

# Impact of Layer Norm on Memorization and Generalization in Transformers, NeurIPS 2025

Rishi Singhal

rsingha4@ncsu.edu

Jung-Eun Kim

jung-eun.kim@ncsu.edu

Computer Science,  
North Carolina State University

# What is Memorization & Generalization?

# Memorization & Generalization

## Memorization

1. Memorizes **specific** samples  
(*atypical, confusing, mislabeled*)
2. Fails to learn **generic patterns**
3. Performs **poorly** on **test** data



→ Dog

Mislabeled, Atypical Example

## Generalization

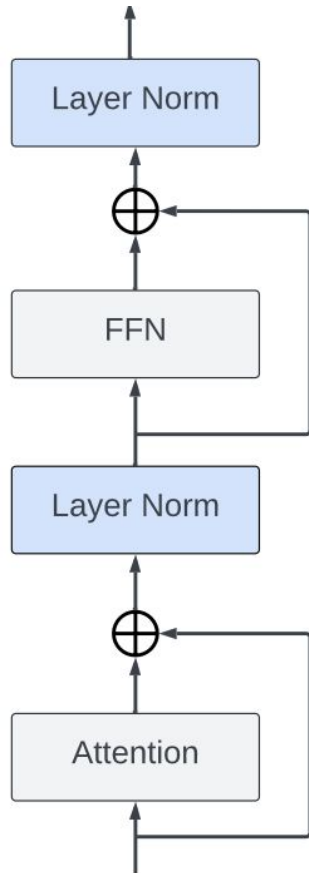
1. Learns from **typical** samples  
(*simple, representative*)
2. Captures **generic patterns**
3. Performs **well** on **test** data



→ Cat

Simple, Typical Example

# Layer Normalization (LN) Layer in Transformer



Original Transformer  
Architecture

$$\text{LN}(x) = w \odot N(x) + b$$

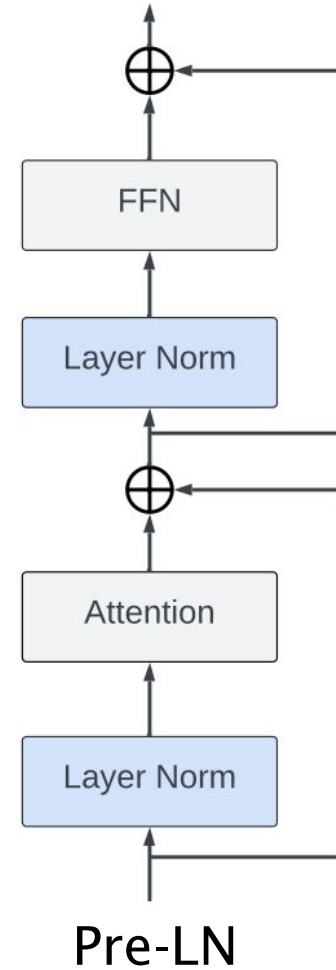
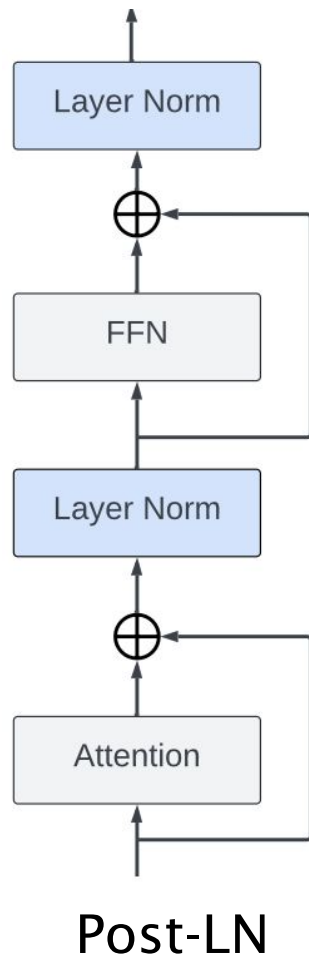
$\text{LN}(x)$ : Layer Normalization operation

$x$ : Input vector of size  $d \times 1$

$N(x)$ : Normalization operation

$w$  and  $b$ : Learnable parameters


# Post-LN vs. Pre-LN Transformer



# Removing LN Parameters to study impact on Memorization and Generalization

To study the impact of LN on memorization and generalization, we remove the learnable parameters ( $w$  and  $b$ ) prior to training, following Xu et. al (2019).

