

Diego Velasquez

Junior ML Engineer

Toronto, ON · diego.velasquez@gmail.com · +1 416-555-0137 · [linkedin.com/in/diegovelasquez](https://www.linkedin.com/in/diegovelasquez)

PROFILE SUMMARY

- **Junior ML Engineer** with **2 years** moving research prototypes into production at a **foundation-model lab**, across **retrieval and reranking models**, specializing in **fine-tuning pipelines**, **retrieval evaluation**, and **reproducible experiments**.
- Hands-on across primary language (**Python**), training framework (**PyTorch** and **PyTorch Lightning**), model hub (**Hugging Face Transformers**), experiment tracking (**Weights & Biases**), serving (**FastAPI**), and containers (**Docker**), with working knowledge of Ray and AWS SageMaker.
- Growing expertise in **embedding-model fine-tuning**, **retrieval evaluation** (recall@k, MRR, nDCG), and **data preprocessing tests**, using **versioned datasets** and **weekly eval reviews** to keep every training run **reproducible from a commit hash**.
- Works day to day with **research scientists**, **senior MLEs**, and **the platform team** under senior mentorship, joining paper reading groups, eval reviews, and release checks for 2 training pipelines and 3 evaluation pipelines.
- MSc thesis on **dense retrieval for long-context QA** built the useful half of the job: **can read a new paper on Monday and have a faithful baseline running by Friday**, with the eval harness to prove it.

TECHNICAL SKILLS

Languages:	Python, SQL, Bash, Git
Deep Learning:	PyTorch, PyTorch Lightning, Hugging Face Transformers, Hugging Face Datasets, embedding models
Fine-Tuning:	LoRA and full fine-tuning of embedding and reranker models, mixed precision, checkpointing
Evaluation:	recall@k, MRR, nDCG, eval harness jobs, pytest for preprocessing, model cards
Data & Tracking:	pandas, NumPy, Weights & Biases, DVC (basics), versioned datasets
Serving & Infra:	FastAPI, Docker, AWS S3, AWS SageMaker (basics), Ray (basics)
LLM Tooling:	RAG pipelines, tokenizers, prompt and eval datasets, LangChain (basics)

EDUCATION

University of Toronto	M.Sc. in Computer Science	Toronto, ON · Sep 2020 - Jun 2022
University of British Columbia	B.Sc. in Computer Science	Vancouver, BC · Sep 2016 - May 2020

WORK EXPERIENCE

Cohere	Junior ML Engineer	Toronto, ON · Sep 2023 - Present
• Built and maintain 2 fine-tuning pipelines in PyTorch Lightning for retrieval embedding models, covering data loaders, checkpointing, and Weights & Biases logging, so a run restarts from any checkpoint in under 5 minutes. • Own the recall@k and nDCG dashboards for retrieval and reranker models inside the internal eval harness, adding 6 evaluation jobs that 2 senior MLEs review every week before a release. • Helped build 3 retrieval evaluation pipelines on Hugging Face Datasets for long-context QA, multilingual recall, and citation grounding, growing the eval set from 4k to 31k labelled queries. • Wrote pytest suites for data preprocessing across 4 tokenization paths, catching 2 silent encoding bugs before they reached a training run and adding the checks to CI. • Automated model-card generation from training metadata in Python under senior review, cutting per-release prep from 4 hours to under 30 minutes.		
Layer 6 AI	Research Intern, then Junior ML Engineer	Toronto, ON · Jun 2022 - Aug 2023
• Reproduced 4 published sequential recommendation baselines in PyTorch, including data loaders, training loops, and offline evaluation on MovieLens-25M and proprietary banking sequences. • Built a scikit-learn baseline harness that 3 research scientists ran as a sanity check before every deep-learning experiment, saving a day of GPU time per failed idea. • Wrapped a transformer reranker in a FastAPI endpoint with input validation, packaged it in Docker, and deployed it to a staging EKS cluster for the product team to try.		