Zian Zeng

Computer Science · University of Maryland, College Park

zianzeng@umd.edu · zian-zeng.github.io · LinkedIn

Reliable language-model agents: action admission, recovery, and evaluation.

EDUCATION

University of Maryland, College Park Aug 2025 - May 2027 (expected)

M.S. in Computer Science · GPA: **3.90**/4.00

University of Hawai‘i at Mānoa Aug 2020 - May 2025

B.S. in Computer Science, Data Science track; Business minor · GPA: **3.88**/4.00

RESEARCH EXPERIENCE

University of Maryland, College Park 2025-present

Graduate research

First author and sole student author on five current manuscripts; led implementation, experimental design, analysis, writing, and revisions.

"ReliaSkill" and "CallAdmit"

Research with Professor Jordan Boyd-Graber · Computational Linguistics and Information Processing (CLIP) Lab

- Developed "ReliaSkill," **independently leading its final redesign** around auditable CALL, REDIRECT, and ABSTAIN decisions, grounded arguments, explicit use boundaries, and side-effect evidence.
- Evaluated the earlier system across **five models and 100 tools**, retaining **120,000 structured-call/routing records**. Controlled offline joint exact match reached **93.4% versus 57.4%** for the strongest compared baseline, and **74.6% versus 51.9%** on an 80-tool held-out API-style split.
- Refocused the study into "CallAdmit" by freezing **3,652 proposed calls** and executing component-removal experiments; showed that **1,432 schema rejections had no marginal decision effect**, distinguishing a check’s activity from its contribution.
- Designed traceable normalization that raised reduced-policy balanced accuracy from **78.9% to 88.4%** on When2Call (exploratory); the selected policy ran **about 23× faster** than the original gate in a single-core CPU measurement.

"EffectTwin"

Research with Professor Ashok Agrawala · Maryland Information and Network Dynamics (MIND) Lab

- Initiated the project independently during summer 2026; built an **exact belief-state solver and a separate certificate checker** to distinguish interface-imposed recovery limits from agent errors. Established **199 recoverable and 100 unrecoverable cases** in the authored finite model.
- Identified which interface changes repair failures: **74 of the 100 unrecoverable cases** remain beyond added fixed-latency reads of stale server state, whereas **durable deduplication repairs all 100** in the catalog.
- Crossed prompt order with schema order to isolate the output protocol: emitting actions before explanations improved **next-action viability by 11.7–14.3 percentage points across two models**. Extended validation checked **1,083 unique cases**, including **372** cross-checked by exhaustive policy enumeration.
- In the authored catalog, the two tested open models achieved at most **29.65% sure recovery** on recoverable cases, versus **99.50% for a fixed contract-aware rule**; sure recovery required success in every compatible hidden execution.

"Restoration Is Not Neutrality"

Research with Professor Furong Huang · Furong Lab

- Designed **three-arm, matched-budget experiments** separating visible-prefix restoration from an invisible generation cap; traced differences to reasoning mode, executed program selection, and program form.
- Audited **60,720 recorded synthesis attempts** and replayed **873 programs** from saved states. All **626 previously executed programs** reproduced their recorded outcome classes; matched actual remaining token allowances and checked final-program extraction.
- Restored Qwen performance with a **one-line ending cue across all 16 primary configurations**, while showing that the cued prefix remained non-equivalent to an invisible cap; established why matched token accounting alone does not make controls interchangeable.

"Correct Statuses, Wrong Targets"

Research with Professor Furong Huang · Furong Lab

- Built paired repair experiments with **432 units per model and four prompt conditions**; token-matched the status-value contrast to separate truthful feedback from checklist structure and preservation reminders in persistent AppWorld states.
- Revealed a preservation–repair tradeoff: truthful indexed feedback reduced joint repair by **21.8 and 15.5 percentage points** across two models, while Mistral's newly lost goal predicates fell from **1,632 to 17**.
- Traced misdirected edits to confusion between goal indices and record identities; **object-keyed statuses removed index-named edits** in pre-specified two-model probes. Reexecuted **183 programs from 72 trajectories**, reproducing states, outputs, writes, and goals.

"The Agent or the World?"

Research with Professor Ashok Agrawala · MIND Lab

- Built a persistent synthetic office environment and crossed carried memory with world persistence; applied end-state and provenance-guarded verifiers to **the same episodes** to distinguish agent work from useful state left by earlier tasks.
- Across **318 runs and nine open-weight models**, measured **6.9–19.6 percentage points** of inflated success credit on exposed tasks and **zero inflation in six structural-control cells**; decomposed the effect into prior-state exposure and unchanged-state reuse.
- Showed that end-state scoring turned **four models' guarded declines into apparent gains**. Audited **16 benchmark settings**, identifying **nine** whose verifiers could credit residual state, with reset protocols protecting seven of those settings.
- Developed trace classification distinguishing legitimate recognition of an already-satisfied goal from failed searches and unsupported completion claims; interpreted provenance checks according to whether the task required new work.

