Ajinkya Kiran Mulay

ML Research Scientist • AI Safety & Interpretability

(765) 409-7857 | ajinkyamulay123@gmail.com | linkedin.com/in/ajinkyamulay | github/thehimalayanleo | Google Scholar

Research Interests

Frontier AI Safety, Interpretability, and Evaluations: I focus on mechanistic understanding of neural networks and building rigorous evaluation frameworks to detect and control harmful behavior in high-capability systems. During my PhD, I developed theoretically grounded sparsification methods under differential privacy and federated learning constraints, achieving near-SOTA performance with only 8–10% of dense parameters. At Meta Integrity, I design and deploy LLM-driven safety systems at scale across Facebook, Instagram, and Threads.

Experience

Meta Platforms, Inc — *Menlo Park, CA*
Senior Research Scientist (AI Safety / Integrity Evals) — Feb 2026 – Present

- Own safety evaluations for integrity LLMs, building automated red-teaming/robustness suites (prompt-injection variants, policy-bypass probes, adversarial scenarios) to prevent regressions pre-rollout.
- Develop user-experience metrics and operationalize them via stratified evaluation and holdouts to understand user retention patterns.
- Translate new attacker patterns into targeted RL evals and related mitigations, closing the loop from incident signal spikes to launch decisions.
- **Stack:** PyTorch, Evals, AI safety, LLM tooling, Red teaming.

Meta Platforms, Inc — *Menlo Park, CA*
Research Scientist — Feb 2024 – Feb 2026

- Designed and deployed LLM systems for large-scale spam/scam detection across Facebook, Instagram, and Threads, achieving **4x** TPR improvements with detection latency under **4 hours**.
- Fine-tuned large (**>70B**) language models with retrieval-augmented safety workflows and production instrumentation to detect spam anomalies.
- Built large-scale behavior monitoring with LLM-assisted labeling and DBSCAN-style clustering to surface coordinated/novel attacker behaviors; curated safety-critical datasets (**>1M** samples) for post-training and fine-tuning.
- **Stack:** PyTorch, PHP, Integrity, Spam.

ECE, Purdue University — *West Lafayette, IN*
Graduate Research Assistant — Jun 2023 – Dec 2023

- Authored *SPriFed-OMP*, a differentially private federated learning algorithm for sparse basis recovery in high-dimensional regimes ($p \gg n$). Achieved accurate recovery with $n = \mathcal{O}(\sqrt{p})$ samples using only 8–10% of model parameters.
- Combined OMP-style coordinate selection with differential privacy noise injection, isolating informative parameter subsets aligned with interpretable functional components.
- **Stack:** PyTorch, Differential Privacy, Federated Learning, Sparse Learning.

Meta (Facebook) — *Menlo Park, CA*
PhD Software Engineering Intern — May 2022 – Aug 2022

- Designed and deployed a modular end-to-end production stack for Federated Semi-Supervised Learning (FSSL) vision tasks.
- Replicated performance benchmarks with FixMatch and SimCLR on real devices.
- **Stack:** C++, TorchScript, Python, PyTorch.

Meta (Facebook) — *Menlo Park, CA*
PhD Software Engineering Intern — May 2021 – Aug 2021

- Developed a fast, scalable private ML algorithm using PCA with differential privacy, outperforming SOTA by **15%** (test accuracy).
- Improved performance-to-privacy trade-off by more than **35%** via varying tree restarts for DP-FTRL.
- **Stack:** Python, PyTorch, Differential Privacy, Federated Learning.

SuperPower Research, Psychological Sciences, Purdue *West Lafayette, IN*
Graduate Research Assistant Aug 2020 – May 2023

- Designed a statistical-power modeling engine (NIH-funded) achieving <5% error while reducing computation by **90%** vs. SOTA.
- Proposed semi-supervised data augmentation and dimensionality reduction methods improving engine stability and power estimation variance.
- **Stack:** PyTorch, R, Bayesian Learning, Hypothesis Testing, Differential Privacy.

Other Research Experience

AI Safety Camp *Remote*
Community Member Jan 2026 – Present

- Collaborate with a cross-disciplinary group modeling frontier AI risk, focusing on how "AI slop" (high-volume, low-signal synthetic content) is created.
- Design and run evaluation frameworks to detect and quantify AI slop across platforms and modalities.
- Conduct interpretability analyses on how slop-like data distributions affect internal model representations (neuron/feature activations).

