KIHO PARK

✉ parkkiho@uchicago.edu 🏠 kihopark.github.io 📍 Chicago, IL

Research Interests

Mechanistic Interpretability, Large Language Models (LLMs), Representation Learning, Information Geometry

Education

University of Chicago **Sep 2021 – Present**
PhD in Statistics *GPA: 4.0 / 4.0*
- Advisor: Victor Veitch
- Committee: Rina Foygel Barber, Ari Holtzman

Seoul National University **Mar 2015 – Jun 2021**
BS in Statistics and Mathematical Sciences *GPA: 4.12 / 4.3 (Summa Cum Laude)*
- Advisor: Sungkyu Jung

Publications

The Geometry of Categorical and Hierarchical Concepts in Large Language Models | *link* **2024**
- **K. Park**, Y. J. Choe, Y. Jiang, and V. Veitch
- **Oral (Top 1.8%)** Presentation at *ICLR 2025*
- **Best Paper Award** at *ICML 2024 Workshop on Mechanistic Interpretability (MI)*
- Poster Presentation at *ICML 2024 Workshop on Theoretical Foundations of Foundation Models (TF2M)*

The Linear Representation Hypothesis and the Geometry of Large Language Models | *link* **2023**
- **K. Park**, Y. J. Choe, and V. Veitch
- Poster Presentation at *ICML 2024*
- Oral Presentation at *NeurIPS 2023 Workshop on Causal Representation Learning (CRL)*

Clustering on the Torus by Conformal Prediction | *link* **2021**
- S. Jung, **K. Park**, and B. Kim
- *Annals of Applied Statistics*, vol. 15, no. 04, p. 1583 - 1603

Multiplication of Integral Octonions | *link* **2016**
- M. S. Kim, Y. Kim, J.-H. Lee, S. Nam, and **K. Park**
- *Journal of Algebra and Its Applications*, vol. 15, no. 08, p. 1650144

Preprint

Incorporating Hierarchical Semantics in Sparse Autoencoder Architectures | *link* **2025**
- M. Muchane, S. Richardson, **K. Park**, and V. Veitch

Research Experience

Victor Veitch's Lab **Sep 2022 – Present**
Graduate Research Assistant, advised by Victor Veitch *University of Chicago*
- Formalized the linear representation hypothesis in LLMs and introduced the concept of a 'causal inner product' that unifies the different notions of linear representations
- Formalized the representations of categorical concepts in LLMs as polytopes, and showed that hierarchical relations between concepts are represented as orthogonality

Netflix **Sep 2024 – Feb 2025**
Internship in Machine Learning & Inference Research Team *New York City*
- Developed a transformer-based point process model for high-dimensional event history data
- Mentored by James McInerney and Michael Lindon; supervised by Nathan Kallus

ML Alignment & Theory Scholar (MATS) **Oct 2024**
Research Scholar *Berkeley*
- Conducted research on the mechanistic understanding of reasoning in LLMs, in collaboration with Bruce W. Lee

Statistical Learning Theory Lab **Jun 2019 – Aug 2020**
Undergraduate Research Assistant, advised by Sungkyu Jung *Seoul National University*
- Created a novel clustering approach on the torus using the conformal prediction framework
- Constructed R code that allows others to apply the clustering method to data on the torus

Talks

Invited Talk at Insight + Interaction Lab	10.03.2025
Invited Talk at Najoung Kim's Reading Group	10.02.2025
Invited Talk at AI in Biomedicine Journal Club	08.28.2025
Invited Talk at Brett Beaulieu-Jones Lab Meeting	07.21.2025
Panel Discussion at KSEA IMPACTs 2025 (Advancing AI)	03.29.2025
Lightning Talk at KSEA UChicago Research Day	03.01.2025
Invited Talk at CMSC 37712 (Machine Learning and AI, invited by Ari Holtzman)	02.20.2025
Lightning Talk at DSI PhD Student Research Day	12.06.2024
Oral Presentation at US-Korea Conference (UKC 2024)	08.22.2024
Oral Presentation at ICML 2024 MI workshop | *link*	07.27.2024
Invited Talk at STAT 37400 (Nonparametric Inference, invited by Claire Donnat)	05.09.2024
Invited Talk at LLM Reading Group Seminar | *link*	03.05.2024
Invited Talk at Topology, algebra, and geometry in Data Science (TAG-DS) Seminar	02.27.2024
Oral Presentation at NeurIPS 2023 CRL workshop | *link*	12.15.2023

Honors and Awards

DAAD AINeT Fellowship, Postdoc-NeT-AI on Explainable AI	2025
Best Paper Award (*ICML 2024 Workshop on Mechanistic Interpretability*)	2024
Best Poster Award 1st Place on Student Research Poster Day (*University of Chicago*)	2024
The Winter 2023 Consulting Cup Team Award (*University of Chicago*)	2023
SNUAA Chicago Scholarship (*Seoul National University Alumni Association in Chicago*)	2022
The First Annual Takintayo Akinbiyi Memorial Award (*University of Chicago*)	2022
Dean's List (*Seoul National University*)	2018 – 2020
National Scholarship for Science and Engineering (*Korea Student Aid Foundation*)	2015 – 2020

Teaching Experience

Teaching Assistant, University of Chicago

• Instructor for Theoretical Statistics Preliminary Examination	Summer 2024, Summer 2023
• STAT 30200: Mathematical Statistics II	Spring 2023
• STAT 32950: Multivariate Statistical Analysis	Spring 2023, Spring 2022
• STAT 23400: Statistical Models and Methods	Fall 2022
• STAT 30040: Statistical Theory and Methods IIa	Winter 2022

Tutor, Seoul National University

• 326.311: Mathematical Statistics I (*Best tutor award*)	Spring 2021
• 326.312: Mathematical Statistics II	Fall 2020

Service

Peer Review

- ICLR 2026, NeurIPS 2025, ICML 2025, ICLR 2025, NeurIPS 2024, ICML 2024 TF2M Workshop, ICML 2024, ICLR 2024, CLeaR 2024, NeurIPS 2023 CRL Workshop, ICML 2023 Workshop SCIS, NeurIPS 2023

Mentoring   **Jun 2024 – Sep 2024**
UChicago Existential Risk Lab Fellowship   *Chicago*

- Mentored Anastasia Wei (undergraduate student at Northwestern University) on interpretability in LLMs
- Conducted research on mechanistic understanding of semantically similar sentences in LLMs

Leadership & Organizations

University of Chicago Korean Graduate Student Association (KGSA)   **Jun 2024 - May 2025**
Vice President   *Chicago*

Korean National Police Agency   **Nov 2016 – Aug 2018**
Auxiliary Police in 43 Company   *Seoul*