Contact

09912963951 (Mobile)
alinkkh9@gmail.com

www.linkedin.com/in/alinikkhah2001 (LinkedIn)
alinikkhah2001.github.io (Personal)

Top Skills

Kubernetes
MLflow
Terraform

Languages

Persian (Native or Bilingual)
French (Limited Working)
Arabic (Professional Working)
English (Native or Bilingual)
German (Limited Working)
Latin (Elementary)

Certifications

Introduction to Machine Learning in Production
Building Systems with the ChatGPT API
ChatGPT Prompt Engineering for Developers
Diffusion Models
Project-Oriented Course In Web Development With Php

Honors-Awards

Iran University entrance exam
IELTS on Computer Academic Band 8

Ali Nikkhah

Machine Learning Researcher / Machine Learning Engineer | Computer Science
Tehran, Tehran Province, Iran

Summary

I am an applied-machine-learning researcher and engineering lead who specialises in transforming state-of-the-art methods into robust, production-grade systems. My mandate involves architecting agentic, multi-modal pipelines that autonomously generate, moderate, and enrich millions of product listings in real time—achieving substantial gains in content quality, regulatory compliance, and operational expenditure.

Academic formation and research trajectory
• Sharif University of Technology, B.Sc. in Electrical Engineering
Grounded my work in signal processing, control theory, and large-scale systems design.
• Trinity College Dublin, Research Assistant in Computer Science
Concentrating on generative vision-language models and autonomous agents while contributing to the Emo Dub Project on emotion-aware machine translation.

Experience

Turquoise Digital - لاتیجید هوزیف

4 months

Senior Data Scientist
September 2025 - November 2025 (3 months)
Tehran, Tehran Province, Iran

Data Scientist
August 2025 - September 2025 (2 months)
Tehran, Tehran Province, Iran

Trinity College Dublin

Machine Learning Researcher – Emotion-Aware Voice Translation
May 2025 - November 2025 (7 months)
Dublin, County Dublin, Ireland

• Conducted pioneering research on emotion-aware speech-to-speech translation as part of the EmoDub Project, focusing on bridging linguistic and emotional fidelity in multilingual voice systems.
• Collaborated with Professor Siobhán Clarke and a cross-disciplinary team to explore multimodal AI approaches integrating speech, emotion, and linguistic features for naturalistic translation.
• Designed and trained deep learning models capable of jointly capturing prosody, tone, and affective cues, ensuring emotional nuance preservation during translation.
• Implemented speech emotion recognition (SER) modules using transformer-based encoders and wav2vec2.0 embeddings, optimizing emotional transfer across languages.
• Developed a multimodal fusion pipeline combining acoustic, semantic, and contextual embeddings for emotion-consistent translation outputs.
• Created and fine-tuned datasets using emotion-annotated speech corpora, ensuring balanced representation and reducing cultural bias in model predictions.
• Evaluated model performance using BLEU, Emotion F1, and MOS (Mean Opinion Score) metrics, achieving significant gains in emotional expressivity and translation accuracy.
• Contributed to research reports, reproducible codebases, and internal publications to advance the lab’s work on affective computing and expressive AI systems.

The University of British Columbia
Ultrasound Report Generation Research Assistant
February 2024 - November 2025 (1 year 10 months)
Tehran, Iran

• Designed and implemented an automated ultrasound report generation system integrating vision-language models to assist clinicians in diagnostic decision-making.
• Applied contrastive and self-supervised learning techniques to align ultrasound image embeddings with clinical text representations, enabling robust cross-modal understanding.
• Developed an image-to-text generation pipeline leveraging transformer-based architectures (e.g., ViT, BLIP, and T5) for structured and explainable medical report synthesis.
• Curated and preprocessed medical imaging datasets (ultrasound scans, annotations, and radiology reports), ensuring data quality, de-identification, and alignment with medical standards (DICOM compliance).

• Enhanced model interpretability by integrating attention visualization and saliency-based explanation modules, improving clinical trust and usability.
• Collaborated with radiologists and researchers to evaluate generated reports against clinical accuracy, terminology consistency, and diagnostic completeness.
• Achieved measurable improvements in report quality metrics (BLEU, ROUGE-L, and Clinical Efficacy Scores), advancing the project's capability for real-world medical deployment.

Robust and Interpretable Machine Learning Lab
Applied Machine Learning Researcher
July 2025 - September 2025 (3 months)
Tehran, Tehran Province, Iran

Digikala
10 months

Al Engineering R/D Lead
January 2025 - August 2025 (8 months)
Tehran, Tehran Province, Iran

• Led the development of Modular Control Pipelines (MCP) and Retrieval-Augmented Generation (RAG) frameworks for scalable enterprise applications.
• Designed and deployed Agentic RAG systems that autonomously coordinate retrieval, reasoning, and response generation across dynamic knowledge sources.
• Built multi-agent orchestration frameworks integrating specialized reasoning agents, enabling context-aware collaboration and adaptive decision-making.
• Architected data retrieval pipelines with hybrid search (vector + keyword) and optimized embedding strategies for improved precision and recall.
• Enhanced system performance and reliability by introducing memory-efficient document indexing, response caching, and knowledge base synchronization.
• Integrated real-time evaluation and feedback loops for continuous learning, model refinement, and context optimization.
• Collaborated cross-functionally with data, infrastructure, and product teams to align AI capabilities with production-level business objectives.
• Drove measurable gains in response accuracy, latency reduction, and operational cost-efficiency through continuous optimization and agentic reasoning loops.

