HENRY HE

(314) 657-5517 | hhernye@gmail.com | linkedin.com/in/henry-he-40a44531b | github.com/HenryWashuHe

Summer 2027 Availability: May 14 – September 7, 2027 | Open to Relocate

EDUCATION

Columbia University **Expected May 2028**
B.A. in Computer Science | GPA: ***3.96*** *New York, NY*
Coursework: Data Structures & Algorithms(A+), Databases, Machine Learning, NLP (A+), Reinforcement Learning (A+)
Leadership: Columbia Space Initiative — Subteam Lead, NASA Lunar Terrain Vehicle AR project (1 of 10 university teams selected to present at NASA Johnson Space Center); Hack WashU — Vice President

EXPERIENCE

ByteDance Jul 2026 – Sep 2026
AI Full-Stack Engineer Intern, Douyin Social & Messaging *Shanghai, China*
- Shipped **in-feed private messaging** for Douyin in **Kotlin/Android**, owning client-side design and implementation; used **AI-assisted development tools** and validated changes through testing and staged rollouts.
- Reused **Quick Reply Panel presenter logic** across view hierarchies, resolving feed-lifecycle mismatches while keeping feature-specific UI and experiment logic out of shared messaging modules.
- Fixed an **experiment-gating regression** for control users; added **remote rollback controls** for staged releases.
- Developed a **Lottie-based widget installation guide** with configurable triggers and display-frequency controls.

Trizen AI May 2026 – Jul 2026
AI Engineer Intern *Hangzhou, China*
- Benchmarked **Fun-ASR-Nano, Paraformer, and SenseVoice** using **word error rate**, medical-terminology recognition, and inference latency to compare clinical transcription quality and accuracy–latency trade-offs.
- Built a **Python speech-data pipeline** processing **500+ hours of clinical audio**, combining voice activity detection, forced transcript alignment, and confidence-based filtering to generate training-ready speech–text pairs.
- Engineered **GPU evaluation and preprocessing tooling for 800M-parameter speech models**, supporting reproducible domain-adaptation, fine-tuning, and deployment workflows for medical speech recognition.

Helios Healthcare Advisors Nov 2025 – Apr 2026
Software Engineer Intern *New York, NY*
- Built **Flask REST APIs and PostgreSQL pipelines** to ingest, normalize, and serve CMS data for **15,000+ U.S. nursing homes**, supporting facility search, comparison, and monitoring.
- **Reduced search latency by 30%** through SQL query tuning, indexing, and pagination; built scikit-learn ranking models that **improved recommendation relevance by 20%**.

PROJECTS

Squeeze — Morgan Stanley Code to Give | GitHub **|** 2nd Place Mar 2026
Next.js | FastAPI | PostgreSQL | AWS Bedrock | LangGraph
- Built a **full-stack AI-powered volunteer platform** with **Next.js, FastAPI, and PostgreSQL**, owning backend APIs and agent integration.
- Developed **LangGraph/AWS Bedrock agents** for campaign content generation and outreach, using structured tool calls, guardrails, and human-in-the-loop approval.

COMET — LLM Inference & Agent Memory Optimization Fall 2026
Python | PyTorch | Mini-SGLang | GCP A100 | Qwen3
- Developing an **LLM systems research framework** to co-optimize **KV-cache memory efficiency and test-time compute** for multi-step agents under latency and memory constraints.
- Extending **Mini-SGLang** to evaluate KV-cache eviction (StreamingLLM, R-KV, AgentKV), checkpoint/restore, and dynamic memory recovery with best-of-N inference.

RESEARCH

EleutherAI Jul 2026 – Aug 2026
Research Fellow, Summer of Open AI Research (SoAR) *Remote*
- Investigated **right-answer–wrong-reason behavior** in LLMs on drug-toxicity benchmarks, extracting **residual-stream activations** and training **linear probes** to study how scientific concepts are represented.
- Evaluated **probe sensitivity and concept separability** across prompts, contexts, and adversarial examples using **Python, PyTorch, and Inspect AI**, testing whether probes track scientific concepts or spurious cues.

Complex Resilient Intelligent Systems (CRIS) Lab Jan 2026 – May 2026
AI Research Engineer *New York, NY*
- Analyzed **LLM representations for mathematical reasoning** using **PCA, normalization, and probing** to quantify concept separability and relate internal representations to problem-solving behavior.
- Designed **cross-format reasoning experiments** spanning natural-language and symbolic problems, with reproducible **Python/PyTorch evaluation on HPC GPU clusters** to compare model outputs and representations.

TECHNICAL SKILLS

Languages: Go, Java, Python, C/C++, Kotlin, JavaScript/TypeScript, SQL
Frameworks: FastAPI, Flask, Node.js, React, Next.js, Android, PyTorch, LangGraph, Playwright
Data & Infrastructure: PostgreSQL, AWS (Bedrock), GCP, Docker, Linux, Git, CI/CD