

Wei Zhang

AI Systems Engineer | LLM Infrastructure Specialist

Mountain View, CA | (650) 555-4277 | wei.zhang@example.com

linkedin.com/in/weizhang-ai | github.com/weizhang-llm

Professional Summary

Results-driven AI Systems Engineer with 6+ years of experience building infrastructure for training and serving large language models. Specialized in distributed computing architectures, memory optimization techniques, and efficient model serving. Proven expertise in scaling AI systems to handle billion-parameter models while maintaining performance and cost efficiency.

Professional Experience

AI Infrastructure Engineer

ScaleAI Technologies, Mountain View, CA (Jul 2021 - Present)

- Designed distributed training infrastructure supporting models up to 70B parameters using tensor and pipeline parallelism.
- Implemented custom memory-efficient attention mechanisms reducing GPU memory requirements by 35%.
- Built inference serving platform handling 10,000+ requests per second with sub-100ms latency.
- Led migration from static batching to dynamic batching, improving inference throughput by 45%.

Machine Learning Engineer

DataScope, San Francisco, CA (Aug 2018 - Jun 2021)

- Developed efficient data loading pipelines for training LLMs, reducing pre-processing bottlenecks by 60%.
- Implemented automated mixed precision training, reducing memory usage while maintaining model quality.
- Created monitoring dashboards for tracking training progress and detecting anomalies in real-time.
- Built continuous integration systems for model evaluation against regressions.

Software Engineer, ML Infrastructure

TechGiant Corp, Sunnyvale, CA (Mar 2016 - Jul 2018)

- Developed and maintained distributed training framework for NLP models.
- Implemented efficient parameter server architecture for model training.
- Created containerized environments for reproducible ML experiments.

Technical Skills

- **Programming Languages:** Python, C++, Go, Rust
 - **Distributed Computing:** PyTorch FSDP, DeepSpeed ZeRO, Megatron-LM
 - **Infrastructure:** Kubernetes, Docker, Slurm, Terraform
 - **Cloud Platforms:** AWS (EC2, S3, SageMaker), GCP (GKE, TPU VM)
 - **Performance Optimization:** CUDA programming, mixed precision, kernel fusion
 - **Serving Systems:** TorchServe, Triton Inference Server, vLLM, Ray Serve
 - **Monitoring & Debugging:** Grafana, Prometheus, PyTorch Profiler
-

Education

Master of Science, Computer Engineering

Stanford University, Stanford, CA (Graduated: Feb 2016)

- Specialization in High Performance Computing - Thesis: “Optimizing Distributed Training for Neural Networks”

Bachelor of Engineering, Computer Science

Shanghai Jiao Tong University, Shanghai, China (Graduated: Jun 2013)

- GPA: 3.92/4.0 - Focus on parallel computing and algorithms

Technical Publications & Presentations

- Zhang, W., et al. (2022). “Memory-Efficient Inference for Large Language Models.” *MLSys Conference*.
 - Zhang, W., et al. (2020). “Scaling Distributed Training for Transformer Models.” *NeurIPS Workshop on Systems for ML*.
 - Speaker at ScaledML Conference 2022: “Building Infrastructure for Trillion-Parameter Models”
-

Certifications

- NVIDIA DLI Certified Instructor - Accelerated Computing (2021)
 - Google Cloud Professional Machine Learning Engineer (2020)
 - Kubernetes Administrator (CKA) (2019)
-

Open Source Contributions

- Contributor to PyTorch core, focusing on distributed training optimizations
 - Developer of “DistributedLLM” - open-source toolkit for efficient LLM deployment
-

Languages: English (fluent), Mandarin Chinese (native)