Jincheng He

jhe516@ucsc.edu | linkedin.com/in/jincheng-he-duke/ | JinchengHeRyan.github.io

Education

University of California, Santa Cruz (Silicon Valley Campus) — **Santa Clara, CA**
Ph.D. Computer Science Engineering, Advised by Prof. Ian Lane — *Sep 2024 – May 2029 (expected)*

- **Research Interests:** Multimodal Conversational Interaction, Full-Duplex SLMs, LLM Hallucination, Speech Processing.
- **Current Research:** Developing Full-Duplex Speech Language Models (SLMs) for real-time interaction. Investigating architectures for handling turn-taking, barge-in interruption, and latency-aware speech generation.

Duke University — **Durham, NC**
M.Eng. Electrical and Computer Engineering — *Aug 2022 – May 2024*

- **Coursework**: Computer Systems, Systems Programming, NLP, Robust Server Software, Quantum Computing.
- **Honors:** Full Scholarship.

Duke University & Duke Kunshan University — **Durham, NC & Kunshan, China**
B.S. Data Science, received from Duke University — *Aug 2018 – May 2022*

- **Coursework**: Computer Architecture, Computer Vision, Complex Analysis, Topological Data Analysis, Signature Capstone.
- **Honors:** Dean's List with Distinction (Spring 2020); Outstanding Peer Tutor.

Working Experience

Samsung Research America — **Mountain View, CA**
Research Intern — *Jun 2026 – Sep 2026*

- **Proposed LI-RAG** (under review at ICLR 2027) to answer complex multi-hop questions accurately at low inference cost. Strong multi-hop systems call an LLM at every retrieval step, which is slow; LI-RAG instead fine-tunes an embedding model (NV-Embed-v2 with LoRA) to imitate how an LLM breaks a question into sub-questions, so it retrieves evidence hop by hop with no LLM calls.
- **Outperformed all retrievers of comparable cost on 4 multi-hop QA benchmarks**, beating the graph-based HippoRAG 2 by up to **8.6 accuracy points**, and matched the LLM-driven IRCoT on most benchmarks while retrieving **~9× faster**, using only 4K training questions and no human-labeled evidence.
- **Built a trainable graph retriever** that re-weights the knowledge graph per query, rather than searching one fixed graph for all queries, as HippoRAG 2 does; implemented differentiable Personalized PageRank to train the edge scorer end to end.

Five9 — **San Ramon, CA**
Natural Language Processing Intern — *Jun 2025 – Aug 2025*

- **Developed two multi-agent architectures**: a **finite state machine flow** and a **stack-based coordination system**, with **hierarchical state management**, **sub-agent invocation**, and **dynamic tool generation**, enabling seamless orchestration and context consistency across multi-turn dialogues.
- **Developed a ReAct (Reasoning + Acting) framework** for both **single- and multi-agent systems**, integrating structured **planning, reasoning, and action cycles**, dynamic tool selection, and **multi-step problem decomposition** with iterative reasoning loops.
- **Integrated RAG-based knowledge retrieval and API simulation**, including context-aware search, dynamic function creation, and knowledge group management, grounding agent reasoning in external data for reliable performance.
- **Built an automated customer simulator** with **dynamic persona generation** and **adaptive system prompts**, producing realistic user behaviors and responses for robust testing of agent performance.
- **Developed a modular evaluation platform** supporting both **real-time and post-conversation analysis**, with **extensible metrics** (response quality, latency, tool usage correctness, etc.) and automated report generation for comprehensive system assessment.

Xiaomi Technology Corporation — **Beijing, China**
Machine Learning Engineer Intern, AI Lab — *May 2022 – Aug 2022*

- Initiated a groundbreaking project, utilizing multi-channel microphone arrays to successfully achieve **multi-directional speech extraction** in in-car call scenarios.
- Developed an automated in-vehicle scene data simulation framework, resolved data scarcity issue. This provided one-click data generation, incorporating varied acoustic parameters and creating a comprehensive multi-channel, multi-speaker, multi-scene dataset.
- Determined in-vehicle microphone array schemes. Using neural network-based beamforming estimation to construct a multi-directional speaker extraction system that effectively eliminated noise and interfering voices, as validated by subjective testing.

Duke University — **Durham, NC**
Research Assistant, Advised by Prof. Jeffrey Krolik — *May 2023 – Aug 2023*

- Advanced speech intelligibility for the hearing-impaired through proposing and implementing a customized speech enhancement model specifically for hearing-aids. This innovative solution addressed a long-standing limitation of conventional hearing-aids.

