RIFAT ERDEM SAHIN

LLM ENGINEER | LARGE LANGUAGE MODEL SPECIALIST

📍 **Location:** London, United Kingdom
🇬🇧 **Citizenship:** British
✉️ **Email:** contact@rifaterdemsahin.com
📞 **Phone:** +44 7848 024173
🔗 **LinkedIn:** linkedin.com/in/rifaterdemsahin
💻 **GitHub:** github.com/rifaterdemsahin
🌐 **Portfolio:** https://rifaterdemsahin.com
📅 **Schedule a Call:** https://calendly.com/rifaterdem/schedule

PROFESSIONAL SUMMARY

Senior LLM Engineer specializing in **Large Language Models, fine-tuning, and production LLM systems**. Deep expertise in designing and implementing enterprise-scale LLM applications with focus on model optimization, inference scaling, and AI safety. Proven track record building LLM solutions that delivered 300% productivity improvements and 30% cost reductions across financial services, healthcare, and real estate sectors. Expert in transformer architectures and responsible AI deployment.

CORE COMPETENCIES

🤖 **LLM Development & Fine-tuning** - Large Language Model fine-tuning with LoRA, QLoRA, and PEFT techniques - Custom model training for domain-specific applications - Model compression and quantization for production deployment - Prompt engineering and few-shot learning optimization

🔧 **LLM Infrastructure & Scaling** - Production LLM inference optimization and scaling - Model serving with TensorRT, vLLM, and TGI (Text Generation Inference) - GPU cluster

management and resource optimization - Cost optimization strategies for LLM inference at scale

---

KEY ACCOMPLISHMENTS

🏆 2025 | IBM | London, UK

Enterprise AI & Hybrid Cloud Transformation - **Challenge:** Lead enterprise-scale hybrid cloud transformation and AI integration for global clients - **Solution:** Architected comprehensive IBM Cloud + Red Hat OpenShift platform with watsonx AI integration and DevSecOps practices - **Impact:** - 35% reduction in operational overhead through AI-driven automation with watsonx - 40% improvement in deployment frequency via DevSecOps and Ansible Automation Platform - Zero-downtime migration of mission-critical workloads to hybrid cloud architecture - **Technologies:** IBM Cloud, Red Hat OpenShift, watsonx AI, Ansible Automation Platform, Terraform, Kubernetes, GitHub Actions

🏆 2024 | Goldman Sachs | Muscat, Oman

Enterprise LLM Platform for Financial Services - **Challenge:** Build scalable LLM platform for financial document analysis and automated compliance - **Solution:** Architected comprehensive LLM platform with fine-tuned models and automated inference scaling - **Impact:** - 300% improvement in document processing speed through optimized LLM inference - 30% reduction in operational costs via efficient model serving - 95% accuracy in financial document classification and analysis - **Technologies:** Transformers, PyTorch, vLLM, CUDA, Kubernetes, MLflow

🏆 2023 | Ypsomed | Switzerland

Medical LLM for Healthcare Documentation - **Challenge:** Develop specialized LLM for medical device documentation and patient interaction - **Solution:** Fine-tuned domain-specific LLM with medical knowledge base integration - **Impact:** - 40% improvement in documentation accuracy through specialized model training - 35% reduction in manual documentation effort via LLM automation - Real-time medical insights enabling improved patient care - **Technologies:** Hugging Face, LoRA, Medical datasets, Azure OpenAI, RAG systems

🏆 2022 | Cushman & Wakefield | London, UK

Real Estate LLM Analytics Platform - **Challenge:** Build LLM system for property market analysis and automated reporting - **Solution:** Implemented fine-tuned LLM with real estate domain knowledge and market data integration - **Impact:** - 50% faster market analysis through automated LLM-powered insights - 90% improvement in report generation speed via natural language processing - Enhanced business intelligence enabling strategic property investment decisions - **Technologies:** GPT models, Fine-tuning frameworks, Real estate datasets, Azure ML

TECHNICAL LLM EXPERTISE

LLM Frameworks & Tools

Models: GPT-4, Claude, LLaMA, Mistral, BERT, T5, Custom transformer architectures
Fine-tuning: LoRA, QLoRA, PEFT, Supervised Fine-Tuning (SFT), RLHF
Libraries: Transformers, PyTorch, TensorFlow, Hugging Face, LangChain, LlamaIndex
Serving: vLLM, TGI, TensorRT-LLM, Triton Inference Server, Ray Serve