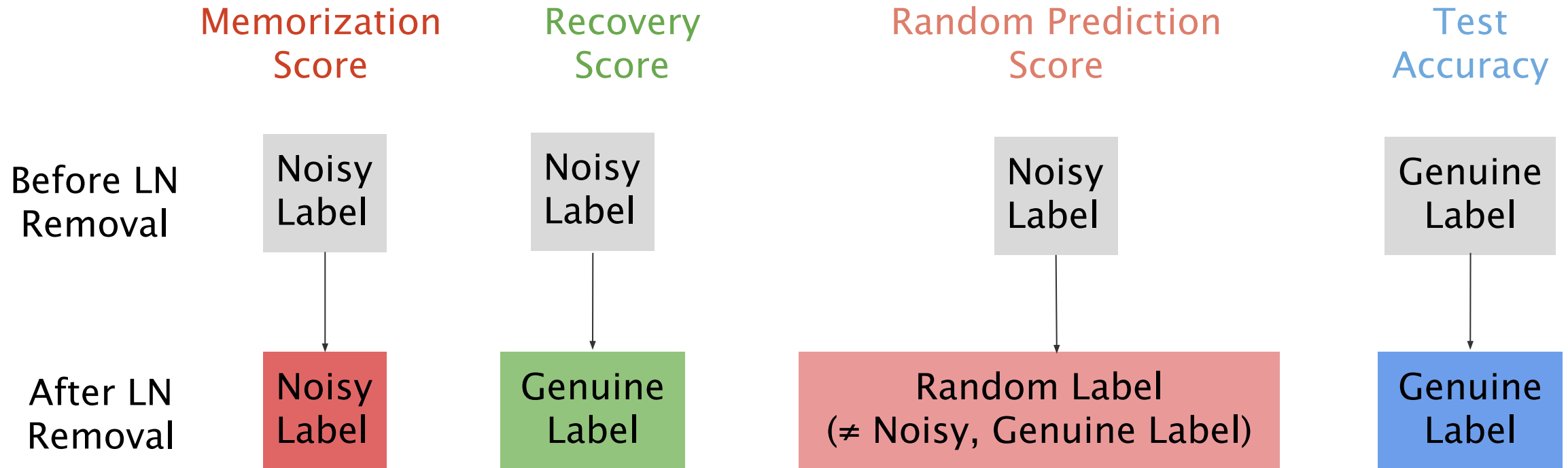
Removing



The diagram shows the word "Removing" in red at the top. Two red arrows originate from the word and point downwards to the parameters  $w$  and  $b$  in the equation below.

$$\text{LN}(x) = w \odot N(x) + b$$

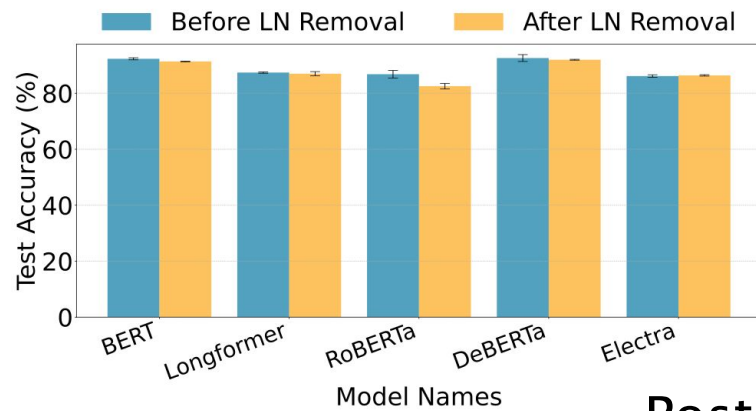
# Metrics: Memorization, Recovery, Random Prediction Score and Test Accuracy



