Bhuvesh **Kumar**

Senior Research Scientist · Ph.D. in Computer Science

kumarbhuvesh@gmail.com | www.bhuveshkumar.com | bhuvesh-kumar | Google Scholar

Education

Georgia Institute of Technology — *Atlanta, Georgia*

Ph.D. in Computer Science (Advisors: Dr. Jacob Abernethy and Dr. Jamie Morgenstern) — *2017 - 2022*

Georgia Institute of Technology — *Atlanta, Georgia*

M.S. in Computer Science (Specialization in Machine Learning, GPA: 4.0/4.0) — *2017 - 2022*

Indian Institute of Technology, Kanpur — *Kanpur, India*

B.Tech in Computer Science and Engineering (GPA: 9.7/10) — *2013 - 2017*

Work Experience

Snap Inc — *Bellevue, WA*

Senior Research Scientist — *Sept. 2026 - Present*

Research Scientist — *Sept. 2024 - Aug. 2026*

- Drive research and development in LLM/VLM-based generative recommendation, multimodal Semantic ID representations and content understanding, and long user-sequence modeling. Co-created the team's internal Python/PyTorch research library and now serve as its sole maintainer and technical lead; it is actively used by every research scientist on the team and 20+ MLEs.
- Developed multimodal Semantic ID content representations now used as features in 10+ production models across Content, Ads, Lens, and Search and as the target representation for Snap's generative retrieval models.
- Advanced LLM-based generative recommendation with Semantic IDs through research and development spanning continued pre-training (CPT), supervised fine-tuning (SFT), Ray-based reinforcement-learning post-training, and fast TensorRT-LLM inference, contributing to multiple production launches.
- Developed cross-domain universal user representations and large-scale long-sequence models using TorchRec and NVIDIA DynamicEmb with sharded embedding tables; the Universal User Embedding Model now provides foundational embeddings used by 15+ personalization models at Snap.
- Co-created GRID, the open-source subset of this platform, and now serve as sole maintainer of its Python library; GRID has 700+ GitHub stars, and its accompanying paper received the CIKM 2025 Best Paper Award.
- Co-developed AutoCDSR and AutoCDSR+ for cross-domain sequential recommendation, improving offline cross-domain recommendation accuracy by more than 20%.
- Co-developed and launched Snap's first unified cross-domain model across Discover, Lens, and Ads, increasing long-form video view time by 19%, Lens play time by 1.76%, and swipe rate by 0.72%.
- Designed and launched the Generalized Unbiased Watch Time model for Snapchat Stories, increasing total watch time by 1.6%; co-designed re-ranking improvements that increased view time by 1.89%, comments by 3.86%, and chat sends by 0.97%.
- Mentor research interns, research scientists, and machine-learning engineers in developing publishable ideas and production-grade systems.

TikTok / TikTok USDS — *Bellevue, WA*

Machine Learning Scientist — *Jan 2023 - Sep 2024*

- Led research and development in TikTok U.S. Core Recommendation, building large-scale sequential user models, ranking systems, and personalization algorithms.
- Introduced and owned a new final re-ranking stage that moved TikTok U.S. recommendation from point-wise item scoring to list/session-level Transformer sequence modeling, increasing total watch time by 2.4%.
- Designed personalized mixing of organic, cold-start, and commercial content and coordinated with horizontal teams to preserve distribution and business goals while optimizing long-term user retention.
- Led the migration of TikTok U.S. Friends, Hashtag, and Social recommendation services from a single data center to a disaster-tolerant multi-data-center design.
- Developed multi-objective ranking frameworks balancing engagement, user satisfaction, content diversity, and platform constraints.
- Ran large-scale experimentation and analysis with Spark, Flink, and Hive and implemented production models with PyTorch and TensorFlow.

Amazon — *Seattle, WA*

Applied Scientist Intern — *06/01/2021 - 08/20/2021*

- Conducted research on fairness and transparency in machine learning under the guidance of Prof. Michael Kearns and Dr. Matthäus Kleindessner
- Introduced a new notion of group fairness for multi-armed bandit algorithms, ensuring equitable treatment of user groups while maintaining system performance
- Developed and analyzed a no-regret algorithm that balances fairness and efficiency, providing theoretical guarantees on long-term outcomes and algorithmic stability
- Implemented and evaluated models using Python, PyTorch, and advanced statistical analysis frameworks to validate fairness-performance tradeoffs

University of Washington *Seattle, WA*
Visiting Research Scientist *01/25/2021 - 04/16/2021*

