Sizhe (John) Zhang

Mobile: +1 (314) 978-5700

Email: sizhe@wustl.edu
Website: https://novaz03.github.io
GitHub Page: https://github.com/novaz03

Research Interests

- **Statistical Machine Learning:** robust/OOD generalization, uncertainty quantification, distribution shift, causal inference.
- **Scientific ML & Geospatial AI:** spatio-temporal statistics, mobility modeling under extremes, physics-/principle-informed learning.
- **Optimization & Computation:** large-scale training/evaluation, HPC/parallel computing (Slurm, Apptainer), numerical methods for AI.
- **Reasoning & LLMs:** data quality and evaluation for theorem proving/chain-of-thought; hybrid probabilistic–neural methods.

Education

- **Washington University in St. Louis (WashU)** — St. Louis, MO
  A.M. in Statistics (GPA 3.95/4.00) — *Aug 2023 – Dec 2024*
  - *Selected coursework:* Theory of Statistics, Nonparametric Estimation, Advanced Machine Learning, Linear Models, Mathematical Foundations of Big Data.
- **University of Cambridge** — Cambridge, UK
  B.A. (Hons) in Natural Sciences — *Sep 2020 – Jun 2023*
  - *Selected coursework:* Statistics, Linear Algebra, Group Theory, Quantum Mechanics, Electromagnetism, Condensed Matter Physics, Special Relativity.

Publications & Manuscripts

- **Manuscripts submitted (details on request):** spatial statistics and scientific ML projects on (i) multimodal human mobility under extreme weather and (ii) clinical driving behavior in preclinical Alzheimer's disease.

Research Experience

- **Geospatial Hazard Research Team (Prof. Nan Lin; collaborators across WashU)** — St. Louis, MO
  Graduate Researcher — *May 2025 – Present*
  - **Geospatial AI for Disaster Response**: Constructed high-resolution, hexagon-based spatio-temporal mobility datasets; integrated POI semantics and graph embeddings to study pre-/post-event behavioral change under hurricanes.
  - **Methods**: Uncertainty-aware clustering and causal analyses for distribution shift; scalable ETL and quality control pipelines for heterogeneous mobility/POI/socioeconomic data.
  - **Impact**: Produced decision-support summaries for evacuation vs. shelter-in-place patterns and critical infrastructure access; documented reproducible pipelines for multi-terabyte data.
  - **Stack**: Python (GeoPandas, PyTorch, scikit-learn), PostGIS, QGIS; HPC (Slurm arrays, Apptainer/Singularity)
- **DRIVES Project (Clinical AI; PI: Dr. Ganesh M. Babulal)** — St. Louis, MO
  Researcher — *2025 – Present*
  - **Clinical Time-Series Modeling**: Analyzed naturalistic driving data to quantify effects of preclinical Alzheimer's disease on mobility/risk; built robust feature pipelines with rigorous validation.
- **Automated Theorem Proving & Reasoning AI** — St. Louis, MO
  Graduate Research Assistant — *Feb 2024 – Present*
  - **Whole-Proof Generation**: Designed proof-tree data structures and intermediate verification to stabilize long-chain reasoning; established metrics for data completeness/consistency.
  - **LLM Data Quality**: Developed versioned pipelines for reasoning datasets (Lean4) with anomaly detection; documented best practices to reduce leakage and spurious correlations.
  - **Stack**: Python, Lean4, Git/DVC; evaluation harnesses for chain-of-thought.
- **Aviation Impact Accelerator (Business & Policy Group)** — Cambridge, UK
  Undergraduate Researcher — *Jun 2022 – Sep 2022*
  - **Techno-Economic Modeling**: Integrated fuel/material flow and capital cost data; built regression-based operating cost models and scenario tools for decision support.
  - **Stack**: MATLAB, R, interactive reporting in Excel/Markdown.

Research-Adjacent Roles

- **Operations & Systems Specialist I (Radiology)** St. Louis, MO
  Washington University School of Medicine *Apr 2025 – Present*
  - **AI Model Development & Deployment**: Co-developed LLM components for patient-facing clinical workflows (retrieval-augmented generation, prompt schemas, safety filters); implemented PEFT/LoRA training pipelines and lightweight adapters for domain adaptation; stood up evaluation harnesses for factuality/hallucination, uncertainty proxies, and regression tests against reference corpora.
  - **Large-Scale Training & HPC**: Orchestrated distributed jobs on Slurm (arrays, GPU/CPU mix), profiled throughput and memory (mixed precision, gradient accumulation), and containerized stacks (Docker/Apptainer) with `lmod` modules for reproducible builds.
- **Teaching Assistant** St. Louis, MO
  Washington University in St. Louis *Jan 2025 – Present*
  - **Courses**: Python for Data Science, Stochastic Processes, Mathematical Foundations of Big Data; led tutorials, office hours, and assessment design.

Selected Projects

- **Mobility Graph Embeddings under Extreme Events**: Node2Vec/LLM-derived POI embeddings on EPSG:5070 hex grids; clustering trajectories over 143-hour windows; evaluation under distribution shift.
- **Automated Theorem Proving & Reasoning AI**: Leakage audits, adversarial paraphrases, influence-function diagnostics, and proof-tree data structures for long-chain reasoning datasets (Lean4); versioned pipelines and evaluation harnesses for chain-of-thought stability.
- **Reasoning Dataset Curation**: Leakage audits, adversarial paraphrases, and influence-function diagnostics for long-chain proof corpora.
- **AIGlucose**: Led the AIGlucose project: an evaluation-first toolkit for plain-language, evidence-linked explanations of glucose patterns. Integrated domain-specific retrieval, span-level evidence display, and a factuality checker.

Technical Skills

- **Languages**: Python, R, MATLAB, C++, SQL, Bash; (working) Lean4.
- **ML/Stats**: PyTorch, scikit-learn, XGBoost; Bayesian modeling, conformal prediction (UQ), causal inference, robust estimation, nonparametrics, time-series, geospatial statistics.
- **Geospatial**: GeoPandas, Shapely, Rasterio, PostGIS, QGIS; EPSG:5070 workflows; hex tiling and network construction.
- **HPC/Systems**: Slurm job arrays, Apptainer/Singularity, CUDA-capable nodes, BeeGFS/CephFS; SSH tunneling, advisory file locks; profiling/monitoring at scale.
- **Data/DevOps**: Git/GitHub, DVC, MLflow/Weights&Biases (exp. tracking), Make/Snakemake; testing & CI basics.
- **Reproducibility**: Containerized pipelines, metadata/versioning, structured experiment logs, documentation and onboarding guides.