Shizhe He

+1 650 250 9154 | shizhehe@stanford.edu | linkedin.com/in/shizhehe | github.com/shizhehe | shizhehe.com

Education

Stanford University Stanford, USA

B.Sc. and M.Sc. in Computer Science (AI Systems), Minor in International Relations; ***GPA: 3.95/4.0*** *Sep 2022 – Jun 2026*

- *Societies*: European Student Association, Club Swim, Sigma Phi Epsilon *VP of Communications*, BASES *VP of Core*
- *Relevant Coursework*: Self-Improving AI Agents, Systems for Machine Learning, Mining Massive Datasets, Deep Learning for Computer Vision, Digital System Design, Principled Entrepreneurial Decisions

Communities I am part of Keel 1.0, Pear Garage 2023/24, NEA Fellows 2025

Work Experience

Beacon Text **San Francisco, USA**

Co-Founder, website *Mar 2024 – Oct 2024*

- Led development of RCS messaging advertising SAAS product, deploying AWS Cloudfront, S3, and Elastic Beanstalk
- Scaled company to **three** corporate design retail partners with an addressable audience of **300.000** contacts

QuantCo **San Francisco, USA**

Data Engineering - Machine Learning Intern *Jun 2024 – Aug 2024*

- Designed custom text embedding models and fine-tuning pipeline for fundraiser descriptions on one of the **largest crowdfunding-platforms**. Fine-tuned **transformer-based embeddings** to encode fundraiser content and quality.
- Built large parts of training & analyses pipeline for new workstream to suggest optimal fundraiser goal during creation. Improved production model using text embeddings by **150%** [Impact numbers protected under non-disclosure agreement]
- Applied feature selection methods on gradient-boosted trees to reduce complexity and training time by **80%** with no significant reduction in model performance across different workstreams

Theros LLC **Remote**

Software Engineering Intern *Jun 2023 – Sep 2023*

- Co-developed **Flask API** for "Jot it Down", a ChatGPT plugin for cross-chat and multi-user memory, deployed on AWS
- Implemented **RAG** system to digest live internal wiki content, github activities and deployment errors.
- Built daily digest and automated emails personalized to each employee based on RAG output.

Check24 Factory GmbH **Munich, Germany**

Data Science Working Student *Jun 2022 – Aug 2022*

- Developed end-to-end machine learning solution for product **popularity ranking**, enhancing sales for energy business unit
- Improved ranking and response times by developing fine-tuned machine learning models and real-time **fastAPI**
- Streamlined business unit operations by deploying ranking service on **AWS** and conducting **A/B testing**

Research/Project Experience

Hazy Research, Computer Science, Stanford University **Stanford, USA**

Student Researcher; Advised by Dan Biderman *April 2025 – present*

- Modeled compressor–predictor pipelines as info-bottlenecks to quantify communication effciency via mutual information.
- Deployed and optimized inference engines for open-source language models (LM) on Modal.
- Submitted under review at ICLR 2026, "An Information Theoretic Perspective on Agentic System Design" establishing design principles for cost-efficient multi-LM systems.

Strom Inc., Hacking 4 Defense (H4D), Stanford University **Stanford, USA**

Critical Minerals Problem; Sponsored by In-Q-Tel *April 2025 – June 2025*

- Developed supply-side solutions to aid US-allied supply of critical minerals essential for semiconductor production, batteries
- Followed Lean Launchpad Methodology, made many briefs and recommendations to govt. officials and industry experts.

Brains in Silicon, Electrical Engineering, Stanford University **Stanford, USA**

Student Researcher; Advised by Saarthak Sarup, Kwabena Boahen *Oct 2024 – April 2025*

- Developed architecture and algorithms for vector similarity search on **neuromorphic chips**

Translational AI Lab, Computer Science, Stanford University **Stanford, USA**

Student Researcher; Advised by Magda Paschali, Ehsan Adeli *Apr 2023 – Sep 2024*

- Investigated contrastive **self-supervised** learning with SO3-equivariance pretext task on 3D brain MRIs using **PyTorch**
- Designed model to understand underlying brain structures, involved in brain foundational model, MLCN 2024 publication

Lab for AI in Medicine, Computer Science, Technical University of Munich **Munich, Germany**

Student Researcher; Advised by Kerstin Hammernik, Daniel Rueckert *Apr 2021 – Jun 2022*

- Investigated **data distribution shift** in MRI reconstruction. Adapted methods to novel dynamic **7T MRI** reconstruction
- Youngest presenter at the 2022 ISMRM-ESMRMB Joint Annual Meeting/Conference in London, abstract, publication

Skills

Languages: German (Fluent), Mandarin Chinese (Fluent), English (Fluent), French (Barely Conversational)

Programming: Python, C, C++, React, Assembly, SQL, Torch, SGLang, AWS, Modal, Git, Verilog, Digital Circuit Design

Interests: How we make incremental improvements feel magical; Swimming, Tennis, Photography, Formula 1 Racing