Kasun Dewage

Ph.D. Candidate in Mathematics | Artificial Intelligence Research | Tensor Methods
KasunTharuka.Dewage@ucf.edu | github.com/Kasun-Dewage

Academic Profile

Ph.D. candidate in Mathematics at the University of Central Florida conducting artificial intelligence research at the intersection of tensor methods, numerical linear algebra, efficient foundation models, and quantitative modeling. Research develops mathematically structured methods for adapting and analyzing transformer models, with applications to parameter-efficient fine-tuning, compression, quantization, interpretability, and financial time series. **First author of six accepted 2026 research papers**, including **CRAFT**, an extremely parameter-efficient PEFT method published as a main-conference paper at COLM 2026; on LLaMA3-8B, CRAFT surpassed the reported LoRA average using about 840× fewer adaptation parameters.

Research Interests

- Mathematical foundations of AI: tensor methods, low-rank approximation, numerical linear algebra, and structured optimization.
- Efficient AI: parameter-efficient fine-tuning, transformer adaptation, model compression, quantization, and attention analysis.
- Foundation models, generative artificial intelligence, and mathematically grounded AI methods.
- Artificial intelligence for financial time series, foundation-model forecasting, and quantitative modeling.

Selected Research Contributions

- **CRAFT / parameter-efficient fine-tuning.** Developed a cross-layer Tucker-3/HOSVD framework that decomposes pre-trained attention weights, freezes the Tucker factors, and trains only small square adaptation matrices. On LLaMA3-8B, CRAFT with 0.068M adaptation parameters achieved 82.9 average commonsense accuracy versus 80.8 for LoRA with 57M parameters—approximately 840× fewer trainable adaptation parameters. At fixed Tucker ranks, the CRAFT adaptation parameter count is independent of model width and depth.
- **Frozen foundation models for financial time series.** Developed hybrid neural–classical correction for a frozen 200M-parameter TimesFM model on 2,011,399 one-minute observations across 10 technology stocks. The best GatedLinear+RF model improved mean per-day correlation from 0.059 to 0.373 (6.4×), reached 0.597 pooled correlation, and used 9× fewer neural parameters than the attention-based corrector.
- **Transformer structure and efficiency.** Led research on attention-weight compressibility, spectral structure, structured head pruning, and component-level quantization sensitivity, resulting in accepted papers at IEEE ICNLP and IEEE ICMLA 2026.

Education

- **Ph.D. in Mathematics**, University of Central Florida — *Expected Fall 2026*
  Advisor: Prof. Marianna Pensky — GPA: 3.7
- **M.S. in Computer Vision**, University of Central Florida — *2024*
  GPA: 3.8
- **M.S. in Pure Mathematics**, Sam Houston State University — *2017*
- **M.S. in Financial Mathematics**, University of Colombo — *2014*
- **B.Sc. in Physical Science**, University of Colombo — *2012*

Sponsored Research

Significant Contributor, NSF Award No. 2310881, *Multiplex Generalized Dot Product Graph Networks: Theory and Applications*, under the direction of Principal Investigator Prof. Marianna Pensky.

Publications, Preprints, and Manuscripts

First-Authored Papers: Accepted and Forthcoming

1. **Kasun Dewage, Marianna Pensky, Shankhadeep Mondal, Suranadi De Silva.**
   LORA-CRAFT: Cross-Layer Rank Adaptation via Frozen Tucker Decomposition of Pre-Trained Attention Weights.
   Main-conference paper, Conference on Language Modeling (COLM 2026).
2. **Kasun Dewage, Suranadi De Silva, Shankhadeep Mondal.**
   Hybrid Neural-Classical Correction for Frozen Time Series Foundation Models: A Comprehensive Ablation Study on High-Frequency Stock Prediction.
   Accepted, IEEE International Joint Conference on Neural Networks / World Congress on Computational Intelligence (IJCNN/WCCI 2026). Code.

