



# Spectral Regularization as a Safety-Critical Inductive Bias

Securing AI with **Fourier Gradient Regularization**  
for enhanced safety and robustness

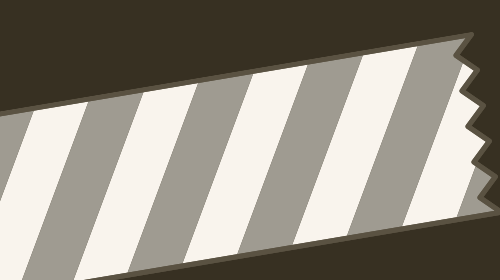
By --- Shivam Dubey  
IIT MADRAS



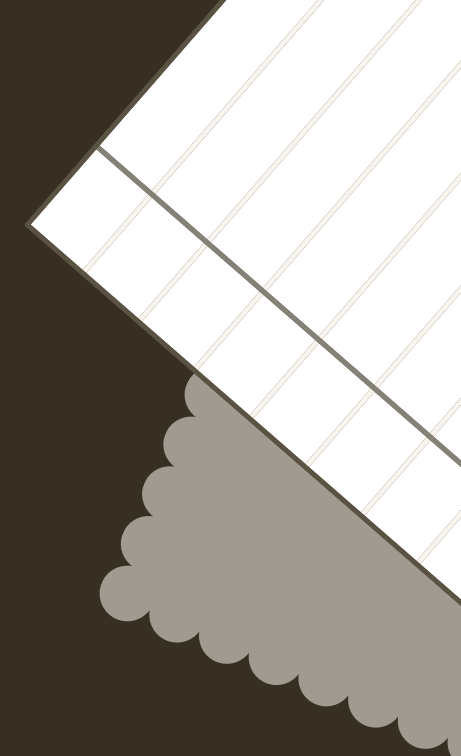
# Introduction

This work explores a novel bridge between a fundamental concept in physics spectral analysis and a critical problem in AI safety: adversarial vulnerability. Modern deep neural networks, despite their capabilities, are notoriously brittle. Their decisions can be completely reversed by tiny, human-imperceptible perturbations to their inputs. We posit that this fragility is not an arbitrary flaw but a direct consequence of the models' inherent "spectral bias." This bias, also known as the Frequency Principle, describes a model's preference for learning simple, low-frequency patterns in data while remaining poorly generalized to high-frequency variations. Adversarial attacks are effective precisely because they operate in this high-frequency "blind spot."

Our work introduces Fourier Gradient Regularization (FGR), a method that reframes the problem from reactive defense to proactive, principled training. Instead of training a model to recognize specific attacks, we alter the training objective itself to instill a safety-critical inductive bias. FGR directly penalizes the model's sensitivity to high-frequency input changes, forcing it to learn smoother, more robust functions that rely on coarse-grained features. This submission outlines the FGR method, presents strong empirical evidence of its success, and discusses its broader implications for AI safety.

