

Zahra Anvari

📍 Santa Clara, CA ✉ zahra.anvari.uta@gmail.com
🌐 LinkedIn 🐙 GitHub 🎓 Google Scholar

Summary

Machine Learning Researcher and Engineer with a PhD in Computer Science specializing in **large language models, multimodal AI, and data-centric AI systems**. Experienced designing **reproducible LLM evaluation frameworks**, generating synthetic training data, and translating research prototypes into production-scale machine learning systems. Research and industry experience spans Document AI, structured information extraction, model evaluation, synthetic document degradation, and robust AI systems for real-world enterprise applications.

Professional Experience

Independent Researcher

Santa Clara, CA

April 2024 – Present

- Leading applied research on **LLM-based Document AI**, with a focus on robust key-value extraction from forms and receipts under realistic OCR noise.
- Designed and implemented a **reproducible LLM evaluation framework** supporting unified prompting, deterministic inference, structured-output canonicalization, and controlled benchmarking across six instruction-tuned LLMs, three document datasets, and three OCR engines.
- Conducted zero-shot and few-shot experiments under clean-text and OCR-degraded conditions, evaluating model performance using Key Recall, Exact Match, and Value F1 and identifying failure modes including hallucination, numeric corruption, and key-value misalignment.
- Investigating hybrid approaches that combine OCR, layout-aware transformers, and LLM reasoning to improve structured information extraction from noisy and layout-dependent documents.

Iron Mountain Inc.

Remote, U.S.

Senior Machine Learning Engineer | July 2021 – March 2024

- Designed and deployed production Document AI pipelines integrating document restoration, OCR, transformer-based NLP, and layout-aware models for structured information extraction from enterprise documents.
- Designed a **synthetic document degradation pipeline** using Python, OpenCV, and OCRODeg to generate multiple levels of blur, noise, skew, fading, bleed-through, and related scanning artifacts for training unpaired CycleGAN document enhancement models, improving document legibility and OCR performance for downstream information extraction.
- Developed transformer-based **information extraction systems** (BERT-based NER) for 30+ entity types, enabling automated parsing of contracts and enterprise documents.
- Prototyped LLM-powered document question-answering workflows using RAG, LangChain, and vector databases, enabling semantic retrieval and context-aware Q&A over contracts and enterprise documents.
- Built **table detection and structure recognition models** for document parsing pipelines, improving detection mAP by > 30%.
- Built evaluation pipelines for document extraction systems, measuring precision, recall, robustness, OCR sensitivity, and downstream task performance to guide model selection and iteration.
- Prototyped rapid NER solutions using GCP Vertex AI AutoML to accelerate client-specific entity extraction and evaluation.
- Led ML projects from research and prototyping through deployment, collaborating with engineering, product, and client stakeholders to translate requirements into production AI solutions and mentoring junior engineers.
- Contributed novel document enhancement methods that led to a provisional patent for improving the quality and readability of degraded enterprise documents.

Wave Computing Inc.

Campbell, CA

Deep Learning Intern | May 2019 – November 2019

- Implemented an enhanced emotion/face detection pipeline for retail analytics, improving accuracy by **16%**.
- Benchmarked super-resolution and face detection methods for performance analysis.

University of Texas at Arlington

Arlington, TX

Research Assistant | Aug. 2016 – May 2021

- Researched deep learning applications for *image enhancement and restoration*, resulting in multiple publications.

ACECR, Sharif University Branch

Tehran, IR

Software Engineer | June 2013 – June 2015

- Implemented several vehicular embedded applications for Intelligent Transportation Systems in C/C++.

Technical Skills

AI & Machine Learning: Large Language Models, Multimodal AI, Vision-Language Models, Retrieval-Augmented Generation, Transformers, Named Entity Recognition, Information Extraction, Synthetic Data Generation, Document AI, OCR, Generative AI, GANs, CNNs

Evaluation & Experimentation: LLM Evaluation, Controlled Experiment Design, Model Benchmarking, Prompt Engineering, Zero-Shot and Few-Shot Evaluation, Robustness Analysis, Failure Mode Analysis, Dataset Curation, Model Validation

Frameworks & Libraries: PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, NumPy, Pandas, OpenCV, LangChain, DETR, YOLOv8

Cloud & ML Infrastructure: AWS (SageMaker, Bedrock, Textract, Lambda, S3), GCP (Vertex AI, AutoML), Microsoft Azure, Docker, Kubeflow

Programming: Python, C/C++, SQL, Git

Collaboration: Cross-functional Collaboration, Client Engagement, Project Leadership, Requirements Gathering, Mentoring

Education

University of Texas at Arlington

Arlington, TX

PhD in Computer Science | 2015 - 2021 | GPA: 4.0 / 4.0

Focus: Deep Learning, Computer Vision, Generative Models | Advisor: Vassilis Athitsos

Urmia University

Urmia, IR

M.S. in Computer Engineering | 2010 - 2013

Shahid Rajaee University

Tehran, IR

B.S. in Information Technology Engineering | 2004 - 2008

Publications

- **From Pixels to Pairs: A Comprehensive Benchmark of LLM-Driven Key-Value Extraction in Noisy Document Settings**, Zahra Anvari and Vassilis Athitsos, In preparation, 2026.
- **Large Language Models for Structured Document Understanding: A Comprehensive Survey of Tasks, Models, and Challenges**, Zahra Anvari and Vassilis Athitsos, TechRxiv, 2025.
- **A Survey on Deep Learning Based Document Image Enhancement**, Zahra Anvari and Vassilis Athitsos, ArXiv, 2021.
- **Enhanced CycleGAN Dehazing Network**, Zahra Anvari and Vassilis Athitsos, In proc. VISAPP, Vienna, Austria, 2021.

- **Evaluating Single Image Dehazing Methods Under Realistic Sunlight Haze**, Zahra Anvari and Vassilis Athitsos, *In proc. ISVC*, San Diego, CA, 2020.
- **A Pipeline for Automated Face Dataset Creation from Unlabeled Images**, Zahra Anvari and Vassilis Athitsos, *In proc. PETRA*, ACM, 2019.