

Shubham Pareek

shubhampareek1211@gmail.com | +1 646-418-8701 | linkedin.com/in/shubhampareek1211 |
github.com/shubhampareek1211 | Portfolio

Education

Columbia University in the City of New York

Master of Science in Data Science

New York, NY

Aug 2025 – Dec 2026

Coursework: Applied Machine Learning, Deep Learning, NLP, Statistical Inference & Modeling, Big Data Analytics

Leadership: Teaching Assistant for Cloud Computing, Introduction to Databases

SRM Institute of Science and Technology

Chennai, IN

Bachelor of Technology in Computer Science and Engineering

Apr 2019 – Jul 2023

Coursework: Data Structures & Algorithms, Object-Oriented Programming, Databases, Operating Systems, Computer Networks

Skills

Languages: Python, SQL, R, C++

ML/AI: PyTorch, scikit-learn, Transformers, RAG, LangChain, Prompt Engineering, LoRA/PEFT, Time-Series Forecasting, Statistical Modeling, Deep Learning

Data / Analytics: PostgreSQL, dbt, Tableau, Power BI, Excel, Spark, Matplotlib

Cloud / DevOps: Google Cloud Platform (BigQuery, Dataform, Looker Studio, GKE, Compute Engine), AWS RDS, Git, GitHub, FastAPI, n8n

Certifications: Google Cloud Associate Cloud Engineer; Google Cloud Professional Cloud DevOps Engineer

Experience

VeYou

New York, NY

Data Engineer Intern

Summer 2026

- Consolidated numerous product, subscription, engagement, attribution, and backend data feeds into a governed BigQuery warehouse and data lake on GCP
- Implemented raw, standardized staging, and analytics-mart Medallion layers with version-controlled Dataform .sqlx models, CI/CD, idempotent loads, and data-quality assertions
- Built internal and external partner-facing KPI dashboards in Looker Studio (Google Data Studio) and enabled governed NL2SQL access over curated lakehouse data

Columbia University Irving Medical Center (CUIMC)

New York, NY

Data Engineer Intern

Mar 2026 – Jul 2026

- Cut PHI re-identification risk across 18+ entity types by building an NLP-based clinical de-identification pipeline
- Enabled secure self-serve access for 5 research teams by deploying governed datasets to AWS with role-based controls
- Ensured reproducible, auditable outputs by automating validation checks and documenting end-to-end data lineage

Weatherhead East Asian Institute, Columbia University

New York, NY

Data Analyst Intern

Sep 2025 – Present

- Achieved 99%+ accuracy on 2,200+ records/day by automating Python ETL pipelines that replaced manual workflows
- Powered forecasting and benchmarking over 6+ months of data by engineering dbt and PostgreSQL data models

Cognizant Technology Solutions

Hyderabad, IN

Programmer Analyst

Jan 2024 – Jul 2025

- Reduced MTTR by 40% across 250+ production cases by debugging GKE cluster failures via root-cause analysis
- Improved reliability by building scripted anomaly detection and automated alerting that cut monitoring overhead

Projects

Resume Griller: LLM Fine-Tuning & RAG Interview Simulator

- Improved answer quality via LLM fine-tuning (LoRA/PEFT) and ChromaDB RAG over a candidate's own resume
- Enabled side-by-side model comparison by deploying the simulator on GCP with runtime model switching

Arcsine: Document Intelligence & Extraction Pipeline

- Eliminated manual data entry by building a 7-stage NLP pipeline that extracts structured records from documents
- Kept records synced end to end by pushing every extracted entity to Notion through its REST API

VisitPrep: Clinical Journaling Web App & NLP Pipeline

- Developed a web app that tracks a patient's post-visit medical journal entries between appointments
- Gave doctors a reviewable patient journey by generating a curated clinical summary from those entries
- Turned free-text health journals into source-grounded clinical signals at 1.00 F1 with a rule-based NLP extractor