NLP Engineer

AI / ML SMB Enterprise

[Company Name] is looking for an NLP Engineer to design and build natural language processing systems that power intelligent features across our products. You will work on problems such as text classification, named entity recognition, semantic search, summarization, and conversational AI. This role combines deep NLP expertise with practical engineering skills to ship production-grade language systems that deliver real business value.

Key Responsibilities

- Design, train, fine-tune, and deploy NLP models for tasks such as classification, extraction, summarization, and search
- Build and maintain data pipelines for text preprocessing, annotation, and feature engineering
- Evaluate and integrate large language models (LLMs) through fine-tuning, prompt engineering, and RAG architectures
- Develop evaluation frameworks including benchmarks, metrics, and human evaluation protocols
- Collaborate with product and engineering teams to define NLP-powered features and ship them to production
- Stay current with NLP research and assess new techniques for practical applicability
- Optimize models for latency, cost, and accuracy in production environments

Required Skills & Experience

- 3+ years of experience building NLP systems in a production environment
- Strong proficiency in Python and NLP/ML libraries (Hugging Face Transformers, spaCy, NLTK)
- Experience with transformer architectures and modern language models (BERT, GPT, T5, LLaMA)
- Hands-on experience with fine-tuning, prompt engineering, and retrieval-augmented generation (RAG)
- Solid understanding of NLP fundamentals: tokenization, embeddings, attention mechanisms, sequence modeling
- Experience with ML experiment tracking and model versioning (MLflow, Weights & Biases)
- Strong software engineering practices: version control, testing, code review, CI/CD

Nice-to-Have

- Experience with vector databases (Pinecone, Weaviate, Milvus) for semantic search
- Familiarity with LLM deployment and serving (vLLM, TensorRT-LLM, Triton)
- Publications or contributions to the NLP research community
- Experience with multilingual NLP or low-resource language challenges
- Background in speech-to-text or multimodal AI systems

Tech Stack

Python PyTorch Hugging Face Transformers spaCy LangChain Pinecone OpenAI API MLflow FastAPI Docker

What We Offer

- Competitive salary and equity package
- Flexible remote or hybrid work arrangement
- Health, dental, and vision insurance
- Annual learning and development budget
- Generous PTO policy

Interview Process

1. Recruiter phone screen (30 min)
2. Technical phone screen with an ML engineer covering NLP fundamentals and system design (45 min)
3. Take-home or live coding exercise: NLP task implementation (3-4 hours estimated)
4. On-site or virtual deep dive: model design, evaluation methodology, and past project walkthrough (3 hours)
5. Final conversation with the hiring manager or ML lead