Syed Muhammad Taha

syedmuhammadtaha.dev@gmail.com | +92 334 0518269
Website | GitHub | LinkedIn

Research Interests

Agentic AI under resource constraints: long-horizon agents that run on small models and limited compute. How should an agent plan, use tools and manage its context over many steps when every model call has a cost, and how can those choices be made reliable rather than hand-tuned? *Foundation from current work:* model compression, early exit, and statistical guarantees for efficiency decisions, most recently in small-model tool calling.

Education

National University of Sciences & Technology (NUST), Islamabad, Pakistan 2022 – 2026
B.S. in Computer Science, School of Electrical Engineering & Computer Science cGPA: 3.5/4.0
Relevant coursework: Linear Algebra, Probability & Statistics, Machine Learning, Deep Learning. *Honors:* Dean's List (5 semesters).

Publications

Peer-Reviewed Conference Papers

[C1] Tayyab Raza, **Syed Muhammad Taha Imam**, Adrian Ulges, Ulrich Schwanecke, Momina Moetesum, Faisal Shafait. *LiteDoc: Distilling Large Document Models into Efficient Task-Specific Encoders.* In *Document Analysis and Recognition – ICDAR 2026*, Lecture Notes in Computer Science, pp. 418–434. Springer, 2026. [DOI]

Under Review

[S1] **Syed Muhammad Taha Imam**, Hafiz Muhammad Ali Zeeshan, Syed Ali Hassan, Stefano Cirillo, Syed Ishtiaque Ahmed, Mohammad Mehedi Hassan. *ESCORT: Escalation with Statistical Certification of Risk Targets for Language-Model Service Brokers.* Submitted to *IEEE Transactions on Services Computing*, 2026.

[S2] **Syed Muhammad Taha Imam**, Hafiz Muhammad Ali Zeeshan, Rohaan Ahmad, Syed Ali Hassan. *AMEND: Agent-Mediated Errata for Noisy Datasets of Function-Calling Benchmarks.* Submitted to *IEEE ICC 2027* (Big Data track).

[S3] Rohaan Ahmad, Hafiz Muhammad Ali Zeeshan, **Syed Muhammad Taha Imam**, Kapal Dev, Haejoon Jung, Syed Ali Hassan. *Flight Plans in Plain Sight: Intent Leakage in Agentic Service Discovery for Low-Altitude Wireless Networks.* Submitted to *IEEE Vehicular Technology Magazine*, 2026.

[S4] Rohaan Ahmad, Hafiz Muhammad Ali Zeeshan, **Syed Muhammad Taha Imam**, Syed Ali Hassan. *Observable Participation: A Passive Aerial Threat to Secure Aggregation in HAPS Federated Learning.* Submitted to *IEEE ICC 2027* (Communication and Information Systems Security track).

In Preparation

[P1] **Syed Muhammad Taha Imam** et al. *Early Exit in a Wireless Foundation Model When the Signal-to-Noise Ratio Must Be Estimated.* In preparation for *IEEE Transactions on Machine Learning in Communications and Networking*, 2026.

Research Experience

Researcher, IPT Lab, SEECS, NUST Jun 2026 – present

- Certified small-to-large model escalation for tool calls [S1]: 67% of requests stay on the small model at a 30% error target, where tuned thresholds broke their target about half the time.
- Agent-based repair of function-calling benchmark keys [S2]: over 1 in 10 BFCL Live and ACEBench items reject a correct call; the repairs raise a 32B model's measured accuracy by up to 12 points.
- Early exit with statistical guarantees in a wireless foundation model [P1]: 14–39% of encoder compute saved at error $\leq \alpha$ with 95% confidence; the guarantee breaks when SNR is estimated, and certifying on estimated groups restores it.
- Co-authored two studies on what encryption fails to hide: agents' service discovery reveals a drone's mission (0.947 balanced accuracy, chance 0.25) [S3]; uplink direction reveals FL participants [S4].

