Open Source Foundation Models for Educational AI

Northeastern University, with its vast human and data capital, is asking for computing resources to allow us to build open-source Foundation Models for Educational AI. AI Skunkworks at Northeastern University is engaged in volunteer open-source AI projects, and stands at the forefront of AI education and application. Our philosophy of "Learn AI by doing AI" underpins initiatives like AI Skunkworks, Teaching & Learning xAI, and AI4ED, where we embrace real-world project-based learning to expand knowledge and portfolios while educators and students build practical, ethical and useful Educational AI tools. Despite our rich resources, the primary challenge in advancing Foundation Model (FM) development has been the limited access to cloud computing resources.

In response to this, and through our expansive network including 13 campuses worldwide and a commitment encapsulated in projects such as Teaching & Learning xAI and AI Skunkworks, we are gearing up for significant contributions to the Foundation Model Development Call for Proposals — Winter 2024. Our focus will include:

- Educational Dialogue System Using Reinforcement Learning with Human Feedback: Aiming to refine educational dialogue systems for enhanced student interaction.
- Adaptive Learning Platform Enhancement Through Distillation with Enhanced Reasoning: Personalizing learning based on individual styles and needs.
- Bias Detection and Mitigation in Student Performance Prediction Models: Ensuring fairness in educational AI applications.
- Efficiency Enhancements in Educational Content Generation: Streamlining the creation of diverse educational materials.
- Foundation Models for Multimodal Learning Experiences: Developing FMs for novel educational modalities like AR and immersive simulations.

Northeastern's unique position, with its extensive network and innovative educational projects, is poised to make significant contributions to the field of educational AI through this proposal. Our initiatives not only aim to enhance the AI curriculum but also strive to tackle the limitations posed by computing resource access, demonstrating our leadership in integrating AI into education and our commitment to advancing foundation model development

Teaching & Learning xAI

The Teaching & Learning xAI project at Northeastern University represents a groundbreaking initiative in higher education, leveraging generative AI to transform teaching and learning methodologies. Led by Connie Yowell, Senior Vice Chancellor for Educational Innovation, and overseen by Project Sponsor David Madigan, this project aims to create a collaborative space for NU students, faculty, and staff to develop and refine AI-driven educational approaches. It underscores Northeastern's leadership in integrating AI into academia and emphasizes the importance of trust and interpretability in AI applications to ensure their ethical integration into educational settings across its global network.

The initiative benefits from the diverse expertise of its leaders. Yowell's experience in innovating learning models, combined with Madigan's analytical and strategic insights from his academic roles, sets a solid foundation for the project. Project Manager Tor Ekstrom brings critical project management skills from healthcare and research, driving the initiative towards making education more personalized and engaging.

As it progresses, the Teaching & Learning xAI project is set to redefine educational paradigms by marrying technology with human creativity, illustrating a future where education is more adaptive, interactive, and impactful.

AI Skunkworks

AI Skunkworks, spearheaded by Nik Bear Brown at Northeastern University, stands as an experiential AI playground with a mission to leverage technology and AI in teaching research, creativity, and entrepreneurial thinking and skills. This innovative experiential AI program aims to blend technology and artificial intelligence to foster research, creativity, and entrepreneurial skills. At its core is the philosophy of "learning AI by doing AI," encouraging students to engage in hands-on projects that apply AI knowledge to real-world scenarios. The program underscores the importance of mentorship by linking experienced AI practitioners with novices for collaboration on open-source projects, thereby enriching the learning experience. It also prides itself on a vast community of thousands of students and mentors, dedicated to AI research and practical applications, further demonstrating its achievements and contributions to the AI field through its extensive repository. This initiative not only equips students for industry challenges but also seeks to address societal issues through AI, marking its leadership role in educational AI integration. A comprehensive list of its accomplishments can be found at AI Skunkworks.

Northeastern University hosts thousands of master's students and faculty who actively engage in volunteer open-source projects, a testament to the university's commitment to experiential learning in artificial intelligence. This participation serves to broaden their knowledge base and enhance their portfolios through hands-on experience with real-world projects.

