

Konstantinos Giantsios

Machine Learning Engineer

📞 (+30) 698-22-33-611 | ✉️ constantinos.giantsios@gmail.com

🌐 [Konstantinos Giantsios](#) | 🌐 [CoGian](#)

SUMMARY

Machine Learning Engineer with 5+ years of expertise specialising in NLP, Recommendation Solutions, Search, and Agentic AI. Proven track record in designing, deploying, and scaling production-grade machine learning models, complex AI architectures, and real-time user profiling platforms.

EDUCATION

Aristotle University of Thessaloniki (AUTH), Thessaloniki, GR 09/2020 - 06/2022

B.Sc. in Data and Web Science, GPA: 9.5/10.0

Thesis: Interpretable Multi-Label Learning in Biomedical Texts, Grade: 9.5/10.0

University of Macedonia (UoM), Thessaloniki, GR 09/2016 - 09/2020

B.Sc. in Applied Informatics, GPA: 8.4/10.0 (top 9%)

Thesis: Detecting toxic comments and minimising unintentional prejudice using neural networks, Grade: 10.0/10.0

PROFESSIONAL EXPERIENCE

Kaizen Gaming, Machine Learning Engineer, Thessaloniki, GR 05/2025 - Present

- Data Pipelines & Agentic CX Chatbot: Engineered robust data pipelines employing Apache Spark and Databricks on Microsoft Azure, processing large-scale interaction data to power an agentic AI Chatbot for customer experience analysis.
- Multi-Agent Copilot Architecture: **Co-designed and implemented a multi-agent sports betting Copilot** using Python, LangChain, and LangGraph, delivering an interactive assistant capable of complex reasoning and resolving nuanced user queries.
- API Deployment & Serving: **Deployed complex agentic AI solutions** as scalable RESTful API services utilising FastAPI, Docker, and Redis, ensuring high availability and low-latency production responses.
- LLM Evaluation & Monitoring: **Established comprehensive evaluation frameworks and dashboards** leveraging LangSmith and RAGAS, enabling continuous monitoring, rapid debugging, and improved reliability of LLM applications.
- Tech stack: Python, Apache Spark, Hugging Face, Neo4j, Docker, FastAPI, Microsoft Azure, Redis, Databricks, LangChain, LangGraph, LangSmith

Independent Consultant, Machine Learning/Data Engineer, Thessaloniki, GR 04/2023 - 09/2023

- Independent project work completed part-time during tenure at Atypon Systems.
- Hybrid Search & RAG Solutions: **Formulated a specialised RAG architecture using Elasticsearch**, enabling highly accurate, hybrid search of influencer profiles for targeted brand campaigns.
- Automated Prompt Engineering: Created a scalable pipeline leveraging OpenAI APIs and Hugging Face models, streamlining the workflow for generating domain-specific educational materials.
- Computer Vision Processing: Designed an automated, end-to-end system for generating professional headshots, harnessing PyTorch and OpenCV, efficiently deployed via FastAPI and Docker.
- Tech stack: Python, PyTorch, Apache Spark, Hugging Face, OpenAI, Elasticsearch, Docker, FastAPI, OpenCV, Pinecone

Atypon Systems, Machine Learning Engineer, Thessaloniki, GR 06/2020 - 05/2025

- Recommendations, Semantic Search & End-to-End RAG: Architected a hybrid retrieval system (Elasticsearch) and integrated LLMs (Gemini, Llama) via LangChain and Self-RAG. **Fine-tuned bi-encoders (MiniLM), boosting recommended article CTR by 100% and search CTR by 20%.**
- Automated Text Tagging: Developed a comprehensive taxonomy and automated tagging system using deep learning Transformers, leading to a **30% increase in Success Search Rate (SSR)** and **reducing manual tagging time by over 90%.**
- Model Inference & MLOps: Orchestrated high-performance model serving utilising **NVIDIA Triton Inference Server and optimised document processing pipelines with ONNX** to ensure scalable, low-latency production inference.

- Real-Time Profiling: Implemented a real-time user profiling system from streaming event data to **generate personalised promotions, driving a 300% increase in new user registrations and a 200% increase in authenticated sessions.**
- Automated Evaluation: **Built testing pipelines to evaluate embedding retrieval** (SciRepEval, BEIR) and continuously measure LLM faithfulness and answer relevancy using the RAGAS framework and LLM-as-a-judge techniques.
- Tech stack: Python, Java, PyTorch, TensorFlow 2.0, Apache Spark, Apache Beam, Hugging Face, PostgreSQL, MongoDB, MySQL, Elasticsearch, MLflow, ONNX, Docker, Flask, FastAPI, Apache JMeter, Google Cloud Platform (GCP), NVIDIA Triton Inference Server, Redis, PubSub, Dataflow, LangChain, LangGraph, LlamaIndex

SKILLS

Programming Languages: Python, Java, Scala

Tools/Frameworks: TensorFlow 2.0, PyTorch, Apache Spark, Apache Beam, scikit-learn, Hugging Face, spaCy, MLflow, ONNX, OpenCV, NVIDIA Triton Inference Server, PubSub, Dataflow, LangChain, LangGraph, LlamaIndex, LangSmith, Langfuse

MLOps & Engineering: REST, Microservices, Git, CI/CD, Maven, Docker/Compose, Apache JMeter, Flask, FastAPI

Cloud & Data: Google Cloud Platform (GCP), Microsoft Azure, Elasticsearch, Neo4j, Redis, PostgreSQL, Pinecone (Vector DB), MongoDB, Databricks

AI & ML: Classical ML (Regression, Clustering, SVMs), Deep Learning (LLMs, Transformers, CNNs), Dimensionality Reduction, Prompt Engineering, Agentic AI

ACHIEVEMENTS & CERTIFICATES

*Google Hash Code: **Achieved 2nd place among 12 University of Macedonia teams and ranked in the top 20% overall in Greece.***

*Kaggle Competition: **Ranked in the top 6% worldwide in the 'Jigsaw Unintended Bias in Toxicity Classification' competition.***

Agentic AI MOOC (Legendary Tier) – UC Berkeley Center for Responsible, Decentralized Intelligence (Fall 2025).

Machine Learning in Production – DeepLearning.AI via Coursera (May 2023).

Deep Learning Specialization (5 Courses) – DeepLearning.AI via Coursera (Oct 2020).

Machine Learning – Stanford University via Coursera (Nov 2019).

LANGUAGES

Greek (Native), English (C2/Proficient)