Fatim Majumder

New York, NY | fatim.majumder@emory.edu | (347) 832-6751

fatimmajumder.github.io | github.com/fatimmajumder | linkedin.com/in/fatim-majumder

Applied Mathematics / Machine Learning PhD Applicant

Scientific Computing | Optimization | Inference | Statistical Learning | AI for Science | Reliable Intelligent Systems

RESEARCH PROFILE

Applied mathematics and machine learning researcher building reproducible computational systems for model evaluation, scientific inference, and uncertainty-aware decision-making. I turn informal empirical ML questions into controlled numerical experiments with explicit data-generating assumptions, artifact provenance, calibrated uncertainty, numerical stability checks, and statistically valid comparisons.

Research direction: reliable AI for science, robust optimization, uncertainty quantification, graph-structured inference, differentiable simulation, and evaluation methods for foundation models. Guiding thesis: modern ML systems should be treated as scientific instruments - calibrated, auditable, stress-tested, and falsifiable, not merely ranked by benchmark scores.

EDUCATION

Columbia University - M.S. in Applied Mathematics

New York, NY | Incoming, Fall 2026

- Planned focus: numerical analysis, scientific computing, stochastic processes, optimization, inverse problems, probabilistic modeling, statistical inference, and mathematical machine learning.

Emory University - B.S. in Computer Science and Mathematics

Atlanta, GA | Expected May 2026

- GPA: 3.98/4.00 cumulative; 4.00 mathematics; 4.00 computer science.
- Selected coursework: real analysis I-II; abstract algebra I-II; linear algebra; multivariable calculus; differential equations; partial differential equations; numerical analysis; numerical linear algebra; iterative methods; numerical optimization; convex optimization; probability; stochastic processes; mathematical statistics I-II; statistical learning; algorithms; theory of computing; machine learning; artificial intelligence; deep learning on graphs; data mining; AI for Science reading seminar.
- Honors and activities: QuestBridge National Match Scholar; Dean's List; Undergraduate Research Distinction; HackDuke Finalist; Selective Technical Leadership Fellow; Varsity Rowing; Algory Capital Head of Research.

RESEARCH INTERESTS

- Mathematical machine learning: statistical learning theory, calibration, conformal prediction, robustness, model selection, representation learning, graph learning, distribution shift, and benchmark validity.
- Optimization and numerical computation: stochastic approximation, adaptive first-order methods, mirror descent, momentum, preconditioning, constrained optimization, inverse problems, and PDE-constrained learning.
- Scientific computing and AI for science: numerical linear algebra, Krylov methods, state-space inference, differentiable simulation, surrogate modeling, uncertainty quantification, active experimentation, and structure-aware neural systems.
- Reliable AI systems: LLM evaluation, tool-use and retrieval benchmarks, agent reliability, artifact lineage, dataset/version control, inference orchestration, auditability, observability, and failure-mode analysis.

RESEARCH AND ENGINEERING EXPERIENCE

Arthur AI - Machine Learning Research Engineering Intern

New York, NY | May 2025 - Aug. 2025

Tools: Python, PyTorch, JAX, Ray, FastAPI, PostgreSQL, Redis Streams, Docker, Kubernetes, AWS, OpenTelemetry, NumPy, pandas, scikit-learn

- Built an experiment-control layer for LLM evaluation, representing each run as a typed object over dataset snapshot, prompt graph, model artifact, tokenizer revision, sampling policy, inference backend, judge rubric, scorer, postprocessor, aggregation rule, and environment hash.
- Scaled benchmark execution to 52,000+ controlled evaluation jobs per month across reasoning, retrieval, tool use, long-context, safety, robustness, multilingual, code, and domain-specific suites while preserving artifact-level reproducibility.
- Raised exact rerun reproducibility from 71.4% to 99.98% by introducing deterministic manifests, immutable dataset hashes, container-pinned evaluator images, seeded decoding controls, schema-versioned prompts, and provenance-preserving result tables.
- Designed asynchronous inference scheduling over heterogeneous model endpoints with adaptive batching, token-budget admission control, retry semantics, circuit breakers, partial-failure isolation, and cost-aware queue priorities; reduced median experiment turnaround from 9.6 hours to 54 minutes.
- Implemented statistically defensible model-comparison tooling: paired bootstrap intervals, stratified randomization tests, multiple-comparison correction, per-slice uncertainty bands, effect-size reporting, regression severity tiers, and run-diff reports.

