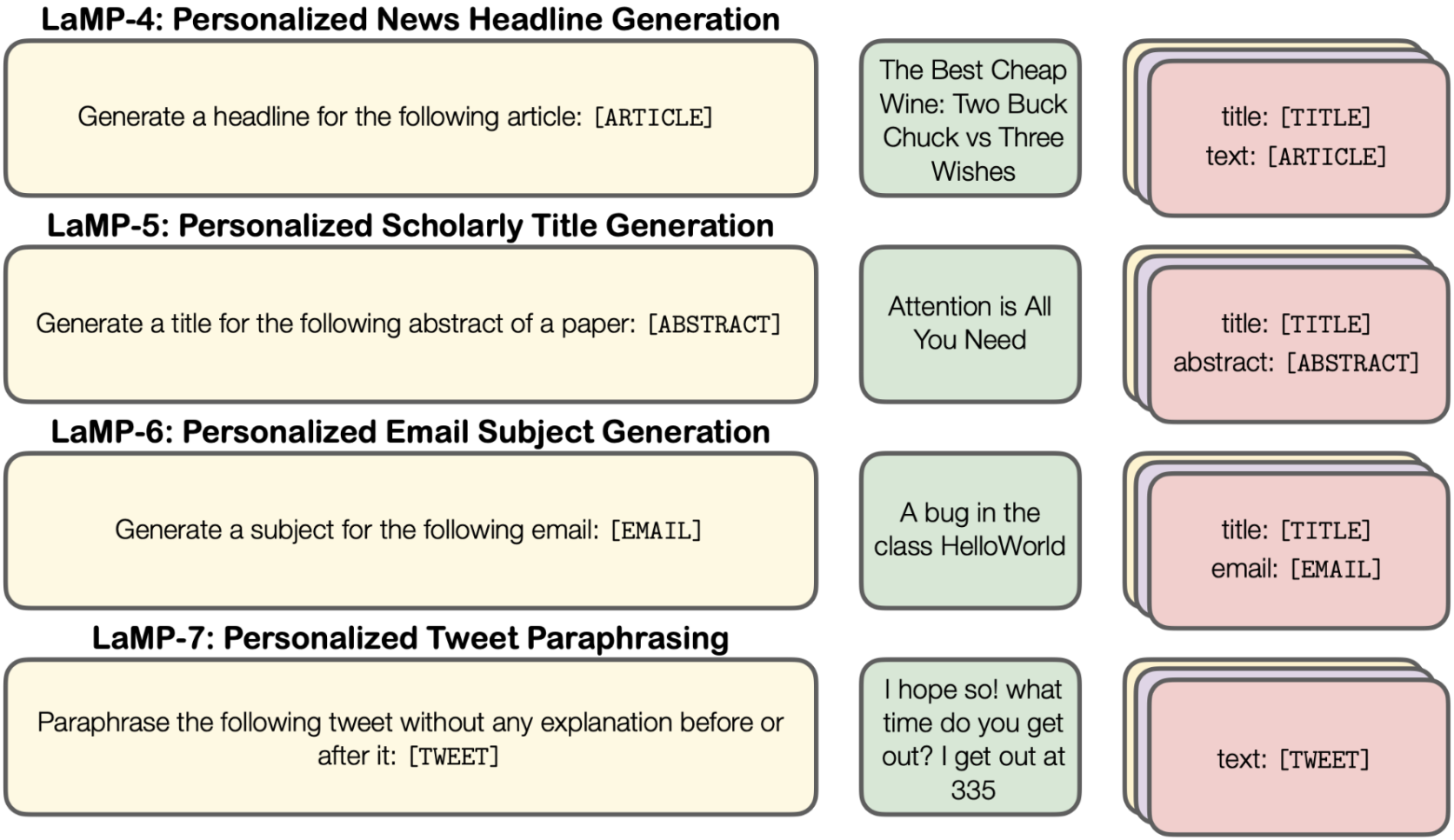
Generation

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

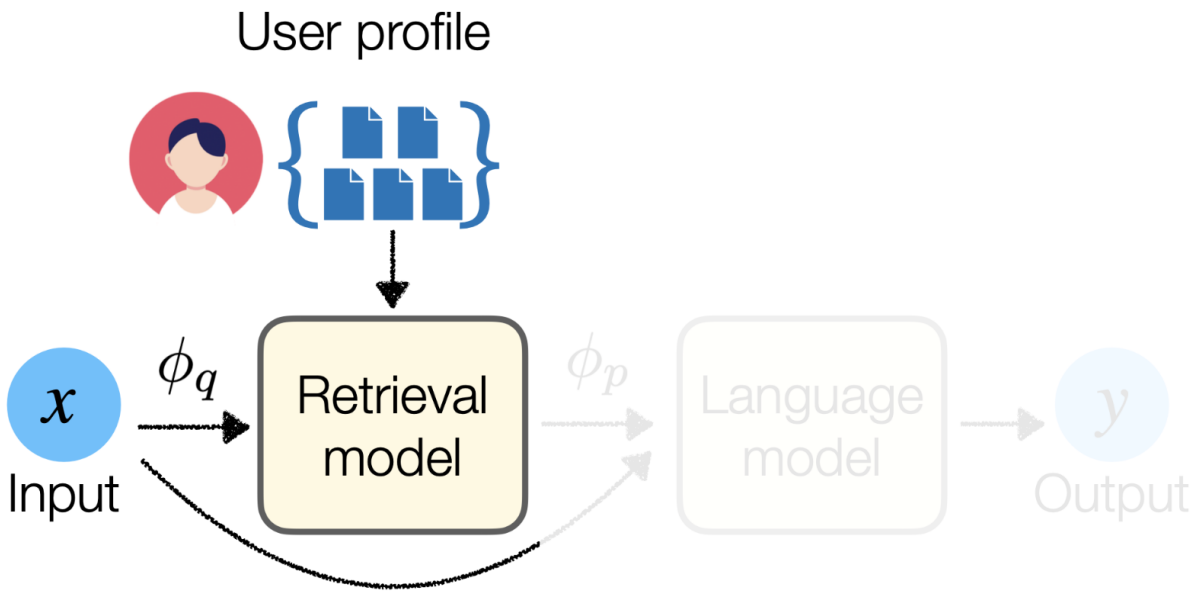
LaMP : classification



LaMP : generation



Retrieval Model



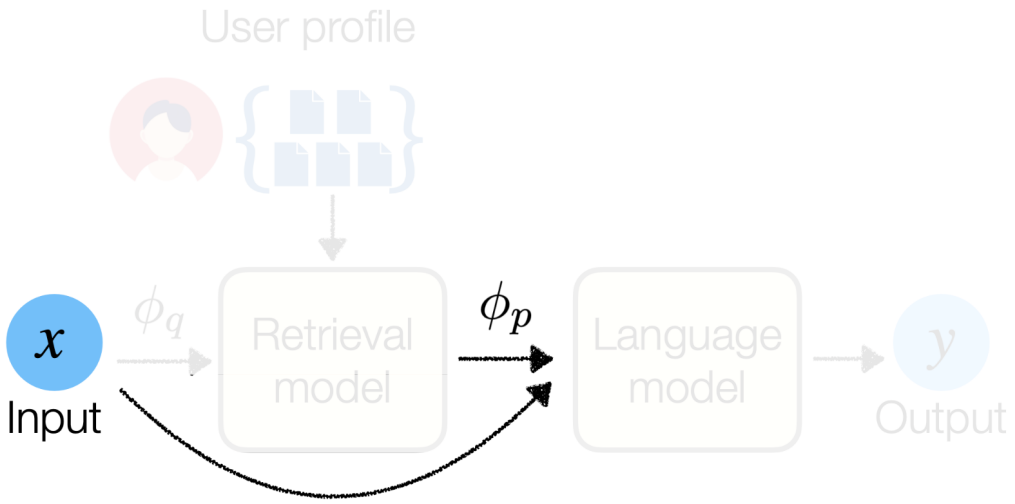
- context length constraint of LLMs
- not all entries within a user profile are necessarily relevant to the specific input at hand

$\rightarrow$ 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 $\mathbf{d} = (d_1, \ldots, d_k)$.

- IPA (In-Prompt Augmentation)