- Conducted research on algorithmic mechanism design and differential privacy to improve fairness, privacy, and efficiency in online economic systems
- Designed simple and incentive-compatible auction mechanisms for revenue maximization in repeated auctions, ensuring robustness against strategic bidder behavior while preserving user privacy
- Applied differential privacy principles to create mechanisms that safeguard sensitive bidder information without sacrificing revenue performance
- Research culminated in publication "Private Mechanism Design via Quantile Estimation" at ICLR 2025, co-authored with Yuanyuan Yang, Tao Xiao, and Jamie Morgenstern

Meta (Facebook) *Menlo Park, CA*
Research Intern, Core Data Science *08/17/2020 - 12/04/2020*

- Conducted research on budget optimization and bid pacing for large-scale online advertising auctions as part of Meta's Core Data Science team
- Developed algorithms for optimal spend plan prediction and automatic bid pacing, designed to dynamically adjust advertiser bids to maximize utility and maintain smooth budget expenditure across time
- Built learning-theoretic models for adaptive pacing and spend rate estimation, combining theoretical guarantees with practical applicability in Meta's ad auction framework
- Conducted extensive experiments on Facebook's real ad auction data, demonstrating significant increases in advertisers' utility and return on ad spend (ROAS) compared to baseline pacing methods
- Research culminated in paper "Optimal Spend Rate Estimation and Pacing for Ad Campaigns with Budgets" (arXiv:2202.05881), co-authored with Jamie Morgenstern and Okke Schrijvers and was presented at INFORMS Annual Meeting 2021

Meta (Facebook) *Menlo Park, CA*
Machine Learning Engineering Intern *05/06/2019 - 07/26/2019*

- Worked on Dynamic Product Ads (DPA) ranking within Facebook's advertising platform, optimizing for Return on Ad Spend (ROAS) and overall ad revenue through advanced deep learning techniques
- Developed and deployed a new deep learning architecture for ads and product ranking that improved prediction accuracy and model efficiency at scale
- Achieved significant increase in ad revenue while simultaneously reducing model memory usage by 60 TB (86% decrease), enabling faster inference and more cost-effective large-scale training
- Collaborated with cross-functional teams across Ads Infrastructure and Applied Machine Learning to integrate the model into Facebook's production pipeline

Georgia Institute of Technology *Atlanta, GA*
Graduate Research Assistant *08/21/2017 - 12/31/2022*

- Conducted doctoral research in machine learning for decision making, focusing on robust algorithms with provable guarantees in adversarial, online, and strategic environments
- Developed and analyzed algorithms using tools from online learning, algorithmic game theory, and differential privacy, advancing theoretical understanding of robustness and adaptivity in intelligent systems
- Published multiple peer-reviewed papers at top international conferences including NeurIPS, ICML, and ICLR, covering incentive-compatible mechanism design, fairness in bandits, and active learning
- Collaborated closely with advisors Dr. Jacob Abernethy and Dr. Jamie Morgenstern on research bridging machine learning, economics, and theoretical computer science
- Served as Teaching Assistant for graduate-level courses: Machine Learning Theory (CS 7545) and Machine Learning for Trading (CS 7646)
- Instructed graduate course "Online Decision Making in Machine Learning (ECE 8803)", designing lectures, assignments, and assessments for Ph.D. students
- Mentored undergraduate and graduate students on theoretical research, experiment design, and publication writing

Awards

2025 **Best Paper Award**, 34th ACM Conference on Information and Knowledge Management (CIKM)
2019 **NeurIPS Travel Award**, 34th Annual Conference on Neural Information Processing Systems
2017 **Chair's Fellowship**, The School of Computer Science, Georgia Institute of Technology
2014-16 **Academic Excellence Award (Dean's List)**, IIT Kanpur
2011 **KVPY Fellowship**, Government of India
2009 **NTSE Scholarship**, Government of India

Research

Peer-Reviewed & Accepted Publications

- **[14] Semantic IDs for Recommender Systems at Snapchat: Use Cases, Technical Challenges, and Design Choices**
  ACM SIGIR Conference on Research and Development in Information Retrieval **(SIGIR 2026)**
  Clark Mingxuan Ju, Tong Zhao, Leonardo Neves, Liam Collins, Bhuvesh Kumar, Jiwen Ren, Lili Zhang, Wenfeng Zhuo, Vincent Zhang, Xiao Bai, Jinchao Li, Karthik Iyer, Zihao Fan, Yilun Xu, Yiwen Chen, Peicheng Yu, Manish Malik, and Neil Shah

