Job Description – AI Engineer

Company Overview

TechVision Solutions is a leading technology company specializing in enterprise AI solutions and intelligent automation. We help Fortune 500 clients transform their operations through cutting-edge machine learning and artificial intelligence technologies.

Position Title

AI Engineer (Senior Level)

Location

San Francisco, CA / Remote (US-based)

Employment Type

Full-Time, Permanent

Role Summary

We are seeking an experienced AI Engineer to design, develop, and deploy production-grade machine learning systems and AI solutions. The ideal candidate will have strong expertise in deep learning, MLOps, and end-to-end model lifecycle management. You will work on challenging problems spanning natural language processing, computer vision, and predictive analytics, transforming cutting-edge research into scalable production systems that deliver measurable business impact.

Key Responsibilities

- **ML Model Development**: Design, develop, and train machine learning and deep learning models for various business applications including NLP, computer vision, recommendation systems, and predictive analytics
- **Model Deployment**: Deploy ML models to production environments using modern MLOps practices, ensuring scalability, reliability, and performance
- **Pipeline Development**: Build and maintain end-to-end ML pipelines including data ingestion, feature engineering, model training, validation, and monitoring
- **Experimentation**: Conduct experiments to evaluate different algorithms, architectures, and hyperparameters to optimize model performance
- **Production Monitoring**: Implement model monitoring, performance tracking, and automated retraining workflows to maintain model accuracy over time
- **Data Engineering**: Collaborate with data engineers to design and implement efficient data processing pipelines and feature stores
- **Research & Innovation**: Stay current with latest AI/ML research and evaluate new techniques for potential application to business problems
- **Cross-functional Collaboration**: Work closely with product managers, data scientists, and software engineers to understand requirements and deliver AI-powered solutions
- **Documentation**: Create comprehensive technical documentation for models, pipelines, and deployment procedures

- **Optimization**: Optimize models for inference performance, including quantization, pruning, and efficient serving strategies

---

Required Skills & Qualifications

Technical Skills - Core ML/AI

- **Machine Learning**: Strong foundation in supervised, unsupervised, and reinforcement learning algorithms
- **Deep Learning**: Hands-on experience with neural networks (CNNs, RNNs, Transformers, GANs) and modern architectures
- **ML Frameworks**: Proficiency with TensorFlow, PyTorch, and/or JAX
- **NLP**: Experience with language models, transformers, embeddings (BERT, GPT, T5, LLaMA)
- **Computer Vision**: Experience with image classification, object detection, segmentation (YOLO, ResNet, Vision Transformers)
- **Model Evaluation**: Expertise in model evaluation metrics, cross-validation, and statistical testing

Technical Skills - Engineering & Deployment

- **Programming**: Expert-level Python; familiarity with C++, Java, or Scala is a plus
- **MLOps Tools**: Experience with MLflow, Kubeflow, Azure ML, SageMaker, or Vertex AI
- **Model Serving**: Hands-on experience with model serving frameworks (TensorFlow Serving, TorchServe, ONNX Runtime, Triton)
- **Containerization**: Proficiency with Docker and Kubernetes for model deployment
- **Cloud Platforms**: Strong experience with at least one cloud platform (Azure, AWS, or GCP) for ML workloads
- **Data Processing**: Experience with data processing frameworks (Spark, Dask, Ray)
- **Vector Databases**: Familiarity with vector databases (Pinecone, Weaviate, Milvus, Azure AI Search)

Data Engineering Skills

- **Feature Engineering**: Strong skills in feature extraction, transformation, and selection
- **Data Pipelines**: Experience building scalable data pipelines using Apache Airflow, Prefect, or similar tools
- **Databases**: Proficiency with SQL and NoSQL databases; experience with data warehouses (Snowflake, BigQuery, Synapse)
- **Data Formats**: Experience with Parquet, Arrow, HDF5, and efficient data serialization formats
- **Big Data**: Familiarity with distributed computing frameworks (Spark, Dask, Ray)

Professional Skills

- **Experience Level**: 5+ years of professional experience in machine learning engineering or AI development
- **Education**: Master's or PhD in Computer Science, Machine Learning, Statistics, Mathematics, or related field (or equivalent experience)
- **Problem-Solving**: Exceptional analytical and problem-solving abilities with strong attention to detail
- **Communication**: Ability to explain complex technical concepts to both technical and non-technical stakeholders
- **Project Management**: Experience managing ML projects from conception to production deployment

