# On-Premises Large Language Models (LLMs)

## Fine-tuning Large Language Models Using Proprietary Data

### ELECOMP Capstone Design Project 2023-2024

## Sponsoring Company:

***Rite-Solutions, Inc.***
One Corporate Place
Middletown, RI. 02842
http://www.rite-solutions.com

## Company Overview:

Rite-Solutions is an award-winning Veteran-Owned Small Business (VOSB) headquartered in Middletown, Rhode Island, established since April 2000. We have a stable and deep corporate history in Rhode Island, with over 380 employees with core competencies including systems engineering, software development, information technology, and cyber engineering. The stability of our business is demonstrated by both our longevity and business backlog – well over $200M in recent contract awards that will be executed over the next five (5) years. We have achieved both state and national recognition; one of our founders, Joe Marino, is an active member of the RI Science and Technology Advisory Council, along with other trusted Rhode Island business and academic leaders.

Rite-Solutions is known for innovation and dedication to the information and decision support needs of our government and commercial customers, a commitment we have coined as a company slogan, *The Information Advantage®.* We have significant US Navy prime contracts in both warfare systems and business systems development and sustainment, as well as in Information Technology infrastructure support and cyber protection. In research and development efforts, we demonstrate a proclivity for our own inventive solutions but also for finding and working with non-traditional partners, including academia and small commercial businesses, whose technologies we adapt to national research objectives in practical and meaningful ways that lead to productization.

Our corporate culture is a major part of our drive to both attract and retain people who make a difference. We take pride in our ethos of the F.E.W. (Friends Enjoying Work). Our ownership and management team work to create an open, creative, and stimulating environment. Our founders embraced the servant-leader concept from company inception, and it continues to guide our management team and approach. We recognize great ideas can and do come from anyone, not just the C-suite. To facilitate ideation in an engaging format, we built and maintain a company-wide platform called Mutual Fun that promotes corporate giving, training, and innovation in technology and our culture in a stock game format. This approach to business nurtures the employer/employee relationship and moves our teams to a model where anyone who wants to can be part of the company's growth and contribute to our success. We have won numerous awards. Each year since 2021, we have been awarded the **Providence Business News "best place to work"** among large organizations (> 150 employees) and received extensive media coverage and academic interest about our culture, including as the subject of case studies by Harvard University and Stanford University, in employee motivation and idea generation techniques. In 2022 and 2023, we had a national organization – Great Places to Work - conduct a survey where our employees could anonymously rate Rite-Solutions. The results were astounding – **96% of our employees say Rite-Solutions is a great place to work** as compared to 57% at a typical U.S. based company.

## Technical Director(s):

**To be determined.**

## Project Motivation:

Interest in Large Language Models (LLMs) such as ChatGPT and Bard is rapidly increasing because Generative AI addresses multiple challenges in diverse domains such as content development, software development, marketing, sales, customer support, sentiment analysis, personal assistant, and others to name a few. Our primary customer (DoD/Navy) has recently formed a task force to better understand the advantages and challenges of using LLMs as well as determine the best use cases for applying them across the DoD. The reason for this increasing popularity is that LLMs significantly increase performance and reduce time to perform a variety of tasks.

There are challenges associated with using LLMs, especially commercial cloud based LLMs like ChatGPT or Bard. Although beta cloud based LLMs are currently free to use, they have limitations on use and, more importantly, problems with security. To be useful for most organizations, LLMs may need to be trained or fine-tuned with proprietary data. Unfortunately, free cloud based LLMs warn against uploading proprietary or sensitive data as it could be le

ked or compromised and, in fact, this has happened. Additionally, LLMs have other limitations in that they do not always provide correct results. Known as AI hallucinations, LLMs may provide results or answers that appear to be correct but are, in fact, wrong.

To address the issue of security, one solution is to host LLMs on-premises where organizations can ensure protection of their intellectual property and digital assets. On a positive note, a large and growing number of open-source LLMs currently exist which could be hosted on-premises IT systems. However, selecting the best open source LLM appropriate for the domain and its use cases can be a challenge. Secondly, training LLMs from scratch may be prohibitively expensive. Fortunately, there are now a wide variety of open source LLMs that could be hosted on-premises that have already been pre-trained and could be fine-tuned with proprietary data. Lastly, since AI hallucinations can occur even if all the training data is correct, organizations need to develop test and evaluation practices to ensure LLM results are used appropriately by end users and stakeholders in the organization.

