Ira A. Fulton Schools of Engineering
Arizona State University

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

EMSOFT24, Oct 1, 2024

Introduction

https://www.linkedin.com/pulse/samsung-galaxy-s24-unleashes-power-google-ai-maria-nedkova-beyqe/

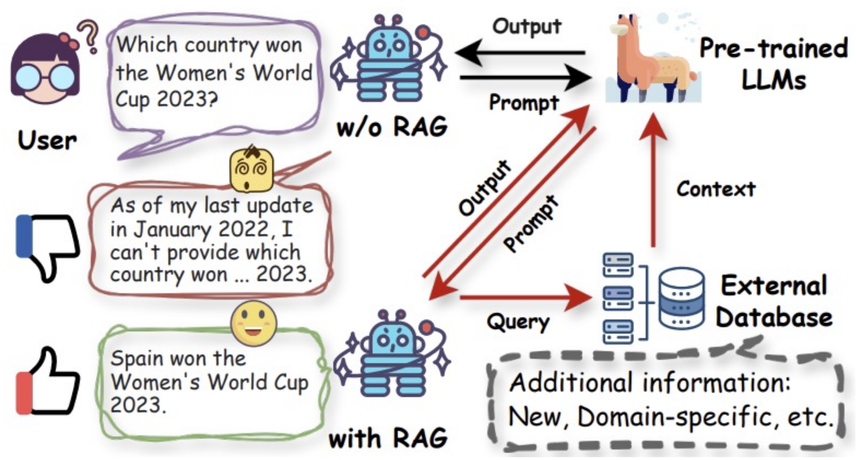
- Cloud-based LLMs have data privacy issues and latency issues with network bandwidths.
- On-device LLMs are free from the above two issues but are under limited computation and memory resources.
- Valuable personal data are generated daily, and utilizing the data efficiently can make on-device LLM smarter.
- Knowledge Graph (KG) and Vector Database (VD) are adequate in managing and analyzing the relation of text.

Research Problem: How can personal data on smartphones be utilized systematically to make LLM smarter?

Background

Retrieval Augmented Generation (RAG)

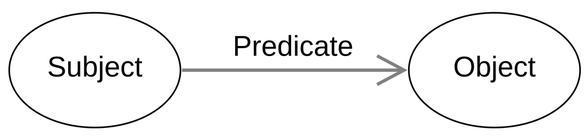
- RAG harnesses external knowledge to augment the quality of the generated content of LLMs*



*Wenqi Fan et al., "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models." KDD'24

Knowledge Graph (KG)

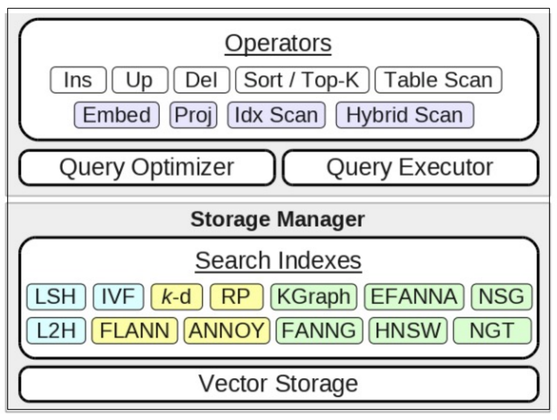
- **KG represents a network of real-world entities** such as objects, events, situations, or concepts and illustrates their relationship.
- Information is usually **stored in a graph database** and **visualized as a graph** structure.
- KG is based on the Resource Description Framework (RDF) triples



Example: "LLM"- "generate"- "texts"

Vector Database (VD)

- Database that **stores data as vectors**, numerical representations of data
- Data is **retrieved based on similarity**.