---

Preferred Skills & Qualifications

- **LLM Experience**: Hands-on experience with large language models, prompt engineering, RAG (Retrieval-Augmented Generation), and fine-tuning techniques
- **Generative AI**: Experience with generative models (Stable Diffusion, DALL-E, GPT-4, Claude, Llama)
- **AutoML**: Familiarity with automated machine learning tools and techniques
- **Edge Deployment**: Experience deploying models on edge devices or mobile platforms (TensorFlow Lite, ONNX, CoreML)
- **A/B Testing**: Experience designing and analyzing A/B tests for ML models
- **Distributed Training**: Experience with distributed training frameworks (Horovod, DeepSpeed, FSDP)
- **Research Background**: Publications at top-tier ML/AI conferences (NeurIPS, ICML, CVPR, ACL, etc.)
- **Open Source**: Contributions to open-source ML projects
- **Certifications**:
  - Microsoft Certified: Azure AI Engineer Associate
  - AWS Certified Machine Learning - Specialty
  - Google Professional Machine Learning Engineer
  - TensorFlow Developer Certificate

Model Deployment Experience Expectations

Candidates should demonstrate proven experience in:

1. **End-to-End Deployment**: Successfully deployed at least 3+ ML models to production environments serving real users
2. **Scalability**: Built systems handling 1000+ requests per second with low latency (<100ms p99)
3. **Monitoring**: Implemented comprehensive model monitoring including drift detection, performance tracking, and alerting
4. **MLOps Practices**: Established CI/CD pipelines for ML, automated testing, and versioning for models and data
5. **Production Debugging**: Diagnosed and resolved production issues related to model performance, latency, and accuracy
6. **Multi-Model Management**: Managed multiple model versions and A/B testing in production
7. **Infrastructure as Code**: Used Terraform, CloudFormation, or similar tools for reproducible ML infrastructure

---

Data Engineering Expectations

The role requires strong data engineering capabilities including:

1. **Data Pipeline Development**: Build robust ETL/ELT pipelines processing TB-scale datasets
2. **Feature Stores**: Design and implement feature stores for consistent feature serving across training and inference
3. **Data Quality**: Implement data validation, quality checks, and anomaly detection in data pipelines
4. **Performance Optimization**: Optimize data processing for cost and performance (partitioning, indexing, caching)
5. **Streaming Data**: Experience with real-time data processing using Kafka, Kinesis, or Event Hubs
6. **Data Versioning**: Implement data versioning and lineage tracking (DVC, Delta Lake, Feast)
7. **Collaboration**: Work effectively with data engineers and data scientists on shared data infrastructure

---

Tools & Technologies

Primary ML/AI Stack

- **ML Frameworks**: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM
- **NLP**: Hugging Face Transformers, spaCy, NLTK, LangChain
- **Computer Vision**: OpenCV, torchvision, Detectron2
- **Experiment Tracking**: MLflow, Weights & Biases, Neptune.ai
- **Model Serving**: TorchServe, TensorFlow Serving, FastAPI, BentoML

Cloud & Infrastructure

- **Cloud Platforms**: Azure (Azure ML, Cognitive Services, AI Search) or AWS (SageMaker, Bedrock) or GCP (Vertex AI)
- **Containers**: Docker, Kubernetes, AKS/EKS/GKE
- **Infrastructure**: Terraform, Ansible
- **CI/CD**: GitHub Actions, Azure DevOps, Jenkins
- **Monitoring**: Prometheus, Grafana, Azure Monitor, CloudWatch

Data Engineering

- **Processing**: Apache Spark, Dask, Ray
- **Orchestration**: Apache Airflow, Prefect, Kubeflow Pipelines
- **Databases**: PostgreSQL, MongoDB, Redis, Elasticsearch
- **Vector DBs**: Pinecone, Weaviate, Azure AI Search
- **Data Warehouses**: Snowflake, Azure Synapse, BigQuery

Development Tools

- **IDEs**: VS Code, PyCharm, Jupyter Lab
- **Version Control**: Git, DVC (Data Version Control)
- **Collaboration**: GitHub, GitLab, Bitbucket

Key Performance Indicators (KPIs)

1. Model Performance Metrics

- **Model Accuracy**: Achieve and maintain target metrics (e.g., F1 score >0.90, AUC >0.85)
- **Performance Improvement**: Improve baseline model performance by 10-15% through iterations
- **Model Reliability**: Maintain model uptime >99.5% in production
- **Inference Latency**: Keep p99 latency below target thresholds (typically <100ms)

2. Deployment & Operations

- **Time to Production**: Deploy models from development to production within 2-3 weeks
- **Model Deployment Success Rate**: >95% successful deployments without rollback
- **Incident Response**: Resolve production issues within defined SLAs (P1: 2 hours, P2: 24 hours)
- **Model Monitoring**: Implement comprehensive monitoring for 100% of production models

3. Business Impact

- **ROI**: Deliver measurable business value (cost savings, revenue increase, efficiency

gains)
- **User Adoption**: Achieve target user adoption rates for AI-powered features (>70%)
- **Customer Satisfaction**: Maintain high satisfaction scores for AI-enabled products (NPS >40)
- **Cost Efficiency**: Optimize inference costs by 20% year-over-year through model optimization

4. Innovation & Quality

- **Experimentation Velocity**: Complete 10-15 experiments per quarter
- **Code Quality**: Maintain >85% code coverage for ML pipelines
- **Documentation**: Document 100% of deployed models with model cards and technical specs
- **Knowledge Sharing**: Present findings at 1-2 internal tech talks or external conferences per quarter

