Mass Fact Editing in LLMs

Project Description. While conventional databases store information in a way that is organized, deterministic, and easy to understand, LLMs store information in a way that is not discernible to humans. This raises the challenge of editing factual information in LLMs. There have been many methods developed for this task, each with its strengths and weaknesses.

This project is about *Mass fact editing in LLMs* i.e., updating LLMs with new memories to replace obsolete information or inject specialized knowledge. Why Locate Facts? We are interested how and where a model stores its factual associations, for two reasons:

i) *To understand huge opaque neural networks.* The internal computations of LLMs are obscure. Clarifying the processing of facts is one step in understanding massive Transformer networks.

ii) *Fixing mistakes.* Models are often incorrect, biased, or private, and we would like to develop methods that will enable debugging and fixing of specific factual errors.

Our work builds upon direct model editing efforts to localize and understand the internal mechanisms within LLMs, i.e., *the internal weights of an LLM are modified to reflect corrected or updated knowledge.* This involves manipulating specific neural network parameters to introduce factual corrections. It allows efficient changes to a model's knowledge.

The successful completion of this project will be to analyze existing works on model editing for dense Large Language Models and to extend these approaches to *Mixture-of-Experts LLMs.*

Project Type. 60% Engineering, 40% Research.

Internship Batch. Batch 1 from May 11 to July 10, 2025.

Duties/Activities. Read and analyze research papers in machine learning (ML). Examine and update source codes on *model editing in LLMs.* Train, fine-tune, and deploy the ML model on large-scale clusters. Implement new ideas to improve the performance and accuracy of the mass-fact editing model, and extend it to *mixture-of-experts.* Write code to collect and process various datasets. Analyze the performance of the model using multiple benchmarks. Write reports and presentations explaining the source code, datasets, and results.

Required Skills. Strong programming skills, especially in Python. Ability to work with existing code. Strong mathematical background at the bachelor's or master's level, including good knowledge of advanced topics in Calculus, Linear Algebra and Statistics.

Preferred Intern Academic Level. B.Sc. (3rd year or 4th year) or MS student.

Learning Opportunities. Students will be exposed to the exciting research in LLMs, such as ChatGPT or Fanar for arabic LLM, and will learn the details of how these models work. Students will learn different model-editing techniques, and how these approaches can be integrated into an LLM with negligible additional computational overhead. Students will gain experience in training and deploying these models on large-scale compute clusters with powerful GPUs.

Expected Team Size. It is preferable to have a team of 2 interns.

Mentors. Dr. Abdelkader Baggag