Shikhar Tuli

Co-founder, CTO at Illumia AI
Senior Research Scientist at Samsung Research
665 Clyde Ave, Mountain View, CA, 94043

stuli@illumia.ai
github.com/shikhartuli
Google Scholar, Homepage

Academic Details

Year	Degree	Institute	CGPA/Percentage
2020-2023	Ph.D. in Elec. and Comp. Eng.	Princeton University	3.9/4.0
2016-2020	B.Tech in Electrical Engineering	Indian Institute of Technology Delhi	9.5/10.0
2016	Class XII, CBSE	Amity International School	96.6%
2014	Class X, CBSE	Amity International School	10.0/10.0

Research Interests

- **Efficient Machine Learning:** Exploring models and hardware architectures for efficient training and inference EML.
- **Artificial General Intelligence:** Mechanistic intepretability MI, neuroscience-inspired AI NI-AI, and neuro-symbolic AI NS-AI.

Professional Appointments

- Senior Research Scientist (On-device AI/ML) at Samsung AI Center, Mountain View. *Jan 2024 - Present*
- Research Intern at Samsung AI Center, Mountain View *May 2023 - Aug 2023*
- Research Associate at CoCoSci Lab, Princeton University. *Jan 2021 - July 2021*
- Research Associate at NAITS Lab, IIT Delhi. *Jan 2020 - July 2020*
- Research Intern at Embedded Systems Lab, EPFL. *May 2019 - Aug 2019*
- Research Intern at DWLCL, IIT Delhi. *May 2018 - Nov 2019*
- Research Intern at CLOUDS Lab, University of Melbourne. *May 2018 - July 2019*
- Founder and CEO. Qubit Inc., Noida. *Jan 2020 - Apr 2022*
- Research Consultant at Coral Telecom Ltd., Noida. *Apr 2016 - Nov 2021*

Awards and Achievements

- Received **Samsung Best Paper Award** for the best paper among all Samsung AI Research centers.
- Received **School of Engineering and Applied Sciences (SEAS) Award for Excellence** at Princeton University.
- Received **Pramod Subramanyan *17 Early Career Graduate Award** at Princeton University.
- Awarded **Ph.D. Fellowship** for the first year of study.
- Received **Rajiv Bhambawale Award for Best B.Tech thesis** at the undergraduate level.
- Awarded **ThinkSwiss Research Scholarship** for a summer internship at Embedded Systems Laboratory (ESL), EPFL under the E3 program.
- Received **Summer Undergraduate Research Award** for outstanding research at undergraduate level.
- Received **Design Innovation Summer Award** (DISA 2017) and **DIT Seed Grant** at undergraduate level.
- Placed among the **Top 7%** of IIT Delhi in the first, second, fifth, and seventh semesters based on academic performance.
- Won **2nd Runners Up**, **Best Mechanical Design Award**, and **Best Technical Report** Cash Prize for Bomb Disposal Robotics National Competition at IIT Kharagpur (December 2016).
- Secured **All India Rank 1624** in Joint Entrance Exam Advanced 2016 among 150,000 candidates.
- Awarded **Chairperson's Trophy** for being the **School Topper**.

Selected Research Projects

Efficient State-space Model Inference Engine EML — Samsung AI Center
Industrial Project — Jan 2024 - Present
Implementing efficient and high-performing state-space models on-device. This requires optimized implementations of the parallel associative scan algorithm on the ARM platform. The proposed novel architecture achieves high prefill and generation speeds (in terms of tokens/s) relative to baselines based on the transformer/SSM architecture. The proposed architecture and inference engine are slated to be released in Galaxy AI 2025. More details upon request.

Transformer Inference Acceleration EML NI-AI — Samsung AI Center
Industrial Project — May 2023 - Jan 2024
Implementing multi-token prediction large language models (LLMs). The proposed models dynamically predict multiple tokens based on their confidence in the predicted joint probability distribution. Designing a lightweight technique to train these models, leveraging the weights of traditional autoregressive counterparts. One of the models in our suite, DynaMo-7.3B-T3, achieves same-quality generated text as the baseline (Pythia-6.9B) while achieving 2.57× speed-up with only 5.87% and 2.67% parameter and training time overheads, respectively. Project webpage link.

Graph Language Models MI NS-AI — Jha Lab, Princeton University
Research Project — Feb 2023 - Present
Formulating graph-language models (GLMs) that combine self-supervision in LLMs and expert-verified knowledge graphs (KGs) that are the gold standard for accurately responding to user requests. The proposed GLM encapsulates concepts from high-quality text into a dense and representative KG, obviating the need for retrieval-augmented generation (RAG). Our proposed framework also includes an intelligent question-answering system that processes user requests and answers questions leveraging the extracted KG. This is a novel regime of training neuro-symbolic models leveraging self-supervision at scale, thus opening new avenues for building responsible and trustworthy artificial general intelligence (AGI) of the future.

