Wenyu Zhang (wenyu.zhang@tamu.edu)
Suphanut Jamonnak (j.suphanut@tamu.edu)
and Regina Ye

Project Title: Leverage knowledge-network and Retrieval-Augmented Generation as assistive chatbot in campus digital twins

Intellectual Merit:
Publicly available Large Language Models (LLMs) such as ChatGPT (OpenAI) [1], Bard (Google), Llama (Meta) [3], and Bing Chat (Microsoft) [4] are trained on vast, complex datasets to support a wide range of tasks. These generative AI systems can produce text in response to user prompts, generating essays, poems, and other forms of content. However, despite their capabilities, LLMs can sometimes provide false information or fabricate data when they encounter gaps in their knowledge. Moreover, LLMs can be vulnerable to manipulation through malicious inputs, leading to unethical or harmful outputs. In academic institutions, the confidentiality of data and access to domain-specific knowledge are critical concerns [5]. It is essential to secure LLMs to protect user privacy and ensure they reliably support institution-specific information. This project addresses these challenges by constructing a domain-specific knowledge-constraints network using Knowledge Graphs (KGs) [6]. These KGs, integrated with Retrieval-Augmented Generation (RAG) [7], will offer fine-grained, accurate information tailored to the needs of academic institutions. Moreover, by incorporating GIS data and digital twin platform, provide information sharing across various departments. For instance, the university registrar could use this system to input course registration data, while event planners could update the academic calendar with future events. This integration will support various university operations by ensuring the availability of accurate and context-specific information.

Border Impact:
We have divided our border impact into two stages: (1) the development stage and (2) the production stage. In the development stage, hosting our Retrieval-Augmented Generation (RAG) application on the High-Performance Research Computing (HPRC) infrastructure provides several advantages. This setup allows us to fine-tune the model to meet specific institutional needs, eliminating the dependency on external services. For academic institutions handling confidential and sensitive information, this approach reduces the risks associated with transmitting sensitive data over the internet. Moreover, hosting the data on HPRC gives us full control over data processing, enabling us to meet specific standards and requirements with precision. In the production stage, the deployment of the RAG application, integrated with a digital twin, will offer a virtual representation of the campus. This not only allows users to interact with real-time events and course information but also provides spatiotemporal data, displaying where and when real-time activities occur. The digital twin will enable enhanced interaction across campus, improving navigation and access to academic information. In conclusion, this project will deploy a locally hosted Large Language Model (LLM) on the HPRC server to power a digital twin platform and real-time assistive chatbot. This system will serve students, faculty, and staff, while preventing user privacy and supporting domain-specific academic information.

References:
[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
[2] Aydın, Ö. (2023). Google Bard generated literature review: metaverse. *Journal of AI*, *7*(1), 1-14.
[3] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
[4] Narayanaswamy, A. (2024). Working with Copilot Using Bing. In *Microsoft Copilot for Windows 11: Understanding the AI-Powered Features in Windows 11* (pp. 93-99). Berkeley, CA: Apress.
[5] Mikolov, T. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
[6] Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In *The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12*(pp. 542-557). Springer Berlin Heidelberg.
[7] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

---

1. List of software's used:
    a. Jupyter Notebooks
    b. Code Editors (VS Code or Vim)
    c. Node.js
    d. QGIS and ArcGIS
    e. Anaconda with AI/ML frameworks (especially Pytorch)
2. Coding languages used:
    a. Python
    b. JavaScript
3. Will this research project benefit from (if yes, please briefly describe why)
    a. Large amounts of memory (1TB or more)?

    Yes, especially for efficiency in handling large models. LLMs require substantial memory and computational resources. It is making them capable of managing the large-scale matrix multiplications and tensor operations needed by these models.

    b. Lots of GPUs – 2, 4, 6, 8, 16 or more.

    Yes, GPUs can handle thousands of operations simultaneously, significantly speeding up tasks such as training and inference in LLMs. It is crucial for heavy computational loads associated with processing large datasets and generating responses in real-time. Furthermore, training an LLM involves adjusting millions of parameters. GPUs can drastically reduce the time required for this training phase compared to CPUs. It enables more frequent updates and refinement to our model.

    c. Lots of CPUs for parallel computing (MPI, OpenMP, etc.)

Yes, Parallel computing with many CPUs is beneficial for processing tasks that involve handling real-time spatiotemporal data in the digital twin and managing large-scale information from multiple campus departments.

d. Support from a web-based workflow engine with automation built in

No.

e. Any others, not listed above

No.

Other questions:

1. Will this project require real-time access to computing resources? HPC resources are typically orchestrated via a queuing software

No.

2. How much storage (GB, TB or PB) will the project need?

Approximately ~ 80 TB