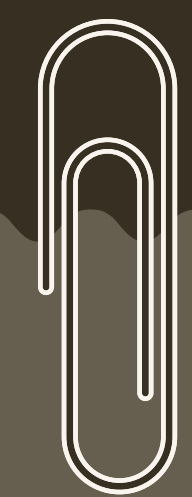


# Understanding AI Brittleness: Key Concepts Explained



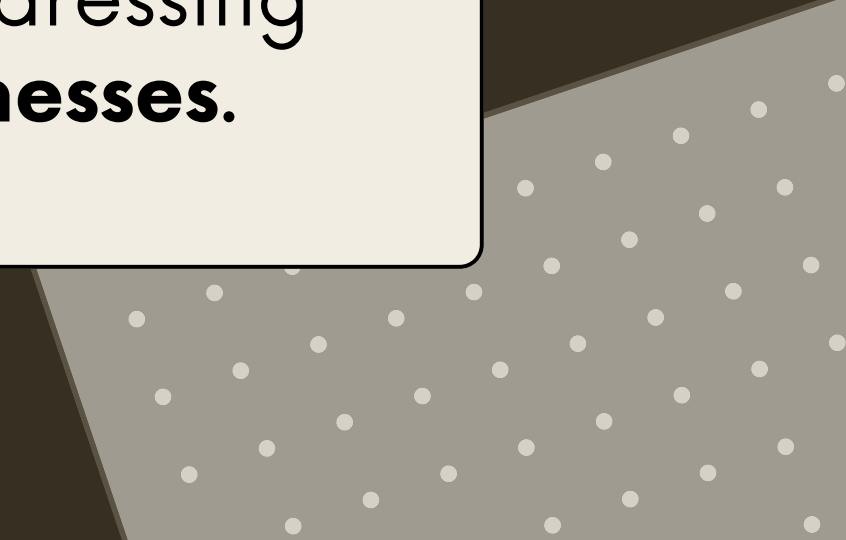
AI systems can be easily misled by **adversarial inputs**, leading to incorrect outputs.

The vulnerability of AI is illustrated through the **cat versus adversarial ostrich** scenario.

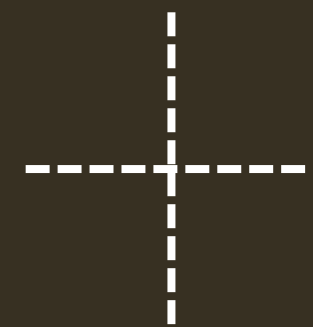


Adversarial examples expose the limitations and **brittleness** of current AI technologies.

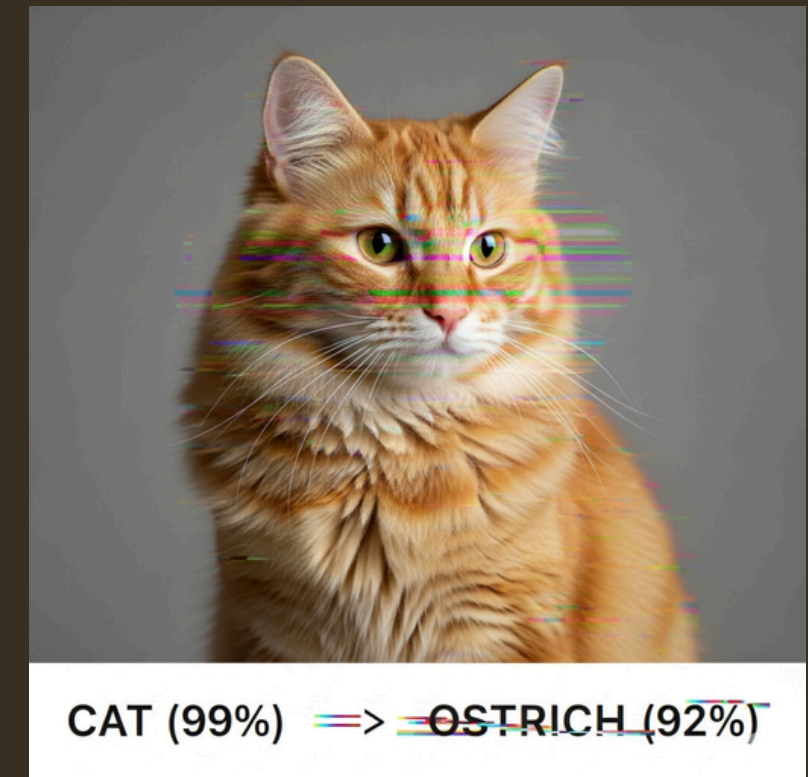
Developing robust AI requires understanding and addressing these **critical weaknesses**.



