Research Statement

My research is motivated by *resourceful modern data infrastructures*: massive and geographically distributed datasets, complex dependence structures, and increasingly powerful but opaque AI systems. Classical statistical methods, which often assume centralized storage, simple dependence, and transparent models, are not designed for these environments. I develop statistical theory, methodology, and algorithms that enable reliable inference in large-scale, decentralized, and AI-augmented systems.

My work is organized around three interconnected themes:

1. **R.1 Efficient distributed learning** for large-scale, multi-site data under communication, privacy, and robustness constraints;
2. **R.2 Calibrated testing procedures for mediation analysis** under complex dependence structures and high-dimensional mediators;
3. **R.3 genAI-enhanced inference**, which leverages black-box predictive models and synthetic data in a statistically principled way.

To date, this agenda has led to **32 manuscripts**, including **16 peer-reviewed publications** in leading statistics journals such as *JASA*, *JCGS*, and *Statistica Sinica*, with additional work under review in *Biometrika* and *Biometrics*. Across these projects, I aim to build a coherent framework where *distributed learning, causal pathway analysis, and genAI-enhanced inference* reinforce one another.

R.1 Efficient Distributed Learning

Modern collaborations in health, environment, and technology often collect data at multiple institutions or devices. Privacy regulations, bandwidth limits, and institutional policies make it difficult or impossible to centralize these data. This raises fundamental questions: how to design *communication-efficient* estimators that approach centralized performance, how to remain *robust* to heterogeneity and adversarial contamination, and how to understand the *statistical–computational trade-offs* in such systems.

A first line of my work studies *divide-and-conquer* estimators that aggregate local estimates to approximate a centralized oracle. A key insight is that naive averaging can be biased when local sample sizes, regularization, or model misspecification differ across nodes. I develop debiasing strategies that correct these discrepancies and show that, under mild regularity conditions, a single round of communication can recover the first-order efficiency of the centralized estimator while still admitting valid uncertainty quantification. This provides a principled recipe for one-shot distributed procedures that are both statistically efficient and communication-light.

A second line addresses *doubly nonsmooth objectives* such as composite quantile regression with ℓ_1 penalties, which are important for robust and high-dimensional learning but

challenging for decentralized optimization. Classical methods typically achieve only sublinear convergence. In my work on decentralized surrogate composite quantile regression, we introduce local smoothing strategies that preserve robustness while enabling gradient-based algorithms with *linear convergence* and finite-sample error guarantees under heavy tails and outliers. This yields scalable and robust sparse learning over peer-to-peer networks without a central coordinator.

A third line studies distributed learning under *heterogeneity and multi-task structure*, where each node corresponds to a related but distinct population. By explicitly modeling cross-node similarity and borrowing strength across tasks, my methods improve efficiency relative to fitting independent models at each site, maintain robustness to corrupted nodes, and provide interpretable measures of similarity and divergence.

Future work in R.1. I plan to extend these ideas in three directions: (i) *federated and privacy-aware inference*, integrating communication budgets and formal privacy constraints into optimality theory; (ii) *adaptive and asynchronous algorithms* that handle stragglers and dynamic network topologies; and (iii) *distributed causal and mediation inference*, linking R.1 with R.2 so that mediation and pathway analysis can operate on federated data held by multiple institutions.

R.2 Calibrated Testing in Mediation Analysis

Beyond association, many scientific questions seek to understand *how* an exposure affects an outcome. Mediation analysis decomposes total effects into pathways operating through intermediate variables. Modern studies often involve high-dimensional mediators (omics, imaging, digital biomarkers) with strong dependence and non-Gaussian features, where classical mediation tools can be misleading.

A core challenge is the *composite null* for each mediator M_j,

$$H_{0j} : \alpha_j \beta_j = 0,$$

which corresponds to the union of submodels ($\alpha_j = 0$ or $\beta_j = 0$). Standard asymptotic tools, which are designed for simple nulls, often yield inflated type I error or overly conservative tests in this setting. My work develops *calibrated testing procedures* that respect the composite null and scale to thousands of mediators.