- Built judge reliability diagnostics for LLM-as-judge workflows using gold preference sets, rubric-level variance decomposition, inter-judge agreement, prompt sensitivity, disagreement clustering, and calibration curves; improved agreement on audited tasks from kappa = 0.43 to 0.76.

- Identified benchmark contamination from prompt-template leakage and near-duplicate examples that inflated a reasoning suite by 5.8 absolute points; added semantic deduplication checks, leakage alarms, and red-team fixtures.

Fullstory - Software Engineering Intern

Atlanta, GA | May 2024 - Aug. 2024

Tools: Python, SQL, Kafka, Airflow, dbt, PostgreSQL, analytical SQL, Docker, GitHub Actions, Grafana, OpenTelemetry

- Built validation infrastructure for analytics pipelines processing 6.4B+ weekly product events, covering schema evolution, event freshness, transformation replay, metric discontinuities, cardinality explosions, join drift, null spikes, and deploy-correlated anomalies.

- Reframed data-quality monitoring as online statistical inference over a changing data-generating process, separating seasonality, instrumentation changes, product launches, upstream lag, bot traffic, and genuine pipeline failures.

- Implemented layered detectors combining robust z-scores, seasonal baselines, Kolmogorov-Smirnov drift tests, population-stability indexes, column-level entropy checks, foreign-key integrity tests, and lineage-neighbor correlation checks.

- Improved alert precision from 61% to 97%, reduced false-positive pages by 82%, and lowered median time-to-detection from 3.7 hours to 6 minutes through adaptive thresholds and ownership-aware routing.

- Designed replay validation that re-executed historical transformations under pinned inputs and compared outputs against current logic to catch non-idempotent code paths, hidden dependencies, and accidental semantic changes.

- Built anomaly attribution tools linking metric shifts to deployment diffs, schema changes, query rewrites, upstream delays, column distributions, ingestion partitions, and ownership metadata; reduced median debugging time from 2.8 hours to 19 minutes.

Algory Capital, Emory University - Head of Research

Atlanta, GA | Sept. 2023 - Present

Tools: Python, pandas, NumPy, SciPy, scikit-learn, XGBoost, statsmodels, cvxpy, SQL, PostgreSQL, DuckDB, Streamlit, FastAPI, Docker

- Led research infrastructure and analyst development for a 30+ member quantitative research organization, establishing a research operating system for hypothesis formation, falsification, peer review, replication, and model-risk documentation.

- Architected a point-in-time research platform spanning ingestion, universe construction, feature engineering, cross-sectional factor testing, walk-forward modeling, portfolio construction, transaction-cost simulation, exposure control, and attribution.

- Increased reproducible research throughput by 8.4x by replacing notebook-only workflows with versioned experiment templates, canonical data loaders, sealed train/test windows, automated tearsheets, invariant checks, and standardized review rubrics.

- Implemented leakage-aware evaluation with purged and embargoed time splits, universe-retention accounting, survivorship-bias checks, delayed feature availability, benchmark alignment, turnover constraints, slippage assumptions, and factor-decay analysis.

- Built portfolio-optimization modules for mean-variance, risk parity, rank-weighted, volatility-targeted, and drawdown-aware portfolios with sector/industry exposure constraints, covariance shrinkage, and transaction-cost-aware rebalancing.

- Added statistical validation layers for multiple testing, White-style reality checks, deflated Sharpe diagnostics, false discovery control, unstable correlation warnings, regime-sensitivity analysis, and backtest-overfitting alarms; produced 18 reviewed research memos and 11 reproducible project repositories.

Georgia Tech / Emory University - Research Assistant

Atlanta, GA | Sept. 2022 - Jan. 2024

Tools: Python, PyTorch, scikit-learn, NumPy, pandas, SQL, Streamlit, MATLAB, medical-imaging preprocessing, calibration analysis, nested cross-validation

- Developed biomedical ML pipelines supporting 1,200+ controlled experimental runs across cohort definitions, preprocessing variants, imaging features, clinical covariates, fusion strategies, calibration procedures, subgroup analyses, and decision thresholds.