5. Collaboration & Leadership

- **Cross-functional Projects**: Successfully collaborate on 2-3 cross-team initiatives per year
- **Mentorship**: Mentor junior ML engineers or data scientists
- **Best Practices**: Contribute to ML engineering best practices and standards
- **Code Reviews**: Participate in 3-5 ML code reviews per week

---

Performance Expectations

First 30 Days

- Complete onboarding and familiarize yourself with existing ML infrastructure and models
- Understand current model architecture, deployment pipelines, and monitoring systems
- Submit at least 2 pull requests (bug fixes, pipeline improvements, or documentation)
- Review and understand data sources, feature engineering processes, and data quality checks
- Build relationships with data science, engineering, and product teams

First 90 Days

- Take ownership of improving or retraining at least one existing production model
- Deploy a new model or significant model update to production environment
- Identify and document technical debt or improvement opportunities in ML infrastructure
- Contribute to ML pipeline optimization or monitoring improvements
- Present findings or progress to the team in a technical review session

First 6 Months

- Lead the development and deployment of a new ML solution from conception to production
- Demonstrate measurable improvement in model performance or operational efficiency
- Establish automated retraining and monitoring for at least one critical model
- Contribute to MLOps practices and infrastructure improvements
- Begin mentoring other team members on ML best practices

First Year

- Successfully deliver 2-3 major ML projects with measurable business impact
- Become a subject matter expert in one or more AI domains (NLP, CV, recommendation systems, etc.)

- Contribute to strategic ML roadmap and architecture decisions
- Published internal best practices or frameworks used by other teams
- Demonstrated leadership in cross-functional initiatives

Ongoing

- Continuously push the boundaries of what's possible with AI technology
- Drive innovation by exploring and implementing state-of-the-art techniques
- Maintain production excellence with reliable, scalable, and performant ML systems
- Mentor and grow the capabilities of the ML engineering team
- Represent the company at conferences, meetups, or through technical publications

---

Compensation & Benefits

- **Salary Range**: $150,000 - $200,000 per year (based on experience)
- **Bonus**: Annual performance bonus up to 20%
- **Equity**: Generous stock options package
- **Health**: Premium medical, dental, and vision insurance (100% coverage)
- **Retirement**: 401(k) with 6% company match
- **Time Off**: 25 days PTO + 12 company holidays + 5 volunteer days
- **Learning**: $3,500 annual professional development budget (conferences, courses, certifications)
- **Remote**: Fully remote or hybrid flexible arrangement
- **Equipment**: Top-tier development hardware (high-end GPU workstations, latest MacBook Pro)
- **Other**: Gym membership, wellness stipend, home office budget ($1,500)

Application Process

1. **Submit Application**: Resume, cover letter, and links to relevant projects/publications
2. **Initial Screening**: 30-minute phone call with recruiter
3. **Technical Assessment**: Take-home ML challenge (4-6 hours, 1 week to complete)
4. **Technical Deep Dive**: 90-minute virtual interview covering:
   - ML theory and algorithms
   - Model deployment and MLOps
   - Data engineering and pipelines
   - Code review of your take-home challenge
5. **System Design Interview**: 60-minute ML system design discussion
6. **Behavioral Interview**: 60-minute interview with hiring manager
7. **Team Fit Interview**: 45-minute conversation with team members
8. **Final Interview**: 30-minute chat with VP of Engineering
9. **Offer and Negotiation**

Timeline: Typically 2-3 weeks from application to offer

What We Offer

- **Cutting-Edge Work**: Work on challenging AI problems with real-world impact
- **Growth**: Clear career progression path from AI Engineer to Staff/Principal Engineer
- **Resources**: Access to high-performance compute resources and latest ML tools
- **Collaboration**: Work with world-class AI researchers and engineers
- **Impact**: See your models serve millions of users and drive business outcomes
- **Culture**: Innovative, collaborative environment that values continuous learning
- **Work-Life Balance**: Flexible schedule and remote-friendly policies

---

Equal Opportunity Employer

TechVision Solutions is committed to fostering an inclusive and diverse workplace. We are an equal opportunity employer and encourage applications from candidates of all backgrounds. We do not discriminate based on race, color, religion, sex, sexual orientation, gender identity, national origin, veteran status, disability, age, or any other protected characteristic.

To Apply: Submit your application at careers.techvision.com/ai-engineer or email your resume, cover letter, and portfolio/GitHub links to with the subject line "AI Engineer Application - [Your Name]"

Application Deadline: Open until filled

Priority Deadline: Applications received by January 15, 2026 will receive priority consideration

Contact: For questions about this position, please contact: - **Dr. James Chen**, AI Engineering Manager - - **Sarah Mitchell**, Technical Recruiter -

We look forward to hearing from talented AI engineers who are passionate about pushing the boundaries of artificial intelligence and creating real-world impact!