Your Name

you@example.com | (555) 210-4487 | github.com/yourname San Francisco, CA

Education

Stanford University — M.S. in Computer Science 2019 – 2021
University of Michigan — B.S. in Electrical Engineering 2015 – 2019

Experience

Machine Learning Engineer 2022 – Present
Anchor AI

- Designed a sparse attention kernel that cut inference latency by 41% on 32k-token contexts.
- Led migration of the training pipeline to a mixed-precision setup, reducing GPU-hours by 28%.
- Mentored two junior engineers and ran the team's weekly paper-reading group.

Software Engineer 2021 – 2022
DataForge Inc.

- Built a feature store serving 200+ models in production with sub-10ms read latency.
- Introduced automated data validation, reducing pipeline incidents by 60%.

Research Assistant 2019 – 2021
Stanford AI Lab

- Co-authored two workshop papers on efficient transformer architectures.
- Implemented a benchmarking suite later adopted by three other labs.

Projects

sparsegrad — open-source library for block-sparse attention training, 800+ GitHub stars.
latency-bench — CLI tool for profiling transformer inference across hardware backends.

Skills

Languages: Python, C++, Rust
ML: PyTorch, JAX, Triton
Infra: Kubernetes, Docker, AWS