Shou-Tzu Han (Debra Han)

GitHub | Personal Website | Google Scholar | Email | Phone

Research Profile

Computer Science Ph.D. student and Graduate Research Assistant at Wayne State University specializing in **trustworthy AI, LLM agent safety, LLM robustness,** and **mechanistic interpretability**. Experienced in designing reproducible GPU/HPC experiments, analyzing internal model behavior, and building evaluation pipelines for language and agent systems. Seeking a research internship focused on reliable and interpretable AI systems.

Research Interests

- **LLM Agent Safety:** Locating and attributing failures in LLM agent reasoning during multi-step, safety-critical tasks, and compiling verified failures into runtime checks.
- **LLM Robustness and Reasoning:** Evaluating reasoning stability under meaning-preserving perturbations and diagnosing performance–confidence mismatches.
- **Mechanistic Interpretability:** Using activation analysis, patching, and ablation to identify internal components associated with model failures.
- **Agentic Retrieval and Scientific Discovery:** Building multi-step retrieval systems for literature comparison, contribution analysis, and research-gap identification.

Publications

- **Shou-Tzu Han**, Rodrigue Rizk, and KC Santosh. *Fragile Reasoning: A Mechanistic Analysis of LLM Sensitivity to Meaning-Preserving Perturbations*. Preprint. arXiv:2604.01639
- **Shou-Tzu Han**. *Novelty-Aware Agentic Retrieval: Comparing Research Contributions Through Structured Multi-Step Reasoning*. Preprint. arXiv:2606.22151
- **Shou-Tzu Han**. *Narrative-Centered Emotional Reflection: An Early Prototype for AI-Supported Emotional Self-Reflection*. Preprint. arXiv:2504.20342

Research Experiences

Graduate Research Assistant
Wayne State University, Trustworthy AI Lab — August 2026–Present
Advisors: Dongxiao Zhu and Sooin Kim

LLM Agent Safety Research

- Research where and why LLM agent reasoning fails during multi-step, safety-critical tasks, spanning action selection, tool use, and decision-making under injected or adversarial context.
- Develop an outcome-grounded auditing method that commits checkable predictions before an action, then uses the real execution result to localize the failing reasoning step and compile it into a runtime check.
- Design fault-injection and component-removal experiments on agent benchmarks to measure whether the method attributes failures to the correct cause.

AI-Supported Mentoring Systems

- Develop AI-supported mentoring systems aimed at improving retention and educational experiences for veteran engineering students.
- Build LLM-based mentoring components that deliver structured, context-aware guidance and relevant academic resources.
- Contribute to system prototyping, experimental design, data analysis, and evaluation of AI-generated support.

Graduate Research Assistant
University of South Dakota, AI Research Lab — January 2026–May 2026
Advisors: KC Santosh and Rodrigue Rizk

LLM Robustness and Mechanistic Interpretability

- Conducted GPU/HPC experiments using PyTorch and Hugging Face Transformers to evaluate LLM reasoning stability under meaning-preserving perturbations.
- Applied activation analysis, patching, and ablation to diagnose reasoning failures; this work resulted in the *Fragile Reasoning* preprint (arXiv:2604.01639).

Research Projects

Fragile Reasoning: A Mechanistic Analysis of LLM Sensitivity to Meaning-Preserving Perturbations

- Evaluated Mistral-7B, Llama-3-8B, and Qwen2.5-7B on 677 paired GSM8K problems, revealing answer-flip rates of 28.8%–45.1% under meaning-preserving perturbations.
- Developed the **Mechanistic Perturbation Diagnostics (MPD)** framework, integrating logit-lens analysis, activation patching, component ablation, and the novel **Cascading Amplification Index (CAI)**.
- Identified localized, distributed, and entangled failure modes, demonstrating substantial architectural differences in failure localization and recoverability.
- Evaluated steering vectors and layer fine-tuning as targeted repair methods using **Python, PyTorch, Hugging Face Transformers, GPUs,** and **SLURM**.

Novelty-Aware Agentic Retrieval: Comparing Research Contributions Through Structured Multi-Step Reasoning

- Built an agentic retrieval system over a 100-paper corpus using six components: query analysis, iterative retrieval, ranking, contribution extraction, comparison, and answer generation.
- Designed structured contribution records and a three-pass comparison agent to identify paper-level overlaps, methodological differences, and research gaps.
- Developed a **problem–method gap matrix** that generates citation-grounded evidence for unexplored combinations within a research corpus.
- Achieved a mean Precision@5 of 0.980, nDCG@5 of 0.739, and 84.0% schema compliance across ten evaluation queries.
- Implemented the system using **Python, GPT-4o, FAISS/Chroma, SentenceTransformers,** and structured prompting.

Technical Skills

- **Machine Learning & LLMs:** PyTorch, TensorFlow, Hugging Face Transformers, OpenAI API, prompt design, model evaluation, reasoning analysis
- **Interpretability & Reliability:** Logit lens, activation patching, component ablation, steering vectors, perturbation analysis, failure diagnosis
- **Agentic AI & Retrieval:** RAG, FAISS, Chroma, BM25, SentenceTransformers, ReAct-style agents, structured prompting
- **Programming & Data:** Python, Java, JavaScript, TypeScript, SQL, Bash, pandas, NumPy, Matplotlib
- **Computing & Development:** SLURM, HPC, CUDA/GPU computing, Linux, Git, Docker, FastAPI, Flask, React, Vercel

Education

Wayne State University, Detroit, Michigan — August 2026–Present
Ph.D. in Computer Science
Graduate Research Assistant, Trustworthy AI Lab

Boston University, Boston, Massachusetts — September 2021–August 2023
Master of Science in Computer Science

Soochow University, Taipei, Taiwan — September 2015–June 2020
Bachelor of Arts in Political Science