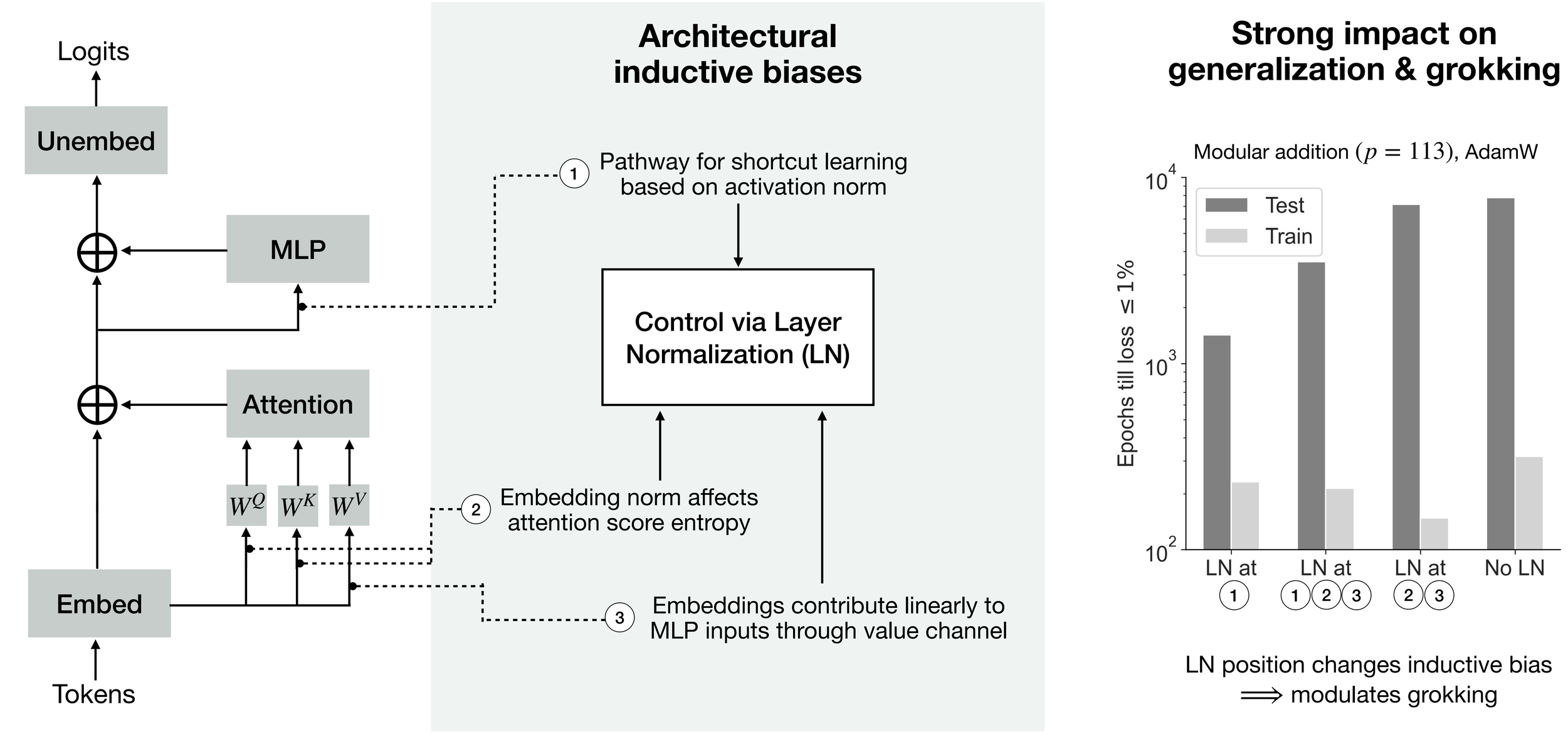
Logits
Unembed
MLP
Attention
W^Q
W^K
W^V
Embed
Tokens
Architectural inductive biases
1
Pathway for shortcut learning based on activation norm
Control via Layer Normalization (LN)
2
Embedding norm affects attention score entropy
3
Embeddings contribute linearly to MLP inputs through value channel
Strong impact on generalization & grokking
Modular addition ($p = 113$), AdamW
Epochs till loss ⩽ 1%
10^4
10^3
10^2
Test
Train
LN at 1
LN at 1 2 3
LN at 2 3
No LN
LN position changes inductive bias
⟹ modulates grokking