
SUMMARY

Senior ML engineer building production LLM systems, RAG/context pipelines, eval harnesses, and model-serving infrastructure. Previously worked on inference serving at Twitter scale across 500+ models and 5M+ QPS.

EXPERIENCE

- **Adobe** Seattle
Senior Machine Learning Engineer *Feb 2023 - Present*
 - **AEP AI Org**: Built LLM assistant and agentic orchestration platform features, including RAG workflows, tool-calling/harness integration, deployment, monitoring, and CI/CD
 - Developed **entity linking** and **business context service** to ground user queries before LLM processing, using embedding models, full-text search, and vector search across millions of entities.[Blog]
 - Implemented LLM evals/regression testing framework that allowed 10+ collaborating teams to merge features simultaneously while maintaining response accuracy and relevance
 - Built debugging workflows for LLM assistant failures across retrieval, context construction, tool calls, and final response generation.
 - Implemented Out-of-scope topic detection ML model for rejecting irrelevant questions, reducing user errors **by 30%**
- **Twitter** New York
Machine Learning Engineer II *Jun 2020 - Jan 2023*
 - Developed and maintained the inference service used by 500+ models across Twitter. Integrated TensorFlow Serving, KFServing/KServe, and NVIDIA Triton into Kubernetes-based serving workflows with CI/CD, monitoring, deployment automation, and model lifecycle tooling.
 - Integrated TensorFlow Serving, KFServing/KServe, and NVIDIA Triton into Kubernetes-based serving workflows with CI/CD, monitoring, deployment automation, and model lifecycle tooling.
 - Worked on model hotswapping functionality and reduced deployment cycle from **4 hours to 10 mins**
 - Integrated Twitter's custom data format into the serving stack for efficient high-throughput inference.
 - Improved model latencies by **upto 30%** via tuning tensorflow parallelism and server-side batching
- **HackerRank** Bangalore
Software Engineer *Jul 2016 - Jan 2018*
 - **Backend Engineer**: Built product and platform features for Ruby on Rails application serving 3M+ developers, including A/B testing, pub/sub feeds, email campaigns, and autocomplete services.
 - Implemented a competitive programming DSL that generated scaffolding code for 19 languages including Erlang, Go, Haskell and Swift

PUBLICATIONS

- **Text-to-SQL Domain Adaptation via Human-LLM Collaborative Data Annotation**: Published at the 30th International Conference on Intelligent User Interfaces. [Paper]
- **Evoschema: Towards Text-to-SQL Robustness against Schema Evolution**: Published at VLDB 2025. [Paper]

PROJECTS

- **CacheMoney (2026)**: Built a mini vLLM-style Python/PyTorch inference engine with paged KV cache, continuous batching, prefix caching, streaming API, and vLLM benchmark comparisons.
- **Huggingface Diffusers Contributor (2022)**: Added tests for attention modules and scheduling pipelines. [Github]
- **Techstars Blackstone Accelerator (2020)**: Received a seed grant for federated learning pilot using PySyft. [Blog]
- **Google Summer of Code (2016)**: Built Python framework to prototype routing algorithms like DFS/BFS, A* for electric vehicle range prediction. [Report]

EDUCATION

- **University of Texas at Dallas** Dallas, TX
Master of Science in Computer Science *July 2018 - May 2020*
- **Birla Institute of Technology and Science** Pilani, India
Msc.(Tech) Information Systems *Jul 2012 - May 2016*

PROGRAMMING SKILLS

- **Languages:** Python, Java, Scala, C++, Javascript, Go
- **ML / LLM:** PyTorch, TensorFlow, RAG, embeddings, LLM evals, Langfuse, batching, vLLM internals
- **Infra / Serving:** Triton, KServe/KFServing, Kubernetes, Docker, Spark, Kafka, Druid, PostgreSQL/vector search, CUDA basics