"RoleFlow"

Collaborative research with Professor Furong Huang · Furong Lab

- Coauthored research on **self-evolving roles and structural adaptation in multi-agent collaboration**. The method exceeded the strongest compared multi-agent baseline by **4.7 points on average across seven benchmarks**.

EARLIER RESEARCH EXPERIENCE

University of Southern California · GLAMOR Lab

June-September 2024

SURE student researcher · Lab PI: Professor Jesse Thomason

- Contributed data-augmentation, decoding, and embedding experiments to neuro-symbolic American Sign Language (ASL) modeling, investigating how linguistic structure can support sign recognition and comprehension with limited labeled video data.
- Coauthored the **NAACL 2025 Findings** ASL Knowledge Graph paper, combining **11 linguistic sources** and evaluating **three video tasks**; the collaborative models achieved **91% isolated-sign recognition accuracy**.

University of Missouri · VIMAN Lab

May 2023-October 2024

REU student researcher; continued collaboration · Professor Prasad Calyam

Graduate mentor: Roland Oruche

- Contributed implementation and experiments to **three papers** on multi-hop question answering and conversational advising, including "Chain-of-Query," multi-hop dependency graphs, and the "Vidura" advisor framework.
- Implemented **four prompting/retrieval comparisons**—direct prompting, zero-shot chain-of-thought, retrieval, and Chain-of-Query—with normalized exact-match and token-F1 evaluation accounting for answer aliases. Worked with pipelines that preserve intermediate evidence, resume unfinished cases, and checkpoint outputs.
- Developed Vidura/KnowCOVID-19 retrieval, summarization, and question-answering workflows for scientific literature; contributed **Flask, Java/Maven and Spring Boot, and Angular** services to the publication-analytics platform.

University of Hawai‘i at Mānoa · Hawaii Digital Health Lab

2023-2025

Student researcher · Lab PI: Professor Peter Washington

- Prepared **first-author research** comparing CNN and Vision Transformer models for autism spectrum disorder (ASD) classification on **547 eye-tracking images from 59 participants**, focusing on whether visual explanations meaningfully represent model decisions.
- Compared **saliency maps, Grad-CAM++, LIME, and attention maps**, examining sparse gaze traces and largely black backgrounds; also contributed to human-in-the-loop crowdsourcing and multimodal autism/ADHD research.

PUBLICATIONS

"Disentangling Complex Questions in LLMs via Multi-Hop Dependency Graphs"

Roland Oruche, Alphaeus Dmonte, Vani Seth, **Zian Zeng**, Yuanxun Zhang, Marcos Zampieri, Prasad Calyam (2025).

Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025). pp. 2243-2252.

"The American Sign Language Knowledge Graph: Infusing ASL Models with Linguistic Knowledge"

Lee Kezar, Nidhi Munikote, **Zian Zeng**, Zed Sehyr, Naomi Caselli, Jesse Thomason (2025).

Findings of the Association for Computational Linguistics: NAACL 2025. pp. 7032-7044.

"Chatbot Dialog Design for Improved Human Performance in Domain Knowledge Discovery"

Roland Oruche, Xiyao Cheng, **Zian Zeng**, Audrey Vazzana, Md Ashraful Goni, Bruce Wang Shibo, Sai Keerthana Goruganthu, Kerk F. Kee, Prasad Calyam (2025).

IEEE Transactions on Human-Machine Systems. 55(2), pp. 207-222.

"Chain-of-Query Prompting for Efficient Small Language Models in Multi-Hop Open-Domain Question Answering"

Rithika Akula, Roland Oruche, **Zian Zeng**, Yuanxun Zhang, Marcos Zampieri, Prasad Calyam (2024).

IEEE International Conference on Big Data 2024, workshop track. Accepted workshop paper.

"CallAdmit: Diagnosing and Efficiently Repairing Over-Abstention in Tool-Call Admission with Fixed-Proposal Certificates"

Zian Zeng, Jordan Boyd-Graber (2026).

20th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2027). Under review.

"Restoration Is Not Neutrality: Auditing Visible Budget Controls in Agent Recovery"

Zian Zeng, Furong Huang (2026).

2027 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2027). Under review through ACL Rolling Review.

"Correct Statuses, Wrong Targets: Truthful Goal Feedback Can Misdirect Stateful Repair Agents"

Zian Zeng, Furong Huang (2026).