# cat versus adversarial ostrich scenario

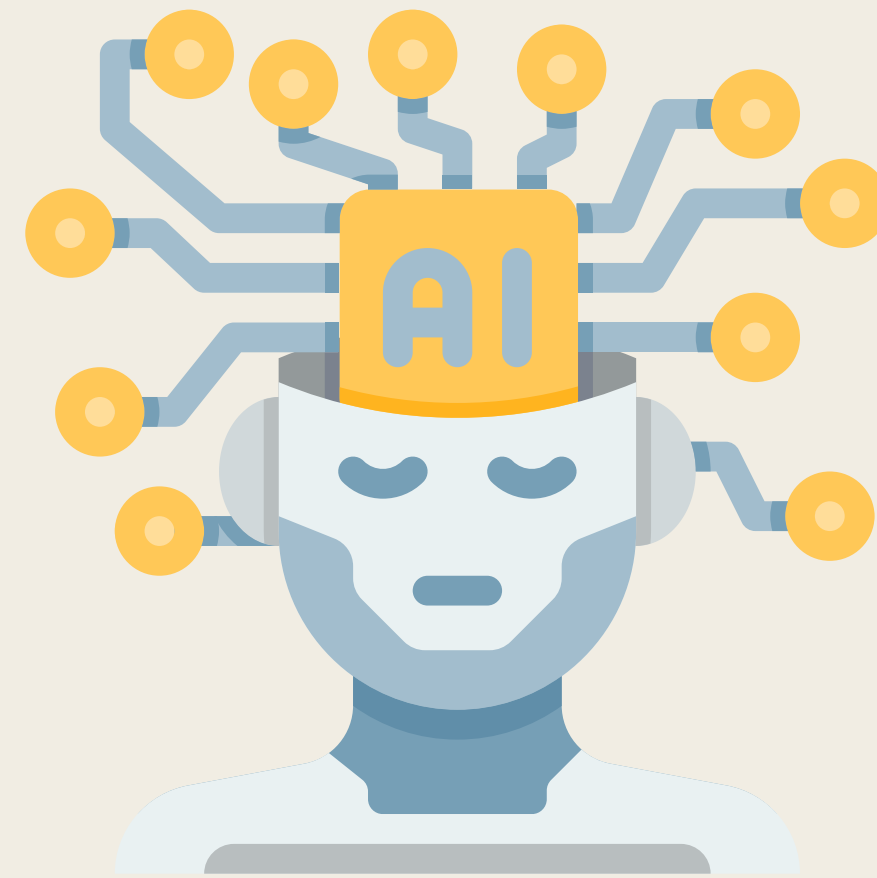


crafted layer of noise



# Understanding Spectral Bias in AI

AI models have a "spectral bias": they easily learn low-frequency signals but are confused by high-frequency noise, which is what adversaries exploit



# Proposed Solution: FGR and Noise Filtering

In this section, we explore how **Fourier Gradient Regularization** addresses **AI brittleness** through effective noise filtering techniques.


$$L_{\text{total}} = J_{\text{CE}} + \lambda \cdot \left\| H \left( F \left( \nabla_x J \right) \right) \right\|_2^2$$

## FGR Formula

The FGR formula is designed to **penalize high-frequency sensitivity**, enhancing the stability of AI models against adversarial attacks.

## Noise Filtering

By implementing noise filtering, we can **reduce unwanted variations** in data, resulting in a more robust AI performance under challenging conditions.

