

Graph-based Retrieval Augmented Generation for Large Language Models

About NIWC Atlantic:

Naval Information Warfare Center (NIWC) Atlantic's mission is to conduct research, development, prototyping, engineering, test and evaluation, installation, and sustainment of integrated information warfare capabilities and services across all warfighting domains with an emphasis on Expeditionary Tactical Capabilities & Enterprise IT and Business Systems in order to drive innovation and Warfighter information advantage.

The intern will research within the Science and Technology (S&T) Department at NIWC Atlantic. The Department's research focuses on (1) the intersection between human and automated decision making (2) agile and effective operations in the radio frequency spectrum and (3) next-generation capabilities that leverage quantum phenomena.

RESEARCH LOCATION: Hanahan, SC

PROJECT DESCRIPTION:

The project aims to (1) harness the power of graph-based data structures to enhance Large Language Models (LLMs) through the development of a graph-based retrieval augmented generation (RAG) system and (2) cultivate expertise in visualizing large-scale graphs to improve the interpretability of the resulting RAG system. By integrating these two domains, this project seeks to elevate the contextual performance and reliability of LLMs while simultaneously providing robust tools for analyzing and visualizing the underlying graph structures used to generate LLM content. The outcomes of this project will be applied to large-scale graph datasets that align with the research interests of the interns and project mentor.

Overall Goal: Develop and visualize graph-based retrieval augmented generation techniques on DoD HPCMP systems applied to large-scale graph datasets.

- Literature review on graph-based RAG for LLMs and large-scale graph visualization techniques.
- Design and implement a graph-based RAG prototype for LLMs based on literature review. Leverage tools including Ollama, HuggingFace, LangChain, and llama.cpp compiled on DoD HPC systems.
- Prototype large-scale graph visualization methods on HPC systems. Integrate results into secure remote desktop (SRD), iLauncher plugin, or equivalent.
- Integrate developed graph-RAG algorithms into the HPC-hosted visualization tool.
- Conduct a case study in a domain (e.g., social networks, biological networks) based on the interest of HIP intern and/or mentor's research interest.
- Document findings and prepare end of internship technical report.

Activities:

- Onboard into the DoD HPCMP environment.
- Implement and evaluate LLMs and various retrieval augmented generation techniques (including, graph-based RAG).
- Develop and optimize graph visualization techniques on DoD HPC systems and make recommendations to the team on future integration.
- Conduct experiments and analyze results on various HPC systems and LLM architectures.
- Collaborate with project mentor and other team members.

Outcomes:

- Technical skills in how to leverage HPC for large-scale data analytics using non-traditional data sources (e.g., graphs).
- Develop technical collaboration and communication skills in a small-team environment.
- Write a final technical report describing summer experience.

Professional Development Tours:

- Participation in NIWC Atlantic "Mission Mindset Tours." Previous tours have included tours of the Charleston Robotics Operations Center (CROC), air-traffic control, Sullivan's Island BLDG 4000, etc.

ANTICIPATED START DATE:

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

QUALIFICATIONS:

Skills:

- Strong programming skills in Python
- Familiarity with machine learning frameworks (e.g., PyTorch, TensorFlow)
- Familiarity with visualization libraries (e.g., D3.js, Plotly)
- Experience with version control (e.g. git)
- Strong analytical and problem-solving skills
- Excellent written and verbal communication skills

Favorable Skills:

- Experience with agile software engineering best-practices
- Knowledge of graph theory and algorithms
- Experience with parallel computing and distributed systems
- Understanding of natural language processing concepts

ACADEMIC LEVEL:

Degree received within the last 60 months or currently pursuing:

- Bachelor's
- Master's
- Doctoral

DISCIPLINE NEEDED:

- Computer, Information, and Data Science
- Engineering
- Science & Engineering-related