AMAN SAINI

+1 (360) 842-6055
saini.aman.personal@gmail.com

EXPERIENCE

Grammarly **Oct 2023 - Present | Vancouver, BC, Canada**
Staff ML Engineer (L5)
Working in the Strategic Research organization to advance Grammarly's LLM capabilities in the writing space. Previously, worked as part of the Core Product team to build new and state-of-the-art Generative AI features. Major projects included:

- Fine-tuning cutting-edge LLMs (Llamas, GPTs) to provide fine-grained writing assistance to Grammarly users.
- Researching and publishing a paper on training multi-task Ukrainian text-editing models. The datasets and models are released on Grammarly's Huggingface page to advance further research in the area.
- Building a python library for Automatic Prompt Optimization (using LLMs) given a seed prompt and target reward metric. The approach is based on the EMNLP 2023 paper published here.
- Building and scaling Synthetic Data Generation pipelines to train various internal LLMs at Grammarly.
- Building high-quality LLM-powered features available to Grammarly Premium users as writing suggestions.
- Instruction-tuning LLaMA3.1 (70B/8B) Teacher models and building evaluation pipelines using LLM-as-a-Judge.

Twitter, Inc. **May 2021 - Jan 2023 | Seattle, WA, USA**
Senior ML / NLP Engineer
Worked for the Natural Language Processing (**NLP**) Signals team in the Central Machine Learning org "Cortex" at Twitter. The team worked on building end to end NLP models and signals used in various product teams at Twitter including Home Timeline, Notifications, Trends etc.

- **Entity Linking** - Worked on building the end to end ML models to link important Named Entities in Tweets to an external knowledge base like Wikipedia. Implemented Transformer based encoder-only models for entity detection, candidate generation pipeline, and entity disambiguation/ranking models. Published a paper to open source the datasets used for building this system in NeurIPS 2022 to support further research. This new system is used internally to link entities on live Tweet traffic and as a signal in other downstream ML models.
- **NER (Named Entity Recognition)** - Worked to replace the old Bi-LSTM based NER model with the SOTA multilingual Transformer style BERT encoder trained using various techniques including - unique subword masking loss, special token level features, weak data labeling using Linking datasets, Teacher-Student knowledge distillation, etc. Offline evaluations showed **13.5% F1-Score** improvements as compared to the old system.
- **Tweet Representation model** - Worked on building a multi-task model trained to perform language, topical and engagement prediction to learn generic Tweet representations. It uses signals beyond the Tweet text and is optimized to work with multiple downstream tasks. Published a paper on this embedding technique in DL4SR(Deep Learning for Search Recommendation) workshop 2022.
- **Twitter Notifications** - Worked on analyzing and improving the quality of seed Tweets used in candidate generation sources to recommend more recent and organically relevant Tweets in Twitter mobile notifications.
- Led the team in multiple projects including adoption of NER, establishing ML code guidelines, setting up BigQuery to Manhattan offline jobs, improving model training speed, etc.

Microsoft Corporation **Jan 2016 - April 2021 | Seattle, WA, USA**
Worked for the Bing ads team in the AI+Research org at Microsoft. The team worked on improving the ad experience, and launching new ad products for Bing users to drive more clicks on the Bing search engine.
Senior ML Engineer

- **Ad Creative Optimization** - Primary owner and lead of the various ad creative algorithms for Dynamic Search Ads. Designed and implemented different extraction (Landing page based) and generation (Transformer based) algorithms to automatically generate ad titles and descriptions for advertiser provided URLs within a 2-day SLA. Responsible for onboarding new markets, supporting new languages, new ad decorations, and improvements in this ad product that brings **1.5% RPM** (Revenue per 1000 searches) in the US market daily. This ad product has been successfully launched in top tier international markets and is delivering **2+% RPM** in all markets.
- **Web Page Similarity Graph** – Worked on building a web page similarity graph within an advertiser domain to perform next link prediction. Trained lightweight student models using Microsoft's TwinBERT architecture to get contextual embeddings of pages along with open-source BERT and USE. User search patterns were mined from the browser to gather data for closely visited pages and the similarity was computed using scalable ANN techniques like HNSW, and NSG to perform next link prediction for the user.
- **Dynamic links with ads** – Worked on recommending website links for an ad copy to drive user clicks and engagement. The techniques involved using MSR's multi-label classification models like DeepXML to categorize

