

Sachin Yadav

Pre-Doctoral Researcher
Machine Learning and Optimization (MLO) Team, Frontier AI Unit
Google DeepMind

✉ sachinyadav7024@gmail.com
🌐 yadav-sachin.github.io
🔗 scholar.google/AhsU8KUAAAAJ

Education

Indian Institute of Technology (IIT) Gandhinagar

2018 - 2022

B.Tech in Computer Science and Engineering

GPA: 9.26/10.0

Mount Abu Public School, Delhi

2017

Central Board of Secondary Education (CBSE) Class XII

Percentage: 96.0%

Work Experience

Google DeepMind

October 2024 - Present

Pre-Doctoral Researcher in Frontier AI Unit

Advisors: Dr. Dheeraj Nagaraj, Dr. Karthikeyan Shanmugam, Dr. Prateek Jain

- Formulated and implemented a novel pre-training objective for diffusion language models, achieving an **8x inference speedup** over state-of-the-art academic baselines, resulting in **adoption by the Gemini Diffusion** team, a **provisional patent**, and a **Spot Bonus** awarded for research contributions.
- Developed a novel framework for unified discrete-continuous diffusion, enabling robust constrained generation:
 - Achieved state-of-the-art **performance** on molecular structures, layout generation, and tabular data.
 - Demonstrated superior logical reasoning on the **SAT problem**, significantly outperforming Autoregressive and Diffusion baselines with robust **scaling behavior** on complex instances.
- Contributing to research on large-scale machine learning efficiency for training and inference.
- Delivered high-impact research driving advancements across DeepMind products, resulting in publications and patents.

Microsoft Research Lab - India

July 2024 - October 2024

Research Intern

Advisor: Dr. Amit Sharma, Principal Researcher

Worked on enhancing the causal reasoning capabilities of large language models by integrating causal analysis tools within an agent-based system.

Microsoft Research Lab - India

July 2022 - July 2024

Research Fellow (AI Resident) in Machine Learning and AI Group

Advisor: Dr. Manik Varma, Distinguished Scientist & Vice President

Developed large-scale machine learning models for extreme multi-label classification (industry-scale search, retrieval and recommendation systems) leading to top-tier publications and impact across Microsoft products.

Samsung Research Institute Bangalore

Summer 2021

Research Intern in Advanced Technology Lab (ATL) Group

Advisor: Dr. Vijay Narayan Tiwari, Director of Software Engineering

Worked on machine learning algorithms leveraging Impulse Radio Ultra-Wideband (IR-UWB) radar sensors for human motion recognition, contributing to the advancement of next-generation Samsung devices.

Talks/Presentations

1. Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval. [Slides]

Presenter: **Sachin Yadav**; Session: Ranking and Retrieval.

ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2024, Barcelona. **Oral Presentation.**

Publications

* - equal contribution

1. Gemini 2.5 Technical Report.

Gemini Team (including **Sachin Yadav**), Google DeepMind

arXiv Preprint, 2025

2. Interleaved Gibbs Diffusion for Constrained Generation.

Gautham Govind Anil, **Sachin Yadav**, Dheeraj Nagaraj, Karthikeyan Shanmugam, Prateek Jain
International Conference on Learning Representations (ICLR) (DeLTa Workshop), 2025

3. On the Necessity of World Knowledge for Mitigating Missing Labels in Extreme Classification.

Jatin Prakash*, Anirudh Buvanesh*, Bishal Santra, Deepak Saini, **Sachin Yadav**, Jian Jiao, Yashoteja Prabhu, Amit Sharma, Manik Varma
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2025.

4. Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval.

Sachin Yadav⁺*, Deepak Saini*, Anirudh Buvanesh*, Bhawna Paliwal, Kunal Dahiya, Siddarth Asokan, Yashoteja Prabhu, Jian Jiao and Manik Varma
⁺ - led the project
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2024. Oral Presentation

5. Deep Encoders with Auxiliary Parameters for Extreme Classification.

Kunal Dahiya, **Sachin Yadav**, Sushant Sondhi, Deepak Saini, Sonu Mehta, Jian Jiao, Sumeet Agarwal, Purushottam Kar, and Manik Varma
ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2023.