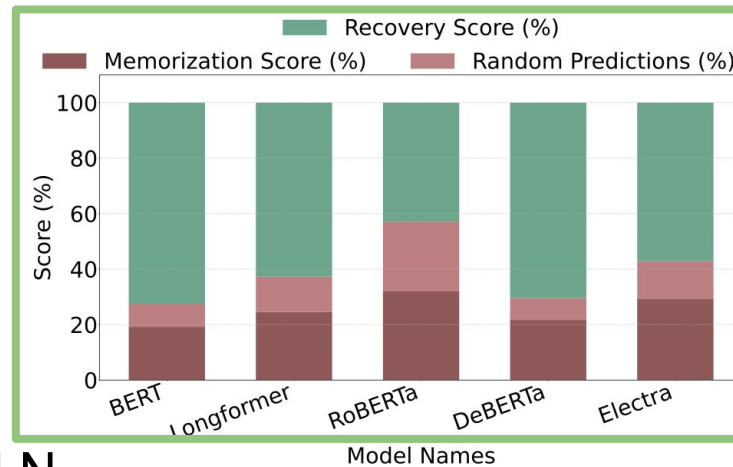
# Impact of LN Removal on Memorization and Generalization



# LN removal mitigates Memorization in Post-LN but degrades Generalization in Pre-LN

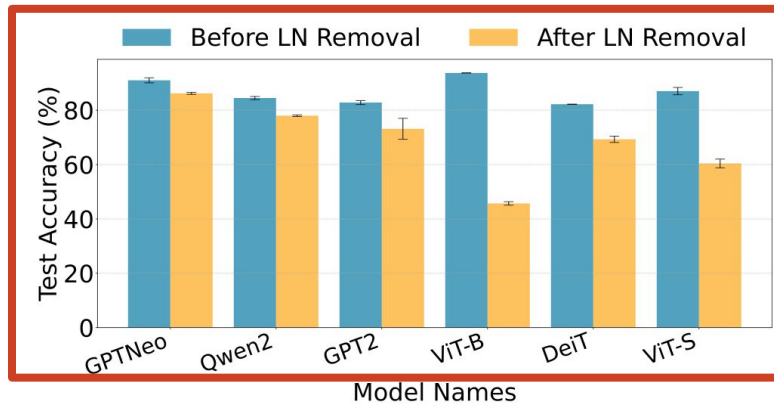


Post-LN

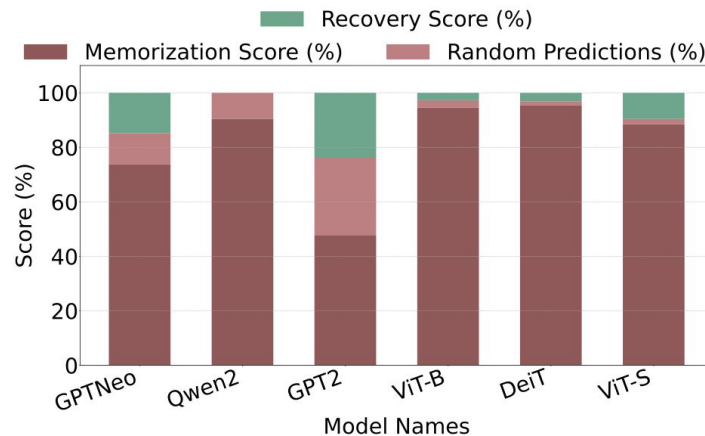


LN removal:

- Mitigates memorization in Post-LN models
- Degrades generalization in Pre-LN models.



