

Anusha Pundir

anushapundirwork@gmail.com | [LinkedIn](#) | [GitHub](#)

SUMMARY

AI-native full-stack engineer who ships production **LLM products** end to end, from **PostgreSQL** schema and typed APIs to containerized deploys on **Google Cloud**. Builds **LangGraph** agents, **MCP servers**, **RAG** pipelines, and validated **LLM-as-judge evaluation**, plus the **Next.js / TypeScript** product surface around them. Owns features from database constraint to live URL.

TECHNICAL SKILLS

AI/ML: LangGraph, LangChain, MCP, RAG, LLM Evaluation, Prompt Engineering, PyTorch, QLoRA, FAISS

Full-Stack: TypeScript, Next.js, React, Node.js, Python, FastAPI, REST APIs, PostgreSQL, Drizzle ORM

Cloud & DevOps: Google Cloud (Cloud Run, Cloud SQL, Cloud Build, Secret Manager), Docker, Git, Vercel

EXPERIENCE

JettyAI - AI Engineering Intern

March 2026 – May 2026

- Architected a **3-layer hybrid content moderation pipeline** for dating platforms, balancing **accuracy, latency, and cost** across text and image moderation.
- Designed and evaluated a **multi-stage moderation workflow** combining classifiers, moderation APIs, keyword matching, caching, LLM-based contextual decisioning, and human review, demonstrating a pathway to limit manual review to only **8-10%** of ambiguous cases.
- Worked on an internal **automated GTM AI tool** to outreach **B2B businesses** with AI solutions, using **LLM-driven personalization** and automated prospect research to accelerate the **sales pipeline**.

PROJECTS

Callie - AI Voice Agent for Appointment Rebooking | [Live](#)

August, 2026

- Built and shipped a **full-stack AI voice agent** (**Next.js 16, TypeScript, PostgreSQL, Retell AI**) that calls customers, negotiates a new appointment time, and commits the rebooking through **live tool calls** before the call ends.
- Made double-booking **impossible at the database layer** with a **PostgreSQL exclusion constraint** (`btree_gist`) over time ranges, proven by concurrent-write tests rather than check-then-write application logic.
- Built idempotent, **HMAC-SHA256-verified webhooks** and **REST tool APIs** with fast-200 async processing, so duplicate or forged deliveries cannot corrupt appointment state.
- Ran post-call extraction through **Claude Haiku** with strict **JSON output** and a retry-then-degrade path, keeping tool-committed outcomes authoritative so a parse failure never overwrites a booking.
- Deployed on **Google Cloud: Cloud Run** and **Cloud SQL** via multi-stage **Docker, Cloud Build** images, and **Secret Manager** credentials, with **Clerk auth, Google OAuth**, and versioned **Drizzle migrations**.

EvalLoop - Self-Improving LLM Agent with Evaluation Harness | [GitHub](#)

June, 2026

- Built a **self-improving LLM agent** on a **LangGraph state machine** that grounds answers in a local **RAG** knowledge base, scores its own output, then critiques and revises in a closed loop with **guaranteed termination**.
- Designed a **hybrid evaluation harness** that runs free **deterministic checks** (grounding, coverage) on the full set and reserves a sampled **LLM-as-judge** for correctness, completeness, and clarity, blending both into one quality score.
- Validated the judge** against human-labeled data before trusting it, reaching **1.00 ranking accuracy** with clean separation between strong and weak answers (mean **0.93 vs 0.30**).

Discourse-MCP - Model Context Protocol Server + Agentic Orchestrator | [GitHub](#)

May, 2026

- Authored a **Model Context Protocol (MCP) server** from scratch (**FastMCP**, stdio JSON-RPC) exposing **11 typed tools** that let any MCP-compatible LLM app search and fetch tech-community threads across **Hacker News, Lobsters, and Reddit**.
- Orchestrated a **LangGraph ReAct agent** that consumes the server as a real second **MCP client** over stdio, binding all 11 tools with a capped reasoning loop as a budget guard.
- Built a **provider-swappable model layer** (Claude Haiku to local Ollama by one env var) and an **LLM-as-judge eval harness** that caught the local 7B model **fabricating URLs and dates** where Claude stayed grounded.

DermWise GenAI - AI Powered Dermatology Assistant | [Website](#) | [GitHub](#)

February, 2026

- Trained **EfficientNet B0** on the HAM10000 dataset (10,000+ images, 7 classes) using two stage fine tuning; achieved 75% validation macro F1 on an imbalanced dataset with **Test Time Augmentation**.
- Fine-tuned TinyLlama 1.1B** Chat using **QLoRA** (4 bit quantization, ~2% trainable parameters) to generate structured clinical reports.

EDUCATION

Bachelor of Technology in Computer Science (*Specialization in Artificial Intelligence*), Bennett University

Expected Sep 2027

AWARDS

- Hackathon Winner:** Won **1st Place at Hackurate** Hackathon and Finalist at a National Hackathon.
- 2nd Place in AI/ML** Track amongst 150 teams for **ResQ** at College Showcase.