Conservatoire national des arts et métiers Mahmoud Mayaleh

US PROJ 2025/2026 **Project # 13**

Lyes Khoukhi & Zakria Abouelhouda

# Implementation and evaluation of efficient methods for fine-tuning large-scale language models (LLM) in a federated learning framework

## Base Specification

The widespread adoption of large language models has transformed natural language processing, yet their deployment on edge devices faces two critical challenges. Models with billions of parameters exceed the memory and computational capacity of individual devices, requiring collaborative training across multiple nodes. Additionally, clients operate on heterogeneous task-specific data such as medical records, financial transactions, or conversational logs, making naive parameter sharing inefficient or harmful. Existing solutions address only one challenge: federated multi-task learning frameworks like MIRA provide task awareness through centralized coordination but cannot split models across devices, while peer-to-peer split learning approaches like TITANIC enable model partitioning but remain task-blind under data heterogeneity.

The primary objective of this project is to design, implement, and evaluate ATLAS, a self-organizing decentralized multi-task learning framework for large language model fine-tuning on heterogeneous edge devices. ATLAS combines gradient-based semantic client clustering with peer-to-peer split learning and cluster-wise adapter aggregation to enable collaborative fine-tuning without centralizing raw data or requiring any device to host the entire model. The framework automatically discovers task-aware semantic neighborhoods of clients with similar workloads and exploits these neighborhoods for efficient routing and aggregation.

The first phase involves comprehensive analysis of federated multi-task learning, split learning, and parameter-efficient adaptation techniques. This includes studying how MIRA leverages centralized task graphs while identifying its bottlenecks, examining TITANIC's peer-to-peer architecture and its limitations under task heterogeneity, and reviewing adapter-based methods like LoRA for decentralized aggregation. The outcome is an architectural specification defining task embeddings, clustering procedures, resource-aware pairing constraints, and hierarchical aggregation mechanisms.

The second phase focuses on implementing a simulation environment instantiating the ATLAS lifecycle. Clients compute gradient-based task embeddings from local data and send compact signatures to a server that performs clustering to identify semantic neighborhoods. The server initializes shared LoRA adapters and returns peer partnership information. During training, clients execute split learning by partitioning model layers according to memory constraints and exchange intermediate activations directly. Robustness mechanisms including intra-cluster re-pairing and inter-cluster fallback handle stragglers and client dropouts.

The final phase conducts systematic experimental evaluation measuring task-specific accuracy, communication volume, convergence speed, computational overhead, and semantic clustering quality. ATLAS will be compared against centralized training, federated multi-task baselines, and task-blind split learning to characterize the benefits of combining semantic routing with peer-to-peer model partitioning and derive actionable deployment guidelines.