

EDUCATION

M.S. in AI for Sciences (Ph.D. research program, '24 – '26) University of Texas at Austin | 2024 – Expected Dec. 2026
Selected UT Austin over 4 full-funded Ph.D. offers, including UCLA.

B.S. in Geophysics (Honors) University of Science and Technology of China | 2020 – 2024
• Nominated for the **Highest honor** for USTC undergrads (one of the two ESS School nominees) Mar. 2024

AI EVALUATION EXPERIENCE

Tencent Hunyuan Internship Offer Qingyun Talent Program, AI for Science Data and Evaluation | Aug. 2026
• Offer received; not joining due to CPT policy. The work below was an independent collaboration.
• Gathered insights through AI events and direct conversations with researchers and industry participants, informing which benchmarks/data source opportunities mattered to the [Hunyuan](#) multi-modal foundation model team.
• Discussed with data vendors to identify potential data-sourcing and collaboration partners. Checked the quality of vendor data samples, delivering reproducible findings and built a reusable pipeline for STEM sample evaluations.

Lead Researcher, Benchmark Radar Jul. 2026 – Present
• Built a [living database and search engine](#) for AI benchmarks with 20,710+ score observations and daily discovery across 37 sources; grew to 300+ users, 264 GitHub stars. [\[Code\]](#) [\[Hugging Face #1 Paper of the day\]](#) [\[arXiv\]](#)
• Attracted 100k+ media views and attention from [Google](#), [NVIDIA](#), [Kimi](#), [Qwen](#), and [Jinju Capital](#). [\[Posts\]](#)

Research Intern, Voice Agent Benchmarking 19Pine.AI (Singapore) | Remote | Jul. 2025 – Aug. 2025
• Built a prototype of PineAgent's production benchmarking system for a personal agent handling high-stakes phone tasks such as customer claims, evaluating RAG accuracy, privacy safety, and task completion based on 1,000+ call sessions.
[Pine later ranked No. 1 on the \$\tau\$ -Voice leaderboard, outperforming GPT Realtime and Gemini Live.](#)

Co-Lead, ScienceIDE AItonomy | Jul. 2026 – Aug. 2026
• Co-led ScienceIDE, which turns scientific codebases into executable environments for agent evaluation, supervised task construction, rubric design, team building, fine-tuning, and reinforcement learning. [\[Paper\]](#)
• Led Earth-system and hydrology track planning, codebase and infrastructure review, and domain-expert coordination; designed scientific-correctness criteria covering numerical equivalence, physical invariants, and multi-step stability.

Lead Researcher, ESM-bench for Earth System Models UT-Austin | Dec. 2025 – Present
• Created 107-task ESM-bench to test physical validity and Fortran syntax correctness, which ordinary coding benchmarks miss; <20% exact match exposed a core weakness in frontier coding agents. [\[Preprint\]](#)

Core Contributor, ASI-Bench Apex Intelligence | Apr. 2026 – Jul. 2026
• Authored 3 Earth Science tasks and strengthened benchmark infra on task and scorer validity. [\[arXiv\]](#)

Core Contributor, MatraIx Simulating the World with 8.3 Billion Persona Agents | Jul. 2026 – Present
• Built a resumable Qwen3.6 vLLM pipeline orchestrating 1.6K GPU-hours to extract 38k unique personas with zero cross-system overlap. 12th of 93 authors for [paper](#) featured by [Forbes Türkiye](#) and [WIRED Czech](#).

Contributor, ResearchClawBench Noah-MP science task. [\[arXiv\]](#) [\[Hugging Face #2 Paper of the day\]](#)

Contributor, AI Safety Audit Infrastructure meridianlabs-ai/inspect_petri | Jun. 2026 – Present
• Landed 2 tested upstream fixes in Petri, an alignment-auditing agent whose transcripts serve as safety evidence: judge-visible auditor prefill attribution and shipped benchmark-seed canary coverage. [\[PR\]](#)

PUBLICATIONS

Selected from 13 papers, preprints, chapters, and workshop papers; see [Google Scholar](#).

- **Lead author**, [What Git Does Not Record: Provenance in AI-Assisted Research](#) (NeurIPS '26 AI-Native Academia Workshop)
- **Lead author**, [Benchmark Radar: A Living Database for AI Benchmarks and Eval](#) [\[Hugging Face #1 PaperOfDay\]](#)
- **Lead author**, [ESM-Bench: Evaluating Whether AI Agents Understand Earth System Model Physics and Code](#)
- **Lead author**, [Noah-Agent: A Multi-Expert Agent Autoresearch Framework for Large-Scale Climate Models](#)
- **Co-Lead**, [ScienceIDE: Turning World's Scientific Codebase into Agent Learnable Environments](#) [\[Hugging Face #2 PaperOfDay\]](#)
- Co-author (12th of 93), [MatraIx: Simulating the World with 8.3 Billion Persona Agents](#) [\[Hugging Face #3 Paper of the day\]](#)
- Co-author, [ResearchClawBench: End-to-End Autonomous Scientific Research](#). [\[Hugging Face #2 Paper of the day\]](#)
- Co-author, [ASI-Bench: At the Dawn of Artificial Superintelligence](#). [\[Website\]](#) [\[Code\]](#)

TECHNICAL SKILLS

- **Evaluation:** benchmark and dataset construction, rubric design, LLM-as-judge, task validators, hidden tests, failure taxonomies, red-teaming, prompt injection and adversarial testing, experiment analysis, evaluation reporting.
- **Engineering:** Python, PyTorch, NumPy, pandas, xarray, SQL, TypeScript/JavaScript, FastAPI, Node.js, Docker, GitHub Actions, AWS/GCP, Linux/HPC.

STARTUP EXPERIENCE

Technical Co-Founder, Tacite.AI Austin, TX | Jan. 2026 – Present
• Designed and built two proactive AI assistants, Tacite APP for individuals and Tacite COS for teams, on a local-first React/Electron/SQLite architecture that keeps user data on-device and supports background scheduling.
• Used test-driven development with hundreds of passing tests and shipped cross-platform macOS and Windows builds.
• Shaped product direction around privacy, responsiveness, local-first workflows, and user control. [\[Website\]](#)