Pre-LN

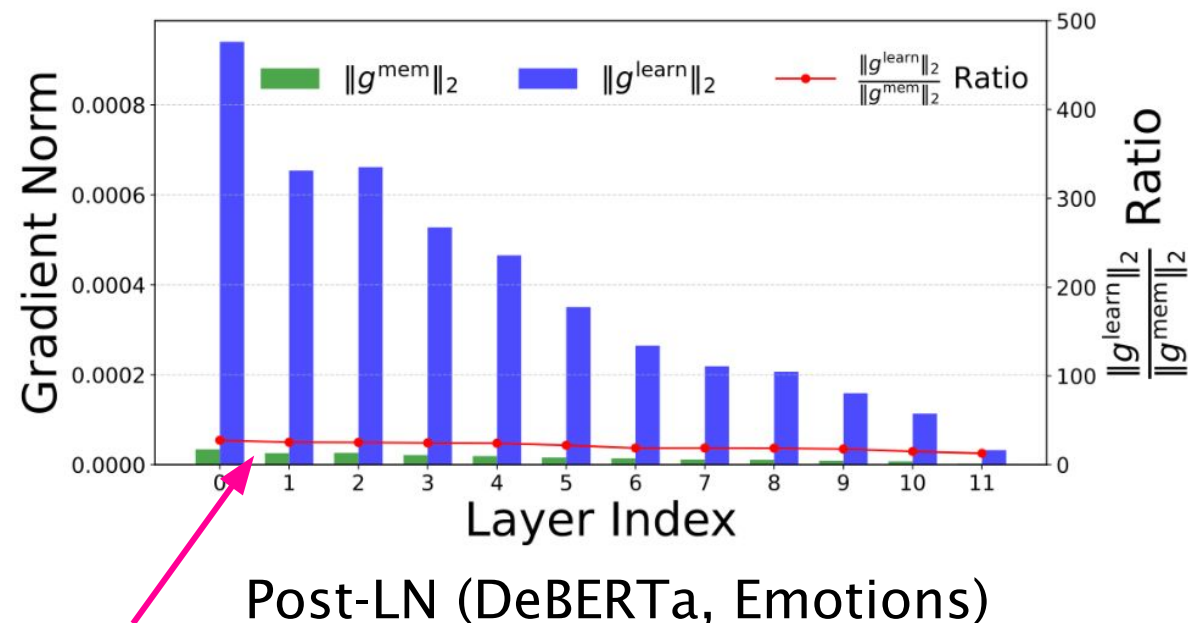
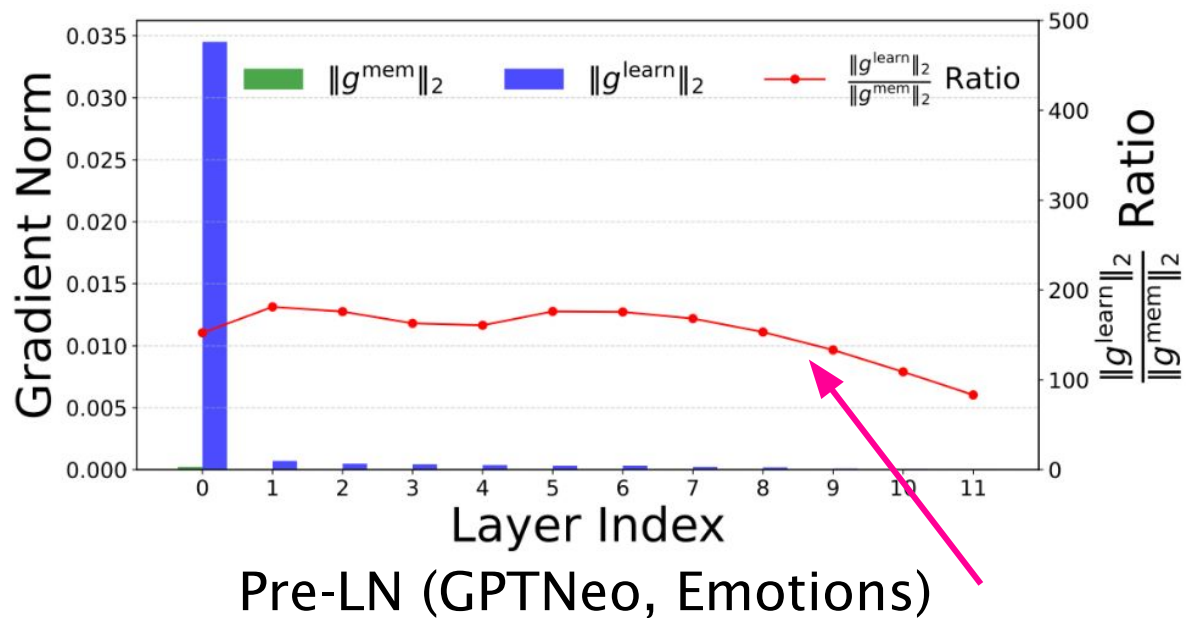


# Overall Trend

If LN Removed	Generalization Intact?	Memorization Mitigated?	Recovery Happens?
Post-LN Model	✓ Generalization Intact	✓ Memorization Mitigated	✓ Genuine Labels Inferred
Pre-LN Model	✗ Generalization Degraded	✗ Memorization Still Present	✗ Negligible Recovery

