GOURAB DEY

(631) 739-7869 | gouravdey96@gmail.com | LinkedIn | Github

EDUCATION

MS (Thesis), Computer Science | Stony Brook University | ***GPA: 3.94/4*** *Aug 2022 - May 2024*
B.Tech, Information Technology | IIIT Allahabad | ***GPA: 9.30/10*** *July 2015 - June 2019*

EXPERIENCE

Amazon AWS, Annapurna Labs (Distributed Training), Software Dev Engineer - AI/ML | Cupertino, USA *June 2024 - Present*

- Enabling distributed training of large scale LLMs (1B-500B+ parameters) on AWS trainium chips by simulating customer models and performance testing throughput on multi-node clusters to ensure scalable and efficient performance.
- Spearheading the Accuracy Engineering subteam to establish and standardize accuracy benchmarks.
- Validating parallelism techniques (DP, PP, TP), analyzing training dynamics (loss curves, gradient norms, parameter norms), and conducting pre-training/SFT evaluations against GPU baseline to ensure training accuracy for large-scale models.
- Technologies: **Python, PyTorch, AWS Neuron, AWS S3, NVIDIA Nemo, Apple AXLearn, CUDA, HuggingFace, Slurm, Docker**

Walmart Global Tech, Software Engineer III | Bangalore, India *Aug 2019 - July 2022*

- Designed, developed and collaborated with a team of 7 members to successfully ship the smart-slotting product for NextGen Fulfillment Centers, fully automating the manual process of slotting and moving freights in warehouses.
- Developed RESTful APIs and implemented 2 new slotting algorithms resulting in a 150% increase in space utilization in the warehouses and reduced shipping time by 50%.
- Implemented highly scalable messaging systems to handle decanting of an average of 250k+ units per market weekly, and to maintain data sanctity across WMS.
- Designed and implemented over 4 new rule engines, improving flexibility, and allowing the product to be used across 6 different markets. Solely led the development of the product for the MFC market.
- Technologies: **JAVA, Spring, Kafka, MS Azure SQL, IBM MQ, Vavr, Drools, SpEL**

RESEARCH

SUNY Research Foundation, Research Assistant | **Advisor: H. Andrew Schwartz** | Stony Brook University *Jan 2024 - May 2024*

- Explored adversarial and residual strategies to develop robust, efficient, and explainable LLMs for authorship attribution.
- Technologies: **Python, PyTorch, DeepSpeed, PEFT, LoRA, RoBERTa, Llama2, Large Language Models**

HLAB, Graduate Researcher | **Advisor: H. Andrew Schwartz** | Stony Brook University *Jan 2023 - May 2024*

- Designed and developed **Fb-GPT** and **Twitter-GPT**, both 1.3B parameter models fine-tuned on 260M tokens of each social media domain, and benchmarked their cross-domain perplexity scores.
- Designed and developed **Socialite-Llama** – the **first** open-source LLM instruction tuned for social scientific tasks. Also released **SocialiteInstructions**, a robust instructions dataset (~203k data points) for Social Scientific NLP tasks.
- Led the scaling efforts of the above models, overcame challenges related to direct fine-tuning on consumer hardware through cutting-edge techniques such as gradient checkpointing, mixed precision training, distributed training (Deepspeed), and PEFT methods (LoRA).
- Engineered solutions to serve LLMs (>7B parameters) at scale and figuring out optimal training strategies for available resources using Deepspeed configurations and quantization techniques (QLoRA).
- Technologies: **Python, PyTorch, DeepSpeed, PEFT, LoRA, QLoRA, GPT-Neo, Alpaca, Llama2, Large Language Models**

Robotics Lab, Graduate Researcher | **Advisor: Michael S. Ryoo** | Stony Brook University *Sep 2023 - Dec 2023*

- Worked with Vision Language Models for Robot Learning and manipulation in Offline Reinforcement Learning settings.
- Pioneered efforts to address and mitigate the challenges posed by distributional shift by actively exploring the integration of additional contextual information, specifically natural language instructions.
- Technologies: **Python, PyTorch, DeepSpeed, PEFT, LoRA, Vicuna, Llama2, Llava, Vision Language Models**

PUBLICATIONS

- **On the Transferability of Causal Knowledge for Language Models [Under Review, COLING 2025]** *Sep 2024*
  Gourab Dey, Yash Kumar Lal
- **Towards Increasing the Social Understanding Capabilities of Large Language Models [Stony Brook University]** *May 2024*
  Gourab Dey
- **Archetypes and Entropy: Theory-Driven Extraction of Evidence for Suicide Risk [CLPsych, EACL 2024, Oral]** *Jan 2024*
  V Varadarajan, A Lahnala, AV Ganesan, **Gourab Dey**, Siddharth Mangalik, AM Bucur, N Soni, R Rao, K Lanning, I Vallejo, Lucie Flek, H.Schwartz, Charles Welch, Ryan L Boyd
- **SOCIALITE-LLAMA: An Instruction-Tuned Model for Social Scientific Tasks [EACL 2024, Oral]** *Oct 2023*
  Gourab Dey, AV Ganesan, YK Lal, M Shah, S Sinha, Matthew Matero, Salvatore Giorgi, Vivek Kulkarni, H. Schwartz

TECHNICAL SKILLS

- **Languages and Frameworks:** Python, Java, C++, SQL, PyTorch, Spring
- **Databases:** MySQL, MS Azure SQL, Cassandra, IBM Db2, AWS S3, Apache Solr
- **Libraries and Tools:** Kafka, Hadoop, Spark, Vavr, IBM MQ, Drools, Airflow, Grafana, OpenCV, Scikit-learn, Scipy, Transformers
- **Miscellaneous:** REST, LLM, VLM, PEFT, LoRA, Deepspeed, HuggingFace, Kubernetes, spaCy, Gensim, Flask, Tensorflow, NLTK