HIMANSHU DHURVE

+91-9699538032 himanshudhurve96@gmail.com linkedin.com/in/himanshu-dhurve-92057b183

SUMMARY

ML engineer with hands-on experience across model development, evaluation, and deployment. From fine-tuning and post-training alignment to production pipelines. Increasingly focused on the gap between models that perform well under evaluation and ones whose internal behaviour can actually be trusted, with recent work in mechanistic interpretability and alignment research. Looking to contribute to teams where getting the measurement right matters as much as the model itself

EDUCATION

San Diego State University — **San Diego, USA**
Master of Science in Aerospace Engineering, GPA: 3.52/4.0 — *2021 – 2024*

Dr. Babasaheb Ambedkar Technological University — **Raigad, India**
Bachelor of Technology in Mechanical Engineering, GPA: 8.57/10.0 — *2014 – 2018*

PROFESSIONAL EXPERIENCE

AI/ML Research Intern, Vizuara AI Labs, India — *Jan 2026 – March 2026*

- Built and evaluated a citation-grounded RAG system for OpenFOAM technical documentation using hierarchical chunking, dense retrieval, and cross-encoder reranking; achieved 4.6/5 overall generation quality and 5.0/5 citation correctness on a custom benchmark. Accpeted at ACL 2026 RAG4Rep Workshop. [OpenFOAM RAG Link]
- Fine-tuned RF-DETR for multi-class Synthetic Aperture Radar (SAR) object detection on a custom dataset, achieving 0.614 mAP and improving detection robustness across heterogeneous scene categories.
- Developed an agentic VLM pipeline that transforms research PDFs into structured technical digests by generating narrative summaries, AI-generated diagrams, and executable Jupyter notebooks through automated multimodal reasoning workflows.

Aerodynamics Research Assistant, San Diego State University, USA — *Aug 2021 – Dec 2023*

- Designed **large-scale data acquisition and analysis pipelines** on an HPC cluster (SLURM); processed high-dimensionality flow field datasets in Python, cutting simulation wall-time by **75%** through parallelization.

TECHNICAL SKILLS

Core Competencies	Mechanistic Interpretability, LLM Evaluation & Benchmarking, Post-Training Alignment
Languages & Tools	Python, C++, PyTorch, TensorFlow, Scikit-learn, OpenCV, Hugging Face, Git
MLOps & Infra	Databricks, MLFlow, DVC, Docker, CI/CD, OpenAI SDK, CrewAI, FastAPI, uv
Research Methods	Dataset Curation, Benchmark Design, Activation Analysis, Linear Probing, Causal Ablation, Model Evaluation

PROJECTS

Geometry of Refusal: Mechanistic Interpretability of Safety-Relevant Representations — *May 2026*

- Replicated and extended Arditi et al. (2024) on **Llama-3.1-8B-Instruct**: extracted per-layer **DIM refusal directions**, confirmed layer 12 as the strongest source via ablation sweeps, and showed that both single-direction and per-layer ablation achieve **100% flip rate** on training prompts against a 0/150 random baseline.
- On JailbreakBench (n=100), ablating using only layer 12's direction, each layer's own direction, or every layer except 12 all reached true **ASR 100%** with coherence intact; refusal scores revealed the nuance; applying layer 12's direction across all layers dropped the score to -3.67, while skipping layer 12 entirely matched the per-layer result (-0.96), suggesting **refusal is distributed across layers** but layer 12's direction transfers unusually well to the rest. Activation addition on harmless prompts confirmed causal sufficiency, pushing avg refusal score $\sim 9\times$ (1.48 $\rightarrow$ 13.46). [Medium Blog]

End-to-End MLOps Pipeline for Customer Churn Prediction — *Apr 2026*

- Built a fully modular production pipeline on 243k rows: **XGBoost** with **Optuna** tuning (30 trials, recall-optimised), **MLflow** experiment tracking and model registry, **FastAPI + Gradio** serving, **multi-stage Docker** build, and **GitHub Actions** CI/CD, achieving **93% recall** at a tuned threshold of 0.30 on an imbalanced churn dataset. [repo]

LLM Evaluation via Multi-Signal Reward Engineering (Medical MCQ) — *Nov 2025*

- Fine-tuned **Qwen3-1.7B** with **GRPO** on structured medical QA using **multi-signal reward functions** (binary correctness, regex format validation, length-calibrated explanation scoring); managed full **SFT + GRPO pipeline** on constrained Kaggle GPUs with 4-bit quantization (LoRA rank 32) and **vLLM-accelerated sampling**, analysed failure modes and output distributions to identify model weaknesses. [repo]

Temporal Signal Stabilization for Video OCR (License Plate Recognition) *Sep 2025*

- Built a real-time detection and recognition pipeline using fine-tuned **YOLO11** and **EasyOCR**; addressed **temporal signal noise** (OCR flickering) with a majority-voting algorithm over a 12-frame rolling buffer. [repo]

Multimodal Benchmarking, Student Misconception Detection (Kaggle MAP) *Aug – Oct 2025*

- Fine-tuned **Qwen2.5-Math-1.5B-Instruct** with **Unsloth** and a compact **MLP classification head** on curated math reasoning misconception benchmarks; improved MAP leaderboard score from 0.885 to **0.925** (+4.5%) by combining LLM semantic understanding with structured output evaluation. [repo]

AI SAFETY ENGAGEMENT

- **BlueDot Impact**, Technical AI Safety Fundamentals -Intensive Cohort (May 2026). Covered mechanistic interpretability, scalable oversight, RLHF failure modes, dangerous capability evaluation, and activation monitoring. Sprint project: *Geometry of Refusal* [Medium Blog] .

PUBLICATIONS

- **Himanshu C. Dhurve** et al. "Decompose, Retrieve, Cite: A RAG Pipeline for Structured Report Generation from Technical Documentation." *RAG4Report Workshop, ACL 2026* (First Author).
- Joseph Katz, **Himanshu C. Dhurve**. "Effect of Surface Texture on the Lift and Drag of Small Spinning Balls." *AIAA Aviation Forum and Ascend 2025*. [Paper Link]

HONORS & CERTIFICATIONS

- Machine Learning Specialization, Stanford University & DeepLearning.AI *Jan 2025*
- Recipient of San Diego State University Master's Research Scholarship *May 2022*

Public Repositories: GitHub **Publications:** AIAA