Fall 2023: COMPSCI 646 - Final Project

Nowadays, people use large language models (LLMs), such as ChatGPT, for getting the answer to their questions, for rewriting or drafting a document, for summarizing a document, and the list goes on. Imagine two users ask the same question from a large language model. They both receive the same response.[1] However, different users have different preferences and different background. In another example, imagine you use LLMs for drafting an email. Different users follow different writing styles in their email.

In your final project, you are going to address this issue and conduct research on "personalizing large language models."

Problem Formulation. Generative language models often take an input x and predict the most probable sequence tokens y that follows x. Personalizing language models can be defined as conditioning the model's output on a user u, represented by a user profile. User profile is defined as the user's historical data, i.e., the past input and personalized outputs produced or approved by the user. Therefore, each data entry consists of three components: an input sequence that serves as the model's input, a target output that the model is expected to produce, and a profile that encapsulates any auxiliary information that can be employed to personalize the model according to the user's profile. Therefore, personalizing LLMs can be formalized as follows: for a given textual input x, the goal is to develop a model M that generates personalized output y for the user u. This can be modeled as $\arg\max_y p(y|x,u)$. The profile for user u is defined as $P_u = \{(x_{u1}, y_{u1}), (x_{u2}, y_{u2}), \cdots, (x_{um_u}, y_{um_u})\}$ where each (x_{ui}, y_{ui}) denotes a pair of input and personalized output for user u.

The LaMP Benchmark There is a public benchmark for the language model personlization tasks, called LaMP. It includes 7 different tasks as follows:

- Personalized Citation Identification
- Personalized News Categorization
- Personalized Product Rating
- Personalized News Headline Generation

[1] There are some randomness in sampling from their output distributions that may lead to different outputs, but if we disable this random sampling they will produce identical output for the same input.

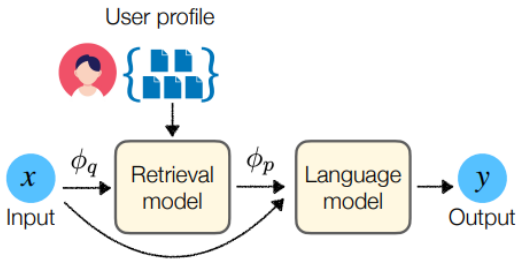

Figure 1: An overview of the retrieval-augmented method for personalizing LLMs. ϕ_q and ϕ_p represent query and prompt construction functions

- Personalized Scholarly Title Generation
- Personalized Email Subject Generation[2]
- Personalized Tweet Paraphrasing

The LaMP benchmark is available at `https://lamp-benchmark.github.io/index.html`. Each group is expected to choose two dataset from this list for their final project. If you are working on this project individually, you can do experiments on only one of these datasets.

Retrieval-Augmentation for Personalizing LLMs There are several ways to personalize a language model: 1) fine-tuning the language model for each user on user's personalized data and 2) prompting the language model with personalized input. The former approach necessitates substantial computational resources, especially for fine-tuning LLMs. Moreover, accommodating personalized language models for each user in scenarios encompassing millions of users necessitates a significant storage capacity. Thus, in a practical context, directing attention towards the development of personalized prompts emerges as a more favorable approach.

In the tasks at hand, each user profile consists of a (potentially large) collection of data points pertaining to the user. Given the inherent context length constraint of many LMs in addition to their efficiency and cost, it is only practical to incorporate a subset of these data points as input prompts. Moreover, it is important to note that not all entries within a user profile are necessarily relevant to the specific task the user aims to accomplish. Thus, we propose the development of solutions based on retrieval augmentation. This framework selectively extracts pertinent information from the user profile that are relevant to the current unseen test case. An overview of the method is shown in Figure 1.

To achieve personalization for a given sample (x_i, y_i) for user u, we employ three primary components: (1) a query generation function ϕ_q that transforms

[2]Access to this dataset requires special permission, so do not use this dataset.

