MARCO ZHUANG

New York, NY | 332-276-7106 | mz572@cornell.edu | https://marcozc19.github.io/

EDUCATION

Cornell Tech (Cornell University), New York, NY   May 2025
Dual MSc in Information Science and Information System | GPA: 4.0
- ▪ **Honors/Awards:** Cornell Merit-Scholarship
- ▪ **Relevant Courses:** Machine Learning Engineering, NLP, Algorithm & Data Structure, Blockchain Applied Data Science

Tsinghua University, Beijing, China   Jun 2023
BEng in Industrial Engineering
- ▪ **Honors/Awards:** government merit-based scholarship (1 of 15 recipients state-wide, 2021)
- ▪ **Relevant Courses:** Machine Learning & Big Data, Intelligent Engineering System, Statistics & Probability, Operations Research

TECHNICAL SKILLS

Coding Language:	Python, SQL, HTML, CSS, R, JavaScript, Hadoop, Spark, C, C++
Other Tools:	Huggingface, Git, Pytorch, Langchain, LlamaIndex, PgSQL, Scikit-Learn, React, MongoDB

EXPERIENCE

Echo3D, Software Engineer Intern, New York, NY   Jun 2024 – Present
- ▪ Designed and built a comprehensive notification system using **AWS** Notification service for new product iteration, including user notification setting and sidebar in the **Angular TypeScript**, and Servlet and Manager in the **Java**, enhancing user control.
- ▪ Integrated and optimize new console item event trigger logic using **conditional propagation**, reducing latency by **400%** .
- ▪ Refined auto-email update feature by adding new email alert triggers to multiple events.
- ▪ Rectified over **7 hot-fix tickets** within frontend and backend services, including API response format irregularities, content display bugs, etc., ensuring a smoother user experience and a more reliable application.

4149AI, AI Engineer Intern, New York, NY   Dec 2023 – Jun 2024
- ▪ Developed and published 7 core hypertasking blueprints for zoom meeting and slack channel using **prompt engineering**.
- ▪ Built in-house **GraphRAG** query tool with Docker, NebulaGraph, LlamaIndex on linux syste**m**.
- ▪ Collaborated with **3** prioritized clients and developed **12** client-specific task blueprints, increased client engagement rate by **33%**.

Bytedance, Product Management Intern, Beijing, China   Mar 2022 – Aug 2022
- ▪ Conducted user profiling on 6 competitors through **200+ user surveys**, concluded **4 general user habits**, finalized UGC toolkit logic and user interface design, achieving **6,000+** monthly active creator within a month.
- ▪ Co-lead 31 expert interviews, compiled 7 feature trends, ideated 15 content ideas, 3 were adopted and launched with the product.
- ▪ Synthesized investment thesis, competitor analysis, and product performance evaluation into the internal pitchbook, resulting in a long-term collaboration and introduction of a leading VR UGC platform onto Pico ecosystem.

PROJECTS

NLP - Generative AI SQL Queries | Python, Hugging Face, PyTorch, Weights & Biases   Apr 2024 – May 2024
LLM and Transformer models for flight database SQL generation
- ▪ **Fine-tuned** a **T5 encoder-decoder** transformer to write valid SQL commands from natural language instructions.
- ▪ Prompt engineered **Gemma 1.1 2B** by providing instructions including the database schema and in-context learning with k-shots.

NLP – Sentiment and Newsgroup Classification | Python, PyTorch   Dec 2023 – Jan 2024
Full lifecycle (data extraction, featurization, model architecting, and training) NLP sentiment analysis and classification project
- ▪ Fine-tuned MLP architecture, hyper-parameters, and features, including regularization in dropout, weight decay, and batch norm.
- ▪ Used data-driven insights to fuel featurization techniques including n-grams, BOW, TF-IDF, word stemming, negation handling, and custom punctuation-based tokenization.

LLM - RepoRover | Python, Langchain, OpenAI, FAISS, Streamlit | [Deployed App]   Dec 2023 – Jan 2024
Developer tool capable of helping users understand GitHub repositories with chat assistant built on top of GPT-3.5 Turbo
- ▪ Built and deployed a chatbot for opensource repositories using streamlit with OpenAI's **GPT-3.5 turbo**.
- ▪ Implemented **retrieval augmented generation (RAG)** using **LangChain** and **FAISS** by chunking file structure and file content.
- ▪ Incorporate **prompt engineering** to complement RAG output for narrowing query scope, enhancing response precision by **20%**.

Machine Learning – MiniTorch | Python, NumPy, Numba, CUDA | [link]   Aug 2023 – Dec 2023
Full re-implementation of the deep learning framework PyTorch
- ▪ Optimized performance of back-end operations by adding **parallelization** and GPU compatibility using **CUDA** and **Numba**.
- ▪ Implemented Tensor auto-differentiation through topological sort on computation graph and chain rule backpropagation.
- ▪ Built and trained CNN including base modules to perform sentiment analysis and image classification.

ADDITIONAL INFORMATION

Languages: English (Native), Mandarin (Native), German (Fluent)

Hobbies: Basketball: Shanghai Jr.NBA MIP 2019; Football: Team Captain 2019; Piano: Level 10 Certification