U.S. ARMY COMBAT CAPABILITIES DEVELOPMENT COMMAND
ARMY RESEARCH LABORATORY
DOD SUPERCOMPUTING RESOURCE CENTER

We're Back!

Spring 2025
Newsletter

bits

Good To Know!

Need Information for ARL DSRC User Services?
dsrchelp@arl.army.mil

Users are reminded to update their current and future HPC requirements in the HPCMP pIE database.

Questions? Contact:
outreach@arl.hpc.mil

We are certainly back to full time onsite work, but as importantly, we are back to full strength with our HPC resources. TI-18 and the delivery of Betty signaled the beginning of big challenges for the ARL DSRC. TI-20 was supposed to provide relief but only made matters worse, for the ARL and ERDC DSRCs and the entire HPCMP more broadly. Betty finally became stable, allocated, and continues to perform well.

ARL DSRC B120.fulll parking lot

TI-22 arrived and brought new hope with Ruth (unclassified) and Marlyn (classified). Snyder, the consideration system provided by HPE in recognition of the delay in stabilizing Betty, was delivered and is now a little sister system to Marlyn. Ruth is allocated and a boon to our unclassified computing. TI-20 is updated with the new WEKA file system, and Jean, the unclassified element, is now allocated. Kay will be moved to classified status shortly providing three systems for rapidly growing classified requirements – Betty, Kay, and Marlyn + Snyder. Fran, our TI-24 unclassified computer, will bring us about 173,000 cores and 64 NVidia H200 GPUs this spring. The Fran test and development system, Punch, is already installed and in use. We are cautiously encouraged that the darkest days are behind us.

As we welcome the new additions on the floor, we say goodbye to SCOUT and Sofia; no more Power platforms at the DSRCs. Because of the Power9 architecture, users were slow to move codes and workflows to SCOUT, but once that hurdle was cleared, SCOUT became a workhorse for machine learning applications and consistently one of the most reliable platforms in the HPCMP. It was a novel concept, housing an HPC in a 53-foot long container and making it mobile.

Our Army Artificial Intelligence Innovation Institute (A2I2) facility is getting close to final acceptance. The first of four planned modules is nearing completion of the initial build-out and will house the new TI-24 system, Fran, in May 2025. Upon launch, the first A2I2 datacenter module will offer an infrastructure with 5 megawatts of generator power and a cooling capacity of 1,400 tons of chilled water. The facility is designed to scale up to 15 megawatts of power and 2,800 tons of cooling capacity per module. This will support the ARL DSRC's mission to continue to provide critical supercomputing resources to the Department of Defense and expand capacity and capability for current and future artificial intelligence and machine learning workloads.

UGM 2025
September 8-11, 2025

The High Performance Computing Modernization Program's (HPCMP) User Group Meeting (UGM) is a forum offering a vital opportunity for our users to connect with Program personnel and resources, and to share their work with others in the DoD science and technology (S&T), test and evaluation (T&E), and acquisition engineering (AE) communities.

Check the UGM website often for updates:
https://ugm.hpc.mil/

Contact UGM at with any questions:
ugm@hpc.mil

Generative AI Models:

Unlocking the Power of Large Language Models

Generative AI models, specifically Large Language Models (LLMs), have revolutionized the way we interact with artificial intelligence. These models have become increasingly popular and versatile, with applications across industries such as customer service, content creation, and data analysis. Many users have expressed interest in an overview of the different ways to interact with LLMs and the resources required for each use case, when it makes sense to use LLMs, and how to get started with them on high performance computing (HPC) systems.

A key challenge in generative AI is to ensure that the generated content is accurate, informative, and relevant to the user's query.

Ways to Interact with LLMs

There are several ways to interact with LLMs, for instance, text-based input/output is one of the most straightforward ways, . This approach is well-suited for applications such as chatbots, virtual assistants, language translation tools, and knowledge-based querying. The use of conversation centric interfaces allows users to interact with LLMs in a natural and intuitive way, using voice or text inputs.

LLMs can also be used for generative tasks, such as text generation, content creation, data augmentation, and planning. These tasks require interaction between a user and an LLM to iteratively work together to generate new text or data based on a given prompt or input.