- **[13] Exploiting ID-Text Complementarity via Ensembling for Sequential Recommendation**
  ACM Conference on Recommender Systems **(RecSys 2026)**
  Liam Collins, Bhuvesh Kumar, Clark Mingxuan Ju, Tong Zhao, Donald Loveland, Leonardo Neves, and Neil Shah
- **[12] CoSearch: Joint Training of Reasoning and Document Ranking via Reinforcement Learning for Agentic Search**
  Conference on Language Modeling **(COLM 2026)**
  Hansi Zeng, Liam Collins, Bhuvesh Kumar, Neil Shah, and Hamed Zamani
- **[11] Training-Free LLM-Based Recommendation with Post-LLM Item Refinement Using Collaborative Signals**
  ACM International Conference on Information and Knowledge Management **(CIKM 2026)**
  Kyungho Kim, Sunwoo Kim, Geon Lee, Shinhwan Kang, Sojeong Kim, Liam Collins, Bhuvesh Kumar, Donald Loveland, and Kijung Shin
- **[10] Sequential Data Augmentation for Generative Recommendation**
  19th ACM International Conference on Web Search and Data Mining **(WSDM 2026)**
  Geon Lee, Bhuvesh Kumar, Clark Mingxuan Ju, Tong Zhao, Kijung Shin, Neil Shah, and Liam Collins
- **[9] Learning Universal User Representations Leveraging Cross-Domain User Intent at Snapchat**
  48th International ACM SIGIR Conference on Research and Development in Information Retrieval **(SIGIR 2025)**
  Clark Mingxuan Ju, Leonardo Neves, Bhuvesh Kumar, Liam Collins, Tong Zhao, Yuwei Qiu, Qing Dou et al.
- **[8] Generative Recommendation with Semantic IDs: A Practitioner's Handbook**
  34th ACM International Conference on Information and Knowledge Management **(CIKM 2025)** **Best Paper Award**
  Clark Mingxuan Ju, Liam Collins, Leonardo Neves, Bhuvesh Kumar, Louis Yufeng Wang, Tong Zhao, and Neil Shah
  Co-creator of GRID (700+ stars on GitHub)
- **[7] Private Mechanism Design via Quantile Estimation**
  Thirteenth International Conference on Learning Representations **(ICLR 2025)**
  Yuanyuan Yang, Tao Xiao, Bhuvesh Kumar, and Jamie Morgenstern
- **[6] Revisiting self-attention for cross-domain sequential recommendation**
  31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining **(KDD 2025)**
  Clark Mingxuan Ju, Leonardo Neves, Bhuvesh Kumar, Liam Collins, Tong Zhao, Yuwei Qiu, Qing Dou, Sohail Nizam, Sen Yang, and Neil Shah
- **[5] Accelerated Federated Optimization with Quantization**
  Data Engineering Bulletin 46, IEEE Computer Society **(IEEE Data Engineering Bulletin 2023)**
  Yeojoon Youn, Bhuvesh Kumar, and Jacob Abernethy
  Also presented as oral at Workshop Federated Learning at NeurIPS 2022
- **[4] ActiveHedge: Hedge meets Active Learning**
  The Thirty-Ninth International Conference on Machine Learning **(ICML 2022)** (Spotlight)
  Bhuvesh Kumar, Jacob Abernethy, and Venkatesh Saligrama
  Also accepted as an Oral Talk at Complex Feedback in Online Learning Workshop, ICML 2022
  Also presented at Adaptive Experimental Design and Active Learning in the Real World Workshop, ICML 2022
- **[3] Observation Free Attacks on Stochastic Bandits**
  Thirty-Fifth Annual Conference on Neural Information Processing Systems **(NeurIPS 2021)**
  Yinglun Xu, Bhuvesh Kumar, and Jacob Abernethy
- **[2] Bridging Truthfulness and Corruption-robustness in Multi-Armed Bandit Mechanisms**
  Incentives in Machine Learning Workshop at ICML 2020
  Yinglun Xu, Bhuvesh Kumar, Jacob Abernethy, Thodoris Lykouris
- **[1] Learning Auctions with Robust Incentive Guarantees**
  Thirty-Third Annual Conference on Neural Information Processing Systems **(NeurIPS 2019)**
  Bhuvesh Kumar, Jacob Abernethy, Rachel Cummings, Jamie Morgenstern, Samuel Taggart
  Preliminary version presented at the Learning in Presence of Strategic Behavior Workshop, EC 2019

