Defining The AI Engineer Role

Alexey Grigorev

February 24, 2026

Today's Agenda

1. Dataset - 895 job descriptions from builtin.com
2. Skills - what companies actually hire for
3. Responsibilities - what AI engineers do day-to-day
4. Use cases - what companies build with AI

The Dataset

895

job descriptions from 590 companies

5,694 responsibilities · 4,525 use cases

Searched for "AI Engineer" on builtin.com, January 2026

builtin

FOR EMPLOYERS JOIN | LOG IN

JOBS COMPANIES REMOTE ARTICLES BEST PLACES TO WORK JOB APPLICATION TRACKER

Better Matches. Better Jobs. Happier You.

Latest Tech Jobs Personalized For You.

Job Title or Keyword: AI Engineer

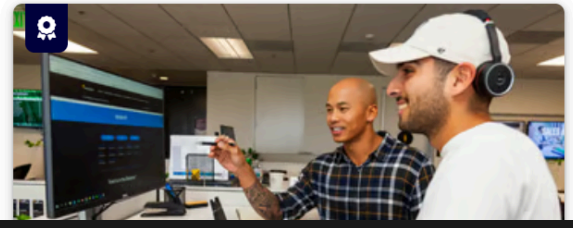
Location: Berlin, State of Berlin, DEU

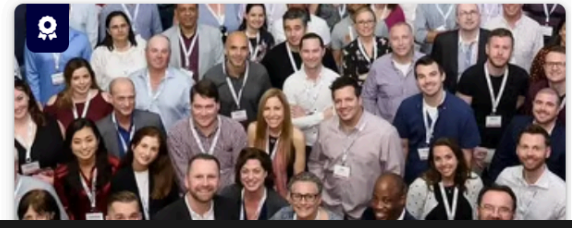
Fully Remote, Hybrid, On Site

SEE JOBS

Explore 111,135 Tech Companies

SEE ALL →





1403	Staff Software Engineer - AI SDK	Temporal Technologies	United States
1404	Staff Software Engineer-AI Agents	Hatch	3 Locations
1405	Senior Software Engineer-AI Agents	Hatch	3 Locations
1406	AI ML Engineer	dv01	USA
1407	Senior Software Engineer, AI Product	Vanta	U.S.
1408	Senior Software Engineer, Applied AI	TLDR	USA
1409	Staff Software Engineer, AI & Automation	Mozilla	US
1410	Software Engineer III/Senior, AI Gateway	ngrok	United States
1411	Sr. Software Reliability Engineer for AI	MixMode	USA
1412	Senior Sales Engineer - AI/ML Advanced Data Architecture - HealthTech	Komodo Health	United States
1413	AI / ML Solutions Engineer	Anyscale	USA
1414	Software Engineer, Ruby - AI Training (Freelance, Remote)	Alignerr	USA
1415	Engineer, DevOps Infrastructure as Code (IaC) - AI Training (Freelance, Remote)	Alignerr	USA
1416	Forward Deployed Engineer – AI Revenue Agents	Outreach	United States
1417	Senior Software Engineer - AI Platform	StubHub	Los Angeles, CA, USA

job-market/_internal/all_jobs.csv

881	Sr. AI Application Developer	Genesys	3 Locations
882	Full Stack Software Engineer - Agentic AI Startup	NinjaTech AI	8 Locations
883	Principal Software Engineer, AI Agents and Tasks	Moon	2 Locations
884	Senior Backend Engineer (Analytics/ AI Personalization)	Tapcart	Santa Monica, CA, USA
885	Senior Software Engineer – Internal AI Tools Lead	Voyager Technologies	3 Locations
886	Software Engineer – Full Stack - Internal AI Tools Team	Voyager Technologies	3 Locations
887	Simulation Engineer / AI Physicist (ACE)	Voyager Technologies	3 Locations
888	Staff Software Engineer, AI/ML Focus	Cambiar Education	California, USA
889	Lead Software Engineer, AI Planning	Divergent	Torrance, CA, USA
890	Full Stack Developer with AI Model Experience (React.js/Node.js/TypeScript)	Bee Techy	Los Angeles, CA, USA
891	Staff AI/ML Engineer	Cambiar Education	California, USA
892	Senior Software Engineer, AI Systems	Subject	Los Angeles, CA, USA
893	Staff Solutions Engineer - Observo AI	SentinelOne	California, USA
894	Principal Full Stack Engineer - AI Product	Re:Build Manufacturing	3 Locations
895	Solution Engineer /SME (AI/Cloud Data Platform)	Alteryx	8 Locations
896	Senior Software Engineer - AI Platform	StubHub	Los Angeles, CA, USA