- Rebuilt cohort construction around patient-level independence, temporal separation, site-aware validation, feature availability, and label-proxy audits, eliminating patient overlap, temporal contamination, and hidden leakage in retrospective splits.

- Improved clinically meaningful held-out AUROC from 0.69 to 0.86 and expected calibration error from 0.18 to 0.045 after leakage removal, standardized preprocessing, model-family sweeps, and systematic ablations across imaging, clinical, and fused representations.

- Implemented evaluation reports with bootstrap confidence intervals, DeLong-style AUROC comparisons, AUPRC, sensitivity/specificity at policy thresholds, decision-curve analysis, subgroup calibration, missingness analysis, and error taxonomy.

- Designed ablation studies to distinguish genuine predictive signal from scanner/site artifacts, missingness shortcuts, preprocessing artifacts, distributional confounding, and label-derived proxies; built reproducibility infrastructure for exact reruns of prior model claims.

SELECTED COMPUTATIONAL RESEARCH PROJECTS

RE-AMP: Generative Audio Robustness Evaluation - Independent / Research Engineering Project

Tools: Python, PyTorch, torchaudio, FastAPI, React, TypeScript, Redis, PostgreSQL, Docker, async workers, statistical dashboards

- Built a full-stack benchmark platform for generative audio models under perturbations, distribution shifts, compression artifacts, prompt variation, acoustic transformations, stochastic decoding differences, and model-version changes.
- Supported 18,000+ controlled robustness runs across 82 perturbation operators and 9 evaluator families, with deterministic manifests over prompts, seeds, model checkpoints, audio transforms, metric configurations, and environment hashes.
- Reduced experiment setup time by 91% through declarative benchmark specifications, reusable dataset/model adapters, typed evaluator configs, cached feature extraction, and structured result storage.
- Designed signal-level, perceptual, and model-based metric panels - including FAD-style distribution distance, embedding similarity, spectral convergence, loudness drift, pitch/chroma stability, clipping artifacts, compression sensitivity, and ranking instability - with bootstrap uncertainty and failure-slice analysis.

Graph-Based Traffic Collision Risk Modeling - Independent / Applied ML Project

Tools: Python, PyTorch Geometric, NetworkX, GeoPandas, road-network topology, SQL, XGBoost, spatial statistics, calibration tooling

- Built a graph ML pipeline over 1.1M+ collision records and road-network topology, representing intersections, road segments, and neighborhoods with spatial, temporal, traffic, structural, weather, and historical-risk covariates.
- Formulated collision risk as heterogeneous spatiotemporal graph prediction and compared GCN, GraphSAGE, GATv2, temporal aggregation, gradient-boosted trees, kernel baselines, and non-graph geospatial models.
- Improved hotspot-ranking AUROC by 23.4 points and top-decile recall by 31% over tabular baselines by incorporating connectivity, neighborhood propagation, centrality, temporal exposure windows, and message passing.
- Implemented leakage-aware validation with temporal cutoffs, geographic buffer zones, held-out corridors, future-information audits, and spatial autocorrelation diagnostics; added calibrated risk maps, ablations, counterfactual edge removal, conformal-style risk sets, reliability curves, and ranking-stability checks.

Robust Kalman Filtering for Vehicle Localization - Independent / Scientific Computing Project

Tools: Python, NumPy, SciPy, MATLAB, C++, stochastic simulation, state-space models, recursive Bayesian estimation, robust statistics

- Implemented a modular state-space simulation environment with configurable dynamics, observation models, sensor dropout, delayed observations, outlier regimes, correlated process noise, measurement misspecification, and asynchronous update schedules.
- Compared standard Kalman filtering with robust variants using innovation chi-square gating, covariance inflation, adaptive noise estimation, Huberized updates, Student-t observation models, residual clipping, and outlier rejection.
- Reduced median localization error by 49% and 95th-percentile error by 44% relative to a standard Kalman filter under mixed dropout/outlier regimes while preserving near-real-time execution at 100 Hz in a C++ loop.
- Analyzed breakdown under covariance misspecification, high-leverage measurements, biased sensors, delayed observations, loss of observability, and ill-conditioned covariance updates; wrote a technical note deriving the linear-Gaussian filter and motivating robust M-estimation variants.

