Velivela Rohit Kumar

+919000476441 | [velivela.rohitkumar@gmail.com](mailto:velivela.rohitkumar@gmail.com) |

 [Velivela](https://www.linkedin.com/in/velivela-rohit-kumar-136357259/) Rohit Kumar |  [rohitvelivela](https://github.com/RohitVelivela) | Hyderabad, Telangana - 530003, India

# Professional Summary

AI/ML Developer with 1 year of professional experience in designing scalable AI solutions and backend systems deployed on AWS. Skilled in building end-to-end LLM applications using tools like **LangChain**, **LangGraph**, **CrewAI**, and **Hugging Face**. Experienced in developing Retrieval-Augmented Generation (RAG) pipelines and orchestrating prompt-based workflows using **LLaMA 3.1 (70B)** and **Flux.dev** for advanced image generation. Proficient in Python, FastAPI, and PostgreSQL, with a strong focus on integrating Generative AI into business-ready products. Eager to contribute to impactful AI systems that combine language understanding, task planning, and multi-modal content generation in cloud-native environments.

# Experience

[**Quantum**](https://scora.io/) AI Global*Dec 2024 - June 2025*

AI/ML Developer *Hyderabad, Telangana*

* + Built Pitch Craft and Cine Sketch Products: AI platforms that generate structured pitch decks and script-to-visual storyboards from plot or screenplay inputs. Used LLaMA 3.1 (70B), Flux.dev on Ada 6000 GPU, and Fireworks API with a FastAPI + PostgreSQL backend for full-cycle generation and editing.
  + Designed and deployed a generative AI platform that transforms user-input plots, genres, and themes into full-fledged pitch decks containing structured slides such as titles, loglines, narratives, and lookbook visuals.
  + Integrated cinematic image generation by deploying the Flux.dev model on a high-performance NVIDIA RTX Ada6000 GPU, enabling real-time, high-resolution visual outputs aligned with scene-level story prompts.
  + Designed and implemented the RACF (Role-Action-Context-Format) prompt framework to drive structured and consistent LLM outputs across scene generation, slide creation, and user prompts. This ensured narrative alignment, model interpretability, and controllable generation at scale.
  + Orchestrated text-to-image workflows via Fireworks API, enabling efficient routing of prompt metadata between backend services and generation models, while ensuring minimal latency and consistent image rendering per slide.
  + Engineered an end-to-end backend system using FastAPI and PostgreSQL, including modules for generation, regeneration, real-time editing, and structured storage allowing users to manage, update, and export projects seamlessly with scene-specific control.
  + Built complete backend architecture including database schema, controller logic, and modular service flow to align with business objectives for project creation, versioning, regeneration, and visual export across products like Pitch Craft and Cine Sketch.
  + Followed MRC (Model-Route-Controller) architecture for both Pitch Craft and Cine Sketch, ensuring clean separation of logic, scalable backend structure, and maintainable codebase. Implemented modular routing, business controllers, and database models for efficient request handling and extensibility.

Quantum AI Global *Jan, 2024 - Dec 20, 2024*

Intern , Machine Learning *Onsite*

* + Developed a multi-modal ML pipeline combining video, audio, and text inputs using models like microsoft/resnet-50 for facial emotion recognition, facebook/wav2vec2-base-960h for speech-to-text conversion, and cardiffnlp/twitter-roberta-base-sentiment for text sentiment analysis.
  + Integrated LLaMA 3 as the core language model to generate adaptive follow-up questions conditioned on previous user responses, using a custom prompt framework for mental health-specific conversational flow.
  + Implemented audio emotion recognition by analyzing vocal tone and confidence levels using superb/hubert-large-superb-er from Hugging Face, enabling deeper emotional state inference through voice patterns.
  + Designed a centralized database schema to store assessment questions, counselor-generated prompts, regional variants, and multi-modal user inputs, facilitating real-time assessment customization through counselor-facing dashboards. Developed the backend infrastructure for the HR automation system, utilizing the Llama 70B model to assess candidate performance in video interviews.
  + Built a report generation pipeline that synthesized outputs from all modalities and rendered prescriptive reports for counselors, combining visualizations, sentiment timelines, and psychological risk scoring all driven by inference results from Hugging Face models and video/audio signal processing.

# 

# Projects

* **RecAgent : Development for Automating Recruiting Process** May 2024 - June 2024 Tools: Python, LangChain, Generative AI, NLP Techniques, Docker, FastAPI, FireworksAI , ffmpeg. [[](https://github.com/ms-130803/GVP-LLMAgents-Team04)]
  + Designed and developed a recruitment automation system called RecAgent, leveraging Generative AI models for intelligent decision-making.
  + Utilized LangChain agents to streamline HR processes, including parsing job descriptions and generating offer letters.
  + Integrated advanced NLP techniques to analyze and match candidate profiles with job descriptions efficiently.
  + Achieved significant improvements in recruitment speed and accuracy by automating complex tasks, minimizing human intervention.

## Intelligent Website Guide

Tools: Python, LangChain, Generative AI, NLP Techniques, Docker, FastAPI, FireworksAI , ffmpeg [[](https://github.com/gunavardhangolagani/Intelligentchatbot)]

* + Developed a chatbot system using Retrieval-Augmented Generation (RAG), achieving improved accuracy in responses by integrating retrieval systems with generative models.
  + Implemented text preprocessing and cleaning pipelines, processing a large volume of text data for meaningful interaction.
  + Created interactive visualizations for debugging and fine-tuning chatbot performance, ensuring usability and reliability
  + Delivered both RAG and non-RAG versions to demonstrate flexibility in application design and model utilization.

# Skills

* **Programming Languages:** Python, R, Java, JavaScript
* **Web Technologies:** HTML, CSS, React, Node.js, Flask, FastAPI
* **Database Systems:** Microsoft SQL Server, MySQL, PostgreSQL, MongoDB
* **Data Science & Machine Learning:** TensorFlow, PyTorch, scikit-learn, Keras, NLP
* **Cloud Technologies:** AWS, Google Cloud, Azure
* **DevOps & Version Control:** Docker, Git, Jenkins, Kubernetes
* **Specialized Area:** Retrieval-Augmented Generation, Deep Learning, Natural Language Processing

## [Generative AI with Large Language Models](https://www.coursera.org/account/accomplishments/verify/656GNJ7WWX6B) *July 2024*

* [**Google IT Support**](https://www.coursera.org/account/accomplishments/professional-cert/1VSUFQDI4XCV)*October 2024*
* [**Google Advanced Data Analytics**](https://www.coursera.org/account/accomplishments/professional-cert/FHIX596F8CTH)*December 2024*
* [**Google IT Automation with Python**](https://certification-link-d.comhttps//www.coursera.org/account/accomplishments/professional-cert/ZQ1HVC5BDTQG)*December 2024*

# Additional Information

**Languages:** English (Proficiency level), Telugu (Proficiency), Hindi

**Interests:** Sports(Cricket)