Gabriel Soto

Jersey City, NJ · gabriel.soto@example.com · (201) 555-0181 · linkedin.com/in/example-gabriel-soto

My Key Qualifications

- **Fraud detection** in production: raised review-queue precision from 0.41 to 0.56 with a model deployed on SageMaker at Marqeta.
- **RAG and LLM applications**: internal assistant on Pinecone and LangChain cut ticket research from 14 to 8 minutes across 1,200 monthly queries.
- **LLM evaluation framework** built with MLflow and OpenAI, lifting approved-response rate from 68% to 84% on a 500-case benchmark.
- Production AWS serving at **430 ms p95** (down from 920 ms) for a risk-scoring API on ECS with CloudWatch monitoring.
- **CI/CD for ML** via GitHub Actions and Terraform, cutting release time from 3 hours to 55 minutes, with standardized MLflow experiment tracking.
- Payments fintech experience at **Marqeta and AvidXchange**, including invoice-risk features that improved recall from 0.63 to 0.74.
- Cross-functional delivery with **product, analytics, and compliance** teams to define experiments, release criteria, and post-launch monitoring.

Experience

AI Engineer — **Marqeta**, card issuing and payments platform

Mar 2023 – Present · 3 yrs 8 mos

- Built and deployed a fraud case-prioritization model for transaction review queues, raising precision at the review cutoff from 0.41 to 0.56 by training gradient-boosted classifiers on transaction and merchant features in Python and Spark and deploying scoring jobs in SageMaker.
- Launched an internal RAG assistant for support and operations teams, cutting average ticket research time from 14 minutes to 8 minutes across 1,200 monthly queries; indexed policy and product content in Pinecone, orchestrated retrieval and prompts with LangChain, and served responses through ECS.
- Implemented an LLM evaluation framework for support reply generation, improving the approved-response rate from 68% to 84% on a 500-case benchmark by defining groundedness and tone checks, tracking runs in MLflow, and comparing prompt versions using OpenAI models.
- Reworked model-serving infrastructure for a risk-scoring API, cutting p95 latency from 920 ms to 430 ms by moving inference to containerized services on ECS, optimizing feature lookups, and adding CloudWatch performance monitoring.
- Introduced GitHub Actions and Terraform into the AI deployment workflow, automating test, build, and infrastructure steps and reducing production release time from 3 hours to 55 minutes.
- Partner with product, analytics, compliance, and platform teams to scope experiments, define release criteria, and run post-launch monitoring for AI services across fraud, support automation, and internal decision workflows.

Sample resume made with Specific Resume · specificresume.com

Machine Learning Engineer — **AvidXchange**, accounts payable automation and payments software

Oct 2021 – Feb 2023 · 1 yr 4 mos

- Built invoice-risk features in Python and SQL for a review model used in payment operations, improving recall from 0.63 to 0.74 on the validation set by retraining with updated class weighting.
- Created a document-triage service for support teams, serving classification outputs through Lambda-backed APIs and Docker, cutting average routing time from 11 minutes to 4 minutes per case.
- Set up MLflow experiment tracking and scheduled retraining support across two ML pipelines, standardizing logging and job definitions and reducing undocumented model runs from 9 per quarter to 1.
- Built feature pipelines and training datasets for production ML models, and supported model deployment, monitoring, and experiment tracking in AWS.
- Worked with product and operations teams to translate workflow pain points into measurable ML use cases.

Education

- **Rutgers University** — Bachelor of Science in Computer Science · 2017 – 2021

Skills

- **AI & Machine Learning:** PyTorch, Spark, SageMaker, MLflow, LangChain, Pinecone, RAG, Fraud Detection
- **Languages & Data:** Python, SQL
- **Cloud & MLOps:** AWS ECS, AWS Lambda, S3, CloudWatch, Docker, GitHub Actions, Terraform