Transformer-accelerator Co-design EML — Jha Lab, Princeton University
Research Project — Aug 2022 - May 2023
Designed and developed a novel HW/SW co-design framework that designs transformers along with the accelerator chip to which it would be mapped. Proposed novel neural architecture search (NAS) techniques along with an expanded suit of models and hardware accelerators for energy-efficient designs. Resultant transformer-accelerator pair achieves 0.3% higher accuracy than BERT-Base and incurs 212× lower latency than an A100 GPU. Accelerator simulator repository link.

Exploration of the Transformer Design Space EML — Jha Lab, Princeton University
Research Project — Dec 2021 - May 2023
Studied the possible design decisions for the transformer architecture along with various training recipes in order to find the best architecture for each task in the natural language processing (NLP) domain. Heterogeneous and flexible architectures have shown to outperform traditional homogeneous and rigid models that have the same set of hyperparameters across all layers in the network. Resultant models achieving similar performance as baselines are 2.6× smaller. The best-performing models outperform baselines with up to 8.9% higher GLUE score. Framework repository link.

Inductive Biases in CNNs and Transformers MI NI-AI — CoCoSci Lab, Princeton University
Research Project — Jan 2021 - Jul 2021
Studied various human inductive biases on common computer-vision models including CNNs and transformers. Trained and evaluated models on the stylized Imagenet dataset to test shape/texture biases. Observed that biases in transformers are more consistent with that of humans. Collaborated with academics from different institutions. Repository link.

Supervised and Unsupervised Spiking Neural Networks NI-AI — Prof. Debanjan Bhowmik, IIT Delhi
Research Project — Aug 2019 - Nov 2019
Simulated supervised and unsupervised spiking neural networks employing STDP learning (thesis link). Implemented code-level and circuit-level simulations of a novel neuromorphic system capable of learning common machine learning benchmarks.

Automated Qubit Design — Houck and Jha Labs, Princeton University
Research Project — Sept 2022 - Present
Developing a machine-learning-based approach to explore high-coherence qubits. We formulate the qubit design as a graph optimization problem. We implement multi-objective optimization to maximize the T_1 and T_2 decoherence times along with the quantum gate speed using machine learning.

Publications

Updated list of publications with software repositories, datasets, and preprint links can be found on my website.

Refereed Conference and Workshop Publications

C9. NAACL '24 — Chi-Heng Lin, Shikhar Tuli, James Seale Smith, Yen-Chang Hsu, Yilin Shen, Hongxia Jin. *SLiM: Speculative Decoding with Hypothesis Reduction.* North American Chapter of the Association for Computational Linguistics, 2024. [acc. rate: 23.2%]. (link).

C8. NAACL '24 — Shikhar Tuli, Chi-Heng Lin, Yen-Chang Hsu, Niraj K. Jha, Yilin Shen, and Hongxia Jin. *DynaMo: Accelerating Langauge Model Inference with Dynamic Multi-Token Sampling.* North American Chapter of the Association for Computational Linguistics, 2024. [acc. rate: 23.2%]. (link).

C7. NeurIPS '21 — Shreshth Tuli, Shikhar Tuli, Giuliano Casale, and Nicholas R. Jennings. *Generative Optimization Networks for Memory Efficient Data Generation.* NeurIPS 2021 - Workshop on ML for Systems. [acc. rate: 9.2%]. (link).

C6. CogSci '21 — Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. *Are Convolutional Neural Networks or Transformers more like human vision?* Annual Meeting of the Cognitive Science Society, 2021. [acc. rate: 18.2%]. (link).

C5. ICONS '20 — Shikhar Tuli and Debanjan Bhowmik. *Design of a Conventional-Transistor-Based Analog Integrated Circuit for On-Chip Learning in a Spiking Neural Network.* International Conference on Neuromorphic Systems, 2020. (link).

C4. ISCAS '20 — Shikhar Tuli and Shreshth Tuli. *AVAC: A Machine Learning based Adaptive RRAM Variability-Aware Controller for Edge Devices.* IEEE International Symposium on Circuits and Systems, 2020. (link).

C3. ASP-DAC '20 — Shikhar Tuli, Marco Antonio Rios, Alexandre Sébastien Julien Levisse, and David Atienza Alonso. *RRAM-VAC: A Variability-Aware Controller for RRAM-based Memory Architectures.* Asia and South Pacific Design Automation Conference, 2020. (link).

C2. Cloudcom '19 — Shreshth Tuli, Shikhar Tuli, Udit Jain, and Rajkumar Buyya, *APEX: Adaptive Ext4 File System for Enhanced Data Recoverability in Edge Devices.* International Conference on Cloud Computing, 2019. (link).

