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 / Learning dynamics / Efficient learning

Research on knowledge retention and plasticity in continual learning, combining gradient/subspace methods, controlled empirical evaluation, and memory-efficient implementation. Additional work spans LLM quantization and hardware-aware inference.

EDUCATION

University of Notre Dame 2024–present
Ph.D., Computer Science and Engineering (third year) · Advisor: Siddharth Joshi

IIIT Bangalore M.Tech., Information Technology (Data Science) 2016–2018

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

RESEARCH EXPERIENCE

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

Continual learning: retention, plasticity, and memory Ongoing research

- Develop and evaluate gradient/subspace methods for fixed-capacity models learning over long task sequences, balancing preservation of prior knowledge with the capacity to learn new tasks.
- Combine mathematical analysis and proofs with empirical comparison of projection, replay, regularization, and distillation approaches; assess forgetting, transfer, plasticity, and resource costs.

Online adaptation and temporal learning Neurocomputing, 2023

First-author work on spiking reservoirs with feedback-driven online updates for forecasting non-stationary time series, connecting temporal representation learning with memory-constrained computation.

QTEA: ternary LLM quantization EMNLP 2026; accepted

Contributed to core-methodology ideation, implemented baselines, and conducted experimental evaluation for sub-2-bit post-training quantization. QTEA reaches an effective **1.7 bits/weight on Qwen3-14B**.

Huff-LLM: lossless LLM inference compression Preprint, 2025

Built PyTorch reference implementations to validate bit-exact decoding in a hardware–software pipeline that keeps LLM weights compressed across the memory hierarchy.

Hardware-aware LLM inference DAC 2026

Developed PyTorch models of microscaled block floating-point computation and hardware non-idealities to evaluate quantization and end-to-end LLM accuracy on a compute-near-memory architecture.

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

- Developed spiking reservoir models for online time-series forecasting and temporal learning; worked on quantized neural networks for acoustic gesture recognition and neuromorphic implementation.
- Developed spiking architectures for event-camera/radar streams; contributed to sparse encoding and Koopman-operator approximation. Co-inventor on **five granted U.S. patents**.

SELECTED RELATED PUBLICATIONS

Full publication list · Granted patents · My name is in bold.

1 **Online time-series forecasting using spiking reservoir**
A. M. George, S. Dey, D. Banerjee, A. Mukherjee, M. Suri.
Neurocomputing 518, 82–94, 2023.

2 **Approximation of Koopman Operator using Spiking Neural Networks**
A. M. George, D. Banerjee, T. Bera, S. Dey.
IEEE Conference on Decision and Control (CDC), 2024.

3 **QTEA: Ternary LLMs with Sparse Residual Salient Weight and By-Column Optimization**
Y. Guo, **A. M. George**, J. Fu, T. Mahmoud, S. Xing, S. Joshi.
EMNLP 2026, Main Conference; Efficient Methods for NLP Track. ***Accepted***.

4 **Foundry-SRAM-Compatible Compute-Near-Memory Macro for Microscaled Block-FP in 12-nm CMOS**
S. Rahman, **A. George**, S. Naga, T. Summe, Y. Guo, S. Joshi.
Design Automation Conference (DAC), 2026.

5 **Huff-LLM: End-to-End Lossless Compression for Efficient LLM Inference**
P. Yubeaton, T. Mahmoud, S. Naga, P. Taheri, T. Xia, **A. George**, Y. Khalil, S. Q. Zhang, S. Joshi, C. Hegde, S. Garg.
arXiv, 2025 (preprint).

TECHNICAL EXPERIENCE

Foundations Linear algebra; machine learning.

Research implementation Python, PyTorch, Avalanche; JAX (Flax/NNX); Hugging Face Transformers.

Continual learning Gradient/subspace methods (GPM, TRGP, SGP, CODE-CL); experience replay; importance regularization (EWC, SI, MAS); distillation (LwF); model merging.

Experimental analysis Retention/forgetting, forward/backward transfer, plasticity, baseline comparison, 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 SFT; policy-gradient methods (PPO, GRPO variants).

Additional knowledge Triton programming model.

RECOGNITION AND SERVICE

IEEE BioCAS 2025 Grand Challenge — team winner, both tracks. Contributed to efficient, robust spiking neural decoders for closed-loop motor control.

Reviewing Nature Communications, Neurocomputing, Neural Networks; IEEE CDC, ICASSP.

ADDITIONAL EXPERIENCE

MiQ Data Scientist Intern 2018; six months
Developed feature-selection services for sparse, imbalanced advertising data using PySpark.

IIIT Bangalore Student Researcher 2016–2017
Proposed an IEEE 754-compatible floating-point addition architecture for neuromorphic systems.

University of Notre Dame Teaching Assistant, Data Structures and Algorithms Fall 2024
Six-month appointment.

Tutorial Co-presenter, Bio-Inspired Neuromorphic Computational Paradigms, IEEE ICASSP 2023