Pranav Singh

Applied AI Engineer | ML Research Engineer
pranav2097@gmail.com| github.com/PS4Emp |https://www.pranavvsingh.com/ | +919971793644

Applied AI/ML engineer with four years of software-engineering experience across healthcare AI, embedded anomaly detection, retrieval systems, and production applications. Currently completing a Master of Artificial Intelligence at UTS with a 6.75/7.00 GPA. My research combines controlled causal experiments, trained-model instrumentation, and closed-loop robotic evaluation to understand memory failures in vision-language-action systems. Experienced in translating research hypotheses into reproducible GPU experiments, auditable evidence, and deployable AI systems.

Education

University of Technology Sydney — Sydney, Australia July 2025 – May 2027(expected)
Master of Artificial Intelligence **GPA: 6.75/7.00**
Relevant coursework: Image Processing, Quantum Computing, Reinforcement Learning, Deep Learning, Data Analytics
Galgotias University — July 2016 – July 2020
Bachelor of Technology in Computer Science and Engineering

Technical Skills

Programming & ML Frameworks: Python, C++, SQL, JavaScript/TypeScript; JAX, PyTorch, TensorFlow/Keras, scikit-learn, Hugging Face Transformers.
Generative AI & Retrieval Systems: Retrieval-Augmented Generation (RAG), agentic AI workflows, tool calling, semantic search, FAISS, ChromaDB, sentence-transformers, embedding pipelines, retrieval evaluation (Recall@K, BEIR), grounding and reliability analysis.
Machine Learning & Research: Deep learning, reinforcement learning, computer vision, CNNs, anomaly detection, time-series modeling, vision-language-action policies, causal intervention experiments, RoboMME, ManiSkill, SAPIEN, OpenCV.
Model Optimization & Edge AI: ONNX, TensorFlow Lite, INT8 quantization, pruning, embedded ML inference, latency and memory profiling, Python/C++ performance optimization.
Backend, Cloud & MLOps: FastAPI, Flask, REST APIs, Docker, AWS (EC2, SageMaker), GitHub Actions, CI/CD, PostgreSQL, Redis, Node.js.

Current Research

MEMTRACE — Causal Memory Interference in Embodied VLA Policies June 2026 - Present
UTS Research | *JAX, RoboMME, MME-VLA, SAPIEN, ManiSkill, Python, GPU Inference*
• Identified downstream RoPE memory-key positions as a causal source of action-level interference in a trained memory-augmented vision-language-action policy while holding the checkpoint, observations, memory content, history length, and inference seeds fixed.
• Designed and audited **81 formal evaluations across 27 robotic episodes and three inference seeds**; a prespecified FAR-alignment intervention reduced action divergence by **0.002948 MAE** on average **(95% CI: 0.001999–0.004049)** and outperformed tested shuffled-position controls.
• Instrumented the trained JAX checkpoint at runtime and verified that changing **432 of 512 memory-key position IDs** preserved the pre-RoPE memory keys and initial queries while changing rotated keys and predicted robot actions.
• Integrated the intervention into online RoboMME inference and completed **45 matched closed-loop rollouts across 15 task episodes**, comparing the native policy, an identity control, and the intervention.
• Found that the intervention activated on **293/293 applicable policy calls but produced zero paired task-success rescues and zero regressions**, demonstrating that a causal action-level correction does not automatically translate into improved robotic task completion.
• Extended the investigation to a TokenDrop memory variant across **27 episodes and three clean-process replications**,

reproducing positional sensitivity while keeping the generalization claim limited to the tested model family.

Recent Research & Technical Projects

Indian Space Research Organization R&AC Centre × NIT Kurukshetra — Research Intern (GNSS / Embedded ML) Nov 2025 – July 2026

- Built an **unsupervised autoencoder** (**277 params; kilobyte-scale footprint**) for **GNSS spoofing detection** on the **TEXBAT** dataset; implemented the full low-compute inference pipeline (**outlier removal → normalization → reconstruction error → adaptive thresholding**).
- Achieved **86.85% TPR @ 2% FPR** on the **most challenging unseen spoofing scenarios (DS7/DS8)** under strict footprint/latency constraints (**448 FLOPs/inference; 0.1 ms per sample** in benchmarks).
- **Evaluated INT8 quantization and pruning strategies** and benchmarked **memory/latency/accuracy tradeoffs** targeting resource-constrained satellite OBCs (**GR740, ARM Cortex-M7**).

