

Ke Zhang

[Email](#) | [Personal Website](#) | [Google Scholar](#) | [LinkedIn](#)

EDUCATION

University of California, Riverside — Ph.D. in Mathematics, Expected May 2028

Advisor: Prof. [Maziar Raissi](#)

Northeastern University — M.S. in Mathematics, May 2023

Southwest Minzu University — B.Econ. in Financial Engineering, June 2021

SELECTED RESEARCH

ToolGate: An Executable Acceptance Pipeline for Tool-Dependent Scientific Benchmark Construction

[arXiv preprint](#) · *First author* · 2026

- Built a pipeline for validating LLM-generated scientific benchmark questions through execution, no-tool screening, and tool-agent validation, retaining 128 unique FEniCSx items from 500 generation attempts.

Beyond Compilation: Evaluating Faithful Natural-Language-to-Lean Statement Formalization

[arXiv preprint](#) · *First author* · 2026

- Investigating human–LLM disagreement in semantic-faithfulness evaluation of natural-language-to-Lean formalizations, treating both as noisy observers and examining how rubric interpretation, repeated judgments, and evaluator selection affect item-level assessments and aggregate estimates.
- Extending a study of eight systems on 400 graduate-level mathematical statements, where compilation exceeded semantic acceptance by 3–29 percentage points and the LLM-based criterion agreed with human-majority judgments on 89.7% of an independently audited sample.

*Earlier related work: [ALA](#), *NeurIPS 2025 LLM Evaluation Workshop*.*

PHREEQC-MCQ-200: A Diagnostic Benchmark for Tool-Augmented Scientific Simulator Agents

[arXiv preprint](#), · *First author* · 2026

- Built PHREEQC-MCQ-200, a 200-question benchmark and sandboxed agent framework for evaluating LLM agents on PHREEQC, a deterministic aqueous-geochemistry simulator.
- Introduced retention diagnostics showing that tool use can improve aggregate accuracy while breaking previously correct items; compared raw-output and TOC-guided access, finding that interface effectiveness varies with model capability.

Agent-Based Code Repair for Gurobi Optimization Models

AAAI 2026 Workshop on AI for Scientific Research (AI4Research) · [Accepted](#) · *First author*

- Built an agentic Gurobi code-repair system and a 26-problem, 260-test benchmark with controlled bug generation and API-grounded retrieval. Used factorial experiments to quantify how tool combinations affect repair performance.

TEACHING

Teaching Assistant, UC Riverside, 2024–2026 — MATH 004L, 005A, 006A/B, 007A/B, 009A/B, 120

Teaching Assistant, Northeastern University, 2022–2023 — Group Theory

TECHNICAL SKILLS

Programming and Tools: Python, C/C++, PyTorch, Lean 4, Mathlib, Docker, PHREEQC, Gurobi/Gurobipy, LaTeX

Methods: AI agents, formal reasoning, AI4Math, AI4Science, benchmark and evaluation design