Sudi Sabet

sudi.sabet@gmail.com -- (650) 243-1405 -- www.linkedin.com/in/sudi-sabet

Summary

AI data engineer with more than 10 years of experience in B2B and B2D application development used by 10k+ users. Accomplished in cloud-based development of AI and computer vision applications as containerized workloads, LLM solutions and chatbot development, as well as data engineering and scalable microservices. Recognized for deep technical knowledge and proven ability to collaborate cross-functionally and drive the software development life cycle end-to-end.

Technical Proficiencies:

Software Development skills: C/C++, Python, Bash, Git, Github, Jenkins, VSCode, Jupyter Lab/Notebook, Docker, Kubernetes
AI Tech Stack: LlamaIndex**,** LLMs, RAG, GraphRAG, Azure AI, Hugging Face, Ragas, Trulens **Data Engineering Tech Stack:** Telegraf, Promtail, Loki, Grafana, Splunk, InfluxDB

Professional Experience:

Senior AI Solution Engineer -- Intel Inc. Santa Clara, CA **Jan 2021 - Present**

AI Engineering:

- Instrumented Intel's OpenVINO integration into TensorFlow C++ code with telemetry capabilities, allowing the team to track unsupported AI model operations and prioritize their support based on telemetry data, which facilitated further optimization of the model.
- Designed and prototyped an in-house evaluation tool using GPT-4 to score LLama2-70B responses against our golden dataset's ground truth. As the result operation cost was reduced by 30%.
- Developed a methodology to measure the accuracy and precision of GPT-4 as an LLM evaluation tool and compared them to those of industry-adopted tools like Ragas and TruLens. This enabled the team to choose the most accurate tool for the automation of chatbot evaluation as the backend LLM continuously improved through retrieval augmented generation (RAG).
- Developed a healthcheck API for multiple different microservices in AI benchmarking platforms, integrated with AWS ECS, enabling automatic restarts during network disruptions. This minimized potential downtime by ensuring continuous microservice communication.
- Led the full development lifecycle of reference computer vision applications, encompassing implementation, testing, containerization, and deployment, while securing legal, security, and open-source approvals.

Data Engineering:

- Developed and configured the integration of an open-source plugin-driven data collector as a microservice within a complex telemetry stack.
- Led a team of 3 developers to empower 1000+ users to monitor system utilization for AI workload optimization through enabling accurate data flow from Telegraf to Promtail, Kafka, and Loki, with visualization on Grafana dashboards.
- Implemented multiple Bash scripts and deployed them as Kubernetes cron objects to send observability metrics to Intel's observability endpoints to enable the product teams to effectively monitor and manage the health of the stack.
- Created an innovative solution using Splunk architecture to automate the sourcing of a web analytics dashboard with data from Hadoop-managed data lakes and Adobe Analytics in the absence of API access. This solution enabled product owners to refine their roadmaps based on user interaction and traction with their cloud products.
- Developed an Analyzer Task Library (ATL) in the .NET framework, ensuring seamless end-to-end integration into the data collection pipeline and successful data delivery to Splunk. Collaborated with teams to enable data transfer to AWS Redshift post-ETL. As the result the team gained visibility into per user system usage.

Computer Vision Engineer -- Nod Labs, Palo Alto, CA **Aug 2018 - Dec 2020**

- Benchmarked TensorFlow Lite models on Google Coral TPUs and PyTorch models on Nvidia GPUs, delivering reproducible performance data. These benchmarks provided the team with reliable reference points for evaluating quantization results as the company shifted focus toward model optimization.
- Utilized Intel's OpenVINO toolkit to optimize and deploy Nod's computer vision model on Intel's Neural Compute Stick 2 (NCS2). This was successfully used on a small robot prototype and presented to Amazon for potential use in their drone package delivery.
- Designed an obstacle avoidance algorithm and integrated voice recognition capabilities on a small robot prototype using LIDAR point cloud data from the Livox C++ SDK. This setup provided valuable ground truth data to evaluate the performance of Nod's Simultaneous Localization and Mapping (SLAM) algorithm.

Frontend Development and Project Management -- Docspera, Sunnyvale CA **Mar 2017 - Feb 2018**

- Built and led the QA team from the ground up, selected and managed TestIO among other options as our QA partner, integrated their platform with our bug tracking system, and authored feature specifications for over 10 Docspera app sections. Migrated the project tracking system from FogBugz to Jira.

Cofounder, Product Manager and Developer -- MagiKite_- Los Altos Hills, CA **Dec 2014 - Mar 2017**

- Designed and developed an Android App for teaching Farsi language.
- Interviewed customers and gathered requirements for the application.
- Launched the application in Google Play and started a partnership with Alborz Farsi School

Education:

Product Management: Transforming Opportunities into Great Product, 2021

Stanford Center for Professional Development

Master of Science (MS) in Chemistry

San Jose State University, San Jose, CA. Master's thesis (292 downloads)

Bachelor of Science (BS) Computer Science

San Francisco State University, San Francisco, CA