Rite-Solutions' goal is to significantly reduce the time, cost, and expertise required to generate draft content needed in the development of proposals and technical studies using in-house data and documentation. Like most organizations, much of Rite-Solutions corporate knowledge and intellectual property is contained in unstructured assets and stored using various formats such as MS Word documents, Adobe pdfs, PowerPoint presentations, text documents, blogs, images, and videos. LLMs provide an opportunity for Rite-Solutions to better identify and utilize its digital assets to generate higher quality content faster with less effort.

## Anticipated Best Outcome:

Rite-Solutions is seeking the development of a locally hosted, open source LLM that uses Rite-Solutions digital assets to address the following use cases:

- Support proposal development efforts such as the generation of draft content for portions of the technical, past experience, and management sections of government proposals.
- Support generation of content for whitepapers, studies, and technical reports.

To mitigate AI hallucinations, an important outcome is the development and demonstration of an approach that can be used to evaluate the results generated by the locally hosted LLM.

Lastly, Rite-Solutions would need the documentation (e.g., build scripts, installation instructions, user manual, test scripts, and open-source software) to replicate the results on Rite-Solutions internal IT infrastructure.

## Project Details:

Rite-Solutions primary goal is to use an LLM to generate proposal content based on our proprietary content. Due to security concerns, the LLM must be hosted on-premises and during normal operation, and cannot access external systems and data sources (i.e., cannot access the internet). The end users (see Figure 1) will consist of Rite-Solutions SMEs from a variety of disciplines (e.g., system, hardware, software, test, cybersecurity, IT analysts and engineers, proposal team management, contracts, cost, etc.) and roles. It is expected that the generated draft content will serve as a starting point in the proposal generation process; generated content will be further edited and enhanced by our SMEs and stakeholders.

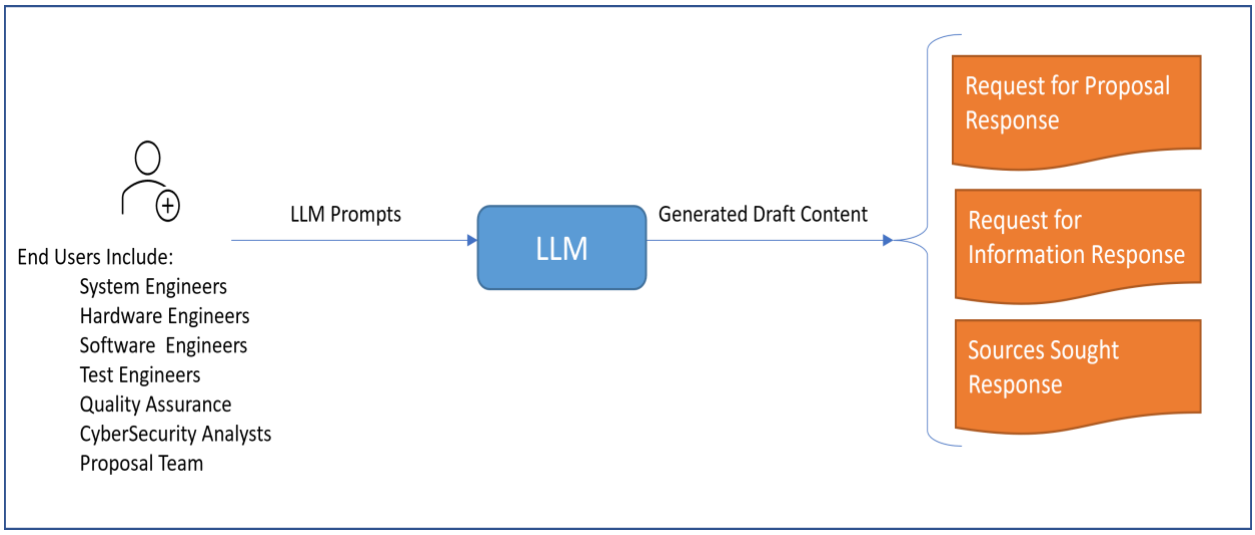


*Figure 1. LLM End User Workflow*

LLMs will be researched to identify candidates that best support this use case. The selected LLM will be fine-tuned by ingesting our proprietary data. Our proprietary data will be contained in documents using a variety of formats such as MS Word (.doc, docx), text (.txt), Adobe PDF, and MS PowerPoint (.ppt, pptx). For this capstone, data can be broadly categorized into several broad areas, namely, technical information, management approach, and past experience.

**Please note**, the capstone team will need to find a public source for sample proposal data to support the fine-tuning of the LLM and demonstrate its efficacy and capabilities. The use of proprietary data to fine-tune the LLM will only be performed by Rite-Solutions staff once the LLM has been hosted within Rite-Solutions' IT infrastructure.

