Anton V.
Senior Senior AI Engineer

SUMMARY

- Senior AI/ML engineer with hands-on experience in GenAI platforms, LLM orchestration, and RAG pipelines;
- Skilled in Python and C# for backend services, chatbot architecture, and AI integration pipelines;
- Built and deployed RAG systems using Azure OpenAI, AI Search, and LangChain;

- Developed Whisper-based transcription pipelines with speaker diarization for business meeting summarization;
- Integrated OpenAI models and multimodal embeddings into workflow automation tools like Jira and Trello;
- Experience in distributed processing with Celery, RabbitMQ, and Redis for asynchronous task execution;
- Deployed scalable AI services in Azure and AWS environments, using Kubernetes, Helm, and Azure DevOps;
- Built real-time event processing and notification systems using Kafka and Apache Camel;
-Designed APIs and backend services for customer support chatbots, voice platforms, and CRM integration;
- Applied prompt engineering, testing (xUnit, pytest), and CI/CD automation across multiple AI initiatives.

TECHNICAL SKILLS

Main Technical Skills	Python, OpenAI, RAG, Azure ML
Programming Languages	C#, C++, Python
AI & Machine Learning	AI, Amazon Machine learning services, AWS ML (Amazon Machine learning services), AWS SageMaker, AWS SageMaker (Amazon SageMaker), Azure AI Vision, Azure OpenAI, FCN, GenAI, GPT, LangChain, LSTM, Nvidia DeepStream SDK, Nvidia Holoscan SDK, ONYX, OpenAI, OpenCV, Pinecone, PyTorch, RAG, TensorFlow, ViT, VITS, Whisper, YOLO NAS
Scala Libraries and Tools	Akka Streams
UI Frameworks, Libraries, and Browsers	Chrome Extensions, Gstreamer

JavaScript Frameworks	Express, Node.js, React
Python Frameworks	Flask
Python Libraries and Tools	PyTorch, TensorFlow
JavaScript Libraries and Tools	Redux, Redux-Saga
Salesforce Ecosystem	Salesforce
Java Frameworks	Spring
Java Libraries and Tools	Spring Security
Data Analysis and Visualization Technologies	Business Intelligence (BI) Tools, Databricks, ML
Databases & Management Systems / ORM	ChromaDB, Cosmos DB, ORM, Redis
Cloud Platforms, Services & Computing	AWS, Azure ML, Nvidia AGX, Nvidia IGX
Amazon Web Services	AWS EC2, AWS Lambda, AWS ML (Amazon Machine learning services), AWS S3, AWS SageMaker, AWS SageMaker (Amazon SageMaker)
Azure Cloud Services	Azure Blob Storage, Azure Functions, Cosmos DB, Databricks, Microsoft Azure API
Collaboration, Task & Issue Tracking	Atlassian Trello, Jira
Message/Queue/ Task Brokers	Celery, Kafka, RabbitMQ
Deployment, CI/CD & Administration	CI/CD, Jenkins
Virtualization, Containers and Orchestration	Docker, Kubernetes
Mail / Network Protocols / Data transfer	JWT
SDK / API and Integrations	JWT, Microsoft Azure API, RESTful API
Platforms	Nvidia, Salesforce
Other Technical Skills	Calypso, Payanote

WORK EXPERIENCE

Lead AI/Machine Learning Engineer, Railway Infrastructure Maintenance Application

Duration: September 2024 - Present

Summary: Developed an advanced application for railway infrastructure maintenance engineers to perform remote inspections of Overhead Line Equipment (OLE). The solution includes a comprehensive video library and player, video analytics, and railway network map visualization, aimed at reducing operator downtime, maintenance costs, and on-track engineer time.

