Complete AI Engineer Roadmap: Full Stack AI Systems

Phase 1: Model Optimization Foundations (6-8 weeks)

Deep Learning Optimization

- **Neural Network Optimization**
  - Gradient descent variants (Adam, AdamW, RMSprop)
  - Learning rate scheduling and warmup
  - Batch normalization, layer normalization, group normalization
  - Regularization techniques (dropout, weight decay, early stopping)
  - Mixed precision training (FP16, BF16)
- **Architecture Optimization**
  - Efficient architectures (MobileNets, EfficientNets, RegNets)
  - Neural Architecture Search (NAS)
  - Depth-wise separable convolutions
  - Attention mechanisms and self-attention
  - Skip connections and residual learning
- **Hardware-Aware Optimization**
  - CUDA programming fundamentals
  - GPU memory optimization and profiling
  - Tensor Core utilization (A100, H100)
  - CPU optimization (SIMD, vectorization)
  - Memory-efficient attention implementations

Model Compression Techniques

- **Quantization**
  - Post-training quantization (PTQ)
  - Quantization-aware training (QAT)
  - INT8, FP16, and dynamic quantization
  - Block-wise quantization for LLMs (GPTQ, AWQ)
- **Pruning**
  - Magnitude-based pruning

    - Structured vs unstructured pruning
    - Gradual magnitude pruning
    - Lottery ticket hypothesis applications
- **Knowledge Distillation**
    - Teacher-student frameworks
    - Feature distillation vs output distillation
    - Self-distillation techniques
    - Progressive knowledge distillation

Tools & Frameworks

- **PyTorch optimization**: TorchScript, torch.compile, profiler
- **TensorFlow optimization**: TF Lite, TensorRT integration
- **ONNX ecosystem**: model conversion and optimization
- **Hardware-specific**: TensorRT, OpenVINO, Core ML

Phase 2: Computer Vision Systems (6-8 weeks)

Advanced CV Architectures

- **Vision Transformers (ViTs)**
    - ViT, DeiT, Swin Transformer
    - Hierarchical vision transformers
    - Efficient attention for vision
    - Hybrid CNN-Transformer architectures
- **Object Detection & Segmentation**
    - YOLO family (v5, v7, v8, v10)
    - DETR and transformer-based detection
    - Mask R-CNN, Segment Anything (SAM)
    - Real-time instance segmentation
- **3D Computer Vision**
    - 3D object detection (PointNet, VoxelNet)
    - Neural Radiance Fields (NeRF)
    - 3D Gaussian Splatting
    - SLAM and visual odometry

CV Optimization & Deployment

- **Real-time Inference**
  - TensorRT optimization for vision models
  - ONNX Runtime optimization
  - Mobile deployment (TF Lite, Core ML)
  - Edge AI hardware (Jetson, Intel NCS)
- **Video Processing**
  - Temporal consistency in video models
  - Video action recognition optimization
  - Real-time video streaming pipelines
  - Frame interpolation and super-resolution

Practical Applications

- Build real-time object detection system
- Optimize NeRF for mobile deployment
- Create efficient video analysis pipeline

Phase 3: Large Language Models & NLP (8-10 weeks)

LLM Architecture & Training

- **Transformer Deep Dive**
  - Multi-head attention mechanisms
  - Positional encodings (absolute, relative, RoPE)
  - Layer normalization variants (RMS, Pre-LN)
  - Activation functions (SwiGLU, GeGLU)
- **Modern LLM Families**
  - Decoder-only models (GPT, Llama, Mistral)
  - Encoder-decoder models (T5, UL2)
  - Mixture of Experts (MoE) architectures
  - State Space Models (Mamba, RetNet)
- **Training Techniques**
  - Pre-training data preparation and filtering

- Instruction tuning and RLHF
- Constitutional AI and safety training
- Parameter-efficient fine-tuning (LoRA, QLoRA, AdaLoRA)

LLM Optimization & Serving

- **Inference Optimization**
  - KV-cache optimization and quantization
  - Speculative decoding and parallel sampling
  - Continuous batching and dynamic batching
  - Memory-efficient attention (Flash Attention, Paged Attention)
- **Model Parallelism**
  - Tensor parallelism and pipeline parallelism
  - Sequence parallelism for long contexts
  - Expert parallelism for MoE models
  - ZeRO optimizer states partitioning
- **Serving Frameworks**
  - **vLLM**: High-throughput LLM serving
  - **TGI**: HuggingFace Text Generation Inference
  - **TensorRT-LLM**: NVIDIA optimized serving
  - **LLaMA.cpp**: CPU-optimized inference
  - **DeepSpeed-MII**: Microsoft inference engine

