Dr. Priya Patel

Senior AI/Machine Learning Engineer

 [email protected] | +1 (650) 321-9876 | linkedin.com/in/priyapatel-ai | github.com/priyapatel-ml

 Palo Alto, CA 94301 | Open to Remote/Hybrid

 PhD in Computer Science (Machine Learning) | Stanford University

---

Professional Summary

Accomplished AI Engineer with 7+ years of experience designing, developing, and deploying production-grade machine learning systems at scale. Expert in deep learning, NLP, computer vision, and MLOps with proven track record of delivering AI solutions that generated $10M+ in business value. Published researcher with 8 peer-reviewed papers at top-tier ML conferences. Specialized in end-to-end ML pipeline development, from research to production deployment serving millions of users. Passionate about building responsible AI systems that solve real-world problems.

Technical Skills

Machine Learning: Deep Learning, NLP, Computer Vision, Reinforcement Learning, Transfer Learning, Few-Shot Learning, Model Optimization
ML Frameworks: PyTorch, TensorFlow, JAX, Scikit-learn, XGBoost, LightGBM, Hugging Face Transformers
LLMs & Generative AI: GPT-4, Claude, Llama 2/3, Mistral, Fine-tuning, PEFT, LoRA, RAG, Prompt Engineering, LangChain
Computer Vision: CNNs, Vision Transformers, YOLO, Faster R-CNN, Segment Anything, Detectron2, OpenCV
NLP: BERT, GPT, T5, Text Classification, NER, Sentiment Analysis, Question Answering, Summarization
MLOps: MLflow, Kubeflow, Azure ML, SageMaker, Weights & Biases, DVC, Model Monitoring, A/B Testing
Model Serving: TorchServe, TensorFlow Serving, ONNX Runtime, Triton Inference Server, FastAPI
Programming: Python (Expert), C++, SQL, Bash
Data Engineering: Apache Spark, Airflow, Kafka, Ray, Dask, Feature Stores (Feast, Tecton)
Databases: PostgreSQL, MongoDB, Redis, Elasticsearch, Pinecone, Weaviate, Azure AI Search
Cloud Platforms: Azure (ML Studio, Cognitive Services, Databricks), AWS (SageMaker, EC2, Lambda), GCP (Vertex AI)
Infrastructure: Docker, Kubernetes, Terraform, GitHub Actions, Azure DevOps
Data Science: Pandas, NumPy, SciPy, Matplotlib, Seaborn, Jupyter, Statistical Analysis

---

Professional Experience

Senior AI Engineer

MetaScale AI | San Francisco, CA | *January 2022 – Present*

Leading AI initiatives for enterprise B2B SaaS platform with 2M+ users

- **Architected and deployed production LLM system** using GPT-4 and Llama 2 with RAG, processing 500K+ queries daily with 92% user satisfaction and <2s latency, generating $3M additional ARR
- **Built end-to-end recommendation system** using deep learning (PyTorch) that increased user engagement by 35% and improved conversion rates by 28%, directly contributing to $5M revenue growth
- **Designed and implemented MLOps infrastructure** on Azure ML with automated retraining pipelines, reducing model deployment time from 2 weeks to 2 days and enabling 50+ model deployments per quarter
- **Developed computer vision system** for document processing using Vision Transformers, achieving 96% accuracy and reducing manual processing time by 80% (saving 1000+ hours monthly)
- **Led migration of ML infrastructure to Kubernetes**, implementing auto-scaling that reduced inference costs by 45% ($200K annual savings) while improving availability to 99.97%
- **Established model monitoring and drift detection** using custom ML observability platform, reducing production incidents by 60% and improving MTTR from 4 hours to 45 minutes
- **Implemented A/B testing framework** for ML models, enabling data-driven model selection and continuous improvement of AI features
- **Built vector database solution** using Pinecone and Azure AI Search for semantic search, improving search relevance by 50% and reducing search latency from 800ms to 120ms
- **Mentored team of 5 ML engineers**, conducting weekly 1:1s, code reviews, and technical design sessions, resulting in 3 promotions within 18 months
- **Optimized model inference** through quantization and pruning, reducing model size by 60% and improving throughput by 3x while maintaining 98% of original accuracy