A key challenge in generative AI is to ensure that the generated content is accurate, informative, and relevant to the user's query. Additionally, commercially available LLMs often lack distinct DoD conversational vocabulary, semantics, and principles. This drives the need to augment models to improve their applicability. Three main approaches exist to accomplish this:

- ***Prompt Engineering:*** If you need to generate specific text or responses based on a well-defined prompt, prompt engineering may be the right choice.
- ***RAG:*** If you need to generate text that is accurate, informative, and relevant to your specific documents and knowledge bases, RAG (Retrieval-Augmented Generation) may be the right choice. In RAG, a retrieval model is used to find relevant documents or information related to the user's query and augment the generative model, providing it with additional context and knowledge needed to produce more accurate and informative text.
- ***Fine-tuning and Training:*** If you need to tailor the LLM to a specific task or domain, fine-tuning and training may be the right choice. This approach is well-suited for applications where the user needs to update the LLM to a very specific use case or industry that is poorly represented in the original training data.

When deciding between these approaches, consider the following questions:

- What is the primary goal of the application? (e.g. generating text, answering questions, providing customer support)
- What is the nature of the input and output? (e.g. conversational, prompt-based, query-based)
- What is the level of complexity and nuance required in the generated text? (e.g. simple responses, detailed explanations, creative writing, niche forms of knowledge only you possess)

When to Use LLMs

LLMs are particularly well-suited for applications such as content generation and automation, where they can be used to generate high-quality content, such as articles, documents, language translation, and presentations. LLMs can be used to analyze large datasets and provide insights and recommendations, making them a valuable tool for data analysis and development. LLMs can also be used to generate code, provide code completion suggestions, and even help with debugging and testing. This can be particularly useful for programmers and developers who need to work with large codebases or complex systems.

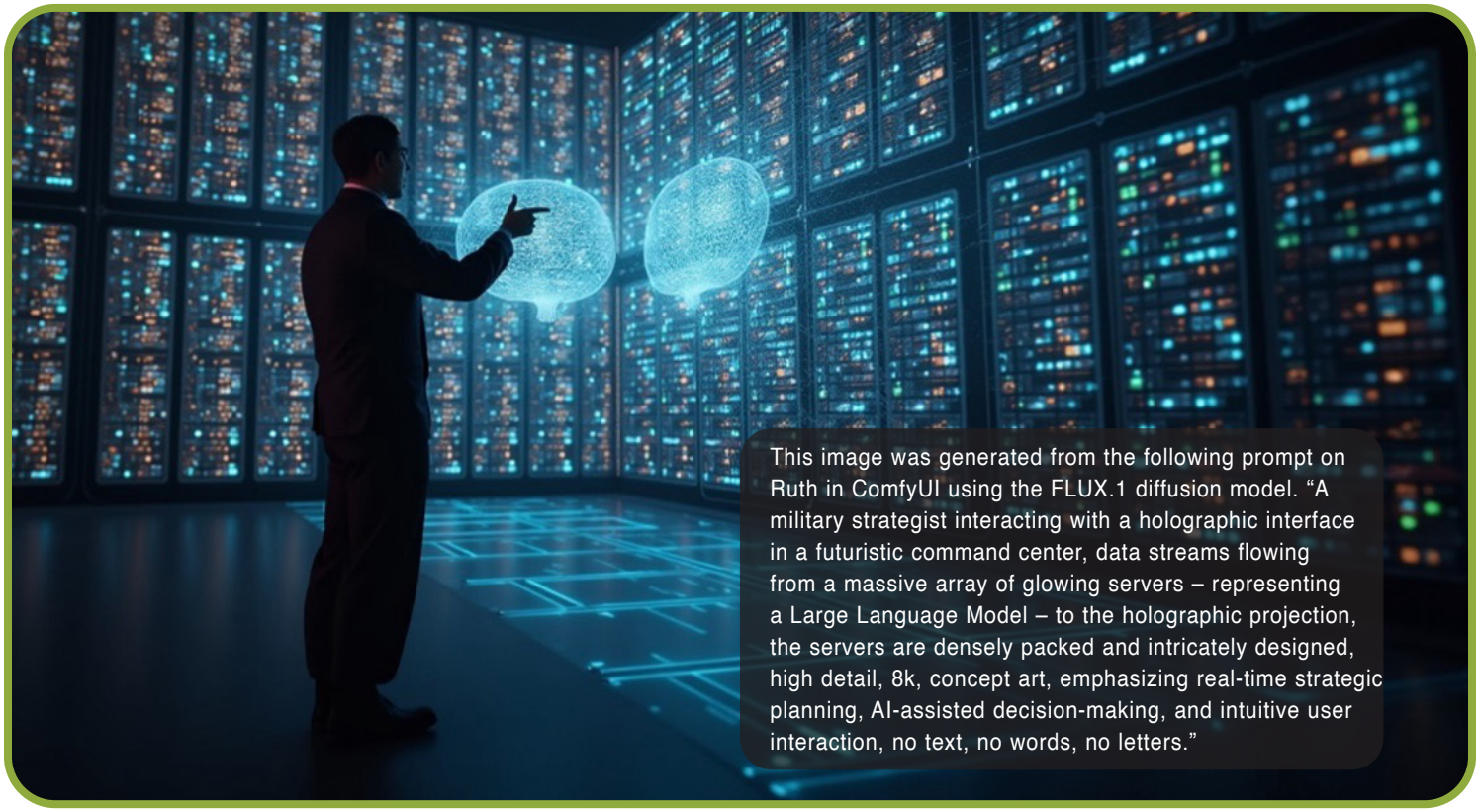
However, there are also scenarios where LLMs may not be the best choice. For instance, LLMs are not suitable for high-stakes decision-making where accuracy and precision are critical or where small errors can have significant consequences. Additionally, LLMs require large datasets and expertise to train and fine-tune, making them less suitable for domains with limited data or expertise.