C1. DAC '19 — Neetu Jindal, Sandeep Chandran, Preeti Ranjan Panda, Sanjiva Prasad, Abhay Mitra, Kunal Singhal, Shubham Gupta, and Shikhar Tuli, *DHOOM: Reusing design-for-debug hardware for online monitoring.* Design and Automation Conference, 2019. (link).

Refereed Journal Publications

J12. TMC '23 — Shikhar Tuli, Niraj K. Jha. *EdgeTran: Device-Aware Co-Search Of Transformers for Efficient Inference on Mobile Edge Platforms.* IEEE Transactions on Mobile Computing, 2023 (link).

J11. TCAD '23 — Shikhar Tuli, Niraj K. Jha. *TransCODE: Co-designing Transformers and Accelerators for Efficient Training and Inference.* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023 (link).

J10. TCAD '23 — Shikhar Tuli, Niraj K. Jha. *AccelTran: A Sparsity-aware Accelerator for Dynamic Inference with Transformers.* IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2023 (link).

J9. JAIR '23 — Shikhar Tuli, Bhishma Dedhia, Shreshth Tuli, Niraj K. Jha. *FlexiBERT: Are Current Transformer Architectures too Homogeneous and Rigid?.* Journal of Artificial Intelligence Research, 2023 (link).

J8. TECS '23 — Shikhar Tuli, Chia-Hao Li, Ritvik Sharma, Niraj K. Jha. *CODEBench: A Neural Architecture and Hardware Accelerator Co-Design Framework.* ACM Transactions on Embedded Computing Systems, 2023 (link).

J7. Nature '22 — Shikhar Tuli, Niraj K. Jha. *DINI: Data Imputation using Neural Inversion for Edge Applications.* Nature Scientific Reports: Special Track on Edge intelligence for the next generation Internet of Things, 2022 (link).

J6. MedRxiv '20 — Shreshth Tuli, Shikhar Tuli, Ruchi Verma, and Rakesh Tuli. *Modelling for prediction of the spread and severity of COVID-19 and its association with socioeconomic factors and virus types.* MedRxiv (2020). link.

J5. IoT '20 — Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, and Sukhpal Singh Gill. *Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing.* Internet of Things (2020). link.

J4. ITL '20 — Shreshth Tuli, Shikhar Tuli, Gurleen Wander, Praneet Wander, Sukhpal Singh Gill, Schahram Dustdar, Rizos Sakellariou, Omer Rana, *Next Generation Technologies for Smart Healthcare: Challenges, Vision, Model, Trends and Future Directions*, Internet Technology Letters. link. **Top downloaded article award** link.

J3. IoT '20 — Sukhpal Singh Gill, Shreshth Tuli, Minxian Xu, Inderpreet Singh, Karan Vijay Singh, Dominic Lindsay, Shikhar Tuli, et al. *Transformative Effects of IoT, Blockchain and Artificial Intelligence on Cloud Computing: Evolution, Vision, Trends and Open Challenges*, Internet of Things, Volume 8. link.

J2. TED '20 — Charu Gupta, Anshul Gupta, Shikhar Tuli, Erik Bury, Bertrand Parvais, and Abhisek Dixit. *Characterization and modeling of Hot Carrier Degradation in N-Channel Gate-All-Around Nanowire FETs.* IEEE Transactions on Electron Devices, 2020. link.

J1. JSS '19 — Shreshth Tuli, Redowan Mahmud, Shikhar Tuli, and Rajkumar Buyya. *FogBus: A Blockchain-based Lightweight Framework for Edge and Fog Computing.* Journal of Systems and Software, Volume 154, 2019, Pages 22-36, link. **Top ten downloaded article of 2019 award** link.

Under Review and Work-in-progress Articles

W7. ICML '25 — Shikhar Tuli, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, *Hydra: Mixture of State-Space Experts is a Multi-Head Attention.* International Conference on Machine Learning, 2025.

W6. NAACL '25 — James Seale Smith, Chi-Heng Lin, Shikhar Tuli, Abhishek Patel, Yen-Chang Hsu, Yilin Shen, Hongxia Jin. *FlexiGPT: Pruning and Extending Large Language Models with Low-Rank Weight Sharing.* North Americab Chapter of the Association for Computational Linguistics, 2025.

W5. ICLR '25 — Chi-Heng Lin, Shangqian Gao, James Seale Smith, Abhishek Patel, Shikhar Tuli, Yilin Shen, Hongxia Jin, Yen-Chang Hsu. *MoDeGPT: Modular Decomposition for Large Language Model Compression.* International Conference on Learning Representations, 2025 (under review; preprint link).

W4. EMNLP '25 — Margarita Belova, Shikhar Tuli, Suma Bhat, Niraj K. Jha. *Graph Language Models: Distilling Reliable Knowledge Graphs from High-Quality Text.* Empirical Methods in Natural Language Processing, 2024.

