Zafarullah Mahmood

Toronto, ON | +1 437-223-7536 | zafar@zafarmahmood.com | zaffnet | zaffnet | zafarmahmood.com

Skills

Technical: Python (Fluent), C/C++ (Familiar), JavaScript (Prior experience), Flask, Sanic, SQLAlchemy, Alembic, Angular
Deep Learning & NLP: PyTorch, NVIDIA NeMo, HuggingFace (transformers, PEFT), spaCy
Generative AI: LangChain, LangGraph, LlamaIndex, RAG Evaluation (RAGAs), Model Serving (vLLM)
MLOps & Cloud: Docker, Kubernetes, Elastic Container Services, Kubeflow, AWS (Bedrock, SageMaker), GCP (Vertex AI)
Databases & Vector Stores: BigQuery, PostgreSQL (pgvector), Firestore, Redis, ChromaDB, FAISS

Education

✦ **Master of Applied Science in Computer Engineering at the University of Toronto** **2023 – 2025**

- **Thesis**: A Fully Generative Counsellor Chatbot for Smoking Cessation and LLM-Based Synthetic Smokers | **GPA**: 4.0/4.0
- Deployed a multi-agent therapeutic chatbot [1] on **AWS ECS**, which helped smokers increase their confidence in quitting smoking by 1.7 (0-10 scale). Improved chatbot's therapeutic quality by 26% by implementing **ReAct** for counsellor behaviour selection. Reduced SLOC by ~60% by re-implementing the chatbot system using **LangChain LCEL**. Gained experience in **ChromaDB**.
- Devised a validation framework for high-fidelity persona installation. Tested the framework's viability by installing human smoker attributes into LLM-based synthetic smoker "doppelgängers".

✦ **Bachelor of Technology in Computer Engineering at Jamia Millia Islamia, New Delhi** **2014 – 2018**

- **Focus**: Natural Language Processing | **Internship**: Indian Space Research Organisation | **CGPA: 8.2/10**

Work Experience

✦ **Natural Language Processing (NLP) Engineer at Dialpad Canada Inc.** **Apr 2019 – Aug 2023**

- Deployed a real-time Spanish-English bilingual sentiment classification model by fine-tuning an English-only model. Achieved target F1 on the Spanish testset without performance loss on original English testset. Used **transformers** and **spaCy**.
- Reduced real-time inference latency by 4% by replacing a sequential punctuation and casing pipeline with a multi-classification-head **BERT** model. Identified and fixed performance bottlenecks in model's custom tokenizer using **py-spy** and **scalene**.
- Learned **Angular** to develop an internal text-to-speech annotation platform. Adopted the iterative co-design paradigm and improved annotator efficiency by 12% by adding features like hotkeys, smart text completion, and dynamic keyword highlighting. Used **Sanic** and **PostgreSQL** for the backend, connected with cloud data sources and sinks, including **BigQuery** and **Cloud Storage** and deployed the application on **Google Kubernetes Engine Deployment**.
- Built a tool to create synthetic speech data from a list of *keywords*: used **Beautiful Soup** for web scraping, **spaCy** for named entity recognition and sentence creation, and **Coqui TTS** to generate speech from sentences. The synthetic dataset was used to train a keyword boosting algorithm [2] and helped improve keyword recognition accuracy on a real test set by 26% relative.
- Built a **Kubeflow** pipeline to automate the generation and verification of pronunciations of newly coined words (e.g., *COVID-19*). The pipeline added ~10,000 words in a year and helped the team beat a benchmark on pronunciation generation [3].

✦ **Data Scientist at Exzeo Software, India** **Jun 2018 – Mar 2019**

- Built a text-based, omni-channel conversational agent using **Google DialogFlow** to help customers file insurance claims.

Publications

[1] A Fully Generative Counsellor Chatbot for Moving Smokers Towards the Decision to Quit. ***ACL Findings***, **2025**.
[2] N-gram Boosting: Improving Contextual Biasing with Normalized N-gram Targets. ***ICASSP***, **2023**.
[3] Avengers, Ensemble! Benefits of ensembling in grapheme-to-phoneme prediction. ***SIGMORPHON***, **2021**.

Projects

✦ **AutoAnnoMI: Automating Annotations of MI Conversations using LLMs** **2023**

- Built an LLM-based tool to label utterances in counselling conversations. Incorporated chain-of-thought reasoning and few-shot examples leading to a ~12% accuracy improvement over baseline.

✦ **Benchmarking Batch Renormalization** **2017**

- Implemented `BatchReNorm1d` module in **PyTorch** and benchmarked its performance on image recognition datasets.

Awards

✦ **Edward S. Rogers Sr. Graduate Scholarship, University of Toronto** **2024**

- Received CAD 20,000 in recognition of outstanding academic accomplishments during Master's studies.

✦ **Best Kernel Award, Kaggle** **2018**

- Won Best Kernel Award among 80 kernels in DonorsChoose.org Application Screening Challenge.

† Eligible to work in Canada without sponsorship.