Research Intern, Bradbury Lab (independent research group, remote) Apr 2025 – Jan 2026

- Reviewed literature on Transformer topology and parameter-efficient fine-tuning (weight sharing).
- Proposed a training-free Tucker-decomposition layer-merging method; analyzed self-attention to assess the feasibility of aligning Query (Q) and Key (K) projections in efficient-by-design architectures.

Undergraduate Research Intern, MachVis Lab, NUST Oct 2025 – Dec 2025

- LLM-based factual verification of generated pathology reports beyond BLEU/ROUGE; curated a challenge set of 100+ gigapixel whole-slide images with real-world artifacts and stain variation.

Undergraduate Research Intern, NUST (remote) Jun 2025 – Sep 2025

- Combined MedSAM with meta-learning for few-shot segmentation of dental radiographs, improving segmentation of scarce tooth-disease classes by 12%.

Teaching, Leadership & Awards

- **Teaching Assistant, Linear Algebra**, NUST (Jan 2024 – May 2024): delivered lectures to 40+ undergraduates on applications of linear algebra in machine learning.
- **Volunteer Tutor, CS & Mathematics (O-Levels)** (Jan 2023 – Jun 2023): one-on-one pro-bono tutoring.
- **Executive**, NUST Science Society; **Team Lead**, Google Developer Student Clubs (GDSC), NUST.
- **Winner**, Epistemy hackathon (70+ participants; awarded a paid internship); **Finalist**, VisionRD Hackathon.
- **Kaggle** (ongoing): Google – The Gemma 4 Developer Agent Competition; Google Quantum AI – Verifiable Quantum Circuit Compilation; Name That LLM.

Professional Experience

AI Engineer, Brainbox Automations Dec 2025 – present

- Deployed a guardrailed AI coaching backend on GCP Cloud Run (FastAPI, Vertex AI) serving 300+ athletes, including minors; cut CI/CD deployment time from about 2 minutes to 10 seconds.
- Built SQL-aware agents for a RAG chatbot over 50+ Supabase tables, halving response time.
- Built chat and voice AI (VAPI) for an Australian automation firm: 200+ daily interactions, 5,000+ calls handled, 80% resolved without a human; summarized hour-long call transcripts with parallel map-reduce on AWS Lambda.

Software Engineer (AI) & Team Lead, Epistemy UK Sep 2025 – Dec 2025

- Built an AI tutoring platform end to end (Nest.js, Next.js, Redis/BullMQ multi-agent queue) with 90%+ test coverage; led a team of two engineers.

Machine Learning Intern, CogniMind AI, NSTP, NUST Feb 2025 – Apr 2025

- Improved VLM extraction accuracy by 20% and retrieval/inference efficiency by 10% (quantization, HNSW); deployed Dockerized Airflow (5+ DAGs) and CI/CD, cutting deployment time by 30%.

Machine Learning Intern, RapidsAI, SINES, NUST Sep 2024 – Dec 2024

- Built a RAG chatbot with multi-model query routing, halving API costs.

Selected Projects

GPT-2 from Scratch: Pretraining, Efficient Inference and Fine-Tuning gpt-2

- Pretrained GPT-2 from scratch in PyTorch on FineWeb-Edu (10B tokens) with FlashAttention.
- Implemented KV caching and speculative decoding; instruction-tuned with SFT on Alpaca.

RAG Tooling: create-rag-app and Agentic-RAG create-rag-app | agentic-rag

- A Python CLI that scaffolds Dockerized RAG apps (13 Jinja2 templates, 3 retrieval strategies), and a LangGraph planner/replan agent evaluated with RAGAS at >95% faithfulness.

Technical Skills

Methods	Pruning, early exit, distillation, LoRA/PEFT, learn-then-test, LLM cascades
ML & Systems	Python, PyTorch, CUDA, FlashAttention, HuggingFace, vLLM, quantization
Engineering	Git, Docker, CI/CD, GCP Cloud Run, AWS Lambda, Airflow, Redis, FastAPI
Languages	English (IELTS 8.0), Urdu (native), German (beginner)