Work-in-Progress: On-device Retrieval Augmented Generation with Knowledge Graphs for Personalized Large Language Models

Chanhee Lee, Deeksha Prahlad, Dongha Kim, and Hokeun Kim
School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, United States
Email: {chanheel, dprahlad, dongha, hokeun}@asu.edu

***Abstract*—On-device large language models (LLMs) have promising benefits, including personalization, enhanced user privacy, and offline operations. We present our work-in-progress approach to on-device LLMs leveraging retrieval augmented generation (RAG) based on personalized on-device knowledge graphs (KGs). We aim to enhance the personalized experience of LLMs while preserving privacy by keeping sensitive data locally.**

I. Introduction and Related Work

As large language models (LLMs) perform exceptionally well in generating realistic conversations, rapid progress has been made in on-device LLMs, as shown in the Samsung Galaxy S24 [1]. On-device LLMs have concrete advantages over external LLM servers in that privacy can be protected entirely within devices. Due to resource limitations, however, gaining personalized responses with LLMs is still challenging.

Previous research uses LLMs for personalized generation or recommendation. Personalized multimodal generation (PMG) [2] converts user behaviors, such as conversations from chat applications, into texts used as prompts to condition the generator. Heterogeneous knowledge fusion (HKF) [3] constructs structured templates for prompt engineering to capture heterogeneous user behaviors and uses LLMs to perform knowledge fusion. Both approaches, however, are almost unusable for on-device due to limited storage and computational resources.

This short paper presents our work-in-progress approach to improving the accuracy of on-device LLMs with retrieval augmented generation (RAG) based on personalized knowledge.

II. Proposed Approach

Our approach comprises two parts, as shown in Fig. 1. Part (1) is the RAG to improve accuracy with personal knowledge. ***Contact-specific Prompt:*** Text data from mobile applications are pre-processed based on contact-based classification formed in the knowledge graph (KG). Our proposed templates gather user preferences for various categories, such as relationships and the personal interests. ***Query Generation:*** Our approach involves generating personalized outputs in response to various inputs, such as user questions, received messages, or emails. We can achieve this through the structured insertion and iterative retrieval of personalized knowledge, a key aspect of our proposed method. This step generates queries or performs embeddings to get vectors depending on input contexts and transfer them to a knowledge manager (KM) shown in Part (2). ***Supervised Finetuning:*** Our RAG approach is a hybrid method that includes LLM fine-tuning complementary to RAG. We plan on using LoRA [4] to update the LLM's weights since full fine-tuning is infeasible in an on-device environment. The proposed fine-tuning is inspired by the hypothesis that personalized LLMs, supervised by user preferences, can also have low intrinsic dimensions.

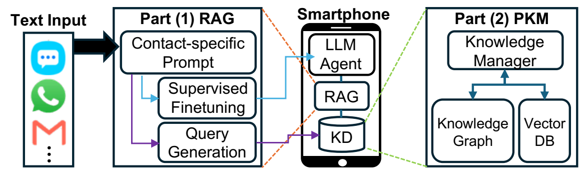

Fig. 1. An overview of our work-in-progress approach.

Part (2) is personal knowledge management (PKM) with knowledge database (KD). The KM inside KD communicates with a KG and a vector database (VD) to organize relationships with others and personal interests of the smartphone owner and maintain historical user preferences, respectively. The KM retrieves personalized knowledge based on data similarity.

III. Implementation and Evaluation

Machine learning compilation (MLC) [5] that supports various platforms and hardware is used as an LLM runtime. The MLC cross-compiles LLMs and KD source code. We are implementing our approach as an Android application that is expected to run on Samsung Galaxy S24 and include Oxigraph [6] and Neo4j Embedded [7] as a KG and a VD.

As experiments, we plan on comparing both Meta Llama2 7b and Google Gemma 2b. We will use Smart Reply data with 8,000 conversations and over 184,000 messages from Kaggle (https://kaggle.com/) and personnel text data collected from Android mobile applications as inputs to populate the KD and generate personalized responses. We will evaluate and compare outputs generated against a few tens of personal questions designed with specific contact information and interest for LLM only, RAG only, and RAG+KD cases.

References

[1] E. Spence, "Samsung looks towards AI for the Galaxy S24," 2023, Forbes.
[2] X. Shen *et al.*, "PMG: Personalized multimodal generation with large language models," in *WWW'24*. ACM, 2024, pp. 3833–3843.
[3] B. Yin *et al.*, "Heterogeneous knowledge fusion: A novel approach for personalized recommendation via LLM," in *RecSys'23*. ACM, 2023, pp. 599–601.
[4] E. J. Hu *et al.*, "LoRA: Low-rank adaptation of large language models," in *ICLR'22*, 2022.
[5] S. Feng *et al.*, "TensorIR: An abstraction for automatic tensorized program optimization," in *ASPLOS'23*. ACM, 2023, p. 804–817.
[6] (2018) Oxigraph. [Online]. Available: https://github.com/oxigraph
[7] (2013) Neoj4 embedded. [Online]. Available: https://github.com/neo4j-contrib/neo4j-mobile-android

2771-571X/24/$31.00 ©2024 IEEE
DOI 10.1109/EMSOFT60242.2024.00006