LLM Operations & Optimization

Optimization: Model quantization, Knowledge distillation, Pruning, TensorRT optimization
Scaling: Distributed inference, Model parallelism, Pipeline parallelism, Load balancing
Monitoring: Model performance tracking, Inference latency optimization, Cost monitoring
Safety: Content filtering, Bias detection, Adversarial robustness, Alignment techniques

Programming & Infrastructure

Languages: Python, CUDA, C++, Bash, SQL
ML Infrastructure: Kubernetes, Docker, NVIDIA GPUs, MLOps pipelines
Cloud Platforms: Azure OpenAI, AWS Bedrock, GCP Vertex AI, RunPod, Lambda Labs
Data Processing: Pandas, NumPy, Apache Spark, Data preprocessing pipelines

---

LLM SPECIALIZATIONS

Custom Model Development

- Fine-tuned LLMs for financial, healthcare, and real estate domains

- Implemented RLHF (Reinforcement Learning from Human Feedback) pipelines
- Built custom tokenizers and preprocessing pipelines for specialized datasets
- Created model evaluation frameworks with domain-specific benchmarks

Production LLM Systems

- Designed high-throughput LLM inference systems serving 1M+ daily requests
- Implemented adaptive batching and caching strategies reducing latency by 60%
- Built cost optimization systems reducing LLM inference costs by 50%
- Created automated model A/B testing and deployment pipelines

LLM Safety & Alignment

- Implemented content filtering and safety guardrails for production LLM systems
- Built bias detection and mitigation frameworks for responsible AI deployment
- Created explainability tools for LLM decision making in regulated industries
- Designed red-teaming frameworks for LLM robustness testing

PROFESSIONAL EXPERIENCE HIGHLIGHTS

Senior LLM Engineer / AI Technical Lead | 2020 - Present
Goldman Sachs, Ypsomed, Cushman & Wakefield

- Led LLM initiatives across finance, healthcare, and real estate sectors
- Designed and implemented LLM systems supporting millions of daily interactions
- Established LLM best practices and production deployment frameworks
- Mentored engineering teams on transformer architectures and LLM optimization
- Delivered LLM solutions achieving 300% improvement in processing efficiency

AI/ML Engineer | 2016 - 2020
Microsoft, Emerson, Various Fortune 500

- Built machine learning solutions for digital transformation initiatives
- Led development of NLP and language model applications
- Established ML frameworks and deployment strategies
- Evangelized AI technologies through technical leadership and training

EDUCATION

🎓 **Bachelor of Science**
Southern New Hampshire University, USA 🇺🇸 | 2013

CERTIFICATIONS & CONTINUOUS LEARNING

📜 **AWS Certified Machine Learning Specialty**
📜 **Azure AI Engineer Associate**
📜 **NVIDIA Deep Learning Institute Certifications**
📜 **Certified Kubernetes Administrator (CKA)**

Continuous Learning:
- Active contributor to LLM and transformer model open-source projects - Regular participant in AI research conferences and LLM community events - Following latest developments in transformer architectures and training techniques - Experimenting with cutting-edge LLM optimization and serving technologies

THOUGHT LEADERSHIP & SPEAKING

- Technical blog on LLM engineering and optimization at rifaterdemsahin.com
- Internal tech talks on LLM fine-tuning and production deployment
- Mentorship in AI/ML communities and LLM engineering teams
- LLM engineering examples and tutorials shared via GitHub

SECURITY CLEARANCES

🔒 **UK SC (Security Check)** - Valid until 2028
🔒 **NATO Clearance** - Valid until 2029
✓ **Background Checks:** Watchdog (2024), Sterling (2019)

AVAILABILITY & CONTACT

Immediate Availability for LLM engineering and AI development roles

📅 **Schedule a Discussion:** https://calendly.com/rifaterdem/schedule
📧 **Email:** contact@rifaterdemsahin.com
📞 **Phone:** +44 7848 024173

SUPPORTING DOCUMENTS

📊 **Technical Portfolio & Presentations:**
https://rifaterdemsahin.com/wp-content/uploads/2025/02/rifaterdemsahinprofilepresentation.v2025.2.pdf

---

References and detailed project portfolios available upon request

---

KEY ACCOMPLISHMENTS

🏆 2024 | Goldman Sachs | Muscat, Oman

Enterprise Llm Engineer Platform - **Challenge:** Implement scalable llm engineer solutions for financial services environment - **Solution:** Architected comprehensive platform with automation and monitoring capabilities - **Impact:** - 300% improvement in operational efficiency through advanced automation - 30% reduction in operational costs via intelligent optimization - 99.9% system availability with automated failover and recovery - **Technologies:** Kubernetes, Terraform, Azure/AWS, Monitoring tools, Automation frameworks