6. Deep Gaussian Processes for Air Quality Inference. Extended Abstract

Aadesh Desai*, Eshan Gujarathi*, Saagar Parikh*, **Sachin Yadav***, Zeel Patel, and Nipun Batra
Young Researchers' Symposium @ Joint International Conference on Data Science & Management of Data (CODS-COMAD), 2023

Research Adoption

‡ - Specific details withheld due to non-disclosure agreements.

1. Gemini Diffusion: State-of-the-art text diffusion model from Google DeepMind[‡]

Formulated and implemented a novel pre-training objective for the text diffusion model, driving substantial improvements in validation metrics. Awarded a **Spot Bonus** for delivering technical advancements to the project.

2. Search Engine: Sponsored Ads [‡]

EMMETT: Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval.
On live traffic from a popular search engine, EMMETT resulted in a 4.2% increase in Click-Through Rate (CTR) and a 0.9% reduction in Quick Back Rate (QBR) for Sponsored Ads.

3. Personalized Ad Recommendations [‡]

EMMETT: Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval.
In live trials on a popular Display Ads Platform, EMMETT achieved a 3.3% improvement in Click-Through Rate (CTR) and a 4.9% improvement in Click Yield for personalized ads targeting users.

Patents

1. **Discrete Diffusion Neural Networks with Planning.** 2025. (Provisional Patent Filed, Google DeepMind).
Inventors: **Sachin Yadav**, Dheeraj Nagaraj, Karthikeyan Shanmugam
2. **Interleaved Gibbs Diffusion for Constrained Generation.** 2025. (Provisional Patent Filed, Google DeepMind).
Inventors: Dheeraj Nagaraj, Karthikeyan Shanmugam, **Sachin Yadav**, Gautham Govind Anil, Prateek Jain
3. **Improved Retrieval of Novel Keywords for Search.** 2024. (Granted, Microsoft, US Patent 12499168).
Inventors: Deepak Saini, Jian Jiao, **Sachin Yadav**, Anirudh Buvanesh, Bhawna Paliwal, and Manik Varma

Software

1. **Gemini Diffusion: State-of-the-art text diffusion model from Google DeepMind.**
Contributor to the Gemini Diffusion Team, Google DeepMind.
2. **TinyGP: Extremely Lightweight Library for Building Gaussian Process (GP) Models in Python.**
Dan Foreman-Mackey, **Sachin Yadav**, Andrew Fowlie, René Tronsgaard, Steve Schmerler, Thomas Killestein.
Stats as of November 2025: ★ 320

Benchmarks

1. ORCAS-800K

Benchmark for mapping user queries on the Bing search engine to the relevant subset of 800K web URLs.

Selected Awards and Honors

- Awarded a Spot Bonus at Google DeepMind for exemplary technical contributions to [Gemini Diffusion](#). 2025
- Awarded the [Ashok Jain Scholarship](#) for merit, given to one B.Tech. student yearly at IIT Gandhinagar. 2021 & 2022
- Selected for ACM-ICPC Regionals programming competition. 2019
- Featured in [Dean's List](#) for semesters 1, 2, and 3 for exceptional academic performance at IIT Gandhinagar.
- Awarded [KVPY Fellowship](#) from the Department of Science and Technology, Government of India. 2016

Selected Projects

Gemini Text Diffusion

Advisors: Dr. Dheeraj Nagaraj, Dr. Karthikeyan Shanmugam, Dr. Prateek Jain, Dr. Brendan O'Donoghue, *Google DeepMind* — [Provisional Patent Filed, 2025]

- Contributed to Gemini Diffusion, a diffusion-based LLM that generates text via iterative noise refinement, matching state-of-the-art performance of large autoregressive models.
- Led innovations in the pre-training objective for Gemini Diffusion, resulting in significant quality gains in text generation.

Interleaved Gibbs Diffusion (IGD): Generating Discrete-Continuous Data with Implicit Constraints

Advisors: Dr. Dheeraj Nagaraj, Dr. Karthikeyan Shanmugam, Dr. Prateek Jain, *Google DeepMind* — [ICLR DeLTA Workshop 2025]

- Introduced Interleaved Gibbs Diffusion (IGD), a generalized Markov chain framework that theoretically guarantees exact reversal of forward processes and flexibly combines discrete and continuous denoisers for robust hybrid data generation.
- Empirically demonstrated significant out-of-the-box improvements in constrained tasks like 3-SAT, showcasing IGD's ability to handle implicit constraints in logical reasoning without domain-specific adaptations.
- Achieved leading performance on molecular structure, layout, and tabular data generation tasks using IGD, relying solely on general modeling strategies rather than task-specific inductive biases.

