Education and Support of Large Language Models in a Research Institution

Juan José García Mesa
Arizona State University
jgarc111@asu.edu

Gil Speyer
Arizona State University
speyer@asu.edu

ABSTRACT

As the capabilities of large language models (LLMs) continue to expand, with more accurate and powerful models being released monthly, researchers and educators are increasingly eager to incorporate these tools into their work. The growing demand for this technology reflects its transformative potential in natural language and its impact on scientific research. However, as more users seek to harness the power of LLMs, the need to provide comprehensive education and scalable support becomes ever more critical. Our institution has recognized this challenge and developed a support framework to educate users through regular educational events, consultations, and project support. To address the growing need for LLM support, we have implemented several key strategies, including deploying Jupyter Lab sessions using Open OnDemand for seamless HPC access and integrating cloud-based solutions via Jetstream2. We provide insights into our approach, detailing how we empower researchers and educators to leverage the capabilities of LLMs in their diverse applications.

KEYWORDS

Large Language Models, OpenOnDemand, Jetstream2

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has sparked significant interest among researchers and educators, driven by the transformative potential in natural language processing and their broad applicability across various scientific disciplines. Our institution has recognized the challenges and opportunities presented by this evolving landscape and developed a support framework that addresses the educational and technical needs of our community. In response to external inquiries and the increasing interest observed during recent presentations about our efforts, we provide an encompassing view of our practices for supporting LLMs, focusing on the strategies and technologies that enable users to harness the full capabilities of these models.

Key technologies that form the backbone of our support infrastructure include Open OnDemand [6], Jupyter Notebooks [8], and Jetstream2 [5]. In addition, we leverage Python packages such as LangChain for building scalable LLM applications, Hugging Face for accessing a vast repository of pre-trained models [10], and Gradio for creating intuitive interfaces for LLM-driven applications [1]. Combined with the technical expertise of maintainers and facilitators from our university staff, these tools create a robust ecosystem that empowers researchers and educators to explore new frontiers in their work.

Here we detail our approach, highlighting the methods and tools we employ to educate users, facilitate LLM adoption, and provide ongoing support for a wide range of projects. By sharing our experiences, we hope to contribute valuable insights to the broader community engaged in similar efforts.

2 TEACHING AND SUPPORT OPPORTUNITIES

2.1 Building Foundational Knowledge

To facilitate the effective use of LLMs, our institution conducts a series of teaching workshops throughout the year designed to introduce users to both the theoretical and practical aspects of these models. These workshops cover:

- An overview of the available HPC resources, how to request resources, and best practices for optimal usage
- Requesting a Jupyter lab session with GPUs and adequate memory using Open OnDemand
- Question-answer chatbot use cases using a Jupyter Notebook
- Retrieval-augmented generation
- How to submit batch queries for large datasets
- Fielding questions and providing guidance

Through these workshops, participants gain the necessary skills to initiate their own projects employing LLMs in their research and teaching activities. In addition, our department organizes day-long events each academic semester to display the research technology tools available and showcase success stories within our community, including a high number of cases that leverage artificial intelligence applications. Researchers from our institution who have benefited from using the available services present their work to the community.

2.2 Consultation and Project Support

Beyond workshops, we engage in collaborations with faculty and researchers, offering specialized support for projects that use advanced LLM capabilities. Our involvement includes:

- Consultation and planning: assisting in the design of experiments and the selection of appropriate models.
- Resource provisioning: allocating resources, including GPUs, ensuring that large-scale LLM experiments are executed efficiently.
- Skill development: ongoing training and mentorship to help users adapt to evolving LLM technologies.

https://doi.org/10.22369/issn.2153-4136/16/1/6

- Execution and results: in limited cases, we work alongside the researcher during the majority of the project and assist in running the experiments.

Often these projects are exploratory and having an experienced facilitator helps advance progress. Furthermore, these commitments ensure that complex LLM-based projects receive the sustained support they need to succeed.

