Leopoldo Garcia Vargas

1063 Rue Rielle, Montréal, Québec. QC H4G 2T2

Mobile: +1 (514) 966-0739 **Email**: lk13.dev@gmail.com **Social**: LinkedIn
Citizenships: Canadian and Mexican

Cover Letter

My educational journey has been defined by a strong foundation in both Pure Mathematics, with a five-year Bachelor's degree featuring a supplementary thesis project, and Data Science, where I specialized in Machine Learning and Artificial Intelligence engineering.

My passion lies in crafting scalable software solutions tailored for the deployment of Machine Learning models in production environments. In today's landscape, the design and engineering of the software components and the infrastructure supporting Machine Learning systems have taken precedence over the algorithms themselves. This shift is primarily attributed to the critical need to monitor the system, assess its performance, and gain insights into the quality of prediction outputs and incoming data. This invaluable feedback loop enables continuous enhancements at various levels, be it algorithmic refinement, fine-tuning of heuristics, or the optimization of supporting software components.

My journey spans over 5 years in the realm of Machine Learning, during which I've had the chance to work in all the roles of the modern Data/ML stack, encompassing Data Scientist, ML Developer/Engineer, and ML Operations Developer. I've also collaborated closely with Data Engineers, contributing to the creation of real-time data ingestion pipelines.

Prior to venturing into the ML/Data landscape, I worked as a professional iOS developer for over 6 years. Additionally, I've embarked on personal startup projects where I've taken charge of conceiving, creating, and deploying backend services and web clients. This diverse background equips me with a holistic perspective and a unique skill set that I'm eager to bring to the table, delivering innovative and impactful solutions in the world of Machine Learning.

Technical Skills

Programming languages	Python 3, Scala (Basics), Java, Objective-C, SQL
Data science	Numpy, Pandas, Scikit-Learn, Keras, PyTorch.
NLP	Spacy, LangChain, NLTK, OpenAI SDK.
Databases	AWS Redshift, Snowflake, AWS DynamoDB.
Microservice technologies	Docker, Kubernetes, Docker Compose.
Infrastructure as code	Terraform
Dependency management	Poetry, Conda, Pip, Maven, NPM.

Version control and CI/CD	Git.Github actions, Jenkins
Web frameworks	Fast API, Flask, Node.js

Languages

Completely fluent in English, French and Spanish.

Experience

Senior Machine Learning Engineer: Generative Q&A, Coveo.

June 2023 – Present. Montreal, Quebec.

I had the privilege of collaborating within a dynamic and fast paced, cross-functional team. This team comprised data scientists, backend developers, UI designers, product specialists, and fellow machine learning engineers. Our collective mission was to give life to an innovative generative question-answering system based on retrieval augmented generation (RAG).

The system we envisioned was meticulously designed to adapt to the unique context of each client. This vision necessitated careful architectural planning to ensure that our creation could be both reusable and cost-effective. In addition, we put a strong emphasis on making it easily monitorable to provide insights into its performance.

A crucial aspect of our system was its ability to efficiently process the documentation specific to each client. This documentation served as the context for our generative answers. By leveraging this context, we aimed to provide highly relevant and tailored responses to address the unique needs of each client. Our combined efforts resulted in a versatile and robust solution, seamlessly tailored to each client's context, and capable of delivering relevant answers.

Main responsibilities:

- Participate in design and architecture discussion with other senior members to effectively conceive a reusable, monitorable and cost effective system.
- Responsible for the observability and monitoring of the training pipeline as well as the serving component. Tools for monitoring: Open search logs, Sentry and OpenTelemetry (with Honeycomb frontend).
- Responsible for the implementation of the API layer to guarantee rapid responses and observability at serving time.
- Responsible for implementing the streaming components to be able to stream answers back to the UI in order to provide a feeling of fluency to the user asking questions.
- Participated in research discussions by reviewing and providing feedback on prompt engineering experiments aimed at uncovering the better ways to query OpenAI to generate tailored answers given the context of each client.

Examples of the system described above can be found here: XERO and Coveo. Go ahead and ask some questions!

Senior Machine Learning Engineer: ML-Operations, Coveo.

September 2021 – May 2023, Quebec.

Key accomplishments:

- Designed a data-driven framework specifically targeted to comprehend and uncover patterns in the development and deployment requirements of multiple machine learning teams. The gathered data was transformed into a scoring rubric, enabling us to assess the MLOps tools available in the market and make an informed decision on selecting a suitable replacement for Coveo's legacy infrastructure.
- Design and oversee the implementation of the chosen MLOps tool (Metaflow) in the Coveo infrastructure.

- Coached several team members with diverse backgrounds, helping them gain proficiency in software best practices as well as Machine Learning concepts and workflows.
- Serve as a conduit through which science seamlessly interacts with software.

Applied Scientist (NLP), Coveo.

November 2019 – September 2021. Montreal, Quebec.

Question Answering / Smart Snippets

Architected and developed the first version of a Question Answering Product from conception to deployment. The goal of this product was to grant search pages the capacity to present a relevant answer directly at the top of the search page when a user performs a search. The scope of this V1 was document snippet retrieval, ie: the system will fetch an preprocessed snippet so that the user can see it at the top of the search page providing relevancy to the client's content. The project was designed under a microservice approach with three main components: Data processing, Model Building and Model Serving, each one of them as Dockerized applications running independently.

The main components of this project were:

- Data processing: Extract documents from the index and process them into snippets into DynamoDB
- Model Building: Fetch the required data and train the model. Model is stored and made accessible to the next step through S3
- Model Serving: Fetch and load the model trained in the previous step and make it ready for inference

Deep Learning Infrastructure (DELI)

Architected, developed and led the team to the implementation of a brand new Deep Learning Python/Microservice based infrastructure. The objective of this infrastructure was to provide the other ML teams with the possibility of training and deploying their DL/ML models without having to build the dedicated software components. DELI would provide already made base Docker containers and libraries to facilitate the training and inference of new models.

