

# Jordan K. Mercer

## Senior Machine Learning Engineer | PhD Candidate (CS)

ML engineer with 8+ years building and deploying production ML systems, with a research track in large-scale retrieval and retrieval-augmented language models. Led distributed training infrastructure for billion-parameter models (9 days to 18 hours), shipped dense-retrieval systems serving 40M requests per month, and published first-author papers at NeurIPS and ICML.

✉ jordan.mercer[at]example[dot]com

🐙 github.com/jordan-k-merc

🌐 linkedin.com/in/jordan-k-merc

## Experience

### Senior Machine Learning Engineer

Novalytic Inc., San Francisco

2021 – Present

- Lead the large-language-model training platform: built a distributed PyTorch + FSDP training stack on Kubernetes that fine-tunes 7B–70B parameter models, taking a 13B model from a 9-day single-node run to 18 hours on a 32-GPU cluster.
- Designed a retrieval-augmented generation service serving 40M requests/month with a P99 latency of 180 ms, using vLLM for inference and a custom dense-retrieval backend with quantized embedding caches.
- Drove the shift from ad-hoc experiment scripts to a reproducible MLOps pipeline (Weights & Biases + DVC + Terraform), reducing experiment onboarding time for new hires from days to under an hour.
- Published 2 first-author papers on retrieval and long-context reasoning (NeurIPS 2024, ICML 2023) while shipping the same research into production.
- Mentored 4 ML engineers; led cross-team design reviews for model deployment, quantization, and inference efficiency.

### Machine Learning Research Engineer

Vertex Systems, Seattle

2019 – 2021

- Built the company's first dense-retrieval search backend, replacing a BM25 baseline and improving top-10 recall by 22% on enterprise document corpora.
- Prototyped an early retrieval-augmented QA system (predating the RAG acronym) that grounded answers against indexed knowledge bases; shipped to 3 enterprise customers.
- Designed and ran distributed training jobs on AWS (EKS + Spot instances), cutting monthly training spend by 35% while preserving throughput.

### Software Engineer

Cascadia Labs, Portland

2017 – 2019

- Built and maintained the backend data pipeline (Python + Airflow + Kafka) serving analytics for a SaaS product with 200k monthly active users.
- Implemented the company's first ML feature: a recommendation ranker that improved click-through rate by 11% in an online A/B test.
- Wrote the internal CI/CD bootstrap for service deployments (Docker + Terraform), adopted by 3 downstream teams.

## Education

2020 – 2025 (expected)

🎓 **Ph.D. in Computer Science** at Northern Cross University, USA.  
Thesis title: *Efficient Large-Scale Retrieval for Neural Search Systems*

2015 – 2017

🎓 **M.Sc. Machine Learning** at Westgate University of Technology, USA.  
Thesis: *Contrastive Representation Learning for Multilingual Sentence Embeddings.*

## Education (continued)

2011 – 2015

📖 **B.Sc. Computer Science** at Westgate University of Technology, USA.

## Technical Skills

|                  |                                                                                                                                                                                                                                                                                                                |
|------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Coding           | 📖 Python, C++ (CUDA), TypeScript, SQL, Bash, $\LaTeX$                                                                                                                                                                                                                                                          |
| Machine Learning | 📖 <b>Areas:</b> Large language models, Retrieval-augmented generation, Dense retrieval, Representation learning, Contrastive learning<br><b>Frameworks:</b> PyTorch, JAX, Hugging Face Transformers, vLLM, Triton, DeepSpeed, FSDP<br><b>Evaluation:</b> A/B testing, causal inference, offline metrics design |
| Infrastructure   | 📖 Docker, Kubernetes, Terraform, AWS (EKS, S3, Spot), GCP (GKE), Airflow, Kafka, Linux                                                                                                                                                                                                                         |
| MLOps            | 📖 Weights & Biases, DVC, MLflow, CI/CD (GitHub Actions, GitLab CI), distributed training on multi-node clusters                                                                                                                                                                                                |
| Languages        | 📖 English (native)<br>German (B2, Goethe-Zertifikat B2)                                                                                                                                                                                                                                                        |

## Research Publications

### Conference Proceedings

- 1 **J. K. Mercer**, L. Huang, and E. Rossi, “Adaptive retrieval-augmented language models for long-context reasoning,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- 2 **J. K. Mercer** and S. Lindqvist, “Efficient dense passage retrieval via quantized embedding caches,” in *International Conference on Machine Learning (ICML)*, 2023.
- 3 H. Chen, **J. K. Mercer**, and D. Navarro, “Contrastive alignment for multilingual sentence embeddings,” in *Association for Computational Linguistics (ACL)*, Association for Computational Linguistics, 2022.

### Under Review (Preprint available upon request)

- 1 **J. K. Mercer**, P. Raman, and T. Okafor, *Scaling dense retrieval with sparse attention priors*, 2025.

## Awards and Achievements

|      |                                                                                                                        |
|------|------------------------------------------------------------------------------------------------------------------------|
| 2024 | 📖 <b>NSF Graduate Research Fellowship</b> – competitive national fellowship for doctoral research in computer science. |
| 2023 | 📖 <b>MLPerf Training Benchmark – Open Division</b> , top-10 submission for distributed language-model training.        |
| 2022 | 📖 <b>Best Reviewer Award</b> , ACL Rolling Review cycle.                                                               |