Aprameya C V

AI Engineer
Bengaluru, India | +91 7892944839 | aprameyacv01@gmail.com | linkedin.com/in/aprameyacv | github.com/Aprameya-C-V

SUMMARY

AI Engineer with ~2 years of experience designing, developing, and deploying production AI systems across enterprise environments. Experienced in building agentic AI applications, RAG pipelines, multilingual and voice AI systems, and deploying and setting up inference for AI models, including LLMs and diffusion models. Hands-on experience spanning model serving, GPU acceleration, cloud-native infrastructure, and scalable AI deployments. Also experienced in technical solution architecture, engineering scoping, and translating complex AI requirements into production-ready systems.

EXPERIENCE

Oracle – Cloud Engineering – AI (L2) | Jun 2026 – Present | Remote / Bengaluru, India

- Technical scoping lead and one of the main technical points of contact for EMEA engagements (GPU, Guardrails, Voice AI, MCP, Agentic Automation), influencing **~$3–4M of pipeline**; direct implementor of the solutions behind **$1.2M+ in annual consumption**.
- Built a **three-agent ServiceNow suite** for a government ministry: natural-language search over **1M+ tickets in under 7 seconds**, hybrid semantic and keyword search over **500K+ multilingual articles**, and ticket creation in **under a minute**.
- Built a **multilingual legal and tax RAG assistant** for a national tax authority in Eastern Europe, letting citizens ask questions in their own language (OCI Enterprise AI, FastAPI, OpenAI SDK, Docker).
- Delivered OCI infrastructure across customer engagements (landing zones, networking, IAM, VM migration, disaster recovery), automated with Terraform.
- Built deployable solution packs (RAG, Agents, Voice AI, MCP, Guardrails, Memory) that save **50–60+ person-hours per engagement**.
- Drove **Codex and Enterprise AI adoption** for **80+ people**, who built 40–50 reusable assets, saving **hundreds of person-days** for the team.

Oracle – Cloud Solution Engineer (L1) | Jul 2025 – May 2026 | Remote / Bengaluru, India

- Won the **Mastery Award** (EMEA TE Recognition Program) for a trademark platform built for a government ministry, delivered as two APIs: (1) a similarity engine over **500K+ images** with multilingual text matching and compliance rules, running thousands of checks a week on a CPU-only machine; (2) an optimized open-source image generation and editing service on NVIDIA CUDA GPUs, driven by a custom ComfyUI module.
- Primary developer of internal AI tools (document generator, story generator, provisioning automation), saving an estimated **1,100+ person-hours**.
- **EMEA Regional Winner and Global Top 15** (of 150+ teams) in Oracle's internal hackathon with a real-time speech-to-speech voice AI.

Oracle – Project Intern | Jan 2025 – Jul 2025 | Bengaluru, India

- Built the Solution Document Generator (LangGraph, OKE); contributed to **50+ documents**, saving **150+ person-hours**.
- Built a **security review agent** used to review **50–60 production go-live projects**, saving **100+ person-hours**.

PROJECTS

Pramana: Calibrated Decision Models | Python, LoRA fine-tuning, Hugging Face

- Fine-tuned 2B and 0.8B models with LoRA on soft gold labels, using a Brier loss term and decoupled calibration, so each answer returns a probability per option and a confidence score that tracks real accuracy. A 0.8B-to-2B cascade holds 80% accuracy across 2,000 decisions.

SpecLane: LLM Serving Under Agent Load | vLLM, speculative decoding, FP8 KV cache, NVIDIA H100, L40S

- Benchmarked 27B FP8 and 9B models in vLLM under agent-style load. Multi-token prediction (generating several tokens per step) more than doubled throughput, from 78 to 175 tokens/s, with no accuracy loss. Also diagnosed and fixed a max-num-seqs CUDA graph capture failure.

HBC-RAG: Agentic RAG for Legal Contract QA | Python, FastAPI, pgvector, hybrid search

- Built an agentic RAG system for legal contract QA: planner and verifier agents under a hop cap, with pgvector HNSW and Postgres full-text retrieval fused by reciprocal rank fusion. Retrieving four passages per hop instead of one raised clause hit rate from 47% to 58% and halved median latency (8.9 s to 4.2 s).

CG-BI: Wind Turbine Blade Inspection | Python, multimodal LLM APIs, confidence-gated routing

- Built a two-tier multimodal cascade on 75 drone images across six defect classes: every image goes to a low-cost model, and the 36% scoring below 0.9 confidence escalate to a stronger one. Reaches 92% critical-defect recall at 47% of the strong tier's cost.

Edge AI Inference Engine | K3s, Helm, Prometheus, llama.cpp, AWS EC2, k6

- Served quantized GGUF models with llama.cpp behind a FastAPI, OpenAI-compatible proxy, Helm-packaged on K3s over AWS EC2. Prometheus metrics via the custom metrics API drive HPA autoscaling; sustains 14–16 tokens/second under k6 load.

SKILLS

RAG & GenAI: Retrieval-Augmented Generation (RAG), Hybrid & Vector Search, Guardrails, Multilingual AI, Voice AI, LangChain, OpenAI SDK

Agentic & Multimodal AI: Agentic Workflows, Multimodal LLMs, LangGraph, Google ADK, MCP, ComfyUI

LLM Serving & Fine-tuning: vLLM, llama.cpp, Speculative Decoding (Multi-Token Prediction), LoRA Fine-tuning

Cloud & Infrastructure: AWS (EC2, Lambda, Amplify), GCP (Vertex AI, Model Garden, Cloud Run), OCI (Enterprise AI, Dedicated AI Clusters, OKE, landing zones, VCN, IAM, DR); Terraform

DevOps & Languages: Linux, Docker, Kubernetes (K3s), Ansible, Prometheus, k6, CI/CD; Python, TypeScript, SQL, FastAPI

EDUCATION

JSS Science and Technology University | B.E., Electronics and Communication | Dec 2021 – Jun 2025
CGPA 8.93; Gold Medal
Design patent; Springer and IEEE publications