Accessing LLMs: Chat vs. API

When working with LLMs, users can access them through a chat interface or by calling an API endpoint embedded in a program. The choice between these approaches depends on the specific use case and requirements of the application. A chat interface is well-suited for applications where the user needs to interact with the LLM in a conversational way, such as in a chatbot or virtual assistant. This approach provides a natural and intuitive way for the user to provide input and receive output. On the other hand, an API endpoint is better suited for applications where the user needs to integrate the LLM into a larger system or workflow, such as in a content generation or data analysis application. The

("AI LLM" continued on page 3)

Distribution Statement A. Approved for public release; distribution is unlimited.



This image was generated from the following prompt on Ruth in ComfyUI using the FLUX.1 diffusion model. "A military strategist interacting with a holographic interface in a futuristic command center, data streams flowing from a massive array of glowing servers – representing a Large Language Model – to the holographic projection, the servers are densely packed and intricately designed, high detail, 8k, concept art, emphasizing real-time strategic planning, AI-assisted decision-making, and intuitive user interaction, no text, no words, no letters."

("AI LLM" from page 2 continued)

API based use cases are an area where HPCMP resources can provide solutions.

Supporting LLM Users on Our Resources

The HPCMP has hundreds of LLM capable GPU nodes across multiple systems available for use. We recognize that LLMs can be a powerful tool for a variety of applications, but they can also be challenging to deploy and use, especially in HPC environments.

To address these challenges, the HPCMP PET program has been working on ways to support LLM users on DSRC resources. For example, PET has provided Mission Project support for a use-case where the data was too complex to automatically generate data for fine-tuning with a rules-based algorithm. Instead of fine-tuning a model, we used RAG to provide context to the LLM running on our compute nodes. However, the LLM output is only as good as the context that is provided to it, and peripheral tasks like preprocessing text from PDFs, creating labeled datasets from text, and retrieving useful context can be challenging. PET provides training webinars on fine-tuning LLMs and building RAG inference pipelines to help users get started, but we recognize that creating training content for unbounded problems like data analysis can be challenging.

In addition, the ARL DSRC data science team has developed several toolsets that customers can use to easily spin up user space LLM applications on HPC systems that span multiple nodes and multiple GPUs. One toolset provides a chatbot user interface similar to ChatGPT that utilizes either a CPU or GPU-based local LLM backend and OpenAI-compatible API to communicate between the user interface and LLM. Another toolset provides the capability to store and retrieve user documents, and to use these documents as additional references when the user asks questions to the chatbot, in a RAG approach. These toolsets have been developed and tested on our systems, and we are continually improving and expanding them to provide clean and easy-to-use solutions for LLM users.

