Soumya Chatterjee

soumyac@stanford.edu | +1 650-642-7039 | linkedin.com/in/soumyach | soumya-ch.github.io

EDUCATION

Stanford University Stanford, CA
M.S. Computer Science (AI Specialization), GPA: 4.075/4.0 Jun 2024

Indian Institute of Technology Bombay Mumbai, India
B.Tech. Computer Science & Engineering, GPA: 9.81/10.0 May 2021

- Graduated 4th in class with Honors for completing additional courses. Minor in Applied Statistics & Informatics.

WORK & RESEARCH EXPERIENCE

Apple | *Machine Learning Engineer* Jul 2024 – Present

- Built an synthetic data generation framework for training agentic search systems reducing data creation time by 90+%
- Used multi-LLM orchestration, embedding-based example selection, compositional trajectory construction, and rejection sampling improving good answer rate and while enabling knowledge sharing across teams owning various tools
- Worked on a refresh for Siri's semantic parsing model conducting model evaluation and serving

Apple | *Machine Learning Engineering Intern* Jun 2023 – Sep 2023

- Created a pipeline for enabling existing translation models to correctly translate unseen terms without human intervention
- Designed prompts to generate sentences with the target term and their translations using large language models (LLMs)
- Used parameter efficient finetuning methods like LoRA to obtain 95+% term translation accuracy without a drop in chrF

Stanford University | *Research Assistant / Graduate Researcher* Sep 2022 – Jun 2023

- Accelerated Training of Protein Structure Detection Models (Representative PR) | *Advisor: Prof. Alex Aiken*
  - Worked on FlexFlow, a distributed training framework that finds optimal machine-specific parallelization strategies
  - Added CUDA/cuDNN operators, Python bindings and other features to support training GNN-based VAE models
- Multi-Distribution Information Retrieval (REML Workshop at SIGIR 2023)
  - Proposed a novel setting of information retrieval from different data distributions, some unseen during training
  - Designed methods for allocating retrieval budget across distributions based on uncertainty giving 8 point higher recall

Google Research | *AI Resident* Jul 2021 – Sep 2022

- Entity Disentangled Language Models
  - Designed a modified BERT-like language model for disentangling factual knowledge from language semantics
  - Pretrained the model by replacing entities with their types and adding entity embeddings in later Transformer layers
  - Preliminary results showed updating facts in our model requires fewer continued pretraining iterations than BERT
- Option Indexing for Hierarchical Reinforcement Learning (Published: AAMAS 2023, Extended Abstract)
  - Proposed a method for efficient re-use of temporally-extended policies (options) from a library of pre-trained options
  - Selected a subset of task-relevant options based on environment affordances and option co-occurrences
- Modeling Sequential Adjustments in Behavioral Policies on Inhibitory Control Tasks (Published: CogSci 2022)
  - Developed a RNN-based model for modeling inter-trial adjustments in human behavioral policies
  - Demonstrated that our method leads to 2x better fits and 10% more reliable re-estimation of behavioral indicators

IIT Bombay | *Undergraduate Researcher | Advisors: Prof. Sunita Sarawagi, Prof. Preethi Jyothi* Jan 2020 – Sep 2021

- Lexically-Constrained Translation using Word Alignments from Transformers (Published: ACL 2022, Oral Presentation)
  - Formulated a novel beam search algorithm incorporating terminology constraints in translation using word alignments
  - Evaluated on 5 language pairs showing an improvement of 1.2 points BLEU and 1.3 points lower alignment error rate
- Hyperbolic Label Embeddings for Hierarchical Multi-Label Classification (Published: EACL 2021)
  - Proposed a novel text classification problem where labels are known to lie in a hierarchy but the hierarchy is unknown
  - Jointly learned a classifier and hyperbolic Poincaré embeddings of labels inferring hierarchy from label co-occurrences
  - Demonstrated classification performance comparable to methods that use the true label hierarchy

Google Research | *Software Engineering (Machine Learning) Intern* May 2020 – July 2020

- Designed a meta-learning approach for estimating human behavioral policies using limited data (Published: AAAI 2021)
- Showed our model captures population-level trends and subject-level variations improving model fit log likelihood by 0.4

AWL Inc. | *Machine Learning Engineering Intern* Dec 2019 – Jan 2020

- Built a Faster R-CNN person and object detector for 360° videos. Obtained 20% higher mAP at 20 FPS on a single GPU
- Proposed a simpler multi-label classification method for faster inference on edge devices which is being used in production

PUBLICATIONS

- Accurate Online Posterior Alignments for Principled Lexically-Constrained Decoding — ACL 2022, Oral Presentation
  Soumya Chatterjee, Sunita Sarawagi, Preethi Jyothi
- Meta-Learning of Dynamic Policy Adjustments in Inhibitory Control Tasks — CogSci 2022
  Soumya Chatterjee*, Aakriti Kumar*, Pradeep Shenoy
- Model-agnostic Fits for Understanding Information Seeking Patterns in Humans — AAAI 2021
  Soumya Chatterjee, Pradeep Shenoy
- Joint Learning of Hyperbolic Label Embeddings for Hierarchical Multi-label Classification — EACL 2021
  Soumya Chatterjee*, Ayush Maheshwari*, Ganesh Ramakrishnan, Saketha Nath Jagarlapudi
- Thermal Face Recognition Based on Transformation by Residual U-Net and Pixel Shuffle Upsampling — MMM 2020
  Soumya Chatterjee, Wei-Ta Chu
- PORTool: Tool-Use LLM Training with Rewarded Tree — Preprint
  Feijie Wu, Weiwu Zhu, Yuxiang Zhang, Soumya Chatterjee, Jiarong Zhu, Fan Mo, Rodin Luo, Jing Gao
- Matching Options to Tasks using Option-Indexed Hierarchical Reinforcement Learning — Extended Abstract, AAMAS 2023
  Kushal Chauhan, Soumya Chatterjee, Akash Reddy, Pradeep Shenoy, Balaraman Ravindran
- Resources and Evaluations for Multi-Distribution Dense Information Retrieval — REML Workshop at SIGIR 2023
  Soumya Chatterjee, Omar Khattab, Simran Arora

SKILLS

- **Languages**: Python, Go, C, C++, Rust, Javascript, SQL, Java, MATLAB, Bash
- **Machine Learning**: PyTorch, Tensorflow, CUDA/cuDNN, JAX, NumPy, Pandas
- **Tools**: Git, Docker, Apache Beam, BigQuery, Gerrit, Bazel, Protocol Buffers, Kubernetes, Ray

SCHOLASTIC ACHIEVEMENTS

- Institute Academic Prize, IIT Bombay for being in top 3 students in the institute — 2018
- All India Rank 235 in the Joint Entrance Examination for Indian Institute of Technology — 2017
- Qualified in the top 300 students in India for Physics and Chemistry Olympiads — 2017