Elena Marquez | Senior AI Engineer | LLM Systems & Machine Learning

500 Terry A Francois Blvd, San Francisco, California, United States, 94158
(415) 555-0148 • elena.marquez.ai@gmail.com • https://elenamarquez.dev
@elena-marquez • @elena-marquez-ai • @elena-marquez-ai

Basics

AI Engineer with 6+ years of experience designing, evaluating, and deploying production machine learning and large language model systems. I turn ambiguous product goals into measurable models, reliable inference services, and evaluation frameworks that teams can trust.

- Built retrieval-augmented generation and agentic workflows for healthcare and enterprise search, improving answer quality while reducing latency and cost.
- Led model evaluation, prompt optimization, and guardrail design across multiple product teams, with a focus on safety, observability, and reproducibility.
- Partnered with product, data, and platform engineers to ship ML features from prototype to production with clear ownership and monitoring.

Education

University of Washington **https://www.washington.edu/**

Master, Computer Science and Artificial Intelligence, Score: 3.9 *Sep 2017–Jun 2019*

- Completed graduate coursework and research at the intersection of machine learning, NLP, and human-centered AI.
- Built reproducible experiments for text classification and retrieval, including error analysis and ablation studies.
- Presented research findings to academic and industry audiences and collaborated across disciplines.

Courses: Machine Learning, Deep Learning, Natural Language Processing, Probabilistic Graphical Models, Distributed Systems, Statistical Inference, Human-Computer Interaction

Work

Aether Labs **https://www.aetherlabs.ai**

Senior AI Engineer *Mar 2023–Present*

- Lead the design and deployment of LLM-powered search and assistant features used by enterprise customers.
- Built a retrieval-augmented generation pipeline with hybrid search, reranking, and citation-aware responses, reducing unsupported claims by 38%.
- Created an offline and online evaluation stack covering relevance, groundedness, latency, and cost, enabling weekly model and prompt releases.
- Partnered with platform engineers to productionize inference services with batching, caching, and observability on Kubernetes.
- Mentored two AI engineers and established review practices for prompt changes, data contracts, and model risk assessments.

Keywords: LLM Systems, RAG, Evaluation, Inference, Mentorship

BrightPulse Health **https://www.brightpulsehealth.com**

Machine Learning Engineer *Jul 2019–Feb 2023*

- Developed clinical risk and triage models using structured EHR data, improving recall for high-risk patients by 17% in prospective validation.
- Built NLP pipelines to extract signals from unstructured notes, with de-identification, labeling, and active learning workflows.
- Deployed batch and real-time inference services on AWS, with data quality checks, model monitoring, and rollback procedures.
- Worked with clinicians and compliance teams to document model limitations, intended use, and bias mitigation steps.
- Automated feature pipelines in Airflow and Spark, cutting weekly model refresh time from two days to four hours.

Keywords: Healthcare, NLP, Risk Models, AWS, Monitoring

VectorForge AI **https://www.vectorforge.ai**

AI Research Intern *Jun 2018–Sep 2018*

- Prototyped sentence embeddings and retrieval baselines for semantic search over technical documentation.
- Ran ablation studies on negative sampling and reranking, documenting findings in an internal research report.
- Contributed reusable evaluation scripts and data preparation notebooks to the research team.

Keywords: Embeddings, Semantic Search, Research

Languages

English: Native or Bilingual Proficiency — **Keywords**: Technical writing, Conference talks

Spanish: Full Professional Proficiency — **Keywords**: Bilingual collaboration

French: Elementary Proficiency — **Keywords**: DELF A2

Skills

Machine Learning: Expert — **Keywords**: PyTorch, TensorFlow, scikit-learn, XGBoost, Statistical Modeling

LLM & Generative AI: Advanced — **Keywords**: Transformers, LangChain, RAG, Prompt Engineering, Fine-tuning, Agents

Data Engineering: Advanced — **Keywords**: Python, SQL, Apache Airflow, Apache Spark, dbt, Snowflake

Cloud & MLOps: Advanced — **Keywords**: AWS, GCP, Docker, Kubernetes, MLflow, Terraform

Software Engineering: Advanced — **Keywords**: FastAPI, TypeScript, REST APIs, CI/CD, Testing, System Design

Awards

AI Platform Innovation Award

BrightPulse Health *Nov 2022*

Recognized for leading the design of a clinical NLP platform that reduced manual chart review time by 30% while maintaining auditability and clinician trust.

Graduate Research Excellence Award

University of Washington *May 2019*

Awarded for research on retrieval and evaluation methods for domain-specific language models, including a published workshop paper and reproducible benchmark.

Certificates

AWS Certified Machine Learning - Specialty **https://aws.amazon.com/certification/certified-machine-learning-specialty/**

Amazon Web Services *Aug 2022*

Google Cloud Professional Machine Learning Engineer **https://cloud.google.com/certification/machine-learning-engineer**

Google Cloud *Jan 2023*

Publications

Proceedings of the Conference on Applied Machine Learning **https://dl.acm.org/doi/10.1145/3612345.3612399**

Evaluation-Driven Development for Retrieval-Augmented Generation *Oct 2023*

- Proposes a practical evaluation loop for RAG systems that combines retrieval diagnostics, groundedness checks, and human review.
- Shares lessons from deploying the framework across customer support and healthcare search use cases.

Journal of Healthcare AI **https://doi.org/10.1016/j.jhai.2021.03.004**

Calibrating Clinical Risk Models Under Distribution Shift *Mar 2021*

- Studies calibration drift in clinical risk prediction models when patient populations and care pathways change.
- Introduces monitoring signals and retraining triggers that helped maintain model reliability in production.

Projects

RAG Evaluation Harness **https://github.com/elena-marquez/rag-eval-ha**

Open-source toolkit for reproducible evaluation of retrieval-augmented generation pipelines *Jan 2024–P*

- Provides pluggable retrievers, rerankers, and judges for measuring relevance, groundedness, and answer completeness.
- Includes synthetic query generation, regression tests, and dashboards for comparing prompt and model changes.

Keywords: RAG, Evaluation, LLM

Clinical Note Summarizer **https://github.com/elena-marquez/clinical-note-summarizer**

Privacy-preserving summarization service for clinician-authored notes *Feb 2022–Nov 2022*

- Built a de-identification and summarization pipeline with human-in-the-loop review and audit logging.
- Evaluated summaries for factual consistency, coverage, and clinician acceptance before a limited pilot.

Keywords: Healthcare, NLP, Summarization

Interests

Open-Source AI: Evaluation, Reproducibility, Agents

Outdoors: Trail Running, Climbing, Cycling

Music: Jazz Piano, Music Production

Volunteer

Women in Machine Learning **https://wiml.org**

Mentor *Jan 2021–Present*

- Mentor early-career engineers and researchers on ML project scoping, interview preparation, and career growth.
- Review workshop submissions and support community events focused on inclusive AI research and practice.