This proposal contains the initial steps to building an Artificial Intelligence Lab based on Generative Pre-trained Transformer (GPT) technology. We will update it with a budget at a later time.

Building an AI Lab

Based on GPT technology

PI: Michael Gendron

Co-PI: Abdullah Oguz

Contents

Introduction

Artificial Intelligence (AI) is rapidly becoming one of the most critical technologies of our time. As AI evolves, it presents numerous opportunities for students to gain practical knowledge and skills while contributing to research and development in the field. Building an AI laboratory that is well-equipped and designed to enhance student learning is an essential step toward preparing students for the future of work.

A significant goal of this lab is to make **New Britain and surrounding communities AI-informed. Students in the AI curriculum will accomplish this** by engaging the community and relaying the info to the general public. Perhaps a required community service course can be part of the lab curriculum. Lastly, as we build the lab, we can garner public support by advertising that CCSU is on the ground floor of AI—and educating the general public is a good start. We should pursue a series of articles in the NB Herald and other local outlets, building community curiosity and interest in CCSU as a resource.

This concept paper presents a step-by-step guide for building an AI laboratory that educational institutions can use. It also outlines a curriculum to provide students with the necessary skills to work in AI. Additionally, this paper provides examples of potential use cases for the lab, highlighting some ways to utilize the lab.

AI is an interdisciplinary domain. We propose creating a lab and, once set up, a multidisciplinary council from the CCSU faculty to guide further use of the AI Lab. Below are our initial steps.

Steps to Create an AI Lab at CCSU

Establishing a GPT-focused AI lab at CCSU can be a great way to promote collaboration, learning, and research among students.

Here are the steps to create a practical lab:

1. **Secure funding and support:** We should contact the university administration, potential sponsors, and funding agencies to secure financial support for the lab. We must highlight the importance and potential impact of Large Language Models (LLMs) such as ChatGPT research on education, business, and society as part of those contacts.
   a. We request that CCSU provide initial funding; after that, the lab should solicit grant funding. These are some examples of potential grand funders:
      i. **National Science Foundation (NSF):** The NSF offers a variety of grant opportunities for researchers and institutions working in science and engineering. The proposers have already contacted NSF contacts to determine the level of funding available. The NSF is already reviewing a draft proposal.
      ii. **National Institutes of Health (NIH):** The NIH offers funding for research and facilities related to health and medical research. If healthcare applications become one of the specializations of the AI lab, the NIH could be a potential funding source.

iii. **Department of Defense (DOD):** The DOD offers national security and defense research funding. If the AI lab deal with applications related to these areas, the DOD could be a potential funding source.

iv. **Private foundations:** Several private foundations fund research and facilities in science and technology. For example, the Chan Zuckerberg Initiative, the Bill and Melinda Gates Foundation, and the Moore Foundation all offer funding opportunities for research and infrastructure. Other potential foundations include Alfred P. Sloan Foundation, Arnold Ventures, Intel Foundation, and IBM Foundation.

v. **Industry partnerships:** It may be possible to secure funding through partnerships with industry organizations interested in the research conducted at an AI lab. These partnerships could include technology companies, healthcare organizations, or defense contractors. OpenAI, the creator of ChatGPT, and Google, the creator of Bard, are two organizations we could approach for this partnership. The PI is already in conversations with Grammarly to determine their interest.

2. **Choose a suitable location:** The lab requires a dedicated campus space. It should have ample room for workstations, collaboration areas, and any necessary hardware.

3. **Equip the lab:** With a location secured and a budget in hand, we can proceed with equipping the lab, including setting up powerful workstations, acquiring necessary software, and providing access to relevant datasets and resources (See the next section for details).

4. **Expertise:** Drs. Gendron and Oguz will provide initial knowledge in artificial intelligence, natural language processing, and machine learning. Their guidance and mentorship are invaluable for students working on GPT-related projects. We will invite other faculty to collaborate after set up.

5. **Encourage interdisciplinary collaboration:** The AI lab must create a cooperative environment to bring together students and faculty from diverse disciplines, including but not limited to computer science, linguistics, psychology, and design. By encouraging this cross-disciplinary collaboration, the lab can leverage broader perspectives and skills, fostering innovative problem-solving and creativity in AI research and application.

6. **Create a supportive community:** A welcoming and inclusive environment is essential where students can share ideas, collaborate on projects, and learn from one another. This environment can help create a strong sense of belonging and motivation.

