Sriram Vadlamani

AI engineer / Developer

Paris, France

Github: sriram-vadlamani ↗

Email: sriram.vadlamani@proton.me ↗

Phone number: +33 7 55 31 20 13

SUMMARY

An engineer with demonstrable ability to develop and deploy artificial intelligence models with cutting-edge generative technologies and make AI accessible for learning and use for the growing needs of the company.

EDUCATION

Msc. Data Science
2022 — 2023 | Openclassrooms (with Centrale Supélec)
La Rochelle, France

Bsc. Computer Science
2018 — 2021 | EPITA
Paris, France

Major: Data structures, algorithms, math and system architecture

Overall grade: 3.5

PERSONAL PROJECTS

Ollamanet
A Load balancer written in `Go` for concurrent LLM inference with Ollama.

OSAIL (OpenSource AI leaderboard)
A local leaderboard for evaluating Large Language Models using Elo rating, written in Golang / templ / HTMX.

Envy
A small CLI tool to encrypt and manage .env files in a repository.

Gemmatino
Finetuning gemma-2b instruct to generate movie plot summaires.

SKILLS

REST/gRPC · Databricks · Azure · Go · Pytorch · Prompt Engineering · NextJS · CI/CD · Tableau · AWS · Tensorflow · MongoDB · NoSQL · MySQL · EDA

EXPERIENCE

Principal AI engineer
July 2024 — September 2024 | Analysisly (Freelance)

› Providing consultancy, and architecture for a generative AI solution to extract information from audio transcriptions to help company get to a pre-seed funding round.
› Implementing a backend for audio transcriptions and information extraction on Azure cloud.

NLP engineer
Dec 2023 — June 2024 | Ryte Corporation

› Analyze patient reviews to model topics and sentiments using state-of-the-art NLP and generative NLP models. Generate review summaries using the most recent LLMs.
› Evaluating the performance of LLMs for generative text and building metrics, monitoring and ingestion pipelines.
› Building and deploying a RAG on Azure Databricks using Python frameworks and custom implementations for query generation, integrating a frontend with React and NextJS. Code reviews and maintainance of python repositories and packages.

NLP engineer
Jul 2022 — Nov 2023 | Raccourci Agency

› Analyze user reviews to model topics and sentiment using state-of-the-art NLP and generative NLP models. Training and deploying text models with CI/CD.
› Develop a temporal application from scratch with Go/Golang and Python. Containerization of AI microservices with temporal workflows. Finetuning encoder/ classification models for better performance in french, by vocabulary injection for masked LMs.

Data analyst
Sep 2021 — Mar 2022 | Freelance

› Demonstrated expertise in understanding data sources, KPIs and creating dashboards with Tableau. Manage resources to make dashboards minimal and deployable.
› Provided customized solutions to clients, including data-driven solutions and data modeling.

HOBBIES AND INTERESTS

Guitar · Chess · Films · Screenwriting

LANGUAGES

English · French · Hindi · Telugu