

Ilia V

Senior Senior AI/ML Engineer

SUMMARY

- Senior AI/ML Engineer with 6+ years in architecting enterprise AI platforms, multi-agent systems, and GenAI workflows, delivering solutions with 94%+ task accuracy and serving 10M+ monthly inferences.
- Expertise in Python, PyTorch, Azure AI ecosystem, LLM/SLM fine-tuning (Phi-3/4, GPT-4), and optimization (vLLM, TensorRT-LLM) reducing compute costs by 45%.
- Proven track record in building scalable MLOps pipelines, real-time AI safety guardrails, and automated evaluation frameworks ensuring SOC2/GDPR compliance.
- Strong background in data engineering with Apache Spark, Kafka, Delta Lake, and hybrid search systems integrating vector databases and semantic ranking for enhanced retrieval precision.

TECHNICAL SKILLS

Main Technical Skills	Python (6 yr.), PyTorch (6 yr.), Azure ML (3 yr.), Kubernetes (3 yr.), Docker (6 yr.)
AI Tools & Assistants	Claude
Programming Languages	C#, Go, JavaScript, Objective-C, Python (6 yr.)
Java Frameworks	Apache Spark (3 yr.)
Scala Frameworks	Apache Spark (3 yr.)
AI & Machine Learning	Azure AI Search (2 yr.), Azure AI Studio, Claude, CrewAI, Flowise (2 yr.), GPT (2 yr.), Hugging Face, Kubeflow, LangChain (2 yr.), LangGraph (2 yr.), LLaMA (2 yr.), Mlflow, OpenCV, Prompt Engineering, PyTorch (6 yr.), Scikit-learn, Spacy (3 yr.), TensorFlow (2 yr.), Triton (2 yr.), vLLM (2 yr.), Xgboost
Python Frameworks	FastAPI (2 yr.)
.NET Platform	Identity Server (2 yr.)
Python Libraries and Tools	NLTK, PyTorch (6 yr.), Scikit-learn, TensorFlow (2 yr.)
Data Analysis and Visualization Technologies	Apache Airflow (3 yr.), Apache Spark (3 yr.), Azure Data Factory, Data visualization, Kibana
Security	Content Security Policy, SOC 2 (4 yr.)
Databases & Management Systems / ORM	Apache Spark (3 yr.), Cosmos DB, dbt (2 yr.), Elasticsearch (3 yr.), ELK stack (Elasticsearch, Logstash, Kibana), Neo4j (2 yr.), OpenSearch, PostgreSQL (4 yr.), Snowflake (2 yr.), SQL

Cloud Platforms, Services & Computing	Azure ML (3 yr.)
Amazon Web Services	AWS LightSail (2 yr.)
Azure Cloud Services	Azure AI Search (2 yr.), Azure AI Studio, Azure Application Insights, Azure Databricks (3 yr.), Azure Data Factory, Azure DevOps (3 yr.), Azure ML (3 yr.), Cosmos DB
Industry Domain Experience	HIPAA (4 yr.)
QA, Test Automation, Security	A/B Testing, CSP
Methodologies, Paradigms and Patterns	Agile, GDPR (4 yr.), microservices, Scrum
Tools for Big Data	Apache Iceberg (2 yr.)
Deployment, CI/CD & Administration	ArgoCD (3 yr.), GitLab CI (2 yr.), Helm (3 yr.), Jenkins
Scripting and Command Line Interfaces	Bash
Platforms	CUDA (3 yr.)
Virtualization, Containers and Orchestration	Docker (6 yr.), Kubernetes (3 yr.), Terraform
SDK / API and Integrations	FastAPI (2 yr.), GraphQL (2 yr.), RESTful API (3 yr.)
Version Control	Git (4 yr.)
Logging and Monitoring	Grafana (3 yr.), Logstash, Prometheus (3 yr.)
Mail / Network Protocols / Data transfer	GRPC
Message/Queue/Task Brokers	Kafka (3 yr.)
Hardware	PHY (2 yr.)
Other Technical Skills	AutoGen (2 yr.), Bicep (2 yr.), Delta lake (3 yr.), EasyOCR, Fine-Tuning SonarQube (4 yr.), LoRa, MLOps (4 yr.), Model Context Protocol (2 yr.), PaddleOCR, Qdrant, QLoRA (1 yr.), Ragas (2 yr.), Semantic Kernel (2 yr.), Tesseract OCR

WORK EXPERIENCE

Senior AI/ML Engineer - Microsoft (Next-generation enterprise AI platforms and autonomous agentic workflows on Azure)

Duration: 08.2025 - present



Summary:

- Architected next-generation enterprise AI platforms and autonomous agentic workflows on Azure, including custom Microsoft Copilot integrations, reasoning pipelines, and automated AI evaluation frameworks for global accounts
- Developed enterprise-grade AI solutions integrating CRM/ERP data streams into conversational interfaces for large-scale user bases

Responsibilities:

- Architected and developed custom Microsoft 365 Copilot Extensions and MCP plugins using Semantic Kernel and Model Context Protocol.
- Integrated enterprise CRM/ERP data streams into daily conversational interfaces for 10K+ active daily users.
- Designed autonomous multi-agent systems and reasoning pipelines with AutoGen and LangGraph enabling dynamic tool calling and automated problem resolution.
- Engineered domain-specific Small Language Model pipelines by fine-tuning Phi-3/4 models via QLoRA on Azure ML.
- Implemented enterprise Red Teaming and real-time AI Safety guardrails using Azure AI Content Safety and Promptflow for compliance and security.

