Ghazaleh Kazeminejad
Senior AI/ML Solution Architect | Former Senior AI Engineer
Beaverton, OR | gh.kazeminejad@gmail.com | 919-961-9650 | LinkedIn

PROFESSIONAL SUMMARY

PhD-trained AI/ML Solution Architect (former Senior AI Engineer) with 8+ years of experience designing and deploying scalable deep learning and generative AI systems. Proven track record translating cutting-edge research into real-world impact across industries including retail, telecom, automotive, entertainment, IT, finance, legal, and eCommerce. Adept in customer-facing technical engagements, cloud-native ML deployment, and development of reusable GenAI assets. Deep expertise in LLM evaluation, RAG pipelines, agentic orchestration, and MLOps best practices. Passionate about building responsible, privacy-aware AI systems that solve complex business problems at scale.

PROFESSIONAL EXPERIENCE

Senior AI/ML Solution Architect
IBM | April 2025 – Present

- Collaborated with AI/ML teams to research, prototype, and evaluate deep learning and generative AI experiments for client-specific scenarios.
- Led direct engagements with Fortune 500 clients to scope PoCs, map pain points to AI solutions, and guide successful adoption paths for GenAI capabilities.
- Designed and delivered customer-facing workshops, demos, and technical briefings that accelerated buy-in and deployment.
- Created internal best-practice guides, reusable code templates, and modular reference architectures reused across global delivery teams.
- Synthesized field feedback to shape roadmap priorities for reusable GenAI modules; acted as conduit between engineering and client-facing stakeholders.

Senior AI Engineer – GenAI Systems & Orchestration
IBM | Oct 2023 – April 2025

- Architected, developed, and deployed production-grade GenAI systems using open-source and proprietary LLMs; integrated RAG pipelines with fine-tuned ranking and filtering layers.
- Developed agentic prototypes, including multi-agent assistants and voice-based copilots, designed to support multi-turn workflows across customer service and internal operations.
- Created containerized LLM stacks using Docker Compose for portable deployment across IBM Cloud, AWS and Azure; enabled rapid PoC development with secure, modular infrastructure.

- Built custom evaluation frameworks using LLM-as-a-judge scoring, hallucination detection, and groundedness testing; implemented observability dashboards with Prometheus + Grafana.
- Presented architectural strategy and experiment outcomes to client executives, product managers, and engineering teams.

NLP Research Scientist
RedShred | Jan 2022 – Oct 2023

- Conducted applied research in neural-symbolic learning, document intelligence, and semantic search in high-stakes domains.
- Fine-tuned LLMs and used with hybrid search to perform RAG on highly domain-specific documents.
- Deployed models to production via Kubernetes and TorchServe; integrated guardrails for PII protection and QA audit compliance.
- Built reusable evaluation workflows to inform model selection and reduce deployment risk.

Data Science Intern
Walmart eCommerce | Summer 2019

- Enhanced product ontology to improve search relevance across eCommerce listings using query-to-product enrichment strategies.
- Built NLP pipelines to extract structured attributes from listing text; improved mapping precision for ranking and recommendations.

Graduate Research Assistant
University of Colorado Boulder | 2015 – 2020

- Researched and invented novel hybrid neural-symbolic architectures for question answering, reasoning, and classification.
- Published 12+ peer-reviewed papers in top NLP venues (ACL, EMNLP, COLING, NAACL); co-authored a book chapter.

EDUCATION

Ph.D. in Natural Language Processing
University of Colorado Boulder
Dissertation Title: Computational Lexical Resources for Explainable Natural Language Understanding
Topics: Transfer Learning (LLM fine-tuning), data augmentation, hybrid knowledge-aware neural-symbolic architectures, Natural Language Understanding
Advisor: Prof. Martha Palmer

B.Sc. in Physics
Amirkabir University of Technology

TECHNICAL SKILLS

- **Languages & Frameworks**: Python, PyTorch, TensorFlow, HuggingFace, Scikit-learn, LangChain, FastAPI
- **Generative AI & LLMs**: RAG, agentic systems, prompt tuning, PEFT/LoRA, hallucination detection
- **MLOps & Infrastructure**: Docker, Docker Compose, Kubernetes, CI/CD, TorchServe, AWS SageMaker (in progress), IBM Cloud
- **Evaluation & Monitoring**: MLflow, Weights & Biases, custom eval harnesses, RAGAS, OpenTelemetry, Grafana, Prometheus
- **Cloud Platforms**: cross-cloud deployment experience (IBM Cloud, AWS, Azure)