JUSTIN S. KANG

@ justin_kang@berkeley.edu justinkang221.github.io

I am a final year PhD student. I have worked on interpretability LLMs, privacy in learning, and data valuation. Some recent projects focus is *attribution* problems: trying to explain which components (input features, training data, etc.) are most important to a model. I have some ongoing projects trying to apply some of these ideas to LLM *alignment*. I have a strong background in signal processing and information theory, and like to view these modern learning problems through that lens, which often leads to unique and exciting solutions. ***I am graduating soon and looking for full-time positions for 2026***.

SELECTED PUBLICATIONS

ProxySPEX: Sample-Efficient Interpretability via Sparse Feature Interactions in LLMs

L. Butler, A. Agarwal*,* ***Justin Singh Kang***** et al.*

- Presented at NeurIPS, 2025. (Spotlight, top 3%) *=equal contribution

SPEX: Scaling Feature Interaction Explanations for LLMs

Justin Singh Kang**, L. Butler*, A. Agarwal*, Y.E. Erginbas, R. Pedarsani, K. Ramachandra, B. Yu*

- Presented at ICML, 2025. *=equal contribution

Learning to Understand: Identifying Interactions via the Mobius Transforms

Justin Singh Kang*, Y.E. Erginbas, L. Butler, R. Pedarsani, K. Ramchandran*

- Presented at NeurIPS, 2024.

The Fair Value of Data Under Heterogeneous Privacy Constraints in Federated Learning

Justin Singh Kang*, R. Pedarsani and K. Ramchandran*

- Presented at NeurIPS Federated Learning in the age of Foundation Models Workshop (Oral), 2023. TMLR, 2024.

Minimum Feedback for Collision-Free Scheduling in Massive Random Access

Justin Singh Kang *and W. Yu*

- Published in IEEE Transactions on Information Theory, Dec 2021.
- 2024 IEEE ITSoc & ComSoc Best Paper Award, chosen among 1000+ papers in Communication and Information Theory IEEE Journals over 3 Years.

EDUCATION

PhD in EECS

With Prof. Kannan Ramchandran

GPA: 4.0, Berkeley Graduate Fellow

UC Berkeley, 2021-2026 (Expected)

MASc Electrical Engineering

With Prof. Wei Yu

GPA: 4.0, Canadian Graduate Scholarship

University of Toronto, 2021

BASc Engineering Physics

GPA: 91.1%, Activities: President of Engineering Physics Student Association

University of British Columbia, 2019

SELECTED AWARDS

- IEEE Information Theory Society and Communication Society Joint Paper Award, 2024
- Meta AI-BAIR Grant, 2023
- Berkeley Graduate Fellowship, 2021
- NSERC CGS-D (3rd nationally in area), 2021
- Trek Execellence Scholarship (5x), 2014-19
- Bycast Award, Entrepreneurship, 2018
- Donald J. Evans Scholarship (2x), 2017-18

SKILLS

Python, Git, Pytorch, Linux, Statistics, Hardware Design Verification

WORK EXPERIENCE

BOSCH AI: Research Intern

May 2025 – August 2025 — Sunnyvale, California

Building data cleaning and auto-labeling pipelines for vision and timeseries data with applications for BOSCH products, including autonomous driving and IoT systems. Data cleaning focused on using ideas from uncertainty quantification and ensemble learning to find label errors. Auto-labeling is a similar problem: start with unlabeled data and leverage a foundation model to label a subset of the data where it achieves low error. Mentors: Suraj Srinivas and Jorge Ono.

Google: Student Researcher

May 2024 – August 2024 — Mountain View, California

Designing explainable embeddings for software programs on Google cloud platforms. Data centers are complex evolving systems, and the resource needs of each program running in them is a constantly evolving time series. Developing embeddings for these timeseries can lead to optimized routing and performance prediction while maintaining human understanding. The research was part of SystemsResearch@Google and the project led by Prof. David Culler and Kun Lin.

Intel: Research and Development Engineering

April 2017 – August 2018 — Vancouver, Canada

Designing really fast and efficient hardware for cutting edge *Intel Optane* storage. This job was an awesome mix of theory and practical engineering that I strive for in my career. The research resulted in US Patent 11146289B2.