

**Extract the main job role and top 5 required skills from the following job description:**

What You’ll Do

Collaborate closely with our research and engineering teams on one of the areas:

Library Development: Architect and enhance open-source Python tooling for alignment, explainability, uncertainty quantification, robustness, and machine unlearning.  
Model Benchmarking: Conduct rigorous evaluations of LLMs and deep networks under domain shifts, adversarial conditions, and regulatory constraints.  
Explainability & Trust: Design and implement XAI techniques (LRP, SHAP, Grad-CAM, Backtrace) across text, image, and tabular modalities.  
Mechanistic Interpretability: Probe internal model representations and circuits—using activation patching, feature visualization, and related methods—to diagnose failure modes and emergent behaviors.  
Uncertainty & Risk: Develop, implement, and benchmark uncertainty estimation methods (Bayesian approaches, ensembles, test-time augmentation) alongside robustness metrics for foundation models.  
Research Contributions: Author and maintain experiment code, run systematic studies, and co-author whitepapers or conference submissions.

General Required Qualifications

Strong Python expertise: writing clean, modular, and testable code.  
Theoretical foundations: deep understanding of machine learning and deep learning principles with hands-on experience with PyTorch.  
Transformer architectures & fundamentals: comprehensive knowledge of attention mechanisms, positional encodings, tokenization and training objectives in BERT, GPT, LLaMA, T5, MOE, Mamba, etc.  
Version control & CI/CD: Git workflows, packaging, documentation, and collaborative development practices.  
Collaborative mindset: excellent communication, peer code reviews, and agile teamwork.

Preferred Domain Expertise (Any one of these is good) :

Explainability: applied experience with XAI methods such as SHAP, LIME, IG, LRP, DL-Bactrace or Grad-CAM.  
Mechanistic interpretability: familiarity with circuit analysis, activation patching, and feature visualization for neural network introspection.  
Uncertainty estimation: hands-on with Bayesian techniques, ensembles, or test-time augmentation.  
Quantization & pruning: applying model compression to optimize size, latency, and memory footprint.  
LLM Alignment techniques: crafting and evaluating few-shot, zero-shot, and chain-of-thought prompts; experience with RLHF workflows, reward modeling, and human-in-the-loop fine-tuning.  
Post-training adaptation & fine-tuning: practical work with full-model fine-tuning and parameter-efficient methods (LoRA, adapters), instruction tuning, knowledge distillation, and domain-specialization.

Additional Experience (Nice-to-Have)

Publications: contributions to CVPR, ICLR, ICML, KDD, WWW, WACV, NeurIPS, ACL, NAACL, EMNLP, IJCAI or equivalent research experience.  
Open-source contributions: prior work on AI/ML libraries or tooling.  
Domain exposure: risk-sensitive applications in finance, healthcare, or similar fields.  
Performance optimization: familiarity with large-scale training infrastructures.

**Main Job Role:**  
Machine Learning Research Engineer (with a focus on model alignment, interpretability, robustness, and open-source tooling for foundation models/LLMs)

**Top 5 Required Skills:**

1. **Strong Python expertise:**
   * Writing clean, modular, and testable code, including experience with open-source development and tooling.
2. **Deep understanding of machine learning and deep learning principles:**
   * Especially hands-on experience with PyTorch and theoretical foundations of modern AI.
3. **Knowledge of transformer architectures:**
   * In-depth familiarity with BERT, GPT, LLaMA, T5, MOE, Mamba, and associated concepts (attention mechanisms, tokenization, training objectives).
4. **Version control & collaborative development:**
   * Proficiency with Git workflows, CI/CD, packaging, documentation, and working in collaborative/Agile teams.
5. **Domain expertise in at least one area (e.g., Explainability, Mechanistic Interpretability, Uncertainty Estimation, Model Compression, LLM Alignment, or Fine-tuning):**
   * Applied experience in at least one specialized subfield, such as XAI methods (SHAP, LIME, LRP, Grad-CAM), circuit analysis, Bayesian uncertainty estimation, quantization/pruning, alignment techniques (RLHF, prompting), or advanced fine-tuning methods.

**Bonus Skills (not in top 5 but valuable):**

* Research publications, open-source contributions, risk-sensitive domain experience, and performance optimization for large-scale ML systems.