

# **Yug D Oswal**

[scholar.google.com](http://scholar.google.com) | [yoswal071@gmail.com](mailto:yoswal071@gmail.com) | [yug-oswal.website.io](http://yug-oswal.website.io) | [linkedin.com/in/yugdoswal](https://linkedin.com/in/yugdoswal) | [github.com/Yug-Oswal](https://github.com/Yug-Oswal)

## EDUCATION

### **VIT Vellore**

*Bachelor of Technology in Computer Science and Engineering, CGPA: 9.51/10*

Vellore, Tamil Nadu, India

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*

- Led funded research on causal induction of reward hacking in LLMs via activation steering; evaluated generalization across 100+ held-out samples using probe-based and LM-as-judge metrics.
- Demonstrated controllable reward exploitation: LLaMA-3B achieved +23.5% probe hack probability and +10.3% LM-judge reward under steering, with effects increasing monotonically up to a critical failure threshold.
- Performed layer-wise mechanistic analysis across layers 11-28, revealing strongest reward-hack signals at mid-layers with attenuation at deeper probes, indicating representational constraints on behavioral control.
- Selected among top 10 of 90+ teams to present lightning talk to leading AI safety researchers and organizations; ICML submission in preparation.

### **Research Intern**

Sep. 2025 – Present

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

*USA, Remote*

- Constructing a highly realistic tractable-distribution 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:  $\leq 6x$  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

---

<b>AI/ML Engineering Intern</b>	Aug. 2024 – Oct. 2024
<i>Bharat Dynamics Limited - Ministry of Defence, India</i>	<i>Hyderabad, India</i>
<ul style="list-style-type: none"><li>Curated a 85,000 sample IR-optical hybrid UAV dataset using MATLAB scripts &amp; the Computer Vision Toolbox.</li><li>Researched and tuned SOTA vision techniques, training strategies, and YOLOv8 for UAV detection and tracking.</li><li>Innovated the first prototype of a multimodal thermal-optical anti-UAV system, successfully tested in 4 field scenarios.</li><li>Architected novel containerization of a unique client-server deployment of the model on air-gapped defence systems.</li></ul>	
<b>ML Engineering Intern</b>	Feb. 2024 – Jun. 2024
<i>WebTiga (renamed Synergetics.AI)</i>	<i>Bangalore, India</i>
<ul style="list-style-type: none"><li>Implemented classical ML POCs and an audio-based car damage classifier for insurance domain clientele.</li><li>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.</li><li>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%.</li></ul>	
<b>Project Lead</b>	Jun. 2023 – Aug. 2023
<i>University of Auckland &amp; Signal Corporation Limited</i>	<i>Auckland &amp; Wellington, New Zealand</i>
<ul style="list-style-type: none"><li>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.</li><li>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.</li></ul>	

## 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

---

<b>Rekindle</b>	Feb. 2023 – Mar. 2023
<ul style="list-style-type: none"><li>Built an assistive memory-support service for Dementia patients, winning 2nd place in the Intel BOLT Hackathon.</li><li>Trained encoder-decoder emotion-extraction models, outperforming baselines on the Google GoEmotions benchmark.</li><li>Designed an interactive life-journal enabling emotion/event-based memory retrieval to support identity continuity.</li></ul>	

## 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