Twisha Mehta

| twisha1.mehta@gmail.com | linkedin.com/in/twisha-mehta1/ | github.com/ahsiwt101

EDUCATION

National University of Singapore (NUS) **Aug 2025 – *May 2029***
Bachelor of Computing in Computer Science, Second Major in Mathematics

Achievements & Hackathons: Top Student in AI Ethics, Apple Spatial AI Hacks (2nd), Acacia AI Launchpad Challenge (Finalist)
Clubs: Google Developer Group @ NUS Computing (Technology Associate), NUS Women in Tech (Exco)
Invite-Only Events: **Jane Street** First Year Trading and Tech Program, **Optiver** Exclusive Insights Day

EXPERIENCE

***Incoming* Generative AI Intern – Razer,** Singapore **Jan 2027** – **Jun 2027**

AI Algorithm Researcher – Huawei, Singapore **Dec 2025** – **Nov 2026**
- Contributed to **DuoRate**, a research paper on per-expert precision allocation for Mixture-of-Experts serving, released on arXiv in Oct 2026 (arXiv:2610.13328); at 2.9 bits per weight it matches 4-bit formats on Qwen3-30B-A3B. Submitted to ICML and CVPR.
- Built a W2A16 post-training quantization pipeline for Qwen3-VL-30B, targeting 8x compression on Ascend 910B NPUs.
- Analyzed MoE expert-routing behaviour across vision-language workloads to guide where precision is spent in compression.
- Integrated LightRAG into a TTD-DR agent pipeline, cutting redundant web searches by 60–80% and LLM calls per query by 40–60% through embedding-similarity gating.

Data Analyst Intern – Foresight Edge, Singapore **Jun 2024 – Jul 2024**
- Built a PowerBI dashboard for sales data across 8 SME clients, enabling real-time profit and revenue tracking for stakeholders
- Developed Python-based sales forecasting and churn models to identify at-risk accounts and support client retention planning.

Software Development Intern – Yasham Software Services Pvt Ltd, India **Oct 2023 – Jan 2024**
- Built a Laravel/MySQL CRM system to replace manual workflows, improving data accuracy and reducing admin overhead.
- Implemented authentication, role-based access control, and transaction processing for internal CRM and inventory operations

PROJECTS & RESEARCH

Inference Gateway – Model-Serving Systems **May 2026 – July 2026**
- Built a from-scratch model-serving gateway (Python/FastAPI) with dynamic batching, async queue, LRU caching, and backpressure; improved throughput 2.7× (65.5→174.5 req/s) and cut p99 latency 4× (792ms→190ms) vs. naive baseline
- Validated graceful degradation under overload (200 concurrent requests: 276/300 failed fast with clean 503s vs. crash) and instrumented Prometheus metrics + CI pipeline via GitHub Actions

CostPerThought – Inference Cost Benchmarking **July 2026 – Aug 2026**
- Built a full factorial benchmark measuring how weight quantization (AWQ-INT4), KV-cache quantization (FP8), and reasoning-token budget caps compose, on DeepSeek-R1-Distill-Qwen-7B served via vLLM on GPU hardware
- Revealed that weight-quant and budget-capping degrade additively while uncalibrated FP8 KV-cache breaks accuracy outright
- Used bootstrap confidence intervals to statistically validate interaction effects across the compression grid

LEADERSHIP

Head of App Development – GIIS Tech Club, 501(c)(3) Non-Profit, Singapore **Dec 2023 – Dec 2024**
- Led app development workshops and organized 10th edition of GIIS Hackathon for 600+ students across 4 countries in APAC.

Cofounder, Director of Technology – InfoAir, Singapore **Jul 2023 – Jan 2024**
- Designed an Arduino-based particulate monitoring prototype; deployed website as part of a student-led social-impact venture.

AWARDS / OLYMPIADS

SOF International Mathematics Olympiad - Bronze Medal, Top 1.1% globally
Global Future-Ready Scholarship – Awarded to the top-performing student in cohort

SKILLS

Languages: Python, Java, C, C++, TypeScript, PHP, JavaScript, SQL, R
Frameworks & Libraries: React, React Native, Node.js, Next.js, Laravel, Flutter, PyTorch
DevOps & Tools: Git, Docker, AWS, Linux, Ascend NPUs