Sarah Chen

AI Research Engineer | Large Language Model Specialist
Boston, MA | (617) 555-8742 | sarah.chen@example.com
linkedin.com/in/sarahchen-ai | github.com/sarahchen-llm

Professional Summary

Innovative AI Research Engineer with 5+ years of experience in developing and optimizing large language models. Specialized in computational efficiency and model distillation techniques. Published researcher with expertise in few-shot learning approaches and multilingual capabilities for LLMs. Strong track record of balancing theoretical research with practical implementation.

Professional Experience

Senior AI Engineer
Lexicon AI Labs, Cambridge, MA (Jan 2021 - Present)
- Led development of a 15B-parameter multilingual language model, improving cross-lingual transfer by 35%. - Implemented knowledge distillation techniques reducing model size by 40% while maintaining 95% of performance. - Designed efficient fine-tuning protocols for domain adaptation, reducing training time by 60%. - Mentored team of 3 junior engineers in advanced NLP techniques and best practices.

NLP Research Engineer
TechFrontier Inc., Boston, MA (Mar 2018 - Dec 2020)
- Developed specialized attention mechanisms for improved long-context understanding in transformer models. - Created custom tokenization strategies for technical and scientific domains, improving domain-specific performance by 22%. - Implemented RLHF (Reinforcement Learning from Human Feedback) systems for aligning model outputs with human preferences.

AI Research Associate
NeuralLabs Research, Providence, RI (Jun 2016 - Feb 2018)
- Conducted research on transformer architecture optimizations for memory efficiency. - Built evaluation frameworks for assessing model capabilities across various NLP tasks. - Published research on efficient pre-training methodologies for language models.

Technical Skills

- **Programming Languages:** Python, C++, Julia
- **ML Frameworks:** PyTorch, JAX, HuggingFace Transformers, Keras
- **Model Architectures:** GPT variants, BERT, T5, PaLM, LLaMA, BLOOM
- **Optimization Techniques:** Quantization, Knowledge Distillation, Sparse Attention
- **Distributed Training:** DeepSpeed, Megatron-LM, PyTorch DDP
- **Evaluation Methods:** MMLU, HELM, BIG-bench, GLUE, SuperGLUE
- **Cloud Infrastructure:** AWS, Google Cloud Platform
- **MLOps:** Docker, Kubernetes, MLflow

Education

PhD, Machine Learning
Massachusetts Institute of Technology (MIT), Cambridge, MA (Graduated: May 2016)

- Dissertation: "Efficient Training Methods for Large-Scale Language Models" - Published 4 papers in top ML conferences (NeurIPS, ICML)

Master of Science, Computer Science
Cornell University, Ithaca, NY (Graduated: Jun 2012)
- Thesis: "Attention Mechanisms for Natural Language Understanding"

Bachelor of Science, Mathematics and Computer Science
University of Michigan, Ann Arbor, MI (Graduated: May 2010)
- Summa Cum Laude, Phi Beta Kappa

Research Publications

- Chen, S., et al. (2022). "Scaling Laws for Efficient Model Distillation." *ICLR.*
- Chen, S., et al. (2020). "Multilingual Transfer Learning with Minimal Supervision." *ACL.*
- Chen, S., et al. (2018). "Memory-Efficient Transformer Architectures." *EMNLP.*

Certifications

- NVIDIA Deep Learning Institute Certified Instructor (2022)
- Google Cloud Professional ML Engineer (2021)

Languages: English (native), Mandarin Chinese (native), French (intermediate)