# Mohammad Palla

Email: mohammadpalla24@gmail.com
Portfolio: palla.co.in
Mobile: +91 7893209830
Github: github.com/mohammad-palla

## Education

- **Chaitanya Bharathi Institute of Technology** — Hyderabad, India
  *Bachelor of Technology - Computer Science Engineering; GPA: 8.40* — *July 2020 - March 2024*
  - **Courses**: Operating Systems, Data Structures, Analysis Of Algorithms, Artificial Intelligence, Machine Learning, Networking, Databases, Full-stack development, mobile app development

## Skills Summary

- **Languages**: Python, C/C++, JavaScript, SQL, Bash
- **Frameworks**: React, NextJS, React Native, Express, Django, Flask, FastAPI, NodeJS
- **Tools**: Docker, Git, Redis, Celery, Celery Beat, PostgreSQL, MySQL, MongoDB, Nginx, Datadog, Mintlify
- **Platforms**: Linux, Web, AWS, GCP, Azure
- **Soft Skills**: Leadership, System Design, Product Thinking, Public Speaking, Time Management

## Experience

- **Graphify (YC S26)** — Remote (Contract)
  *Performance & Systems Engineer* — *Aug 2026*
  - **4.45x Faster Indexing – 77.5% Lower Build Time**: Took a cold index of the Bun monorepo (25,725 files) from **10m27s to 2m21s**, a **77.5% reduction**, converting an overnight/CI-blocking job into an interactive-latency one and unblocking per-commit indexing as a product capability, with byte-identical graph output on every run.
  - **58% Less RAM and 67% Less CPU per Index**: Cut **peak memory 5.53 GiB → 2.30 GiB (-58%)** and **CPU burned per index by 67%**, while lifting parallel utilisation from 129% to 188%. Brought large-monorepo indexing within reach of memory-constrained CI runners and laptops, cutting compute cost per index by roughly two-thirds.
  - **Erased a 3.9-7.9x Competitive Gap**: An internal head-to-head had the product indexing **3.9-7.9x slower** than the leading commercial code indexer on identical input. After this work the two run at **wall-clock parity on 42% less peak memory and 15% less CPU**, removing the benchmark as a competitive liability.
  - **Native Rust Extraction Engine across 10 Languages**: Architected and shipped a Rust (PyO3 + tree-sitter) AST extraction kernel replacing the Python hot path for 10 languages, fail-open per file so any kernel fault degrades to the Python extractor. Proved **byte-identical to the incumbent across ∼57,000 files in ∼30 open-source repos** with zero divergences.
  - **Profiling-Led Optimization, Not Guesswork**: Attributed the serial extraction tail with py-spy and deleted **5.2M redundant Path constructions and 5.4M lstat syscalls**; one loop-invariant set() rebuild accounted for **71% of a single pass (50.6s → 4.7s)**. Cut the phase-3 resolution tail **58%** and landed a 3-4x file-detection speedup merged upstream.
  - **Verification Engineering & Adversarial Review**: Built differential-parity, fingerprint and injection-test harnesses to prove the Rust and Python paths identical, then hardened the engine through **∼90,000 fuzz probes** (malformed UTF-8, deep nesting, hostile identifiers) with zero panics. Grew the suite 5,302 → 5,530 tests, surfacing a latent cache-invalidation defect.
- **Vaani AI Research** — Bangalore, India
  *Founding Engineer (Full-time)* — *Oct 2025 – Present*
  - **Voice AI Infrastructure & Latency Optimization**: Architected and built production-grade, end-to-end AI voice agents deployed over telephony and WebRTC, owning the entire software and agentic stack from the ground up as founding engineer. Benchmarked voice AI providers, implemented streaming to minimize latency, and built the full conversational pipeline including multiple function tool-calling integrations.
  - **RAG Pipeline Engineering for Voice AI**: Re-engineered the retrieval-augmented generation pipeline purpose-built for voice, cutting inference cost by 50% and end-to-end latency to sub-300ms while sustaining 90% retrieval accuracy for real-time conversational use.
  - **Multi-Agent Transfer System**: Designed and built a multi-agent transfer system enabling in-call hot-swaps of persona, prompt, and tool configuration within a single session, with per-session transfer caps to bound cost and prevent runaway loops.
  - **Developer Platform: API, SDK & Documentation**: Designed and shipped a public API and SDK to productize the platform for third-party developers, authoring end-to-end developer documentation using Mintlify. Owned the complete frontend (React) and backend (FastAPI) powering the platform.
  - **Core Features & API Optimization**: Designed and implemented end-to-end callback and campaigns systems following deep stakeholder requirement analysis. Optimized backend APIs by up to 80% through efficient query design and architectural improvements.
  - **Schema & System Design**: Modeled nearly the entire database schema and backend system design. Designed complex concurrency and agent management systems, and built a robust auth and middleware layer.
  - **Post-processing & Async Infrastructure**: Built the post-processing pipeline from scratch, leveraging Celery, Celery Beat, and Redis for task queue optimization and scheduled job management.