7. **Develop a curriculum (See Below):** An AI-focused curriculum must be designed that covers the fundamentals of AI, NLP, and machine learning, as well as advanced topics specific to LLMs and AI, such as fine-tuning, data collection, and evaluation methods.

8. **Offer workshops and training sessions:** Regular workshops, seminars, and training sessions led by experts in the field will complement the curriculum and provide practical, hands-on learning experiences. These sessions can cover the technical and ethical aspects of LLM development and deployment.

9. **Establish a Council:** An interdisciplinary council can be created to enhance innovation and collaboration from across CCSU that can maximize the use of the lab. This council can provide oversight and strategic direction for the lab.
10. **Promote research opportunities:** Students should be encouraged to participate in research projects, publish papers, and present their findings at conferences. This promotion can help establish the lab's reputation and contribute to the broader LLM and AI research community.
11. **Monitor progress and adapt:** Metrics should be developed to evaluate the lab's performance and adjust as needed regularly. These adjustments may include updating equipment, refining the curriculum, or changing research priorities based on new developments in the field.

Necessary Technology

Below is a more detailed description of the equipment and resources needed to set up an AI lab:

1. **Hardware:**
    a. **High-performance computers:** Powerful workstations with the latest CPUs and GPUs should be provided to handle intensive tasks like model training and fine-tuning. We should ensure these computers have sufficient RAM and storage for large datasets and temporary files.
    b. **Servers and storage:** A local or cloud storage solution should be set up to host datasets, models, and other resources. This storage allows for easier collaboration among lab members and ensures data security.
    c. **Networking**: High-speed internet and a secure Wi-Fi network should facilitate seamless communication and data transfer between workstations, servers, and the cloud. Build two ring network; research for PII data and open for teaching.
    d. **Peripherals**: Workstations equipped with necessary peripherals like high-resolution monitors, ergonomic keyboards, and mice to create a comfortable student working environment.
2. **Software:**
    a. **Operating systems:** Given CCSU's predominant use of Windows and macOS computers, the lab should primarily use Windows and macOS PCs, with preference given to Windows systems. The lab will consider using virtual machines if Linux becomes necessary for specific tasks. This strategy would allow Linux environments to run on existing hardware, providing the flexibility to use Linux's advanced capabilities for specific AI research tasks without a dedicated Linux machine.
    b. **AI frameworks and libraries:** The lab should ensure access to prominent AI frameworks and libraries like TensorFlow, PyTorch, and Hugging Face Transformers, which are vital for developing LLM models.
    c. **OpenAI API access:** The lab should arrange secure access to the OpenAI APIs, allowing experimentation with ChatGPT and other AI models offered by OpenAI.

d. **Development tools:** The lab needs to install Integrated Development Environments (IDEs) like Visual Studio Code, Anaconda, PyCharm, or Jupyter Notebooks to streamline coding and debugging processes.

e. **Collaboration tools:** The lab should adopt version control systems like Git and collaboration platforms like GitHub or GitLab to manage code repositories and foster teamwork.

f. **Data visualization and analysis tools:** The lab should incorporate software tools like Matplotlib, Seaborn, or Tableau, which are essential for visualizing and analyzing research data.

3. **Datasets and resources:**

a. **Access to datasets:** The lab should ensure access to publicly available and proprietary datasets pertinent to LLM research. These datasets may include text corpora, conversational data, and domain-specific collections.

b. **Data collection tools:** The lab should provide tools and resources necessary for web scraping, data annotation, and other data collection methods to help students build custom datasets for their projects.

c. **Documentation and tutorials:** The lab should maintain a library with relevant documentation, tutorials, and educational materials to facilitate students' learning about LLM and related technologies.

Development of Curriculum

The interdisciplinary council should develop a comprehensive AI lab curriculum covering foundational and advanced AI, NLP, and machine learning topics. These could be courses from existing ones at CCSU, new techniques, and a new degree. Below is a detailed outline of the critical areas to include in the curriculum:

1. **Introduction to Artificial Intelligence (AI):**
   a. History and Evolution of AI
   b. Types of AI (Narrow AI, General AI, Superintelligence, and LLMs)
   c. AI applications across various industries

2. **Fundamentals of Natural Language Processing (NLP):**
   a. Basic Concepts in Linguistics
   b. Tokenization, stemming, and lemmatization
   c. Part-of-speech tagging and named entity recognition
   d. Syntax and dependency parsing
   e. , Sentiment analysis and text classification