Quant Research Lab - Independent / Research Infrastructure Project

Tools: Python, FastAPI, pandas, NumPy, SciPy, cvxpy, SQL, Docker, GitHub Actions, Streamlit, experiment manifests, backtesting

- Built a public-facing factor research and backtesting demo covering point-in-time ingestion, signal construction, portfolio formation, evaluation, transaction-cost modeling, benchmark comparison, and attribution.
- Implemented a factor library for momentum, value, quality, volatility, liquidity, seasonality, residualization, and cross-sectional normalization, with each module tied to explicit assumptions and availability windows.
- Generated reproducible tearsheets covering rank information coefficients, factor returns, turnover, drawdown, exposure decomposition, covariance sensitivity, deflated Sharpe diagnostics, transaction-cost stress tests, and regime breakdowns.
- Designed the project as a teaching artifact: each chart links to assumptions, data windows, transformations, validation constraints, code commits, and falsification checks.

TECHNICAL NOTES, PRESENTATIONS, AND RESEARCH PRACTICE

- Technical notes: stochastic optimization; numerical linear algebra; graph learning; mirror descent and momentum; diffusion models and ML foundations; Gaussian processes with derivative matching; robust Kalman filtering; foundations of stochastic control.
- Selected presentations: diffusion models; derivative-matching Gaussian processes; robust Kalman filtering; stochastic control; SGD, momentum, and Adam; mirror descent; graph neural networks for risk modeling; benchmark design for LLM evaluation; calibration and uncertainty for biomedical ML.
- Mentorship: designed analyst onboarding modules on probability, linear algebra, ML evaluation, optimization, leakage detection, ablation studies, and reproducible research workflows.
- Research practice: writes LaTeX notes emphasizing derivation, geometric intuition, computational experiments, and failure cases; standardizes dataset snapshots, preprocessing manifests, seed control, environment capture, artifact hashes, statistical uncertainty, and falsification logs.

HONORS, AWARDS, AND LEADERSHIP

- QuestBridge National Match Scholar - nationally selective scholarship recognizing academic achievement, leadership, resilience, and intellectual promise.
- Dean's List, Emory University - sustained high academic performance across mathematics, computer science, and interdisciplinary research commitments.
- Undergraduate Research Distinction - recognition for sustained research engagement, rigorous methodology, reproducible experimentation, and scholarly contribution.
- HackDuke Finalist - finalist recognition for interdisciplinary technical prototyping, system design, applied problem-solving, and rapid execution under constraints.
- Selective Technical Leadership Fellow - selected for technical leadership, mentorship, software execution, research communication, and project development.
- Varsity Rowing - collegiate athletics; disciplined training, endurance, teamwork, high-performance execution, and resilience under sustained pressure.

TECHNICAL SKILLS

Languages: Python, C++, C, SQL, TypeScript, JavaScript, Java, Bash, MATLAB, LaTeX

Machine Learning and AI: PyTorch, JAX, PyTorch Geometric, scikit-learn, XGBoost, statsmodels, Ray, model evaluation, LLM-as-judge calibration, graph ML, multimodal ML, transformer inference, retrieval evaluation, robustness testing, ablation studies, uncertainty estimation

Applied Mathematics and Scientific Computing: numerical linear algebra, Krylov and iterative methods, preconditioning, convex and numerical optimization, stochastic approximation, state-space models, Kalman filtering, Monte Carlo simulation, Bayesian inference, statistical learning, spectral methods, graph algorithms

Optimization and Inference: gradient methods, mirror descent, momentum/adaptive methods, constrained optimization, covariance shrinkage, hierarchical comparisons, bootstrap inference, randomization tests, calibration, conformal-style uncertainty, experimental design, model selection

Data and Research Infrastructure: PostgreSQL, MySQL, DuckDB, Redis, Kafka, Airflow, dbt, ETL/ELT, feature validation, dataset versioning, artifact lineage, experiment tracking, reproducible replay, structured logging, lineage graphs

Backend and Systems: FastAPI, Docker, Kubernetes, AWS, Linux, Git, GitHub Actions, CI/CD, distributed queues, asynchronous orchestration, caching, batching, observability, profiling, reliability engineering

Evaluation and Reliability: benchmark design, metric validation, regression detection, prompt/dataset lineage, statistical comparison, confidence intervals, run-diff tooling, audit logs, failure analysis, incident triage, data-quality testing, model-risk documentation