Krystian Safjan

Data Scientist · Technical Leader · Researcher

✉ ksafjan@gmail.com ⌂ https://safjan.com izikeros 🎓 Krystian Safjan

Summary

Senior AI/ML Engineer specializing in LLMs and RAG applications with 8+ years of experience in AI and 18 of professional experience. Proven expertise in building secure, enterprise-scale AI systems serving 100,000+ users. Skilled in driving technical innovation, mentoring teams, and establishing MLOps practices. Experienced in developing production-grade AI solutions, leading cross-functional teams, and driving technical innovation in cutting-edge AI technologies. Accomplished researcher with publications, patents, and open-source contributions.

Experience

AI Lead & Data Scientist

Ernst & Young — Nov 2019 - Present

- Led the development and deployment of an enterprise-scale AI chatbot assistance system serving 100,000+ global users, managing a cross-functional team of seven senior engineers.
- Designed and implemented hybrid search architecture combining MongoDB BM25 and vector search, with custom ranking algorithms and multilingual support.
- Created comprehensive evaluation framework including and WebApp for gamified feedback collection system - used for evaluation of prompt engineering and RAG optimization techniques.
- Knowledge base document chunking optimization reducing token usage per query by 37%
- Fine-tuned SLMs on domain-specific data to improve response accuracy from 10% to 67% and reduce hallucination rates. Optimized fine-tuning hyperparameters, customizing loss functions. Authored method for synthetic dataset generation.
- Successfully prepared system for production, implementing security measures and passing InfoSec/DR audits while coordinating multiple stakeholders.
- Led development of document intelligence services, applying NLP and Computer Vision for document classification and understanding. Using Poly-cloud - deliberate selection of specific cloud providers for specialized tasks (e.g. Speech services from Google Cloud).
- Created document understanding applications with OpenAI LLMs via LangChain for information retrieval.
- Provided technical advisory for the team and initiated technical discussions.

Data Scientist & Technical Leader

Nokia — Feb 2019 - Nov 2021

- Directed development of an advanced log classification system for 4G/5G Base Stations, employing sentence clustering techniques to clean inconsistent data and reduce noise. Classifier was used in production in recommender system for routing bug report to one of the 2000+ teams.
- Established robust MLOps practices through comprehensive model lifecycle management, leveraging tools like Jupytext, GitLab, and MLflow for version control, collaboration, and monitoring.
- Architected and deployed neural networks featuring LSTM and word embeddings using TensorFlow and FastText, implementing data drift detection and mitigation protocols.
- Led and mentored a five-person team using OKR methodology, setting strategic research directions while fostering growth of junior team members and students.
- Drove technical excellence through rigorous validation of methods and results, ensuring high-confidence solutions through systematic questioning and thorough analysis.

System AI Architect for Cloud Base Stations

Nokia — Sep 2018 - Feb 2019

- Designed AI architecture for Cloud mobile networks, covering data storage and flows.
- Coordinated work on Nokia Cloud RAN BTS Reference Architecture specification.

Data Scientist, Web app developer, Entrepreneur

LaData — Apr 2017 - Aug 2018

- Developed comprehensive NLP pipelines for document classification, encompassing data cleaning, feature engineering.
- Created and delivered a prototype machine learning text classification system for Wilabs, incorporating end-to-end processing capabilities. The system was used to warn users about potential tax risks in case they reported non-creative work for tax deduction.
- Designed and implemented web applications using Django framework.
- Handled DevOps including Docker containers, orchestration, CI with GitLab, and QA.

Senior Research Engineer, 5G

Nokia Bell Labs — 2015 - 2017

- Leveraging historical data, developed machine learning-based enhancements to predict blockages and mitigate interference in millimeter-wave systems, improving their overall reliability.
- As the leader of a work package in the EU-funded mmMAGIC project, drove technical architectural initiatives while coordinating contributions from ten diverse partners spanning industry, telecommunications operators, and academia.

Software and System Architect, Radio Research Engineer

Siemens, Nokia Siemens Networks 2006 - 2015

Education

M.Sc. in Telecommunications

Wrocław University of Technology 1999 - 2004

Thesis in Computer Vision: “Face image-based person recognition”

M.Sc. in Social Science

University of Wrocław 1999 - 2004

Thesis in Communication and Data Mining: “Characteristics of Internet Mediated relation establishment – a study of dating service”

Skills

Technical

- **Generative AI:** RAG, Agents, OpenAI API, LoRA, PERFT
- **NLP:** Text embeddings, Text classification, Text clustering, Text summarization, Topic modeling
- **ML:** Supervised/Unsupervised ML, Deep Learning, Transformers, LSTM, Autoencoders, GAN, MLOps
- **Developer Tools:** Git, Docker, Scrum, Jira, Jupyter Notebooks, Linux, Bash, SQL, Azure, Google Cloud, ADO, Jenkins, GitHub Actions, Poetry
- **Libraries:** LangChain, LlamaIndex, RAGAS, Scikit-Learn, MLFlow, Pandas, Numpy, PyTorch, LIME, SHAP, Pytest, FastAPI, Django, Flask, Unsloth, NLTK, Spacy, Gensim
- **Languages:** Python, Bash, SQL, MATLAB

Soft

- Guiding employees to grow and develop their skills
- Ability to look at things from various perspectives and see the bigger picture
- Eagerness to explore the unknown
- Effective communication skills
- Teaching/mentoring experience: Nokia Lecturer at Wrocław University of Technology; co-organizer of the GenAI guild in EY, fostering AI knowledge sharing and collaboration.

Publications and Patents

- 11 Conference papers, 6 IPRs, Chapters in 3 books, 1 Journal Paper (details on Google Scholar)
- eBook: MLOPs Interview Preparation - Questions and Answers, solving problems in machine learning production use cases

Certifications

- “Neural Networks and Deep Learning” certificate from deeplearning.ai (2TMZFJMHKZF)
- Microsoft Data engineering exam DP-201: “Designing an Azure Data Solution”

Open Source Contributions

most popular PyPI Packages (where I’m author and maintainer):

- **Count tokens**: Count tokens in text file, counting tokens used by OpenAI models (4k downloads/month)
- **Trend classifier**: Library for automated signal segmentation, trend classification, and analysis (762 downloads/month)
- **git-commits-graph**: Display plot of changes in repo - count of lines or changed lines (261 downloads/month)
- **rankflow**: Library for plotting multiple ranks evolved over processing steps - draw a rankflow (155 downloads/month)

. . . and 4 more

Professional Writing

Most popular articles on safjan.com:

- *“Understanding Retrieval-Augmented Generation (RAG) empowering LLMs”* (Sep 2023)
- *“From Fixed-Size to NLP Chunking - A Deep Dive into Text Chunking Techniques”* (Sep 2023)
- *“MLOps Scorecard – How Advanced Is Your Organization in Implementing MLOps Processes?”* (Jan 2023)
- *“Techniques to Boost RAG Performance in Production”* (Nov 2023)