Responsibilities:

- Model Development & Fine-Tuning:
  - Fine-tuned YOLO NAS computer vision models to accurately detect critical objects such as wires, pantographs, and arcing.
  - Enhanced model performance to operate effectively at high-speed conditions (200-280 km/h) using cameras with at least 120 fps
- Deployment & Integration: Deployed optimized models into the ONYX framework for seamless integration. Transitioned from Nvidia Holoscan to Nvidia DeepStream SDK to achieve real-time processing and improved inference speeds.
- Analytic Engine Development:
  - POC Phase: Set up an initial analytic pipeline using Nvidia Holoscan SDK to validate core functionalities.
- Advanced Pipeline Implementation:
  - Developed a robust analytic pipeline with Nvidia DeepStream SDK and GStreamer. Implemented the pipeline in C++, incorporating custom elements for filtering and object measurement.
  - Created a camera calibration plugin using OpenCV to ensure accurate distance estimation. Developed algorithms for deep estimation to measure object distance and size, enhancing defect detection accuracy.
- Challenges & Solutions:
  - High-Speed Processing: Ensured real-time object detection and defect identification by optimizing camera frame rates and deploying efficient inference pipelines.
  - Performance Optimization: Overcame performance limitations of Nvidia Holoscan by migrating to Nvidia DeepStream, resulting in significant improvements in processing speed and reliability.
  - Accurate Distance Estimation: Developed custom plugins and algorithms to accurately measure distances and sizes of detected objects, facilitating precise maintenance actions

Technologies: Python, C++, PyTorch, OpenCV, YOLO NAS, ONYX, Nvidia Holoscan SDK, Nvidia DeepStream SDK, GStreamer, Databricks, RabbitMQ, Nvidia IGX, Nvidia AGX.

Lead AI/Machine Learning Engineer, NDA

Duration: August 2024 - September 2024

Summary: Led the design and implementation of GenAI-based solutions to automate document flows, support regulatory compliance, and optimize SDLC processes. Focused on using LLMs, multimodal embeddings, and Azure AI services to increase productivity and reduce operational bottlenecks.

Lead Engineer, GenAI Process Optimization for SDLC

Summary: Integrated GenAI into the software development lifecycle to ensure alignment with evolving industry regulations and reduce compliance overhead.

Responsibilities:

- Developed a chatbot for daily document-related inquiries, eliminating manual lookups and document exchanges;
- Integrated Azure Document Intelligence to extract structured data from technical documentation;
- Built RAG-based retrieval and prompt pipelines to support contextual LLM output;
- Collaborated with Azure platform teams to track SDK updates and integrate new AI service features.

Outcomes:

- Improved regulatory compliance workflows;
- Reduced compliance maintenance costs by ~10%.

Challenges & Solutions:

- **LLM output variability:** Solved with detailed prompt templates;
- **Azure token limitations:** Addressed via parallel request batching and load balancing.

Technologies: Python, C#, LangChain, Azure OpenAI, Azure Document Intelligence, Azure Functions, Flask, Redis, Celery, RAG pipelines.

AI Engineer, GenAI Document Generator

Summary: Built a document generation pipeline that compiles new documentation using existing internal content, leveraging large language models and multimodal embeddings.

Responsibilities:

- Used Azure AI Vision and Azure Blob Storage for image-text embedding and document input management;
- Generated structured regulatory and release documents using LLMs;
- Designed a parallel processing infrastructure to optimize generation times.

Outcomes:

- Increased document generation speed by 20–30%;
- Reduced release cycle timelines through automated documentation.

Challenges & Solutions:

- **Multimodal SDK instability:** Handled through dynamic logic and fallback mechanisms;

- **Prompt consistency:** Achieved through reusable template libraries.

Technologies: Python, Azure OpenAI, Azure AI Vision, Azure Blob Storage, Flask, ChromaDB, Celery, Redis, CI/CD.

AI Lead, Asistme AI

Duration: January 2024 - August 2024

Summary: Led the development of a voice-to-action AI platform using LLMs and Whisper to improve meeting productivity and streamline business operations. Focused on multi-language transcription, summarization, search, and workflow integration.

AI Lead, AI Meeting Assistant Platform

Summary: Built a Chrome Extension–based platform to capture, transcribe, and summarize business meetings using OpenAI Whisper and LLMs.

Responsibilities:

- Led AI team and orchestrated platform development from architecture to delivery;
- Implemented Whisper + Payanote stack for accurate transcription with speaker diarization;
- Applied LLMs to summarize meetings and generate action items and insights;
- Built search functionality for fast retrieval of past meeting insights;
- Integrated system outputs into Jira, Trello, and HR tools for automated workflow updates.

