Joao Gabriel de Souza Pinto

joao_g2000@hotmail.com

LINKS

Portfolio, Github, Linkedin

PROFILE

I am João, a dedicated Data Scientist and Machine Learning enthusiast from São Paulo, Brazil. My professional foundation is built upon robust experience in Industrial Data Analysis and SEO, further enriched by advanced studies in Big Data at PUC-PR. As a pivotal member of the HAILab research team, I specialize in the development of Large Language Models (LLMs), particularly focusing on biomedical and clinical data data.

EMPLOYMENT HISTORY

Oct 2023 — Present

NLP Researcher, HAILab- Health Artificial Intelligence Lab

Remote

As part of the HAILab team, I spearheaded the training of large language models (LLMs) on biomedical and clinical data.

Utilizing the Vertex-AI platform, I have conducted training sessions employing techniques such as Lora, qLora, and Full training, aimed at developing a model capable of generating clinical texts. My expertise encompasses proficiency in training models, using a spectrum from single A100 GPU utilization to harnessing parallelism across multiple GPUs. Additionally, I have actively participated in the complete model inference pipeline, ensuring a smooth transition from training to deployment.

Technical cases:

1. Pre-training: Using the LLama-Factory library in conjunction with Huggingface Accelerate, I conducted the pre-training of the LLama 7B base model with new data(Clinical and Biomedical)...
2. Fine-tuning: After injecting new knowledge into my model, I performed fine-tuning on biomedical QA datasets using Supervised Fine-Tuning techniques...
3. Inference: I actively participated in the model inference process... Combining various techniques and evaluating the model using benchmark approaches such as the mmlu dataset and creating scripts to test multiple models simultaneously and assess response performance.

Skills: Fine-tuning,Supervised Fine-Tuning,Pre-Training,Vertex-AI,LLama,GCP,Inference,evaluation,LLM.

Sep 2022 — Present

SEO Data Scientist, Cadastra

São Paulo

As a member of the Data team specializing in SEO, I have actively participated in projects centered around API integration, web crawler-based data collection, and report automation. Additionally, I developed tools using Large Language Models (LLMs). As a matter of fact, I executed automation using Selenium in EC2 instances, aiming to extract data from various sources such as Google Search Console, Google Trends, site: searches, robots.txt files, and logs. The data obtained was then securely stored on AWS S3.

Technical cases:

1. GPT API: I developed an automated dashboard that utilizes the GPT-3 model to generate text based on data gathered from Crawlers and applications such as ScreamingFrog, Ahrefs, and GA4. I designed and implemented the entire data pipeline and API trigger. Currently, this dashboard is being utilized by 32 clients.
2. Automation using AWS: I implemented automation using Selenium on EC2 instances to scrape data from various sources such as Google Search Console, Google Trends, site: searches, robots.txt files, and logs. The scraped data was then securely stored in AWS S3
3. RAG Q/A: Leveraging cutting-edge LLM tools, I triumphed at the company hackathon. My project used Langchain, Chromadb, Streamlit, and ChatGPT to create an application where it was possible to ask questions/answer to the company's training PDFs. I constructed a vector database with diverse PDFs on departments, roles, timekeeping, and other new hire resources. This project, showcasing my expertise in LLM and vector database, empowers new employees with intuitive search capabilities for a seamless onboarding experience.

Skills:Langchain,Chromadb,Streamlit,Chatgpt,Multi-agents,LLM,EC2,S3,Crawler,RAG,QA,llamaindex

EDUCATION

Jan 2021 — Jan 2024

BTech, Big Data Technology and Analytics, PUCPR- Pontifical Catholic University of Parana

Londrina

After an arduous 3.0 years, I achieved a bachelor's degree in Big Data and Analytical Intelligence from the Pontifical Catholic University of Paraná (PUCPR).