Al Engineer
November 2024 - January 2025 (3 months)
Tehran, Tehran Province, Iran

• Led the development of Modular Control Pipelines (MCP) and Retrieval-Augmented Generation (RAG) frameworks for scalable enterprise applications.
• Designed and deployed Agentic RAG systems that autonomously coordinate retrieval, reasoning, and response generation across dynamic knowledge sources.
• Built multi-agent orchestration frameworks integrating specialized reasoning agents, enabling context-aware collaboration and adaptive decision-making.
• Architected data retrieval pipelines with hybrid search (vector + keyword) and optimized embedding strategies for improved precision and recall.
• Enhanced system performance and reliability by introducing memory-efficient document indexing, response caching, and knowledge base synchronization.
• Integrated real-time evaluation and feedback loops for continuous learning, model refinement, and context optimization.
• Collaborated cross-functionally with data, infrastructure, and product teams to align AI capabilities with production-level business objectives.
• Drove measurable gains in response accuracy, latency reduction, and operational cost-efficiency through continuous optimization and agentic reasoning loops.

L3S Research Center
Video Motion Classification Research Assistant
April 2023 - February 2024 (11 months)
Hannover, Lower Saxony, Germany

• Conducted a comprehensive literature review on cutting-edge video action and motion classification techniques, focusing on advancements in spatiotemporal representation learning and transformer-based architectures.
• Designed and implemented texture-free motion classification models, leveraging optical flow representations, 3D CNNs, and transformer encoders to isolate movement dynamics from environmental noise.
• Pioneered methods to improve domain generalization under varying textures, lighting, and camera conditions, contributing to more robust and adaptable video understanding pipelines.
• Developed a data preprocessing and augmentation pipeline for motion-based datasets, optimizing temporal sampling and frame-level consistency.
• Benchmarked experimental models against standard baselines, achieving state-of-the-art performance on the UCF-101 and HMDB-51 datasets with improved accuracy and reduced overfitting.
• Collaborated closely with senior researchers to analyze cross-domain generalization and interpret model attention patterns for explainable motion recognition.

• Documented research findings in detailed reports and contributed to the preparation of manuscripts and presentations for academic dissemination.

Sharif University of Technology
3 years 1 month

Teaching Assistant -- Natural Language Processing (Graduate Course)*
August 2023 - January 2024 (6 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Generative Models (Graduate Course)*
August 2023 - January 2024 (6 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Deep Learning (Graduate Course)*
January 2023 - August 2023 (8 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Analog Electronics
August 2022 - January 2023 (6 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Machine Learning
August 2022 - January 2023 (6 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Data Science
January 2022 - August 2022 (8 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Circuit Theory
August 2021 - January 2022 (6 months)
Tehran, Tehran Province, Iran

Teaching Assistant -- Engineering Probability and Statistics Course
January 2021 - August 2021 (8 months)
Tehran, Tehran Province, Iran

HomaCloud || ۱۴۰۰
10 months

MLOps
August 2021 - February 2022 (7 months)
Tehran, Tehran Province, Iran

• Designed and implemented scalable MLOps pipelines for multi-node training and continuous model deployment across distributed environments.
• Developed automated CI/CD workflows integrating Docker, Kubernetes, and GitHub Actions, streamlining model versioning, retraining, and API deployment.
• Researched and benchmarked machine learning frameworks (PyTorch, TensorFlow, and ONNX Runtime) to optimize compatibility and performance across the platform.
• Built a high-performance model serving infrastructure using FastAPI and TensorFlow Serving, achieving low-latency inference for real-time client applications.
• Implemented data validation, monitoring, and logging systems to ensure reproducibility, model reliability, and traceability throughout the ML lifecycle.
• Collaborated with backend and data engineering teams to integrate AutoML components, accelerating experimentation and deployment cycles.
• Contributed to the design of infrastructure-as-code (IaC) templates, improving reproducibility and scalability for internal ML workflows.
• Achieved measurable improvements in deployment speed, system uptime, and resource utilization, enhancing both developer productivity and model performance at scale.

MLOPs intern
May 2021 - August 2021 (4 months)
Tehran, Tehran Province, Iran

• Conducted extensive research on machine learning libraries like PyTorch and TensorFlow, enhancing our integration strategies.
• Developed MLOps workflows for multi-node model training, improving scalability and deployment automation.
• Built a high-performance model serving infrastructure, ensuring seamless API integration for clients and optimizing model performance.

Mapna-MD2
7 months

Full Stack Engineer
January 2021 - May 2021 (5 months)
Tehran, Tehran Province, Iran

Full Stack Engineer Intern
November 2020 - January 2021 (3 months)
Tehran, Tehran Province, Iran

Education

Sharif University of Technology

Bachelor's degree, Electrical and Electronics Engineering · (2020 - 2024)

Atomic Energy High School

High School Diploma, Mathematics · (January 2017 - February 2019)