Sudhanshu Agrawal

+1 (310) 500-8571 Palo Alto, California sudhagra@stanford.edu sudhanshuagrawal.com Google Scholar

Education

Stanford University **Sept 2026 – June 2028**
M.S. Computer Science

University of California, Los Angeles **Sept 2019 – June 2023**
B.S. Computer Science *(Magna cum laude)* B.S. Mathematics *(Cum laude)* **GPA: 3.92**

Academic Research

- UCLA Computer Science: undergraduate research, advised by Professor Aditya Grover.
- UCLA Mathematics: undergraduate research, advised by Professors Levon Nurbekyan and Samy Wu Fung.

Work Experience

Qualcomm : ***ML Research Engineer*** **2023 – 2026** *(San Diego)*

- **LLM efficiency research:** Conducted research on new techniques to accelerate the generation speed of LLMs via speculative decoding. 4 conference papers published, and 12 inventions with multiple US/global patent applications pending.
- **Research commercialization:** Innovations deployed on Qualcomm devices for public demos and commercial products.

Qualcomm : ***ML Engineering Intern*** **Summer 2022** *(San Diego)*

- **Novel *Python* profiler:** created a profiling tool for deep learning applications, capable of timing code function-by-function, line-by-line, and profiling multiprocessing applications.

SonicJobs : ***ML Engineering Intern*** **Summer 2021** *(Remote)*

- **Synthetic data-set generation:** created a pipeline to scalably generate over 100,000 *HTML/CSS* webpages with varying contents and styles, solving the challenge of obtaining a diverse, labeled training data-set for a computer vision model.

Reliance Jio : ***ML & Data Science Intern*** **Summer 2020** *(Remote)*

- **Predicting the properties of hydrocarbons:** developed predictive models leveraging lasso regression, ridge regression, support vector regression, recursive feature elimination, linear discriminant analysis, and 1-D convolutions.

Publications

Speculative Decoding for Diffusion LLMs **2026**

- **Agrawal, Sudhanshu**, et al., "Structuring The Future: Diffusion LLM Speculative Decoding via Calibrated Draft Graphs." **ICML 2026 FoGen and SPIGM Workshops**.

Representations in AR-LLMs vs dLLMs **2026**

- Raghavv Goel*, Risheek Garrepalli*, **Sudhanshu Agrawal**, et al., "A Comparative Analysis of Layer-wise Representational Capacity in AR and Diffusion LLMs." **ICLR 2026 Workshop on Science for Deep Learning**.

Efficient Speculative Decoding via Vocabulary Dimensionality Reduction **2025**

- Goel, Raghavv, **Agrawal, Sudhanshu**, et al., "VOCABTRIM: Vocabulary Pruning for Efficient Speculative Decoding in LLMs." **ICML 2025 Workshop on Efficient Systems for Foundation Models**.

Early Draft Stopping for Speculative Decoding of LLMs **2024**

- **Agrawal, Sudhanshu**, Jeon, Wonseok, and Lee, Mingu, "AdaEDL: Early Draft Stopping for Speculative Decoding of Large Language Models via an Entropy-based Lower Bound on Token Acceptance Probability." **NeurIPS 2024 Efficient Natural Language and Speech Processing Workshop.**

Foundation Models for Offline Black-Box Optimization **2023**

- Nguyen, Tung, **Agrawal, Sudhanshu**, and Grover, Aditya, "Expt: Synthetic pretraining for few-shot experimental design." **Advances in Neural Information Processing Systems 36 (2023): 45856-45869 (NeurIPS 2023)**.

Efficiently Solving High Dimensional Non-local Mean Field Game Problems **2022**

- **Agrawal, Sudhanshu**, Lee, Wonjun, Fung, Samy W., Nurbekyan Levon, "Random features for high-dimensional nonlocal mean-field games." **Journal of Computational Physics 459 (2022): 111136**.