Yash Bonde

Product Builder — AI research — Startup veteran — Project Artha

I have 0 → 1 startup experience in AI research, consulting, and product development. Passionate about using AI to solve hard problems in fundamental sciences and languages.

bonde.yash97@gmail.com — GitHub — yashbonde.com — LinkedIn — Blogs

Skills

Programming Languages: Python, Go, TypeScript, SQL
ML Frameworks: PyTorch, TensorFlow, LangChain, Hugging Face, OpenAI API, Anthropic Claude, Google Gemini
ML domain: Neural Networks, Large Language Models, Context Engineering, AI Agents, Reinforcement Learning
Software Stack: NextJS, Postgres, gRPC, Clickhouse, Cypher

Work Experience

Software Engineer — Ema Unlimited

March 2025 – Present, Bangalore

- Post-sales implementation for several F50 clients with end-to-end lifecycle from **discovery to delivery**. Solutions included chatbots (with **100K+ users**) and resume matching for **12K+ candidates**.
- Built 2 internal tools reducing effort by **multiple hours/week/employee** across multiple teams
- Implementation of CVE-One internal AI for post-sales team to automate boring parts of project management, used to track project updates.

Head of Research — Tune AI

December 2020 – February 2025, Chennai, Bangalore & San Francisco

Joined Tune AI (prev. NimbleBox AI) as a founding researcher where I experienced the complete 0 → 1 startup journey—from ideation through PMF discovery to landing enterprise contracts with top organizations. Led research on AI algorithms and enterprise solutions while mentoring new technical talent, and like any startup, jumped into sales and product development as needed. The company is backed by Accel, Together Fund, and other leading investors.

Key Achievements as a Startup

- Successfully delivered **multiple enterprise projects** from ideation to production with combined revenue of **$140K+**
- Improved experience for **600K+ users** of Tune Chat by adding **Chain of Thought (CoT)** and several other context engineering techniques
- Customers fine-tuned **100+ models** on Tune Studio and using adapters
- Among **first companies to deploy Meta Llama 2** in production within 24 hours of release
- Deployed **Meta Llama 3 in production within 1 hour** of release; recognized by Meta as India partner
- Conducted several events for Tune AI, building the Bangalore AI/ML community

AI research

Led AI solutioning working directly with Abu Dhabi F1 organizer (Ethara), world's largest scientific contents product (Clarivate), and Intel. Projects became **biggest revenue drivers** for Tune AI.

- **Led two teams** totaling 8 people
- Architected large data processing pipeline running inference on **100K+ documents/day** for extraction task with **96%** accuracy.
- Developed AI Agents reducing sales TAT from 14 days → 5 minutes by auto generating **200+ slide long PPT presentation** and is highly personalized for each prospect and potential event, following the design guidelines. Works from **inside MS Teams** to answer any question via chat interface.
- Researched **context engineering systems** ensuring 100% grounded AI results using hierarchical storage.
- Build a novel transformer model to run AlphaGo style **Monte Carlo Tree Search** based algorithm for autoregressive tasks. It was trained to perform both classification and regression. Implemented by writing custom kernels for **Nvidia Triton & vLLM**.
- Solutioned LLM training with **375GB+ data**
- Wrote **whitepaper with Intel** on their framework OpenVino achieving **20x faster Mask-RCNN on CPUs**.

Product Development

Engineered multiple features and tools including Blob (AI Agent + APIs), ChainFury (CoT prompting backend + workflow builder UI), Silk (distributed code execution engine build on jupyterkernel), Armoury (RAG system in Go), etc.

Mentor & Judge — Hack MIT 2024 & PennApps XXV

September 2024, MIT Cambridge & University of Pennsylvania

- Gave technical workshops on latest AI research and Tune Studio
- Mentored teams building AI products for first responders, fashion designing, education, etc.

AI Consultant — NPAW, Spain

December 2020 – March 2021, Remote

Research and develop a Grafana plugin agent that converts user input in natural langauge to charts. The novel solution used a decision tree to parse the query parameters based on prompting. Deployed **model sharded GPT-2 1Bn on 2 Nvidia-3090 GPUs** to maximise the context length for each input query.

ML Engineer — Shipmnts

July 2019 – November 2020, Ahmedabad

- ML solution to convert unstructured business data to structured knowledge using classical ML techniques applied to document extraction. Pioneered implementation of neural networks for language modeling (translation task) like problems.
- Maintained and developed on data standardization services which used rules for abnormality detection and capturing data for finetuning
- Worked with **Maersk & CMA-CGM** delivering PoCs for clients in Europe, APAC, and LATAM
- Involved in product design, development, and customer interaction

Internships

Summers, 2016 – 2018, Nagpur, Pune & Bangalore

- Improved anomaly detection accuracy by **8%** by adding dynamic trajectory clustering on live traffic data, running on cameras placed throughout the city to monitor traffic and emit events such as jumping signals, ambulance location, etc.
- Developed a faster algorithm by combining the watershed algorithm with the image segmentation model **fine-tuned on India specific road data** which reduced the data labeling effort.
- Developed FAQ Q/A implementing Facebook's **Memory Networks**
- Implemented NLP pipelines for question-answering and text classification tasks
- Designed Python library for creation of infographics in ERP Solutions using ggplot2, pandas, numpy
- Automated data visualization workflows where the report generation time was **reduced by 60%**

Personal Projects

Project Artha

Started July 2025 [Ongoing] — website

Building world's largest digital museum for ancient Indian literature. Developed app, curating and digitizing books, editing and compiling the digital Encyclopedia.

AI Researcher — KS2 Labs

November 2020 – August 2021, Remote, India

- Research on reinforcement learning agents playing chess without perfect board state representation based on the idea that humans can drive without knowing millimeter level precision — YouTube
- New research directions for weather modeling using ground-based sensor data instead of satellites — GitHub

Open Source Projects

tuneapi: Swiss knife Python package for building LLM applications (Python & TypeScript). GitHub

vriksham: Tree-based conversation storage interface with Cypher implementation in Go. Used as data storage in Tune AI to perform tree rollouts. GitHub

nbox: Official Python package for NimbleBox exposing APIs as CLIs. GitHub

Education

National Institute of Technology, Raipur (NIT, Raipur)

B. Tech. in Electronics and Telecommunication — *May 2015 – May 2019*

Built AI-powered Indian sign language detector (Texas Instruments challenge 2018) and transformer network for speech-to-text (Microsoft MSAIC challenge 2018).