GABRIEL BO

gabebo@stanford.edu | (469) 878-0892 | linkedin.com/in/gabriel-bo | github.com/gabrielkmbo | gabrielbo.com

EDUCATION

Stanford University B.S. Computer Science – Artificial Intelligence & Systems Track | GPA: 4.0/4.0 | Graduating: May, 2027

Relevant Courses: Parallel Computing, Compilers, Computer Network Programming, Convex Optimization, Data Structures & Algorithms (A+), Operating System, Computer Architecture, Linear Algebra & Multivariable Calculus (A+), Matrix Theory, Discrete & Continuous Math (A+), Statistics Probability, Physics (A+), Deep Learning ML, NLP, Reinforcement Learning

EXPERIENCE

DriveWealth LLC – *Quantitative Developer* | New York, NY May 2025 – August 2025

- Architected an ultra-low-latency multi-threaded trading cash-posting engine on **Unix, TCP,** and **AWS OpenShift (K8s)** with **C++, Java/Spring Boot 3, Kafka,** and **Debezium CDC**—streaming in **1 million+ order events at 150 ns E2E** to deliver real-time **PnL for quants and brokers**.
- Orchestrated a zero-blind-spot observability and auto-scaling platform by wiring **Prometheus, Telegraf metrics, and Grafana** dashboards into Jenkins-driven Docker/Helm CI-CD, cutting maintenance costs and elevating **equity trade returns in simulations by 35% across 200+ pods**.
- Engineered **unit-tests, integration tests**, and **Bash** data-extraction pipelines on **DynamoDB and MySQL** that accelerated quant trade updates.
- Empowered traders with a "shadow market" simulator that replays tick-level equities fills, extending DriveWealth's trade capabilities to mutual funds, fixed-income, and dividend reinvestments, **cutting onboarding from T+1 to T-0** and unlocking **12% more deployable capital**.

Stanford Artificial Intelligence Lab (SAIL) – *HazyResearch & Scaling Intelligence Research Associate* | Stanford, CA September 2024 – Present

- Leading a first-authored paper on LLM agentic systems routing with a team of **5+ other undergraduate students**, submitted to **ICLR 2026**.
- Building **kernels, ML systems**, and models to improve **test-time compute** and ML scaling advised by Professor Chris Ré, and Azalia Mirhoseini
- Collaborated on **Weaver (accepted: Neurips 2025)** to make verifiers with weak reward models, pioneering as the **3rd method of LLM scaling**.
- Optimized machine learning pipelines by leveraging **embedded systems, hyperparameter tuning, LoRA, and CoT techniques** with Hugging Face and PyTorch to scale NLP input documents up to **32K tokens long**, improving performance by **23.3% with 90x fewer parameters**.
- Engineered scalable pipelines on H-100/A-100 GPUs, GCP, AWS, and Stanford clusters for database and Huggingface integration.

AfterQuery (YC W25) – *Machine Learning Researcher and Software Engineer* | San Francisco, CA December 2024 – Present

- Leading a research and engineering team on tasks to improve benchmarking on pre-trained LLMs that uses **post-training techniques** such as **GRPO, PPO, Q-Learning, and reward learning (RLAIF)** that improves finetuned model by **+12% for OpenAI, Llama, Gemini, and Claude** tools.
- Generated synthetic and hand-written datasets by using autoencoders, PCA methods, feature extraction, R, and Stata to assist more than **100+ companies** in accelerating the scaling of agentic applications– published for **Neurips 2025 and ICLR 2026**.

PocketChange Digital – *Technical Co-Founder* | San Francisco, CA November 2023 – January 2025

- Founded a startup to simplify gift card liquidation & improve fractional trading, securing **$50K** in initial funding, **receiving YC W25 interviews**.
- Developed an end-to-end fintech solution encompassing a trading algorithm, API-based financial SaaS for corporate integration, JWT and SSO security, a custom **ML retail recommendation engine**, B2B data analytics and financial markets capabilities, and a Stripe payment system.
- Built a full-stack app (Python Django/Firebase/AWS) ensuring secure, scalable transactions that was downloaded by **1,000+ students**.

TriTech Software – *Software Engineer Intern* | Plano, TX June 2023 – September 2023

- Collaborated in an **Agile environment** with Gitlab, Jenkins, GCP cloud to work on the software to file premium tax for insurance companies.
- Revamped RESTful APIs to manage user's options using **Kotlin, Spring Boot and PostgreSQL**, experienced in full-stack web development.

ACADEMIC PROJECTS

Step-Wise Policy for Rare-tool Knowledge (SPaRK) – *Client: CS224R & Scaling Intelligence (Outstanding Custom Project)* June 2025

- Conceived and spearheaded "SPaRK" — an offline-RL framework that teaches Llama-3.1 8B to choose **rare but high-utility tools** — generating **12.5k synthetic trajectories**, fine-tuning with **PPO & QLoRA**, and boosting **MMLU-Pro accuracy from 22% to 40.8%** (82% relative lift).

GPT Meets Graphs and KAN Splines: Frameworks on Multitask Fine-Tuned GPT-2 – *Highlights: CS224N Best Default Project* March 2025

- Integrated LoRA with GPT-2 using PyTorch and Hugging Face to fine-tune multi-task NLP models (sentiment analysis, paraphrase detection, sonnet generation) with optimized self-attention (**RoPE encoding, multi-head transformers**) on **A100/H100 GPUs** via GCP.
- Benchmarked advanced architectures by combining **KAN and GAT with LoRA/DORA, PyTorch, and RLHF**, demonstrating expertise in GPU computing and scalable transformer development, achieving 55.2% accuracy on SST, 99% on CFIMDB, and ~90% on paraphrasing **(2nd in class)**.

LEADERSHIP AND HONORS

- Jane Street Guts++ (Harvard MIT Math Tournament) Champion 2025
- 4X American Invitational Mathematics Examination (AIME) Qualifier 2020, 2021, 2022, & 2023
- 2X National Finalist in International Extemporaneous Speaking (National Speech and Debate Association) 2022 & 2023

COMPUTER SKILLS

Java, Python, Kotlin, C++, C, React, Node.js, AWS, GCP, Pytorch, Tensorflow, SQL, Firebase, Flutter, Javascript, Typescript, TCP, NumPy, Tailwind, Git, EC2, CI/CD, LangChain, CoT, Transformers, CUDA GPU, Linux, vLLM, MongoDB, Huggingface, Distributed Systems, Lambda, Kubernetes, HPC