ktwu@utexas.edu
ktwu01 ktwu01

Koutian Wu

(+1) 737-420-8027
Open to relocation

Education

M.A. 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; featured by UT Austin Graduate Studies.

B.S. in Space Sci and Tech (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 [Code] [Hugging Face #1 Paper of the day] [arXiv] | Jul. 2026 – Present
- Built a living database and search engine for AI benchmarks with 44,978+ score observations and daily discovery across 40 sources; 351 GitHub stars and 100k+ media views.
- Used by or drew attention from researchers and investors at Google, NVIDIA, Amazon, IBM, Kimi, Qwen, CMU, Harvard, MIT, and Jinqiu Capital. [Posts]

Research Intern, Voice Agent Benchmarking 19Pine.AI (Singapore) | Remote | Jul. 2025 – Aug. 2025
- Built a prototype of PineAgent's production evaluation pipeline to supply scored feedback for RLHF, scoring retrieval accuracy, privacy safety, and task completion across 1,000+ high-stakes phone sessions, including customer claims. *Pine later ranked No. 1 on the τ-Voice leaderboard, outperforming GPT Realtime and Gemini Live.*

Co-Lead, ScienceIDE AItonomy | Jul. 2026 – Aug. 2026
- Contributed to ScienceIDE, which turns scientific codebases into executable agent environments whose verified interaction trajectories support supervised fine-tuning (SFT), reinforcement learning (RL), and evaluation. [Paper]
- Co-designed scientific tasks and correctness criteria; led Earth-system and hydrology track planning, infrastructure review, and domain-expert coordination.

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, Simulating the World with 8.3 Billion Persona Agents MatrAIx | Jul. 2026 – Sep. 2026
- Built a resumable Qwen3.6 vLLM pipeline for synthetic persona data on H100/A100 GPUs (1.6K GPU-hours), with schema validation, grounding checks, and deduplication; extracted 38k unique personas with zero cross-system overlap. 12th of 93 authors for paper featured by Forbes and WIRED Czech.

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

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: Scientific Codebases as Agent Learnable Environments*, submitted to ICLR 2027
- Co-author (12th of 93), *MatrAIx: Simulating the World with 8.3 Billion Persona Agents*, under review at *Nature*
- 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:** RL, Post training evaluation, 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]