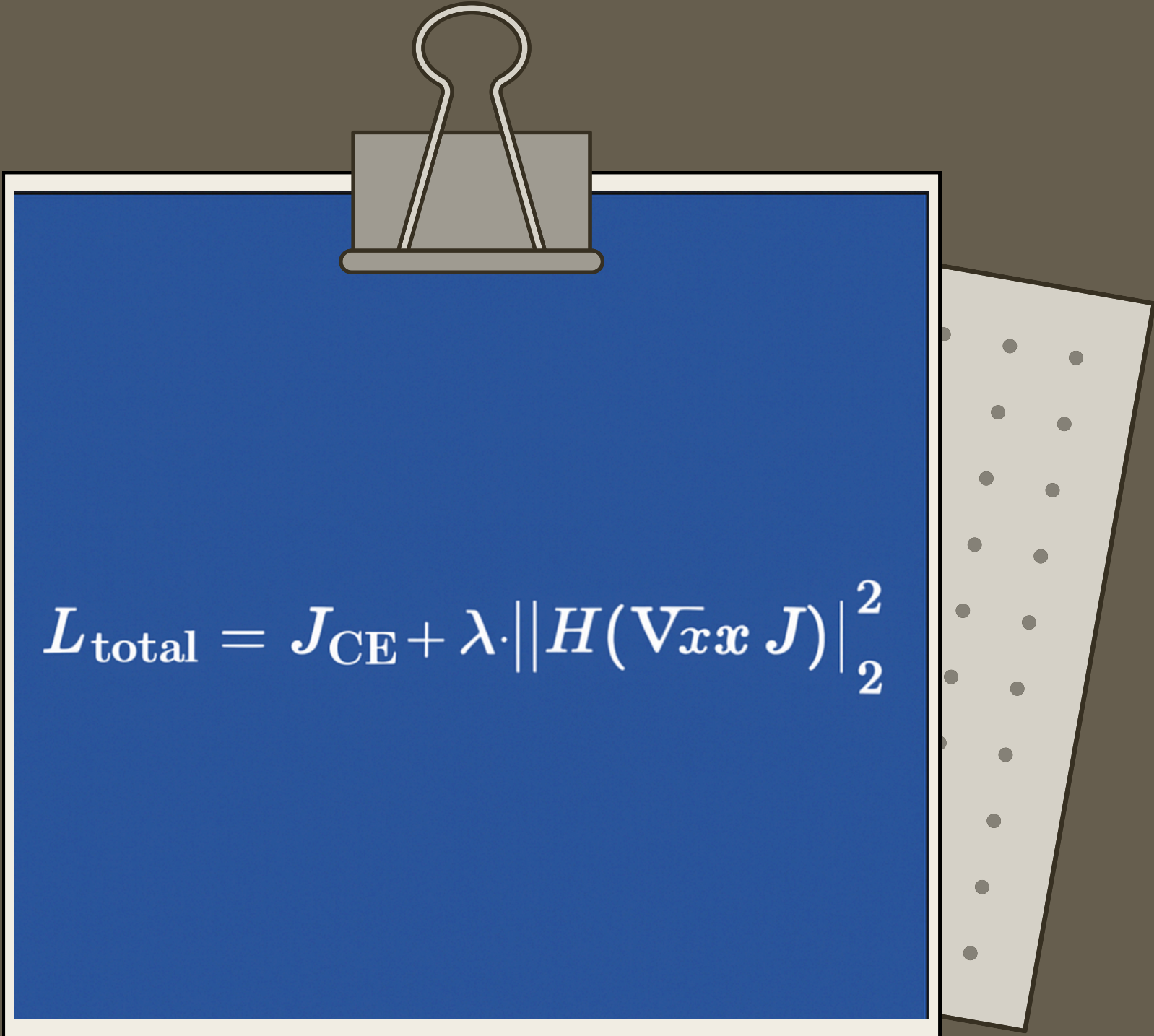
## High-Frequency Penalty

The penalty for high-frequency sensitivity ensures that models maintain **consistent performance**, improving reliability in safety-critical applications.

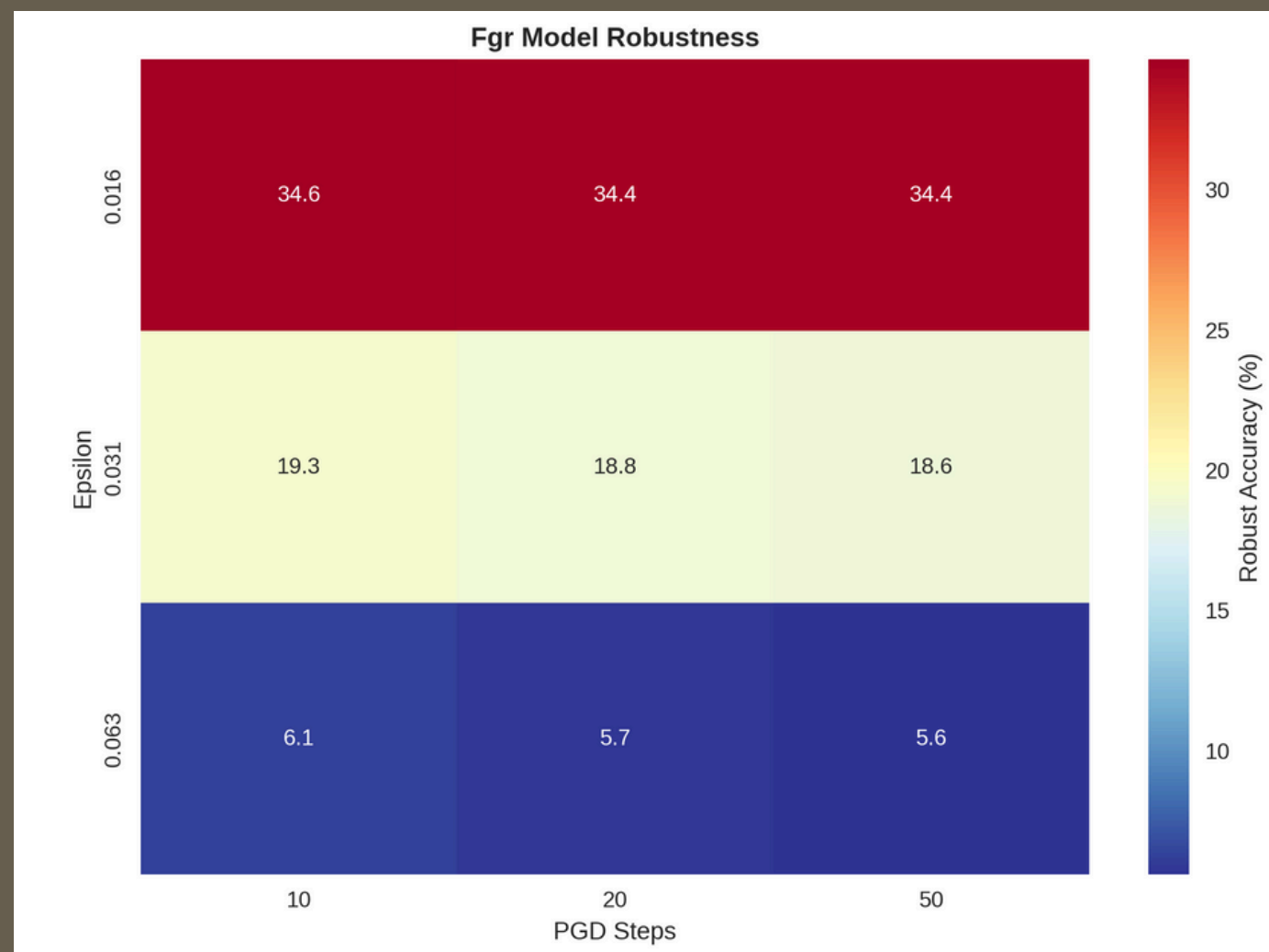
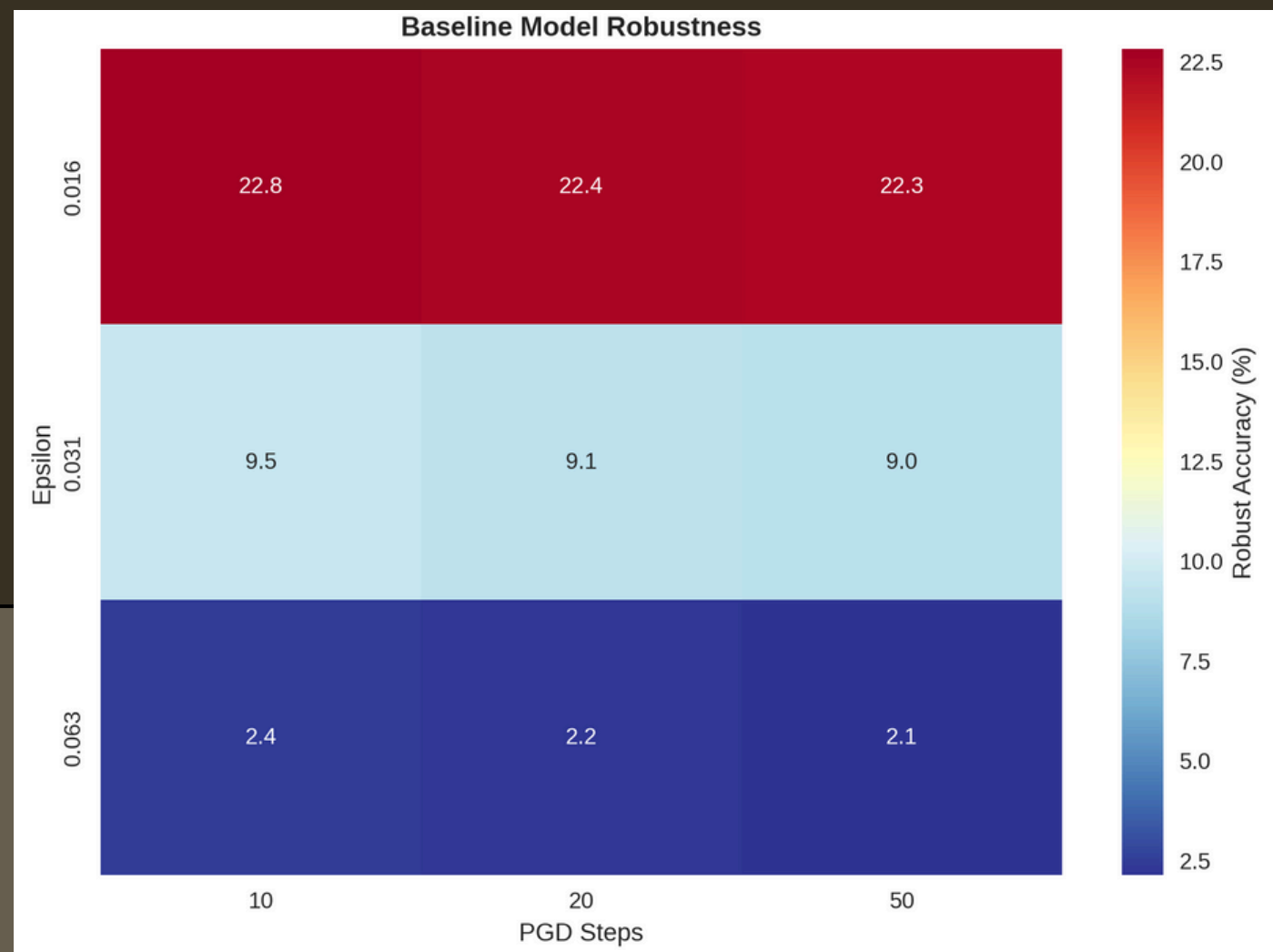


