Yug D Oswal

scholar.google.com | yoswal071@gmail.com | yug-oswal.website.io | linkedin.com/in/yugdoswal | github.com/Yug-Oswal

Education

VIT Vellore — Vellore, Tamil Nadu, India

Bachelor of Technology in Computer Science and Engineering, ***CGPA: 9.51/10*** — *Sep. 2022 – Jul. 2026*

Research Experience

Research Fellow — Sep. 2025 – Present

Supervised Program for Alignment Research, with Shivam Raval (Harvard University) — github/SPAR — *USA, Remote*

- Conducting funded research in mechanistic interpretability and AI safety, designing chain-of-thought (CoT) control methods to evaluate and enhance oversight robustness.
- Engineered phase-dependent steering-vector mixtures to induce complex LLM behaviors and elicit CoT-output divergences for stress-testing CoT monitors.
- Induced reward hacking via targeted steering, achieving 99% hacking-classification accuracy on Llama-3.2-1B against the School of Reward Hacks benchmark.

Research Intern — Sep. 2025 – Present

New York University, with Prof. Ravid Shwartz-Ziv — *USA, Remote*

- Constructing a highly realistic tractable-distribution 'perfect' dataset with normalizing flow models like Apple's TarFlow.
- Created oracle queries for posteriors, marginals, and log-probabilities over the flow-generated dataset to enable ground-truth evaluation of learning methods.
- Analyzing representation optimality and aleatoric/epistemic uncertainties across Bayesian, VIB, and related models.

Research Intern — Apr. 2025 – Present

William & Mary, with Prof. Jindong Wang — github/conditioning-reasoning — *USA, Remote*

- Formulated the idea of conditioning LLM reasoning using corresponding outputs, inspired by long-term action-thought feedback loops in human cognition.
- Formalized this into a scalable statistical framework using a KL-divergence-based loss for self-alignment of LLM CoT.
- Achieved 25-35% task-accuracy increase and reduced bias over base DeepSeek Qwen 1.5B on the Bias Benchmark for QA.
- Established a bias-monitoring setup and integrated causal activation patching to evaluate mitigation effectiveness and trace how biased reasoning propagates into model outputs.

Preprints & Research Papers

Loss Switching, Novel Classification and Regression Losses

Yug D Oswal, PI: Mathew Mithra Noel — *Submitted to ESWA, 2025*

- Conceived a gradient-based loss scheduling method, loss switching, complementing new statistically optimized losses.
- Demonstrated accelerated convergence through learning curve analysis, achieving $\geq 3\%$ top-1 accuracy gain on ImageNet by tuning loss-switching with novel classification losses.
- Drafted and executed robustness studies by inducing asymmetric outlier distributions, demonstrating $\geq 1.4\%$ RMSE improvement across 4 regression benchmarks.

Cone-Class of Activations: More Learning, Less Neurons

Yug D Oswal, PI: Mathew Mithra Noel — *arXiv, 2024*

- Hypertuned cone activations computing hyperstrip representations, establishing their suitability for classification heads.
- Orchestrated all experiments, achieving $\geq 4.6\%$ accuracy gain on ImageNet with 46.4% parameter reduction in VGG19.
- Established efficiency-accuracy trade-offs: ≤ 6x neuron compression yields $\approx 2\%$ drop for cone vs. $\approx 8\%$ for ReLU.

Computationally Efficient Quadratic Neural Networks

Yug D Oswal, PI: Mathew Mithra Noel — *arXiv, 2023-2025*

- Designed and implemented vectorized forward/backward matrix algorithms, enabling efficient parallelism and resolving the core computational bottleneck in QNNs.
- Developed $O(n^2)$ reduced-parameter RP-QNN variants and ablated both RP-QNNs and vectorized QNNs to evaluate expressiveness-efficiency tradeoffs.

Professional Experience

AI/ML Engineering Intern — Aug. 2024 – Oct. 2024
Bharat Dynamics Limited - Ministry of Defence, India — *Hyderabad, India*

- Curated a 85,000 sample IR-optical hybrid UAV dataset using MATLAB scripts & the Computer Vision Toolbox.
- Researched and tuned SOTA vision techniques, training strategies, and YOLOv8 for UAV detection and tracking.
- Innovated the first prototype of a multimodal thermal-optical anti-UAV system, successfully tested in 4 field scenarios.
- Architected novel containerization of a unique client-server deployment of the model on air-gapped defence systems.

ML Engineering Intern — Feb. 2024 – Jun. 2024
WebTiga (renamed Synergetics.AI) — *Bangalore, India*

- Implemented classical ML POCs and an audio-based car damage classifier for insurance domain clientele.
- Led the ML lifecycle - from data curation and model training to API development and deployment - for all AI pipelines of a humanoid, speech-capable agent supporting de-addiction therapy.
- Engineered, dockerized, and deployed agentic workflows, guardrails, context-aware chat history, and RAG pipelines for fine-tuned LLMs, with real-time integration into client-used services, reducing latency by 53%.

Project Lead — Jun. 2023 – Aug. 2023
University of Auckland & Signal Corporation Limited — *Auckland & Wellington, New Zealand*

- Spearheaded an international team to resolve 5 real-world issues in Signal's threat intelligence system, improving real-time threat prediction capabilities for executive clients in New Zealand.
- Formulated a scalable pipeline that incrementally clusters live threat data streams, extracts landmarks via NER, and geocodes them into precise coordinates to deliver automated, location-aware threat reports for high-profile clients.

Leadership & Recognition

- **Research Fellow**, Supervised Program for Alignment Research (Fall 2025)
- **Board Member & R&D Head**, Computer Society of India Student Chapter:
  - Directed operations, technical strategy, and cross-team coordination for a 100-member premier student chapter.
  - Mentored junior members in ML & research, guiding them through complex topics, projects, and career development.
  - Organized large-scale events including *Rural Outreach* (40+ rural students), *Riddler* (1000+ participants, 25+ countries), *Lasertag* (1000+), and *Init with CSI* ML workshops (200+).
- **Intel Developer Spotlight Feature**, for my work on Rekindle, a project aiding dementia patients.

Projects

Rekindle — Feb. 2023 – Mar. 2023

- Built an assistive memory-support service for Dementia patients, winning 2nd place in the Intel BOLT Hackathon.
- Trained encoder-decoder emotion-extraction models, outperforming baselines on the Google GoEmotions benchmark.
- Designed an interactive life-journal enabling emotion/event-based memory retrieval to support identity continuity.

Technical Skills

Technical: Python, R, Tensorflow, PyTorch, OpenCV, HuggingFace (transformers), Keras, Scikit-learn, C, C++, SQL, Git, Java, JavaScript, Node.js (Express), Flutter (Dart), Firebase, MongoDB, Redis, Docker
Certifications: EDA and Data Visualization (Scaler), Machine Learning Specialization, Deep Learning Specialization, DeepLearning.AI Tensorflow Developer Professional Certificate, Advanced Techniques in Tensorflow