After deduplication · job-market/_internal/all_jobs_dedup.csv

```yaml
job_id: 1393425
title: Applied AI Engineer & Researcher
company: Speechify
location: USA
work_type: FULL_TIME
level: Expert/Leader
skills:
  - Python
  - PyTorch
  - TensorFlow
company_size: 96 Employees
description: |
  The mission of Speechify is to make sure that reading is
  never a barrier to learning.

  Over 50 million people use Speechify's text-to-speech
  products to turn whatever they're reading - PDFs, books,
  Google Docs, news articles, websites - into audio, so they
  can read faster, read more, and remember more. Speechify's
  text-to-speech reading products include its iOS app, Android
  App, Mac App, Chrome Extension, and Web App. Google recently
  named Speechify the Chrome Extension of the Year and Apple
  named Speechify its App of the Day.
```

1393425_Speechify_Applied_AI_Engineer_Researcher.yaml

```yaml
company:
  name: Speechify
  stage: Series B+
  focus: Text-to-speech and accessibility tools for reading
position:
  title: Applied AI Engineer & Researcher
  ai_type:
    type: ml-first
    reasoning: |-
      The role focuses on building text-to-speech and image
      generation models using deep learning frameworks. The core
      work involves developing TTS systems and NLP/CV models,
      which aligns with traditional ML engineering rather than
      building LLM/agent applications or supporting AI
      infrastructure.
  responsibilities:
  - Research and implement state-of-the-art techniques in NLP, TTS, or CV with focus
    on image generation
  - Build and develop human-sounding AI speech models
  - Deploy NLP or TTS models to production at large scale
  - Manage engineers and grow the research & development team
  use_cases:
  - Text-to-speech conversion for various content types (PDFs, books, articles, websites)
  - Generating human-like AI speech voices
  - Image generation for content
  - Accessibility tools for reading and learning
  skills:
    genai: []
    ml: [TensorFlow, PyTorch]
```

1393425_Speechify_Applied_AI_Engineer_Researcher.yaml

Three Types of "AI Engineer"

Type	Jobs	What they do
AI-First	621 (69.4%)	Build RAG, agents, LLM features
AI-Support	255 (28.5%)	Build platforms, infra, tooling
ML	16 (1.8%)	Traditional ML rebranded

The key question: does this role work ON AI, or NEAR it?

Research vs Applied

Type	Jobs	%
Applied/Production	856	95.6%
Research	39	4.4%

Research	Applied
Novel algorithms, SOTA	Implement existing models in production
Publish papers	Ship AI features to customers
Model architecture design	Build applications with AI APIs

The market wants people who ship, not people who publish

PART 1

Skills

role/02-skills.md

Analysis

- job-market/analysis.ipynb - data analysis notebook
- role/02-skills.md - full skills writeup

RAG Is the #1 Pattern

35.9%

of all jobs mention RAG

Connect LLMs to your data (documents, databases, knowledge bases)

AI Engineers Are Full-Stack

93.1%

of roles need skills beyond just GenAI

- Frontend skills - 195/621 (31.4%)
- Backend skills - 308/621 (49.6%)
- Full-stack (both) - 134/621 (21.6%)

Only 1.4% of roles expect pure GenAI work

Top Skills Demanded

Skill	Jobs	%
Python	738	82.5%
AWS	359	40.1%
RAG	321	35.9%
Docker	277	31.0%
Prompt engineering	260	29.1%
CI/CD	262	29.3%
Kubernetes	260	29.1%
LLMs	227	25.4%
TypeScript	209	23.4%

GenAI Skills

Skill	Jobs	%
RAG	321	35.9%
Prompt engineering	260	29.1%
LLMs	227	25.4%
LangChain	168	18.8%
Agents	129	14.4%
OpenAI API	78	8.7%
LangGraph	72	8.0%
LlamaIndex	52	5.8%
Anthropic API	49	5.5%

GenAI Framework Ecosystem

Framework	Jobs	%
LangChain	168	18.8%
LangGraph	72	8.0%
LlamaIndex	52	5.8%
CrewAI	28	3.1%
AutoGen	17	1.9%

LangChain dominates. LangGraph growing fast for agents.

How Much ML Do You Need?

64.3%

of AI-First roles require some ML knowledge

ML Skill	Jobs	%
PyTorch	165	26.6%
Fine-tuning	159	25.6%
TensorFlow	93	15.0%
Embeddings	81	13.0%
Model evaluation	69	11.1%

Practical ML (PyTorch, fine-tuning, embeddings), not deep research

Fine-Tuning Is Overhyped

Level	Jobs	%
Primary FT responsibility	25	4.0%
Secondary/occasional	94	15.1%
No FT mentioned	502	80.8%

Focus on RAG and agents first.
Learn fine-tuning for domain-specific roles (healthcare, finance, legal).