Details of the AI Skunkworks project can be found here https://github.com/nikbearbrown/AI_Skunkworks

The AI for Education project (AI4ED)

The AI for Education Project (AI4ED) at Northeastern University, as part of AI Skunkworks, exemplifies the integration of AI into education, aiming to make learning more adaptive, interactive, and student-centered. This initiative focuses on creating open-source tools with AI technologies like Large Language Models (LLMs) and Generative AI, bridging the gap between research and practical application. A key feature is enabling institutions to customize their LLMs, offering autonomy and reducing costs. The AI for Education Mentor tool, for instance, provides personalized teaching experiences and analytics.

This project seeks to revolutionize education by making it more tailored to individual needs, employing AI to enhance teaching and learning. It spans various disciplines, aiming to make education more efficient

and accessible through AI. Informed by educational technology, AI, and human-computer interaction research, it applies personalized learning theories to adapt content to each learner's pace and style.

The project's goals include enhancing expertise in AI educational tools, fostering international collaborations, and innovating curriculum development. Expected outcomes include the development of AI tools, enhanced curriculum offerings, scholarly publications, and international partnerships. In essence, AI4ED aims to transform educational paradigms, inviting collaboration to create a future where AI and education evolve together for transformative learning experiences.
Details of the AI for Education (AI4ED) project can be found here
https://github.com/nikbearbrown/AI4ED

Northeastern's Network

Northeastern is in a singular position to work on a diverse range of applications of AI to Education. Northeastern University operates a network of 13 campuses across the U.S., U.K., and Canada, offering a range of unique experiences, opportunities, and connections to enrich and inform learning and research. Additionally, the university's academic offerings are extensive, encompassing more than 290 majors, demonstrating the university's comprehensive academic structure.

Foundation Models for Educational AI Projects

Educational Dialogue System Using Reinforcement Learning with Human Feedback

- Objective: Develop an FM that improves dialogue systems in educational contexts, capable of engaging students in meaningful conversations, providing tutoring, and answering factual questions related to course content.
- Approach: Use reinforcement learning with human feedback to refine the model's ability to understand and respond to student inquiries accurately, reducing incorrect or nonsensical answers, and asking clarifying questions in cases of ambiguity.

Adaptive Learning Platform Enhancement Through Distillation with Enhanced Reasoning

- Objective: Enhance adaptive learning platforms by integrating an FM that personalizes educational content and assessments based on individual learning styles, performance, and preferences.
- Approach: Employ model distillation techniques to create lightweight versions of complex FMs capable of enhanced reasoning, making them suitable for real-time educational applications without compromising on the quality of content personalization.

Bias Detection and Mitigation in Student Performance Prediction Models

- Objective: Develop and integrate models that predict student success and identify at-risk students early, with built-in mechanisms for detecting and mitigating bias.

- Approach: Leverage novel datasets and training methods to build FMs that can impartially analyze student data, ensuring fairness and equity in educational interventions and support mechanisms.

Efficiency Enhancements in Educational Content Generation

- Objective: Create an FM that generates high-quality, personalized educational content, including lecture notes, quizzes, and summaries, with a focus on efficiency enhancements across the training and hosting lifecycle.
- Approach: Explore novel training methods and scaling laws to develop models that can generate diverse educational content efficiently, reducing the computational resources required for training and hosting.

Foundation Models for Multimodal Learning Experiences

- Objective: Develop FMs that support novel modalities in education, such as interactive simulations, augmented reality (AR) for hands-on learning, and image-based content analysis.
- Approach: Utilize foundation models in novel domains like biology and engineering to create immersive and interactive learning experiences, enhancing student engagement and understanding of complex concepts.

Using AWS to Build Educational Dialogue System Using Reinforcement Learning with Human Feedback

This project aims to create a Foundation Model (FM) that enhances dialogue systems for educational purposes, ensuring they can engage students effectively. Here's a detailed approach utilizing AWS tools:

1. Project Setup and Data Collection

- Amazon S3: The AI for Education project (AI4ED) should use Amazon Simple Storage Service (S3) to store and organize its training data, including transcripts of educational dialogues, feedback data, and more.