Data Scientist, Plusgrade.

September 2018 – August 2019. Montreal, Quebec.

Worked with a team of data engineers to design and implement real time data pipelines using the AWS ecosystem. The pipelines' objective was to ingest business events as they were processed by the company's applications, transform them as required for later consumption and store them into a data warehouse. Once the data was stored as desired it was to be used for dashboarding and BI purposes.

The main components of the pipelines were:

- Event publisher framework; built to instrument the company's applications so events could be published to a data stream.
- Data stream; AWS Kinesis was used as a stream in order for the publisher framework to publish events to. The stream would then deliver the raw events to AWS S3.
- Event transformer; AWS Lambda function that would be triggered by the raw events stored on S3 (by Kinesis) and accordingly transform them and insert as desired into the data warehouse.
- Data warehouse; AWS Redshift DBMS, columnar storage which would be updated by

the event transformer and store our tables ready for OLAP operations (dashboards and BI).
- Dashboard client; Used Apache Superset to be the client which will fetch the tables from the data warehouse and display the desired dashboards.

Developed an anomaly detection model that was more accurate than the company's current implementation. The model was built using Poisson's law over a time series of events.

Machine Learning Engineer, Playster.
January 2018 – July 2018. Montreal, Quebec.

Introduced the use of Machine learning techniques to build predictive tools for improving data business analytics and for maximizing the company's revenue. Worked closely with BI manager to design and implement concrete pipelines from conception to deployment.

Key accomplishments*:*

- Built and deployed anomaly detection model based on users' behavior in order to identify potential 'super users' within 10 days of interaction with our digital products. This model ended up saving the company money by identifying super users and allowing the segmentation team to take action on time. It was deployed as a dockerized application on an AWS instance to be accessed as a service. I developed an API to be able to communicate with and monitor the model.
- Built and deployed affiliate profitability classifier model ensemble to detect fraudulent affiliates, which allowed us to pinpoint the desirable affiliates, in order to take the necessary actions to optimize acquisition costs. The model was delivered to the BAs and BIs to be executed as a script to create predictions whenever enough affiliate data was collected.

Senior iOS Developer, Playster.
October 2016 – January 2018. Montreal, Quebec.

In charge of the core engine of the app; database, download ecosystem, web services and media players.

- Proposed and implemented different architectural patterns and best practices to improve the health of the app's code in general and the stability of the system.

- Rebuilt completely 3 out of 4 media players (book reader, audiobook player, and movie player) to provide a better user experience and to optimize the performance of the application.

- Playster app.

Mobile Developer, Yellow Pages Group.
June 2014 – July 2016. Montreal, Quebec.

Worked on two of the main mobile applications; YP Shopwise and YP Dine, for which I had the chance to work with two different teams of developers, designers, POs, growth hackers, and QAs.

Developed and maintained these products working with different native frameworks and APIs

while using advanced techniques in Objective-C. Worked closely with POs and designers to create and invent new product features through powerful and responsive user interfaces in order to cover our users' needs.

Both Apps have been featured several times in different categories by Apple.

- YP Shopwise app & YP Dine app.

---

iOS Developer & Founder, HaiKu Development & Technology.
Mexico-Canada — 2013-2014.

Founded a small consulting startup with two partners. We developed and sold mobile applications for both iOS and Android devices to companies in Mexico and Canada. Led the team of mobile developers and designers to create simple, reliable and efficient mobile apps.

Senior iOS Developer, Ascenzo Consulting S.A. de C.V.
June 2013 – March 2014. Mexico City Area, Mexico

Working with a team of several developers and UX Designers to build a mobile application for Servicio de Administración Tributaria (SAT) http://www.sat.gob.mx/ . Which manages the taxing and billing service in Mexico.

App Store iOS Developer, Apple App Store.
Mexico & Canada — 2009-2014

Developed an encryption app (Cifra app) which helped protect user's messaging data by using a custom created base64 ciphering algorithm.

Developed an app to illustrate the results of my graduation thesis (Finite fields app). It displays the elements of a certain type of mathematical algebraic set and allows the user to operate such elements.

---

iOS SDK vlogger, Youtube
Mexico-Canada — 2010-2014

Back in 2010 I started a Youtube channel (in Spanish) in which I taught Object Oriented programing using Objective-C and different methods to develop iOS apps.

- Link to channel: http://www.youtube.com/user/LKronecker

Education

Mathematician
Universidad Nacional Autónoma de México — 2007-2013

My degree takes the student through all three main areas of Mathematics and the branches that connect them. It is based on giving the student the chance to understand the ideas behind mathematical theorems and postulates.

Publications:

- Thesis project: Hash Function Families

Machine Learning Specialization

University of Washington — 2016-2017

The specialization included the following courses:

- Machine learning Foundations: A case study approach
- Machine learning: Regression
- Machine learning: Classification
- Machine learning: Clustering and retrieval

More information : https://www.coursera.org/specializations/machine-learning

Artificial Intelligence Nanodegree

Udacity — 2017-2018

- Introduction to artificial Intelligence
- Deep Learning and Applications
- Natural Language Processing Specialization

More information : https://www.udacity.com/course/ai-artificial-intelligence-nanodegree--nd898

Interests

Reading scientific research papers and books, studying the nature of the mind and reality through meditation and juggling.

As a side project I teach meditation and yoga: http://timenow.ca/

Repository

https://github.com/lkronecker13

References

Will be provided upon request