Production/Ops Skills

Skill	Jobs
Docker	277
CI/CD	262
Kubernetes	260
MLOps	107
Terraform	104

50.2% of AI-First roles require production/ops skills

Evaluation Is the Differentiator

39.6%

of AI-First roles explicitly require evaluation skills

- LLM-as-judge systems
- Golden datasets for RAG performance
- Hallucination detection
- Drift monitoring

Anyone can build a chatbot. Companies hire people who can measure if it works.

Databases

Database	Jobs
Vector databases (general)	97
PostgreSQL	83
Pinecone	53
Redis	43
Weaviate	41

Vector DBs are the new requirement. PostgreSQL still essential.

Cloud Platforms

Cloud	Jobs	%
AWS	359	40.1%
Azure	214	23.9%
GCP	205	22.9%

Programming Languages

Language	Jobs	%
Python	738	82.5%
TypeScript	209	23.4%
Java	133	14.9%
Go	101	11.3%
SQL	88	9.8%

The AI Engineering Stack

Layer	Technologies
Application	React, Next.js, FastAPI
AI orchestration	LangChain, LangGraph, LlamaIndex
LLM APIs	OpenAI, Anthropic, local models
Vector databases	Pinecone, Weaviate, pgvector
Infrastructure	Docker, K8s, AWS/GCP/Azure

PART 2

Responsibilities

role/03-responsibilities.md

The Dataset

5,694

responsibilities extracted from 895 job descriptions

Building AI Systems

Very common

- Design and develop RAG systems for knowledge retrieval
- Build agents using LangChain, LangGraph, CrewAI
- Create LLM-powered applications and features
- Build evaluation frameworks and testing systems
- Develop prompt and template libraries

Core challenge: translating business problems into reliable, scalable AI systems

Productionizing AI

Very common

- Deploy AI software with testing, QA, and monitoring
- Deliver production-ready APIs and microservices
- Take ownership from concept through shipped feature
- Participate in on-call rotation
- Build infrastructure for model serving and inference

Core challenge: making probabilistic AI systems reliable enough for production

Evaluation and Quality

Very common

- Design evaluation frameworks and testing harnesses
- Implement bias assessment and safety guardrails
- Build observability tooling and dashboards
- Monitor for performance, drift, and safety
- Hallucination detection with citations and faithfulness metrics

Core challenge: defining meaningful metrics for non-deterministic systems

Using Provider APIs

Very common

- Integrate OpenAI, Anthropic, and other provider APIs
- Handle errors, retries, and fallbacks
- Optimize prompt design for specific models
- Manage costs through token tracking and caching

Core challenge: building reliable apps on third-party APIs while managing costs and rate limits

RAG and Retrieval

Common

- Implement RAG with vector databases and semantic search
- Document processing and chunking (PDFs, Word, HTML, audio)
- Hybrid search with re-ranking and query rewriting
- Knowledge graphs and enterprise data integrations
- Context window management

Core challenge: accurate semantic retrieval at scale with domain-specific terminology

Data Processing

Common

- Build data pipelines for document processing, indexing, retrieval
- Work with proprietary datasets for fine-tuning and RAG
- Data preprocessing and transformation systems
- Dataset curation, versioning, and quality checks

Core challenge: ensuring data quality at scale while handling diverse formats

Collaboration

Common

- Work with product teams on roadmaps
- Cross-functional collaboration with engineering and design
- Gather business requirements and translate to technical solutions
- Mentor junior engineers
- Maintain documentation and knowledge bases

Core challenge: bridging technical and non-technical communication

Infrastructure and Platforms

Common

- Architect scalable distributed systems
- Build GPU clusters and inference infrastructure
- Contribute to AI platform tooling and Kubernetes ecosystem
- Create platforms that other engineers use to build AI
- Security and compliance infrastructure

Core challenge: building flexible platforms that accommodate rapid AI evolution

Agents and Agentic Workflows

Common

- Design agentic workflows for multi-step tasks
- Configure tool-calling functions, agent memory/state
- Build autonomous systems that execute complex tasks
- Communication protocols for hierarchical agent systems

Core challenge: reliable multi-step planning with graceful failure handling

Less Common Responsibilities

Responsibility	Frequency
Working with customers	Uncommon
Frontend and UI	Uncommon
Performance optimization	Uncommon
Self-hosting models	Uncommon
Fine-tuning models	Uncommon
Experimentation and research	Uncommon
Security and compliance	Rare

