GAUTAM VR

gautamvr55@gmail.com | +1 (646) 238-7264 | New York, US | linkedin.com/in/gautam-vr | github.com/gautamvr | gautamvr.github.io

EDUCATION

Columbia University | MS – Specialization in ML/AI - New York, NY, US *[GPA: 4.07/4.00]* *Sept 2024 - May 2025*

- **Course work:** *Applied ML, Applied Deep learning, Modern GenAI, Big-Data, Cloud Computing, DL on EDGE, ML on Cloud.*
- **Research Assistantship** - *RAG & Vector Database* | **Lead Teaching Assistant** for *ML Networking & NxtGen Cloud Computing.*

Vellore Institute of Technology University | B. Tech. in ECE – Chennai, India *June 2015 – May 2019*

- **Robotics Team:** Engineered robots coordinating with interdisciplinary teams with ML for International Robotics Competition.
- **Published research paper -** *IEEE publication citation : ["Assistance For Visually Impaired Using Finger-Tip Text Reader Using Machine Learning" 2019 11th International Conference on Advanced Computing (ICoAC), 2019].*

TECHNICAL SKILLS

Languages	Python, Java, C#, C++, SQL, JavaScript
Frameworks	AWS, GCP, PyTorch, TensorFlow, Pandas, Scikit-learn, PySpark, WPF, .NET, node.js.
Databases	MSSQL, PostgreSQL, VectorDB, MongoDB, Dynamo DB, RDS

WORK EXPERIENCE

Instalily.ai – Vertical AI Agents | Software Engineer 2 - AI/ML – New York, NY, US **May 2025 – Present**

- **Ensemble Healthcare** – Built ML models for COB denial prediction with 91% recall that resulted in $1.2M cost reduction.
- Laid the infrastructure for modelling & scoring pipelines using AzureML, leading to seamless deployment for 7 clients.
- **MCP Platform -** Integrated RBAC OAuth2.1 authorization framework internally for enhancing security within AI modules.
- **SRS Distribution** – Developed and owned 3 vital AI modules used in the Cosailor – AI powered sales agent that generated incremental revenue of $47M for SRS construction distributor. *Built modules using LLMs for 'TODO tasks generation from notes', 'email generation with personality index' – focused toward sales representatives with 87% adoption rate.*

AIrlitz | ML Cloud Engineer Intern – New York, NY, US *Sept 2024 – Jan 2025*

- Implemented end-to-end image ingestion workflow with AI object detection by fine-tuning CLIP with 77% accuracy in AWS.
- Performed cost analysis vs. performance for AWS services, to improve the pipeline speed 3x times & cut down costs by 80%.

Intuit | Software Engineer-2 – Bangalore, India *Oct 2022 – Sept 2024*

- Worked in development of ProSeries TY22,23 Professional Tax products for US, which has revenue over $170 million.
- Implemented AWS server-less pipelines & released 11+ versions to 50k+ customers in US, performed data analysis in Splunk.
- *Filed patent* and honoured for integrating *GenAI* platform for reducing 60% of tax preparers' time, using LangChain & RAG.

Philips Healthcare | Software Engineer-2 – Bangalore, India *Sept 2021 – Oct 2022*

- Carried out the ML Clinical pipeline development, by modelling ML models with data ingested from Clinical Data Lake (CDL).
- Led the end-end development cycle of 3 major features from high level system design to development & code reviews.
- Awarded with the best innovative project for prototyping Medical image segmentation (*PyTorch)* to identify tumour anomalies in medical images by incorporating *transfer learning & distillation.*

Philips Healthcare | Intern + Software Engineer-1 – Bangalore, India *Jan 2019 – Sept 2021*

- Owned & delivered 4 major components using object-oriented programming along with design diagrams and requirements.
- Implemented 6 features in C# while adhering to SOLID principles using BDD (SpecFlow), ensuring code quality TICS > 95%.
- Fixed 3 major Change Requests to address critical customer feedback & investigated 18+ P1,P2 defects in MRI R10 release.

KEY PROJECTS |*More on* GitHub|

Q-BIT Model Inference analysis – GCP, *Jetson Nano GPU, PyTorch, QLoRa, BitNet, HuggingFace* *[Mar 2025]*

- Memory-efficient & computationally optimized fine-tuning approach for GPT2 with data parallel training using 2 GPU nodes.
- Combined *BitNet (1-bit transformers) & QLoRa (quantized adapters)* to deploy on Jetson Nano GPU for inference analysis.

GamezOn – *AWS, JWT, REST, Docker, Content-based Recommender system* *[Dec 2024]*

- A web application for match matching in games, with 7 microservices using AWS services using 3+ deployment strategies.

SpaceX Landing Analysis – *IBM Cloud, SQL, Plotly, Folium, Scikit-learn, Random Forest, XGBoost* *[Feb 2023]*

- A complete ML pipeline that predicts the landing outcome of the SpaceX's Falcon 9 using previous outcome data features.
- Performed Data Collection, EDA, Data Visualisation, Predictive analysis with ensemble of models & hyperparameter tuning.

SkimLit – Skims research papers & provides a detailed summary using NLP – with *RNN (GRU, LSTM) & Conv1D* models. *[2022]*

Document Analyzer – An AI pipeline with *GCP, AutoML models* to extract the key data from receipts to store in *BigQuery*. *[2022]*