Saeid Alavi

✉ saeid.alavi@mail.utoronto.ca | 🌐 saeidalavi.me | salavina | salavina

Machine Learning Engineer with +3 years of work/research experience in NLP, GenAI, and MLOps projects

Education

Biomedical Engineering, Specialized in Machine Learning — *MASc* Sep 2020 - Aug 2022
Publications in Top ML Conferences – Conference Awards GPA: 4.00
University of Toronto, Toronto, ON

Electrical Engineering, Minor in Computer Science — *BESc* Sep 2016 - Aug 2020
Gold Medalist – Graduated with Distinction – Dean's Honor List GPA: 4.00
University of Western Ontario, London, ON

Work Experience

Vector Institute — *Machine Learning Associate* Sep 2023 - Dec 2023

Worked toward democratizing the use of building codes via AI-based chat copilot.

- Developed an LLM-based building code RAG chatbot for Trax.co by incorporating state-of-the-art LLMs, advanced prompt engineering, vector databases, LLM agents, Azure ML studio, and Azure Promptflow. Demo Chatbot can be accessed here or via Trax.co.
- Collaborated in an **Agile/Scrum** team of 3 developers to optimize the performance and deliver new features.

University Health Network (UHN) — *Machine Learning Scientist* Sep 2022 - Aug 2023

Proposed a novel dataset & benchmark for evaluating creative problem solving & Artificial General Intelligence (AGI) in LLMs.

- Conducted EDA and Python-based web scraping for data acquisition.
- Fine tuned transformer based NLP models such as BERT using HuggingFace library and PyTorch for downstream tasks.
- Performed prompt engineering and fine-tuning of LLMs for benchmark comparison.
- Applied unsupervised learning methods such as constrained K-means clustering to classify sentence embeddings.
- Collaborated in a team of 5 researchers, contributing to open source Github and HuggingFace repositories as well as published the results in NeurIPS conference [1].

The Entrepreneurship Hatchery — *ML Project Lead* Dec 2021 - Dec 2022

Worked on project AlphaWit: AI-based platform That Makes SLPs' Workflow More Efficient

- Scraped SLP contacts from internet using Beautiful Soup and Selenium for interview setup.
- Conducted interviews with SLPs to identify their pain-points and needs to come up with a thorough business plan.
- Collaborated in developing a mobile app using React Native to assist SLPs with in-session clinical assessment.
- Utilized AWS cloud services such as Lambda, S3, EC2, and RDS to run machine learning models and extract real-time features from audio/video data.
- Managed a group of 5 employees.

University of Toronto — *Machine Learning Researcher* Sep 2020 - Aug 2022

Worked on developing deep learning algorithms for automatic detection of Neurological Diseases.

- Adapted and integrated state-of-the-art deep learning algorithms for accurate speech recognition, repetition detection, and temporal segmentation via Huggingface and Pytorch framework.
- Trained/fine-tuned transformer based Computer Vision models for accurate facial landmark detection and repetition count in videos.
- Worked remotely on a startup working towards automatic detection of Parkinson Disease using in-house-developed algorithms which resulted in receiving $200K grant.
- Pulished the results in top tier IEEE conferences and Journals [2–6].

Technical Projects

UHN — *End-to-End Kidney Tumor Segmentation Project with Cloud Deployment* Jan 2024 - March 2024

- Created a Flask app that performs Kidney tumor segmentation. The dataset consists of medical imaging resources and was obtained from Kaggle KITS challenge.
- Fine-tuned YOLO 5 using PyTorch and HF transformers, managed the model lifecycle using MLFlow, tracked the large data files using DVC, built a CI/CD pipeline using GitHub actions, and Deployed the final web app on Azure web services instance.

UHN — *Medical Chatbot Project with Quantized Llama-2, Mixtral, & ChainLit* Jan 2024 - Feb 2024

- Created a chatbot that can run on CPU and answer medical questions.
- Embed, Chunked & Stored data in Pinecone vector database and connected it to quantized LLMs through Langchain. Built the chat interface using Chainlit. All models are open-source.

Personal Website — *3D Digital Avatar Portfolio Agent with LLM Backend* Dec 2023 - Ongoing

- Developed a 3D digital clone capable of performing RAG and answering questions regarding resume with cloned voice. Front-end was developed using React and Three.js. Back-end was developed via OpenAI API and Azure custom TTS.
- First version of the CV chatbot was developed on Google Dialogflow CX using intent and entity recognition connected to a knowledge base vector database. The base chatbot can be accessed here.

VirtualSLP Startup — *Web App for Temporal Segmentation via Speech Recognition* Jan 2023 - May 2023

- Created a web app using Flask that records audio/video and performs temporal segmentation via fine-tuned speech recognition models.

Responsibli.ai Demo — *AI for Investment* May 2023 - July 2023

- Implemented a Streamlit app using fine-tuned Bert-based model (Roberta-financial-news-sentiment) for Investment Materiality prediction.

Skills

Python, JavaScript, SQL, React, Pytorch, HuggingFace, Transformers, Git, DVC, MLFlow, Terraform, Matplotlib, Pandas, Numpy, Scikit-learn, Spark, OpenCV, Azure, GCP, AWS, Docker, Kubernetes, THREEJS, Flask, Streamlit, Chainlit, Langchain, LLamaIndex, Openai. LaTeX

Professional Certificates

- Machine Learning (Stanford)
- Deep Learning Specialization (Deeplearning.AI)
- Customer Experiences with Contact Center AI - Dialogflow CX Specialization (Google Cloud)

Non-refereed Contribution

- Performed peer review for NeurIPS Dataset and Benchmarks Track 2023
- TAship for the course APS 106H1S – Fundamentals of Computer Programming
- Presented the paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" at IATSL.
- Presented the paper "AST: Audio Spectrogram Transformer" at Computer Vision Taati Lab.

Selected Publications

[1] S. Alavi Naeini, R. Saqur, M. Saeidi, J. Giorgi, and B. Taati, "Large language models are fixated by red herrings: Exploring creative problem solving and einstellung effect using the only connect wall dataset," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[2] S. A. Naeini, L. Simmatis, D. Jafari, Y. Yunusova, and B. Taati, "Improving dysarthric speech segmentation with emulated and synthetic augmentation," *IEEE Journal of Translational Engineering in Health and Medicine*, 2024.

[3] S. A. Naeini, L. Simmatis, D. Jafar, D. L. Guarin, Y. Yunusova, and B. Taati, "Automated temporal segmentation of orofacial assessment videos," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022, pp. 01–06.

[4] S. A. Naeini, L. Simmatis, Y. Yunusova, and B. Taati, "Concurrent validity of automatic speech and pause measures during passage reading in als," in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 2022, pp. 01–06.

[5] L. Simmatis, S. Alavi Naeini, D. Jafari, M. K. Y. Xie, C. Tanchip, N. Taati, S. McKinlay, R. Sran, J. Truong, D. L. Guarin *et al.*, "Analytical validation of a webcam-based assessment of speech kinematics: Digital biomarker evaluation following the v3 framework," *Digital Biomarkers*, vol. 7, no. 1, pp. 7–17, 2023.

[6] G. Almog, S. Alavi Naeini, Y. Hu, E. G. Duerden, and Y. Mohsenzadeh, "Memoir study: Investigating image memorability across developmental stages," *Plos one*, vol. 18, no. 12, p. e0295940, 2023.