Mistral 7B LLM Deployment and API Interaction - Analytical Data Systems

Get ready to dive into the world of generative AI with an exciting project that combines the power of the Mistral 7B language model (LLM) with Google Cloud Platform (GCP).

Over the course of 5 weeks, you'll gain hands-on experience in setting up and deploying a private Mistral 7B LLM on GCP. Each week presents new challenges as you set up the infrastructure, develop deployment scripts, interact with the model by creating a secure API layer, and integrate it with Google Colab notebook that queries the LLM.

You'll work with cutting-edge technologies like Python, LlamaIndex, and Google Cloud Dev Ops, leveraging GCP services such as AI Platform and Vertex AI. By the end of the project, you'll have developed a fully functional pipeline for fine-tuning and deploying the Mistral 7B LLM, along with a secure API layer and Colab integration.

This project offers a unique opportunity to gain valuable skills in cloud computing, model deployment, API development, and secure access patterns. You'll collaborate with peers, learn from experienced mentors, and tackle real-world challenges in the field of AI and cloud computing.

If you're ready to take your skills to the next level and make a meaningful contribution to the world of generative AI, then this project is perfect for you. Let's embark on this exciting journey together and unlock the potential of the Mistral 7B LLM API!

Technologies: Python, Google Colab, LlamaIndex, Mistral 7B, Google Cloud Dev Ops

Overall Project Goal: To build a reusable python library, a secure API that allows authenticated access to the LLM, and dev ops pipeline that allows a user to create, configure, and deploy the Mistral LLM on a Google Cloud VM.

IP Rights: Analytical Data Systems will own all rights to deliverables, including code, data, prompts, and documentation. Feel free to use the know-how to advance your career.

Project Scope:
Week 1:
- Familiarize yourself with the Mistral 7B LLM, setup, hardware, and deployment requirements.
- Build a Colab notebook that queries the LLM from the colab environment using llamaindex.

Week 2:
- Develop a set of GCP build and deployment scripts to automate resource provisioning and the Mistral 7B model deployment on a Google Cloud VM.
- Create a dedicated network segment within the GCP project to isolate the Mistral 7B deployment from other resources and ensure network security.
- Configure appropriate firewall rules and access controls to restrict unauthorized access to the LLM infrastructure in the script.

Week 3:
- Develop a secure API layer that wraps the Mistral 7B model and exposes its functionalities through well-defined endpoints.
- Test the API layer.
- Implement monitoring and logging mechanisms to track the calls in and out of the model that are written to a log file.

Week 4:
- Deploy the Mistral 7B model and API on GCP.
- Implement authentication and authorization mechanisms to ensure that only authorized clients can access the API and interact with the LLM.
- Integrate the Mistral 7B API with Google Colab, allowing users to securely authenticate and make API calls to the deployed model.

Week 5:
- Develop a set of example Colab notebooks that demonstrate how to interact with the Mistral 7B API, showcasing various use cases and providing code snippets for common tasks.
- Conduct thorough testing and validation of the end-to-end pipeline, including the fine-tuned model, API layer, and Colab integration.
- Optimize the deployment for performance, scalability, and cost-efficiency, considering factors such as auto-scaling, caching, and resource allocation.

Throughout the project, you will gain hands-on experience in deploying large-scale language models on GCP, fine-tuning models for specific domains, and building secure API layers for model serving. You will also learn to work with Google Colab and develop interactive notebooks that leverage the power of the Mistral 7B LLM.

The project will span 5 weeks, with each week dedicated to specific tasks and milestones. Regular check-ins and progress updates will be conducted to ensure smooth progress and address any challenges encountered along the way.

By the end of this project, you will have developed a fully functional pipeline for fine-tuning and deploying the Mistral 7B LLM on GCP, along with a secure API layer and Colab integration. This experience will equip you with valuable skills in cloud computing, model deployment, API development, and secure access patterns.