





ADITYA KHEDEKAR

Pune, Maharashtra

☎ +91-9422395205 ✉ adityakhedekar98906@gmail.com  [Linkedin Profile](#)  [twitter](#)  [Github](#)  [Leetcode](#)

Education

Dr.Dy Patil College of Technology, Pune, Maharashtra

Aug. 2018 – Jul 2022

B.E in Electronics and Telecommunication

CGPA: 8.0

Projects

Cpu-Scheduling-Deep-Reinforcement-learning | [Source Code](#)

Oct 2024

- Developed a CPU scheduling model using **Proximal Policy Optimization (PPO)** within a **custom OpenAI Gym environment** to simulate process arrivals and executions. The model optimized process **turnaround time**, achieving significant improvements over traditional **round-robin** scheduling algorithms
- Employed a feed-forward actor-critic architecture within the **PPO algorithm**, where the actor-network assigns process priorities using a **multivariate probability distribution**, and the critic network evaluates state values, ensuring optimal policy learning for minimizing turnaround time.
- Implemented **Generalized Advantage Estimation (GAE)** measures how much better or worse taking a particular action is compared to the average, ensuring robust **temporal-difference** learning Integrated with a **clipped surrogate objective** This ensures that the new policy doesn't deviate too far from the old policy, by clipping the probability ratio between the new and old policies, ensuring stable learning and efficient convergence.

Shakespear-LLM-GPT | [Source Code](#)

Sep 2024

- Implemented a **Transformer-based Bigram model** using **PyTorch** for **next-character prediction** with a **decoder-only architecture**, optimized for **autoregressive text generation**, similar to models like **GPT** (Generative Pre-trained Transformer).
- The model has a **block size of 256** with **6 parallel attention heads**. Each block consists of a **multi-head self-attention layer** that captures different types of relationships between characters and subsequently combines them to create **richer text representations**.
- Multinomial sampling** is used instead of **top-k** or **nucleus sampling** for diverse generations. **Masked self-attention** is employed to prevent tokens from attending to future positions, ensuring that each character only **attends to preceding characters**. predicting the next character based solely on the past.

Experience

InTouchAPP

May 2024 – May 2024

AI-Engineer (Consultant)

(Pune) Remote

- Designed and developed an **end-to-end architecture** to integrate an advanced **retrieval-augmented generation (RAG)** feature as part of an experiment to connect **private data** with a **large language model (LLM)**
- I worked on a subset of **sales data** to implement a **natural language Q/A bot** for managers and the CEO and made key decisions on the choice of **database (Qdrant DB)**, **LLM (GPT-4)**, and **embedding model (OpenAI's text-embedding-ada model)** with respect to **Self-hosting**, **Cost**, **Scalability**, **reliability**, and **model evaluations**
- Key features I implemented** include **query translation**, **reranking**, **advanced message-based chunking**, and improvements to **accuracy**, achieving approximately **90 percent**

Burni Consulting

Jul 2024 – Jul 2024

AI-Engineer (Consultant)

(Spain) Remote

- Worked on **ideation and designing system architecture**, while researching various **speaker diarization techniques** to separate audio based on individual speakers in the recording.
- Complied with the EU's **Data Protection Act (GDPR)** by implementing **Whisper.rn on-device**, eliminating the need for cloud data storage.
- Researched **Medical-LLM** for summarization and drug suggestion, and learned how to **quantize models** and deploy them to edge devices using **Qualcomm AI Hub**.

Technical Skills

Languages: Python, JavaScript, Typescript,

Databases: MongoDB, Firebase, Supabase, Redis, Qdrant, Pinecone, Neo4j, Drizzle

Front-End: React.js, Next.js, Redux/Redux-toolkit, React-router, TailwindCss

Back-end: Express.js, Socket.Io, Pub/Sub, Kafka, FastApi, Django, Pydantic, Puppeteer, Celery, Rabbitmq, Webhook

AI/ML: Pytorch, Numpy, Matplotlib, Langchain, HuggingFace, Open-AI-API, Fine-tuning, OCR, Weights & Biases

Devops: Kubernetes, Kubeflow, Airflow, Docker, Nginx, Aws/Ec2/Ecs/Ecr/IAM, GitHub Actions, Grafana, Prometheus

Testing: Kali-Linux, Wire-Shark, Burp Suite, Postman, Jest, Web/API-pen-testing

Trainings

Coding Ninjas

Feb 2023 – Feb 2024

Full Stack Web Development

- [Back-End](#) | [Front-End](#)