Siddharth Suresh

Santa Clara, CA | +1 (661) 857-2957 | ssures23@ucsc.edu | linkedin.com/in/siddharth-suresh-01924451 |
siddharth.github.io/siddharth.github.io

Experience

Bosch Sunnyvale, CA

NLP and LLM Researcher (Intern) *Oct 2024 – Present*

- Built a **deep research-based scientific discovery system** enabling LLMs to investigate and verify claims using internet-scale retrieval, evidence synthesis, and structured re-ranking for automated knowledge discovery.
- Designed structured **tool topologies** for task composition and multi-tool reasoning, improving function-calling accuracy by ∼7% on the BFCL benchmark.
- Built a **generative SEO pipeline** to influence LLM search ranking through structured content rewrites and **prompt–output evaluation** using learned model ranking heuristics.

Seezo.io Bengaluru, India

Founding AI Engineer (Preseed to 7M) *Nov 2023 – Jun 2024*

- Built the orchestration layer for **human-in-the-loop LLM agents** for AppSec vulnerability detection using **LangGraph**, coordinating automated analysis and expert review; reduced triage time by 40%.
- Created a **retrieval-based analysis module** to review code and architecture docs, producing context-aware insights and fix suggestions, raising detection accuracy by 35% on high-severity issues.
- Implemented an **MCP server** connecting the agent system to **CI/CD and GitHub PR workflows**, delivering real-time vulnerability alerts and remediation actions.

MetaForms Bengaluru, India

Head of Technology (Preseed to 9M) *Mar 2023 – Oct 2024*

- Post-trained **Vicuna-13B → LLaMA-2-13B** using **LoRA (PEFT)** with **RLHF and DPO** for reward-aligned, safety-focused fine-tuning, improving refusal accuracy on sensitive prompts by ∼30%.
- Applied **ToolFormer-style fine-tuning** to teach API usage patterns and parameterization, enabling controlled interactions with external services (WolframAlpha, Google Maps).
- Built **distributed inference pipelines** using **Ray Serve** and **TensorRT-LLM** for large-scale model deployment.

Draup Bengaluru, India

Senior Data Scientist *Oct 2020 – Mar 2023*

- Owned pre-training pipeline of a job-postings LM (**GPT-J**): engineered the data pipeline and led distributed training using **PyTorch + DeepSpeed**.
- Served distilled models with **Ray Serve + ONNX Runtime**, reducing p95 latency by 30% and increasing throughput by 1.5× while cutting model size by 40% and retaining 98% accuracy.

Chefling Bengaluru / San Francisco

Founding Machine Learning Engineer *Mar 2019 – Oct 2020*

- Built a **content-based recipe recommender** using **Word2Vec ingredient embeddings (4k+)** to construct a flavor graph for recipe personalization.

SmartBeings (formerly WooHoo) Bengaluru, India

AI Engineer *Jan 2018 – Mar 2019*

- Built a **Dialogflow-style enterprise chatbot platform** in **Python/Django** with REST APIs and on-prem deployment.
- Implemented **XGBoost-based intent classification** and **custom entity training** with cloud training jobs from user-provided examples.

Patents & Publications

Generative Engine Optimization Framework for Brand Visibility in LLM-Based Recommendations
Patent No. 2025/8008 — Bosch Global Research *2025*

- **Filed:** Oct 2025
- Framework optimizing LLM-driven recommendations for brand visibility through generative query rewriting and alignment signals.

From Heterogeneous Food Data to a Provenance-Aware Ingredient–Nutrient–Recipe Index
Patent No. 2025/8007 — Bosch Global Research *2025*

- **Filed:** Oct 2025
- Provenance-aware LLM pipeline integrating ingredient, nutrient, and recipe data from heterogeneous sources for nutrition verification.

Tool Topology Enhanced Large Language Models: Tool-Based Reasoning Capabilities
Patent No. 2025/7289 — Bosch Global Research *2025*

- **Filed:** Sep 2025
- Introduces structured tool-topology design enabling multi-tool reasoning, compositional planning, and error-tolerant execution.

Education

University of California, Santa Cruz (UCSC) Santa Cruz, CA
M.S. in Natural Language Processing *Sept 2024 – Dec 2025 (Expected)*

- Relevant Coursework: Probabilistic Graphical Models, Transformer Architectures & Scaling Laws, Statistical NLP, Reinforcement Learning from Human Feedback, Interpretability & Alignment, Distributed Model Training.

SRM Institute of Science and Technology Chennai, India
B.Tech. in Computer Science and Engineering *June 2013 – May 2017*

- Relevant Coursework: Machine Learning, Deep Learning, Data Structures, Operating Systems, Distributed Systems.

Skills

Languages: Python, C++, Rust (only vibe-coded), PySpark

Databases: PostgreSQL, MongoDB, Qdrant, Pinecone

Agentic Frameworks: LangGraph, AutoGen

Agent Orchestration: Multi-agent coordination, memory management, context routing

Deep Infrastructure: Ray Serve, vLLM, TensorRT-LLM, DeepSpeed

Training Frameworks: PyTorch, PEFT (LoRA/QLoRA), ONNX Runtime

Retrieval & Memory Systems: BM25, ColBERT, SPLADE, FAISS, Annoy, HNSW, ScaNN, ReRankers (Cross-Encoder, MonoT5)