One component focuses on constructing test statistics and resampling schemes that remain valid under dependence and weak signals, yielding accurate p-values for the product structure $\alpha_j \beta_j$. I design high-sensitivity joint-significance and max-type tests that control family-wise error or false discovery rate while retaining power for sparse signals. Another component introduces *copula-based* and *quantile* mediation frameworks. By separating marginal distributions from dependence via copulas, we can model heavy tails and skewness, incorporate network or spatial structure among mediators, and move beyond mean effects to *quantile-level mediation*, which captures heterogeneous pathways across the outcome distribution.

I also work on *streaming mediation inference*, where test statistics and calibration are

updated as data accrue. This enables online monitoring of mediation pathways in mobile health and digital intervention studies, supports early stopping rules for expensive data collection, and aligns naturally with the distributed settings studied in R.1.

Future work in R.2. I plan to extend mediation methods to nonlinear and dynamic systems (longitudinal and networked data), to multi-exposure and multi-omics contexts where pathways form cascades, and to federated environments where mediators and outcomes are stored across institutions. These developments will provide mechanism-based insight in genomics, environmental epidemiology, and digital health.

R.3 genAI-Enhanced Inference: Black-Box Predictors and Synthetic Data

Generative models and large language models (LLMs) have made *black-box prediction* and *synthetic data generation* routine in applied work. They excel at capturing complex patterns but are often opaque and do not directly yield valid inference. A central question is:

> When, and how, can black-box predictions or synthetic data be used to improve classical statistical inference *without compromising validity*?

My research tackles this question in two complementary components. First, I develop *diagnostics and tests* to evaluate whether synthetic data and black-box outputs preserve the structures needed for inference on target parameters. This includes testing whether synthetic samples preserve relevant distributional and dependence features, and assessing whether black-box risk scores or embeddings maintain conditional relationships required by downstream analyses. These tools act as *gatekeepers* that determine when genAI components are adequate for a given inferential goal.

Second, once adequacy is established, I design procedures that *fuse real and synthetic information* to gain efficiency. Examples include two-step estimators that use synthetic data for informative initialization or control variates and semi-supervised frameworks where black-box predictions reduce variance or enhance power for detecting weak signals. A unifying theme is to treat synthetic data and black-box outputs as *auxiliary statistics*, carefully integrated so that asymptotic normality and valid confidence intervals are preserved.

In ongoing work, I propose *statistically principled stopping rules* for genAI training and fine-tuning, drawing on sequential monitoring and information-theoretic ideas. These rules aim to prevent overfitting and instability in high-capacity generative models, quantify uncertainty in evaluation metrics derived from LLMs, and provide safeguards when genAI is embedded in sensitive scientific or policy pipelines.

Future work in R.3. I plan to develop frameworks for inference based on LLM-generated features, summaries, or pseudo-observations; study robustness under distributional shift when AI systems are updated over time; and release open-source `R` packages (building on

my existing `SIT`, `abima`, and `MarginalMaxTest`) implementing genAI-enhanced inference methods.

Vision and Fit

Across R.1–R.3, my long-term vision is to establish a unified statistical framework for **trustworthy distributed learning, causal mediation, and genAI-enhanced inference**. Concretely, over the next decade I aim to: develop optimality theory and algorithms for privacy- and communication-constrained distributed inference; build scalable mediation and pathway tools for complex multi-omics and digital-health data; and design statistically principled interfaces between classical inference and LLM-based systems, including diagnostics, calibration, and real–synthetic fusion.

These efforts align closely with the Department of Statistics and Data Science at Tsinghua University, which sits at the intersection of optimization, statistics, and artificial intelligence. By combining rigorous theory, scalable algorithms, and open-source software, I aim to contribute to Tsinghua's leadership in foundational statistics and its applications to large-scale data science and AI.