# Arun M. George

Ph.D. student, Computer Science and Engineering — University of Notre Dame
ageorg23@nd.edu · nexus-aeon.github.io · Scholar · GitHub · LinkedIn

**Continual learning** / **Quantization** / **Efficient LLMs**

Developing memory-efficient continual-learning methods and compact LLM representations, with hardware constraints informing algorithm design, optimization, and evaluation.

## EDUCATION

**University of Notre Dame** 2024–present
Ph.D., Computer Science and Engineering
Advisor: Siddharth Joshi · Intelligent Microsystems Lab

**International Institute of Information Technology Bangalore** 2016–2018
M.Tech., Information Technology; specialization in Data Science

**Mahatma Gandhi University, Kerala** 2011–2015
B.Tech., Computer Science and Engineering

## RESEARCH EXPERIENCE

**University of Notre Dame** 2024–present
Research Assistant, Intelligent Microsystems Lab

**Continual learning under memory constraints** Ongoing research

Develop and evaluate learning methods for fixed-capacity models over changing task sequences. Focus on knowledge retention, stability–plasticity trade-offs, and bounded auxiliary memory so that models remain able to learn new tasks.

**QTEA: ternary LLM quantization** EMNLP 2026; accepted

Contributed to core-methodology ideation, implemented baselines, and conducted experimental evaluation for QTEA. The method combines ternary weights, structured sparse residual corrections, and column-wise optimization, reaching an effective 1.7 bits/weight on Qwen3-14B.

**Huff-LLM: lossless inference compression** Preprint, 2025

Built PyTorch reference implementations to validate bit-exact decoding in a hardware–software compression pipeline. The work keeps LLM weights compressed across the memory hierarchy to reduce storage and data movement.

**Mixed-signal LLM inference** DAC 2026

Developed PyTorch models of microscaled block floating-point computation and hardware non-idealities for end-to-end LLM evaluation. Connected quantization and numerical-format choices with the accuracy of a compute-near-memory architecture.

**Efficient and robust neural decoding** BioCAS 2025

Contributed to spiking neural decoders for closed-loop motor control. Our team won both tracks of the IEEE BioCAS 2025 Grand Challenge.

## EARLIER RESEARCH AND INDUSTRY EXPERIENCE

**TCS Research** 2018–2024
Researcher, neuromorphic computing and online learning

- Developed spiking reservoir models for online time-series forecasting and temporal representation learning, supporting learning from evolving data with compact recurrent architectures.
- Worked on quantized neural networks and neuromorphic implementation for acoustic gesture recognition, connecting model design with specialized execution hardware.
- Developed spiking architectures for event-camera and radar streams; contributed to sparse spike encoding and spiking approximations of Koopman operators. Co-inventor on five granted U.S. patents.

**MiQ** 2018; six months
Data Scientist Intern

Developed feature-selection services for sparse, imbalanced advertising data using PySpark.

**IIIT Bangalore** 2016–2017
Student Researcher

Proposed an IEEE 754-compatible floating-point addition architecture for neuromorphic systems.

## TECHNICAL EXPERIENCE

**Core frameworks** Python, PyTorch, Hugging Face Transformers; JAX (Flax/NNX); Avalanche.

**Continual learning** Gradient/subspace projection (GPM, SGP, CODE-CL); experience replay; importance-based regularization (EWC); distillation (LwF); model merging.

**CL evaluation** Retention and forgetting, forward/backward transfer, plasticity, and memory/computation budgets.

**Quantization and adaptation** PTQ/QAT; low-bit and microscaled formats; torchao; LoRA/PEFT.

**Inference and systems** vLLM; NVIDIA Nsight; PyTorch DDP/FSDP; Linux, Git.

**LLM training knowledge** Supervised fine-tuning (SFT); policy-gradient methods (PPO, GRPO variants).

**Additional knowledge** Triton programming model.

## RECOGNITION

**IEEE BioCAS 2025 Grand Challenge** 2025
Team winner of both tracks: efficient and robust closed-loop neural decoding.

## TEACHING AND TUTORIALS

**Teaching Assistant, University of Notre Dame** Fall 2024
Data Structures and Algorithms; six-month appointment

**A Tutorial on Bio-Inspired Neuromorphic Computational Paradigms** 2023
IEEE ICASSP. Co-presented with Manan Suri and Sounak Dey.

## PROFESSIONAL SERVICE

**Journal reviewing** Nature Communications; Neurocomputing; Neural Networks.

**Conference reviewing** IEEE CDC; IEEE ICASSP.

## PUBLICATIONS

### Journal and conference papers

01 **QTEA: Ternary LLMs with Sparse Residual Salient Weight and By-Column Optimization**
Yipin Guo, **Arun M George**, Jie Fu, Tareq Mahmoud, Sixue Xing, Siddharth Joshi.
*EMNLP 2026, Main Conference; Efficient Methods for NLP Track.* ***Accepted***.

