

Institute of Artificial Intelligence

UNIVERSITY OF CENTRAL FLORIDA

FINAL ORAL EXAMINATION

OF

Md Kowsher

M.Sc., Stevens Institute of Technology, NJ, USA, 2023

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

(COMPUTER SCIENCE)

16 October 2026, 11 A.M.

DISSERTATION COMMITTEE

Professor Chen Chen, *chen.chen@crcv.ucf.edu*
Professor Ozlem Garibay, *ozlem@ucf.edu*
Professor Ali Siahkoohi, *alisk@ucf.edu*
Professor Rui Xie, *Rui.Xie@ucf.edu*

DISSERTATION RESEARCH IMPACT

Adapting large pretrained models has become central to machine learning in language, vision, and multimodal settings. A single model is trained once and then adjusted for many tasks, such as text understanding, question answering, image classification, and video understanding, which support applications like search, recommendation, and digital assistants. Progress in this area has been driven by these pretrained models and by methods that fine-tune only a small part of them. However, these methods are mostly found by trial and error, and the field still lacks a clear reason for why such small changes are enough, where in the model the useful signal sits, or how much of a model a task really needs to change. A further problem is cost: even efficient methods can use a large amount of memory and time.

This dissertation shows that a pretrained model already holds far more useful information than any adaptation uses, and that this information sits in certain parts of the model rather than spread evenly. We first introduce a way to measure how much task-useful information each layer carries, and show that more of it means better performance. We then show that a model can be adapted by training only a small slice of its existing weights, with no new parameters. We extend this to models with many experts, replacing heavy per-expert parts and routers with lightweight pieces. Finally, we show that the training signal sits in only a few token positions, so gradients can pass through just those to save memory while keeping accuracy.

SELECTED PUBLICATIONS (h-index: 19, total citation: 1516)

1. **Where You Backpropagate Matters: Token Hypothesis for Memory-Efficient Fine-Tuning,** M Kowsher, and C Chen, in Conference on Neural Information Processing Systems (NeurIPS), 2026.
2. **Monkey Jump: MoE-Style PEFT for Efficient Multi-Task Learning,** NJ Prottasha, M Kowsher, CN Yu, C Chen, and O Garibay, Conference on Empirical Methods in Natural Language Processing (EMNLP), 2026 (Main).
3. **LiME: Lightweight Mixture of Experts for Efficient Multimodal Multi-task Learning,** M Kowsher, H Mansoor, NJ Prottasha, O Garibay, V Zhu, Z Ji, and C Chen Forty-third International Conference on Machine Learning (ICML), 2026 (Spotlight).
4. **SliceFine: The Universal Winning-Slice Hypothesis for Pretrained Networks,** M Kowsher, AO Polat, EM Ardehaly, M Salehi, Z Ghiasi, P Murali, and C Chen Forty-third International Conference on Machine Learning (ICML), 2026.
5. **FlowNIB: An Information Bottleneck Analysis of Bidirectional vs. Unidirectional Language Models,** M Kowsher, NJ Prottasha, S Xu, S Mohanto, O Garibay, N Yousefi, and C Chen The Fourteenth International Conference on Learning Representations (ICLR), 2026.
6. **RoCoFT: Efficient Finetuning of Large Language Models with Row-Column Updates,** M Kowsher, T Esmaeilbeig, CN Yu, C Chen, M Soltanalian, and N Yousefi The 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025 (Main).
7. **Predicting Through Generation: Why Generation Is Better for Prediction,** M Kowsher, NJ Prottasha, P Bhat, CN Yu, M Soltanalian, I Garibay, O Garibay, C Chen, and N Yousefi The 63rd Annual Meeting of the Association for Computational Linguistics (ACL), 2025 (Main).

