EMNLP 2024

# Sprout: Green Generative AI with Carbon-Efficient LLM Inference

**Baolin Li**, Yankai Jiang,
Vijay Gadepally, Devesh Tiwari

LVX VERITAS VIRTVS
Northeastern University

MIT LINCOLN LABORATORY

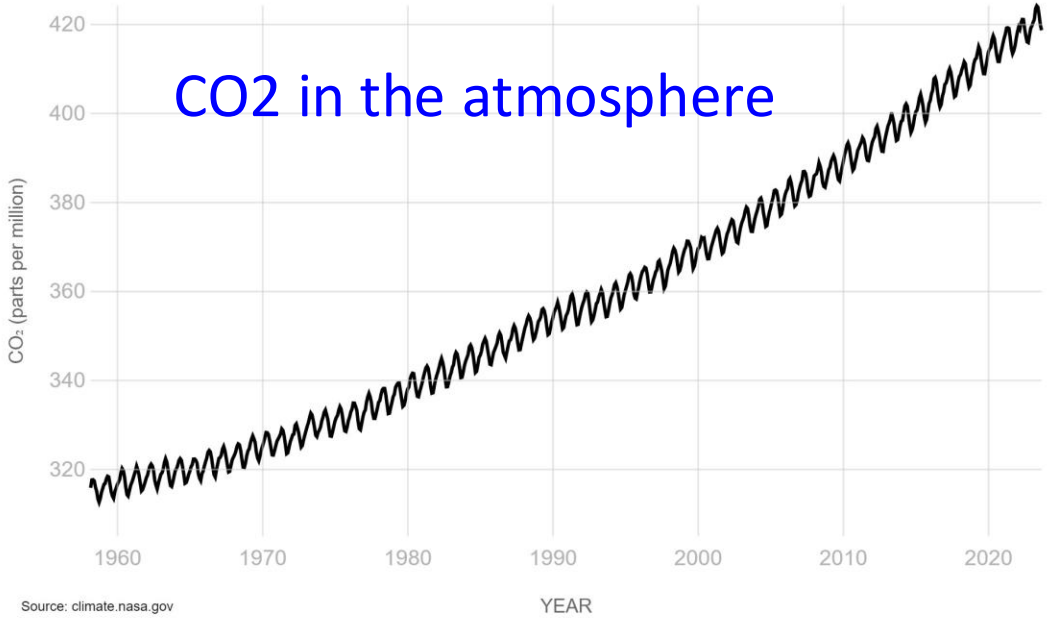
# Reducing Carbon Emission Is of Critical Importance

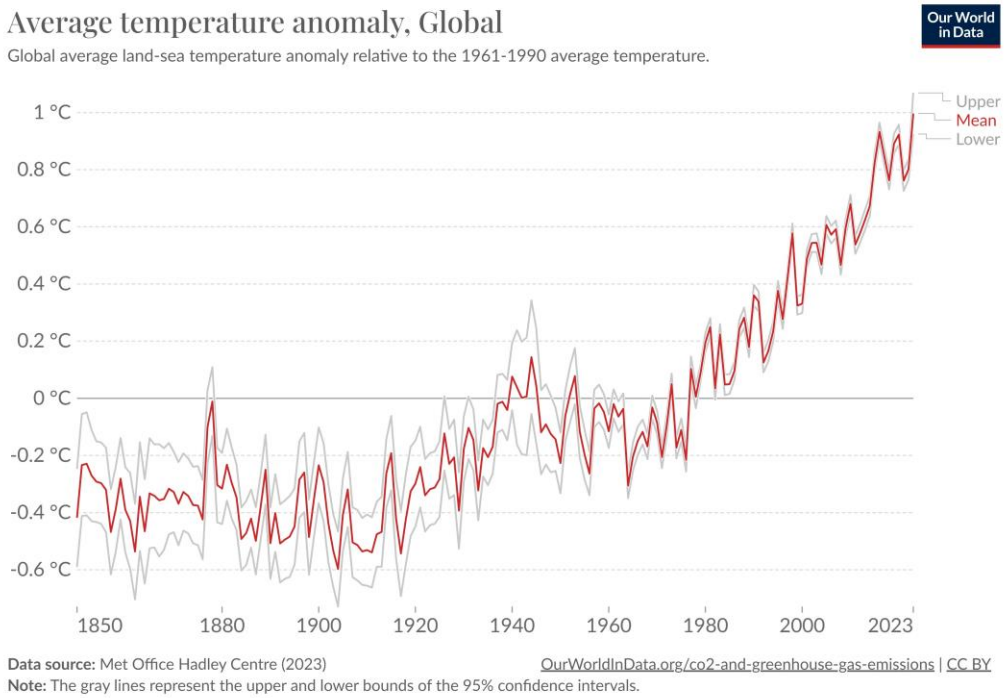
The Washington Post
*Democracy Dies in Darkness*

