




Prakash Kumar

+91 7479433835 | ✉ prakashope@gmail.com |  Prakash |  Prakash |  India

EXPERIENCE

BNG Advanced Mobile Solutions

Sep, 2022 – Present

Software Engineer Full Stack

Gurugram, India

- Led **R&D on Generative AI solutions** as part of the **core AI team**, leveraging **LLaMA, vLLM, and Stable Diffusion** to deliver impactful business use cases.
- Architected llm module for Generative AI products, following **SOLID principles**, achieving **80%+ test coverage**.
- Oversaw deployment of LLM module handling **5M+ daily requests** across **5 OPCOs**.
- Led Development of **POCs and demos** for **50+ operating companies**, driving AI adoption and streamlining business processes.

Google Summer of Code - Intel OpenVINO

Mar, 2024 – Aug, 2024

Contributor |  Final Report |  Completion Certificate |  All PRs

Remote

- Improved OpenVINO Node.js binding by implementing essential tensor Ops and model APIs.
- Developed key Node.js API samples, including **OCR** and **Vision Background Removal**.
- Discovered and worked on a critical infer request bug causing failures when model weights lacked a default output layer name.

Pepcoding Education Private Ltd

Aug, 2021 – Sept, 2022

Product Engineer & Mentor

Noida, India

- Led the development of **nados.io Career Page**, serving **10K+ users** and boosting engagement by **15%**.
- Optimized component rendering, reducing **re-renders by 30%** and improving overall load times by **25%**.
- Designed a web development curriculum and mentored **200+ students**, boosting completion rates by **20%**.

FNNDSC/REDHAT CHRIS PROJECT

Mar, 2022 – May 2022

Opensource Contributor |  All Contrubutions

remote

- Reported Issues and Fixed numerous bugs in **ChrisUI** (UI dashboard of the Chris Project).
- **Implemented Mixins** to add Content length in response headers for raw file requests

EDUCATION

Birsa Institute of Technology Sindri , Dhanbad

Aug, 2018 – Apr, 2022

Bachelor of Technology in Electronics and Communication Engineering **CGPA – 7.73**

Dhanbad, India

PROJECTS

EVA - AI Voice Assistant Platform | FastAPI, Socket.io, NextJS, Redis, MySQL

- Reduced **TTFB by 40%** with a custom chunking logic, enhancing response time.
- Boosted **user engagement by 40%** with a custom image generation pipeline using **BLIP** and **Stable Diffusion**.
- Implemented **custom phonetic SSML** for accurate pronunciations across multiple languages.
- Migrated codebase to *asyncio*, reducing **CPU usage by 40%** and increasing **throughput by 50%**.
- Deployed and optimized *SOTA model inference* (vLLM, Stable Diffusion, BLIP), leveraging **NVIDIA MPS** for GPU optimization.

SKILLS

Languages: Python, Javascript / Typescript

Frameworks/Tools: FastAPI, NextJS, LangChain, ChromaDB, ReactJS, MySQL, Git, Linux, Docker