Aprameya C V

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

Summary

AI engineer at Oracle taking enterprise and public-sector AI from discovery to production on OCI across EMEA: scoping lead on [about $3–4M of pipeline], builder of the solutions behind [around $1M in annual consumption], and SME for OCI Enterprise AI and agentic AI.

Experience

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

- Scoping lead and technical point of contact for EMEA engagements (GPU, guardrails, voice AI, MCP, agentic automation) [about $3–4M of pipeline; around $1M 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, so citizens can ask questions in their own language (OCI Enterprise AI, FastAPI, OpenAI SDK, Docker).
- Delivered extensive **OCI infrastructure** across customer engagements: landing zones, cloud networking, storage, IAM, VM migration, and disaster recovery, automated with Terraform.
- Built deployable **solution packs** (RAG, agents, voice, MCP, guardrails, memory) saving 50–60+ person-hours per engagement; drove **Codex adoption** for 80+ people who built 40–50 reusable assets.

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

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

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

- Built the Solution Document Generator (LangGraph, OKE); contributed to **50+ documents, saving 100+ man hours**.

Projects

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

- 2B and 0.8B models: LoRA fine-tuned on soft gold labels with a Brier term and decoupled calibration, so each answer carries a probability per option plus a confidence score. A 0.8B to 2B cascade holds **80% accuracy on 2,000 decisions**.

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

- Benchmarked 27B FP8 and 9B models under agent-style load with vLLM, one flag change at a time. Multi-token prediction **more than doubled throughput** (78 to 175 tokens/s) with no accuracy loss; fixed a max-num-seqs CUDA graph capture failure.

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

- Agentic RAG over legal contracts: planner and verifier calls under a hop cap, with pgvector HNSW plus full-text retrieval fused by reciprocal rank fusion on Postgres. Four passages per hop instead of one raised clause hit 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*

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

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

- Quantized GGUF serving via llama.cpp behind a FastAPI OpenAI-compatible proxy; Prometheus metrics through the custom metrics API drive HPA scaling, Helm-packaged on K3s over AWS EC2. **14 to 16 tokens/second** sustained under k6 load.

Skills

Agents & GenAI: ADK, LangGraph, LangChain, Deep Agents, OpenAI SDK, MCP; RAG, vector search, multilingual and voice AI
LLM Inference & Training: vLLM, speculative decoding, FP8 KV cache, CUDA, NVIDIA GPUs, LoRA fine-tuning, Hugging Face
Cloud: AWS (EC2, ECS, Lambda); Google Cloud (Vertex AI, Model Garden); OCI (Enterprise AI, Dedicated AI Clusters, OKE)
OCI Infrastructure: Landing zones, networking (VCN), storage, compute, IAM, VM migration, disaster recovery, Terraform
DevOps & Languages: Linux, Docker, Kubernetes, Helm, CI/CD; Python, TypeScript, SQL

Education

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