Siddhartha Bhat

6362417789 work.siddharthabhat@gmail.com Siddhartha SBhat2615

EXPERIENCE

Avaamo.ai **Jan 2024 - Present**
Machine Learning Engineer *Bangalore, Karnataka*

- Built and shipped a self-hosted **vLLM** inference service on **Kubernetes**, serving an **FP8**-quantized 27B model through an OpenAI-compatible streaming API; scaled to **100K+ requests/day**, cut LLM spend **86%** ($7K→$1K/month), and improved throughput per GPU by **67%** through batch-size and KV-cache tuning.
- Improved retrieval recall by **20%** on identifier- and keyword-heavy queries with a hybrid **Milvus** dense retrieval + **Elasticsearch** BM25 pipeline using weighted **RRF**,& denormalized attribute filtering.
- Reduced query-time inference latency by migrating reranker and embedding serving from **ONNX Runtime** to **NVIDIA Triton**, using semaphore-gated **gRPC** connection pooling, asynchronous execution, 1ms-window dynamic batching, & priority isolation between query and ingestion workloads, achieving up to **40**× latency reduction.
- Lowered LLM serving cost with exact and **semantic caching** of generated answers and a citations-only generation mode that cut output tokens by **60%**; de-duplicated concurrent identical requests through atomic **Redis** stream replay.
- Built a pluggable document extraction pipeline across **20+ PDF/HTML parsers**, preserving section hierarchy with context-aware, token-budget chunking; prototyped **LLM**-based structure correction for malformed documents.
- Automated **LLM-as-a-judge** regression testing across customer deployments, evaluating correctness and semantic faithfulness across model-specific structured outputs and parameters, cutting manual QA effort by **70%**.
- Implemented **KEDA**-based autoscaling for **Celery** workers across four **Kubernetes** clusters, scaling replicas from **Redis** queue depth to match ingestion backlog and prevent scaling thrash.
- Improved query understanding with **LLM**-based query correction, **NER**-driven spell correction, and alternate-query generation, capping alternate fan-out by retriever score before reranking to reduce cross-encoder compute.
- Built a natural-language **text-to-SQL** platform across **PostgreSQL**, **MySQL**, and CSV/Excel sources; used **LLM**-guided table filtering, foreign-key relationship graphs, and generated table/column descriptions.
- Contributed to a real-time streaming **ASR** service over **WebSocket** using **faster-whisper** and **Qwen3-ASR**, with a **wav2vec2** verifier to suppress hallucinated transcripts during silence.
- Contributed to a **PII** detection service covering **40+ entity types**, migrating from **GLiNER** to **LLM**-based extraction and adding **Redis** caching to reduce redundant inference.

Samsung R&D Institute India **Sep 2023 – Dec 2023**
PRISM Intern *Remote*

- Fine-tuned **CodeT5** for Text-to-Code generation on CVE security patches, improving **CodeBLEU by 2%** over baseline; built and cleaned a comprehensive multi-task training dataset through EDA and data-quality analysis.
- Received a **Certificate of Excellence**, ranking among the top five projects.

IoTracX Pvt. Ltd. **Jun 2023 – Aug 2023**
AI Intern *Manipal, Karnataka*

- Developed a **YOLOv8** aerial detection system for locating people in forest-fire zones, achieving **93.4% test accuracy** on drone imagery; optimized and deployed the model as **TensorFlow Lite** on an **NVIDIA Torizon** edge device via **Docker** for real-time inference.

SKILLS

Languages: Python, SQL, C/C++
ML/AI: PyTorch, Hugging Face, vLLM, NVIDIA Triton, ONNX Runtime, Azure OpenAI, Vector Search, Reranking
Frameworks/Data: FastAPI, Django, Celery, Milvus, Elasticsearch, PostgreSQL, MySQL, Redis, Langchain
Infrastructure: Kubernetes, Docker, KEDA, AWS (EC2, S3, SQS), Prometheus, Grafana, CI/CD, gRPC

EDUCATION

Manipal Institute of Technology, **2020 - 2024**
BTech in Computer Science & Engineering (CGPA: 8.81/10.0) *Manipal, Karnataka*

AWARDS & LEADERSHIP

Robotics and Circuits **Sep 2021 – Oct 2023**

Coding Subsystem Lead *Manipal*

- Global Top-12 finish in the 2022 International Virtual RobotX Competition, developing a real-time perception system.
- Project Lead for the 2nd-ranked team (42 teams) at MIT Techfest; led and mentored a 25-member computer vision team.