Aaron Pham

linkedin.com/in/aarnphm aarnphm.xyz +1(415)-535-8841

Experience

BentoML

Machine Learning Engineer

05/2023 - Present
Remote | Toronto, ON
05/2022 - 04/2023
On-site | San Francisco, CA

Authored and lead *OpenLLM* - Run any open-source LLMs, such as Llama, Mistral, as OpenAI compatible API endpoint in the cloud.
Implemented state-of-the-art inference optimisation, such as continuous batching, quantisation, sliding context windows using *PyTorch*, optimised attention kernels in v0.4.
Working on vLLM's structured decoding. LLM Inference on BentoCloud

Technologies: Python, PyTorch, OpenAI Triton, BentoML, Transformers, vLLM

Software Engineer

08/2021 - 04/2022
Remote | Hamilton, ON
01 - 08/2021
Remote | Hanoi, Vietnam

Maintained and contributed features improvement to HuggingFace's ecosystem, including *optimum, transformers, accelerate, PEFT*. Integrated *NVIDIA Triton Inference Server* with BentoML.
Designed and implemented ML framework runners support in BentoML 1.0.
Added Docker image releases management, provides supports for multiple Linux OS.
Proposed an improved CI system. Added integration tests and **Python structured typing** to BentoML.
Improved internal dynamic batching algorithms that increase general throughput by **2.2x**.
Mentored Fall 2021 Batch MLH Fellows and added label supports on BentoCloud in Go.

Technologies: Python, Pyright, GitHub Action, Bash, Go, AWS EC2, PyTorch,

Project

avante.nvim

github
maintainer
09/2024 - Present

A neovim plugin designed to emulate the behaviour of the Cursor AI IDE.
Implemented bounding UI popover to improve QOL (#29)
Added support for lazy setup for better load time improvement (#14)
Added Rust crates for `.avanterules` templates (#466)
Working on tool-use and function-calling support for multiple LLMs providers.

Technologies: Rust, Lua, neovim, tokio

Quartz

github and website
maintainer
01/2024 - Present

A fast, batteries-included *static-site generator* that transforms Markdown content into fully functional websites.
Improved performance of graph interaction with HTML5 Canvas (#1328).
Added support for PDF in popover modal (#913). Implemented font-fetching before runtime for faster build-time (#817)
Enhanced search experience with telescope-style layout (#722, #774, #782).

Technologies: preact, Typescript, SASS, unified.js

OpenLLM

github
mle at bentoml
06/2023 - Present

OpenLLM allows developers to run any open-source LLMs (Llama 3.3, Qwen3, Phi4 and more) or **custom models** as **OpenAI-compatible APIs** with a single command.
It features a built-in chat UI, state-of-the-art inference backends, and a simplified workflow for creating enterprise-grade cloud deployment with Docker, Kubernetes, and BentoCloud.
Include vertical support with ML ecosystem, including LangChain, LlamaIndex, etc.

Technologies: Python, vLLM, NextJS

incogni.to

demo and posts
New Build' 24

An event platform that curates for those yearning to be seen for who they are, not what they can "*sell*"
Implemented a RAG pipeline for *recommendation system* based on users preferences and interests, with command-r-plus-08-2024, deployed with vLLM and BentoML.
Added **query-based semantic search** with Cohere Rerank. Implemented dashboard with `shadcn/ui` and Next.js

Technologies: Next.js, vLLM, BentoML, shadcn/ui

BentoML

github
swe at bentoml
05/2022 - 04/2023

BentoML is a Python library for building online serving systems optimised for AI apps and model inference.
Build **Model Inference APIs**, Job queues, LLM apps, Multi-model pipelines, and more!
BentoML automatically generates Docker images, ensure reproducibility, and streamline deployment.
Leverage built-in optimisation such as *dynamic batching, model parallelism, multi-stage pipeline, multi-model inference graph orchestration.*

Technologies: Python, ASGI, NumPy, Pydantic, HTTPX, OpenTelemetry

onw

github
hack the north 2021 finalist

A real-time navigation tool that harnesses the power of community to make community safer.
Trained and fine-tuned a **Gaussian Mixture Model** to find the *safest* path between different locations, trained on past assault data provided by Toronto Police Department from 2015 onwards.
Integrated with **Google Maps API** to show heat maps of avoided zone.
Implemented designs from Figma in React Native, Expo, AWS Fargate

Technologies: AWS Fargate, Figma, React Native, TypeScript, Apache Spark MLlib, GraphQL, Expo

Education

McMaster University

graduating 12/2025

Bachelor of Engineering and Management Software Engineering Co-op
Mentor/Judge at DeltaHacks XI