Felix Brenner Senior MLOps Engineer

New York, NY · felix.brenner@gmail.com · +1 212-555-0178 · linkedin.com/in/felixbrenner

PROFILE SUMMARY

- **Senior MLOps Engineer** with **7 years** running **model serving and inference platforms** at **Hugging Face** and an enterprise search company, focused on **GPU inference** and **LLM serving**.
- Works daily with **Python** and Go, **vLLM** and **NVIDIA Triton**, **Kubernetes** on AWS and Google Cloud, Terraform, Argo CD, **Ray**, and Prometheus with Grafana.
- Strongest in **autoscaling on GPU load**, **quantization and batching**, **LLM evaluation before rollout** and **canary releases**, with every serving change tested against a recorded traffic replay first.
- Works with the **open-source library teams**, product engineers and support on customer endpoints, and runs the **incident review** after any serving outage.
- Mentors **4** engineers on serving performance, wrote the team's **GPU capacity playbook**, and reviews every change to the shared inference Helm charts.

TECHNICAL SKILLS

Languages & Scripting:	Python, Go, Bash, SQL
ML Platforms & Tracking:	MLflow, Weights & Biases, Amazon SageMaker, Vertex AI
Pipelines & Orchestration:	Ray, Argo Workflows, Apache Airflow, Kubeflow Pipelines
Model Serving:	vLLM, NVIDIA Triton Inference Server, KServe, ONNX Runtime, TensorRT-LLM
Feature & Data Tools:	Apache Spark, Delta Lake, DVC, Hugging Face Datasets
Cloud, Containers & CI/CD:	Kubernetes, Amazon EKS, GKE, Karpenter, Terraform, Helm, Argo CD, GitHub Actions
Observability & Drift:	Prometheus, Grafana, OpenTelemetry, Evidently

EDUCATION

Technical University of Munich — **M.Sc. in Informatics** — Munich, Germany · Oct 2017 to Sep 2019

WORK EXPERIENCE

Hugging Face — **Senior MLOps Engineer** — New York, NY · Mar 2023 to Present

- Moved dedicated LLM endpoints from a custom server to **vLLM** on **Kubernetes** with continuous batching and prefix caching, raising output tokens per GPU by **2.6x** at the same p95 latency.
- Set up autoscaling on queue depth and GPU use with **Karpenter** and scale-to-zero for idle endpoints, cutting monthly GPU spend on the shared fleet by **34%** with no drop in uptime.
- Built a release gate that runs each new model version against a fixed evaluation set and a replay of recorded traffic, then shifts **5%** of requests as a canary before full rollout.
- Added **Prometheus** metrics for time to first token, tokens per second and error rate per endpoint, with **Grafana** alerts that cut the time to detect a stuck endpoint to **4 minutes**.
- Wrote **Terraform** modules for GPU node pools on AWS and Google Cloud, so a new region for customer endpoints now takes **2 days** to bring up instead of 3 weeks.

Yext — **Machine Learning Engineer** — New York, NY · Nov 2019 to Feb 2023

- Moved search ranking and answer extraction models to **NVIDIA Triton** with ONNX export and dynamic batching, cutting p95 latency from 180 ms to **65 ms** for customer search pages.
- Built **Kubeflow Pipelines** retraining for **14** language-specific ranking models on features from a shared Spark job, with MLflow tracking and an evaluation step that blocked **3** weaker models from release.
- Added weekly drift reports on query and click data with **Evidently**, which caught a tokenizer change that had cut answer accuracy by **6 points** in German search.