Advanced NLP Applications

- **Retrieval Augmented Generation (RAG)**
  - Dense retrieval with bi-encoders
  - Hybrid search (sparse + dense)
  - Multi-hop reasoning over documents
  - RAG evaluation and optimization
- **Function Calling & Tool Use**
  - Structured output generation
  - Tool selection and orchestration
  - Multi-step function execution

- Error handling and retry mechanisms

Phase 4: Multimodal AI Systems (6-8 weeks)

Vision-Language Models

- **Multimodal Architectures**
  - CLIP and its variants (OpenCLIP, SigLIP)
  - Vision-language transformers (ViLT, BLIP-2)
  - Large multimodal models (GPT-4V, LLaVA, Flamingo)
  - Multimodal fusion techniques
- **Applications**
  - Visual question answering (VQA)
  - Image captioning and dense captioning
  - Visual reasoning and commonsense QA
  - Document understanding (LayoutLM, Donut)

Audio & Speech Systems

- **Speech Recognition**
  - Whisper and its variants
  - Wav2Vec and self-supervised speech
  - Real-time streaming ASR
  - Multilingual and code-switching ASR
- **Speech Synthesis**
  - Neural TTS (Tacotron, FastSpeech)
  - Voice cloning and few-shot TTS
  - Real-time voice conversion
  - Emotional and expressive speech synthesis

Video Understanding

- **Video-Language Models**
  - Video captioning and summarization
  - Temporal action localization
  - Video question answering

    - Multi-modal video retrieval
- **Video Generation**
    - Text-to-video models (Sora-like systems)
    - Video editing with language
    - Temporal consistency in generated videos
    - Real-time video manipulation

Phase 5: AI Agents & Reasoning Systems (8-10 weeks)

Agent Architectures

- **Single Agent Systems**
    - ReAct (Reasoning + Acting) patterns
    - Planning and execution frameworks
    - Memory systems (episodic, semantic, working)
    - Multi-step reasoning chains
- **Multi-Agent Systems**
    - Agent communication protocols
    - Collaborative task decomposition
    - Consensus mechanisms and voting
    - Hierarchical agent organizations

Advanced Reasoning

- **Symbolic Reasoning**
    - Neuro-symbolic integration
    - Logic programming with LLMs
    - Constraint satisfaction problems
    - Formal verification of AI systems
- **Causal Reasoning**
    - Causal inference in AI systems
    - Counterfactual reasoning
    - Interventional queries
    - Causal discovery algorithms

Agent Frameworks & Tools

- **Development Frameworks**
  - **AutoGen**: Multi-agent conversations
  - **LangGraph**: State-based agent workflows
  - **CrewAI**: Role-based agent teams
  - **Agency Swarm**: Hierarchical agent systems
- **Tool Integration**
  - Code execution environments
  - Web browsing and scraping
  - Database querying and manipulation
  - API integration and orchestration

Phase 6: Generative AI & Creative Systems (6-8 weeks)

Image Generation

- **Diffusion Models**
  - DDPM, DDIM, and sampling strategies
  - Stable Diffusion architecture and optimization
  - ControlNet and adapter techniques
  - Real-time image generation (LCM, Turbo)
- **Advanced Generation Techniques**
  - Inpainting and outpainting
  - Style transfer and artistic generation
  - 3D-aware image generation
  - Video-to-video and image-to-video

Code Generation

- **Code LLMs**
  - CodeT5, StarCoder, WizardCoder
  - Code completion and suggestion
  - Bug detection and fixing
  - Code explanation and documentation

- **Programming Agent Systems**
  - Automated testing generation
  - Code review and optimization
  - Multi-file code generation
  - Software architecture planning

Content Generation

- **Text Generation**
  - Long-form content generation
  - Structured document creation
  - Creative writing assistance
  - Technical documentation automation
- **Audio Generation**
  - Music generation and composition
  - Sound effect synthesis
  - Podcast and audiobook creation
  - Voice-based content adaptation

Phase 7: Production AI Infrastructure (10-12 weeks)