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

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

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

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

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

Experiments : Time-based

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

Appendix : Fusion in Decoder

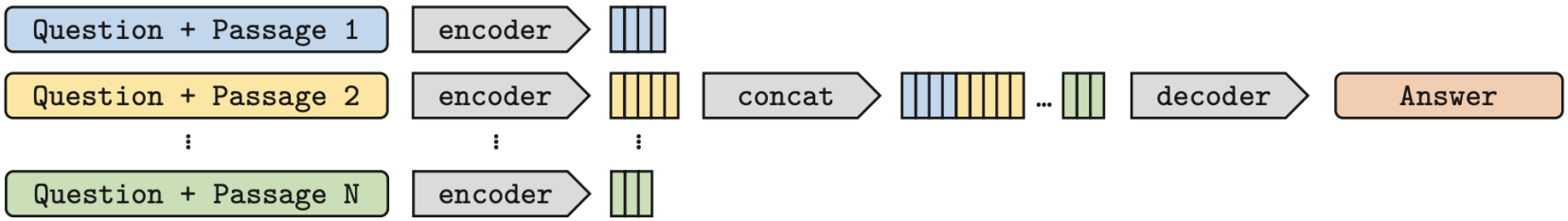


Figure 2: Architecture of the Fusion-in-Decoder method.

Figure: Architecture of the Fusion-in-Decoder method