3 FRAMEWORK

3.1 Institution-Supported Web Interface

Open OnDemand, an NSF-funded open-source HPC portal, plays a crucial role in our user support strategy. This web-based interface simplifies the use of HPC resources that require or are enhanced by a graphic interface, creating an accessible option for users to launch Jupyter lab sessions and run LLMs.

To further streamline user experience and system efficiency, we provide pre-downloaded models. This approach reduces the need for users to fetch and load models individually, which not only simplifies their workflow but also helps us, as maintainers, to troubleshoot issues more effectively. Additionally, pre-downloading models reduce the memory footprint on the system, ensuring a more stable and efficient environment for all users. In line with this effort to streamline resources, we recommend the use of permanent databases that are created once and queried multiple times, using technologies such as ChromaDB.

We also offer existing materials that implement a user-friendly interface to run question-and-answer chats [9] and retrieval-augmented generation [4]. By lowering the barrier to entry, OpenOnDemand enables a wider range of users to take advantage of our LLM support infrastructure.

3.2 Command Line Interface for Batch Queries

Jupyter notebooks provide intuitive and interactive access to LLMs. However, this approach lacks scalability. For users with command-line experience, we provide Python and sbatch template scripts to process LLM queries in parallel.

3.3 Cloud Computing with Jetstream2

Our institution also leverages cloud-based resources through Jetstream2, a national science and engineering cloud funded by the NSF and made accessible through the ACCESS [2] and Campus Champions [3] programs. Jetstream2 offers 8 petaFLOPS of supercomputing power, designed to simplify data analysis and support AI-driven research. It provides researchers with on-demand access to interactive computing resources, including a library of virtual machines and shared software for creating customized research environments. This user-friendly platform enables researchers to build personalized virtual machines or private computing systems. Key benefits include:

- Flexibility: provisioning virtual machines tailored to the specific needs of LLM projects, including custom configurations for GPU and memory.
- Availability: reduced waiting time for the compute nodes.
- Usability: Jetstream2 allows the creation of pre-defined images that users can deploy without requiring setup.
- Collaboration: facilitating collaborative efforts across institutions through shared access to cloud-based environments.

Offered as a specialty request, the integration of Jetstream2 with our existing HPC framework provides a flexible, scalable solution that complements our on-premise resources. Notably, this resource provides the means to operationalize proof-of-concept projects initially developed on Sol, ASU's flagship supercomputer [7]. Additionally, Jetstream2 offers the capability to provision sustained instances that are continually accessible, ensuring the long-term stability of research tasks.

3.4 Infrastructure

A well-known issue in supporting artificial intelligence applications is the resource-intensive nature of the workflows that translates to a high demand for resources, especially GPUs. The combined request for these resources in a large research institution can often create bottlenecks that result in long waiting times. At our institution, we have implemented several measures to alleviate this situation that focus on improving efficiency. We advise our users to choose the most fitting node partition and quality of service flag to increase the pool of possible nodes where their jobs can land. Specifically for GPUs, we offer the ability to specify different models (i.e., Nvidia A30 vs Nvidia A100 40GiB vs Nvidia A100 80GiB vs first available). We also leverage multi-instance GPU (MIG) slices, which facilitate efficient resource utilization and work well for quantized models. When applied to our Nvidia A100 GPUs, the 20GiB MIG slices allow comfortably running LLMs with up to 8 billion parameters.

4 DISCUSSION

Users of LLMs range from novices wanting to evaluate their potential applications for their research to experts training foundational models or fine-tuning existing ones. The former presents a use case where education and plug-and-play materials are key, offered by our regularly schedule workshops and by-appointment consultations. The latter, and ultimately all cases present a heavy demand for resources. We mitigate this by taking advantage of MIG slices, a curated hierarchy of node and GPU specifications, and offering consultations to researchers wanting to optimize their computations.

