Work-in-Progress: On-device Retrieval Augmented Generation with Knowledge Graphs for Personalized LLMs

EMSOFT'24

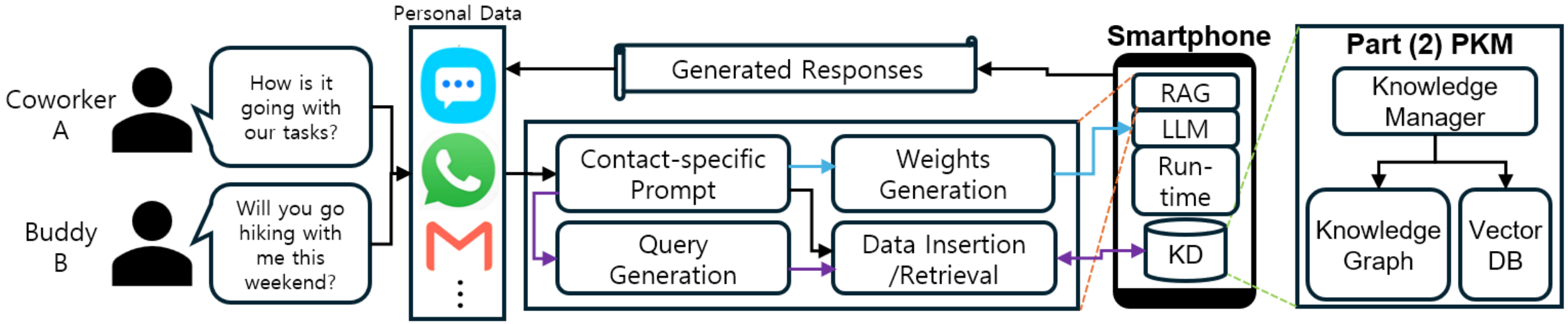
Chanhee Lee (chanijjani@gmail.com), Deeksha Prahlad, Dongha Kim, and Hokeun Kim
Arizona State University

Personalized On-device LLM

- Motivation
  - Limitations in Cloud-based LLMs
    - Data privacy issue & Network delay
  - On-device LLMs
    - **Free** from privacy issues and network delay, but limited resources
- Research Problem
  - How can personal data generated daily on smartphones be utilized to make LLMs smarter under resource constraints?

Proposed Approach

- Hybrid On-device *Retrieval Augmented Generation (RAG)* with *Knowledge Graph (KG)* and *VectorDB (VD)*
  - Hybrid = *RAG* + *Fine-tuning* with *Low-Rank Adaptation (LoRA)*
    - Underlying hypothesis: Personalized LLMs may have low intrinsic dimensions.



Evaluation Plans

- Setup

Target Device	Target Model
Android Smartphone (*Samsung Galaxy S24*)	*Llama2 7b* & *Google Gemma 2b*

- Implementation

On-device LLM Run-time	On-device KG & VD	On-device Fine-tuning
MLC LLM	*Oxigraph / ObjectBox*	*LoRA* or *MeZO**

- Evaluation Criteria
  - Dataset: Generation for various data domains/sources
    - Chatting, Calendar, E-mail
  - Metrics: Calculation of answer accuracy

*Memory-efficient zeroth-order optimizer

References

- Edward J Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” ICLR'22
- https://github.com/mlc-ai/mlc-llm
- https://github.com/oxigraph/oxigraph
- https://github.com/objectbox/objectbox-java
- https://github.com/princeton-nlp/MeZO

- + Contacts: https://labs.engineering.asu.edu/kim/

Arizona State University

Ira A. Fulton Schools of Engineering

KIM Lab