02 **Foundry-SRAM-Compatible Compute-Near-Memory Macro for Microscaled Block-FP in 12-nm CMOS**
Samir Rahman, **Arun George**, Shehab Naga, Thomas Summe, Yipin Guo, Siddharth Joshi.
*Design Automation Conference (DAC), 2026*.

03 **GRU-RVCIM: Efficient Implementation of Gated Recurrent Unit in RISC-V - Compute In Memory co-designed for Edge Applications**
J Chithambara Moorthii, Deepak Verma, **Arun M George**, Dighanchal Banerjee, Sounak Dey, Arijit Mukherjee, Arpan Pal, Manan Suri.
*IEEE Embedded Systems Letters, 2025*.

04 **Efficient Closed-Loop Neural Decoding for Motor Control of non-Human Primates with Spiking Neural Networks**
Thomas Summe, **Arun George**, Siddharth Joshi.
*IEEE Biomedical Circuits and Systems Conference (BioCAS), 2025*.

05 **Approximation of Koopman Operator using Spiking Neural Networks**
**Arun M. George**, Dighanchal Banerjee, Titas Bera, Sounak Dey.
*IEEE Conference on Decision and Control (CDC), 2024*.

06 **StereoGest-SNN: Robust Gesture Detection with Stereo Acoustic Setup Using Spiking Neural Networks**
Andrew Gigie, **Arun M George**, A Anil Kumar, Sounak Dey, Arpan Pal.
*IEEE International Symposium on Circuits and Systems (ISCAS), 2023*.

07 **Online time-series forecasting using spiking reservoir**
**Arun M. George**, Sounak Dey, Dighanchal Banerjee, Arijit Mukherjee, Manan Suri.
*Neurocomputing 518, 82–94, 2023*.

08 **Efficient Time Series Classification using Spiking Reservoir**
Sounak Dey, Dighanchal Banerjee, **Arun M. George**, Arijit Mukherjee, Arpan Pal.
*International Joint Conference on Neural Networks (IJCNN), 2022*.

09 **EchoWrite-SNN: Acoustic Based Air-Written Shape Recognition Using Spiking Neural Networks**
**Arun M George**, Andrew Gigie, A Anil Kumar, Sounak Dey, Arpan Pal, K Aditi.
*IEEE International Symposium on Circuits and Systems (ISCAS), 2022*.

10 **Efficient Optimized Spike Encoding of Multivariate Time-series**
Dighanchal Banerjee, Sounak Dey, **Arun M. George**, Arijit Mukherjee.
*Neuro-Inspired Computational Elements (NICE), 2022*.

## PUBLICATIONS / CONTINUED

11 **A Reservoir-based Convolutional Spiking Neural Network for Gesture Recognition from DVS Input**
**Arun M. George**, Dighanchal Banerjee, Sounak Dey, Arijit Mukherjee, P. Balamurali.
*International Joint Conference on Neural Networks (IJCNN), 2020*.

12 **Application of Spiking Neural Networks for Action Recognition from Radar Data**
Dighanchal Banerjee, Smriti Rani, **Arun M. George**, Arijit Chowdhury, Sounak Dey, Arijit Mukherjee, Tapas Chakravarty, Arpan Pal.
*International Joint Conference on Neural Networks (IJCNN), 2020*.

13 **IEEE 754 floating-point addition for neuromorphic architecture**
**Arun M. George**, Rahul Sharma, Shrisha Rao.
*Neurocomputing, 2019*.

### Workshop papers

W1 **Surrogate Gradient Descent Based Spiking Intrusion Detection System for Edge Devices: A Performance Study**
Dighanchal Banerjee, **Arun M. George**, Urmi Jana, Sounak Dey.
*ECML PKDD 2025 Workshops · Proceedings, 2026*.

W2 **Efficient Optimized Spike Encoding for Spiking Neural Networks**
Dighanchal Banerjee, Sounak Dey, **Arun M. George**, Arijit Mukherjee.
*OPT 2020 · NeurIPS workshop, 2020*.

### Preprints

A1 **Huff-LLM: End-to-End Lossless Compression for Efficient LLM Inference**
Patrick Yubeaton, Tareq Mahmoud, Shehab Naga, Pooria Taheri, Tianhua Xia, **Arun George**, Yasmein Khalil, Sai Qian Zhang, Siddharth Joshi, Chinmay Hegde, Siddharth Garg.
*arXiv (preprint), 2025*.

## GRANTED U.S. PATENTS

Co-inventor on five grants; assignee: Tata Consultancy Services.

G1 System and method for real-time radar-based action recognition using spiking neural network (SNN). US12347159B2; granted Jul 1, 2025.

G2 System and method for acoustic based gesture tracking and recognition using spiking neural network. US12326979B2; granted Jun 10, 2025.

G3 Method and system for optimized spike encoding for spiking neural networks. US12307360B2; granted May 20, 2025.

G4 Acoustic system and method based gesture detection using spiking neural networks. US11960654B2; granted Apr 16, 2024.

G5 System and method of gesture recognition using a reservoir based convolutional spiking neural network. US11256954B2; granted Feb 22, 2022.