BlueDot Impact Technical AI Safety Course *Remote*
Participant 2025

- Completed structured coursework on technical and governance dimensions of AI safety, covering alignment, interpretability, and catastrophic risk from frontier systems.

OpenMined *Remote*
Community Member Mar 2020 – Mar 2023

- Explored the relationship between Differential Privacy and Adversarial Robustness; quantified DP/FL impact on real-world systems (FedPerf).
- **Stack:** PyTorch, PySyft, Git.

Education

Purdue University *West Lafayette, IN*
PhD in Electrical and Computer Engineering Aug 2018 – Aug 2024

- Advised by Prof. Xiaojun Lin | GPA: 3.6/4.0
- Thesis: Developed private and non-private sparse learning algorithms with provable convergence under extreme sample scarcity, achieving near-SOTA accuracy with $\leq$10% of dense parameters.

IIT Hyderabad *Hyderabad, India*
B.Tech (with Honors) in Electrical Engineering Aug 2014 – May 2018

- Advised by Prof. Bheemarjuna Reddy | GPA: 8.88/10
- Research Focus: Inference-aware game-theoretic framework for unlicensed LTE and Wi-Fi bands.

Skills

Frontier LLM Systems	Fine-tuning $\geq$70B models, RAG-based safety workflows, alignment evaluation.
Efficient ML	Sparse training, parameter reduction, structured optimization, extreme low-sample regimes ($n \sim \sqrt{p}$).
Safety & Integrity	Large-scale spam/scam detection, high-TPR safety filters, behavior anomaly detection.
Theory	Convergence proofs, DP-SGD, private PCA, sparse recovery, rational iteration for optimizer orthogonalization.
Kernel Engineering	Triton, block-sparse FlashAttention, CSR-format sparsity masks, H100 kernel optimization.
Technical	PyTorch, Triton, C++, Golang, TorchScript, PHP, distributed ML pipelines.

Projects

Halley-Gram-Muon: Cubic Convergence for Optimizer Orthogonalization *Remote*
Independent Research Jan 2026 – Present

- Replaced Newton-Schulz polynomial in GramMuon (used in Kimi K2, GLM-5) with a Halley rational iteration in Gram space achieving cubic convergence. Pareto-dominant over GramMuon T=5 on WikiText-103 and enwiki8: better model quality and 16–29% faster wall-clock.
- **Stack:** PyTorch, Triton, language model training.

Activation Sparsity as a Scheming Signal — Apart Research AI Control Hackathon *Remote*
Hackathon | Project Link Mar 2026

- Mechanistic, output-independent scheming monitor: scheming agents produce more uniform MLP activations (lower Gini) across 15/18 layers. Threshold detector achieves AUROC $= 0.745$ with no output access ($p = 0.0003$, Cohen's $d = 0.701$), consistent across Qwen2.5 1.5B and 3B.
- **Stack:** PyTorch, Qwen2.5, mechanistic interpretability, activation probing.

Block-Sparse Causal Attention Kernel — Paradigm Attention Kernel Challenge *Remote*
Competition | GitHub Link Apr 2026 – Present

- Triton kernel for block-sparse causal attention on H100 (CSR mask, `head_dim=128`). Applied OMP column-reuse intuition: KV-block reuse tiling across query row groups sharing support in sliding-window/banded patterns.
- **Stack:** Triton, PyTorch, H100, block-sparse FlashAttention.

Sparse Structured Agent for Financial Document QA — Sentient Arena OfficeQA *Remote*
Competition | Challenge Link Mar 2026 – Present

- Agentic reasoning system for multi-hop QA over Treasury Bulletin corpora, grounded in Sanskrit Kāraka grammar: query roles (agent, object, instrument) map to structured retrieval steps, with OMP-based decomposition identifying the minimal supporting document set.
- **Stack:** Goose, Python, OMP-style retrieval, agentic reasoning, Sanskrit Karaka grammar.

Journal Publications

- Ajinkya K Mulay, Xiaojun Lin. "SPriFed-OMP: A Differentially Private Federated Learning Algorithm for Sparse Basis Recovery." *Transactions of Machine Learning*, Purdue University.
- Ajinkya K Mulay, Anand Basawade, Bheemarjuna Tamma, Anthony Franklin. "DFC: Dynamic UL-DL Frame Configuration for Improving Channel Access in eLAA." *IEEE Networking Letters*, IIT Hyderabad.

