Milestone Deliverable Review Report

Deep Funding Round: 3

Project code: DFR3-RFP4

Project title: Memory-augmented LLMs: Retrieving Information using a Gated Encoder LLM (RIGEL)

Milestone number: 5

Milestone deliverable: The final project (including code, databases, and the hierarchical compression model) has been uploaded to Google Drive.
https://drive.google.com/drive/folders/1JhN0-lBGHVjara8YNTzDfRnnayeKj0Oq

Code by itself has been uploaded to GitHub.
https://github.com/mlabs-haskell/rigel

Date: 12/02/2025

Status: Accepted

Feedback (Why accepted, why rejected?):
The MLabs team successfully integrated the compressed context vectors from the previous milestone with Llama. The pipeline enables users to submit text completion queries through Rigel, where Llama processes them alongside context vectors drawn from a neural network pipeline. These vectors pass through a hierarchical compression module, retrieving relevant Wikipedia-based context before being reinjected into Llama. Users interact with this process by calling the "generate" function in the rigel.ipynb notebook.

A key challenge was determining the optimal transformer stack position for extracting and injecting context vectors. Initially, the team expected the mid-point but found better results using early-stage vectors. This required re-evaluating the hierarchical compression model to maintain performance. The final system enables calls to incorporate external knowledge beyond its training data. The complete project, including code and models, has been uploaded to Google Drive.

An important element not included in the deliverables is the model checkpoint file for Llama, as the development team does not have the rights to distribute it. However, the checkpoint file can be downloaded following the instructions in the Readme file, which also provides key setup details.

If rejected, suggested changes: