

# VEDANT PURI

vedantpuri@cmu.edu | <https://vpuri3.github.io/>

## SUMMARY

Designs transformer architectures with explicit attention to scaling, memory, and communication structure. Developed FLARE, enabling million-token regimes on a single GPU. Implements new architectures in PyTorch and Triton. Background spans high-performance computing, numerical analysis, and computational fluid dynamics.

**Research Interests:** Efficient attention, language modeling, scientific machine learning.

## EDUCATION

<b>Carnegie Mellon University</b> <i>Ph.D Mechanical Engineering.</i> Advisors: Levent Burak Kara, Yongjie Jessica Zhang	Jan 2022–Present
<b>University of Illinois Urbana-Champaign</b> <i>B.S. Engineering Mechanics, B.S. Mathematics.</i>	2015–2019

## SELECTED RESEARCH CONTRIBUTIONS

<b>FLARE: Fast Low-Rank Attention Routing Engine</b>   <i>Efficient attention architectures</i>	2025
• Derived a flexible low-rank reformulation of self-attention via latent routing.	
• Reduced quadratic complexity of global communication in self-attention to linear complexity.	
• Demonstrated scaling to 1M tokens on a single H100 GPU, attaining over 200× speedup over vanilla self-attention.	
<b>FLARE for Language Modeling (Ongoing dissertation work)</b>   <i>Decoder-only architectures</i>	2025–Present
• Extending FLARE to decoder-only next-token prediction with causal attention.	
• Enabling adaptive attention state size to control memory and compute during training and inference.	
• Implementing fused Triton kernels for causal training, prefill, and decode.	
<b>Equation-based PDE modeling with neural fields</b>   <i>Hybrid data + physics methods</i>	2025
• Introduced smooth neural fields as nonlinear spatial ansatz functions in equation-based reduced-order modeling.	
• Retained physics-based Galerkin time evolution while learning expressive low-dimensional representations.	
• Attained 200× speedup over full-order simulations in transport-dominated regimes.	

## EXPERIENCE

<b>Carnegie Mellon University</b>   <i>Research Assistant</i>	Jan 2022–Present
• Phase field simulations of lithium dendritic growth in solid-state batteries.	
• Turbulence closure modeling with differentiable physics.	
<b>Julia Computing</b>   <i>Intern Engineer</i>	Apr 2021–Nov 2021
• Built numerical solvers for scientific machine learning ecosystem in Julia.	
<b>Carnegie Mellon University</b>   <i>Research Assistant</i>	Sep 2020–Jan 2021
• Developed differentiable geometry representations and meshing algorithms.	
<b>Argonne National Laboratory</b>   <i>Research Assistant</i>	Mar 2020–Sep 2020
• Executed large-scale simulations of turbulent airflow over urban landscapes on supercomputers.	
<b>Argonne National Laboratory</b>   <i>Research Assistant</i>	May 2018–Jul 2018
• Executed high-fidelity fluid dynamics simulations and analyzed turbulent statistics for closure modeling.	
<b>National Center for Supercomputing Applications</b>   <i>Intern</i>	Sep 2017–May 2018
• Numerical simulation of spacetime metric for gravitational wave simulations.	

## PUBLICATIONS

- Puri, V. et al., *FLARE: Fast Low-Rank Attention Routing Engine*. arXiv 2025. (Under review)  
Puri, V. et al., *SNF-ROM: Projection-based nonlinear reduced order modeling with smooth neural fields*. JCP 2025.  
Shankar, V., Puri, V., et al., *Differentiable physics closure modeling for Burgers' turbulence*. MLST 2023.

## AWARDS

<b>World Conference on Computational Mechanics</b>   <i>Best poster in fluid dynamics</i>	2024
<b>University of Illinois</b>   <i>Theoretical and Applied Mechanics Merit Award</i>	2019

## TECHNICAL SKILLS

- Machine Learning Systems:** PyTorch, Triton, mixed-precision/distributed training, causal language modeling  
**Numerical Computing:** Numerical analysis, scientific computing, linear algebra, finite elements, spectral methods  
**Programming Languages:** Python, Julia, C, Fortran 77, MATLAB, UNIX, L<sup>A</sup>T<sub>E</sub>X  
**Modeling Domains:** Transformer architectures, neural operators, reduced-order modeling