Preprints

- **[2] LLM-Based Generative Retrieval for Snapchat Content Recommendation**
  Preprint, 2026
  Liam Collins, Jiwen Ren, Donald Loveland, Bhuvesh Kumar, Clark Mingxuan Ju, Xuan Guo, Mo Li, Alvin Hou, Yi Cui, Peng Yang, Jian Wang, Saud Afzal Shafi, Nga Than, Ruiming Lu, Wenfeng Zhuo, Dongheng Li, Lili Zhang, Mingtao Zhang, Jinchao Ye, Vincent Xue, Chunhui Zhu, and Neil Shah
- **[1] Optimal spend rate estimation and pacing for ad campaigns with budgets**
  arXiv preprint arXiv:2202.05881 **(Presented at INFORMS 2021)**
  Bhuvesh Kumar, Jamie Morgenstern, and Okke Schrijvers

Software Packages

- **[1] GRID: Generative Recommendation with Semantic IDs**
  2025 Release **700+ stars on GitHub**
  Developers: Clark Mingxuan Ju, Liam Collins, Bhuvesh Kumar, and Leonardo Neves

Talks and Presentations

- Doordash, Ads Gen AI Workshop, 2025**, Generative Recommendation with Semantic IDs**
- Spotlight Talk at ICML 2022**, ActiveHedge: Hedge meets Active Learning**
- ReALML workshop at ICML 2022**, ActiveHedge: Hedge meets Active Learning**
- Complex Feedback in Online Learning at ICML 2022**, ActiveHedge: Hedge meets Active Learning**

- Incentives in ML at ICML 2020, **Bridging Truthfulness and Corruption-Robustness in Multi-Armed Bandit Mechanisms**
- INFORMS Annual Meeting 2021, **Optimal Spend Rate Estimation And Pacing For Ad Campaigns With Budgets**
- JHU CS Theory Seminar 2021, **Optimal Spend Rate Estimation And Pacing For Ad Campaigns With Budgets**
- Learning in Presence of Strategic Behavior at EC 2019, **Learning Auctions with Incentive Guarantees**

Academic & Professional Service

Reviewer / Program Committee Member

- Association for Computational Linguistics (ACL), 2026
- International Conference on Learning Representations (ICLR), 2025, 2026
- Conference on Neural Information Processing Systems (NeurIPS), 2020, 2021, 2025
- Conference on Learning Theory (COLT), 2020, 2021, 2023
- Annual Symposium on Discrete Algorithms (SODA), 2021, 2023
- International Conference on Machine Learning (ICML), 2022
- Journal of Machine Learning Research (JMLR), 2021
- International Conference on Algorithmic Learning Theory (ALT), 2018, 2020

Seminar Organizer

- Algorithms, Combinatorics, & Optimization Student Seminar, Georgia Institute of Technology , 2020-2022

Competition Judge

- Judge DubHacks, University of Washington's Hackathon, 2025
- Robotics Judge at Regeneron International Science and Engineering Fair (ISEF), 2022

Professional Memberships

- Member, Association for Computing Machinery (ACM)
- Member, Institute of Electrical and Electronics Engineers (IEEE)

Teaching

Course Instructor

Online Decision Making in Machine Learning (ECE 8803) – Georgia Tech *Fall 2022*

Teaching Assistant

Machine Learning for Trading (CS 7646) – Georgia Tech *Spring 2022, Summer 2022*

Machine Learning Theory (CS 7545) – Georgia Tech *Fall 2018, Fall 2019*

Introduction to Computing (ESC 101) – IIT Kanpur *Fall 2016, Spring 2017*

Leadership & Community

Student Leadership

- Founding Vice President of the School of Computer Science Graduate Student Association, Georgia Tech (2021-2022)
- Student Nominee in Faculty Hiring Committee at School of Computer Science, Georgia Tech (2020-2021)
- Coordinator of the Programming Club, IIT Kanpur (2015-2016)
- Leader of Science Coffeehouse, IIT Kanpur (2016-2016)

Community Service

- Organizer of ML at GT Social Hour, Georgia Tech (2021)
- Organizer of the College of Computing Happy Hour, Georgia Tech (2018-2019)
- Co-organizer of the PhD visit weekend, School of Computer Science, Georgia Tech (2020)
- Mentor at HackGT Hackathon, Georgia Tech (2018, 2019)