

Language Modeling Objective Functions

Autoregressive LM

Causal / Left-to-Right



$$P(w_t | w_1, \dots, w_{t-1})$$

→ Predict **next token**

⚡ **Causal masking** (uni-directional)

✍ Good for **generation** tasks

📝 GPT-style models

Masked LM

Bidirectional Context



$$P(w_{\text{masked}} | \text{context})$$

🎯 Predict **masked tokens**

↔ **Bidirectional** context (left+right)

🧠 Better **understanding**

BERT-style models

Training Approach

Self-supervision enables learning from unlimited text using **cross-entropy loss**

Architecture Fit

Different objectives suit different architectures and **downstream tasks**