Outcomes:

- Improved project workflow automation and stakeholder visibility;
- Enabled real-time conversation analysis in call centers;
- Reduced HR review times through automated feedback summarization.

Technologies: Whisper, Payanote, OpenAI LLMs, Chrome Extension, Python, Jira, Trello, HR tool integrations.

ML Engineer, Impressit.io

Duration: June 2023 - August 2024

Summary: Delivered AI and ML components for the Gnetwork project, focused on improving customer interaction, automation, and scalable AI infrastructure through NLP, data engineering, and API design.

ML Engineer, Gnetwork AI Platform

Summary: Developed AI-powered chatbot and backend infrastructure, enhancing user engagement and automating support workflows.

Responsibilities:

- Designed and deployed a fine-tuned OpenAI-based customer service chatbot;

- Built NLP analysis tools using LangChain pipelines for better user input processing;
- Managed vector-based semantic search with Pinecone for real-time context retrieval;
- Created Node.js REST APIs and integrated Salesforce CRM and Cosmos DB;
- Built the custom Calypso Framework and ORM tool for Node.js to optimize backend development.

Outcomes:

- Increased customer satisfaction through intelligent query resolution;
- Streamlined CRM workflows and improved backend maintainability.

Technologies: OpenAI, LangChain, Pinecone, Node.js, REST API, Cosmos DB, Salesforce CRM, Calypso Framework, ORM tool.

ML Engineer, Infopulse

Duration: May 2023 - June 2023

Summary: Short-term engagement on a confidential media analysis initiative using large-scale ML infrastructure, speaker embedding, and disinformation detection. Worked with audio parsing, cloud pipelines, and vector search to enable scalable and intelligent media processing.

ML Engineer, Media Monitoring Platform (NDA)

Summary: Developed components for large-scale speech and media analysis with Whisper and OpenAI models.

Responsibilities:

- Parsed and fine-tuned 600,000+ video files using Whisper for improved speaker embedding;
- Managed Azure ML infrastructure with 50 virtual machines for distributed audio processing;
- Stored vectorized embeddings in Pinecone DB and supported Cosmos DB integration for structured metadata;
- Designed GPT-based prompt logic for advanced text markup and classification;
- Built a custom model for disinformation detection in news media.

Technologies: Whisper, Azure ML, Pinecone, Cosmos DB, OpenAI (GPT), Python.

ML Engineer / Data Scientist, Infopulse – JERA Project

Duration: January 2021 - May 2023

Summary: Designed and deployed predictive and computer vision models for a major industrial client. Worked with time-series data and sensor monitoring to improve reliability and forecasting accuracy.

Data Scientist, Industrial AI Models for JERA

Summary: Built time-series forecasting and defect detection models using deep learning on AWS SageMaker.

Responsibilities:

- Developed LSTM and FCN-based models for predictive analysis on sensor data;
- Built a ViT-based computer vision model for real-time defect detection;
- Used SageMaker for scalable training and deployment;
- Applied forecasting and image classification architectures for safety-critical use cases.

Technologies: Python, PyTorch, TensorFlow, LSTM, FCN, ViT, AWS SageMaker.

Software Developer, Sixt

Duration: October 2019 - January 2021

Summary: Full-stack development of microservices and web applications in the car-sharing and mobility sector. Built backend services, security layers, and responsive PWA applications.

Software Developer, Mobility Platform Services

Summary: Developed secure, scalable microservices for car rental and sharing workflows.

Responsibilities:

- Built REST APIs using Java Spring and Node.js Express;
- Integrated Kafka and Salesforce for event and customer flow management;
- Configured API security with Spring Security and JWT;
- Developed React-based PWA with Redux and modern ES8 syntax;
- Applied TDD and BDD practices with Jenkins CI/CD pipeline;
- Deployed and maintained systems on AWS (S3, EC2, Lambda) and Kubernetes.

Technologies: Java Spring, Node.js, Kafka, Docker, Kubernetes, React, Redux, Spring Security, JWT, Jenkins, AWS (S3, EC2, Lambda).

EDUCATION

Master, National Maritime University

Faculty: Transport technologies and systems

September 2014 - December 2020