MohammadJavad AghababaieBeni

75 Waterworks Rd, Birmingham, B169AL
mj.babaie@gmail.com | 07508487847

Applied AI / ML Engineer with a strong foundation in machine learning and hands-on experience building, deploying, and operating production-grade AI systems. Specialises in NLP and LLM-based applications, including Retrieval-Augmented Generation (RAG), model evaluation, and reliability guardrails. Experienced in delivering scalable AI solutions using Python, FastAPI, Docker, CI/CD, and cloud platforms (GCP, Azure), with end-to-end ownership from experimentation to monitored production. Holds the right to work in the UK on a Graduate Visa valid until February 2027; open to remote, hybrid, or on-site roles UK-wide.

GitHub: https://github.com/Muh76
LinkedIn: https://www.linkedin.com/in/mohammadbabaie/

Professional Experience

Outlier AI Remote/US   **10/24 - Present**
LLM Evaluation Analyst

- Evaluated and scored 600+ LLM responses across legal QA, general knowledge, and creative tasks, identifying reasoning errors, hallucinations, and knowledge gaps.
- Designed and authored data science–focused coding tasks (data preprocessing, modeling, evaluation, and error analysis) to assess LLM reasoning quality and alignment with real-world analytical workflows.
- Implemented structured evaluation frameworks and content guardrails, improving response correctness by ~20% across subsequent model iterations.
- Collaborated with AI engineers by delivering detailed evaluation reports and actionable feedback, directly informing data and model improvements.
- Increased evaluation coverage by 30% by designing edge-case and stress-test scenarios, strengthening robustness across diverse input distributions.

Behineh Bana Zagros Isfahan/Iran   **02/20 - 06/23**
Data Scientist

- Analyzed operational, cost, and project datasets across multiple construction and industrial initiatives to support planning and data-driven decision-making.
- Built regression and time-series models to forecast project timelines, material demand, and costs, reducing manual estimation effort by ~25%.
- Developed automated data processing pipelines and analytical reports using Python and SQL, cutting recurring reporting time by ~30%.
- Designed and maintained Power BI dashboards tracking KPIs (progress, cost variance, resource utilisation) for both technical and non-technical stakeholders.
- Cleaned, structured, and maintained large-scale datasets covering procurement, schedules, and supplier performance.
- Applied statistical analysis to uncover patterns, anomalies, and operational risks, supporting measurable improvements in project execution.

Education

Aston University   **09/23 - 10/24**
MSc in Computer Science

Azad University of Sharekord   **09/16 - 01/20**
BSc in Computer Science

Key Skills

- Programming Languages: Python, LangChain, TensorFlow, PyTorch, scikit-learn, Pandas, NumPy, SQL, Git
- LLMs & NLP: RAG Pipeline, Hugging Face Transformers, Guardrail Techniques, Prompt Engineering, Text Classification
- Machine Learning: Supervised Learning (Regression, Classification, Ensemble Methods), Neural Networks, Feature Engineering, Clustering (KMeans, DBSCAN), Model Calibration
- Backend & APIs: REST APIs, FastAPI, Pydantic, API Auth (JWT), Async Endpoints
- MLOps & Deployment: Docker, MLflow, DVC, CI/CD (GitHub Actions)
- Monitoring & Evaluation: Prometheus, Grafana, Evidently, Experiment Tracking, RAGAS
- Cloud & Analytics: GCP, Azure ML, Power BI
- Web & Interfaces: Streamlit, JavaScript/ TypeScript, Next.js

Projects

Legal Q&A RAG Chatbot (UK Legal Domain)   09/25 - Present

- Production-ready RAG system over 131k+ legal document chunks
- Hybrid retrieval: BM25, FAISS, OpenAI embeddings, RRF, cross-encoder reranking
- FastAPI backend with role-based access (Solicitor / Public)
- Sub-3s average latency with 680+ embeddings/sec throughput
- 15–20% retrieval accuracy improvement via reranking
- 40% hallucination reduction using domain filtering and citation guardrails
- End-to-end deployment: Docker, PostgreSQL, Streamlit

Autonomous Data & Compliance Officer (Multi-Agent AI System)   10/25 - 12/25

- Multi-agent AI system for automated GDPR/HIPAA compliance auditing
- Built with Google ADK and Vertex AI (Gemini)
- Orchestrated 6 specialised agents for PII detection, regulatory mapping, and violation classification
- Presidio-based granular PII detection and redaction
- ChromaDB-backed RAG for regulation grounding
- 85%+ precision in violation detection with 2–3× speedup vs manual audits
- Designed to reduce audit time by ~95% with full schema coverage
- CI/CD-ready deployment with Docker, Cloud Run, and distributed tracing

Contract Review & Risk Analysis System (CUAD)   04/25 - 08/25

- NLP system for automated clause segmentation, contract classification, and risk scoring
- Built on the CUAD dataset with production-grade evaluation
- Micro-F1 0.837 and 97.4% contract-type accuracy
- Sub-500 ms inference latency per contract
- End-to-end pipeline: TF-IDF + Logistic Regression and DistilRoBERTa with probability calibration
- Model serving via FastAPI with batch and dashboard interfaces
- Deployed on GCP Cloud Run with Dockerised services
- MLOps practices: DVC/MLflow tracking, group-by-contract splits
- ROI analysis showing break-even at ~14 contracts/month (≈6.7-month payback)