Research Experience & Selected Projects

DuplexTutorBench: A Benchmark for Full-Duplex Spoken AI Tutors **Jul 2026**
Accepted at IEEE SLT 2026, First Author

- **Built the first benchmark for full-duplex spoken AI tutors**, in which any real-time speech model (GPT Realtime, Gemini Live, PersonaPlex, MoshiVis) tutors a simulated student on MathVista math problems.
- **Designed a two-track evaluation suite** that scores speech quality with Audiobox Aesthetics, compares turn-taking timing against human conversation (Map Task corpus) using word timings from Whisper and forced alignment, and rates teaching with an LLM judge on 8 pedagogical dimensions. **Benchmarking showed that no current model does both**: real-time APIs teach well but take turns unnaturally, while full-duplex models time their turns much more like humans but teach poorly.

TEXT2SPACE: Enhancing LLM Reasoning Capabilities via Structured Output Learning **Dec 2025**
Accepted at COLM 2026

- **Co-developed "TEXT2SPACE"**, a novel spatial reasoning benchmark that utilizes ASCII visualizations to decouple representation failures from reasoning errors, identifying a "Read-Write Asymmetry" where LLMs interpret spatial layouts effectively but struggle to reconstruct them.
- **Fine-tuned Qwen3-30B models** using LoRA and Hugging Face TRL on NVIDIA A40 GPUs, implementing "Learning to Draw" (text-to-ASCII) objectives to enforce the internalization of verifiable spatial "cognitive maps" during training.
- **Demonstrated reasoning transfer** by validating that auxiliary training on structured spatial construction generalizes to improve downstream text-only reasoning accuracy by ~9% and enhances performance on external benchmarks like STEPGAME.

Multilingual Hallucination Detection in LLM Outputs (SemEval-2025 Task 3: Mu-SHROOM) **Jan 2025**
SemEval 2025, The 19th International Workshop on Semantic Evaluation ***(Best Paper Award)***

- **Co-developed a system that pinpoints hallucinated spans in LLM answers across 14 languages**: it retrieves evidence through web search, uses an LLM to flag unsupported facts, and maps each fact back to its exact character span in the answer.
- **Ranked #1 on average among 40+ teams** and top-2 in 11 of 14 languages, after tuning prompts automatically with MiPROv2 and combining multiple systems.

Deep Learning-Based Target Speech Separation and Speech Recognition **Jun 2019 – May 2022**
Research Assistant, Advised by Prof. Ming Li *Speech and Multimodal Intelligent Information Processing (SMIIP) Lab*

- **Developed an original Automatic Speech Recognition (ASR) system** using Python and Numpy, integrating **MFCC extraction, dynamic time warping, HMM, and GMM** for precise speech template matching.
- Collaborated with Xiaomi researchers to extract and integrate target pitch into the separation model, pioneering **two joint training strategies** that improved the Target Speaker Separation (TSS) model.
- **Achieved a 5.85 dB Signal-to-Distortion Ratio (SDR) improvement** (21% relative increase over baseline) on the LibriSpeech dataset by designing an LSTM-based pitch extraction model integrated with MultiBlock RNNoise (MBRNN) via joint training.
- Secured competitive summer research funding in 2021, and contributed to a published paper at **Odyssey 2022**.

IEEE ICASSP Signal Processing Cup 2022 - Synthetic Speech Attribution **Mar 2022**

- Proposed an end-to-end multi-system fusion solution with ResNet backbone and ranked #6 among 25+ teams.
- Performed Gaussian disturbance with MATLAB and train/dev mismatching strategy to augment data for robustness.

Publications

- Shiyuan Huang, Li Liu, **Jincheng He**, and Leilani H. Gilpin, "**Learning to Draw ASCII Improves Spatial Reasoning in Language Models**," in *Third Conference on Language Modeling*, 2026
- Sicong Huang, **He, Jincheng**, Shiyuan Huang, Karthik Raja Anandan, Arkajyoti Chakraborty, and Ian Lane, "**UCSC at SemEval-2025 Task 3: Context, Models and Prompt Optimization for Automated Hallucination Detection in LLM Output**," in *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri, Eds., Vienna, Austria, July 2025, pp. 1981–1992, Association for Computational Linguistics **(Best Paper Award)**
- **Jincheng He**, Yuanyuan Bao, Na Xu, Hongfeng Li, Shicong Li, Linzhang Wang, Fei Xiang, and Ming Li, "**Single-Channel Target Speaker Separation Using Joint Training with Target Speaker's Pitch Information**," in *Odyssey 2022*, 2022, pp. 301–305
- Yuanyuan Bao, Yanze Xu, Na Xu, Wenjing Yang, Hongfeng Li, Shicong Li, Yongtao Jia, Fei Xiang, **Jincheng He**, and Ming Li, "**Lightweight Dual-channel Target Speaker Separation for Mobile Voice Communication**," in *Proc. of NCMMSC 2021*, 2021

Skills

Computer: Python, Verilog, FPGA, HDL, Java, Agile/DevOps, SQL, Docker, MIPS, Assembly, Matlab, Flutter, C++, Linux
Machine Learning Frameworks: PyTorch, Keras, Tensorflow, LangChain, SpeechBrain, Hugging Face