Teacher Forcing vs. Autoregressive Decoding

Teacher Forcing

Training

How it Works

- Uses **ground-truth** target tokens as input
- Model receives correct previous token
- No error accumulation during training

✓ Advantages

+ Faster convergence
+ Stable gradients

Autoregressive

Inference

How it Works

- Uses model's **own predictions** as input
- Each prediction feeds into next step
- Errors can compound over sequence

✓ Advantages

+ Matches inference mode
+ More realistic training

+ Easier to train
+ Parallel processing

✗ Disadvantages

- Exposure bias problem
- Train/test mismatch
- Less robust at inference
- Can't handle own errors

+ Better generalization
+ No exposure bias

✗ Disadvantages

- Slower training
- Unstable gradients
- Error accumulation
- Sequential processing

📊 Quick Comparison

Aspect	Teacher Forcing	Autoregressive
Input Source	Ground truth	Model predictions
Usage Phase	Training	Inference
Training Speed	Fast ⚡	Slow 🐌
Robustness	Lower 📉	Higher 📈