Early Research Work

- Ajinkya K Mulay, Xiaojun Lin. "Humming-Bird: A Forward-Backward Based Differentially Private Federated Learning Algorithm for Sparse Basis Recovery." *In Review*, Purdue.
- Ajinkya K Mulay, Xiaojun Lin. "Humming-Bird+: Batched and General Differentially Private FL Algorithms for Sparse Basis Recovery." *In Preparation*, Purdue.

Workshop Presentations

- Ajinkya Mulay, Sean Lane, Erin Hennes. "PowerGraph: Using neural networks and principal components to multivariate statistical power trade-offs." *AI for Science Workshop, ICML, July 2022* (Non-Archived). SuperPower Lab, Purdue.
- Ajinkya Mulay, Sean Lane, Erin Hennes. "Private Hypothesis Testing for Social Sciences." *Theory and Practice of Differential Privacy Workshop, ICML, July 2022* (Non-Archived). SuperPower Lab, Purdue.
- Rakshit Naidu, Harshita Diddee, Ajinkya Mulay, Aleti Vardhan, Krithika Ramesh, Ahmed Zamzam. "Towards Quantifying the Carbon Emissions of Differentially Private Machine Learning." *Socially Responsible ML Workshop, ICML, July 2021* (Non-Archived). OpenMined.
- Ajinkya Mulay, Tushar Semwal, Ayush Agrawal. "FedPerf: A Practitioners' Guide to Performance of Federated Learning Algorithms." *Pre-Registration Experiment Workshop, NeurIPS, December 2020* (Archived). OpenMined.

Honors & Awards

2025 Winner, Grokipedia track, XAI Hackathon, CA, USA.
2023 Graduate Research Assistantship, ECE Department, Purdue.
2020 Graduate Research Assistantship, SuperPower Group, Purdue.

2018 Winner and World Finalist, Microsoft Imagine Cup Japan National Final (Emergensor Startup).
2018 Winner, Third Business Plan Competition, University of Tokyo.
2017 Two-Year Graduate Teaching Assistantship, ECE Department, Purdue.
2017 India-Japan Engineering Program Research Scholarship, University of Tokyo.
2016 Undergraduate Teaching Assistantship, IIT Hyderabad.
2016 Special Recognition and 8th Rank for Young Team, IEEE Signal Processing Cup.
2014 Academic Excellence Award, IIT Hyderabad.
2010 National Talent Search Examination (NTSE) Scholar, Govt. of India.

Invited Talks

2023 Using neural networks and principal components to optimize multivariate statistical power trade-offs. *Modern Modeling Methods Conference* (Accepted; unable to attend).
2023 Privacy of Noisy SGD. *ML Theory, Cohere for AI.*
2022 How to promote open science under privacy. *Psychological Sciences Department, Purdue University.*
2022 PowerGraph: Using neural networks and principal components to multivariate statistical power trade-offs. *International Meeting of the Psychometric Society* (Accepted; unable to attend).
2021 Graphing multivariate statistical power manifolds with Machine Learning. *MCP Colloquium, Purdue University.*
2020 FedPerf: A Practitioners' Guide to Performance of Federated Learning Algorithms. *NeurIPS Pre-Registration Workshop.*

Teaching & Mentoring

Jan–May 2022 Mentoring Undergraduate Students for the Anvil Co-Founder AI Matching Platform Development.
Aug 2019–May 2020 Graduate Teaching Assistant, ECE 27000 — *Introduction to Digital Design*, Purdue.
Aug 2018–May 2019 Graduate Teaching Assistant, ECE 20002 — *Electrical Engineering Fundamentals II*, Purdue.

Other Services

2026 Attendee, EA Global, San Francisco, CA.
2024 Reviewer: NeurIPS, ICML TF2M Workshop, ICLR, DMLR, TMLR, ICLR Tiny Papers, Privacy Preserving AI Workshop at AAAI.
2023 Reviewer: NeurIPS, AAAI, ICML Tiny Papers, FAccT, ISIT, IJCAI, CHIL. Top Meta-Reviewer, AAAI Representation Learning Workshop.
2022 Reviewer: CHIL. Open Source: OpenMined, Gradio by HuggingFace. Professional Grant Reviewer, Purdue.