Technologies: Python, PyTorch, Azure OpenAI (GPT-4o), Phi-3/4, Llama 3, Semantic Kernel, AutoGen, LangGraph, Model Context Protocol (MCP), Azure Databricks, Delta Lake, dbt, PostgreSQL (pgvector), Neo4j, Azure AI Search, Azure ML, vLLM, AKS, Docker, Promptflow, Ragas, Azure DevOps, Bicep, SOC2/HIPAA/GDPR

Senior AI/ML Engineer - EPAM (Generative AI systems and agentic workflows for enterprise clients)

Duration: 06.2023 - 07.2025

Summary:

- Designed and deployed cutting-edge Generative AI systems specializing in high-performance LLM applications, RAG architectures, and AI safety for enterprise clients
- Delivered scalable document ingestion and indexing pipelines supporting hybrid search with strict metadata and access control

Responsibilities:

- Designed and implemented GenAI and agentic workflows using LangChain and LangGraph for multi-step document analysis and technical support automation.
- Built scalable document ingestion and indexing pipelines supporting hybrid keyword and vector search with metadata, versioning, and ACL enforcement.
- Improved data quality and retrieval precision by introducing validation re-ranking and provenance tracking.
- Optimized end-to-end data serving latency and throughput achieving sub-1.2s P99 latency for over 500K daily requests.
- Operated Kubernetes-based data services with monitoring, alerting, and access control including RBAC and PII handling.

Technologies: Python, PyTorch, FastAPI, Llama 2/3, GPT-4, LangChain, vLLM, TensorRT-LLM, Hugging Face Transformers/TRL, Docker, Kubernetes (EKS/GKE), ArgoCD, Helm, Snowflake, Snowflake Cortex, Apache Iceberg, Apache Spark, PostgreSQL, MongoDB, Weights & Biases, Evidently AI, Vertex AI, AWS SageMaker, GCP, Airflow, Vector DBs, Prometheus, Grafana, Git, SOC2/HIPAA/GDPR



Senior ML Engineer (NLP) - Luxoft (AI-powered contract analysis and multimodal RAG system)

Duration: 07.2021 - 06.2023

Summary:

- Developed AI-powered contract analysis system and optimized multimodal retrieval-augmented generation (RAG) systems integrating vehicle manuals to improve search efficiency
- Implemented NVIDIA Triton inference server with dynamic batching to reduce latency for high-volume queries

Responsibilities:

- Implemented NVIDIA Triton inference server with dynamic batching for low-latency inference.
- Developed AI-powered contract analysis system using DeBERTa and spaCy for clause extraction.
- Built automated MLOps pipelines enabling weekly model updates across multiple client deployments.
- Optimized multimodal RAG system integrating vehicle manuals with GPT-3.5 for technician search.
- Built feature stores and batch/streaming pipelines using Kafka and Airflow for consistent training data.

Technologies: Python, PyTorch, BERT/DeBERTa, GPT-3.5, spaCy, NVIDIA Triton, CUDA, Docker, Kubernetes, Azure ML, Kafka, Airflow, Delta Lake, Databricks, PostgreSQL, Elasticsearch, Java, Oracle, FinBERT

ML Engineer/Data Engineer - Lionbridge (Multilingual NER pipelines and distributed text preprocessing framework)

Duration: 02.2020 - 07.2021

Summary:

- Built multilingual named entity recognition (NER) pipelines for legal document processing and designed distributed text preprocessing frameworks to reduce annotation costs
- Developed custom quality control models to flag translation errors and automated data ingestion from various sources

Responsibilities:

- Built multilingual NER pipelines supporting 15+ languages including low-resource dialects.
- Designed distributed text preprocessing framework to reduce annotation costs for AI training datasets.
- Developed BERT-based quality control models to flag translation errors.
- Automated data ingestion from REST APIs and SaaS platforms using Airflow and custom Python services.

Technologies: Python, PyTorch, TensorFlow, BERT, XLM-R, spaCy, Hugging Face, Apache Spark, AWS SageMaker, AWS Glue, AWS Redshift, Docker, Kubernetes, Airflow, PostgreSQL, Elasticsearch, MLflow, REST API, GitLab CI/CD



EDUCATION

- **Vistula University**

Master's degree of Computer Science, GPA: 4.85
2019-2024