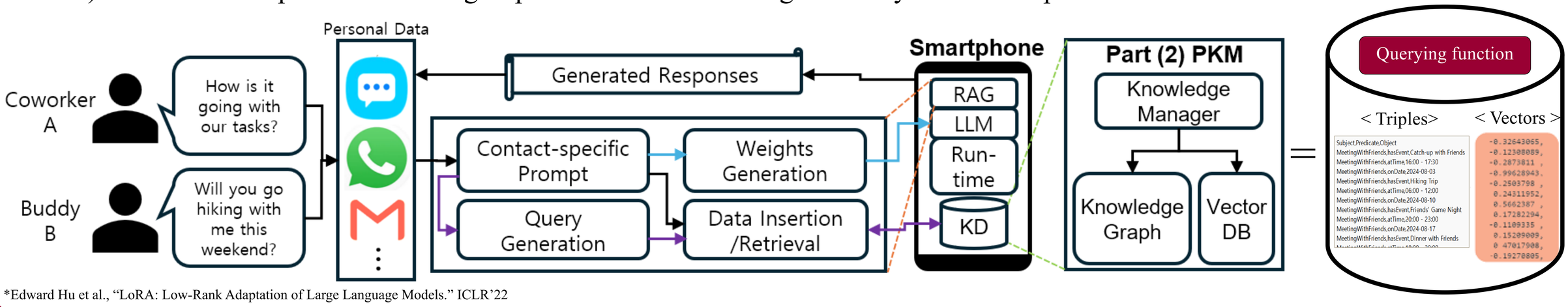
<Overview of Vector Database Management System** >

**James Jie Pan et al., Vector Database Management Techniques and Systems." SIGMOD-Companion '24

Proposed Approach

Hybrid On-device RAG with Personal Knowledge Management using KG and VD

- Hybrid On-device RAG = 1) On-device RAG + 2) On-device Fine-tuning
  - 1) On-device RAG: Merge personal data extracted by generated queries with contact-specific prompts and feed them into a LLM
    - Iterative and graph-based knowledge extraction (IGKE) with Personal Data Management (PDM)
  - 2) On-device Fine-tuning: Integrate personal knowledge into the on-device LLM using LoRA* (or derivative-free fine-tuning.)
- Personal Data Management: Classify personal data and provide Create/Read/Update/Delete operations from KG and VD
  - 1) KG data: Store exact relations of personal information as triples obtained by pre-processing various text data
  - 2) VD data: Store personal data as groups for each contact using similarity of vector representation



*Edward Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models." ICLR'22

Experiments

Setup

- LLM models are compiled, loaded, and executed by *MLC LLM* on an Android smartphone.
- KG and Fine-tuning are implemented with JNI & Java.

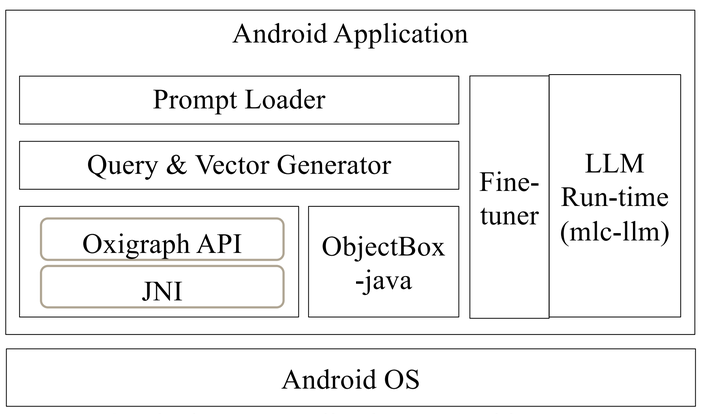
Target Device	Target Model
Android Smartphone (*Samsung Galaxy S24*)	*Llama2 7b* & *Google Gemma 2b*

On-device LLM Run-time	On-device KG & VD	On-device Fine-tuning
MLC LLM[1)]	*Oxigraph*[2)]/*ObjectBox*[3)]	*LoRA*/*MeZO*[4)]

1) https://github.com/mlc-ai/mlc-llm
2) https://github.com/oxigraph/oxigraph
3) https://github.com/objectbox/objectbox-java
4) https://github.com/princeton-nlp/MeZO

Implementation

- RAG and fine-tuning are implemented as modules in an Android application.



< On-device Personalized LLM Architecture >

Evaluation

- Dataset – Generated data for various domains (*Entertainment, Business, Sports, Family,* etc.)
- Comparison – {*Performance(Latency), Accuracy*} × {*LLM-Only, (LLM+RAG), (LLM+RAG+Fine-tuning)*}

Future Work

- Expand to multi-modal data
  : Personalized emoticon/voice

EMBEDDED SYSTEMS WEEK