ANKIT DAHAL

https://linkedin.com/in/ankitdahal | https://dahal.ai | https://github.com/dankit
Open to Relocation

PROFESSIONAL SUMMARY

Software Engineer with 5 years of experience transitioning into Applied AI/ML work. Hands-on practitioner in LLM pretraining, supervised fine-tuning (LoRA/PEFT), retrieval-augmented generation, agentic systems, and distributed GPU training. Strong production engineering foundation in scalable, high-availability systems with a proven track record of delivering measurable business impact through optimization and cross-functional technical leadership.

AI / ML PROJECTS (JULY 2025 - PRESENT)

Independent research and engineering - more projects and details at https://dahal.ai

Adversarial AI Agents Research - AI Safety

- Researching adversarial AI methodologies to better understand astroturfing and disinformation capabilities. Outline available on website - ideas include computer use agents (qwen 3.5-35b-a3b), teacher model (qwen 3.5-397b-a17b), reinforcement learning, synthetic data, vLLM, playwright browser harness

Large Language Model Pretraining

- Designed and pretrained a **~450M-parameter** dense transformer on **10B tokens** (fineweb-edu) using **8x A100 GPUs** with PyTorch distributed data parallel (DDP), custom training loops, and Chinchilla-optimal scaling. Implemented RoPE, RMSNorm, SwiGLU, GQA, KV-cache, and Flash Attention.

Agentic Retrieval-Augmented Generation

- Built agentic RAG over **250K+ pages and 3M+ embeddings** of legal documents: hybrid search (Elasticsearch BM25 + bi-encoder + ChromaDB), reciprocal rank fusion, reranking (cross-encoder), conversational and search agents with planning/self-triage. Finetuning with Google cloud platform (kubernetes and compute engine), custom document chunking, agent/retrieval evals, distributed training, and hnsw index tuning.

Llama 3.1 8B Instruction-based Supervised Fine-Tuning

- Fine-tuned Llama 3.1 8B base model via **LoRA/PEFT** (Unsloth), achieving **52% improvement** on IFEval (200/834 to 305/834) with ~**$10 compute**. Also conducted quantization experiments (4-bit, 8-bit, BF16) and evaluated on tinyMMLU and IFEval.

PROFESSIONAL EXPERIENCE (MAY 2021 - JULY 2025)

Senior Software Engineer | QCI (Consulting company)

Promoted to Senior after 1 year 9 months. Client: https://www.brownells.com, a major e-commerce retailer, 10M+ users.

- Led platform modernization serving **10M+ users**, coordinating across **6+ teams**; reduced downtime by **90%** and increased revenue and customer growth by **~20%**.
- Identified critical availability gaps in the payment gateway provider, reaching escalation to c-suite and leading cross-functional replacement with rapid turnaround, preventing up to **$500K/day** in potential revenue loss.
- Cut core API traffic by **68% (10Ms requests/day)** through Redis caching and telemetry-driven analysis, significantly reducing infrastructure costs and API latency across multiple services and teams.
- Architected message queue infrastructure (pub/sub), enabling asynchronous, fault-tolerant workflows. i.e. processing **1M+ orders** totaling **$200M+/year**.
- Improved query performance **10x** via full-text search indexing in SQL Server across **100M+** log records, eliminating timeouts and improving developer productivity, along with general database indexing and query performance tuning.
- Modernized CI/CD pipelines for **80+ applications** across **10+ teams**, introducing automated test coverage gates and reusable templates that reduced release errors and improved deployment velocity.
- Built **10+ internal tools** and automations, reducing manual error rates and enabling non-technical teams to self-serve without engineering team dependencies.
- Established **data governance** through better data modeling, data validation, and alerting, to maintain **compliance with the Bureau of Alcohol, Tobacco, Firearms and Explosives (ATF).** Prevented millions of dollars in losses due to potential fines and business interruption.
- Reduced false-positive fraud review cases from **18-20% to ~1%** via Cybersource/Visa decisioning, decreasing customer support burden and accelerating order fulfillment from **4 to 3 days**.
- Mentored two junior engineers, and assisted with the creation of an internal knowledge base to establish a culture of knowledge sharing and continuous improvement across different teams.

TECHNICAL SKILLS

AI / ML: PyTorch, LLMs, Pretraining, SFT, LoRA/PEFT, Quantization (bitsandbytes, unsloth), LLM Evaluation (IFEval, MMLU), Reinforcement Learning, AI Agents, Huggingface, Transformers, Natural language processing (NLP)
ML Infra: Distibuted data parallel (DDP), Docker, GCP (Compute Engine, Kubernetes Engine), Lambda Cloud, ChromaDB, Elasticsearch, RAG Pipelines, Weights & Biases, Torch Elastic, vLLM
Languages: Python, Java, C#, SQL, JavaScript, Angular
Backend: Azure, .NET, REST APIs, Microservices, Pub/Sub, Redis, NoSQL, CI/CD, Terraform, High-Availability Architecture

EDUCATION

Bachelor of Science in Computer Science — University of Iowa, Iowa City, IA | 2016-2020