Krishna Penukonda

Singapore, SG krishna@penukonda.me https://penukonda.me
Krishna Penukonda tasercake

Building scalable agentic systems for enterprise ecommerce workflows at Hypotenuse AI, owning the engineering pipeline end-to-end. Core strengths include Full Stack Development, Generative AI, LLMOps, and Site Reliability Engineering. Previously developed computer vision models for instance segmentation and real-time object detection systems at Qritive and Red Dot Robotics.

Experience

Lead Software Engineer *January 2021 — Present*

Hypotenuse AI (YC Summer '20)

First full-time hire, owned engineering end-to-end at Hypotenuse AI. Architecting and scaling AI-powered content and image pipelines, cloud infrastructure, and DevOps workflows while mentoring the team and working with product to drive reliable, high-impact releases.

- Led architecture and scaling of AI content generation systems to 1M+ users with P99 latency <500ms and >99.999% availability, owning reliability, cost, and roadmap trade-offs.
- Mentored and onboarded 25+ engineers; set coding/ops guidelines and paired on designs used by multiple product lines.
- Drove team-wide developer velocity: -70% CI/test runtime and -50% pull request turnaround time via pipeline redesign, parallelism, caching, and review automation.
- Cut AWS compute cost ~30% through workload profiling, S3 offload, instance mix, and autoscaling policy updates
- Built and productized LLM-backed features (RAG with <2s retrieval, real-time guardrails), improving factuality and boosting user trust
- Implemented LLM guardrails and feedback loops with fine-tuning pipelines (Python/SQL/Jupyter), improving factual accuracy and safety of generated content.
- Designed LLM prompt-chaining framework for translation, structured data extraction, and brand-voice copywriting, enabling consistent, scalable outputs across customer use-cases.
- Scaled multilingual content generation (30+ languages) via LLM pipelines, enabling international expansion.
- Shipped image-generation service to 10K+ MAUs; cut P99 latency from 180 s to <20 seconds
- Drove full-stack architecture and implementation across React/TypeScript front-end, FastAPI/Python services, and Rust performance modules
- Implemented Redis GCRA / sliding-window rate limiting system that curbed abuse and improved conversion metrics.
- Instituted unified observability (metrics/tracing/logs) that reduced MTTA <60s and MTTR <1h; led incident reviews and set reliability standards across the team.
- Built one-click CI/CD (GitHub Actions → Elastic Beanstalk/ECS) with sub-10-min builds and deployments.
- Owned SaaS platform foundations (billing, access control, anti-abuse systems) ensuring security, scalability, and compliance for enterprise customers.
- Designed polyglot persistence for product search (OLTP + ElasticSearch) to keep <10ms read paths while enabling complex queries at scale

Computer Vision Engineer *October 2018 — December 2018*

Qritive

Developed Object Detection and Instance Segmentation models for use on gigapixel-scale medical images

- Technologies: Python, Keras, TensorFlow, OpenCV, OpenSlide

Computer Vision Engineer *May 2018 — September 2018*

Red Dot Robotics

Developed real-time Object Detection and Tracking models for deployment on autonomous vehicles

- Technologies: Python, Keras, TensorFlow, OpenCV

Education

Singapore University of technology and Design (SUTD) *September 2020*

Bachelor of Engineering in Information Systems Technology and Design

Skills

Core Skills: Generative AI, LLMOps, Full Stack Development, Site Reliability Engineering, Prompt engineering, Technical Leadership, DevOps, Cloud Architecture

Technologies: Python, TypeScript, React.js, FastAPI, PostgreSQL, Celery, DynamoDB, Redis, AWS, OpenTelemetry, GitHub Actions, Vector Databases, Git, Docker, PyTorch, Next.js, Rust