Soham Ray

917-753-6259 | sr2259@cornell.edu | linkedin

SUMMARY

With over six years of experience in machine learning and engineering, I have spent the last 6 months working at Sierra under Dr. Karthik Narasimhan (Princeton) and three years at ASAPP under Dr. Kilian Weinberger (Cornell). My strengths include taking projects from ideation to production, designing experiments with clear evaluation and swift iteration, scalably training/finetuning models, writing clean production code, and effectively communicating results across teams.

EDUCATION

Cornell University Ithaca, NY
Masters in Computer Science | GPA 3.91/4.20 May 2020
Courses: Large Scale Machine Learning, NLP, Advanced Machine Learning, Computational Linguistics, Algorithms

TECHNICAL SKILLS

python, pytorch, langchain, aws ec2, aws athena, tensorflow, langgraph, openai, matplotlib, R, git, sql, java, huggingface

EXPERIENCE

Sierra.ai | Customer Service AI San Francisco, CA
Senior Research Engineer Mar 2025 - Present

- Tau2 Bench: Built a benchmarking framework for evaluating foundational models agentic capabilities in interactive customer service domains where customers can also take actions [mentioned in the gpt-5 keynote and cited by OpenAI, Anthropocene, Google and others]
- Multilingual Transcription Benchmarking: Currently leading the effort on building an agent oriented multilingual benchmark for transcription, for the leading CX AI Agents company with millions of calls and supporting 30+ languages.

ASAPP | Customer Service AI New York, NY
Senior Research Engineer Nov 2021 - Mar 2025

- Led a team of 4 to improve AI agent performance by identifying customer pain points, addressing them by fine tuning, re-architecting and evaluating the system for an enhanced customer experience
- Pushed to production a new *trigger mechanism for human intervention* in a human-in-the-loop AI customer service agent, significantly reducing errors, customer conflicts, and escalation; decreasing uncaught errors from 75% to 16%.
- Designed a graph-based architecture for ASAPP's LLM-powered chatbot, *transitioning from monolithic to targeted task-based prompts*. This approach increased configurability and robustness by leveraging agentic systems.
- Led a team to develop ASAPP's QA system, i.e., *training models for binary disposition forms and post-call summaries*. Finetuned open-source models (3B to 70B), surpassing GPT-4 performance according to human annotations.
- Led research and adoption of *generative language models for a next utterance recommendation system*: ASAPP's then primary chat augmentation product. Boosted usage scores by 10% on average across customers in A/B tests
- Developed a system to create *enriched conversation summaries using GPT-4 for concise summarization and entity extraction* through prompt engineering and regex, tailored to customer specifications.
- Researched and *developed both generic and dedicated phrase autocomplete retrieval models* using GPT-2, increasing exact match scores by approximately 20% while maintaining coverage, across 7 customers, measured by A/B tests.
- Built an agent-coaching system that *raised the floor of agent performance by suggesting senior agent responses* live, using FAISS and retrieval models, improving response quality by 17%, as per human annotation.

Afiniti | Customer Service AI Washington DC
Data Scientist Aug 2020 - Nov 2021

- Developed statistical models to behaviorally pair customers and agents using linear optimization and ML methods
- Capitalized on changing trends in customer support call routing data to boost revenue gain from ~ 0.5% to ~5% (~$500,000/month)

ACHIEVEMENTS AND SERVICE

- [*patent pending*]: Developed SAGA: a system for conversational AI agents to learn from live agent supervision, de-risking new customer lanches and facilitating AI agent training in partially observable markov decision processes
- [*Hackathon winner*] Sphinx: Led my team to build a unified QA model trained on multi-task multi-customer data that could be horizontally integrated across ASAPP products, as opposed to a model per task/customer.

- [*Mentorship*] Theorem Proving: Mentored undergraduates with Prof. Claire Cardie to predict the next tactic (proofstep) in theorem proving using gpt2, trained on cornells' arxiv and NuPRL data stores
- [*patent pending*] Key Value Extraction System: Built a system to detect, prompt, and extract custom sets of entity, entity value pairs from a conversation to facilitate multi-turn automation in task-oriented dialogue systems
- [*patent pending, Interspeech 2024 arxiv*] Sample-Efficient Diffusion for TTS, *Justin Lovelace, Soham Ray, Kwangyoun Kim, Felix Wu, Kilian Weinberger*
- *[ICLR 2025 submitted]* Tau2 Bench: Evaluating Agents in Interactive Environments, Victor Barres, Honghua Dong, *Soham Ray*, Karthik Narasimhan

RECOMMENDATIONS

- Kilian Weinberger, Principal Scientist ASAPP, Professor of Computer Science at Cornell University link
- Ryan Macdonald, Chief Scientist ASAPP, Ex Google 14 years link

ADDITIONAL EXPERIENCE

Cybage | Machine Learning Engineer Intern *Feb 2019 ─ May 2019*

- Developed software to analyze sentiment and abstracted the product domain so it can be used for any feedback analysis task
- Improved classification accuracy by 11% on inplace system, and scored an accuracy of 81% on data from un-trained domains

ASquared IoT | Machine Learning Engineer Intern *July 2018 ─ May 2019*

- Developed a web application in python to classify welding sound files based on quality with a projected accuracy of 71%
- Constructed denoising models using statistical and ML methods to improve projected classification accuracy to 84%

Exadatum | Big Data Intern *July 2018 ─ Dec 2018*

- Developed a production-ready real-time streaming ETL data pipeline, with pluggable ML model for sentiment analysis
- Designed data visualization dashboards illustrating various KPIs of a Fortune 100 Company

Selected Projects *Aug 2019 ─ May 2020*

- distilBertNQ: Developed a training script for Google's Open Domain Question Answering Task using HuggingFace's distilBert model, low precision GPU computing, and a custom optimizer to understand and answer a question given a raw wikipedia page. F1-score: 59%
- Deceptive Opinion Spam Classification: Constructed an opinion spam classifier to weed out fake reviews written by real people. Explored usage of language models and naive bayes classifier using different linguistic and preprocessing techniques. Accuracy: 91%
- Metaphor Detection with Sequence Labeling Models: Implemented pos tagging and metaphor detection on a given text corpus. Explored feature engineering, hidden markov models and feedforward neural networks to get the observation probability matrix. F1 Score: 65%
- Air Quality Modeling (Master's Project): Designed data pipelines and ML models to predict hotspots hyper-locally and observe spatio-temporal patterns in air pollution in global metropolitan areas to help curb rising air pollution levels worldwide.