32nd International Conference on Computational Linguistics (COLING 2027). Under review through ACL Rolling Review.

"EffectTwin: Separating Recovery Limits from Output-Protocol Effects in Tool-Using Language Models"

Zian Zeng, Ashok Agrawala (2026).

2027 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2027). Under review through ACL Rolling Review.

"The Agent or the World? Verifier Inflation in Persistent-World Agent Evaluation"

Zian Zeng, Ashok Agrawala (2026).

2027 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2027). Under review through ACL Rolling Review.

"RoleFlow: Self-Evolving Multi-Role Agentic Collaboration via Structural Adaptation"

Yi Ge, Pan Wang, Xuehao Cui, **Zian Zeng**, Xu Hu, Heng Huang, Furong Huang (2026).

2027 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL 2027). Under review through ACL Rolling Review.

PROFESSIONAL EXPERIENCE

Honolulu Fire Department

January-May 2023

Software engineering intern

- Designed an incident/performance dashboard for fire chiefs, the Fire Prevention Office, and the Public Information Office. Integrated **three systems: computer-aided dispatch (CAD), records management (RMS), and geographic information (GIS)** through C# MVC APIs, SQL Server, and Angular/ASP.NET Core, with interactive charts and maps.
- Built a scheduled **C#/SMTP major-incident alerting tool** with structured incident details and automatically generated reports for Public Information Office staff, including email access on mobile devices at an incident scene; automated execution with Windows Task Scheduler.
- Developed **Python/Outlook inquiry analytics** to count media inquiries, measure response times, and retrieve messages by keyword, making the communications workload inspectable alongside operational incident data.

Blue Planet Energy

December 2018-July 2019

IT assistant

- Contributed to monitoring-dashboard and control-interface prototypes for **high-voltage solar-battery systems**, including external display panels and legacy web-data extraction using **Python, Raspberry Pi, and PLCs**.
- Maintained cloud-based inventory and sales records; continued with the company in a part-time role following the internship.

TEACHING AND MENTORING

Windward Community College

Supplemental Instruction Leader — *December 2020-December 2022*

- Led weekly Java/introductory computer science study sessions, group learning activities, and office hours for **40+ students across six semesters**, helping students practice course concepts collaboratively.

Math lab tutor — *January-December 2022*

- Provided individualized mathematics tutoring through Calculus II for **20+ students across three semesters**; independently operated the summer math lab for **2.5 months**.

SELECTED PROJECTS AND LEADERSHIP

"ORBIT"

2026

Founder and Team Leader · Faculty guidance: Professor Ashok Agrawala (MIND Lab)

- **Led product vision, architecture, implementation, and team coordination** for a student-support prototype integrating academic planning, workload, daily logistics, and campus resources. Guided delegation and integration through a UMD xFoundry competition entry and bilingual public showcase.
- Built a **Flutter/Dart system with five specialist agents**, shared student context, and structured skill outputs. Kept workload calculations and workflow control deterministic while using Gemini or local Gemma/Ollama for synthesis; connected Canvas/Google through a **Node.js proxy with caching and request deduplication**.
- Diagnosed nonempty but unfinished Gemini replies and implemented **incomplete-response detection with local-model fallback**. Preserved provider/fallback reasons in response traces and added regression tests so returned text alone would not count as successful recovery.
- Designed **student-controlled data access and feedback**: explicit consent for live context, quiet hours, adjustable nudges, source-backed resource cards, and feedback linked to agent traces. Logged activated roles, data sources, model path, and latency to make recommendations inspectable.
- Evaluated the prototype with **40 synthetic student profiles** and built a readiness dashboard linking fixture coverage, feedback, and audit records. Designed a **proposed 30-day pilot** tracking planning activity, resource use, and self-reported workload manageability.
- **Led the 75-page business plan**, guiding teammates' writing and chart creation and connecting technical milestones to pilot scope, institutional adoption, operating costs, and a funding roadmap. Proposed staged university-funded deployment to support accessible student use.

"Hybrid Edge-Cloud Agentic Reasoning" February 2026

Google DeepMind and Cactus Compute Global Hackathon

- Designed a hybrid tool-calling strategy combining **deterministic extraction, on-device FunctionGemma, and Gemini cloud fallback**, balancing correctness, latency, and local execution. **Ranked #2 on the Washington, DC leaderboard at competition close** in an event spanning five host cities.
- The team system combined argument normalization and schema validation with intent-coverage checks before accepting local calls; unresolved requests went to cloud inference with execution-source and latency information retained.

