

MAYANK KUMAR

AI Engineer · Backend & Cloud Infrastructure

contact@mayankkr.co · linkedin.com/in/mayankkr6 · github.com/mayankkumar-cmd · India (open to remote)

SUMMARY

AI engineer who ships production LLM systems and owns the infrastructure they run on. Built a real-time speech-to-speech training platform and a permission-aware MCP knowledge server spanning 150+ SharePoint sites, alongside Django and FastAPI services on AWS. Also co-authored the organisation's AI acceptable-use policy and runs its Microsoft 365 and Entra ID estate — product, platform and governance in one seat. Microsoft Certified: Azure Solutions Architect Expert (AZ-305).

TECHNICAL SKILLS

AI / ML OpenAI & Anthropic APIs, OpenAI Realtime API, RAG, Model Context Protocol (MCP), LangChain, LangGraph, vector databases & embeddings, semantic search, topic modelling, sentiment analysis, SFT & GRPO fine-tuning, AI guardrails, Azure AI Foundry, Copilot Studio

Backend Python, Django, FastAPI, Celery, Redis, PostgreSQL, REST APIs, React, Next.js, WebRTC & WebSockets, Microsoft Graph API, OAuth / OBO token flows

Cloud & DevOps AWS (EC2, AMIs, Auto Scaling Groups), Azure, Nginx hardening, Gunicorn, Docker, GitLab CI/CD, performance tuning & cost optimisation

Enterprise IT Microsoft 365, Entra ID, Intune, Exchange Online, SSO (Entra ID, Okta), MFA & conditional access, GDPR / privacy-by-design, AI governance & acceptable-use policy

EXPERIENCE

Associate Engineer | Voluntās Advisory

Feb 2025 – Present

AI product engineering

- **Designed and built a real-time voice-based leadership training platform end to end** — Next.js frontend, OpenAI Realtime API for speech-to-speech conversation, WebRTC audio streaming with WebSocket fallback, Supabase/PostgreSQL, Better Auth. Includes a scenario-management engine (case studies, AI characters, difficulty tiers, learning objectives), gamified call simulation and post-session feedback.
- **Built an MCP server over the SharePoint Search and Microsoft Graph APIs** that turned ChatGPT Enterprise into a governed internal knowledge engine across 150–200 company sites, exposing search, list, folder-browse and query-expansion research tools. Used on-behalf-of delegated tokens so every retrieval respects the requesting user's own SharePoint permissions.
- Exposed that knowledge layer as a **reusable MCP endpoint** now consumed by other internal AI agents, rolled out in access tiers under a governed release.
- **Delphi** (Django platform hosting multiple product lines) — built the comment-analysis service behind MWQ, the employee-survey analytics product: a FastAPI ML service doing sentiment analysis and topic modelling/tagging over vector and semantic search.
- **Outlook add-in for email-to-structured-data extraction** — React frontend with a FastAPI backend, authenticated via Entra ID nested app authentication (NAA) tokens.
- Shipped a range of internal AI tooling: custom GPTs for business processes, RAG systems, a Teams-integrated time-tracking agent, document-review and transcript-analysis tools, and a tender-automation platform.

Backend, cloud & DevOps

- **Diagnosed a dashboard latency regression as sync-view serialization under ASGI** — Django runs thread-sensitive sync code in a single shared thread, so 6–7 concurrent AJAX calls queued instead of running in parallel, costing ~600 ms each. Split the deployment into a WSGI path for HTTP and an ASGI path for WebSockets, routed by NGINX, removing the queueing entirely.
- **Cut AWS infrastructure cost 20–30%** through right-sizing and a revised deployment model, while improving headroom under load.
- Architected and operate AWS workloads on EC2, AMIs and Auto Scaling Groups behind hardened Nginx; implemented custom Gunicorn and Celery services for distributed task processing.
- Optimised application performance via Redis caching, ORM query tuning and elimination of N+1 patterns, static asset compression and CDN integration.
- Containerised services with Docker and maintain GitLab CI/CD pipelines.

AI governance & enterprise IT

- **Co-authored the organisation's AI acceptable-use policy** — approved tooling, permitted and prohibited data inputs, and per-tool access tiers. Moved the company onto enterprise-tier plans (ChatGPT Enterprise, OpenAI API, Claude Team/API) so company data is excluded from model training, with usage monitoring and auditing.
- **Designed a two-layer data-protection model** for the knowledge base: a deterministic blocklist of restricted SharePoint sites applied at retrieval time, then a metadata-only filtering stage run through an open-source model in a controlled cloud environment (Azure AI Foundry / Fireworks), so contracts and salary data never reach third-party AI providers.

- Kept AI-assisted assessment scoring deterministic and rules-based rather than delegating high-stakes judgement to a model — human-in-the-loop and explainability by design.
- Own the full Microsoft 365 and Entra ID estate: identity and access management, MFA and conditional access, Intune, Exchange Online, SSO (Entra ID, Okta), compliance settings and SaaS governance.
- As part of a **two-person team**, **advised a 4,000-employee organisation** on enterprise AI adoption — drafting AI policy and acceptable-use rules, designing the training and rollout approach, and building custom agents in Microsoft Copilot Studio (e.g. lead enrichment).

Associate Technical Engineer | Kyndryl India Pvt. Ltd.

Sept 2023 – Jan 2025

- Implemented and managed Azure virtual machines and networks; troubleshoot and resolved production system issues.
- Integrated operations with ServiceNow for incident and change management; optimised resource performance and cost, and maintained process documentation.

Backend Developer (Remote) | VirtualTX

Mar 2023 – Jul 2023

- Developed and maintained web applications in Yii and Laravel; contributed to the design of a language-learning game platform.
- Built a Chrome extension with annotation and transcription features for speech-data collection.

Cloud Engineer Intern (Remote) | RankSoldier Pvt. Ltd.

Jan 2023 – Mar 2023

- Deployed and managed cloud solutions and supported a cloud infrastructure migration; improved web application performance tenfold.

PROJECTS

Teaching a 100M-parameter model to play Wordle

- Fine-tuned BabyLLaMA (100M) with supervised fine-tuning and GRPO reinforcement learning to generate constraint-satisfying five-letter guesses from Wordle feedback.
- Analysed failure modes where the model produced plausible-looking guesses that violated earlier feedback — a reward-hacking pattern where it optimised for the reward signal instead of reasoning through constraints.

G-Zero — Logitech mouse controller over the HID++ 2.0 protocol · github.com/mayankkumar-cmd/G-Zero

- Python/Tkinter Windows utility that talks directly to mouse hardware over the raw vendor-defined USB channel, resolving HID++ feature indices dynamically from the device's root feature table so it works across models rather than one.
- Replaces a 400 MB vendor client that runs persistent background telemetry; writes settings to onboard memory and exits, with zero resident CPU or memory cost. Three tagged releases with a standalone executable.

CERTIFICATIONS

- Microsoft Certified: Azure Solutions Architect Expert (AZ-305)
- Microsoft Certified: Azure Administrator Associate (AZ-104)
- Microsoft Certified: Azure AI Fundamentals (AI-900) · Azure Fundamentals (AZ-900)
- Cisco CCNA: Introduction to Networks · MATLAB Onramp

EDUCATION

Master of Computer Applications (MCA) | Amity University, Noida

2023

Bachelor of Computer Applications (BCA) | Amity University, Noida

2021