PRAGUN **ANANDA**

website | github | linkedin | pragun.ananda@gmail.com

EXPERIENCE

Google **_June 2022 - Present_**

ML Software Engineer - Core Search Infrastructure

- Onboarded several information retrieval product teams (Search, Youtube, Lens, Research) onto our team's vector database, search index, and vector similarity search infrastructure which stores >100B embeddings.
- Onboarded team's deep embedding models onto our hosted embedding inference platform and assessed TPU resource requirements for daily inference jobs used to maintain freshness for recommendation quality.
- Prototyping a deep learning-based optimization to Google Search query time using a Mixture-of-Experts based model architecture to predict the location of webpage embeddings in the Search index. Applied research from this paper on "learned index structures".
- Taught ML classes to >200 Google engineers on topics like LLMs, reinforcement learning, image understanding, recommendation systems, and deep learning.

Research Engineer - Cloud AI Research **_Aug 2023 - Present_**

- Re-implemented LLM interpretability research and ran long-document retrieval experiments for Cloud AI customers using the explainability value scoring algorithm. See accompanying blog post and Google Research Github repo.
- Built a model-agnostic abstraction for the algorithm to support open-source LLMs like LLaMa3 and Gemma hosted on Google Cloud's Vertex AI infrastructure. Demoed to the Deepmind Interpretability team to explore opportunities to apply the research to Gemini.
- Scaled explainability value generation by speeding up large model inference (10B params) using techniques like speculative decoding, in-place encoder resampling, and FlashAttention to optimize for the GPU.

University of Virginia Computer Science **_Aug 2019 - Jan 2022_**

Research Assistant - Joint research b/w NLP and Adversarial ML research groups

- Assisted in EMNLP published research on using graph theory to improve the robustness of mis-labeled natural language datasets. Trained and tested BERT performance on augmented data.

EDUCATION

University of Texas at Austin **_June 2024 - Present_**

M.S. in Artificial Intelligence

- **Courses:** Deep Learning, Natural Language Processing, Reinforcement Learning, Machine Learning

University of Virginia **_Aug 2018 - May 2022_**

B.A. in Computer Science, B.A. in Statistics

- **Courses:** Databases, Computer Architecture, Networking, Time Series, Linear Algebra, Probability
- **Accomplishments:** Distinguished Researcher, Head Algorithms TA, Data Structures TA

SKILLS

- **Languages:** Python, C++, Java, SQL
- **Technologies:** TensorFlow, Pytorch, JAX, HuggingFace, GCP, AWS, Kubernetes, Spark