2. Data Preparation and Processing

- AWS Glue: The AI for Education project (AI4ED) can employ AWS Glue for data preparation and loading its datasets into a suitable format for training. AWS Glue can help clean, normalize, and transform data to improve the quality of the training dataset.
- Amazon SageMaker Ground Truth: The AI for Education project (AI4ED) should utilize SageMaker Ground Truth to label its data effectively, creating accurately labeled training datasets for the dialogue system, using both machine learning and human annotators.

3. Model Development and Training

- Amazon SageMaker: The AI for Education project (AI4ED) should use Amazon SageMaker for developing and training its reinforcement learning model. SageMaker provides a fully managed environment that can scale to the project's needs.
    - SageMaker RL: SageMaker's reinforcement learning environments and pre-built algorithms can be utilized by the AI for Education project (AI4ED) to train its model, incorporating human feedback loops to refine responses.
    - SageMaker Neo: To optimize its model for efficient deployment, the AI for Education project (AI4ED) can use SageMaker Neo, which compiles models to run more efficiently on target hardware, such as AWS Trainium chips.

4. Human Feedback Loop Integration

- Amazon Augmented AI (A2I): The AI for Education project (AI4ED) should integrate Amazon Augmented AI to easily implement human review and feedback into its ML workflow, critical for refining the model based on human assessments of its dialogue capabilities.

5. Deployment and Scaling

- Amazon EC2 with AWS Trainium Chips: The AI for Education project (AI4ED) should deploy its trained model on EC2 instances powered by AWS Trainium chips for cost-effective, high-performance inference, optimized for machine learning workloads.
- Amazon Elastic Kubernetes Service (EKS) or Amazon SageMaker: For managing deployments and scaling, the AI for Education project (AI4ED) can consider using Amazon EKS or SageMaker's hosting capabilities, offering robust solutions for deploying and managing containerized applications.

6. Monitoring and Iteration

- Amazon CloudWatch: The AI for Education project (AI4ED) should use Amazon CloudWatch to monitor the application's performance and system health, enabling iterative improvements based on real-world usage data.
- AWS Step Functions: The AI for Education project (AI4ED) can automate its workflow for retraining and redeploying the model with AWS Step Functions, allowing for seamless updates to the dialogue system as new data and feedback are collected.

By following this approach and leveraging AWS's comprehensive suite of services, the AI for Education project (AI4ED) can develop an Educational Dialogue System that is scalable, efficient, and responsive to the needs of students and educators, integrating human feedback throughout the development and deployment phases to ensure the system remains relevant and effective in educational contexts.

Using AWS to Build Adaptive Learning Platform Enhancement Through Distillation with Enhanced Reasoning

To develop an Adaptive Learning Platform Enhancement with Distillation and Enhanced Reasoning for The AI for Education project (AI4ED), leveraging AWS's comprehensive tools and services can provide an effective pathway. This initiative aims to integrate a Foundation Model (FM) to personalize educational content dynamically. Here's a structured approach using AWS tools:

1. Initial Data Collection and Management

- Amazon S3: The AI for Education project (AI4ED) should utilize Amazon Simple Storage Service (S3) for storing and organizing diverse educational content, including student performance data, learning materials, and feedback.

2. Data Labeling and Preparation

- Amazon SageMaker Ground Truth: The AI for Education project (AI4ED) can use SageMaker Ground Truth for labeling the data accurately. This tool facilitates the creation of high-quality training datasets by combining machine learning and human judgment, crucial for developing personalized learning experiences.

3. Model Development and Training

- Amazon SageMaker: For the development and training of the foundation model, the AI for Education project (AI4ED) should leverage Amazon SageMaker. It offers a fully managed service that simplifies the process of building, training, and deploying machine learning models at scale.
    - Model Distillation: Within SageMaker, the AI for Education project (AI4ED) can implement model distillation techniques. This involves training a smaller, more efficient model (the "student") that mimics a larger, pre-trained model (the "teacher") to ensure the distilled model retains the capability for enhanced reasoning with reduced computational demands.

