GPT Pre-training: Causal Language Modeling

Autoregressive Prediction Process

1 The → cat — *Predict next token*

2 The cat → sat — *Use previous outputs*

3 The cat sat → on — *Continue generation*

Training Data
Diverse internet text: **Common Crawl, WebText**

Unidirectional Attention
Enables **efficient generation**

BPE Tokenization
Byte-pair encoding for vocabulary

No Explicit Fine-tuning
GPT-3: **Direct inference** capable

Pattern Learning

Massive Scale

Efficiency

Learns patterns, facts, and reasoning from data

Unlocks in-context learning abilities

Fast generation through causal masking