**CLIMATE** Environment Weather Climate Solutions Climate Lab Green Living Business of Climate

## World is on brink of catastrophic warming, U.N. climate change report says

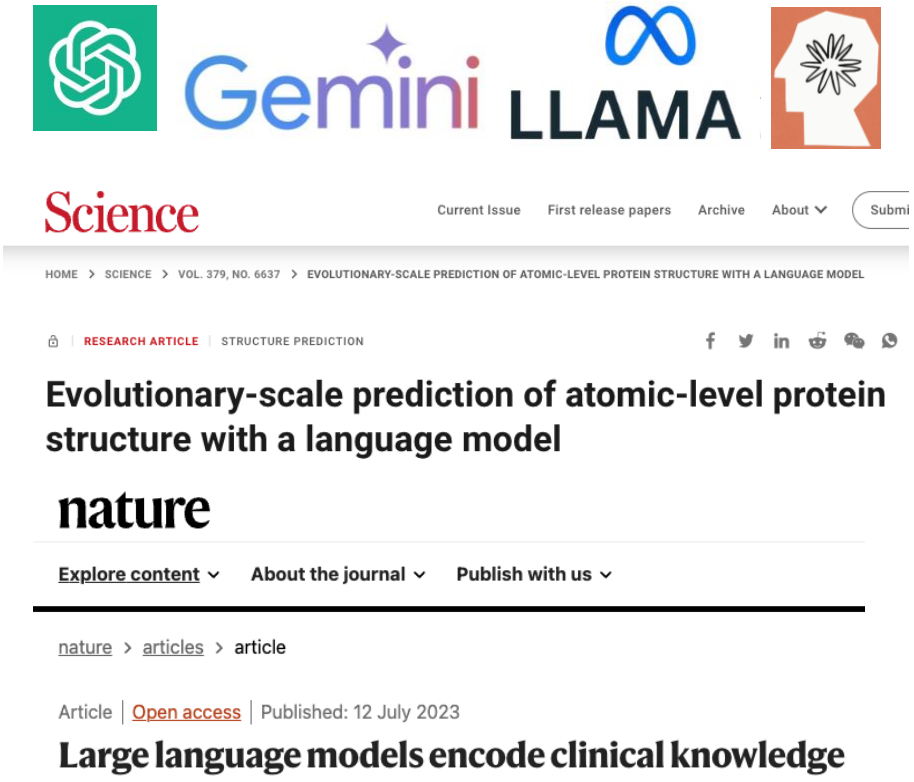
A dangerous climate threshold is near, but 'it does not mean we are doomed' if swift action is taken, scientists say



