jaijoshi1003@gmail.com
+1 (404) 706-9143

Jai Joshi

Current Location: Atlanta, Georgia, USA
https://www.linkedin.com/in/jj0310
https://www.github.com/JaiJoshi123

Education

Georgia Institute of Technology Aug 2023 – May 2025
Master of Science, Computer Science, Atlanta, Georgia ***GPA: 3.88/4.00***

- **Coursework**: Conversational AI, Human and Machine Learning, Grad Algorithms, Machine Learning, Big Data Systems, Linear Programming, Computer Vision, Human-Robot Interaction, Ubiquitous Computing

Sardar Patel Institute of Technology Aug 2019 – May 2023
Bachelor of Technology, Information Technology, Mumbai, India ***CGPA: 9.72/10***

- **Coursework**: Data Structures, Algorithms, Object-Oriented Programming, Advanced Database Management Systems, Blockchain, Artificial Intelligence, Computer Communications & Networks, Distributed Computing.

Work Experience

MicroStrategy Tysons Corner, Virginia
Software Engineer Intern | TypeScript | Spring Boot | Azure | PostgreSQL | AWS | LLMs | Backend *May 2024 – August 2024*

- Designed and implemented auto-completion feature for search engine within MicroStrategy's analytics platform utilizing vector space embeddings and efficient in-memory indexing to accelerate and refine query suggestions.
- Enhanced search response times by **20x** using a caching mechanism to preemptively retrieve and validate SQL from semantically similar past inquiries, ensuring accuracy and faster responses.
- Developed a Cube recommendation engine using Retrieval-Augmented Generation (RAG) for efficient metadata management, facilitating precise Cube identification to power MicroStrategy's Auto Dashboard and BI features.
- Designed and implemented an API in Spring Boot for the telemetry service, enabling secure data retrieval between two microservices.
- Implemented a planner using LangGraph to solve complex queries during an internal hackathon aimed at creating the next version of the product.

PricewaterhouseCoopers LLP Mumbai, India
Technical Intern | Java | MySQL | Oracle ERP Implementation Tools *Jan 2022 – Jun 2022*

- Spearheaded multiple Oracle ERP implementation projects, demonstrating adept management of BI reporting systems for generating and scheduling reports. Streamlined Oracle HCM extracts to facilitate seamless auto data migration.
- Engineered intricate technical OIC integrations, resulting in a remarkable 10% surge in automation of client-side processes.
- Assessed crucial performance metrics on an ad-hoc basis, leveraging cutting-edge big data analytics tools such as PowerBi.

Technical Skills

Concepts: LLM, RAG, Self-Attention, LoRA, Neural Network Quantization and Pruning, Agile Development, REST API
Programming Languages: C, C++, Java, JavaScript, Python, HTML, CSS, SQL, Solidity, Dart, Kotlin (Jetpack Compose).
Frameworks & Other Technologies: Flutter, MongoDB, MySQL, Firebase, Tensorflow, REST Api, Keras, Supervised / Unsupervised learning, Node.js, Express.js, React, Django, Docker, Kubernetes, Jenkins, JIRA, Alteryx, Tableau, Hadoop, Linux, Android Studio.
Nanodegree: Web Development Bootcamp (Udemy), Applied AI with Deep Learning (Coursera), Flutter & Dart Bootcamp (Udemy).

Projects

- **Multi LLM Agent Debate Network**: Developed a multi-agent framework leveraging Retrieval-Augmented Generation (RAG) and LangGraph to classify cognitive engagement markers sin online discussions, achieving 78% accuracy (5% improvement over single-agent methods). Innovations included dynamic few-shot prompting, vLLM-based latency optimization, and a debate mechanism to enhance classification accuracy. The system demonstrated scalable automation for assessing critical thinking in collaborative learning environments.
- **KGInPaint: Image Inpainting with Interactive Scene Graphs**: Developed a user-friendly dashboard for KGInPaint, enabling image uploads, scene graph interaction, and object removal/replacement with inpainted results. Designed a lightweight Relation Transformer (RelTR) for efficient triplet detection, outperforming traditional SGG models. Implemented a DETR-inspired encoder-decoder for scene graph generation and integrated Meta's SAM and HuggingFace's inpainting model for high-quality, efficient inpainting.
- **Visual Question Answering AI ChatBot**: Engineered a browser extension featuring a LLaMA-2 LLM chatbot that answers questions related to uploaded images using a multi-modal attention model. Image features are extracted from a bottom-up attention model (ResNet-50), while textual extraction relies on a BERT model. Employed Parallel Co-Attention for simultaneous image and question attention.
- **DeCluttering Research Assistant Tool**: Created a user-centric web dashboard with full-stack development, incorporating the REST framework, which operates on 3 core principles: BERT algorithm for information summarization, Latent Dirichlet Allocation algorithm (natural language processing) for text classification, and collaborative filtering, leveraging predictive modeling, for recommending relevant articles to expand the user's knowledge.

Research Publications

[1] **"Cataract Detection by Leveraging VGG-19 Classification Model on Retinal Images,"** 2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT).
[2] **"Sign Language Certification Platform with Action Recognition using LSTM Neural Networks,"** 2022 International Conference on Computing, Communication, Security and Intelligent Systems (IC3SIS).

Achievements

- Awarded **First Merit Prize** for being the undergraduate class topper of the I.T. branch with a GPA of 9.95/10 on 22nd April 2022.
- Achieved **First position** in **Kakushin 6.0**, the largest software hackathon organized by **Nomura group**, on September 2022.

Leadership Positions

- **Alpha Microsoft Learn Student Ambassador**: Held several webinars and activities to encourage myself and my juniors to learn new and cutting-edge industry technologies such as Microsoft Azure cloud,.Net 5, and so on.