EMMETT: Extreme Meta-Classification for Large-Scale Zero-Shot Retrieval

Advisor: Dr. Manik Varma, *Microsoft Research India* — [KDD 2024]

- Developed EMMETT, an innovative framework for synthesizing classifiers for novel items in zero-shot retrieval, enhancing both accuracy and efficiency.
- Introduced IRENE, an effective instantiation of EMMETT suited for large-scale deployments, achieving up to 15% improvement in zero-shot retrieval accuracy and seamlessly integrating with existing Siamese encoders.
- Formulated a theoretical framework to guide algorithm and training strategy design for large-scale zero-shot retrieval, ensuring robust generalization across diverse applications.
- Empirically demonstrated the effectiveness of IRENE through comprehensive experiments and online A/B testing in a major search engine, resulting in a 4.2% increase in ad click-through rate.

DEXA: Deep Encoders with Auxiliary Parameters for Extreme Classification

Advisors: Prof. Purushottam Kar and Dr. Manik Varma, *Microsoft Research India* — [KDD 2023]

- Identified semantic gap and data paucity issues in Extreme Classification settings, leading to suboptimal encoder training.
- Proposed DEXA, a lightweight alternative that enhances Encoder embeddings with extra auxiliary parameters on the label side, addressing the identified training issues.
- Conducted comprehensive experiments with various architectures, showcasing DEXA's modularity within existing XC solutions, all achieved with minimal time and memory overheads during training.
- Attained accuracy improvements of 6% and 15% on public and proprietary benchmarks respectively.

Language Models with Causal Analysis Tools

Advisor: Dr. Amit Sharma, *Microsoft Research India*

- Integrated LLMs with causal analysis tools in an agent-based framework for problems across multiple levels of reasoning.
- Enhanced accuracy by training LLMs to effectively utilize multiple tools for causal inference tasks in natural language.

Adaptive Sparse Training of Large Networks on GPUs

Advisor: Prof. Anirban Dasgupta, IIT Gandhinagar

- Explored adaptive sparse training of large neural networks using approximate nearest neighbor search algorithms (HNSW and IVF-PQ) to accelerate training on CPU hardware.
- Proposed a customized variant of adaptive sparse training specifically designed for GPU-based training.

Teaching, Services and Responsibilities

- *Teaching Assistantship* - Computer Science and Engineering, IIT Gandhinagar

Introduction to Data Science - Prof. Anirban Dasgupta

Spring 2022

- * Developed collaborative open-source course notes in an online format with interactive code demos. [GitHub]
- * Designed data-driven assignments enabling students to perform interactive data analyses through thematic storytelling. [Submissions]
- * Worked alongside the professor to facilitate engaging discussions, design assignments, and support students' learning objectives.

- *Event Organizer* - Annual Coding Hackathon of IIT Gandhinagar.

2019 & 2020

Organized and managed technical problem statements in the hackathon, coordinating with participants, sponsors, and mentors to ensure successful events.

- *Founding Core Member* - Group for Algorithms and Sport Programming (GRASP), IIT Gandhinagar.

2019

Co-founded the student club dedicated to competitive programming, organizing workshops and intra-collegiate competitions.

- *Event Organizer* - D'Code - Competitive Programming Contest, Amalthea Tech Summit, IIT Gandhinagar

2019

Designed contest problems and organized the programming competition as part of Amalthea, the Annual Tech Summit of IIT Gandhinagar.

Skills and Courses

Programming Languages: Python, C++, C, JavaScript, MATLAB, SQL

Libraries and Frameworks: JAX, PyTorch, Flax, PaxML, Node.js

Relevant Courses: Advanced Machine Learning, Data Science, Data Structures and Algorithms, Discrete Mathematics, Linear Algebra, Probability and Statistics, Differential Equations, Operating Systems, Computer Organization and Architecture, Databases, Modeling and Simulation of Complex Systems, Methods of Social Science Experiments