Haoling (Brian) Pu

haolingp@andrew.cmu.edu | (734) 834-5987 | github.com/HaolingPu | linkedin.com/in/haoling-pu/

EDUCATION

Carnegie Mellon University *Pittsburgh, PA*
Master of Science in Artificial Intelligence and Innovation, School of Computer Science (GPA: 4.00/4.00) May 2027
Relevant Coursework: AI Engineering, Deep Learning, Multimodal ML, NLP, AI Agents, LLM Systems, Generative AI

University of Michigan - Ann Arbor *Ann Arbor, MI*
Bachelor of Science in Computer Science & Data Science (GPA: 3.95/4.00) May 2025

SKILLS

LLM & Agent Engineering: Agentic Workflows, Tool Calling, RAG, LangChain, vLLM
ML, Systems & Cloud: PyTorch, CUDA, MLflow, TensorFlow, Slurm, Kubernetes, Docker, AWS, Megatron-SWIFT
Programming Languages: Python, C++, C#, SQL, JavaScript, Go, Java

WORK EXPERIENCE

Google | ***Software Engineering Intern****, Los Angeles, CA* *05/2026 - 08/2026*

- Engineered a **NotebookLM-style agentic knowledge system** that automatically generates, serves, and continuously self-maintains a team engineering knowledge base for a **20+ engineer team**
- **Deployed** the end-to-end agentic pipeline as scheduled distributed jobs, automating recursive source discovery, delta ingestion, **Gemini** synthesis, and automated change delivery across **1,000+** engineering docs
- Built a semantic auditing and self-maintenance layer to detect stale or inconsistent knowledge and automatically propagate source updates, reducing outdated-document defects by **90%**
- Developed the team-wiki system using a **multi-agent engineering workflow** (EGM), orchestrating AI agents for architecture design, TDD implementation, code review, adversarial design critique, and automated debugging
- Productionized the system for **cross-team adoption at Google**, exposing it through a reusable **AI agent skill** that delivers source-cited answers, accelerating onboarding and knowledge discovery

CMU Language Technologies Institute, Li Lab | ***Research Assistant****, Pittsburgh, PA* *10/2025 - Present*

- Developed a reference-free, future-consistent **decoding algorithm** for simultaneous speech translation using **Gemma** and **Qwen** future predictions to prevent premature target emissions, achieving COMET parity with a reference-based baseline
- Engineered a resumable, distributed vLLM/Slurm synthesis and inference pipeline across up to 24 **NVIDIA L40S** GPUs, orchestrating parallel generation of streaming translation trajectories
- Generated training data trajectories from 40K GigaSpeech utterances and **LoRA**-fine-tuned Qwen3-Omni with **Megatron-SWIFT**, achieving a **+5.4 BLEU** improvement through SimulEval-based evaluation on ACL En-Zh testset

WeRide | ***Backend Development Intern****, Shanghai, CN* *05/2025 - 08/2025*

- Reduced Knative cold-start latency by 88% (12.2s to 1.5s) by building a service warming pool with symlink-based resource mapping and **Kubernetes** state recovery backed by PostgreSQL
- Productionized a containerized REST service with Bazel testing and CI/CD, and scaled **Prometheus/Grafana** health monitoring across a 120-GPU cluster, cutting scan time by 60% while maintaining 99.9% uptime

UMSN Pregnancy App | ***Machine Learning Engineer Intern****, Ann Arbor, MI* *05/2024 - 08/2024*

- Led a 6-person team to launch a HIPAA-compliant AI prenatal care platform serving 20+ expectant mothers
- Integrated a GPT-4o-powered RAG agent with n8n to generate personalized care summaries and nursing tasks; built Redis/JWT authentication that reduced login latency by 60% and database load by 70%

PROJECT EXPERIENCE

NVIDIA MLSys 2026 Competition | ***Sparse Attention on Blackwell*** *02/2026 - 04/2026*

- Developed a Blackwell-optimized **CUDA** kernel for DeepSeek-V3 sparse attention, applying **FlashAttention**-style IO-aware fusion to reduce **KV-cache** traffic and accelerate long-context LLM inference
- Optimized Tensor Core execution, memory pipelining, parallel decomposition, and workload autotuning on **NVIDIA B200**; passed 23/23 attention workloads at ~57 us, achieving 22-35x speedup over the PyTorch reference

CMU Advanced NLP | ***Hybrid Retrieval-Augmented Generation System*** *01/2026 - 03/2026*

- Built an end-to-end **RAG** pipeline from scratch, converting heterogeneous HTML and PDF sources into chunked embeddings and searchable FAISS/BM25 indexes
- Engineered hybrid retrieval with Reciprocal Rank Fusion and local open-source LLM inference for domain-specific factual QA, with automated evaluation across retrieval relevance and answer quality