Rafid K.

778-918-1785
rafidka@gmail.com
rafidka
rafidka
rafid.ai

About Me

- **Machine Learning Engineer** specializing in **search, retrieval, and large language models (LLMs)**.
- Currently at **Cohere**, building and improving **Compass**, an enterprise-grade retrieval platform powering RAG pipelines.
- Hands-on experience with **ML model development**, **information retrieval evaluation**, and **LLM-as-a-judge** techniques.
- Research + product contributions across **multi-modal models**, **synthetic data generation**, and **retrieval pipeline optimization**.
- Published **2 papers** on multi-modal learning and deployed ML-powered systems at scale.
- Deep learning practitioner with extensive experience in **PyTorch**.
- Proficient in ML-centric programming languages including **Python**, with additional experience in **C++** for high-performance components.
- Passionate about bridging **machine learning research and real-world products**, with an open-source track record (e.g., `microtorch`, `Alkitab`, `pat-cli`).

Experience

Member of Technical Staff @ Cohere

October 2024 – Present

Python, Machine Learning, Synthetic Data, IR Evaluation, LLM-as-a-Judge, Data Ingestion

Vancouver, Remote

- Designed and implemented the **Compass end-to-end evaluation framework**, including retry mechanisms, checkpointing, and integration with parser/ingestion; enabled scalable IR experiments with LLM-based evaluation.
- Shipped multiple integrations: **web2compass** for RAG-enabled documentation, MCP tooling for North, metadata enrichment for TOC-based PDFs, and **code2compass** for semantic search over source code.
- Improved **Compass Parser** efficiency for vision-language models, reducing memory usage by 50% in image-to-markdown tasks.
- Led development of **Compass SDK V2** with async support, addressing major customer needs and improving ingestion pipelines.
- Implemented a **Postgres-based job queue** to replace Celery for long-running jobs, improving reliability and testability.
- Built the **Compass Asset Service**, enabling storage and retrieval of assets (e.g., PDF page rendering) to support async ingestion, parsing, and downstream RAG pipelines.
- Contributed to the open-source ecosystem by adding **Cohere support to Pydantic-AI**, enabling structured AI workflows with Cohere models.

Senior Software Development Engineer @ AWS

November 2018 – October 2024

Java, Python, TypeScript, Golang, ECS, Lambda, Step Functions, DynamoDB, RDS, CloudWatch, CloudFormation, CDK

Vancouver, BC

- Founding engineer on **Amazon Managed Apache Airflow (MWAA)**, leading design and launch at re:Invent 2020; scaled the service to 15+ global regions.
- Led the modernization of the **Distributed Job Scheduler** backend using Elasticsearch, powering large-scale scheduling across Amazon.

- Mentored and grew engineers to mid- and senior-level roles.
- Drove **AI-Ops initiatives**, implementing automated health monitoring, event filtering, and recovery systems to improve production reliability.
- Championed open source efforts, launching the **amazon-mwaa-docker-images** repository on GitHub.

Software Development Engineer, Mobile Shopping @ Amazon
April 2017 – November 2018

Java, Android, AWS Cloud — *Vancouver, BC*

- Delivered platform services for the Amazon Mobile Shopping App, including analytics and secure customer migration infrastructure used by millions of retail customers.

Senior Software Engineer, FactSet Europe Limited
August 2011 – April 2017

C#, Python, JavaScript, MS SQL Server, IIS, Elasticsearch — *London, United Kingdom*

- Designed and built a **workflow orchestration platform** for financial reporting, including backend data storage and execution engine.
- Implemented a **real-time alerting system** and integrated Elasticsearch to power workflow search and monitoring.
- Collaborated with global teams to deliver scalable, client-facing financial software.

Web Developer, Ninua, Inc.
September 2008 – August 2009

PHP, Python, MySQL, Google App Engine — *Amman, Jordan*

- Contributed to **NetworkedBlogs**, a social media app with hundreds of thousands of users; optimized infrastructure and migrated services to Google App Engine.

Team Programmer, LEAD Technologies, Inc.
September 2006 – September 2008

C++, .NET Framework, Imaging SDKs — *Amman, Jordan*

- Developed and maintained imaging modules, including reverse engineering of AutoCAD DWG/DXF file formats.

Publications

Alexa Visual Item Selection (AVIS) Dataset, *Amazon Computer Vision Conference*
2023

- Created a comprehensive dataset containing visual-attribute-centric interactions, including user utterances, scene and product images, and associated metadata. Collected data via Amazon Mechanical Turk and released it to the internal Amazon community. Benchmarked the performance of the open-source multi-modal model ALBEF, demonstrating a 19% relative accuracy improvement for visual item selection after fine-tuning. Designed a low-latency system architecture for training, deployment, and inference of the model.