5 CONCLUSION

Supporting large language models at our institution, from proof-of-concept to sustained instances, requires a multifaceted effort that combines educational initiatives, sustained project collaboration, and a wide array of advanced computational resources. By leveraging tools such as Open OnDemand and Jetstream2, commonly and easily implemented in centralized systems, and an active outreach program that educates users about efficiency best practices, we empower our users to explore the full potential of LLMs in their work, contributing to the broader academic and research community.

ACKNOWLEDGMENTS

The authors acknowledge Research Computing at Arizona State University for providing computing and storage resources that contribute to the deployment and maintenance of the systems described within this paper [7].

REFERENCES

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild. arXiv:cs.LG/1906.02569 https://arxiv.org/abs/1906.02569

[2] Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. 2023. ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support. In *Practice and Experience in Advanced Research Computing 2023: Computing for the Common Good (PEARC '23)*. Association for Computing Machinery, New York, NY, USA, 173–176. https://doi.org/10.1145/3569951.3597559

[3] Marisa Brazil, Dana Brunson, Aaron Culich, Lizanne DeStefano, Douglas Jennewein, Tiffany Jolley, Timothy Middelkoop, Henry Neeman, Lorna Rivera, Jack Smith, and Julie Wernert. 2019. Campus Champions: Building and sustaining a thriving community of practice around research computing and data. In *Practice and Experience in Advanced Research Computing 2019: Rise of the Machines (Learning) (PEARC '19)*. Association for Computing Machinery, New York, NY, USA, Article 78, 7 pages. https://doi.org/10.1145/3332186.3332200

[4] Juan Jose Garcia Mesa and Gil Speyer. 2024. Supplementary materials for Enhancing the application of large language models with retrieval-augmented generation for a research community. https://doi.org/10.5281/zenodo.13328766

[5] David Y. Hancock, Jeremy Fischer, John Michael Lowe, Winona Snapp-Childs, Marlon Pierce, Suresh Marru, J. Eric Coulter, Matthew Vaughn, Brian Beck, Nirav Merchant, Edwin Skidmore, and Gwen Jacobs. 2021. Jetstream2: Accelerating cloud computing via Jetstream. In *Practice and Experience in Advanced Research Computing 2021: Evolution Across All Dimensions (PEARC '21)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3437359.3465565

[6] Dave Hudak, Doug Johnson, Alan Chalker, Jeremy Nicklas, Eric Franz, Trey Dockendorf, and Brian L McMichael. 2018. Open OnDemand: A web-based client portal for HPC centers. *Journal of Open Source Software* 3, 25 (2018), 622.

[7] Douglas M. Jennewein, Johnathan Lee, Chris Kurtz, Will Dizon, Ian Shaeffer, Alan Chapman, Alejandro Chiquete, Josh Burks, Amber Carlson, Natalie Mason, Arhat Kobwala, Thirugnanam Jagadeesan, Praful Barghav, Torey Battelle, Rebecca Belshe, Debra McCaffrey, Marisa Brazil, Chaitanya Inumella, Kirby Kuznia, Jade Buzinski, Sean Dudley, Dhruvil Shah, Gil Speyer, and Jason Yalim. 2023. The Sol Supercomputer at Arizona State University. In *Practice and Experience in Advanced Research Computing (PEARC '23)*. Association for Computing Machinery, New York, NY, USA, 296–301. https://doi.org/10.1145/3569951.3597573

[8] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, et al. 2016. Jupyter Notebooks–a publishing format for reproducible computational workflows. In *Positioning and power in academic publishing: Players, agents and agendas*. IOS press, https://ebooks.iospress.nl/doi/10.3233/978-1-61499-649-1-87, 87–90.

[9] Dhruvil Shah, Gil Speyer, and Jason Yalim. 2023. Centralized provisioning of large language models for a research community. In *Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis (SC-W '23)*. Association for Computing Machinery, New York, NY, USA, 704–707. https://doi.org/10.1145/3624062.3624147

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:cs.CL/1910.03771 https://arxiv.org/abs/1910.03771