HIGH PERFORMANCE COMPUTING MODERNIZATION PROGRAM

RESEARCH PROJECT #: HPCMP-HIP-25-005

Digital Design and Manufacturing Using Large Language Models

About AFRL:

Air Force Research Laboratory (AFRL) is a scientific research organization operated by the United States Air Force Materiel Command. AFRL is dedicated to leading the discovery, development, and integration of aerospace warfighting technologies, planning, and executing the Air Force science and technology program, and providing warfighting capabilities to United States air, space, and cyberspace forces.

As part of AFRL's Materials and Manufacturing Directorate, the Digital Manufacturing Research Team studies the intersection of digital & simulation capabilities and advanced manufacturing technologies, working in areas including human-machine teaming, machine intelligence, and process-informed design. Motivated by recent advances in manufacturing processes such as additive and robotics such as collaborative robotics, this project will support the team's interests in ensuring effective teaming between humans and AI/machine learning architectures to support rapid, beneficial adoption of new designs and techniques.

RESEARCH LOCATION: Wright-Patterson AFB, OH

PROJECT DESCRIPTION:

Optimal mechanical design and manufacturing promise broad improvements to Air Force capabilities, including lightweighting and efficient operations during manufacturing scale-up. Despite advances in optimization, control, and machine learning, these techniques often still require some degree of human interaction or guidance to be of practical benefit. This project will investigate the application of natural language processing--for example, large language models (LLMs)--as a tool to bridge the gap between an optimizer/AI/controller and its human user.

Leveraging HPC resources for large-scale interaction with computationally expensive transformer networks, efforts could include prompt engineering parameter studies to uncover best practices for language models involved in control algorithms for robotic manufacturing and assembly tasks. Alternatively, benchmark tests for design/manufacturing-specific applications could also be developed, with available models tested on HPC hardware. Ultimately, this project will help illuminate the extent to which and/or methods by which language can be used to help humans interact with intelligent systems or cognitively challenging processes.

This project will consist of three phases. Depending on exactly how the research takes shape, the third phase may be completed by onsite staff following the end of the intern's short tenure.

Phase 1: Implement large language model(s) on HPC hardware:

The intern will identify an open source LLM of interest, research how to use sharing platforms such as Hugging Face, and (presumably using an interactive HIE job on an MLA node), set up a Python-based pipeline for querying the model. Estimated completion time: 2 weeks.

Phase 2: Perform prompt engineering parameter study or model benchmarking:

Using the pipeline from Phase 1, the intern will perform large scale (~thousands of queries) prompt engineering and benchmarking exercises, first creating the prompt series using automated techniques, then submitting the prompts, and analyzing results. Estimated time: 6-8 weeks.

Phase 3: (If appropriate) Demonstrate resulting best practices on RX problem:

Using the best prompting techniques or the best model identified in Phase 2, example problems generated by the Digital Manufacturing Research Team will be solved, for example, prompting LLM-based robot control schemes. Demos in the CAM

Intern activities will align with Phases 1 and 2 of the Project Plan. During Phase 1, this will include researching the mechanics of accessing the weights of open-weights transformer neural networks, and the logistical details of using HPC hardware to make inferences using the weights. Related to this will be an introductory tutorial on HPC usage.

The majority of the time will be spent on Phase 2. This will include brainstorming with RX personnel about methods of generating large sets of prompts in a manner that appropriately probes the edges of model performance related to RX applications in robotics and design. Similar development will also be required for assessing the accuracy of the results relative to desired outputs. Finally, performing the planned parameter / prompt engineering study will take place, with closing weeks devoted to results visualization and presentation.

ANTICIPATED START DATE:

May 2025 – Exact start dates will be determined at the time of selection and in coordination with the selected candidate.

QUALIFICATIONS:

The ideal candidate will be currently pursuing a degree in a field such as engineering, computer science, or machine learning, and have interest in applying digital tools to engineering problems. Demonstrated interest in natural language processing is ideal. Previous coding is experience required, with Python preferred.

ACADEMIC LEVEL:

Degree received within the last 60 months or currently pursuing:

- Master's
- Doctoral

DISCIPLINE NEEDED:

- Computer, Information, and Data Sciences
- Engineering