

Kaushik Sanjay Prabhakar

(773) 299-5736 | kprab@uic.edu | [LinkedIn](#) | [GitHub](#)

Education

University of Illinois Chicago (UIC)

MSc, Computer Science

May 2027

Chicago, IL

Vellore Institute of Technology (VIT)

BTech, Computer Science and Engineering

Jul 2024

Vellore, India

- **Achievements:** Abdul Kalam Award Recipient

Technical Skills

- **Programming & Scripting Languages:** Python, C++, JavaScript, TypeScript, Solidity, Lua, Bash, SQL, Jax
- **Web Technologies:** FastAPI, gRPC, RESTful APIs, GraphQL, React, Next.js, Node.js, Flask, Streamlit
- **AI, ML & MLOps:** PyTorch, Scikit-Learn, NumPy, Pandas, Seaborn, Plotly, HuggingFace, Weights & Biases, vLLM, LangChain, Airflow
- **AI Engineering:** LLM Fine-Tuning (LoRA/QLoRA), RAG Pipelines, Multi-Agent Systems (CrewAI, OpenClaw, Claude Code), Apache Kafka
- **Databases:** NoSQL (MongoDB, Neo4j, Pinecone, Redis), PostgreSQL, MySQL, Snowflake Data Cloud, SQLite, Supabase
- **Cloud & DevOps:** Docker, AWS (EC2, EKS, SageMaker), GCP, DigitalOcean, Terraform, Kubernetes, GitHub Actions, CI/CD, Cloudflare

Experience

Apart Research

Jul 2024 - Feb 2025

AI Research Fellow

- Engineered distributed training pipeline for Llama 3 models using PyTorch, Python, and Unsloth, optimizing GPU memory allocation to reduce training latency by 40%.
- Designed end-to-end data processing (ETL) pipelines using Python and pandas to aggregate, format, and serialize large-scale datasets into structured JSON for direct LLM fine-tuning, saving ~20 developer hours per iteration cycle.

AmberFlux EdgeAI

Jan 2024 - Jul 2024

AI Development Intern

- Engineered serverless full-stack Generative AI applications on Snowflake Data Cloud using TypeScript, optimizing prompt context windows and prompt caching to reduce token processing time and deliver 40% lower latency generative responses.
- Evaluated operational costs and scalability of 7+ agentic AI workflows (AutoGen, CrewAI) and built a React-based dashboard adopted by senior executives that drove resource allocation decisions and revealed a 150% cost premium over local developer interns.

Zbyte

May 2023 - Jul 2023

Software Development Intern

- Developed high-performance backend microservices in C++ for a Web3 digital asset distribution platform using async query processing and caching, optimizing RESTful API response times by 45% and ensuring secure ledger integrity for over 10,000 daily transactions.
- Established a comprehensive CI/CD testing strategy with Docker, writing automated unit and integration tests in HardHat to achieve 95% code coverage prior to public production release.

Projects

Geospatial Analysis of Chicago Urban Food Deserts | [Code](#)

- Implemented predictive machine learning pipelines in Python using scikit-learn (Random Forest, Gradient Boosting), achieving a ROC-AUC of 0.871 and 87.1% accuracy in classifying urban food desert status across 800+ census tracts.
- Addressed severe class imbalances in demographic datasets by implementing SMOTE (imbalanced-learn) and successfully increased the predictive recall of the at-risk minority class by 129% (from 31% to 71%) for downstream ML models.
- Designed and integrated geospatial data pipelines using GeoPandas to ingest, clean, and fuse 5+ disparate municipal datasets to enable longitudinal analysis of SNAP participation from 2015 to 2025.
- Developed an interactive, data-driven web dashboard utilizing D3.js to dynamically simulate variable income thresholds (ranging from \$20k to \$130k) and visualize real-time geospatial impacts.

K2-Chat: Distributed LLM Inference Platform | [Code](#)

- Developed a self-hosted, multimodal (text, audio, vision) AI platform serving concurrent user sessions, persistent conversation history, and completely eliminating 3rd-party API dependencies.
- Reduced infrastructure costs by 70% and inference cold-start times by 98% by optimizing a PyTorch C++ inference engine with quantization and lazy-loading, running 3+ AI models (SmolLM, Gemma, Whisper) on low-tier 4GB AWS EC2 nodes via HuggingFace caching.
- Engineered a fault-tolerant, low-latency microservices architecture using Docker, FastAPI, and gRPC, reducing data payload overhead by ~60% and ensuring 99.9% uptime through automated CI/CD pipelines via GitHub Actions.
- Instrumented full-stack observability using OpenTelemetry and Jaeger across 100% of request paths, pinpointing thread-blocking bottlenecks via flame graphs to optimize streaming Time-To-First-Byte (TTFB).
- Accelerated software development lifecycle leveraging Claude Code as an AI-assisted programming tool, delivering a functional prototype in 2 days and the complete production-ready architecture in 8 days, reducing estimated engineering time by over 80%.
- Deployed immutable IaC (Terraform) and containerized Kubernetes clusters, accelerating environment provisioning by 40% while securing backend nodes with a Zero-Trust perimeter via Cloudflare Tunnels.

Publications

- Review paper on Chatbots & Conversational Interfaces. *140+ citations, Co-Author, April 2024, Computer Science Review (Elsevier)*
- Tic Translation Using Transformer Attention. *Co-Author, April 2023, Quest Journals*
- A study of Biases in Common Face Recognition Datasets. *Co-Author, August 2021, Research Gate*