DISSERTATION

FROM INFORMATION STRUCTURE TO EFFICIENT ADAPTATION: PRINCIPLED FOUNDATIONS FOR PRETRAINED MODELS

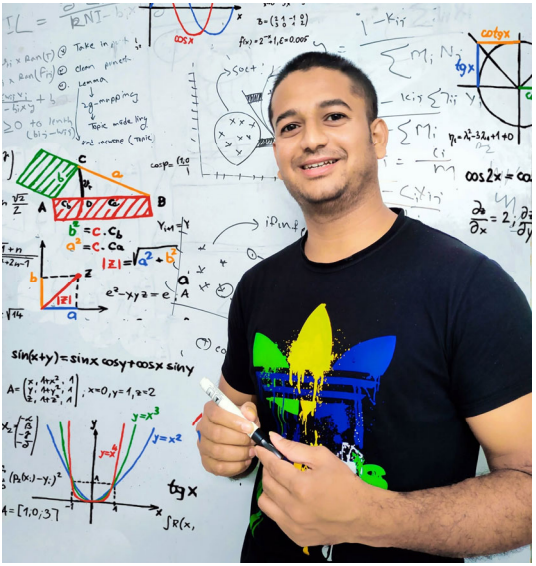
Large pretrained models are now the starting point for most work in language, vision, and multimodal learning, and the main cost has shifted from training a model to adapting it to many tasks. Methods that fine-tune only a small part of a model make this cheaper and work well, but the field still lacks a clear reason for why such small changes are enough, where in the model the useful signal sits, or how much a task really needs to change. This dissertation argues that a pretrained model already holds far more task-useful information than any adaptation uses, and that this information is concentrated in certain parts of the model rather than spread evenly. Building on this idea, we present methods that find where the useful signal sits and change only that, across four parts of a model: its layers, its weights, its experts, and its token positions.

Our first contribution is FlowNIB, a method that measures how much task-useful information each layer of a model carries about both the input and the task. Unlike earlier approaches, which measure these quantities separately and on scales that cannot be compared, FlowNIB places them on a single, comparable path, and shows that layers carrying more task-useful information lead to better performance. It also explains a long-standing observation: why models that read text in both directions understand context better than those that read in one. FlowNIB gives the lens that the rest of the dissertation uses.

While FlowNIB shows that the useful signal is already present and concentrated, it does not by itself change how we adapt a model. Next, we introduce SliceFine, which adapts a model by training only a small slice of its existing weights while freezing the rest, adding no new parameters. We show that almost any slice works, because the pretrained weights are balanced and the backbone already carries the task-useful signal, so a small slice is enough to match much heavier methods while using less memory and time.

SliceFine adapts along the weights, but it applies the same change to every input, which is wasteful when the data is diverse, as in multimodal multi-task settings. Next, we propose LiME, which gives different inputs different treatment without the usual cost. Instead of copying a full adapter for each expert and adding a learned router, LiME shares one small module, adjusts it with lightweight per-expert vectors, and routes inputs using information the model has already computed. This matches or beats heavier mixture-of-experts methods while training several times fewer parameters.

The methods above reduce how many parameters we train, but they still pass gradients through every token, so memory grows with the length of the input. Finally, we present the Token Hypothesis, which shows that the training signal is carried by only a few token positions. Running a full forward pass but passing gradients through just those positions matches full training across many tasks and models while cutting memory, and it works on top of parameter-efficient methods. Together, these contributions give one simple account of efficient adaptation: find where the useful information sits, and change only that.



Md Kowsher

1995	Born in Chattogram, Bangladesh
2015-2019	B.Sc, Noakhali Sc. & Tech. University, Bangladesh
2019-2020	AI Engineer Intern, NKSoft, Dhaka, Bangladesh
2020-2021	AI Engineer, Hishab Ltd, Dhaka, Bangladesh
2022-2022	Applied Scientist Intern, Amazon, Seattle, WA
2022-2023	Alexa Prize Researcher, Amazon, USA
2021-2023	M.Sc., Stevens Institute of Technology, NJ
2024-2024	Math & Algo. Intern (LLM), Nokia Bell Labs, NJ
2025-2025	ML Intern, Meta, Menlo Park, CA
2026-2026	AI/ ML Intern, Apple Inc., Cupertino, CA
2024	Ph.D., University of Central Florida, FL
2026	Research Scientist, Meta, Menlo Park, CA