Visual Item Selection with Voice Assistants, *ACM Web Conference*
2023

https://www.amazon.science/publications/visual-item-selection-with-voice-assistants

- Developed a multi-modal visual shopping experience for Amazon Alexa, enabling natural language interactions with visual content on screen. Designed and implemented a system architecture for model fine-tuning and deployment on Amazon Echo devices with screens. Achieved high accuracy in the "Visual Item Selection" task using open-source models like CLIP and ALBEF, with significant improvements through fine-tuning. Launched the technology as an Alexa Skill, available in the Alexa Skills store.

Advanced Composition in Virtual Camera Control, *International Symposium on Smart Graphics*
2011

https://link.springer.com/chapter/10.1007/978-3-642-22571-0_2

- Developed an advanced camera control system for 3D content, incorporating aesthetic rules and conventions from visual media like film and television. Focused on automating the composition process by precisely positioning elements within a shot. Selected and specified rating functions for various compositional rules, using optimization techniques to determine the best camera configuration. Implemented image processing methods to evaluate rule satisfaction in real-time, enhancing the overall quality of 3D content rendering.

Open Source Projects

microtorch
2024

Python, PyTorch https://github.com/rafidka/microtorch

- A minimal deep learning framework built from scratch in Python, showcasing the fundamentals of **autograd, optimizers, and neural network modules**. Intended as both a learning tool and a lightweight framework for experimentation.

neuroscout
2024

Python, LLMs, NLP https://github.com/rafidka/neuroscout

- An **LLM-powered tool** to analyze **NeurIPS paper abstracts** and help researchers efficiently discover relevant topics for poster sessions.

pydantic-ai
2024

Python, Pydantic, LLMs https://github.com/pydantic/pydantic-ai

- Contributed to the popular **pydantic-ai** library by adding **Cohere LLM support**, enabling users to seamlessly use Cohere models in structured AI pipelines.

amazon-mwaa-docker-images
2024

Python, Shell, Docker, Docker Compose https://github.com/aws/amazon-mwaa-docker-images

- I single-handedly wrote the Docker Images used by Amazon MWAA to run Apache Airflow 2.9 and beyond. The code was written from the ground up, employing modern programming methodologies and best practices, and putting great emphasis on code quality and maintainability. The code currently serves thousands of Amazon MWAA environments.

image-search
2023

Python, PyTorch, CLIP https://github.com/rafidka/image-search-with-human-language

- An innovative application that enables users to search for images using natural language queries. Employed a multi-modal language-vision model (CLIP) to combine the embedding spaces of images and text, which enable users to describe the images they are looking for like they are describing it to another human being. Contributed to the project's development, documentation, and community engagement to refine features and improve overall functionality.

logc
2023

Python, NLP, LLM https://github.com/rafidka/logc

- A shell tool for clustering logs based on the textual content of the log. Useful to quickly discover patterns in large amount of system logs. It employs clustering algorithms from Machine Learning, along with a Large Language Model (BERT) to intelligently group similar logs together.

github-downloader
2022

TypeScript https://github.com/rafidka/github-downloader

- An open-source tool designed for downloading repositories from GitHub. Implemented features such as automated retry, repository cleaning, and speedy download to reduce the overall time it takes to download a large number of repositories. The tool employs a user-friendly command-line with self-help to make it easy for the user to use the tool without a lot of manual reading.

boto3async
2021

Python, AWS https://github.com/rafidka/boto3async

- An asynchronous wrapper for AWS's boto3 library, automating the creation of async methods. This solution addresses boto3's lack of native async support, allowing for non-blocking API calls, which enhances scalability and efficiency of cloud-based applications by enabling high-concurrency operations. It generates the clients and methods automatically, hence when boto3 supports new services or add add new APIs to existing clients, support for them get automatically added to boto3async.

Alusus Programming Language
2021

C++, LLVM https://github.com/Alusus/Alusus

- I was one of the lead developers of Alusus programming language.
- I single-handedly wrote the initial code generation component, which is responsible for taking an Abstract Syntax Tree (AST) and generating machine code.
- The code is written in C++ and employs the LLVM Compiler Infrastructure.

Education

Newcastle University
2009 – 2011

Master of Philosophy, Computer Science *Newcastle, UK*

- A 2-year research program in Computer Science
- Thesis: *Advanced Composition in Virtual Camera Control*, focusing on automating virtual camera control based on cinematographic composition rule using Mathematical Optimization techniques and Computer Graphics techniques.

Baghdad University
2002 – 2006

Bachelor of Science, Computer Engineering *Baghdad, Iraq*

- A 4-year Bachelor program in Computer Engineering.
- Graduation Project: *Computer Algebra System*. Implemented a Computer Algebra system capable of solving mathematical equations symbolically.