

BILAL SHARIF

AI and Full Stack Engineer | Python | Go | Django | FastAPI

bilalbilal1005@gmail.com | +92 314 4249717 | github.com/bilalwork31-cyber | linkedin.com/in/bilal-sharif-developer | bilalsharif.work

Lahore, Pakistan (UTC+5) | Remote worldwide. I work the hours the role needs, including a full US or EU day, and I set my schedule to the team

SUMMARY

Senior full stack engineer who takes full ownership of a product: backend, frontend and the decisions in between, not just the ticket in front of me. Five years shipping Python/Django and Go services behind React and Next.js, with applied AI built to actually reach production: agentic LLM workflows, RAG and fine tuned models. I build with Claude Code daily, using custom workflows, skills and MCP tooling to ship faster with a testing pipeline that catches regressions before a reviewer has to. Author of Cauli, an open source Rust task queue that runs 32 times the throughput of Celery.

TECHNICAL SKILLS

Languages: Python, Go, TypeScript, JavaScript, Rust (PyO3), SQL, Bash

AI and LLM: LangChain, LangGraph, CrewAI, AutoGen, Anthropic Claude API, OpenAI SDK, tool calling, multi agent orchestration, RAG (pgvector, Pinecone), embeddings, evals, SLM fine tuning (LoRA, PEFT), vLLM, structured outputs

ML and Generative: PyTorch, scikit-learn, Hugging Face Transformers and Diffusers, classical ML, CNNs, transformers, OCR, FLUX, LTX-2.3 and Qwen self hosted on RunPod GPUs (A100, H100), ComfyUI, cost per inference optimization

Backend: Django, Django REST Framework, FastAPI, Celery, TaskIQ, Django Channels, WebSockets, gRPC, REST, Pydantic, event driven architecture

Frontend: React, Next.js (App Router, SSR, ISR), TypeScript, Tailwind CSS, Zustand, Redux, React Query, real time dashboards

Data: PostgreSQL, Redis, MySQL, pgvector, PgBouncer, query optimization, indexing, double entry ledgers

Cloud and DevOps: AWS (EC2, S3, RDS, Lambda, CloudFront), RunPod, Docker, Caddy, Nginx, GitHub Actions, DigitalOcean, Linux administration, custom autoscaling, observability, Playwright, pytest, load testing

EXPERIENCE

Senior Software Engineer

Sep 2024 to Present

DarkBytes

Lahore, Pakistan

- Own the Django/DRF and Go backend behind ChatsX, a unified Meta inbox sustaining 10,000 requests per second at peak. Cut p95 latency from 900 ms to 220 ms through query optimization, multi layer caching and PgBouncer connection pooling.
- Rearchitected the real time WebSocket layer (Django Channels, Daphne) for 50,000 concurrent connections, bringing response time down from roughly 5,000 ms to roughly 100 ms.
- Designed fault tolerant async pipelines processing 300,000 background jobs, with dead letter queues, idempotent retries, priority routing and admin facing observability.
- Built production agentic LLM systems on LangGraph and CrewAI against the Claude and OpenAI APIs: planner, executor and reviewer graphs with tool calling, structured outputs and eval harnesses, automating internal operations and support triage.
- Fine tuned small language models and deployed self hosted image and video generation (FLUX, LTX-2.3, Qwen) on RunPod GPU infrastructure, removing recurring third party inference costs.
- Replaced managed cloud load balancing and autoscaling with a self built Caddy stack, including a custom load balancer, autoscaler and process manager, cutting DevOps spend and vendor lock in while keeping zero downtime deploys.

Senior Software Engineer

Sep 2023 to Aug 2024

VinnyAndCo

Remote

- Built a white label ecommerce platform on Django and React, serving multiple brands from one multi tenant codebase with per tenant theming, configuration and data isolation.
- Built the internal operations dashboard in React and Next.js, giving the team real time visibility into orders, inventory and brand performance across every tenant.
- Built a RAG system with hybrid retrieval and reranking over the product catalog, cutting manual brand strategy research by up to 70 percent and lifting recommendation relevance over a plain vector search baseline.
- Owned the multi tenant architecture end to end: tenant provisioning, data isolation at the database layer and shared infrastructure that scaled to new brands without a redeploy.

- Cut API query time by 40 percent by fixing a bad query planner path, eliminating N+1 queries, adding composite indexing and connection pooling, and cut CI/CD deployment time from 25 minutes to 7 using GitHub Actions and containerized builds on a DigitalOcean VPS.
- Built a Pinecone and OpenAI RAG pipeline for document intelligence, saving 30 analyst hours per month.
- Implemented card and crypto payment integrations (Stripe, NOWPayments) with a shared authentication layer, covering checkout, webhook verification and payment reconciliation.
- Rebuilt internal tooling login with Playwright browser automation: cookie injection and session based auto login that kept internal dashboards authenticated without manual re-login, plus systemd units, cron jobs and health check scripts for deployment.

PROJECTS AND OPEN SOURCE

Cauli: High Throughput Task Queue for Django and FastAPI

github.com/bilalwork31-cyber/Cauli

- Open source distributed task queue with a Rust core exposed to Python through PyO3. Drop in Celery alternative with async and sync workers, a Redis broker and tunable I/O concurrency.
- Benchmarked at 27,060 tasks per second against 7,812 for TaskIQ and 826 for Celery prefork on identical hardware, a 32 times improvement over Celery. Methodology and reproduction steps published in the repository.

Veriport: Regulated Drug and Alcohol Testing Platform

veriport.ca

- Compliance critical Canadian SaaS for third party administrators, collection service providers and medical review officers. Django codebase of roughly 500,000 lines on AWS.
- Built DOT-regulated and non regulated test workflows, chain of custody, laboratory and eCCF integrations, medical review, audit trails and role based access control, under Canadian privacy law and AES-256 requirements for private medical data.

UmbrellaX: Prediction Market Platform

umbrellax.org

- Polymarket-style prediction market covering market creation, order matching, position management and settlement.
- Built the trading core: order book, double entry ledger, wallet and liquidity management, with real time market data pushed over WebSockets to a React and Next.js frontend. Atomic settlement, idempotent transactions and full reconciliation of balance movements.

DroidFlow: Android Device Automation Platform

droidflow.app

- Desktop and web platform for orchestrating fleets of connected Android devices, built on Django, Node.js, Appium, scrapy, ADB and WebSockets.
- Visual n8n-style graph builder for authoring automation workflows with XPath actions and a plugin system, plus live screen mirroring and low latency device control giving operators real time visibility and manual override.

fancydocx: Pure Python DOCX to HTML Converter

pypi.org/project/fancydocx

- Published PyPI package with zero third party dependencies, no LibreOffice and no Word. Renders tables, multi column layouts, shapes, floating images, headers and footers into a single self contained HTML file with fonts and images inlined, and ships a batch CLI. Replaced a paid .NET dependency in production.

ADDITIONAL HIGHLIGHTS

- Reverse engineered a deprecated vendor WebSocket protocol and shipped a replacement Python client in under three days, restoring a revenue critical integration.
- Identified and fixed a Django remote code execution vulnerability before it reached production.

EDUCATION
