

ARJUN BAJPAI

New York, NY | +1 929-420-9174

arjunbajpaiwork@gmail.com | linkedin.com/arjun-bajpai | github.com/antailbaxt3r | arjunbajpai.dev

EDUCATION

New York University

Master of Science in Computer Science

Sep. 2025 – May 2027

New York, NY

BITS Pilani, K. K. Birla Goa Campus

Bachelor of Engineering in Computer Science

Aug. 2018 – Jun. 2022

Goa, India

EXPERIENCE

New York University, AI Centre of Excellence

AI Fullstack Developer

Jun. 2026 – Present

New York, NY

- Architect and deploy retrieval-augmented generation (RAG) systems for the Pilot GenAI initiative using LangChain and pgvector, lifting retrieval precision@5 by 12% across enterprise knowledge bases.
- Integrate fine-tuned LLMs (GPT-4o, Claude 3.5, Llama 3) into core SaaS applications through FastAPI and Node.js services, sustaining sub-300ms initial response latency for teams across multiple university departments.
- Implement semantic caching in Redis for recurrent prompt queries, trimming API token spend and improving response latency by 60%.

Microsoft

Software Engineer

Jul. 2022 – Jul. 2025

Hyderabad, India

- Shipped backup and restore orchestration features for Azure BCDR services, protecting SQL and SAP workloads on Azure VMs across large-scale production environments.
- Devised a secondary job-validation pipeline that eliminated false restore failures, raising restore SLA compliance to 99.8% and preventing inconsistent state propagation across distributed backup services.
- Engineered an automated machine learning remediation agent that resolved 800+ security vulnerabilities across 120+ repositories, reducing manual triage effort by 95% and strengthening compliance readiness.
- Led high-availability backup support for SAP ASE, giving enterprise customers 24/7 resilient recovery, and migrated 80+ deprecated BCDR pipelines onto modernized infrastructure aligned with secure SDLC standards.
- Directed recovery during a ransomware incident affecting 80+ customer VMs, designing retention-bypass workflows and sequencing restores to achieve full workload recovery within strict operational timelines.

Software Engineer Intern

Jun. 2021 – Sep. 2021, Jan. 2022 – Jun. 2022

- Delivered Partial Restore for Azure Backup, shortening incremental restore duration by 40%, and parallelized Full Sync execution to lower microservice load and operational cost by 25%.

PROJECTS

Contextly — Local RAG Application | *Next.js, TypeScript, FastAPI, PostgreSQL, pgvector, Celery*

Apr. 2026

- Built a document-grounded question answering application that parses PDF and text files and returns answers with citations pointing to exact source passages, with the full RAG pipeline running locally.
- Developed asynchronous ingestion backed by Redis and MinIO S3-compatible object storage, with nomic-embed-text embeddings and an ms-marco-MiniLM-L-6-v2 cross-encoder reranker.

Erica AI — GraphRAG Course Tutor | *Python, Neo4j, MongoDB, Ollama, GraphRAG*

Sep. 2025

- Designed a Graph-based RAG course tutor with a multi-stage pipeline for ingestion, knowledge graph construction, and scaffolded answer generation spanning 1300+ RAG concepts.
- Constructed a hybrid storage layer combining MongoDB for multi-source ingestion with Neo4j knowledge graphs modeling prerequisite relationships, backed by local GPU inference for privacy-preserving answers with structured citations.

TECHNICAL SKILLS

Languages: Python, C#, Java, TypeScript, JavaScript, SQL

AI/Machine Learning: RAG, GraphRAG, LangChain, LLM Fine-Tuning, PyTorch, Ollama, Scikit-learn, NumPy, Pandas, OpenCV, CNN, RNN

Backend/Distributed Systems: .NET, Node.js, FastAPI, REST API Design, Microservices, High Availability, Celery

Frontend: React, Next.js, Tailwind CSS

Databases/Storage: PostgreSQL, pgvector, SQL Server, SAP HANA, SAP ASE, MongoDB, Neo4j, Redis, Azure Blob Storage, MinIO

Developer Tools: Microsoft Azure, Docker, Kubernetes, Git, CI/CD, Linux, Agile/Scrum