LLMs as On-demand Customizable Service

Souvika Sarkar[1] , Mohammad Fakhruddin Babar[2] , Monowar Hasan[2] , Shubhra Kanti Karmaker[1]
[1] Big Data Intelligence (BDI) Lab, Auburn University
[2] School of Electrical Engineering & Computer Science, Washington State University

Motivation

- LLMs, despite their remarkable advantages, face significant challenges due to their immense size.
- These models require substantial computational resources, often lacking on local devices, hindering their accessibility and customization.
- Our solution aims to enhance LLM accessibility and utility through the following aspects: Hierarchical Organization of Knowledge, by structuring LLMs hierarchically, we distribute vast knowledge across layers based on language, application domains, and sub-domains.
- This architecture will offer, **Enhanced Customization**, **Efficient Resource Management, Scalability.**

Multi-Layer LLM Architecture - Workflow

a. The interacts with a Virtual Assistant, specifying requirements for their application.

b. The virtual assistant consults a Language Model Recommender System to recommend the most suitable model, considering user specifications and resource constraints.

c. The user clones the recommended model and fine-tunes it on their goal task using local devices.

d. Continual learning allows users to update the model with new data, ensuring it remains relevant and accurate over time.

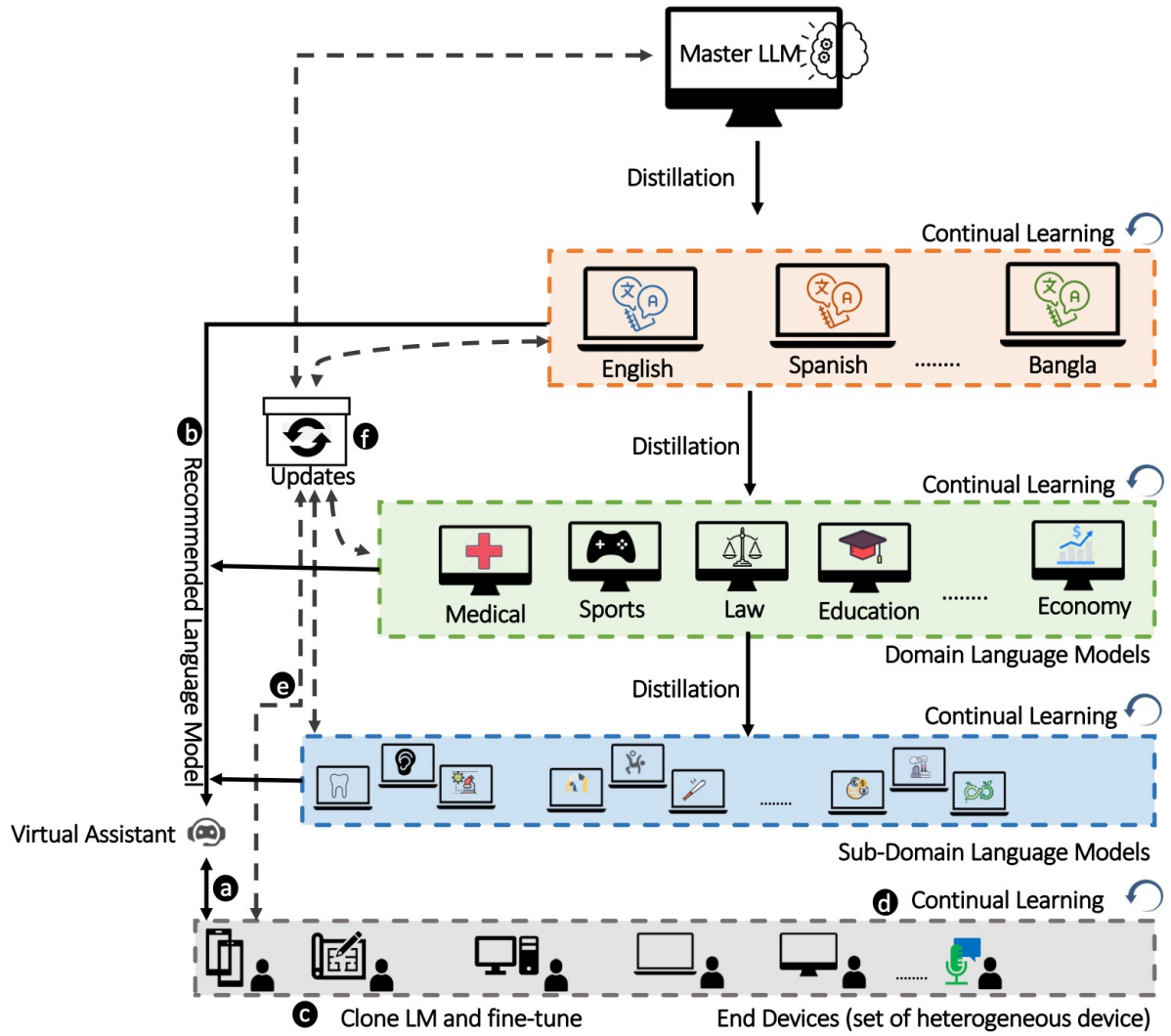
e. Peer language models are notified of updates or fine-tuning, ensuring consistency across the system.

f. Knowledge transfer mechanisms takes place for sharing of new information between language models, both upstream and downstream, enhancing the overall knowledge base of the system.

Challenges and Deployment Issues

- Challenge 1: How to identify the most suitable language model?
- Challenge 2: How to coordinate continuous updates?
- Challenge 3: How to prevent the loss of previously learned knowledge?
- Challenge 4: When should we update the parent language model?
- Challenge 5: What if a node is malicious?

Proposed Architecture



Conclusion

- This work presents a "layered" LLM architecture to tackle the challenges of deploying LLMs in practical, real-world applications.
- We believe this concept can serve as a steppingstone for implementing an open-source, customizable LLM architecture, which will foster a wider adoption of LLMs across various platforms and applications, unlocking their potential for agile performance.