Guanghao(Jack) Su

http://jacksuuu.github.io/jacksuuu-portfolio/ Email : 18922443765jack@gmail.com
https://github.com/JackSuuu Mobile : +44 07421023765
https://www.linkedin.com/in/jack-suuuu

Experience

- **Perception Software Engineer** Edinburgh University Formula Student
  EUFS *Sep 2025 - Present*
  - **Perception Pipeline Acceleration**: Leveraged *HipSYCL(AdaptiveCpp)* to accelerate a 3D perception pipeline, focusing on the parallel optimization of computationally intensive PCL modules (e.g., noise filtering, clustering, projecting point cloud)
- **Research Intern** Huawei
  Edinburgh Huawei Research Institute *Jun 2025 - Aug 2025*
  - **SeverlessLLM**: Designed and implemented the migration of the ServerlessLLM distributed system to the Huawei Ascend 910B3 NPU, transitioning from a CUDA-based architecture to the Huawei CANN framework. Achieved a 6x to 8x improvement in LLM loader time performance compared to the PyTorch loader.
- **Research Assistant** University of Edinburgh
  Informatics MLSys Research Lab *Apr 2025 - May 2025*
  - **SeverlessLLM**: Researched efficient serverless inference strategies for Mixture-of-Experts (MoE) language models
- **Platform Software Engineer** Edinburgh University Formula Student
  EUFS *Sep 2024 - May 2025*
  - **EUFS Testing Platform**: Designed a modular software architecture for managing vehicle testing data, utilizing TypeScript state transition features to streamline data handling, and integrated python matplotlib with graphical visualization enhancements for real-time analysis.
  - **ROS2 Integration**: Automated ROS2 package downloads boosted download speed by 40%, streamlining workflows and enhancing team productivity.
- **Research Assistant** University of Edinburgh
  Informatics ML Research Lab *Sep 2024 - Apr 2025*
  - **AI4Whisky**: Implemented an interface for a whisky carbon emission calculator, iterating and mainlining model hyper-parameter to enhance accuracy and provide actionable insights.
  - **AI Doctor Agent**: Built an AI chatbot using Nerif framework and SambaNova Cloud, offering real-time medical assistance, facilitates more accurate guidance for users by utilizing log probabilities within the API

Projects

- **Transformer Model (53M) Implementation in Apple MLX**: Engineered a 53M-parameter GPT-2 architecture from scratch using Apple's MLX framework. Implemented Pre-LayerNorm transformer with 8 layers, achieving 27.5K tokens/s training throughput and 169 tokens/s inference on Apple M2 Pro. Trained on 10M tokens of TinyStories dataset for 20K iterations (loss: 0.758). Model available on Hugging Face.
- **IntelliH1**: Developed a *cognitive humanoid navigation framework* combining LLM planning with Unitree RL policy controller for H1 robot. Architected 4-layer system integrating natural language understanding, C++ optimized perception (pybind11, less 10ms LIDAR processing), A* path planning, and adaptive motion control. Solved critical navigation challenges: speed parameter propagation, heading error correction, and autonomous stopping (1.2m tolerance). Achieved stable navigation at configurable speeds (0.5-3.0 m/s) in *MuJoCo* simulation.
- **JasOS Kernel**: JasOS-Kernel is a Unix-inspired, bare-metal operating system kernel built from scratch as a personal project, designed to run on ARM architecture using the QEMU emulator. It showcases core OS concepts such as memory management and process scheduling.
- **ContextGPT**: Developed a context-based AI agent by integrating the GROQ API, the LangChain framework and the Chroma vector database to enhance the interpretation capabilities of files. Optimized the processing and retrieval of PDF materials, enabling efficient and interactive learning experiences. This solution improved accuracy and response speed, streamlining how users engage with academic content.

- **AnalyzeGPT**: Developed an AI agent for cryptocurrency market analysis and trading strategy generation. Utilized FastAPI for the backend and React for the frontend, enabling real-time market updates and personalized insights. Implemented voice handling using Twilio's Programmable Voice API and Grok's Voice Mode , facilitating interactive user engagement.
- **Battle Infinity**: Engineered a browser-based 2D fighting game utilizing Kaboom.js, incorporating Brain.js to implement a neural network that enables AI opponents to learn and adapt to player strategies through reinforcement learning, thereby enhancing gameplay complexity and user engagement.
- **Papering**: Developed a MacOS Python application with a PyQt6 GUI to automate the retrieval of A-Level exam papers and mark schemes. Utilized web scraping techniques to dynamically generate URLs, enabling users to access desired documents directly through their browser, thereby streamlining the study process.

Education

- **University of Edinburgh** Edinburgh, UK
  Bachelor of Science Honor in Computer Science *Sep 2023 - Present*
  - **Relevant Courses**: Object-Oriented Programming, Computer System, Algorithms and Data Structure, Foundation in Data Science, Computational Cognitive Science, Operating System, Natural Language Processing

Competitions

- **Hack the Burgh XII**: Huawei AI & openEuler OS challenge track *first-place* winner
- **UKIEPC Spring 2025**: Achieved *top 1* ranking in UK ICPC competitive programming contest.
- **UKIEPC 2024**: Ranked in the top 50% in a national-level competitive programming contest.
- **LeetCode Weekly Contest**: Achieved top 50% ranking

Skills and Certifications

- **Languages**: Python, C/C++, Java, HTML/CSS, JavaScript, Typescript, SQL, Haskell, MIPS, Latex
- **Skills / Framework**: Git, Pytorch, TensorFlow, Sckit-learn, React, React Native, Next.js, Streamlit, Numpy, Pandas, Matplotlib, Seaborn, ROS2, Docker, AWS, Azue ML, Google Cloud, Vercel
- **Certification**: *Coursera Machine Learning Specialization*