"Quiz App" January-May 2024

Technical Lead and Requirements Lead · Undergraduate capstone

University of Hawai'i at Mānoa · Sponsor: Professor Edoardo Biagioni

- Led technical development and requirements for a **three-person capstone team**, building the Flutter/Dart student app for a classroom quiz system with a companion instructor desktop app.
- Implemented photo-based attendance capture, persistent student profiles, and real-time answer submission; refactored a **singleton WebSocket connection** to share network state across mobile screens and support multiple student clients.
- Worked through multiple Bluetooth libraries and Android 12+ permission issues, then shifted BLE/TCP prototypes to **LAN-based WebSockets**. Integrated location capture and improved connection-error handling to support the classroom workflow across devices.

"Saber Quest: Echoes of the Fallen" November-December 2023

Project Lead and Co-developer · University of Hawai'i at Mānoa

- **Originated the game concept and led project framing, resource selection, and task allocation** for the two-person team, coordinating the work from initial design through gameplay integration.
- Co-developed a **two-person VR action game** in **Unreal Engine 4.26 and Blueprints**; implemented lightsaber glow, activation animation and audio, controller interactions, level integration, and the main menu.
- Built player health, damage feedback, and death-screen systems; resolved build failures, saber/player alignment issues, and animation/audio bugs while integrating gameplay across multiple maps.

"App Review Sentiment Analysis" May 2023

- Compared **Gaussian Naive Bayes, multinomial Naive Bayes, and BERT** for sentiment classification of Google Play reviews of the Medium app, examining classical probabilistic approaches alongside transformer-based modeling.

"Study Partner" December 2022

Project Lead and Co-developer · University of Hawai'i at Mānoa

- **Initiated the idea and led project framing, resource discovery, and task allocation** for the five-person team, organizing development around students' needs for finding peers and coordinating study sessions.
- Co-developed a **Meteor/React and MongoDB** application with a **five-person team** to help computer science students organize in-person study groups, with course-based peer-support profiles, session scheduling, a calendar, and moderator reporting.
- The team documented feedback from **five community members** and **12 end-to-end test scenarios**; the project used GitHub Actions with ESLint/TestCafe and was deployed on DigitalOcean.

HONORS AND CERTIFICATIONS

Dean's List, six semesters (Fall 2021–Spring 2024), University of Hawai'i at Mānoa; **AI4ALL** Discover AI, Apply AI, and Change Maker Certificates of Competence (2022, 2023); **CRLA Certified Level 1 Tutor** (2022).

TECHNICAL SKILLS

Programming: Python, Java, C/C++, C#, TypeScript/JavaScript, SQL, Dart, HTML/CSS.

Machine learning and adaptation: PyTorch, PyTorch Lightning, Hugging Face Transformers, PEFT/LoRA, scikit-learn; bf16 training, gradient checkpointing, custom action-learning objectives.

LLM inference: vLLM, Ollama, bitsandbytes; 4-/8-bit quantization, tensor-parallel GPU inference, Gemini and OpenAI-compatible model APIs.

AI agents: Tool/function calling, JSON Schema validation, constrained decoding, multi-agent orchestration, tool/skill routing, memory and context management, hybrid local/cloud inference, failure-aware fallbacks.

Retrieval and grounding: Retrieval-augmented generation (RAG), multi-query evidence retrieval, query decomposition, structured context integration, source-grounded answer generation.

Vision and interpretability: CNNs, Vision Transformers, saliency and attention maps, Grad-CAM++, LIME.

Evaluation and verification: Agent benchmark design, paired counterfactual experiments, matched-budget and matched-seed ablations, deterministic replay, execution-trace analysis, exact policy enumeration, independent certificate checking.

Data and statistics: NumPy, pandas, SciPy, Matplotlib, seaborn; paired and cluster bootstrap, permutation/sign tests, sensitivity analysis.

Research infrastructure: Slurm, Hydra, Linux, Git; reproducible GPU experiments, checkpointing and resumable execution, data/model provenance, runtime validation.

Web and backend: Node.js, Angular, React, ASP.NET Core, Flask, Spring Boot, Meteor, Pydantic; REST APIs, WebSockets, SQLite, SQL Server, MongoDB; caching and request deduplication.

Mobile and real-time systems: Flutter, WebSockets, LAN client-server networking, Bluetooth Low Energy prototyping, Android permissions, camera/location APIs, SharedPreferences; Raspberry Pi and PLCs.

Game development and VR: Unreal Engine, Blueprints, UMG/Widget UI, Unity; VR controller interaction, gameplay state, materials, animation and audio integration.

Software testing and delivery: pytest, TestCafe, ESLint, GitHub Actions; automated end-to-end testing, continuous integration, DigitalOcean deployment.

Languages: English, Mandarin, Cantonese.