# Why Targeting Large Language Models

## The next frontier of Datacenter workloads




## Challenges to Environments

**Deep Model**

- Billions to trillions of parameters

**Heavy Computation**

- Attention calculation between token pairs
- autoregressive generation pattern

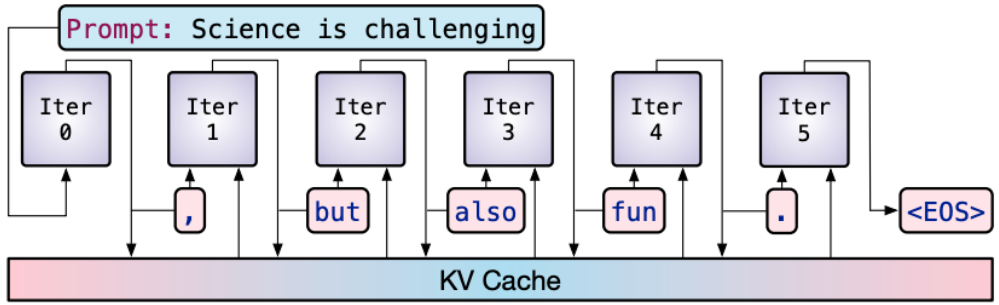
→ Significant Carbon Emission

## Carbon Footprint

$$C_{\text{req}} = CO_2^{\text{Intensity}} \cdot E_{\text{req}} + \frac{CO_2^{\text{Embed}}}{T_{\text{life}}} \cdot T_{\text{req}}$$

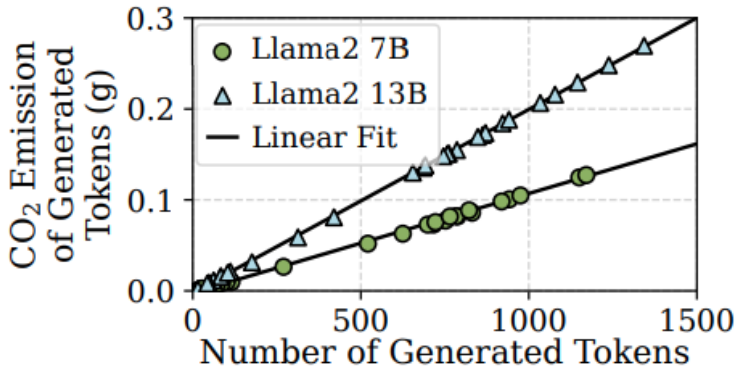
# Auto-Regressive Generation Dictates the LLM Inference Carbon Footprint

## Text tokens are generated iteratively




- The KV cache stores intermediate context tensors of previously generated tokens
- Allowing LLMs to efficiently generate significantly more tokens than input prompt

## Inference carbon is dictated by generated tokens




Instead of using smaller models with compromised context learning capabilities, can we let a larger model generate fewer tokens?

# Using Generation Directives to Guide LLM Token Generation

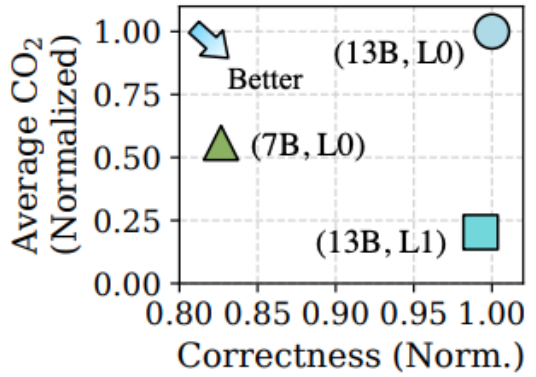
Applying generation directives to user prompts

*Definition 1:* A **generation directive** is an instruction/guidance associated with a prompt input that dictates the manner in which a generative language model produces tokens for the given input. Each **generation directive level** specifies a pre-defined text sequence that acts as this guiding instruction.

```
<prompt> How old is the Earth approximately?
(A) 50,000 years (B) 300 million years
(C) 4.5 billion years (D) no one knows

<generation directive L0 (default)> Based on a
variety of geological and astronomical
evidence, including …. While …, the scientific
consensus is (C): 4.5 billion years old.

<generation directive L1 (brief)> (C). The
Earth is approximately 4.5 billion years old.
```

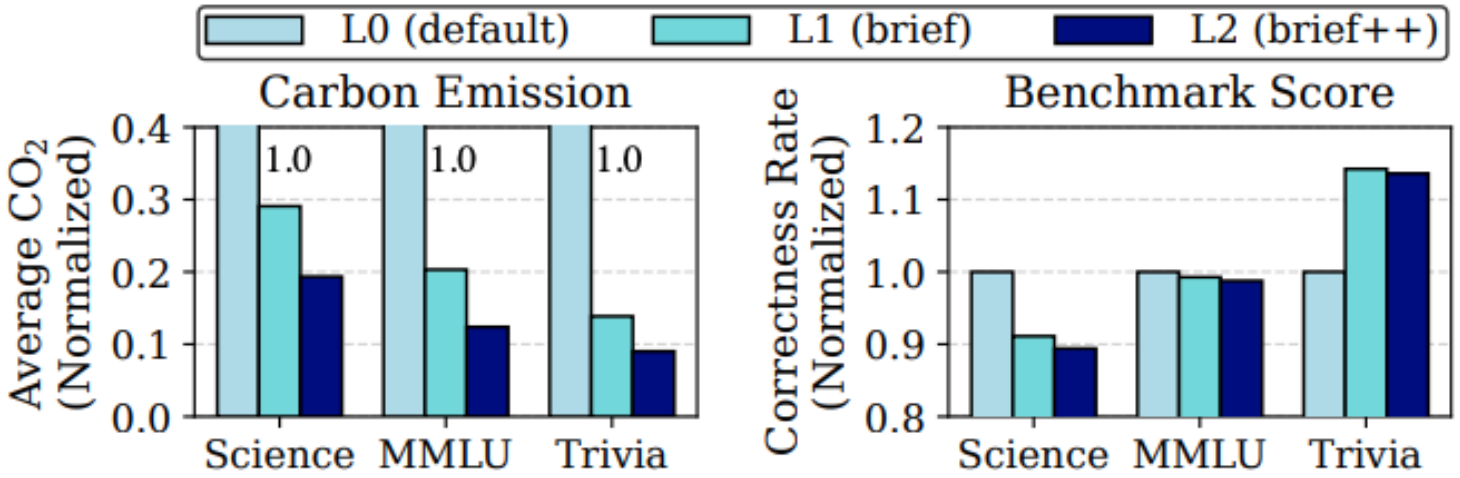



With less token generation, the LLM can still answer the user question correctly

Using generation directives, larger models can even save more carbon than smaller models while maintaining high generation quality

# Complex Interactions between Carbon and Content Quality

Carbon and quality sensitivity to generation directive varies across tasks




Three sets of tasks

- Science knowledge (biology/physics/chemistry)
- Massive multitask language understanding (MMLU)
- Trivia questions

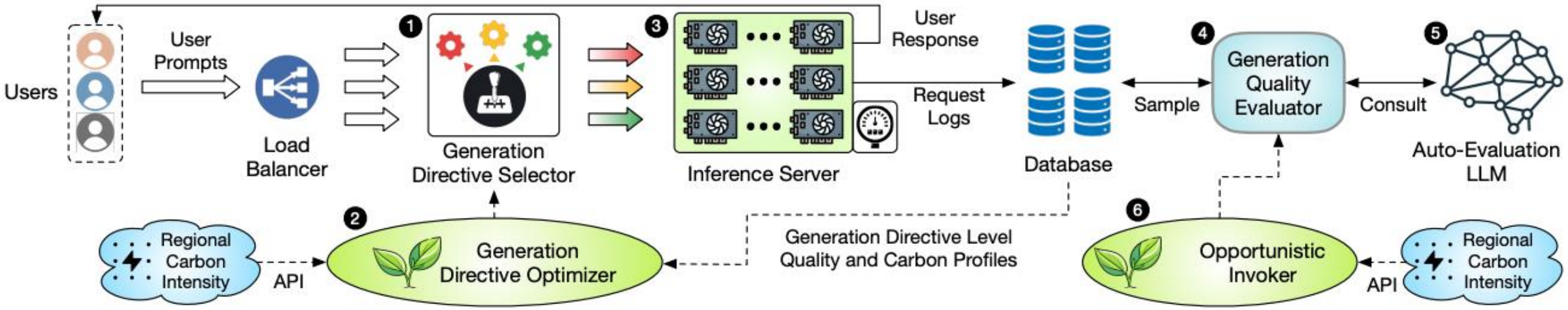
Directives may decrease accuracy in complex, multi-step reasoning tasks while enhancing accuracy when responses are directly inferable from the prompt or learned context.

**Important to collect generation quality feedback and adjust based on carbon & quality.**

# Sprout System Overview

**Configure the generation directive selection according to carbon intensity and generation quality feedback**

Minimize inference carbon footprint while guaranteeing generation quality




**Generation directive selection**

**Quality feedback collection**

# Sprout Design Details

## Generation directive selection

Carbon footprint of an inference request

$$C_{\text{req}} = CO_2^{\text{Intensity}} \cdot E_{\text{req}} + \frac{CO_2^{\text{Embed}}}{T_{\text{life}}} \cdot T_{\text{req}}$$

Objective: minimize expected carbon

$$f(\mathbf{x}) = k_0 \cdot \mathbf{e}^T\mathbf{x} + k_1 \cdot \mathbf{p}^T\mathbf{x}$$

Probability of selecting each directive level

Constraints: quality

$$\mathbf{q}^T\mathbf{x} \geq (1 - \frac{k_0 - k_0^{\min}}{k_0^{\max} - k_0^{\min}} \cdot \xi) \cdot q_0$$

Other constraints

$$\forall i,\ 0 \leq x_i \leq 1$$

$$\sum_{i=0}^{n-1} x_i = 1$$

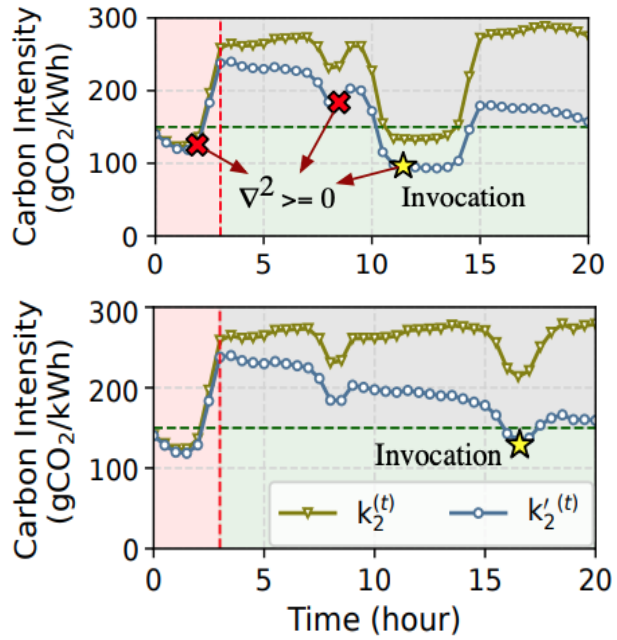
Linear Programming!