4. Enhancing Model Reasoning and Personalization

- Amazon SageMaker Neo: To further enhance the efficiency of the distilled model, the AI for Education project (AI4ED) can utilize Amazon SageMaker Neo. This tool optimizes the model to run more efficiently on diverse hardware, including AWS Trainium chips, without sacrificing performance or accuracy.

5. Deployment and Real-Time Application

- AWS Elastic Compute Cloud (EC2) with AWS Trainium Chips: For deploying the distilled model, the AI for Education project (AI4ED) should use EC2 instances powered by AWS Trainium chips. These instances are optimized for machine learning inference tasks, offering the necessary computational power for real-time educational applications.

6. Continuous Learning and Model Improvement

- Amazon Augmented AI (A2I): To continuously improve the model based on real-world feedback, the AI for Education project (AI4ED) can integrate Amazon Augmented AI. This service allows easy incorporation of human reviews into the ML workflow, crucial for personalizing educational content effectively.

7. Monitoring and Analytics

- Amazon CloudWatch: The AI for Education project (AI4ED) can use Amazon CloudWatch for monitoring the deployed models and applications. It provides detailed insights into application performance, enabling timely adjustments and enhancements.

By employing this structured approach and utilizing AWS's suite of machine learning and AI services, the AI for Education project (AI4ED) can effectively enhance adaptive learning platforms. This initiative will not only personalize educational content and assessments more dynamically but also ensure the scalability and efficiency of real-time educational applications, thereby significantly improving learning outcomes.

Using AWS to Build Bias Detection and Mitigation in Student Performance Prediction Models

To build Bias Detection and Mitigation in Student Performance Prediction Models for The AI for Education project (AI4ED) utilizing AWS tools, a detailed approach is required to ensure fairness and equity in educational interventions. Here's how AWS services can support the development of such models:

1. Data Collection and Storage

- Amazon S3: The AI for Education project (AI4ED) should use Amazon Simple Storage Service (S3) to securely store and organize large volumes of student performance data, including grades, participation metrics, and feedback, ensuring data privacy and accessibility.

2. Data Preparation and Bias Detection

- Amazon SageMaker Data Wrangler: For preprocessing and cleaning the data, the AI for Education project (AI4ED) can utilize Amazon SageMaker Data Wrangler. This tool helps in identifying potential biases within the dataset through its analysis features.
- AWS Glue: To automate data preparation tasks and transform data into a machine learning-compatible format, AWS Glue can be employed, facilitating efficient data handling and processing.

3. Model Development and Training

- Amazon SageMaker: The AI for Education project (AI4ED) can develop and train bias detection models using Amazon SageMaker, which provides a broad set of built-in algorithms and support for popular machine learning frameworks.

- Bias Detection Algorithms: Within SageMaker, the project can leverage pre-built bias detection algorithms or develop custom algorithms to identify and mitigate bias in student performance data.

4. Bias Mitigation

- Amazon SageMaker Clarify: To detect and mitigate bias, Amazon SageMaker Clarify can be used. It provides insights into the model's predictions, helping to understand and reduce bias by identifying which attributes in the data contribute most to the predictions.

5. Deployment and Monitoring

- Amazon EC2 Instances with AWS Trainium Chips: For deploying the trained models, the AI for Education project (AI4ED) should use EC2 instances powered by AWS Trainium chips, optimized for machine learning inference tasks, ensuring high performance and cost-efficiency.
- Amazon CloudWatch: For ongoing monitoring of the model's performance and bias metrics, Amazon CloudWatch can be utilized. It enables the project to track operational metrics and logs, ensuring the model continues to perform fairly and accurately.

6. Continuous Improvement and Feedback Loop

- Amazon Augmented AI (A2I): To continuously improve the model and further mitigate bias, the AI for Education project (AI4ED) can integrate human reviews into the machine learning workflow using Amazon Augmented AI. This allows for manual review of predictions and adjustments based on human judgment.

By leveraging these AWS tools and services, the AI for Education project (AI4ED) can effectively develop and integrate models that predict student success while ensuring bias detection and mitigation. This approach not only promotes fairness and equity in educational interventions but also aligns with the project's commitment to ethical AI development and use.

Using AWS to Build Efficiency Enhancements in Educational Content Generation

To build Efficiency Enhancements in Educational Content Generation for The AI for Education project (AI4ED) using AWS tools, a comprehensive approach is vital for creating a Foundation Model (FM) that not only generates high-quality, personalized educational content but also focuses on efficiency. Here's an expanded detail on how to utilize AWS services for this objective:

1. Data Collection and Management

- Amazon S3: The AI for Education project (AI4ED) should use Amazon Simple Storage Service (S3) for storing vast amounts of educational content, including raw educational materials, lecture notes, and quizzes. S3’s scalability and data management features make it ideal for handling large datasets required for training efficient FMs.

2. Data Labeling for Personalization

- Amazon SageMaker Ground Truth: To ensure the educational content generated is personalized, The AI for Education project (AI4ED) can leverage Amazon SageMaker Ground Truth. This service provides accurate data labeling to train models on recognizing various educational needs and preferences.

3. Model Training and Development

- Amazon SageMaker: SageMaker will be central to developing and training the FM. It supports a wide range of machine learning algorithms and frameworks suitable for generating educational content.
    - Model Optimization: Use Amazon SageMaker Neo to optimize the model for different hardware, ensuring the best performance across AWS Trainium chips for both training and inference phases.

4. Efficiency Enhancement

- Utilizing AWS Trainium Chips: For training the FM, The AI for Education project (AI4ED) should use EC2 instances powered by AWS Trainium chips, designed to provide high performance at lower costs.
- Exploration of Scaling Laws: The project should explore scaling laws within the training process, using Amazon SageMaker's capabilities to adjust model size and training data volume to find the most efficient training methodology.

5. Content Generation and Hosting

- AWS Lambda and Amazon API Gateway: Once the FM is trained, AWS Lambda, in conjunction with Amazon API Gateway, can be used to host and serve the model. This setup allows for generating and delivering personalized educational content on demand without provisioning or managing servers.

6. Monitoring and Optimization

- Amazon CloudWatch: The AI for Education project (AI4ED) should implement Amazon CloudWatch for monitoring the operational health and performance of the FM. Insights gained can drive further optimizations in content generation efficiency.
- Amazon SageMaker Model Monitor: To continuously improve the quality and efficiency of the generated educational content, SageMaker Model Monitor can be used to detect and remediate deviations in model performance over time.

By integrating these AWS tools and services, The AI for Education project (AI4ED) can effectively achieve its objective of creating an FM capable of generating diverse, high-quality, personalized educational content efficiently. This approach not only optimizes the computational resources required for training and hosting but also ensures the scalability and adaptability of educational content generation to meet varying needs.

Using AWS to Build Foundation Models for Multimodal Learning Experiences

To develop Foundation Models (FMs) for Multimodal Learning Experiences for The AI for Education project (AI4ED), leveraging AWS's ecosystem provides a robust foundation for innovation. This initiative aims to integrate FMs into various educational modalities, such as interactive simulations, augmented reality (AR), and image-based content analysis, to foster immersive learning environments. Here's a detailed strategy utilizing AWS tools:

1. Dataset Collection and Management

- Amazon S3: The AI for Education project (AI4ED) should start by storing diverse datasets on Amazon S3. These datasets could include educational content in various formats (text, images, videos) and domain-specific data (biology, engineering) for training the FMs.

2. Data Labeling and Augmentation

- Amazon SageMaker Ground Truth: To ensure high-quality training data, the AI for Education project (AI4ED) can use SageMaker Ground Truth. This service facilitates the labeling process with machine learning and human annotators, crucial for multimodal data.

3. Model Training and Development

- Amazon SageMaker: SageMaker is central to training and developing the FMs. It supports an array of machine learning frameworks that are conducive to multimodal learning model development.
    - Utilizing Pre-built Algorithms: SageMaker provides pre-built algorithms that can be leveraged for creating FMs capable of processing multimodal data.

4. Enhancing Model Performance

- Amazon SageMaker Neo: Post-training, the AI for Education project (AI4ED) can use SageMaker Neo to optimize the FM for efficient deployment on diverse hardware, including devices supporting AR and interactive simulations.

5. Deployment for Interactive Learning

- AWS Lambda and Amazon API Gateway: To deploy the FMs, the AI for Education project (AI4ED) can use AWS Lambda for serverless computing, allowing the models to be invoked in real-time. Amazon API Gateway can serve as the front door for applications to access data, business logic, or functionality from the FMs.

6. Augmented Reality and Simulations

- Amazon Sumerian: For creating AR experiences and interactive simulations, the AI for Education project (AI4ED) can explore Amazon Sumerian. It allows developers to build highly immersive and interactive scenes that can be integrated with the FMs for an enriched educational experience.

7. Monitoring and Analytics

- Amazon CloudWatch: To monitor the deployed models and applications, the AI for Education project (AI4ED) should implement Amazon CloudWatch. It provides detailed insights into application performance, enabling timely adjustments.

8. Iterative Improvement with Human Feedback

- Amazon Augmented AI (A2I): For continuous improvement, the AI for Education project (AI4ED) can integrate human feedback into the learning model workflows using Amazon A2I. This ensures the FMs remain relevant and effective in delivering personalized and immersive learning experiences.

Through this comprehensive approach, leveraging AWS's AI and ML services, The AI for Education project (AI4ED) can successfully develop and deploy FMs for multimodal learning experiences. This strategy not only enhances student engagement and understanding of complex concepts but also sets a new standard for interactive and immersive educational technologies.

Summary

The AI for Education project (AI4ED) at Northeastern University, leveraging its extensive network and educational data, faculty mentorship, and student engagement in building foundational models for education, seeks AWS's support for computing resources. This collaboration aims to advance foundation model development through open-source contributions to educational AI. If awarded, the computing resources will enable significant advancements in projects like Educational Dialogue Systems, Adaptive Learning Platforms, and Multimodal Learning Experiences, enhancing AI curriculum and addressing the computational resource limitations, thereby solidifying Northeastern's leadership in integrating AI into education.

For AWS, supporting the creation of open-source foundational models through the AI for Education project (AI4ED) at Northeastern University presents a unique opportunity to enrich the AI and ML community. It fosters innovation, drives the development of cutting-edge technologies, and enhances AWS's reputation as a leader in supporting educational advancements and research in AI. This collaboration not only showcases AWS's commitment to education and open-source projects but also extends the capabilities and reach of AWS technologies, potentially leading to new use cases and improvements in AWS's AI and ML services.

Budget

The following is a rough budget of how the AWS Promotional Credits would be spent across the five projects.

Educational Dialogue System Using Reinforcement Learning with Human Feedback: $60,000

- Heavy use of Amazon SageMaker for model training and real-time feedback loops.
- Amazon EC2 instances, potentially utilizing GPU or high-compute instances.

Adaptive Learning Platform Enhancement Through Distillation with Enhanced Reasoning: $50,000

- Amazon SageMaker for model distillation processes.
- AWS Lambda for deploying lightweight models for real-time educational applications.

Bias Detection and Mitigation in Student Performance Prediction Models: $40,000

- Amazon SageMaker and Amazon SageMaker Ground Truth for model training and data labeling.
- Use of Amazon S3 for extensive data storage and management.

Efficiency Enhancements in Educational Content Generation: $50,000

- Amazon EC2 and AWS Trainium chips for model training and content generation.
- Amazon SageMaker Neo for model optimization for various deployment environments.

Foundation Models for Multimodal Learning Experiences: $50,000

- Use of Amazon Sumerian for AR and simulation-based learning experiences.
- Amazon SageMaker and EC2 instances for training multimodal foundation models.