3. **Basics of Machine Learning (ML):**
   a. Supervised, unsupervised, and reinforcement learning
   b. Standard ML algorithms and techniques
   c. Feature engineering and data preprocessing
   d. Model evaluation and validation
   e. , Overfitting, underfitting, and model selection
   f. Application of ML principles to the training of LLMs

4. **Introduction to Deep Learning:**

a. Artificial neural networks and their components
b. Activation functions, loss functions, and optimization algorithms
c. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)
d. Basics of transfer learning

5. **Advanced NLP with Deep Learning:**
   a. Word embeddings and pre-trained language models
   b. Transformer architecture and attention mechanisms
   c. State-of-the-art models like Google's BERT (Bidirectional Encoder Representations from Transformers) and Bard, OpenAI's ChatGPT, and their variants.

6. **ChatGPT-specific topics:**
   a. Overview of the GPT architecture and its evolution (GPT-1, GPT-2, GPT-3, GPT 3.5, and GPT-4)
   b. Preprocessing and tokenization for ChatGPT
   c. Fine-tuning techniques for domain-specific applications
   d. Model evaluation metrics and best practices
   e. , Techniques for controlling and improving model output

7. **Ethics and Responsible AI:**
   a. Bias, fairness, and Transparency in AI Models
   b. Privacy and data protection concerns
   c. AI safety and robustness
   d. Responsible deployment and use of AI systems
   e. , Regulatory and policy considerations

8. **Practical applications and projects:**
   a. Building LLM-powered chatbots, virtual assistants, and Q&A systems
   b. Content generation, summarization, and translation
   c. Sentiment analysis, emotion detection, and personality modeling
   d. Domain-specific applications in healthcare, finance, law, and more

9. **Research and innovation:**
   a. Current research trends and challenges in AI, NLP, and ChatGPT
   b. Identifying research gaps and opportunities
   c. Methods for Conducting and disseminating research

Faculty would deliver the curriculum through lectures, seminars, hands-on workshops, and self-paced learning materials. By covering these topics, students will gain a solid understanding of LLMs and their related technologies, preparing them for research, development, and innovation in the field.

Project Ideas for Students in an AI Lab
Students can undertake various projects in an AI lab, from developing practical applications to conducting research. Here are some project ideas that can help students build their skills and contribute to the field:

1. **Chatbot development**: Creating a chatbot for a specific domain, such as customer support, mental health counseling, or tutoring, by fine-tuning GPT models on relevant datasets.
2. **Virtual assistants**: Developing a virtual assistant with advanced capabilities like a multi-turn conversation, context awareness, and personalized recommendations using GPT models.
3. **Summarization and content generation**: Building tools to automatically summarize long articles or generate content for blogs, social media, or marketing campaigns using GPT models.
4. **Machine translation:** Implementing a system leveraging GPT models to translate text between different languages, improving translation quality and handling idiomatic expressions.
5. **Sentiment analysis:** Fine-tune GPT models to perform sentiment analysis on user-generated content such as reviews, tweets, or forum posts and visualize the results.
6. **Emotion and personality detection:** Train GPT models to recognize and respond to different emotions or personality traits in a text to create more empathetic and personalized chatbot interactions.
7. **Question-answering systems**: Developing a system using GPT models, providing accurate and relevant responses to user queries on a specific topic or domain.
8. **Code generation and programming assistance:** Utilizing GPT models to generate code snippets, suggest improvements to existing code, or offer debugging services based on natural language inputs.
9. **Bias and fairness analysis**: Investigating and analyzing biases in responses generated by GPT models and proposing methods to mitigate them, ensuring a more fair and inclusive AI system.
10. **Controllable text generation:** Experimenting with different techniques to control the output of GPT models, such as prompt engineering, fine-tuning, or rule-based systems, to generate text that meets specific requirements or constraints.
11. **Multimodal applications:** Exploring ways to integrate GPT models with other modalities, such as images, audio, or video, to create innovative applications like image captioning, video summarization, or voice assistants.
12. **Privacy-preserving NLP:** Researching and developing techniques to ensure user privacy and data protection when using GPT models, such as differential privacy or federated learning.
13. **Evaluation and benchmarking**: Designing and conducting experiments to evaluate the performance of GPT models in various tasks, comparing them to other models, and identifying areas for improvement.

These projects can help students develop a deeper understanding of LLMs and GPT models and their capabilities while contributing to advancing the AI field. The lab will encourage students to collaborate, share their findings, and present their work at conferences, workshops, or publications to foster innovation and knowledge exchange.

