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

Classes: Data Structure & Algorithms (A+), Operating System (A), Comp Arch (A), Linear Algebra & Multivariable Calculus (A+), Matrix Theory (A), Discrete & Continuous Math (A+), Probability (A), Physics (A+), ML (A), Computer Vision (A), NLP (A)

EXPERIENCE

DriveWealth LLC – *Incoming Post-Trade Software Engineer (Quant Developer)*; New York, NY May 2025 – August 2025

PocketChange Digital – *Technical Co-Founder;* San Francisco, CA November 2023 – Present

- Leading a startup focused on simplifying gift card liquidation & trading, securing over $50K in initial funding, receiving YC W25 and Techstars NYC final round interviews (top 1% of 10,000+ start up applicants), and signing LOIs with 3+ CEOs (McDonald's, Dunkin, PDQ).
- Developed an end-to-end 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 capabilities, and a Stripe payment system.
- Built a full-stack app (Dart/JS frontend, Python Django/Firebase/AWS backend) with Docker, Redis, ensuring secure, scalable transactions.

Stanford Artificial Intelligence Lab (SAIL) – *NLP Group Research Assistant* | Stanford, CA September 2024 – Present

- Developed RAG and long-context retrieval models at SAIL's NLP group, collaborating with advisors Professor Christopher Ré, and Azalia Mirhoseini (Hazy & Scaling Lab) on projects featured in publications such as Building Long-Context Retrieval Models with LoCo and M2-BERT.
- Optimized NLP and ML pipelines by leveraging hyperparameter tuning, SGD, LoRA, and CoT techniques with Hugging Face and PyTorch to enhance model performance and to build benchmarks and evaluations of existing retrieval (RAG and long-context) models.
- Engineered scalable pipelines on H-100/A-100 GPUs, GCP, AWS, and Stanford clusters for database integration.

Stanford University Graduate School of Business – *Software Research Intern*; Stanford, CA May 2024 – September 2024

- Analyzed 5,000+ survey entries (150+ variables) in R, Stata, and Python, hand-crafting and visualizing chi-squared tests, regressions, and confidence intervals to assess public opinion's impact on Supreme Court decisions.
- Fine-tuned LLM models in Python on Google Cloud, processing 1,500+ files (court documents, appeals) to classify cases by specific justices.

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.
- Developed RESTful APIs to manage user's options using Kotlin, Spring Boot and PostgreSQL, experienced in full-stack development.

ACADEMIC PROJECTS

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.

LegalEval Large Language Model: Understanding Legal Texts – *Client: Semantic Evaluation ACL* January 2024

- Created a custom PyTorch data loader that improved LLaMA2 accuracy by 15% for log probability analysis used in SemEval 2023 LegalEval.
- Implemented model parallelism techniques using Hugging Face Accelerate and DeepSpeed ZeRO 3 to cut training time by 30% on A100 GPUs.

ACM Convolutional Neural Network Bird Classifier – *Client: Google BIG-Bench* November 2023

- Led the development of a custom bird classification model using PyTorch, achieving 87% accuracy across a dataset of 3,113 bird images.
- Implemented data augmentation techniques and GAN image generation to reduce overfitting by 20%, improving generalization on data.

LEADERSHIP EXPERIENCE AND ACTIVITIES

New Age Learning & Be The Light Youth – *Co-Founder, Instructor,* & *Executive Director of Expansion*; USA Fall 2020 – 2023

- Coached more than 30 students in speech and debate that resulted in over 3 national finalists and 12 state finalists in California.
- Led and organized a 501(c)(3) non-profit, raising over $175,000 to support children in need of educational assistance in Texas.

HONORS

- 4X American Invitational Mathematics Examination (AIME) Qualifier 2020, 2021, 2022, & 2023
- 2 Time National Finalist in International Extemporaneous Speaking (NSDA) 2022 & 2023
- AWS Certified Cloud Practitioner 2024

ADDITIONAL INFORMATION

Computer Skills: Java, Python, Kotlin, C++, C, React, Node.js, AWS, GCP, Pytorch, Tensorflow, SQL, Dart, Javascript, Typescript, iOS, NumPy, Deep Learning, NLP, Tailwind, Git, EC2, CI/CD, RAG, CoT, Transformers, GPUs, Graph Attention, Linux, LLM, Machine Learning, Databases