Sequence Models: Practical Implementation Checklist

Essential considerations before deploying your model

Data Preparation

- [x] **Tokenization**
  Choose appropriate tokenizer (BPE, WordPiece, SentencePiece)
- [x] **Vocabulary Size**
  Balance between coverage and efficiency (typically 10K-50K)
- [x] **Special Tokens**
  Define PAD, UNK, BOS, EOS tokens correctly
- [x] **Sequence Length**
  Set max length based on data distribution (e.g., 128, 256, 512)

Model Architecture

- [x] **Hidden Dimensions**
  Choose appropriate size (256, 512, 1024) based on task
- [x] **Number of Layers**
  Balance depth and training time (2-12 layers typical)
- [x] **Attention Mechanism**
  Implement proper attention with correct masking
- [x] **Dropout Rates**
  Add dropout for regularization (0.1-0.3 typical)

Training Strategy

- [x] **Teacher Forcing**
  Use during training, consider scheduled sampling
- [x] **Batch Size**
  Optimize for GPU memory (32-128 typical)

Implementation Details

- [x] **Padding & Masking**
  Properly handle variable-length sequences
- [x] **Loss Calculation**
  Ignore padding tokens in loss computation

Learning Rate

Use warmup + decay schedule (e.g., 1e-4 to 1e-5)

Gradient Clipping

Prevent exploding gradients (clip norm 1.0-5.0)

Inference Strategy

Choose beam search, greedy, or sampling decoding

Evaluation Metrics

Track BLEU, perplexity, or task-specific metrics

Quick Reference: Typical Hyperparameters

Hidden Size	Batch Size	Vocab Size	Learning Rate
256-1024	32-128	10K-50K	1e-4 ~ 1e-3