Joybrata Sarkar

Full Stack AI/ML Engineer

+91 8637543959 · joybrata.official@gmail.com · LinkedIn · GitHub · Portfolio

Professional Summary

Full Stack Engineer with expertise in AI-powered applications and high-performance distributed systems. Led development of real-time conversational platforms serving 10K+ users with sub-50ms latency. Proven track record shipping production AI features using LLMs, RAG systems, and modern web technologies. Combines strong system design skills with product engineering experience to deliver scalable solutions from concept to deployment.

Technical Skills

- **AI & Machine Learning:** Google Vertex AI | LangChain | LangGraph | Silero VAD | Wav2Vec2 | PyTorch | NLP | RAG Systems | Turn Detection
- **Languages & Frameworks:** Python | JavaScript | TypeScript | FastAPI | Angular | React | Node.js | Django
- **Programming Fundamentals:** Data Structures & Algorithms | System Design | Problem Solving | API Design
- **Infrastructure & Systems:** Redis | Celery | Docker | Microservices | CI/CD | MongoDB | WebSocket | Distributed Systems

Professional Experience

Full Stack AI/ML Engineer **Dec. 2024 – Present**
Impacteers *Onsite*

AI Interview Platform (Flagship Project) - Leading 4-Engineer Team

- **Challenge:** Build AI system that conducts natural, human-like interviews autonomously
- **Solution & Impact:** Led 4-engineer team to engineer breakthrough conversational AI platform with real-time processing:
  - Collaborated with stakeholders to **determine user requirements** and translate business needs into technical specifications
  - Contributed to **system design documents** and identified cross-module dependencies for scalable architecture
  - **Real-time speech processing pipeline** achieving **sub-50ms response latency** with **Silero VAD**, **model optimization**, and **WebSocket streaming**, implementing smart turn detection to distinguish between natural speech pauses and completed thoughts, preventing AI interruptions while eliminating response delays
  - Implemented **smart turn detection** using **open-source Pipe-cat adapted for real-time AI interview agent** for conversation endpoint prediction with **sub-50ms inference**
  - Implemented **intelligent conversation orchestration** via **LangGraph state machine** with **Vertex AI Gemini 2.0**
  - Integrated **dynamic question generation, real-time scoring (0-10 scale)**, and **adaptive followup logic**
  - Architected **distributed system supporting 100+ concurrent connections** with < **50ms WebSocket routing**, **multi-layer Redis architecture** (broker + state + pub/sub), **0.1s polling intervals**, and **pattern-based resource cleanup** with distributed locking
  - Optimized **tensor reuse and memory management** achieving <**100ms audio processing** and **200MB per worker** efficiency

Enterprise AI Module Development

- Developed and deployed **3 production AI modules** integrated into enterprise Impacteers platform
- Serving **10K+ active users** with scalable RESTful APIs designed for multi-platform deployment
- Established **microservices architecture** with **error isolation**, **message-level fault tolerance**, and **automatic resource cleanup** across WebSocket, Celery, and Redis
- Implemented **worker specialization** with dedicated audio/video task queues and **comprehensive monitoring**

Additional AI Systems

- **Candidate Scoring Engine:** Built automated job-fit evaluation using lexical, phonetic, and semantic similarity algorithms with **LLM pipeline generating recruiter-ready reports**
- **Career Guidance RAG System:** Engineered production-ready system (GitHub) with **three-phase architecture**, **advanced RAG patterns**, and **persistent conversation history**

Software Engineer – Frontend **Dec. 2022 – Nov. 2024**
Supersourcing *Onsite*

Production SaaS Platform Development

- Shipped **3 production SaaS platforms**: AI interview system, ATS, vendor management tool
- Implemented **NgRx state management, lazy loading, component-level caching** for performance optimization
- Enhanced user experience under heavy load while ensuring maintainability and scalability

Technical Leadership & Team Growth

- **Technical Leadership:** Led front-end development, participated in **design reviews**, and collaborated with **cross-functional teams** on technology decisions
- **Team Growth:** Conducted **technical interviews and hired developers**; performed **code reviews** to maintain quality standards and improve team efficiency

Interactive AI Interview Platform

- **Challenge:** Create seamless multi-candidate interview experience with real-time features
- **Solution:** Built comprehensive platform with **integrated proctor mode, dynamic question generation**
- **Features: Real-time sentiment analysis** and **concurrent multi-candidate support** using **Redis/FastAPI WebSocket** (Demo Link)
- Designed **Redis-based session management** improving state tracking and reducing latency by **60%**

Featured Projects

AI-Driven Developer Search Engine | GitHub *2024*

- Built **NLP-powered candidate search** using **LLM processing and TF-IDF retrieval**
- Implemented **cosine similarity matching and MongoDB indexing** with **dynamic query refinement**

Multimodal Video Query RAG System | GitHub *2024*

- Built **multimodal video retrieval** returning **precise timestamped segments** via semantic search
- Used **BLIP image captioning, Whisper transcription, Sentence Transformers, and FAISS indexing**
- Automated pipeline with frame extraction achieving **sub-second query responses** across large datasets

Education

- **Master in Computer Application**, *Bhilai Institute of Technology, Chhattisgarh* *2020 – 2022*
- **Bachelor of Computer Application**, *KLE Society's S Nijalingappa College, Bengaluru, Karnataka* *2017 – 2020*