# Understanding loss function

where JCE is the cross-entropy loss, F is the 2D Fourier Transform, H is a high-pass filter, and  $\lambda$  is the regularization strength. This penalty forces the optimizer to find solutions in the model's weight space that correspond to smoother functions in the input space.

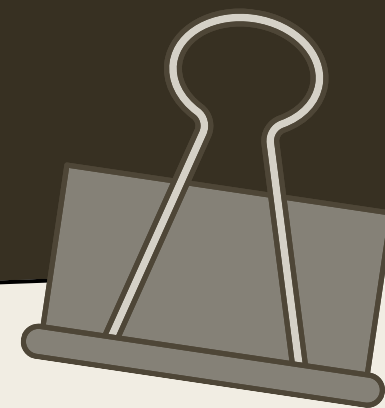

$$L_{\text{total}} = J_{\text{CE}} + \lambda \cdot \|H(V_{xx} J)\|_2^2$$

Visual representation of low- and high-frequency influences on AI models.



# Improved Robustness with FGR

The results demonstrate a **significant enhancement** in AI robustness through Fourier Gradient Regularization (FGR). The heatmap illustrates how FGR effectively mitigates vulnerabilities by penalizing high-frequency sensitivity. This **innovative approach** leads to a more resilient AI model capable of withstanding adversarial attacks while maintaining performance. As evident from the data, FGR has the potential to transform how we secure AI systems, ensuring their reliability and safety in critical applications.



**Thank you for your  
attention!**