Other Student and Faculty Activities in the Lab

In addition to working on projects, students can engage in various activities in the AI lab to enhance their learning experience, foster collaboration, and contribute to the broader AI community. Here are some ideas:

1. **Open House and other CCSU Activities:** Leveraging this lab as a centerpiece for CCSU open house events and admitted students' days will significantly enhance our appeal to potential and incoming students. By actively exhibiting the initiatives proposed here, we will attract more students to our institution and instill in them the confidence that CCSU is a forward-thinking, innovative choice for their academic journey. The long-term impact of this strategy will positively affect enrollment numbers and the overall reputation of our university, creating a robust return on investment that justifies funding this initiative.

2. **Business Analytics Activities:**
   a. **Data analysis:** Using ChatGPT with business data can help analyze and interpret the information. Identifying patterns, trends, or anomalies in the data may be possible.
   b. **Predictive modeling:** ChatGPT can help build predictive models based on historical data. Training the model with relevant information can predict future outcomes, such as sales, customer behavior, or market trends.
   c. **Decision support:** ChatGPT can provide insights and suggestions based on the data provided when faced with complex business decisions. It can help weigh different options, assess risks, and evaluate potential outcomes.
   d. **Data visualization:** ChatGPT can assist in creating meaningful visual representations of data, such as charts, graphs, or dashboards. It can help us understand and communicate complex information more effectively.
   e. **Natural language interfaces:** ChatGPT can serve as a natural language interface for business analytics tools and platforms. It can understand queries in plain English and provide responses or retrieve relevant information from analytics systems.

3. **Workshops and seminars:** Participating in or organizing workshops, discussions, and guest lectures to learn about the latest developments in AI, NLP, and ChatGPT, as well as explore practical applications and ethical considerations.

4. **Reading groups**: Form groups to discuss and analyze recent research papers, books, or articles related to GPT models and their associated fields, promoting a deeper understanding of state-of-the-art.

5. **Hackathons and coding sessions:** Organizing hackathons, coding sessions, or project-based competitions to encourage teamwork, problem-solving, and rapid prototyping of ideas using LLMs.

6. **Collaboration with other labs and institutions:** Establishing partnerships with other university labs, research institutions, or industry partners to share knowledge, resources and collaborate on joint projects.

7. **Community outreach:** Conduct outreach programs to educate the public about AI, NLP, and LLMs, such as workshops for local schools, presentations at community events, or volunteering to help non-profits implement AI solutions.
8. **Mentoring and peer tutoring:** Offering mentoring and peer tutoring programs to help students with different skill levels and backgrounds learn from each other, fostering a supportive and inclusive environment.
9. **Participation in conferences and events:** Attending or presenting at AI, NLP, or LLM-focused conferences, workshops, and symposia to network with other researchers, learn about new developments and share insights from lab projects.
10. **Open-source contributions:** Contributing to open-source AI and NLP projects, such as AI frameworks, libraries, or tools, to improve skills, give back to the community, and build a portfolio.
11. **Blogging and documentation:** Writing blog posts, tutorials, or creating video content about lab projects, insights, or experiences to share knowledge and inspire others in the AI and ChatGPT community.
12. **Networking and professional development:** Engaging in networking events, industry talks, or webinars to connect with professionals in the field, seeking potential internships or job opportunities, and staying informed about industry trends.

By participating in these activities, students can enhance their skills, knowledge, and professional network while contributing to the growth and development of the LLMs and AI community.

Exploring the Applications and Advantages of an AI Lab:
Enhancing Research, Learning, and Collaboration
The AI lab can facilitate research and workshops in many areas, including:

1. **Natural language processing research**: GPT models can be a benchmark for natural language processing tasks such as text generation, question answering, and language translation. Researchers can fine-tune the model on their specific dataset to improve its performance and gain insights into the workings of neural language models.
2. **Educational tool:** ChatGPT can be integrated into a learning management system (e.g., Blackboard) or a virtual classroom to provide personalized learning experiences to students. ChatGPT can answer questions, explain, and generate examples based on course material. This tool, which will require development, can help students understand the concepts better and engage with the material more interactively.
3. **Virtual assistant:** A virtual assistant (VA) from ChatGPT for administrative tasks such as answering frequently asked questions, scheduling appointments, and providing information about university services. This VA can free up time for administrative staff to focus on more complex tasks and improve the efficiency of the university.
4. **Research tool:** ChatGPT can generate synthetic data for research purposes. This generation can help overcome privacy concerns associated with using accurate data and allow researchers to explore different scenarios without the risk of harming individuals or

businesses. For example, researchers can use ChatGPT to generate realistic medical records, financial data, or customer reviews.

