

GPT Training Objective: Mathematics

Autoregressive Language Modeling Objective

$$\mathcal{L}_{AR} = - \sum_{t=1}^T \log P(w_t | w_1, \dots, w_{t-1})$$

Each
position

Cross-entropy
loss

Conditional on
past only

Causal
masking

Example: "The cat sat on"

Predict next: $P(\text{the}|\text{past}) = 0.88$

$\text{Loss} = -\log(0.88) = 0.13$