## **How an LLM Works** The LLM processes LLM the tokens and predicts the next token repeatedly Transformer The attention Input Output Tokeni-Layers mechanism zation **Text** determines how **Text** much importance Attention to give each token Input text A token is is split generated The feedforward **Feedforward** into smal ler pieces network further for each step processes the and tokencalled are transforinformation tokens The LLM processes med into the tokens and the output tat predicts the next token repeatedly