# Gradients Analysis

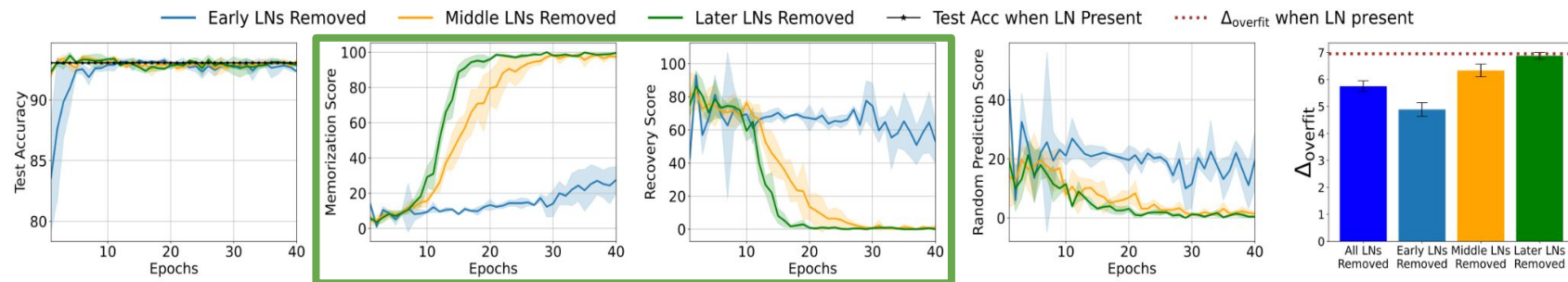
# Gradients Explain Why



Learning gradient norms are much larger than memorization gradient norms in Pre-LN models, in comparison to Post-LN.

# Which layers LNs are Critical: Early, Middle or Later?

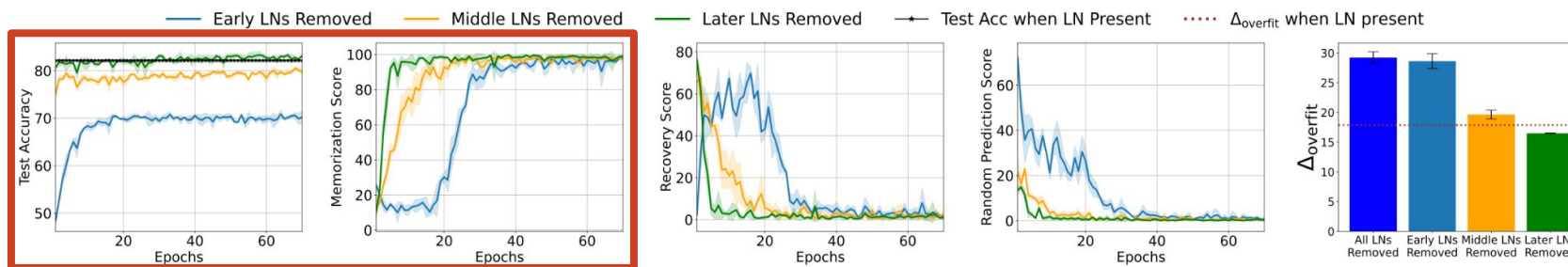
# Early Layers LNs are Most Critical



Post-LN (DeBERTa, Emotions)