Most Common Words in Responsibilities

Word	Mentions
build	684
deploy	565
design	551
teams	493
product	464
implement	432
develop	407
models	405
collaborate	403

The language emphasizes action: build, deploy, design, collaborate

PART 3

Use Cases

role/04-use-cases.md

The Dataset

4,525

use cases extracted from 895 job descriptions

AI-First roles: 3,177 (70.2%) · AI-Support roles: 1,259 (27.8%)

Automating Manual Workflows

696 mentions (15.4%)

- Automate business workflows across Salesforce platforms
- IT operations automation through AI agents with stateful memory
- Autonomous supply chain orchestration
- Process millions of issues for hundreds of customers at scale

AI solution: agents that execute multi-step workflows autonomously

Internal Operational Efficiency

519 mentions (11.5%)

- Early signal detection of emerging risks and threats
- Secure, compliant AI infrastructure for enterprise requirements
- Automated reasoning and evaluation of insurance claims
- Fraud detection and risk assessment for financial services

AI solution: enterprise-grade AI systems for internal operations

Finding Information in Company Data

360 mentions (8.0%)

- Enterprise knowledge retrieval from internal documents
- Medical literature search over millions of articles
- RAG-based system for financial information and guidance
- Knowledge Graph RAG for complex enterprise data

AI solution: RAG and semantic search over proprietary data

Answering Customer Questions at Scale

312 mentions (6.9%)

- Real-time personalized customer experiences
- AI system that understands phone conversations
- Conversational AI for patient interactions and follow-up care
- Process customer inquiries without human intervention

AI solution: customer-facing AI with access to company knowledge

Deploying AI to Production Reliably

219 mentions (4.8%)

- Low-latency production inference systems
- AI inference as a service on edge GPUs worldwide
- Scalable AI-powered UIs at production scale
- Large-scale language model deployment with complex networking

AI works in notebooks but fails in production - latency, scalability, cost

Making Decisions from Data

163 mentions (3.6%)

- Transform complex financial data into actionable insights
- AI-driven health data analysis for improved outcomes
- Predictive signals through AI-enhanced research
- Intelligent event data analysis for marketing workflows

AI solution: AI-powered data analysis and insights

Ensuring AI Quality and Safety

141 mentions (3.1%)

- Real-time content integrity and safety detection
- Automated reasoning and evaluation for accuracy
- Sensitive data handling with appropriate safeguards

AI systems can hallucinate, produce unsafe content, or behave unpredictably

More Use Cases

Problem	Mentions	%
Personalizing user experiences	128	2.8%
Creating content at scale	118	2.6%
Helping developers write code	60	1.3%
Handling specialized domain knowledge	38	0.8%

Domains Served

Domain	Mentions
Finance	340+
Healthcare	232+
Education	181+
Cybersecurity	177+
Legal/Regulatory	157+
Manufacturing	57+
Retail/E-commerce	40+

Key Takeaways

- 69.4% of roles are AI-First - RAG, agents, LLMs
- 93.1% need skills beyond GenAI - it's a full-stack role
- RAG (35.9%) is the #1 pattern, automation (15.4%) is the #1 use case
- Evaluation (39.6%) is the emerging differentiator
- AI Engineers are builders first - build, deploy, monitor
- Python mandatory (82.5%), then AWS, Docker, K8s

BONUS

Learning Paths

learning-paths/

Core Learning Path

1. Foundation - LLMs, RAG and search
2. RAG use cases - FAQ assistants, document Q&A, different data sources
3. AI agents - function calling, tool use, agent frameworks
4. Testing - tests for agents, LLM-as-judge, cost tracking
5. Monitoring - logging, tracing, OpenTelemetry, Grafana
6. Evaluation - offline eval, synthetic data, prompt optimization
7. Production - deployment, guardrails, end-to-end project

Skills by Priority

Must have	High value
Python	Agent frameworks
Prompt engineering	TypeScript
RAG patterns	FastAPI
Cloud (AWS/Azure/GCP)	Kubernetes
Docker	CI/CD

Transition Guides

From	Difficulty	Key advantage
ML Engineer	Easiest	Replace model call with API call
Data Engineer	Smooth (3-4 mo)	Pipelines directly relevant to RAG
Backend Engineer	Smooth (2-3 mo)	Add AI on top of engineering
Data Scientist	Moderate	Evaluation is your superpower
Frontend Engineer	Longer path	Full-stack end-to-end advantage

github.com/.../learning-paths/

AI Engineering Field Guide

github.com/alexeygrigorev/ai-engineering-field-guide

⭐ Star the repo to keep an eye on updates

Newsletter: alexeyondata.substack.com