ASHWIN GUPTA
+91 79966 01575 · ashwingupta3012@gmail.com · LinkedIn · GitHub · ashwingupta.dev
Backend & AI Systems Engineer — Distributed Systems, Real-Time Inference, Scalable ML Platforms

AI Engineer with 3+ years building and operating production-grade LLM and distributed ML systems across BFSI and enterprise platforms, delivering measurable latency, throughput, reliability, and cost improvements at scale.

SELECTED PRODUCTION IMPACT — HSBC Real-Time Conversational Analytics

- **Led a 4-engineer team** designing and delivering the enterprise SIP integration stack across Packer-automated GCE workloads and core signaling services (SBC → STT → LLM inference); authored **LLD and voice orchestration architecture**, reducing post-call documentation from **10–15 min → 2–3 min** per interaction.
- **Drove architectural refactor** from thread-based media service to **asyncio + uvloop**, defining the concurrency model, code standards, and review practices across the team; eliminated GIL contention and increased per-VM capacity from **20 → 140–160 calls**, sustaining **1,600+ concurrent sessions** with **<2s E2E transcription delay** and **<5% packet loss**.
- **Optimized infrastructure density**, enabling migration from n2-standard-32 → c2-standard-8 and reducing projected compute from **~$118K → ~$8K/month (~$1.3M annualized)** while improving transcript length **30–40% under load**.
- Architected and deployed a **cross-stack observability and log-correlation layer** over GCP Logging APIs, reconstructing **250K+ log lines in <5s** and reducing **MTTR from ~1–2 hrs → ~5 min**, eliminating manual cross-service debugging.

PROFESSIONAL EXPERIENCE

AI Engineer — Coforge · Jun 2024 – Present
- **Architected an LLM-driven Azure infrastructure intelligence system**, reducing topology analysis turnaround from **2–3 days → ~2–3 hours** through automated extraction and validation.
- **Designed and shipped** a **production RAG**-based contract intelligence pipeline achieving **~96% numerical extraction accuracy**, accelerating sales and customer support workflows by enabling real-time contract query resolution.

Data Scientist — Gida Technologies · Jan 2023 – May 2024
- **Built multilingual RAG systems (163+ languages)** with structured retrieval pipelines achieving **~97% factual accuracy**, enabling real-time vehicle intelligence and sales-support query automation.
- **Developed AI-powered developer tooling suite** (AI CMS, no-code chatbot builder, API utility engine) enabling multilingual content generation and automated cURL-to-20+ language API conversions.
- **Designed real-time weighted graph-based skill recommender**, improving recommendation accuracy by **~30%** and achieving **sub-50ms latency on NVIDIA T4 GPU** under testing load.

RESEARCH & OPEN-SOURCE SYSTEMS

PageIndexOllama **— Provider-Agnostic Runtime Refactor for Tree-Based RAG**
- Replaced OpenAI-tied inference with **provider-routed runtime abstraction**, enabling fully offline Ollama execution while preserving core tree-based reasoning
- Added **response/finish-reason normalization layer** to stabilize recursive traversal across model providers
- Externalized prompts into **registry-driven governance system** and introduced **bounded async concurrency**, improving reproducibility and local-model throughput
- Expanded e2e/integration validation to harden provider variability and reduce regression risk

controla — **Local-First Inference Control Plane (Open Source)**
- Built a provider-agnostic **orchestration layer for routing and managing LLM workloads** across CPU/GPU environments
- Implemented **dynamic model routing, request prioritization, and bounded concurrency** to maintain latency under load
- Designed **resource-aware scheduling** (VRAM/CPU) with fallback handling for large-context and multi-model workloads
- Added execution tracing and telemetry to **debug multi-step inference pipelines** and isolate failures

Publication — NCISCT 2022 - Generating MCQs using Graphs and Language Models

CORE SKILLS

Backend & Systems: Distributed systems, microservices, REST APIs, async & event-driven processing, real-time pipelines, concurrency & performance optimization, caching (Redis), observability, fault tolerance
Languages: Python (asyncio, concurrency), SQL, C/C++, Bash, Linux
Data & Infra: Database design (SQL/NoSQL), ETL/streaming, FastAPI, FAISS/ANN, vector search
AI/ML Systems: LLM integration & deployment, RAG, evaluation & monitoring, agentic workflows, LoRA/QLoRA
Cloud & DevOps: AWS (working), GCP, Azure, Docker, Kubernetes, CI/CD, Terraform, Packer Automation

EDUCATION

Executive Diploma in MLOps & Generative AI — **IIIT Bangalore (2025–2027)**
B.E., Mechanical Engineering — **BMS College of Engineering (2019–2023)**