Key Achievements: - Received “Innovation Award” for LLM-powered chatbot that deflected 40% of support tickets, saving $800K annually - Published internal ML best practices guide adopted across organization (200+ engineers) - Invited speaker at internal AI summit with 500+ attendees

Technologies: PyTorch, TensorFlow, Hugging Face, GPT-4, Llama 2, LangChain, Azure ML, Kubernetes, Docker, FastAPI, PostgreSQL, Pinecone, Spark, Airflow, MLflow

---

Machine Learning Engineer

TechCorp Analytics | Mountain View, CA | *July 2019 – December 2021*

Developed ML solutions for predictive analytics and business intelligence platform

- **Built predictive maintenance ML model** using XGBoost and LSTM networks that reduced equipment downtime by 30% for manufacturing clients, saving $2M annually
- **Developed NLP pipeline** for sentiment analysis and entity extraction from customer feedback, processing 100K+ documents monthly with 89% F1 score
- **Created automated feature engineering system** that reduced feature development time by 70% and improved model performance by 15%
- **Deployed 15+ ML models to production** on AWS SageMaker, implementing comprehensive monitoring, logging, and alerting infrastructure
- **Designed time-series forecasting models** using Prophet and LSTMs for demand prediction, achieving MAPE of 8% and enabling $1.5M in inventory optimization
- **Implemented distributed training pipeline** using Horovod on multi-GPU clusters, reducing training time from 48 hours to 6 hours
- **Built data preprocessing pipelines** using Apache Spark processing 10TB+ of data daily with 99.9% reliability
- **Established ML code review standards** and testing frameworks achieving 85% code coverage across all ML projects

Key Achievements: - Promoted from ML Engineer to Senior ML Engineer in 18 months based on exceptional contributions - Led successful migration of 20+ models from TensorFlow 1.x to TensorFlow 2.x with zero downtime

Technologies: Python, TensorFlow, PyTorch, Scikit-learn, XGBoost, AWS SageMaker, Apache Spark, Docker, Jenkins, PostgreSQL, Elasticsearch

---

AI Research Engineer

Stanford AI Lab | Stanford, CA | *September 2017 – June 2019*

Conducted research on deep learning for computer vision and NLP during PhD

- **Developed novel attention mechanism** for multi-modal learning (vision + text), improving state-of-the-art by 7% on VQA benchmark
- **Published 5 papers** at top conferences: 2x NeurIPS, 1x CVPR, 1x ACL, 1x ICML (500+ citations total)
- **Built open-source framework** for visual question answering used by 2000+ researchers (GitHub: 3.5K stars)
- **Collaborated with Google Research** on transfer learning project resulting in CVPR publication
- **Mentored 3 PhD students** and 8 MS students on research projects
- **Received Best Paper Award** at NeurIPS 2018 workshop for work on few-shot learning

Technologies: PyTorch, TensorFlow, CUDA, Python, NumPy, Computer Vision, NLP

Notable Projects & Open Source

Vision-Language Transformer (VLT) | *Research Project*

- Developed state-of-the-art model for image captioning and VQA achieving 85% accuracy on MS-COCO
- Paper accepted at NeurIPS 2018 with 200+ citations
- Open-sourced implementation: github.com/priyapatel-ml/vlt (□ 3.5K stars)

Production MLOps Template | *Open Source*

- Created comprehensive MLOps template for Azure ML with CI/CD, monitoring, and retraining
- Includes best practices for model versioning, experiment tracking, and deployment
- Adopted by 500+ organizations: github.com/priyapatel-ml/mlops-template (□ 2.1K stars)

Real-Time Object Detection System | *Side Project*

- Built edge-deployed object detection system using YOLOv8 running on NVIDIA Jetson
- Optimized model achieving 60 FPS on edge device with 91% mAP
- Used for smart retail analytics application

Publications & Research

1. **Patel, P.**, Chen, L., & Zhang, M. (2018). “Attention-based Multi-Modal Learning for Visual Question Answering.” *NeurIPS 2018.* **[Best Paper Award]**

2. **Patel, P.** & Kumar, R. (2019). “Few-Shot Learning for Medical Image Classification.” *CVPR 2019.*

3. **Patel, P.**, Lee, S., & Wang, J. (2020). “Efficient Neural Architecture Search for Edge Devices.” *ICML 2020.*

4. **Patel, P.** et al. (2021). “Robust Fine-tuning of Large Language Models.” *ACL 2021.*

5. **Patel, P.** & Anderson, K. (2023). "Monitoring and Mitigating LLM Drift in Production." *arXiv preprint*.