Scalable ML Systems

- **Distributed Training**
  - Data parallelism (DDP, FSDP)
  - Model parallelism strategies
  - Gradient synchronization optimization
  - Fault-tolerant training systems
- **Model Serving Infrastructure**
  - Multi-model serving architectures
  - Auto-scaling based on demand
  - Load balancing strategies
  - Circuit breakers and fallback systems

MLOps & LLMOps

- **Experiment Management**
  - Model versioning and lineage
  - Hyperparameter optimization at scale
  - A/B testing for ML models
  - Continuous integration for ML
- **Monitoring & Observability**
  - Model performance monitoring
  - Data drift and model drift detection
  - Real-time alerting systems
  - Performance optimization feedback loops

Cloud & Edge Deployment

- **Cloud Platforms**
  - AWS SageMaker, Bedrock, and Lambda
  - Google Cloud Vertex AI and Cloud Run
  - Azure ML and Cognitive Services
  - Multi-cloud deployment strategies
- **Edge Computing**
  - Mobile AI optimization (iOS, Android)
  - IoT device deployment
  - Federated learning systems
  - Privacy-preserving edge AI

Phase 8: AI Safety & Ethics (4-6 weeks)

AI Safety

- **Alignment & Control**
  - Constitutional AI principles
  - Reward modeling and RLHF
  - AI safety via debate
  - Interpretability and explainability
- **Robustness & Security**
  - Adversarial attack detection

- Prompt injection prevention
- Model poisoning defense
- Differential privacy in AI

Responsible AI

- **Bias & Fairness**
  - Bias detection and mitigation
  - Fairness metrics and evaluation
  - Demographic parity considerations
  - Algorithmic auditing processes
- **Privacy & Compliance**
  - Data privacy regulations (GDPR, CCPA)
  - Model transparency requirements
  - Consent management in AI systems
  - Audit trails and compliance reporting

Essential Tools & Technology Stack

Core Development

- **Frameworks**: PyTorch, TensorFlow, JAX, Hugging Face
- **Optimization**: TensorRT, ONNX, OpenVINO, TVM
- **Serving**: FastAPI, Triton, TorchServe, Ray Serve
- **Training**: DeepSpeed, FSDP, Colossal-AI, FairScale

Infrastructure

- **Orchestration**: Kubernetes, Ray, Dask, Airflow
- **Storage**: MinIO, S3, GCS, vector databases
- **Monitoring**: Prometheus, Grafana, W&B, MLflow
- **CI/CD**: GitHub Actions, GitLab CI, Jenkins

Specialized AI Tools

- **LLM**: vLLM, TGI, LangChain, LlamaIndex
- **Vision**: OpenCV, YOLO, Detectron2, MMDetection

- **Audio**: Whisper, TTS libraries, audio processing
- **Agents**: AutoGen, LangGraph, CrewAI

Practical Project Roadmap (12 months)

Months 1-2: Optimization Mastery

- Optimize computer vision models for mobile deployment
- Implement efficient attention mechanisms
- Build custom CUDA kernels for specific operations

Months 3-4: Multimodal Systems

- Create vision-language model for document analysis
- Build real-time video understanding system
- Implement speech-to-text with speaker identification

Months 5-6: LLM Applications

- Deploy optimized LLM serving infrastructure
- Build RAG system with vector database
- Create function-calling agent with tool access

Months 7-8: Agent Systems

- Develop multi-agent research assistant
- Build code generation and review system
- Create automated workflow orchestration

Months 9-10: Production Systems

- Design fault-tolerant AI service architecture
- Implement comprehensive monitoring and logging
- Build cost-optimized inference infrastructure

Months 11-12: Advanced Applications

- Create end-to-end generative AI application
- Implement federated learning system
- Build AI safety and monitoring dashboard

Career Success Metrics

Technical Excellence

- Deploy models with <100ms latency at scale
- Achieve >90% cost reduction through optimization
- Build systems handling millions of requests/day

System Architecture

- Design resilient multi-modal AI architectures
- Implement effective model lifecycle management
- Create reusable AI infrastructure components

Business Impact

- Deliver measurable ROI from AI implementations
- Enable new product capabilities through AI
- Reduce operational costs through automation

This comprehensive roadmap covers the full spectrum of AI engineering - from low-level optimization to high-level applications, ensuring you can handle any AI system from concept to production.