Junlin (Lionel) Hu

lionelh@andrew.cmu.edu | linkedin.com/in/lionelhu

EDUCATION

Carnegie Mellon University — **Pittsburgh, PA**
Master of Science in Machine Learning — *Dec. 2027*

University of California, Los Angeles — **Los Angeles, CA**
Bachelor of Arts in Linguistics and Computer Science; Bachelor of Science in Cognitive Science — *Aug. 2025*
Minor in Data Science Engineering | GPA: 3.91/4.00

SKILLS

Programming: Python, Bash, C/C++, JavaScript
Machine Learning: PyTorch, Hugging Face Transformers, NumPy, Pandas, scikit-learn
LLM Training and Inference: vLLM, flash-linear-attention, Unsloth, DeepSpeed
Web and Tooling: React, Node.js, MongoDB, Git, LaTeX
Languages: Mandarin (Native), English (Fluent), Vedic/Classical Sanskrit (Limited Reading Proficiency)

RESEARCH EXPERIENCE

Language Understanding and Synthesis Laboratory, University of California, Los Angeles — **Los Angeles, CA | Remote**
Research Assistant | Advisor: Professor Nanyun (Violet) Peng — *Jul. 2025 – Oct. 2026*

- *Hybrid-attention distillation:* Distilled Qwen3-4B and Granite-4.1-3B into hybrid recurrent-attention students on a single L40S GPU with flash-linear-attention, engineering staged weight transfer and KL-based logit distillation for multilingual learning
- Discovered a full-attention layer placement that beats the canonical layout of SOTA hybrid LLMs in multilinguality under identical distillation, cutting KL divergence to the teacher by **18–26%** and cross-entropy gap by **34–53%** across models on 26 languages
- *Multilingual MoE routing:* Designed the "token-level routing-consistency" metric exposing non-overlapping distribution of domain-specific expert subnetworks in open-source models; steered routing logits using vLLM for consistent **1–2%** relative accuracy gains on MGSM and MMLU (medical)

Neural Engineering and Computation Laboratory, University of California, Los Angeles — **Los Angeles, CA**
Machine Learning Research Assistant | Advisor: Professor Jonathan Kao — *Jun. 2025 – Jun. 2026*

- Engineered *LightBeam*, a GPU CTC beam-search decoder replacing the baseline's graph-represented n-gram language model with a lightweight LLM rescorer finetuned with Unsloth, cutting decode memory usage **30-fold** to 10 GB RAM and 6 GB VRAM
- Implemented a time-masked Transformer encoder within a 3 GB VRAM budget, utilizing dynamically chunked attention and constrained left context to flexibly serve low-latency streaming and full-context decoding from one checkpoint
- Achieved **5.10%** and **2.21%** word error rate on the Brain2Text 2024 and 2025 benchmarks, state of the art at submission, with statistically significant gains (10 seeds) over baseline decoding
- Built *Brainaudio*, a PyTorch-based Python package unifying preprocessing, training, and evaluation across brain-to-text datasets

INTERNSHIP EXPERIENCE

Shanghai Artificial Intelligence Laboratory — **Shanghai, China**
Research and Development Intern | OpenGV Group | Manager: Dr. Zhiyong Wu — *Aug. 2024 – Oct. 2024*

- Built reproducible test suites with Android Studio for evaluating agent assistance in mobile environments as a part of *AgentStore*, a platform which integrates heterogeneous agents into generalist computer assistants
- Engineered the vision modality for the desktop computer-use agent *"Friday"(OS-Copilot)* by pairing optical character recognition with OS accessibility-tree parsing, which formulates the GUI elements into structured, model-readable representation

PROJECTS

Rubato: Music Social Networking Application — **Los Angeles, CA**
Full-Stack Web Application, University of California, Los Angeles — *Jan. 2024 – Apr. 2024*

- Built a full-stack application with React, Node.js, MongoDB, and Passport.js, designing interfaces for and implementing the post functionality, *mood* tagging, and *trending* feed, with a YouTube-based ingestion fallback

PUBLICATIONS

Bandarkar, Lucas*, **Junlin Hu***, Chenyuan Yang, Mohsen Fayyaz, and Nanyun Peng. "Multilinguality in Hybrid Attention LLMs." *arXiv* 2609.35378 (2026). Web. [Under review; Co-first author]

Feghhi, Ebrahim*, **Junlin Hu***, Nima Hadidi, and Jonathan C. Kao. "LightBeam: An Accurate and Memory-Efficient CTC Decoder for Speech Neuroprostheses." *Proceedings of Interspeech* (2026). Web. [Co-first author]

Bandarkar, Lucas, Chenyuan Yang, Mohsen Fayyaz, **Junlin Hu**, and Nanyun Peng. "Multilingual Routing in Mixture-of-Experts." *International Conference on Learning Representations* (2026). Web.

Jia, Chengyou, Minnan Luo, Zhuohang Dang, Qiushi Sun, Fangzhi Xu, **Junlin Hu**, Tianbao Xie, and Zhiyong Wu. "AgentStore: Scalable Integration of Heterogeneous Agents as Specialized Generalist Computer Assistant." *Findings of ACL 2025* (2025): 8908–8934. Web.