Experience

Adsys — SDE / Technical Product Manager April 2023 – July 2025

- Developed an end-to-end computer-vision pipeline for five-part white blood cell (WBC) differential classification, collaborating with clinical teams to curate and label 3,200+ microscopy images and implementing stain normalization, augmentation, class balancing, and five-fold cross-validation.
- Evaluated transfer-learning CNN architectures, including ResNet50, achieving approximately **97% overall classification accuracy** across five WBC categories; conducted class-wise error analysis using confusion matrices, ROC-AUC, and precision-recall metrics.
- Built a cell detection and localization pipeline and integrated it into an offline-first diagnostic application supporting image capture, preprocessing, on-device inference, local storage, and result generation.
- Re-engineered the application from React Native to native Android, improving responsiveness by approximately **40%**; migrated performance-critical image processing and inference components from Python to C++, reducing end-to-end inference latency by approximately **50%** and runtime memory consumption by approximately **30%**.
- Integrated the application with a portable finger-prick CBC device, collaborating with hardware, firmware, and clinical teams to support blood collection, analysis, and five-part differential results in under **10 minutes** across **15+ clinics**.

Halo Health — Founder & CTO Nov 2022 – Jun 2023

- Architected and built a healthcare management platform for clinics and small hospitals, emphasizing workflow optimization and system scalability.
- Reduced patient wait times by **30%** and improved engagement by **20%** with optimized workflows.
- Architected and scaled multi-tenant healthcare platform to support 30000+ concurrent users with sub-second page load times, integrated with Google Sheets and mobile clients.
- Designed end-to-end data flows and automation pipelines to streamline patient intake, scheduling, and reporting.

Pocketfin — SDE May 2021 – Nov 2022

- Scaled fintech apps (Kamao App, DOP Agent App) to 10,000+ active agents, implementing secure payment rails, OTP-based authentication, and basic fraud detection (velocity checks, geo-fencing)
- Increased daily active users (DAU) by 15% and generated $3K+ MRR by optimizing onboarding flows and adding gamification features (referral bonuses, leaderboards)
- Designed backend architecture (Node.js, PostgreSQL) and REST APIs for partner integrations; contributed technical documentation and system design presentations that helped secure $150K funding
- Led healthcare product development (0→1) as Product Owner, coordinating requirements, designing backend (appointment system, user management), and deploying public-facing website

Lumiq — Data Science Intern Jan 2021 – Apr 2021

- Improved bank credit-risk model accuracy by 10 percentage points, from 72% baseline logistic regression to 82% ensemble performance, using feature engineering and hyperparameter tuning.
- Built an end-to-end ML pipeline for credit-risk scoring, covering preprocessing, training, evaluation, and batch scoring integration through a REST API.
- Engineered financial risk features including credit-utilization ratios and payment-velocity signals to improve loan-application assessment.

Projects

RobustVDB — Retrieval Reliability & RAG Evaluation —

• Built a FAISS-based dense retrieval layer that adds result-level explanations, confidence signals, robustness metrics, and hard-query detection for RAG-style systems.

• Implemented retrieval diagnostics including keyword overlap, matched terms, confidence grading, Recall@K, and Robustness-δ@K to make retrieval failures more inspectable.

• Benchmarked hard-query detection on BEIR SciFact and FiQA, finding that a margin-based Clarity signal separated unstable queries more effectively than mean-distance heuristics.

CacheSentry Lite — KV-Cache Stability Auditor for LLM Systems

Python, LLM Systems, Algorithms, JSONL, Unit Testing

- Built a Python CLI that analyzes multi-turn LLM request traces for prompt instability that prevents KV/prefix-cache reuse, using deterministic prompt rendering and token-level longest-common-prefix analysis.
- Implemented character-to-token span mapping to attribute the first divergent token to application fields, with detection rules for UUIDs, timestamps, request IDs, random identifiers, and tool-schema drift.
- Estimated raw and block-aligned reusable tokens and generated Markdown/JSON audit reports with CI-compatible severity checks.
- Detected a synthetic UUID-based cache failure at token 7, reducing the stable-prefix ratio to 10.3% and eliminating all reusable 16-token cache blocks.

Multi-Agent Reinforcement Learning for Negotiation & Strategic Planning | Python, PyTorch, PPO, Imitation Learning, AWS SageMaker, Catanatron

Academic Research Project | UTS | 2026

- Contributed to a multi-agent RL system combining opponent modelling, behaviour cloning, Maskable PPO and simulation-based planning in a partially observable, four-player strategy environment.
- Executed and evaluated three rounds of expert iteration on AWS SageMaker, using Monte Carlo play-outs and learned build-value ranking; improved the checkpoint-selection win rate from **14.2% to 32.5%** against AlphaBeta-based opponents.
- Conducted imitation-learning experiments with planning-aware observations and benchmarked agent performance across multiple opponent configurations, identifying limitations in behavioural cloning for long-horizon decisions.
- Supported reproducible evaluation and artifact management for a final agent achieving **28.3% win rate on 2,000 unseen boards**, outperforming the PPO + offer-chooser baseline by **12.2 percentage points**.

Publications & Recognition

Singh, P. (2025). *Proficient Detection of Leukemia Blood Test Enabled by Deep Learning.* 7th IEEE International Conference on Signal Processing, Computing and Control (ISPCC), pp. 180–184.

Fundraising: Secured **$250K** across Adsys and Pocketfin

UTS: Dean's list

Certifications

Stanford Machine Learning (Coursera, 99.66%)

Data Structures and Algorithms in C (Udemy)