# Nishal Thomas

nishal.thomas44@gmail.com • nishal14.github.io • Nishal14

## Education

**St. Mary's Sr. Sec. School** — **New Delhi, India**
*Class 11* — *Ongoing*
*Class 10 – 91%* — *2025*

## Research and Technical Projects

**Step-Free Arithmetic Transformers**
*Mechanistic Interpretability Study*

- Designed and trained compact transformer models (0.66M–10M parameters) to solve arithmetic expressions under final-answer-only supervision
- Conducted mechanistic interpretability experiments including targeted attention head ablation, linear probing, and out-of-distribution evaluation
- Identified attention heads specialized for structural processing and demonstrated their causal role in model generalization
- Showed that parenthesis depth is linearly decodable from internal activations, revealing structured representations inside the model
- Built a fully reproducible research pipeline including dataset generation, training infrastructure, and interpretability analysis tools
- Authored a beginner-friendly technical blog explaining mechanistic interpretability concepts and project findings: Link

**Grokking Feasibility Boundaries in Transformers**
*Empirical Study of Generalization Dynamics*

- Systematically investigated grokking in small transformer models by varying output-space overlap in modular arithmetic tasks
- Designed controlled experimental regimes spanning interpolation, boundary, and extrapolation settings to map where grokking emerges and fails
- Analyzed internal model dynamics using representation rank and attention entropy to detect algorithmic phase transitions
- Demonstrated a sharp feasibility boundary: delayed grokking appears only within a narrow interpolation regime and collapses under structured overlap constraints
- Built a reproducible experimental framework with automated training pipelines, post-training analysis, and comparative visualization tools
- Released a fully documented open-source repository with final plots and experimental summaries

**Continuum (Extension)**
*Epistemic Drift Monitoring System*
- Built a Chrome extension (Manifest V3, React + TypeScript) that monitors large language model conversations for contradictions and reasoning instability in real time
- Designed a backend reasoning engine (FastAPI) implementing cumulative drift accumulation, longitudinal stance tracking, and structural dependency analysis
- Engineered a hybrid escalation architecture where lightweight heuristics operate continuously and selectively escalate high-drift states to K2 Think V2 for authoritative verification
- Modeled conversations as structured commitment graphs with lifecycle tracking, contradiction edges, and stability scoring
- Developed an animated visualization panel displaying drift metrics, structural state transitions, and K2 verification outcomes with confidence scores
- Integrated async K2 verification with graceful fallback, enabling performance-optimized local analysis (¡200ms) with strategic reasoning escalation

**Smart Traffic Light System**
*Integrated Computer Vision + Embedded Controls*
- Built an intelligent traffic control system combining real-time computer vision with embedded hardware for traffic signal automation
- Developed a Python pipeline using YOLOv11 to detect vehicle density from webcam input and determine optimal light phases
- Implemented bidirectional serial communication with an Arduino Uno to control LED-based traffic lights in multiple modes (automatic, manual, emergency)
- Designed and programmed an Arduino sketch for robust physical signal sequencing and emergency behavior
- Integrated smoothing and decision logic to stabilize detection and minimize flickering during live operation
- Packaged project in an open-source repository with full documentation, hardware wiring, and software setup instructions

## Awards

**2025**: **New Delhi YMCA Award for Academic Excellence** — Recognized for outstanding performance in Class 10 Board Examinations.

**2025**: **Science Subject Topper — Class 10** — Awarded for highest performance in science at St. Mary's Sr. Sec. School.

**2025**: **Perfect Score in Artificial Intelligence (100/100)** — Achieved full marks in Class 10 Board Examination; recognized by school with academic distinction award.