# Bill Qian

+1 (317) 993-2669 bill.qian@yale.edu billqian

## Education

**Yale University** New Haven, CT
*MS in Computer Science, GPA: 3.93* *Aug. 2022 – May 2026*
*Courses: Systems, Deep Learning, NLP, Operating Systems, Computer Architecture, Algorithms, Game Engines*
*Awards: USA Computing Olympiad – Platinum Qualifier (**Top 300** USA), USA Mathematics Olympiad Qualifier (**Top 200** USA), Google CodeJam Round B, USCF Chess Rating: 1980+ (Top 1000 USA Collegiate)*

## Work Experience

**R&D Software Engineering Intern** Jun. 2024 – Nov. 2024
*The New York Times* *New York City, NY*
- Developed AI tag generation framework utilizing OpenAI API and SpaCy, completely replacing external vendor.
- Researched and developed LLM proofreading tools using fine-tuning and RAG. Utilized by 2000+ journalists with positive feedback.
- Deployed custom high-performance LLM and retrieval software on GCP to handle 200k+ requests/day.
- Optimized Firestore database queries, indexing strategies, and application architecture, resulting in a 40% reduction in data retrieval times.
- Designed 40+ RESTful APIs to integrate newly developed frameworks with existing content management systems

**Research Assistant** Dec. 2022 – Present
*Gerstein Laboratory, Yale University* *New Haven, CT*
- Developed distributed machine learning training system using AWS, Azure, and Yale clusters. Used for large-scale model training, inference, and scientific workloads. Tools include NCCL, Kubernetes, PyTorch.
- Developed software to build, train, and evaluate large language models for research. Used Python, PyTorch, and C/C++ for low-level SIMD optimizations on Metal and CUDA
- Published 2 research papers on LLM agents and code generation, with 230+ citations

**Software Engineering Intern** Jun. 2023 – Dec. 2023
*DeepMedia AI* *Oakland, CA*
- Created translation model with 10% better ROUGE score than existing solutions. Used PyTorch and Transformers
- Designed algorithms for the analysis of 2TB+ of multi-modal (text, video, audio) data. Used Python, Rust, SQL
- Scaled single-GPU app to 8 GPUs with over 7x performance improvement. Enabled new real-time deepfake streaming and translation feature. Utilized NCCL, FFmpeg, WebRTC, in addition to previous tools
- Implemented a caching mechanism that reduced API response times by 40%, significantly improving user experience.

**Software Engineering Intern** Jun. 2021 – Aug. 2021
*Thomas Ho Company Ltd.* *New York City, NY*
- Used Python and TensorFlow to apply neural networks to assess performance of mortgage loan pricing model. This work is being published in the *Handbook of Financial Technology, Statistics, Econometrics, and Risk Management*

## Research Publications

**ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs** Aug. 2023
- Developed novel methodologies, including Chain-of-Thought and DFS, applying LLMs to real-world solutions
- **Accepted** at ICLR 2024 | 4500 GitHub stars | in collaboration with **WeChat AI**

**BioCoder: A Benchmark for Bioinformatics Code Generation with Contextual Pragmatic...** Jun. 2023
- Xiangru Tang*, **Bill Qian***, et al. | Used Python, Java, Docker, and RabbitMQ, among others
- **Accepted** at ISMB 2024 | in collaboration with **Google DeepMind** researchers

## Additional Experiences

**Yale RoomAdvisor** | *Project Lead* Jan. 2023 – Present
- Led 20+ student developers in implementation of RESTful APIs (Express.js), iOS/Android apps (Flutter), and a web app (React). Oversaw the entire product and full-stack software development lifecycle.

## Technical Skills

**Languages**: Python, C, C++, Java, Go, Scala, JavaScript, SQL, CSS, HTML, PHP, Bash
**Frameworks & Libraries**: Node.js, React, TensorFlow, PyTorch, Kafka, Linux/UNIX, Redis,CUDA, Metal
**Developer Tools**: AWS, Azure, Google Cloud, Kubernetes, Docker, HuggingFace, Git, CI/CD, Virtualization