Chris Brousseau

(801) 919-6319 | chrisbrousseau304+jobs@gmail.com | Provo, UT

Experience

JPMorganChase

VP - Staff Machine Learning Engineer | 12/2023 - Present
I built a platform for Data Scientists to be able to deploy models easier while allowing us to quicker make those models usable for various production environments. My main evangelism was using an LLM to approximate a complex function (such as data extraction), then using a combination of knowledge distillation and clever sampling to teach a smaller model to replicate the large model's performance.
Partial Stack - Terraform, AWS, Datadog, LlamaCPP, vLLM, DSPy, Outlines, Guidance, Transformers, PyTorch, ONNX
This saves money and increases ROI in every direction. We save money on engineering time because the engineers are enabled with repeatable, opinionated, and streamlined systems within the platform to work quicker. We also save a large percentage (between 30 and 60%) on compute, by paying to finetune much smaller models that replicate the LLM results without needing to pay for a hypercluster to train the larger model.

Utah State University

Adjunct Professor | 01/2025 - Present
I teach a special topics course on LLMs in Production for graduate students in the Math and CS department of USU.

Manning Publications

Author | 02/2023 - 12/2024
I co-authored the book LLMs in Production with Matthew Sharp. We wrote this book based on our experiences at large multinational corporations taking LLMs and making successful products out of them. As opposed to other books in the space our book focuses on actual code and practical applications rather than theory and supposed business best practices. We sold more than 1700 copies in early access (the physical copy is done but not out at the time of writing), and the GitHub repository with 48 stars containing all the code from the book can be found here: https://github.com/IMJONEZZ/LLMs-in-Production.

Nerd United

Machine Learning Engineer | 07/2023 - 11/2023
I used my experience as a Data Scientist to inform building an AutoML platform to automate the process of training and evaluating models, all the way from ingesting and transforming data on

to deployment. I also lead the company's DevRel initiatives, using my personal brand to add validity to company events like teaching people to train and use diffusion models for text-to-image purposes.
Partial Stack: AWS, Sagemaker, Airflow, LlamaCPP, Ollama, Transformers, Diffusers, Automatic1111, bitsandbytes, xformers, DeepSpeed, Modelbit

Mastercard

Lead Data Scientist - International NLP | 11/2021 - 07/2023
I trained and deployed models that created a language and bank-agnostic categorization, entity resolution, and comparison space.
Partial Stack: Transformers, Pytorch, Optuna, Jupyter, AWS, Sagemaker, EC2, ECR, Lambda, Docker
During the building of this product, we secured 2 patents, and based on the tech we built the sales team was able to secure a multi-million dollar deal with a credit reporting agency for Open Banking with banks using MC as a payment processor.

Projects

National Basketball Association

I created a system for real-time multilingual interpretation and voice cloning for use by national sports broadcasting. I completed this as a 1-month contract. This system enabled audiences to hear commentators like Shaquille O'Neal and Charles Barkely in near-perfect French, keeping continuous delay under 6 seconds (total live broadcasting buffer). I also completed the drive for on-prem self-hosting of all services to eliminate ChatGPT or DeepL latency.
Partial Stack: Redis, Bespoke Translation Agent, Speculative Decoding, Whisper v3 large, Bespoke Voice Cloning Text-to Speech Agent

Education

Brigham Young University

Linguistics | 04/2020
Graduating with a BA in Linguistics and a minor in Localization while concurrently serving as an Adjunct Professor for Czech and Slovak 330, I pursued an intensive course load at the intersection of language and technology. An immersive study abroad in Moscow solidified my Russian and localization technology experience, but it was a graduate-level seminar (Python and Machine Translation), that fundamentally redirected my academic focus. My persistent argument for the relevance of computational courses to my degree culminated in this pivotal experience, which introduced me from first principles to how linguistic analysis can directly inform AI models. This foundation, grounded in my undergraduate studies of semantics and web information technologies, is what I really want to advance within this program.
I was honored as a Presidential Leadership Committee member in 2019 based on my achievements for the school when in Russia.