each page to a list of predefined queries and using that to form the page-page connections. The technique was further enhanced by using an Ads RoBERTa model trained on landing page features to provide relevance scores. This technique was mainstreamed in the US market with **+1.5% CTR** (Click Through Rate) gains.

- **Landing page Summarization** – Worked on transformer-based techniques like BERTSUM to do extractive summarization of landing pages and created short assets from the page to be used as ad title snippets.

Senior Software Engineer

- **Ad Creative infrastructure** – Primary owner and technical lead of the infrastructure that powered the ingestion of automated ad copies in Bing ads. The workflow involved support for different ad copy algorithms, editorial service for generated ad copies, azure integration, caching and other optimizations to handle the scale, and seamless integration with Microsoft's workflow execution engines like Sangam. This infrastructure is currently in production and is being used by other teams in Bing for automatic ad-creative generation scenarios.
- **Personalization** – Designed and implemented ways to personalize Bing ads using the user views, clicks, and conversion signals collected across Microsoft ad products.
- **App Install Ads** – Implemented the backend for App Install ads, which promotes the installation of apps on Bing ads. This product resulted in a significant increase in the app downloads on Bing.
- **Generative work** – Implemented the algorithms to add 2 lines of description to sitelinks, which increased the clickability of sitelink extensions on Bing ads significantly. Expanded the generic display URLs and descriptions for ads with specific and more relevant data from the landing page to increase ad clicks.
- **Other projects -** Shortening of product ad titles, supporting chat bot annotation, etc.

Microsoft India Development Center **July 2013 - Dec 2015 | Bangalore, India**

Software Engineer

Worked for the Rich Ads experience team under Relevance and Revenue team (RnR) in Bing Ads. The team worked on the online infrastructure and creating rich ad experiences for all non-US markets.

- **Ads-Composition -** Designed and coded an entirely new infrastructure to process all the ads and decorations in one place before serving on Bing. Parallelization and modularization made the code robust, debuggable and readable which resulted in tremendous improvements in latency and agility of experimentation.
- **Related Product Annotation –** Implemented the E2E flow to show relevant products from Product ads corpus for a retail intent query. This increased the clicks on both the text ad and related product links.
- **Online Infrastructure work -** Worked on making Bing ad infrastructure robust by analyzing the critical paths that contribute to latency. The built online utilities are being used in other online components.

PUBLICATIONS

- Spivavtor - An Instruction Tuned Ukrainian Text Editing Model
- TweetNERD - End to End Entity Linking Benchmark for Tweets

INTERNSHIPS

Amazon Jan - June 2013 | Bangalore, India

Worked for the Amazon Pricing team to predict the shipping price for products sold on amazon.com. Implemented the E2E service called 'Pricing Attribute Prediction Service' using Amazon Simple workflow service (SWF) to automate the ML model building process.

University of Victoria, British Columbia, Canada May - July 2012 | Victoria, Canada

Worked to replace Java with a bidirectional language "Boomerang" in a high confidence medical data device known as "MirthConnect" for correctness of transformations and to guarantee well-behaved channels.

Indian Institute of Remote Sensing, IIRS, India May - July 2011 | Dehradun, India

Worked for the *'Weather Research and Forecast Modeling'* team on deploying WRF models on Linux systems to recreate past events to check correctness and use it for real-time weather prediction in future.

EDUCATION

B.E (Bachelor of Engineering) in Computer Science | **CGPA - 9.54/10** | Aug 2009 - July 2013
Birla Institute of Technology and Science (**BITS, Pilani**) India

Class 12th (CBSE)- 94.6% (2009) , Class 10th (CBSE) - 95.2% (2007) | D.A.V Public School