5. **Student Assistant:** ChatGPT can be a valuable resource for students in a university lab, offering a range of applications that can improve research efficiency, learning, and collaboration. Here are some potential applications:
   a. **Research assistance:** ChatGPT can help students conduct literature reviews, find relevant information, and develop research questions. Students can ask ChatGPT-specific questions related to their research interests and receive appropriate information and guidance.
   b. **Learning tool:** ChatGPT can be used as a learning tool to supplement course material or help students understand complex topics. Students can ask ChatGPT questions related to their coursework and receive additional information or explanations.
   c. **Writing assistance:** ChatGPT can help students improve their writing skills. Students can use ChatGPT to generate topic ideas, refine thesis statements, and receive feedback on their writing.
   d. **Language learning:** ChatGPT can be used as a language learning tool, helping students practice reading, writing, and speaking skills. Students can converse with ChatGPT in a foreign language and receive feedback on their language proficiency.
   e. **Virtual study partner:** ChatGPT can be a virtual study partner for students, helping them review course materials, ask questions, and even provide quiz questions.

6. **API Programming:** API programming can be a valuable skill for students in a university lab, particularly for those studying management information systems, business analytics, computer science, or related fields. **Overall, API programming can be a valuable skill for students in a university lab, offering a range of applications that can improve software development, data analysis, web development, IoT, and collaboration.** Here are some potential applications of API programming:
   a. **Software development:** API programming is essential to the 21st software application development for personal projects, as part of the coursework, and in industry. Students can learn how to design and implement APIs to connect different components of an application.
   b. **Data analysis:** Accessing and analyzing data from external sources is facilitated by API programming. Students can learn how to use APIs to retrieve data from social media platforms, government databases, and other sources for research or analysis purposes.
   c. **Web development:** Web development depends on API programming to communicate with servers and other online services. Students can learn how to use APIs to retrieve data, authenticate users, and perform other tasks within a web application.
   d. **Internet of Things (IoT):** IoT interaction is often performed through API programming to build applications that interact with IoT devices, such as smart

home devices or wearables. Students can learn how to use APIs to control and access data from these devices.

    e. **Collaboration:** Integration with other tools is often best performed utilizing API programming to develop applications communicating with other devices and services. Students can learn how to build APIs that enable communication and data sharing between different applications.

7. **Building a Private, Localized LLM for Document Interaction:** A unique project regarding a localized LLM can be initiated inside the AI Lab. There are two steps to building this localized LLM.

    a. **Step 1: Preparations for knowledge base and vector database**
- Compiling a list of documents to be used as our knowledge base.
- Breaking these large documents into smaller chunks. These chunks can consist of approximately 500 words.
- Using an embedding model to create an embedding for each document chunk.
- Establishing a vector database that houses all the embeddings of the documents.
- When documents are added to the knowledge database in the future, we should update our vector database correspondingly.

    b. **Step 2: Local LLM Workflow**
- A user inputs a prompt into the user interface.
- The application employs the embedding model to generate an embedding from the user's prompt, then sends it to the vector database.
- The vector database returns a list of document chunks relevant to the user's prompt based on the similarity of their embeddings.
- The application then creates a new prompt combining the user's initial prompt and the fetched document chunks as context and forwards it to the LLM.
- The LLM then generates a response with references from the context documents. This response is displayed on the user interface, along with its sources.

Budget

The IT department is preparing the budget under George Claffey's direction.

Proposed Timeline

Fall 2023	
Claffey, Gendron, Oguz	Interdisciplinary Committee
- Gendron and Claffey finalize planning and present to deans, provost, and president for final approval. - Claffey and the IT department secure room, furniture, and hardware. - Gendron and Oguz oversee the building of the lab installation of hardware/software. - Create draft proposals for the new curriculum.	- Create an interdisciplinary committee to oversee the AI Curriculum - Gendron will be the initial chair. - Create curriculum for submission in November. - Once courses and curriculum are approved, create course shells. - Marketing starts for the new program.
Spring 2024	
Build IVAI LLM for the CCSU Website (proof of concept).	Plan for the new program
Fall 2024 - Start to offer the new program.	