- **Product & Cross-functional Contributions**: Delivered strong product inputs that improved UI/UX and boosted user engagement. Wore multiple hats including Forward Deployed Engineer when supporting sales and customer success.

- **FindaTutor.ch** Remote (Switzerland)
  *Full-Stack Engineer (Full-time)* *March 2025 – Oct 2025*
  - **Legacy Code & Database Migration**: Led the full migration of the legacy codebase and database from MySQL to PostgreSQL, ensuring data integrity and continuity throughout the transition.
  - **New Product Version**: Owned end-to-end development of the new platform version built on Node.js and React, leading the entire development lifecycle from architecture decisions to production delivery.
  - **Complex Feature Development**: Designed and built end-to-end tutoring booking and calendar management systems for tutors and students, involving complex scheduling logic, availability management, and conflict resolution.

- **Confidential Client (YC F24)** Remote (Contract, under NDA)
  *Voice AI Engineer* *March 2025*
  - **Voice Agent Platform Improvement**: Engaged on a one-month contract to improve the company's production voice AI agent setup built on Vapi, auditing the existing configuration and reworking agent design, system prompts and function tool-calling to make conversations more reliable.
  - **Latency & Conversation Quality**: Tuned STT / LLM / TTS provider selection and streaming behaviour within Vapi to reduce end-to-end response latency, and improved turn-taking and interruption handling to cut dropped and misrouted calls.

- **Sigiq** Gurugram, Haryana
  *Full-Stack Engineer (Full-time)* *Sept 2024 – March 2025*
  - **AI Voice Agent System**: Architected and implemented an end-to-end AI Voice Agent system for EdTech, integrating Deepgram (STT), Anthropic tool-calling (LLM orchestration), ElevenLabs and Cartesia (TTS) into a low-latency, production-grade pipeline. Led development of the frontend, integrating streaming audio, real-time state handling, and intuitive UI feedback loops.
  - **Backend Scalability and Data Consistency**: Built and optimized a scalable backend stack using Django, PostgreSQL, Redis, and Celery. Resolved critical database migration conflicts between two Django services sharing the same schema, clearing significant tech debt.
  - **Observability and DevOps**: Implemented Blue-Green deployments from scratch for zero downtime, Dockerized backend services, configured NGINX for routing, integrated Datadog for observability, and migrated infrastructure from AWS to Azure.
  - **Frontend Modernization & Architecture**: Refactored a legacy frontend into a fully functional Progressive Web App using Tailwind CSS, and established a new frontend repository and code architecture for a greenfield project with modular design best practices.

- **One Immersive** IIIT, Hyderabad
  *SDE Intern* *April 2024 – July 2024*
  - **ArenaExplore.com**: Took full stack ownership of ArenaExplore.com and led its development end-to-end, building the frontend in NextJS (Zustand, Tailwind CSS) and backend with Node/Express and MongoDB/Firebase.
  - **Mobile & Deployment**: Integrated Unity authentication into the server for the Unity mobile app for seamless cross-platform access, and configured AWS, Nginx, and pm2 to deploy both frontend and backend to production.

- **maths.ai** Remote
  *SDE Intern* *June 2023 – April 2024*
  - **Full-Stack & Mobile Development**: Worked with stakeholders and a senior developer to build a mobile app, backend, and web app; built the mobile app in React Native, learning and shipping the MVP in 1 week with no prior mobile experience.
  - **Generative AI Features**: Developed all generative AI features and built a low-cost, low-latency Retrieval-Augmented Generation (RAG) app with Llama-Index and Groq to improve response speed, serving both the mobile and NextJS web app via a common Node/Express backend.

## Projects

- **Prove It – prove-it.palla.co.in**: A GitHub PR reviewer that executes code instead of just reading a diff: clones the PR branch into an isolated Solari sandbox, installs dependencies, runs the existing test suite, generates and runs a targeted repro/verification script with Gemini Flash, and returns a verdict grounded in what actually happened when the code ran. Tech: React/Vite, Vercel serverless functions, GitHub REST API, Gemini Flash, Solari sandboxes.
- **pyapimon – PyPI Package**: Published Python middleware for API monitoring and analytics that plugs into FastAPI, Django, and Flask to capture request/response data, store it locally, and provide a rich CLI (stats, dashboards, ASCII graphs, LLM-powered insights) for exploration.
- **Petrol Pump Management System**: (Work in progress) Built a custom CRM for a petrol pump. Designed the system and schema from scratch with robust, fault-tolerant architecture. Tech: NextJS, NodeJS, MongoDB, Express
- **RAG App with Groq & Llama-Index (no OpenAI)**: Built a low-cost, low-latency RAG microservice without OpenAI API keys by diving into Llama-Index source code. Documented at blog.palla.co.in. Tech: Python, Flask
- **MindTherapy - Therapy Booking App**: Built a full-stack end-to-end web app to book appointments for patients and view schedules for doctors, with separate auth flows. Tech: NextJS, NodeJS, MongoDB, Express
- **Hyderabadi Lang**: A toy programming language forked from open source bhailang and published as an npm package