W3. Nature '24 — Shikhar Tuli, Shashwat Kumar, Niraj K. Jha, Andrew A. Houck. *GraphQ: High-coherence Qubit Design using Active Graph Search.* Nature Communications, 2024.

W2. CogSci '24 — Shikhar Tuli, Niraj K. Jha. *GiT: Can learning from good-old English grammar make Transformers more human-like?.* Cognitive Science, 2024.

W1. JAIR '24 — Shikhar Tuli, Niraj K. Jha. *BREATHE: Second-Order Gradients and Heteroscedastic Emulation based Design Space Exploration.* Journal or Artificial Intelligence Research, 2024 (under review; preprint link).

Patents

- *DynaMo: Why Predict Just One Token at a Time?.* Shikhar Tuli, Chi-Heng Lin, Yen-Chang Hsu, Yilin Shen, Hongxia Jin. Filed at the US patent office. Sep 7, 2024.
- *Hardware-software co-design for efficient transformer training and inference.* Shikhar Tuli, Niraj K. Jha. Filed at the US patent office. July 24, 2023, App. no.: 63/528,445.
- *Graph Language Models: Distilling reliable knowledge graphs from high-quality text.* Shikhar Tuli, Margarita Belova, Suma Bhat, Niraj K. Jha. Filed at the US patent office. November 16, 2023. App. no.: 63/599,846.

- *Low cost air purification system.* Shikhar Tuli, Shreshth Tuli, Sujeet K. Sinha. Filed at the Indian patent office. August 2, 2017, App. No.: 201711027523.
- *Combination Lock with limited trial and resetting mechanism.* Shikhar Tuli, Shreshth Tuli, Harshit Abrol, Shivang Dwivedi, Saujanya Chaudhary, Kargil Singh, Sivanandam Aravindan. Filed at the Indian patent office. August 10, 2017, App. no.: 201711028520.

Reviewing

I have served as a reviewer for many journals and conferences. See my Publons profile at this link.

- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (6)
- Wiley: Software Practices and Experience (4)
- IEEE Transactions on Evolutionary Computation (1)
- IEEE Transactions on Emerging Topics in Computing (1)
- IEEE Transactions on Industrial Informatics (1)
- International Conference on Machine Learning (1)
- Annual Meeting of the Cognitive Science Society (1)
- Conference on Information Sciences and Systems (1)

Teaching Experience

Department of Electrical and Computer Engineering, Princeton University:

- Machine Learning for Predictive Data Analytics. **Head T.A**. *Sep 2021 - Dec 2021.*

Department of Electrical Engineering, Indian Institute of Technology Delhi:

- Introduction to Electrical Engineering. **T.A**. *Jul 2019 - Nov 2019.*

Courses

- **Electrical Engineering:**
  Computer Architecture, Digital Electronics, Machine Learning and Intelligence, Analog Electronics, Physical Electronics, Power Electronics, Communication Engineering, Control Engineering, Engineering Electromagnetics, Signals and Systems, Electromechanics, Circuit Theory, IC Technology*, MOS VLSI Design*, Neuromorphic Engineering*, Mixed-Signal Circuit Design*, Compact modeling of Semiconductor Devices*, CMOS RF IC Design*, Digital Signal Processing†, Embedded Computing†.
- **Computer Science, Mathematics, Physics, and Cognitive Science:** Data Structures and Algorithms, Probability and Stochastic Processes, Calculus, Linear Algebra, Principles of Semiconductors, Computer Vision†, Machine Learning and Pattern Recognition†, Natural Language Processing†, Reinforcement Learning†, Probabilistic Models of Cognition†, Dynamics in Cognition†.

*Graduate-level course at IIT Delhi, †Graduate-level course at Princeton University

Technical Skills

- **Programming Languages:** Python, MATLAB, Java, C/C++, Verilog, RTL, x86 and ARM assembly, Verilog-A, PEL, OpenCL, HTML, R.
- **Frameworks:** PyTorch, Tensorflow, Keras, OpenCV, CUDA, Git, Xilinx Vivado, AnSYS HFSS, Synopsys Design Compiler, Capo Floor-planner, CACTI/FinCACTI, NVMain, NVSim, Keysight EasyEXPERT, Keysight IC-CAP, Altium Designer, Eagle, PSIM, Origin Pro, Adobe Photoshop, Adobe Illustrator, Arduino, Solidworks, Cinema 4D.

Positions of Responsibility

- **Technical Executive** at Makerspace: Design and Innovation Centre at IIT Delhi.
- **Coordinator** at Sportech '17: Sports fest at IIT Delhi.

Other Interests

Endurance running (first 5k in 2021, 10k and half marathon in 2022), rock climbing (v3/v4 level), lawn tennis, street jazz and hip-hop dance, graphics designing, poster making, video editing, and poem writing.