PARAS KARKI

Full Stack Developer | Gen AI & Agent Workflows | IT Infrastructure & Cloud

Toronto, ON · Open to relocation across Canada · paraskarki.com@gmail.com · karkiparas8@gmail.com · linkedin.com/in/paras-karki

SUMMARY

Full-stack developer with a BSc in Computer Science and a year on a production AI document platform, where I moved from testing into troubleshooting, cloud and GPU infrastructure, user access and AI engineering. I build Python and React applications, LLM and vision-model extraction pipelines, RAG search and agent workflows, and I run the Azure, Google Cloud, AWS and Akamai servers they depend on. Hands-on PC troubleshooter and builder, trained in Active Directory, with nearly two years of front-line customer service.

TECHNICAL SKILLS

Software: Python (Django, FastAPI, Flask), TypeScript, JavaScript, React, Next.js, PostgreSQL, SQL, REST APIs

Gen AI and agents: frontier and local LLMs and VLMs, prompt engineering, RAG, embeddings (pgvector), Azure AI Search, agent workflows (LangGraph), OCR pipelines, model evaluation

Cloud: Microsoft Azure (Document Intelligence, AI Foundry / OpenAI, AI Search, Blob Storage), Google Cloud, AWS (S3, IAM, Bedrock), Akamai Cloud

GPU and model deployment: GPU VM instances (NVIDIA L4, RTX 6000 Pro), self-hosted inference, local model deployment, port tunnelling

DevOps: Azure DevOps (YAML pipelines), Docker, Kubernetes (working knowledge), SonarQube, Git

Identity and directory: Active Directory (users, groups, OUs, Group Policy), Microsoft Entra ID, Microsoft Graph, Microsoft 365 administration, AWS IAM

Desktop and hardware: Windows 10/11 installation, configuration and recovery; blue screens and boot failures; drivers and devices; PC builds from parts; hardware diagnostics; printers and peripherals

Networking: IP addressing, DNS, ports and firewall rules, SSH, connectivity troubleshooting

Testing and tools: Postman, Jira, manual accuracy verification, log analysis, Linux command line, Citrix Workspace, SAP, technical documentation

EXPERIENCE

Full Stack Developer · **Mavericks Tech Solutions Inc.** · Toronto, ON — Sep 2025 – Present

Co-op Sep – Dec 2025; full-time since Feb 2026

TESTING AND TROUBLESHOOTING

- Started in QA: tested document uploads and APIs end to end in Postman, checking every extracted field by hand against the source files, because the models cannot reliably check their own output
- Wrote user stories and logged, reproduced and tracked defects in Jira; re-tested after each prompt and model change to catch regressions
- Troubleshoot platform failures, often overnight during releases, by tracing application and server logs; diagnosed payload and field-mapping failures and unreachable model endpoints, and documented each fix
- Explain issues, causes and fixes clearly to teammates and my senior engineer, and flag blockers early

AI AND SOFTWARE

- Build and maintain a multi-service document-extraction platform for insurance documents (Python: Django, FastAPI, Flask) on Azure that turns emailed documents into structured, review-ready records stored in PostgreSQL
- Reached about 98% field-level accuracy on dense, multi-version forms by combining Azure Document Intelligence with OpenAI vision models and refining prompts against real documents
- Led model evaluation to cut AI service costs: benchmarked 5+ open-source OCR and vision-language models on rented GPUs, then shipped a hybrid setup that matches each document type to the most accurate, cost-effective model
- Built search over archived documents with embeddings in PostgreSQL (pgvector), Azure AI Search indexing and a RAG answer layer; evaluated AWS Bedrock Knowledge Bases for document Q&A
- Prompt engineering to normalize extracted data into the target schema for downstream systems

- Built the React / Next.js (TypeScript) review screen that shows extracted fields beside the source PDF, and a step-by-step policy submission flow that saves each step to a policy system's API

INFRASTRUCTURE AND ACCESS

- Stood up self-hosted GPU inference on Akamai Cloud for development and production; ran evaluations on Google Cloud NVIDIA L4 and RTX 6000 Pro instances
- Set up and maintain cloud virtual machines on Azure, Google Cloud and Akamai Cloud; configure SSH, ports and firewall rules and resolve connectivity issues
- Manage user access: add users to Microsoft 365, manage Microsoft Entra ID, and integrate Entra ID and Microsoft Graph into platform services; provision AWS IAM users with scoped permissions
- Work within an Azure DevOps CI/CD flow: branches and pull requests, YAML pipelines, Dockerfiles for containerized deployment, and SonarQube quality gates

Independent Technical Support · Toronto, ON — Ongoing

- Diagnose and repair desktops and laptops: blue screens, boot failures and missing boot drives, battery drain, and driver and device faults
- Install, reinstall and recover Windows 10/11 with files backed up first; speed up slow machines with cleanup, updates and SSD upgrades
- Build desktops from parts (CPU, motherboard, RAM, GPU, SSD, power supply, cooling and case) and run hardware health tests
- Set up printers, monitors, docks and other peripherals; recommend laptops and desktops to match each person's needs, and explain every fix in plain language

Clerk · **Loblaw Companies Limited** · Toronto, ON — Oct 2024 – Present (part-time)

- Help customers every shift, including frustrated or rushed ones: listen, resolve the issue on the spot, and escalate to management when authorization is needed
- Train new colleagues step by step on routines, equipment and standards; cross-trained across four departments and cover shifts at short notice
- Use SAP for receiving, goods transfers and inventory count verification

PROJECTS

Agent workflow: research-to-publish blog pipeline · personal project — 2026

A LangGraph workflow that finds and scores trending topics, researches them, then drafts, illustrates and publishes posts to WordPress. LangGraph, Azure OpenAI (GPT-5 mini / nano), PostgreSQL + pgvector, Tavily, Gemini image generation, WordPress API

Active Directory administration · course and home lab — 2026

Windows domain administration: users, groups and OUs, Group Policy, domain join, password resets, account unlocks, and onboarding and offboarding accounts. Windows Server, AD DS, DNS, Group Policy, ADUC

Portfolio website · personal project — 2026

Interactive portfolio with live demos of my work (document extraction, retrieval, GPU hosting) and a 3D robot scene. Next.js, TypeScript, Tailwind CSS, Spline

EDUCATION

Ontario College Graduate Certificate, Full Stack Software Development (Co-op) · Lambton College — 2024 – 2026

BSc Computer Science · University of Wolverhampton — 2019 – 2022

References available on request