

Hiep "Henry" Nguyen

(+1) (682) 365-5307 · Dallas - Fort Worth Metroplex, TX

hiepnguyentcu@gmail.com · [linkedin.com/in/hiepnnnguyentcu/](https://www.linkedin.com/in/hiepnnnguyentcu/) · github.com/hiepnnnguyentcu · hiepn.dev

EDUCATION

Texas Christian University

M.S. Computer Science

B.S. Computer Science, Minor in Mathematics — **Magna Cum Laude, GPA: 3.73/4.0**

Awards: Louise Dilworth David Scholar · Boller Award Finalist · Andrews Honors Scholar · Kathryn McDorman Honors Scholar · Dean's List

Relevant Courses: Operating Systems · Computer Networks · Parallel Computing · Distributed Systems · Data Structures & Algorithms

Certification: AWS Certified Cloud Practitioner

Fort Worth, TX

May 2028

May 2024

SKILLS

Languages: Python, Go, Java, CUDA, Triton, C++, TypeScript, SQL

Distributed Systems & Infra: Apache Kafka (EOS), FastAPI, Spring Boot, Celery, Redis, Debezium, Stripe

ML Platforms & Core AI: PyTorch, GRPO, RLHF/DPO, SLURM, Weights & Biases, Langfuse

Databases & Data: PostgreSQL, MongoDB, Snowflake, dbt, Airflow, Elasticsearch

DevOps & Cloud: Docker, Kubernetes, KEDA, OpenTofu (Terraform), Prometheus, AWS, Datadog

EXPERIENCE

Texas Christian University

Graduate Student Researcher | Multi-agent Systems & Post-training (RLHF & GRPO)

Fort Worth, TX

May 2026 - Current

- Built an RL pipeline using Qwen2.5-Coder to auto-generate CUDA/Triton kernels, achieving **2.4x** speedups over PyTorch baselines
- Increased KernelBench fast@1 accuracy from **14% to 58%** by fine-tuning the solver via GRPO, designing a reward function that gated execution correctness and penalized runtime latency
- Scaled multi-node RL training across a **16-GPU NVIDIA H200** cluster by building an automated, high-throughput rollout and trace-collection harness using SLURM, Docker, and Python

Alma (Series A startup)

Software Engineer | LLM pipelines & Billing Core

Palo Alto, CA

Dec 2025 - May 2026

- Engineered an async document verification pipeline using Gemini VLM, FastAPI, and Celery, processing **4K+** monthly uploads with real-time Langfuse tracing for LLM latency and reliability
- Cut database read latency **10x** under strict multi-tenant RBAC rules by migrating expensive relational queries to precomputed JSONB snapshot tables with event-driven updates
- Scaled B2B billing engine to process over **\$500K/month** in recurring revenue by architecting a custom Stripe-backed subscription platform with complex tier proration and automated invoicing

CBRE Group Inc. (Fortune #128)

Software Engineer II | Mobile Applications - Facility Management

Richardson, TX

March 2025 - Dec 2025

- Engineered a Spring Boot BFF handling **2M+** daily requests, cutting P95 latency by **65%** using Redis caching with a tag-based, Kafka-driven invalidation pipeline for real-time mobile consistency
- Reduced push notification latency **40%** for **50K+** users by building a Kafka pipeline processing **1M+** daily events with exactly-once semantics, structured dead-letter queues, and batched FCM delivery
- Architected an offline-first mobile sync engine and GraphQL layer handling **3M+** daily work orders using MongoDB Realm, earning the **CBRE CDTO Excellence Award** as the primary engineer

Software Engineer I | Operational Reporting - Facility Management

June 2024 - March 2025

- Cut global reporting latency from **24 hours to under 5 minutes** and eliminated pipeline failures across **12** downstream integrations by engineering Debezium and Snowpipe CDC data pipelines
- Consolidated telemetry across **20+** microservices for EMEA, APAC, and US units by designing Snowflake star schemas and orchestrating automated dbt and Airflow analytics pipelines

Software Engineer Intern | Request Platform

June 2023 - June 2024

- Processed **2M+** daily maintenance requests with sub-millisecond overhead by engineering a Spring Boot, Hibernate, and PostgreSQL validation microservice to execute real-time business rule checks

FPT Software Ltd. | Software Engineer Intern - Hanoi, Vietnam

April 2022 - July 2022

- Led a 4-intern team to ship a **30K-user** employee portal (React Native, GraphQL, Node.js) with automated CI/CD

MISA JSC. | Software Engineer Intern - Hanoi, Vietnam

April 2021 - July 2021

- Saved **10+** hours/week by building a Python/Selenium invoice automation service deployed on Docker/AWS ECS

PROJECTS

Jarvis | Wearable AI Memory for Meta Ray-Ban Glasses | [GitHub](#)

April 2026 - July 2026

- Built a real-time Meta Ray-Ban speaker ID pipeline using Resemblyzer embeddings over WebSockets to map live audio to contacts
- Devised an Apache AGE knowledge graph in PostgreSQL using LLM SPO triple extraction to surface context-aware contact briefings

MiniAttention | CUDA Attention Kernel | [GitHub](#)

April 2026 - June 2026

- Wrote a tiled CUDA attention kernel, raising arithmetic intensity from **7.5 to 131 FLOPs/byte** for compute-bound execution
- Verified FP32 numerical correctness against PyTorch and confirmed **99.8%** in-kernel GPU execution via Nsight profiling

Bibo.AI | Multilingual Children's Book Studio | [Demo](#)

Feb 2025 - Dec 2025

- Engineered a concurrent Go and Celery/Redis asset engine to orchestrate parallel text, image, and voice generation
- Implemented a production hybrid search engine combining Elasticsearch and vector databases for semantic content recommendations