Hidden Score To next layer **Hallucination Detected**  $m \times m$  $m \times d$ Transformer Block dxmYes Score  $\log \det \left( \mathbf{\Sigma}^2 \right) = \log \prod^m \sigma_i^2$ >= **Layer Normalize** threshold No **Hallucination** Hidden Feedforward Layer  $\mathbf{H}^{\mathbf{T}}$  $\left[ \mathbf{\Sigma}^{2} = \mathbf{H}^{\mathbf{T}} \mathbf{H} \right]$ Activations **Not Detected Attention Score Hallucination Detected** Layer Normalize  $m \times m$ Yes Score  $\log \det(Ker_i) = \sum_{i=1}^{m} \log Ker_i^{jj}$ >= Self-Attention Layer threshold  $|Ker_i|$ No Hallucination Lower-triangular **Not Detected** Self-Attention Kernel Map

Prompt: Explain how neural networks work in Layman's terms Response: A neural network is a computer program designed to mimic how the human brain works ...

<u>Legend</u>:

m = number of tokens in prompt and response

concatenation d = size of hidden dimension