## Quality feedback collection

Urgency-adjusted carbon-aware offline quality evaluation

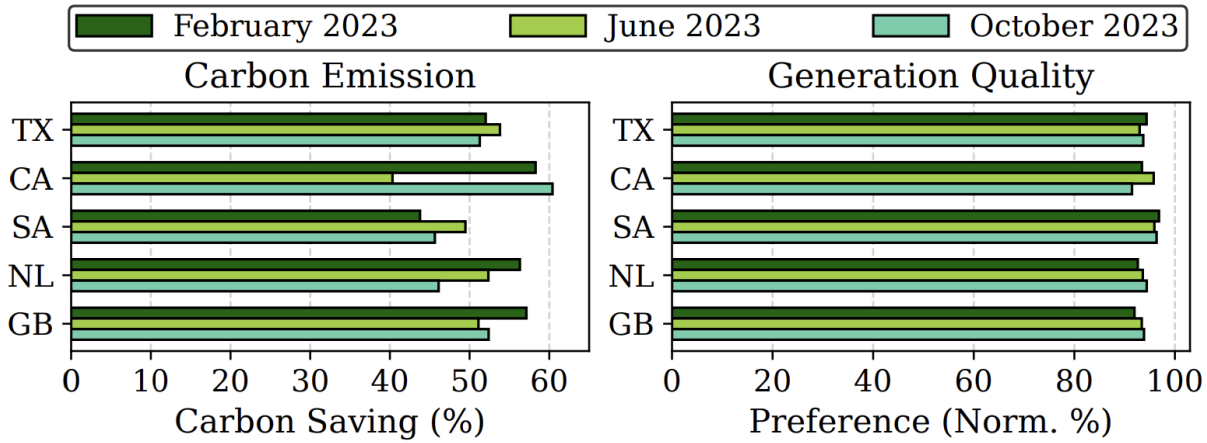
$$k_2'^{(t)} = e^{-\beta(t-t_0)} \cdot k_2^{(t)}$$

$CO_2^{Intensity}$ of auto-eval LLM

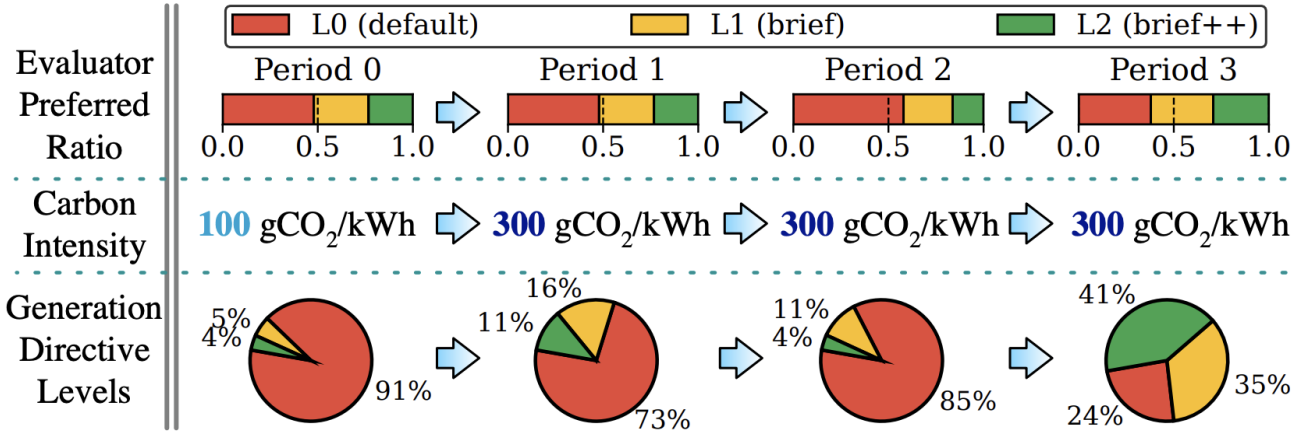


# Sprout significantly reduces carbon emission while guaranteeing generation quality

**Saves carbon emission by up to 60% with guarantee on auto-evaluator's generation preference**




The selection of generation directive levels responds to carbon intensity and user task shifts



# Sprout Summary of Key Contributions

Sprout is the first carbon-aware LLM inference system.

Sprout actively configures the generation directives to achieve a balance between carbon emission and generation quality under varying carbon intensity.

Sprout highlights ML efficiency from a carbon perspective, directly relating LLM operation to environmental impact.

**Contact**

Baolin Li

li.baol@northeastern.edu

This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001, and United States Air Force Research Laboratory Cooperative Agreement Number FA8750-19-2- 1000. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering, or the United States Air Force. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein