

Yug D Oswal

scholar.google.com | yoswal071@gmail.com | yug-oswal.website.io | linkedin.com/in/yugdoswal | 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 - Jun. 2026

PREPRINTS & RESEARCH PAPERS

Computationally Efficient Quadratic Neural Networks

Yug D Oswal, PI: Mathew Mithra Noel

arXiv, 2023-2025

- Resolved the computational bottleneck in higher order ANNs, enabling expressive yet efficient models.
- Designed $O(n^2)$ complexity reduced parameter RP-QNNs and vectorized forward/backward matrix algorithms.

Cone-Class of Activations: More Learning, Less Neurons

Yug D Oswal, PI: Mathew Mithra Noel

arXiv, 2025

- Introduced hyperstrip representations enabling exponentially smaller yet performant foundation models.
- Achieved $\geq 4.6\%$ Top-1 accuracy increase on ImageNet while reducing parameters by 46.4% in VGG19.

Loss Switching, Novel Classification and Regression Losses

Mathew Mithra Noel, Arindam Banerjee*, **Yug D Oswal***, Geraldine Bessie, Venkataraman MN

arXiv, 2024

- Proposed novel loss functions and scheduling strategies accelerating convergence and final model performance.
- Presented regression losses robust to noise and distribution shifts, improving RMSE by $\geq 1.4\%$ in 4 benchmarks.

RESEARCH EXPERIENCE

Research Fellow

Supervised Program for Alignment Research, with Shivam Raval (Harvard University)

Sep. 2025 – Present

USA, Remote

- Researching mechanistic interpretability and AI safety, focusing on CoT monitorability.
- Developing control knobs that regulate CoT-output correspondence to stress-test and enhance CoT monitors.

Research Intern

New York University, with Prof. Ravid Shwartz-Ziv

Sep. 2025 – Present

USA, Remote

- Designing a tractable-distribution dataset using normalizing flows to investigate true optimality of diverse models.
- Analyzing representation optimality and generalization across Bayesian, VIB, and other learning methods.

Research Intern

William & Mary, with Prof. Jindong Wang

Apr. 2025 – Present

USA, Remote

- Leading research to develop novel algorithms utilizing a KL divergence-based loss for alignment of CoT in LLMs.
- Achieving 25-35% task accuracy increase and mitigated bias on the religion subset of Bias Benchmark for QA.

PROFESSIONAL EXPERIENCE

AI/ML Engineering Intern

Bharat Dynamics Limited - Ministry of Defence, India

Jul. 2024 – Oct. 2024

Hyderabad, India

- Curated a dataset for UAV detection using MATLAB scripts and the Computer Vision Toolbox.
- Researched and tuned SOTA vision techniques, training strategies, and YOLOv8 for UAV detection and tracking.
- Developed the first prototype of a UAV detection & tracking system, successfully tested in 4 field scenarios.
- Engineered unique edge and web migratable model deployments on isolated systems following defence policies.

ML Engineering Intern

WebTiga (renamed Synergetics.AI)

Feb. 2024 – Jul. 2024

Bangalore, India

- Developed classical ML POCs and an audio-based car damage classifier for insurance domain clientele.
- Led the ML lifecycle, from data curation, model training, to API creation and deployment, for all AI pipelines of a humanoid speech-capable autonomous agent serving 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

University of Auckland & Signal Corporation Limited

Jun. 2023 – Aug. 2023

Auckland & Wellington, New Zealand

- Led 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.
- Developed 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 chapter operations and led technical and managerial initiatives.
 - Mentored junior members in ML and research, providing exposure to advanced topics.
 - Organized large-scale events including *Riddler* (1000+ participants, 25+ countries), *Lasertag* (1000+), and *Init with CSI* ML workshops (200+).
- **Intel Developer Spotlight Feature**, in recognition of my work on Rekindle, a hackathon-winning project aiding Dementia patients.

SELECTED PROJECTS

- | | |
|--|-----------------------|
| Rekindle | May 2023 – May 2023 |
| <ul style="list-style-type: none">• Developed a service to aid Alzheimer's and Dementia patients, winning 2nd place in the Intel (BOLT) Hackathon.• Trained custom encoder-decoder NLP models with novel loss functions and distributed strategies, outperforming baselines on the Google GoEmotions benchmark. | |
| Inducing Complex Behaviours in LLMs | Aug. 2025 – Sep. 2025 |
| <ul style="list-style-type: none">• Inducing complex behaviors in LLMs to stress-test and improve AI safety oversight mechanisms.• Developing phase-dependent steering vector mixtures to robustly probe behaviors and elicit CoT-output divergences. | |
| Bias Monitor for Safety Tuned LLMs | Aug. 2025 – Sep. 2025 |
| <ul style="list-style-type: none">• Built a bias monitoring setup to evaluate mitigation effectiveness in LLMs tuned with novel alignment algorithms.• Identified alignment suppression and bias amplification patterns, revealing hidden disparities in model behavior.• Integrated causal activation patching to trace how biased reasoning propagates into model outputs. | |

TECHNICAL SKILLS

Technical: C, C++, Python, SQL, Git, Tensorflow, PyTorch, Keras, Java, JavaScript, Node.js (Express), Flutter (Dart), Firebase, MongoDB, Redis, Docker, Scikit-learn, R

Certifications: EDA and Data Visualization (Scaler), Machine Learning Specialization, Deep Learning Specialization, DeepLearning.AI Tensorflow Developer Professional Certificate, Advanced Techniques in Tensorflow