A secondary goal of the LLM will be to use it to support related proposal activities such as responding to Request For Information (RFI) and Sources Sought requests by the Government. The Government often requests information from industry prior to the release of an RFP for a variety of reasons. Whatever the reason, the content of these documents is often similar to RFPs in many respects. As in the case of an RFP, use of a LLM to rapidly develop a draft RFI or Sources Sought request would be extremely helpful to determine our response as well as identify potential gaps in our capabilities.

**Software Tasks:**

The proposed approach should include the following tasks:

- Identify open source LLMs, preferably pre-trained to support content generation.
- Evaluate the identified LLMs using a Pugh matrix; the evaluation criteria and weights should be jointly developed with the Rite-Solutions mentors. Note: It is possible that the evaluation results in multiple LLM candidates similar enough that the criteria/weighting needs to be changed, and/or require that multiple LLMs be installed and further evaluated.
- Present to Rite-Solutions the recommended LLM(s).
    - If a pre-trained LLM is recommended, identify open-source datasets that could be used to fine-tune the LLM.
    - If a pre-trained, open source LLM is not recommended, provide an approach that identifies the source of publicly available software needed to train the LLM, a test approach to evaluate the LLM, resources required to perform the training, and an estimate of the time to perform the training with the identified resources.
- Install and configure the recommended LLM; it would be preferred if the LLM could be installed in a Docker container.
- Create software needed to ingest data into the LLM in various formats (e.g. .docx, pptx, pdf, text, image); use the software to ingest the data and perform the training.
- Document the approach (e.g., metrics, data to be collected, types of tests to be performed) to evaluate the quality of the LLM results.
- Perform in-depth tests of the LLM and collect metrics.
- As part of the evaluation of results, perform “Prompt Engineering”. Prompt Engineering should provide additional context to improve results or to simplify the user interface (e.g., “as the Cybersecurity SME, show me the steps we used to perform ….”); Incorporate the prompts and re-evaluate the results.

- Generate user manual, installation guide, prepare the version description document that describes the open-source software versions, build scripts, installation scripts, and tests
- Generate preliminary Final Report that describes what was performed, lessons learned, challenges, recommendations for Rite-Solutions review; incorporate review comments, and generate final report.

## Composition of Team:

1 Computer Software Engineer; 1 Data Science Engineer (2 CPE)

## Skills Required:

**Software/Data Science Engineering Skills Required:**

- Software design, development, integration, and test
- Ability to understand, evaluate, and test LLMs
- Ability to identify and install open-source software solutions
- Ability to generate and/or update scripts to support activities (e.g., build, installation, test, etc.)
- Ability to generate user manual, installation documentation, design documentation
- Experience with Linux; understanding containers would be desirable.

## Anticipated Best Outcome's Impact on Company's Business and Economic Impact

Content development for proposals and other documentation is usually expensive and often requires Subject Matter Experts (SME) with the requisite expertise to generate appropriate content. However, time and speed are often the scarce resource in these efforts. The ability to quickly generate a significant portion of a draft proposal at proposal startup using Rite-Solutions' intellectual property would be a game changer. This would allow SMEs to focus their efforts on those parts of the proposal where we don't currently have content or where we need to improve the past content to satisfy the requirements of the Government Request for Proposals (RFP). Moreover, recognizing the challenges of past knowledge management efforts, our expectation is that the use of LLMs will significantly simplify knowledge management of our IP while providing ease of access and use by members of our staff.

## Broader Implications of the Best Outcome on the Company's Industry:

Generative AI and LLMs are nascent but rapidly evolving technologies. Research and the use of these technologies have continued to accelerate in the past several years. New state-of-the-art LLMs are continually being designed and evaluated to address specific domains and use cases. Because of this, industry and Government's understanding of the advantages and limitations of LLMs have not necessarily kept pace with research. Nonetheless, early results of LLMs are very exciting and promising even though it has not been perfected. As recently as August 2023, the DoD has formed a Task Force (dubbed Task Force Lima) to better understand where and how LLMs could be used across DoD. We believe that our use case where draft content is generated by LLMs but evaluated by SMEs to eliminate AI hallucinations will be a common approach implemented by many organizations.

Rite-Solutions also recognizes that the training of LLMs using billions of parameters often costs in the millions of dollars and is beyond the ability of most organizations to fund those efforts. As a result, the approach proposed in this capstone to use pretrained models fine-tuned with proprietary data is a more affordable option for a wider variety of use cases, domains, and organizations.