Early Layers LNs removal (blue line plot) are the most critical in

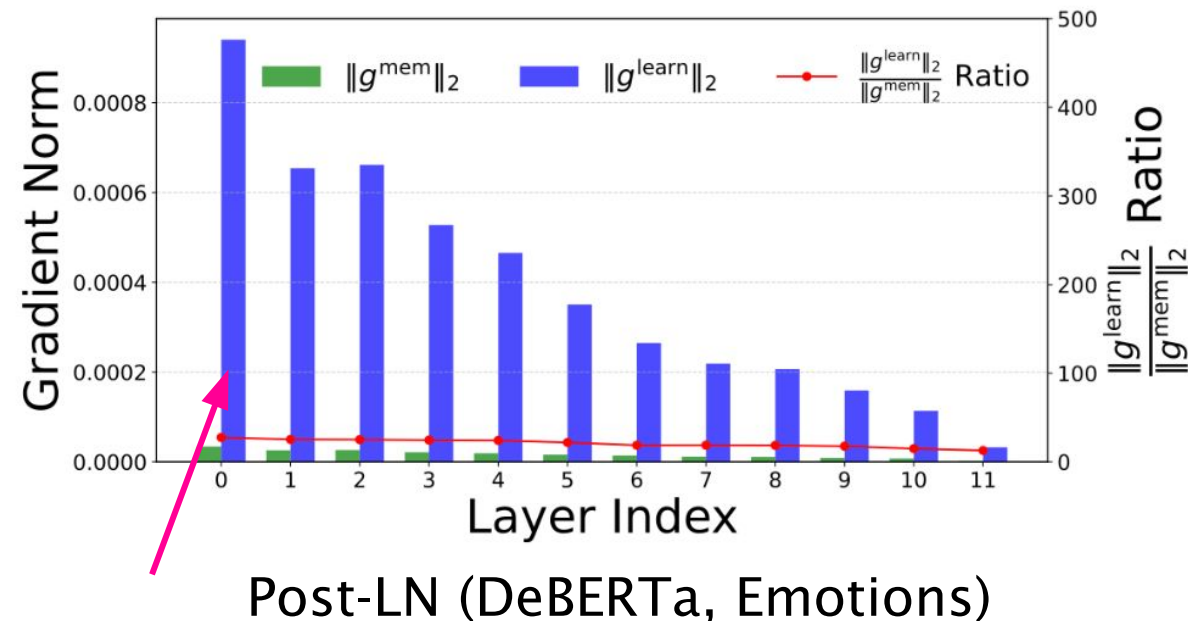
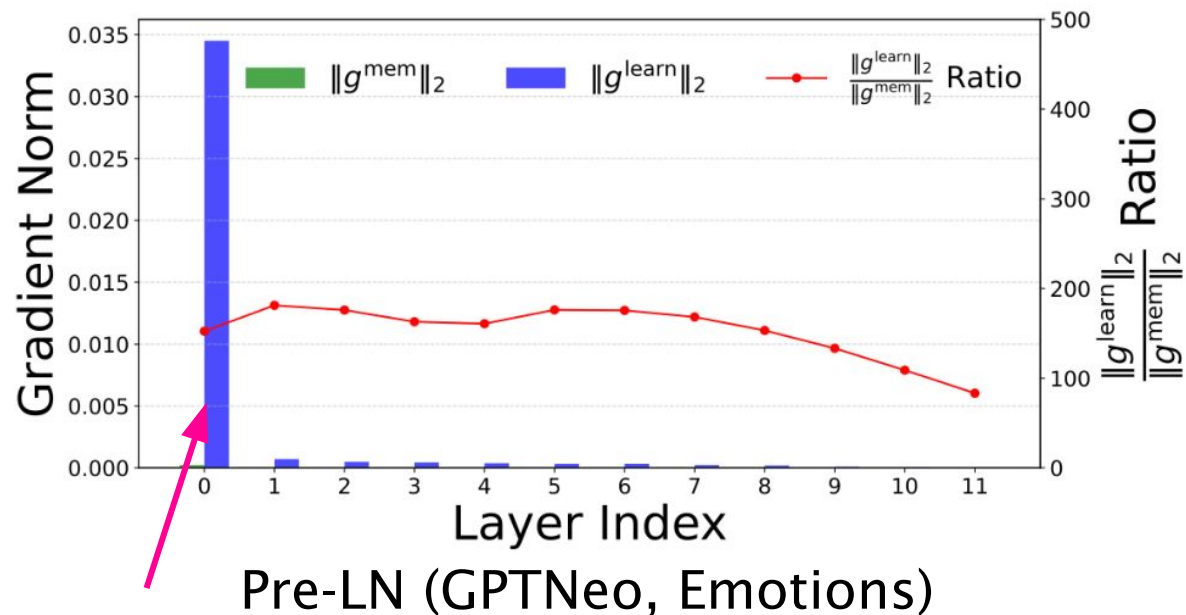
- Mitigating memorization in Post-LN models
- Degrading generalization in Pre-LN models



Pre-LN (DeiT, UTK-Face)

# Gradients Analysis

# Gradients Explain Early LNs Impact



Early layers LNs exhibit higher gradient norms than middle and later layers.



# Take-Home Points

- Layer Normalization (LN) critically shapes both generalization and memorization in Transformers, but the influence varies with LN placement.
- Removing LN parameters mitigates memorization in Post-LN models but degrades generalization in Pre-LN models.
- Early layers LNs are the most influential, and gradient analysis reveals distinct LN effects across the Pre- and Post-LN models.

# Thank you