Senior Machine Learning Engineer — Generative AI / LLM Delivery (US)

Company: Amazon (AWS)

Role family: MLE | Seniority: Senior | Location: United States (multiple hubs)

Source (public posting): https://www.amazon.jobs/jobs/2921687

Cluster demand (ground-truth): {"MLE": 0.75, "SWE": 0.15, "DS": 0.1, "QD": 0.0, "QR": 0.0}

Reconstruction method: normalized sections + paraphrased responsibilities/qualifications for QA use.

ROLE SUMMARY
You will fine-tune and adapt large language models for real customer use cases and ensure they run efficiently in production environments. The role blends applied ML with pragmatic engineering delivery.

KEY RESPONSIBILITIES
• Fine-tune and evaluate modern LLMs for domain-specific tasks, including data prep and evaluation design.
• Optimize models for efficient inference (latency, throughput, and cost) on cloud accelerators.
• Build repeatable pipelines for training, evaluation, and deployment, including experiment tracking.
• Collaborate with customers and internal teams to translate problem statements into workable ML solutions.
• Document decisions and trade-offs; communicate results to technical and non-technical stakeholders.

REQUIRED QUALIFICATIONS
• 5+ years building software systems; strong Python expertise.
• Hands-on experience training and deploying deep learning models.
• Familiarity with prompt/finetuning workflows, evaluation, and safety considerations.

PREFERRED QUALIFICATIONS
• Experience serving models at scale and optimizing inference stacks.
• Experience with LLM toolchains and model observability.