Qian QU

Palo Alto, CA | eequq99@gmail.com | github.com/quq99 | quq99.github.io | linkedin.com/in/qian-qu

Summary

Machine Learning Engineer at Amazon (6+ years; Ads Foundation Model, Annapurna Labs, **Amazon Rufus**) building LLM agents, GPU kernels (**#1** on NVIDIA SOL-ExecBench), PyTorch on Trainium, and LLM data and serving systems.

Work Experience

Machine Learning Engineer — Amazon, Palo Alto, CA

LLM agents, GPU kernels, ML frameworks, LLM data & serving — *Mar 2020 – Present*

Ads Foundation Model — *Sep 2026 – Present*

- **CompoundingKernel** (arXiv 2611.19230, first author): conceived, designed and built a self-improving LLM agent system that writes NVIDIA B200 kernels (CUDA, Triton, CUTLASS) from PyTorch, with no reference kernels.
- **Leaderboards: #1** on the L1, L2 and Quantization boards of NVIDIA SOL-ExecBench (as of Nov 11, 2026); correct kernels for all **235** problems; **4.40**× geomean speedup over PyTorch on the 209 public ones.
- **Search design:** a per-problem archive of every correct kernel and its lineage, an LLM controller that picks a parent or new design for parallel child agents, and correctness checks outside the model.
- **Scale & results: 1,000+** concurrent agent sessions; one warm restart raised evaluated candidates per hour **3.6**×; at equal wall time the controller beat a UCB-style tree search on **35 of 38** problems.

Annapurna Labs: PyTorch on AWS Trainium — *Dec 2025 – Sep 2026*

- **TorchNeuron (torch-neuronx):** core team member enabling native PyTorch on Trainium, open-sourced (Apache 2.0) at the PyTorch Conference in Oct 2026; PyTorch code runs on Trainium with a one-line change.
- **Operators, tests & models:** code contributor for PyTorch ATen operators (torch-mlir), upstream PyTorch tests, accuracy and Hugging Face model onboarding; LLM training, evaluation and distributed training on Trainium.

Amazon Rufus (now Alexa for Shopping): Foundation Model & Data — *Apr 2023 – Dec 2025*

*Early member of the team behind the conversational shopping assistant used by **millions** of customers on the Amazon app.*

- **Pre-training data pipeline:** led and launched a refreshable pipeline processing **trillions** of tokens (AWS EMR Serverless, Airflow) to transform, clean, de-duplicate and PII-filter product catalog, reviews, community Q&A, A+ and OCR data; cut manual effort from **1 week** per dataset to nearly **no human touch**.
- **Customer data system:** led and designed the system for model release gating, evaluation, traffic replay and RLHF, meeting privacy, security, governance and compliance requirements, and drove cross-team integration.
- **LLM playground & serving:** launched two FastChat playgrounds for LLMs and prompts, used for **demos to the Amazon CEO**; customized vLLM: **4**× faster inference, tensor-parallel serving, 10 to 50–100 chat rounds, no OOM.

Brand Classifier: ML for the product catalog — *Mar 2020 – Apr 2023*

- Launched ML brand classification at **2,000** TPS in **12** marketplaces; raised UK recall **57%** to **72%** at **95%** precision.

Publications (arXiv, 2026)

Qian Qu, et al. *CompoundingKernel: Self-Improving from Zero to First.* 2611.19230; Project page
R. Wen, . . . , **Q. Qu**, . . . *Plumbline: When the Loss Lies, Validating LLM Training Correctly.* 2611.13393
Y. Yan, . . . , **Q. Qu**, . . . *Sandman: A Sandbox Service for Production Agentic Reinforcement Learning.* 2610.39711
L. Fan, . . . , **Q. Qu**, . . . *ATTUNE: Agentic Throughput Tuning for LLM Training.* 2610.17528

Education

University of Virginia, *M.S. Computer Science, GPA: 3.9/4.0* — *2018 – 2019*
Nanjing University, *M.S. Electronic and Communication Engineering, GPA: 4.38/5.0* — *2015 – 2018*
Nanjing University, *B.S. Electronic Information Science and Technology, GPA: 4.26/5.0,* ***Top 15%*** — *2011 – 2015*
Awards: First Prize of Postgraduate Scholarship (twice, 2015 & 2017); Second Prize of the People's Scholarship (2012)

Skills

Languages: Python, C/C++, CUDA, Triton, Java, Scala, Shell script, SQL, TypeScript
AI & ML: LLM agents, Claude Code, PyTorch, torch-mlir, AWS Trainium, distributed training, CUTLASS, vLLM
Data & Tools: PySpark, Pandas, Docker, Linux, AWS (Bedrock, SageMaker, EMR, Airflow, Lambda, S3, DynamoDB)