Yixin Huang

(774)737-4480 yixinhuang48@gmail.com San Diego, CA

EDUCATION

University of California, San Diego, CA
Master of Science in Computer Science Sep 2025 - Present
GPA: 4.00/4.00

University of California, Berkeley, CA
Bachelor of Arts in Computer Science and Applied Mathematics Aug 2021 – May 2025
GPA: 3.86/4.00

HIGHLIGHTED EXPERIENCES

Hao AI Lab May 2025 – Present
Research Assistant

- Integrated Doom environment into the Lmgame Bench framework, enabling LLM agent evaluation with harnesses such as memory and perception modules
- Authored a technical blog post demonstrating how to use Lmgame Bench as an evaluation framework for both single-player and multi-player LLM agents
- Built VideoScience-Bench, a scientific video evaluation benchmark assessing physical and chemical correctness of video generation models with VLM-as-a-Judge scoring (CVPR submission)
- Integrating Sokoban and Tetris game environments into Nvidia's NeMo Gym, enabling large-scale reinforcement learning training and evaluation

Perceptis AI Jun 2025 – Aug 2025
Software Engineering Intern

- Built an LLM-based pipeline that converted natural language slide descriptions into consulting-grade PowerPoint decks, significantly reducing HITL team effort
- Designed and implemented internal MCP tooling using FastMCP, enabling seamless integration of Claude with tools like custom RAG search over company-specific data
- Conducted exploratory supervised fine-tuning (SFT) on Gemini-2.5 Flash via Vertex AI using Slideworks consulting slides to evaluate potential improvements in model performance for automated slide generation

Emotect AI Mar 2025 – May 2025
Machine Learning Engineering Intern

- Led data curation and schema design for a proprietary emotional classification dataset, incorporating Core Memory Events (CME), Emotion Lexicon (EL), and contextual attributes like valence, intensity, and time decay
- Built a pipeline using LLMs and web scraping to generate and clean thousands of emotional memory samples across diverse cultural and psychological contexts
- Enhanced emotional reasoning of LLaMA model by fine-tuning with Unsloth and LoRA adapters, resulting in improved accuracy and nuanced emotional interpretation across different categories
- Used vLLM inference for evaluation to analyze model responses, identifying gaps in emotional reasoning

Sky Computing Lab – UC Berkeley May 2024 – May 2025
Undergraduate Research Assistant

- Engineered a scalable data generation pipeline for Openfunctions-v3, producing thousands of multi-turn function-calling training data using LLM-based programs with advanced prompt engineering, demonstrating the effectiveness of synthetic data generation for model training
- Spearheaded data cleaning and refinement for BFCL-v3, an open-source leaderboard for evaluating multi-turn and multi-step function-calling abilities of LLMs, accompanied by a widely-read blog post that gained recognition from tech companies including Meta, Salesforce, and Alibaba
- Enhanced LLM evaluation by refining LLM-as-a-Judge to align more closely with human preferences, contributing to the development of Arena-Hard-v2, a rigorous benchmark for assessing LLMs' conversational abilities

XpBrand AI May 2024 – Aug 2024
Software Engineering Intern

- Spearheaded the development of a buffer system for queuing LLM requests and responses for the XpBrand POC application, aimed at assisting executives in their decision-making process
- Utilized FastAPI, Streamlit, RabbitMQ, Docker, and PostgreSQL to build and deploy the buffer system

SKILLS

Python, C++, Java, Rust, SQL, Bash, JavaScript, PyTorch, TensorFlow, Scikit-learn, Transformers, LangChain, vLLM, Unsloth, SGLang, Prompt Engineering, Supervised Fine-Tuning (SFT), Distributed Training (3D Parallelism), RAG, NLP, Deep Learning, LLMs, LLM Agent, CUDA, MapReduce, RPC, Spark, Docker, RabbitMQ, FastAPI, FastMCP, Streamlit, PostgreSQL, Linux, Git, Distributed Systems, Operating Systems, Computer Networks, Algorithms, Database Systems