Yucheng (Leonora) Zhu

+1 (267) 881 5179 yuchengz@seas.upenn.edu Linkedin

EDUCATION

University of Pennsylvania **Philadelphia, United States**

Master of Computer and Information Technology *Aug 2024 – May 2026*

- **Relevant Courses:** Machine Learning, Machine Perception, Interactive Computer Graphics, Physical Intelligence: Science and Systems

University College London **London, United Kingdom**

BSc Psychology and Language Science *Sep 2020 – Jun 2023*

- **Grades:** First Class Honors
- **Relevant Courses:** Advanced Statistical Research Methods, Introduction to deep learning for speech and language processing

SKILLS

ML & AI: LLM Benchmarking (vLLM, TensorRT-LLM, llmPerf), Speculative Decoding, Mixture-of-Experts (MoE), Foundation Models (SAM, MiDaS, FoundationPose), Multimodal Fusion (EEG/fMRI/MRI)

Data & Research: Feature Engineering, Neuroimaging (FreeSurfer, fMRIPrep), Model Explainability, Statistical Modeling

Programming & Systems: Python, C, C++, CUDA Kernel Programming, PyTorch, TensorFlow, PySpark, SQL, Bash, Git, Docker, Linux

EXPERIENCES

Amazon Web Services - SageMaker Inference **LLM Efficiency & Benchmarking**

Software Development Engineer Intern *Sep 2025 - Nov 2025*

- **Engineered GPU-aware orchestration** for vLLM using **tensor-parallel scheduling, synchronized serving rounds, and dataset-aware job packing**, maintaining **OTPS/TTFT accuracy** while **reducing end-to-end runtime by 50%** for large-scale runs
- **Minimized GPU idle time** via **KV-cache warm reuse, parallel server bring-up**, and **round-based dispatch**, improving **GPU-time utilization** across multi-GPU clusters under diverse workload shapes and concurrency levels
- **Enhanced inference efficiency** using **kernel-level profiling, micro-batch tuning, and fair workload allocation heuristics**, enabling **higher compute density** and more consistent **memory footprint** across benchmarking workloads
- **Automated scalable benchmarking workflows** with **ECS-driven config sweeps** and **Lambda-triggered vLLM workloads**, providing **hands-free evaluation** for parallel model releases and deployment variants
- **Standardized benchmark schemas** for prompt length, token limits, and model configs, ensuring **cross-model comparability** and reproducible performance metrics across benchmarking experiments

Apple **Apple Pay**

Data Analytics Intern *Mar 2024 - Aug 2024*

- **Engineered large-scale behavioral datasets** in **Python (Pandas, PySpark)** to analyze **human–device interaction patterns** across 100M+ Apple Pay transactions, supporting intent inference modeling
- **Developed interpretable ML pipelines** using **XGBoost and PyTorch** to predict **user engagement and anomaly likelihood**, integrating model explainability metrics for fairness and transparency
- **Collaborated with applied ML teams** to define **human-interpretable evaluation metrics** that bridge model accuracy with user behavior consistency, advancing human-centered model design

Alibaba Group **Alibaba Digital Media & Entertainment Group**

User Experience Intern *Nov 2023 - Mar 2024*

- **Analyzed large-scale user interaction logs** across Alibaba's e-commerce platform using **Python and SQL**, extracting behavioral features to support **recommendation and engagement models**
- **Collaborated with ML researchers** to evaluate personalization algorithms through **A/B testing and human feedback loops**, refining ranking models based on **user satisfaction metrics**
- **Designed interpretable UX metrics** connecting **human decision patterns** (clicks, scrolls, dwell time) with algorithmic outputs, improving alignment between model behavior and real-world user intent

RESEARCH

University of Pennsylvania – Rehabilitation Robotics Lab (GRASP Lab) **Infant Toy Interaction Project**

Research Assistant *Jul 2025 - Sep 2025*

- **Developed a multimodal behavior analysis pipeline** combining **pose estimation, gaze tracking, and force modeling** from synchronized RGBD videos to quantify infant–toy interactions
- **Applied foundation models (MiDaS, SAM, FoundationPose)** for **depth, segmentation, and 6D pose estimation**, enabling 3D motion reconstruction and **transformer-based intent prediction**
- **Linked spatiotemporal pose dynamics** to **motor coordination metrics**, uncovering early cognitive and motor development patterns from real-world motion data

Stanford University – Computational Neuropsychiatry Project **Deep Learning for Psychiatric Subtyping**

Research Assistant *Jul 2025 - Sep 2025*

- **Developed deep learning models for patient subtyping**, extracting **latent neural embeddings** from EEG, fMRI, and MRI to identify data-driven psychiatric phenotypes
- **Built multimodal fusion pipelines** integrating **EEG + cognition + MRI**, enabling shared-representation learning of **neurofunctional biomarkers** across modalities
- **Applied autoencoder and contrastive learning frameworks** to capture cross-subject variability, advancing explainable models for early disorder prediction