Citations: 500+ | **h-index**: 8 | **Google Scholar**: scholar.google.com/priyapatel

---

Certifications

- **Microsoft Certified: Azure AI Engineer Associate** | Microsoft | *Issued: March 2023*
- **AWS Certified Machine Learning – Specialty** | Amazon Web Services | *Issued: November 2022*
- **TensorFlow Developer Certificate** | Google | *Issued: August 2021*
- **Deep Learning Specialization** | Coursera (Andrew Ng) | *Issued: 2019*

Education

Doctor of Philosophy (PhD) in Computer Science

Stanford University | Stanford, CA | *2015 – 2019* - **Specialization**: Machine Learning, Computer Vision, Natural Language Processing - **Dissertation**: "Multi-Modal Deep Learning for Visual Understanding" - **Advisor**: Professor Michael Zhang (renowned ML researcher) - **GPA**: 3.95/4.0

Master of Science in Computer Science

Stanford University | Stanford, CA | *2013 – 2015* - **Focus**: Artificial Intelligence and Machine Learning - **Thesis**: "Transfer Learning for Low-Resource Image Classification" - **GPA**: 3.9/4.0

Bachelor of Technology in Computer Engineering

Indian Institute of Technology (IIT) Bombay | Mumbai, India | *2009 – 2013* - **First Class with Distinction** - **GPA**: 9.2/10.0

Key Achievements & Impact Metrics

- □ Deployed **20+ ML models** to production serving **10M+ users** with 99.95% uptime
- □ Delivered AI solutions generating **$10M+ in business value** through increased revenue and cost savings
- □ Reduced ML model deployment time by **85%** (2 weeks → 2 days) through MLOps automation
- □ Improved inference efficiency by **60%** through model optimization and quantization
- □ Published **8 peer-reviewed papers** at top ML conferences with **500+ citations**
- □ Mentored **15+ engineers** across multiple organizations with 5 promotions to senior roles
- □ Open source contributions with **5K+ GitHub stars** across projects
- □ Saved **$1M+ annually** through infrastructure optimization and efficient ML operations

---

Speaking & Teaching

- **Guest Lecturer** - Stanford CS229 (Machine Learning) - 2023, 2024
- **Speaker** - NeurIPS 2023 - "MLOps Best Practices for Enterprise AI"
- **Keynote** - Bay Area ML Meetup (1000+ attendees) - "From Research to Production

ML”
- **Workshop Instructor** - PyTorch Training Workshop (500+ participants)
- **Technical Blogger** - Medium (15K+ followers) - AI/ML technical articles

Professional Affiliations

- Member, Association for Computing Machinery (ACM)
- Member, IEEE Computer Society
- Reviewer for NeurIPS, ICML, CVPR, ICLR conferences
- Mentor, AI4ALL (promoting AI education for underrepresented groups)

Awards & Honors

- **Best Paper Award**, NeurIPS 2018 Workshop on Meta-Learning
- **Stanford Graduate Fellowship**, 2015-2017
- **Grace Hopper Conference Scholarship**, 2016
- **IIT Bombay Institute Silver Medal**, 2013

Languages

- **English**: Fluent
- **Hindi**: Native
- **Gujarati**: Native

Interests

AI Ethics & Responsible AI, Technical Mentorship, Open Source Contribution, Hiking, Classical Indian Music

Portfolio: priyapatel.ai | **Blog**: medium.com/@priyapatel-ml | **Scholar**: scholar.google.com/priyapatel

References available upon request