Aryan Gupta

London, United Kingdom | +44 750-750-9180 | aryanguptaa2022@gmail.com | linkedin.com/in/mraryangupta

EDUCATION

Richmond American University London, United Kingdom | Bachelor of Computer Science **Jan'23 - Apr'26 (Expected)**
Courses: AI/ML, Web Technology, Database System, Computer Architecture, Algorithms, Cloud Security **GPA:3.7/4.0**

Springdales School, New Delhi, India | High School, Central Board of Secondary Education (CBSE) **Aug'06 - Jun'21**
Courses: Computer Science, Mathematics, Business Studies, Accounting, Economics, English **Percentage: 95%**

SKILLS

Programming Languages: Python, Java, C/C++, JavaScript, Swift, SQL, HTML5, CSS3
Frameworks, Tools & Development: Django, React.js, iOS & Web Development, MongoDB, Redis, REST APIs, Git, LangChain
AI/Cloud Technologies: AWS, Vector DBs - Pinecone, Weaviate, LLMs, RAG, Agentic AI, Performance Profiling & Engineering

EXPERIENCE

Applied AI Software Engineering Inten | Mphasis, London, United Kingdom Jun'25 - Sept'25

- Benchmarked 10 LLMs across latency, cost, and task complexity; engineered intelligent routing framework with Redis-based semantic caching reducing inference costs by 70%.
- Automated LangChain evaluation pipelines improving benchmarking speed and reproducibility while maintaining sub-50ms latency in production.

AI/ML Research Assistant – Dr. Jane Norris | Richmond American University London, UK Feb'25 – Dec'25

- Developed Persona AI framework integrating socio-technical systems (STS) for ethical AI governance aligned with UN.
- Led STS-informed AI ethics research for Warwick BELAP, bridging computer science with critical technology studies

Software Development Engineer Inten | Commudle, New Delhi, India Jun'24 – Aug'24

- Learned SDLC best practices to deliver responsive web features for 250,000+ developers globally using HTML5, CSS3, JavaScript, & contributed UI/UX components for event features supporting hackathons & workshops like GDG DevFest.

PROJECTS

Context-Aware AI Assistant with Embedded Persona, Self-initiated project.

- Built conversational AI using Ollama with custom persona embedding and RAG pipeline using vector databases for efficient context management within 8K token budgets.
- Integrated Google Text-to-Speech SDK for voice interaction; engineered context pruning strategies to maintain coherence on resource-constrained laptop hardware.

CPU-Based LLM Inference & Quantization, Self-initiated project - inspired by Intel Labs.

- Benchmarked LLM inference across FP32, BFloat16, INT8 precisions with/without Intel AMX instructions, revealing memory bandwidth as primary performance bottleneck.
- Profiled performance using Intel VTune, Emon analyzing CPI and LLC miss rates, validating that data movement - not computation - dominates AI inference costs.

Safest Path Navigation for Late-Night Commuters, IC Hack'25 - Hackathon winning project.

- Developed heuristic routing algorithm computing 'public presence scores' from Google Places API data (foot traffic, lighting, commercial activity) using bounding box segmentation.
- Applied K-Nearest Neighbors clustering to Metropolitan Police crime data, fusing time-of-day patterns with real-time context to generate holistically safer routes.

PUBLICATIONS

- Analyzing the compression Function of newly proposed SHA-256 and 64-Bit Architecture – SN Computer Science Journal
- Early Warning Systems to Predict Student Attrition: A cross Domain Framework – Book Chapter, NIPA Publishers
- Predicting Curie Temperature & Magnetic Anisotropy in Quantum Materials for Efficient Computing – MRS Spring 2026
- Data-Driven Screening of Magnetic Materials for Aerospace Applications – TMS 2026, San Diego, CA

ACTIVITIES

Dean's Scholar 2023-25 | President South Asian Student Society, SGA Treasurer & Senate, Peer Mentor & TA (500+ students)