Sambit Kumar Barik

LinkedIn: www.linkedin.com/in/sambit-kumar-barik
Github: github.com/Sam-364
Email: sambitbarik70@gmail.com
Mobile: +91-7608095335

Education

- **National Institute of Technology, Rourkela** Rourkela, India
  Bachelor of Technology - Mining Engineering *December 2020 - August 2024*
  Courses: *Machine Learning, Deep Learning, Data Structures, Algorithms, Databases, Operating Systems, Compiler Design*

Skills Summary

- **Languages**: **Python**, **Bash**, **Rust**, **C++**
- **AI/ML Stack**: **PyTorch**, **TensorFlow/Keras**, **ONNX**
- **LLM/RAG/Agentic Frameworks**: **Transformers**, **LangChain**, **LlamaIndex**, **CrewAI**, **DSPy**, **Firecrawl**, **JinaAI**
- **Vector DB**: **Qdrant**, **Milvus**, **Weaviate**
- **Computer Vision**: **OpenCV**, **Pytesseract**, **TensorRT**, **OpenVINO**
- **Infrastructure/Observevility**: **Docker**, **vLLM**, **Triton**, **CUDA/cuDNN**, **Ollama**, **llama.cpp**, **W&B**
- **Development**: **Git**, **VSCode**, **Colab**
- **Deployment**: **Linux**, **AWS**, **CI/CD**

Experience

- **Skylark Labs** Pune, Maharashtra, India
  Machine Learning Engineer - I (Full-time) *January 2024 - Present*
  - **Low Code Platform (LCP) - Multi-Modal AI Pipelines**: Engineered **8+ production-grade AI pipeline solutions** using open-source LLMs (Llama3.2b, Llava-1.5-7B, nllb-200-3.3B) for multi-modal processing including speech-to-text, image-to-text, video-to-text, knowledge graph generation, sentiment analysis, and relevance scoring with **95%+ accuracy** across all pipelines
  - **High-Performance Inference Optimization**: Architected migration from Ollama local inference to **vLLM on Nvidia Triton Inference Server**, implementing tensor parallelism and dynamic batching to achieve **3.2x throughput improvement** and **40% latency reduction** while supporting **concurrent processing of 50+ requests**
  - **Rust Server Migration - Performance Engineering**: Led complete server migration from Python to Rust for in-house Kepler AI platform, achieving **360% performance boost (5 FPS → 23 FPS)** and **60% memory usage reduction**. Integrated **100% of existing Python AI solutions** using PyO3 bindings and implemented dynamic configuration management with **zero-downtime deployment**
  - **Advanced Model Quantization & Cross-Platform Optimization**: Researched and implemented precision quantization strategies (FP16, BF16, INT8, INT4) for custom object detection models, achieving **4.8x faster inference**, **65% memory footprint reduction**, and **3.2x faster model load times** while maintaining **100% detection accuracy**. Optimized models for GPU/CPU/TPU/NPU using ONNX Runtime, TensorRT 8.6, and OpenVINO 2023.3
  - **Containerized GPU-Accelerated Deployment**: Developed Docker containerization for Rust-based inference pipeline with CUDA/cuDNN support, enabling **horizontal scaling across 10+ GPU instances** and reducing deployment time from **45 minutes to 8 minutes** with automated CI/CD integration
  - **Vision-Language-Action (VLA) Model Development**: Spearheading development of custom VLA model for robotics manipulation inspired by OpenVLA and LeRobot architectures. Designed end-to-end ML pipeline with custom PyTorch DataLoader supporting **multi-modal data ingestion (RGB, depth, proprioception)** and implemented efficient batching collater for **training datasets exceeding 100GB**
- **Rupeek Finance** Bengaluru, Karnataka, India
  Data Scientist Intern (Full-time) *June 2023 - August 2023*
  - **Multi-Algorithm Fraud Detection System**: Engineered sophisticated fraud detection pipeline using **ensemble of 3 ML algorithms** (Random Forest, XGBoost, Gradient Boosting) on **75,000+ credit reports**, achieving **96.8% precision**, **91.3% recall**, and **F1-score of 0.94** while reducing **false positives by 23%**
  - **Predictive Analytics & Default Risk Assessment**: Built multi-class classification models using Logistic Regression, Support Vector Machines, and Deep Neural Networks with attention mechanisms, achieving **AUC of 0.93** for default prediction and reducing loan approval errors by **34%**, directly impacting **11,000+ gold loan applications monthly**

Projects

- **FinRexent - AI-Powered Financial Investment Agent (LLM, Multi-Agent Systems, Real-time Data Processing)**: Engineered sophisticated financial investment agent using **Ollama with Llama 3.1-8B** for Indian stock markets (NSE/BSE), implementing **real-time news crawling from 4+ sources**, advanced technical analysis with **8+ indicators** (RSI, MACD, Bollinger Bands), and persistent memory system for investment tracking. Achieved **comprehensive risk assessment** with automated stop-loss recommendations and portfolio diversification. **Tech:** Python, LangChain, Firecrawl, SQLite, Technical Analysis Library, Multi-source Data Integration
- **Inferno — High-Performance LLM Inference Engine (Production-Grade, OpenAI-Compatible)**: Developed a production-grade LLM inference engine integrating advanced techniques from vLLM and SGLang for scalable, low-latency language model serving; implemented **PagedAttention and RadixTree-based prefix caching for memory-efficient KV cache management, continuous batching and speculative decoding** to boost throughput, and OpenAI-compatible REST API for seamless client integration. Engine supports multi-model inference (Llama, Mistral, Qwen, Phi) with INT8/FP8 quantization and streaming generation, enabling high-performance deployment across CPU/GPU environments. **Tech:** Python, Rust/C++ internals, REST API, quantization, high-throughput inference systems.