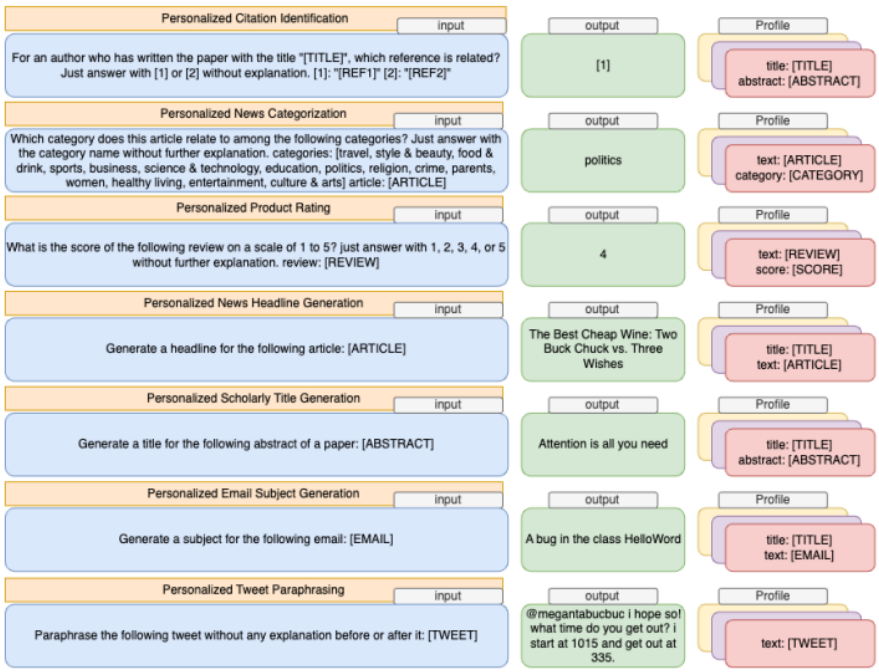

Figure 2: Input and output examples from LaMP.

the input x_i into a query q for retrieving from the user u's profile, (2) a retrieval model $\mathcal{R}(q, P_u, k)$ that accepts a query q, a user profile P_u and retrieves k most pertinent entries from the user profile, and (3) a prompt construction function ϕ_p that assembles a personalized prompt for the user u based on input x_i and the retrieved entries. Consequently, the input (prompt) $\bar{x_i}$ for the language model is derived using the following formulation:

$$\bar{x_i} = \phi_p(x_i, \mathcal{R}(\phi_q(x_i), P_u, k)) \tag{1}$$

where we use $(\bar{x_i}, y_i)$ to train or evaluate the language models. We can imagine various implementation of $\mathcal{R}$. Read this paper to learn about them: `https://arxiv.org/abs/2304.11406`.

Figure 2 demonstrates the data format for all tasks. Each of them has its specific format of queries, documents and "profile". For example, title and abstract of all the papers from an author are defined as an user profile of this author in personalized citation identification.

You are expected to conduct research on the retrieval aspect of this model.

1 Some notes about the project

Before you decide on your research question, consider the following points:

- You are expected to submit a proposal for your project which accounts for 10% of your final project grade. In your proposal, you must introduce a novel research question and for your final project you are expected to conduct experiments to address that research question. Choosing the right research question is important. Your research question must be novel, important, and well-motivated. If you have a research question in mind, ask yourself: if I find the answer to this question, does it matter to anybody or any application?

- If you think the research questions studied in the LaMP paper (cited earlier) is not studied carefully, feel free to reuse those questions, but your answer to the questions are expected to go beyond the LaMP paper.

- Is your proposed research feasible to complete as a course project? Before you actually start to implement, carefully estimate the time and computing resources you need and make sure you have the required resources.

Our grading will mainly focus on these aspects:

- Do you clearly state your research question and describe why it is important? For example, you can provide a real-world scenario and explain why your research is helpful in practice.

- Does your work actually answer your research question? your methodology and experimental results should answer the research question you define.

- Is related work well studied? We expect that you study the literature and write a careful "Related Work" section.

- Is your report well written and well presented? You can use appropriate equations, figures, tables, and any visualization to present your work more clearly.

- The performance improvement over baselines is not required. We encourage you to try your best to improve your methods for better performance. That being said, even if you observe no improvement compared to the baseline, it is perfectly acceptable as a course project. What matters is making sure that your method is correctly implemented and is accurately studying the proposed research question.

- You can finish the project independently or as a group of two. For one person, you need to validate your work on at least one dataset within the LaMP benchmark. For a two-person team, you need to validate on at least **two** datasets.

2 Proposal and Final Report

You submit brief answers to a form (think of it as a Google Form) as your proposal. The questions in the form will focus on:

1. Research Question: What is the research question you aim to answer?
2. Data: What datasets from LaMP you will use in your project?
3. Prior work: Examine the papers citing or cited in the papers tackling your problem/dataset of interest: What is the state of the current published research on this problem? What are the open questions? You are asked to mention the top three most similar papers to your project idea.
4. Method design: What is your proposed method for tackling the problem? Why do you believe this method will work well? What makes this method novel or interesting? At the time of the proposal, a high-level answer to this part is sufficient; no details are needed.

The final report should look like a short conference paper! It must elaborate on all the questions you answered for the proposal and describe your experiments and results. It needs to be in PDF format and be at most 4 pages, including any content except references, using the current ACM SIG two-column conference format. Suitable LaTeX, Word, and Overleaf templates[3] are available from the ACM Website[4] (use "sigconf" proceedings template for LaTeX and the Interim Template for Word). When you submit your final report, you are asked to submit your code.

3 Tentative Schedule

- Proposal due date: Nov. 20th
- Final report due date: Dec. 8th

[3] https://www.overleaf.com/gallery/tagged/acm-official
[4] https://www.acm.org/publications/proceedings-template