🏆 2023 | Ypsomed | Switzerland

Healthcare Technology Platform - **Challenge:** Scale llm engineer capabilities for medical device and IoT environments - **Solution:** Implemented robust platform with real-time processing and analytics - **Impact:** - 40% improvement in system performance through optimized architecture - 35% reduction in operational overhead via automation - Real-time

insights enabling improved healthcare delivery - **Technologies:** Cloud platforms, Monitoring systems, Automation tools, Analytics platforms

🏆 2022 | Cushman & Wakefield | London, UK

Real Estate Technology Platform - **Challenge:** Modernize llm engineer infrastructure for property management analytics - **Solution:** Built cloud-native platform with automated processes and real-time analytics - **Impact:** - 50% faster processing through optimized platform architecture - 90% improvement in system reliability via automated monitoring - Enhanced business intelligence enabling strategic decision making - **Technologies:** Azure services, Analytics platforms, Automation frameworks, Monitoring tools

---

TECHNICAL EXPERTISE

Core Technologies

Platforms: Kubernetes, Docker, AWS, Azure, GCP
Automation: Terraform, Ansible, GitLab CI/CD, GitHub Actions
Monitoring: Prometheus, Grafana, ELK Stack, Azure Monitor
Programming: Python, Bash, PowerShell, YAML, JSON

Specialized Skills

Infrastructure: Infrastructure as Code, Container orchestration, Cloud architecture
Security: Zero-trust architectures, Compliance frameworks, Identity management
Operations: CI/CD pipelines, Monitoring and alerting, Incident response
Integration: API design, Microservices, Event-driven architectures

---

PROFESSIONAL EXPERIENCE HIGHLIGHTS

Senior LLM Engineer / AI Solutions Architect | January 2025 - Present
IBM | London, UK

- Architecting hybrid cloud transformation using IBM Cloud and Red Hat OpenShift for Fortune 500 enterprises

- Implementing watsonx AI solutions for intelligent automation, reducing operational overhead by 35%
- Leading DevSecOps transformation with Ansible Automation Platform and Terraform IaC across multi-cloud environments
- Delivering zero-downtime Kubernetes cluster migrations for mission-critical financial and government workloads
- Building enterprise CI/CD pipelines with GitHub Actions and Jenkins serving 500+ developers globally

Senior Llm Engineer / Technical Lead | 2020 - 2025
Goldman Sachs, Ypsomed, Cushman & Wakefield

- Led llm engineer initiatives across finance, healthcare, and real estate sectors
- Designed and implemented solutions supporting millions of daily transactions
- Established best practices and automated frameworks
- Mentored engineering teams on modern technologies and practices
- Delivered solutions achieving 300% improvement in operational efficiency

Llm Engineer | 2016 - 2020
Microsoft, Emerson, Various Fortune 500

- Built enterprise solutions for digital transformation initiatives
- Led cross-functional teams designing scalable applications
- Established frameworks and implementation strategies
- Evangelized modern technologies through technical leadership

---

EDUCATION

🎓 **Bachelor of Science**
Southern New Hampshire University, USA 🇺🇸 | 2013

CERTIFICATIONS & CONTINUOUS LEARNING

📜 **Microsoft Certified Solutions Architect Expert**
📜 **AWS Certified Solutions Architect Professional**

📜 **Azure Solutions Architect Expert**
📜 **Certified Kubernetes Administrator (CKA)**

Continuous Learning:
- Active contributor to open-source projects - Regular participant in technical communities - Following latest developments in technology - Experimenting with emerging tools and practices

---

SECURITY CLEARANCES

🔒 **UK SC (Security Check)** - Valid until 2028
🔒 **NATO Clearance** - Valid until 2029
✓ **Background Checks:** Watchdog (2024), Sterling (2019)

AVAILABILITY & CONTACT

Immediate Availability for llm engineer roles

📅 **Schedule a Discussion:** https://calendly.com/rifaterdem/schedule
📧 **Email:** contact@rifaterdemsahin.com
📞 **Phone:** +44 7848 024173

SUPPORTING DOCUMENTS

📊 **Technical Portfolio & Presentations:**
https://rifaterdemsahin.com/wp-content/uploads/2025/02/rifaterdemsahinprofilepresentation.v2025.2.pdf

---

References and detailed project portfolios available upon request