3. **Kasun Dewage, Suranadi De Silva, Shankhadeep Mondal.**
On the Compressibility of Fine-Tuned Attention: A Tensor Decomposition Study of PEFT Weight Structure.
Accepted, IEEE ICNLP 2026.
4. **Kasun Dewage, Marianna Pensky, Suranadi De Silva.**
Magnitude Profile Pruning: Calibration-Free Structured Attention Head Removal for Transformer Compression.
Accepted, IEEE ICMLA 2026.
5. **Kasun Dewage, Marianna Pensky, Suranadi De Silva.**
Component Type, Not Reconstruction Error, Predicts Attention Quantization Sensitivity.
Accepted, IEEE ICMLA 2026.
6. **Kasun Dewage, Marianna Pensky, Suranadi De Silva.**
Spectral Outliers Reveal Dominant Learned Structure in Transformer Attention.
Accepted, IEEE ICMLA 2026.

First-Authored Manuscripts Under Review

1. *Causal Head Selectors Agree with Each Other but Not with Attention Heuristics.*
2. *A Behavioral Taxonomy of Attention Heads That Transfers Across Transformer Language Models.*

Collaborative Paper

1. Shankadeep Mondal, Ram Narayan Mohapatra, **Kasun Dewage**. *Refined Upper Bounds for the Numerical Radius via Weighted Operator Means.* Under review at *Acta Scientiarum Mathematicarum.*

Academic Appointments and Teaching Experience

- **Teaching Associate / Instructor of Record**, University of Central Florida *2020–Present*
  - Instructor of record for Calculus II, Calculus III, and Linear Algebra; taught up to 8 credit hours per semester.
  - Reported student-evaluation averages consistently between 4.6 and 5.0 out of 5.0.
  - Designed lectures, assessments, review materials, and structured problem-solving activities.
- **Full-Time Lecturer / Pool Faculty**, Sam Houston State University *2017–2020*
  - Taught up to seven mathematics courses per semester, including College Algebra and undergraduate service mathematics courses.
  - Contributed to curriculum delivery, student support, and departmental service.
- **Graduate Teaching Assistant**, Sam Houston State University *2015–2017*
  - Supported undergraduate mathematics instruction through teaching, grading, tutoring, and review sessions.

Scholarly Presentations

- **Hybrid Neural-Classical Correction for Frozen Time Series Foundation Models**. IEEE WCCI/IJCNN 2026.
- **On the Compressibility of Fine-Tuned Attention: A Tensor Decomposition Study of PEFT Weight Structure**. ICNLP 2026.

Professional Service

- **Session Chair**, Time Series and Temporal Modeling, IEEE WCCI/IJCNN 2026.
- **Reviewer**, IEEE International Joint Conference on Neural Networks (IJCNN 2026).
- **Reviewer**, IEEE International Conference on Machine Learning and Applications (ICMLA).

Applied and Quantitative Experience

- **Independent Quantitative Finance Researcher / Systematic Trading** *2018–Present*
  - Designed and evaluated systematic trading and risk-management strategies using mathematical modeling, statistical analysis, and computational methods.
  - Built quantitative models for market-regime analysis, forecasting, and benchmark-based evaluation.
- **Banking Adviser**, Standard Chartered Bank *2014–2015*
  - Provided financial advisory services to clients.

Technical Skills

- **Programming**: Python, PyTorch, MATLAB.
- **Mathematical Methods**: Tensor decomposition, low-rank modeling, numerical analysis, statistical computation, linear algebra.
- **Artificial Intelligence**: Transformers, foundation models, generative AI, PEFT/LoRA, compression, pruning, quantization, time-series forecasting.
- **Teaching Areas**: Calculus II–III, Linear Algebra, Differential Equations, Undergraduate Statistics.

Awards and Professional Memberships

- Outstanding Faculty Award Nominee, Sam Houston State University, 2019.
- Member: IEEE; International Neural Network Society; American Mathematical Society.