What is an API Endpoint?

An API (*Application Programming Interface*) endpoint is a specific URL where an API receives requests for data or functionality, acting as the entry point for clients to interact with the API and access its resources.

DoD employees have access to several chat-based services such as:

- **CamoGPT**
  https://camogpt.army.mil/camogpt
- **NIPR GPT**
  https://niprgpt.mil/
- **Ask Sage**
  https://www.asksage.ai/

Conclusion

LLMs are powerful tools that can be used in a variety of ways, from text-based input/output to fine-tuning and training. By understanding the strengths and limitations of LLMs and choosing the right use case and resources, DoD scientists and engineers can unlock the full potential of these models and achieve success. Whether you're looking to automate content generation, improve customer service, or gain insights from data, LLMs are definitely worth considering and HPCMP resources are available to assist you in accomplishing your mission.

HPC Training:

To see the latest schedule and enroll, visit the HPCMP PET Training webpage:
training.hpc.mil

Webinars:
Upcoming events. (Webcast Enrollment Required)

- **HPC "New User" Training (2-day workshop)** Apr 22-23
- **AI/ML Inference Crash Course:** *From Single-GPU to Multi-Node Serviing* May 8
- **Adapting Retrieval-Augmented Generation to Domain Data** May 13 & 15
- **HPCMP New Account Orientation** May 14

If you have suggestions, problems, or need training informaition /assistance Email:
HPCtraining@hpc.mil

Contact us for more information:
www.arl.hpc.mil

email:
outreach@arl.hpc.mil

Directors Message:
Sophia Paros

Welcome to spring 2025, though the Groundhog from our northern neighbor forecasts six more weeks of winter. This is the windiest winter I remember, with so many power outages. Those facts aside, it is nearly unimaginable to recall that we were in COVID-19 mode for nearly five years, working remotely and having little in-person contact with each other. In my one day per week or so onsite up to now, my office in the 1944 vintage Army Research Laboratory Building 120 was quiet, cold, and less clean than in previous years. Beginning the week of 10 February, it was bright (all the building lights were on) and bustling; the beginning of our new normal. I work best when I am with smart people, challenging them and being challenged by them, feeding off each other's ideas and enthusiasm. I welcome ideas, suggestions, and feedback to help improve, grow, or push the ARL DSRC to become better, and to fit the needs of anyone and everyone who wants to use it.

As you can read elsewhere in this Newsletter, we are finally reaching a new stable equilibrium with our computing resources, unclassified and classified. Betty from TI-18 is consistently available with high utilization. We are in discussion with the HCPMP about extending the life of Betty for one to two years. TI-20, the Liqid systems, Jean and Kay, are updated with the WEKA file system, dedicated PSF nodes are being created for Jean, and it is in allocation. TI-22, Ruth and Marlyn, underwent successful acceptance testing and are now in allocation, along with Snyder, the consideration system for Betty. The TI-24 unclassified system, Fran, will arrive in the May but its test and development system, Punch, is already here and in use. Fran will be the first system in our new A2I2 building and the first system in the HPCMP with the latest Nvidia H200 GPUs, 64 of them. Get your ML applications ready.

Speaking of applications, Army utilization is at an all-time high across the Program. Army users are consuming their hours for high impact projects at all 5 Centers. The ARL DSRC went for too long with too few systems available and I appreciate your patience during that time. Now we are back with a vengeance, more systems than ever before on the floor, and users are responding voraciously. If you need an increase in your allocation, talk to your S/AAA. If you have a new application and want advice on which platform is best suited for it, what software is available, or if related work is being done at one of the DSRCs, talk to our Customer Success Team. Engage with us and send us examples of your successes as we build the next HPC Review highlighting the use of HPCMP resources.

The ARL DSRC is adjusting to our new way of operating. We are now responsible for day-to-day system administration of Jean, Kay, and Fran, and future TIs. Speaking of which, we have not chosen a name for TI-26. We could go back to naming TIs after Army weapon systems, or we could continue to honor people who have played significant roles for ARL and/or computing. Or we could go a different direction. I'd like to hear your ideas. What or who should we commemorate with TI-26 and future systems?

3D rendering of our newest platform, Fran, due in May 2025, and the last commemorating the original six ENIAC programmers.