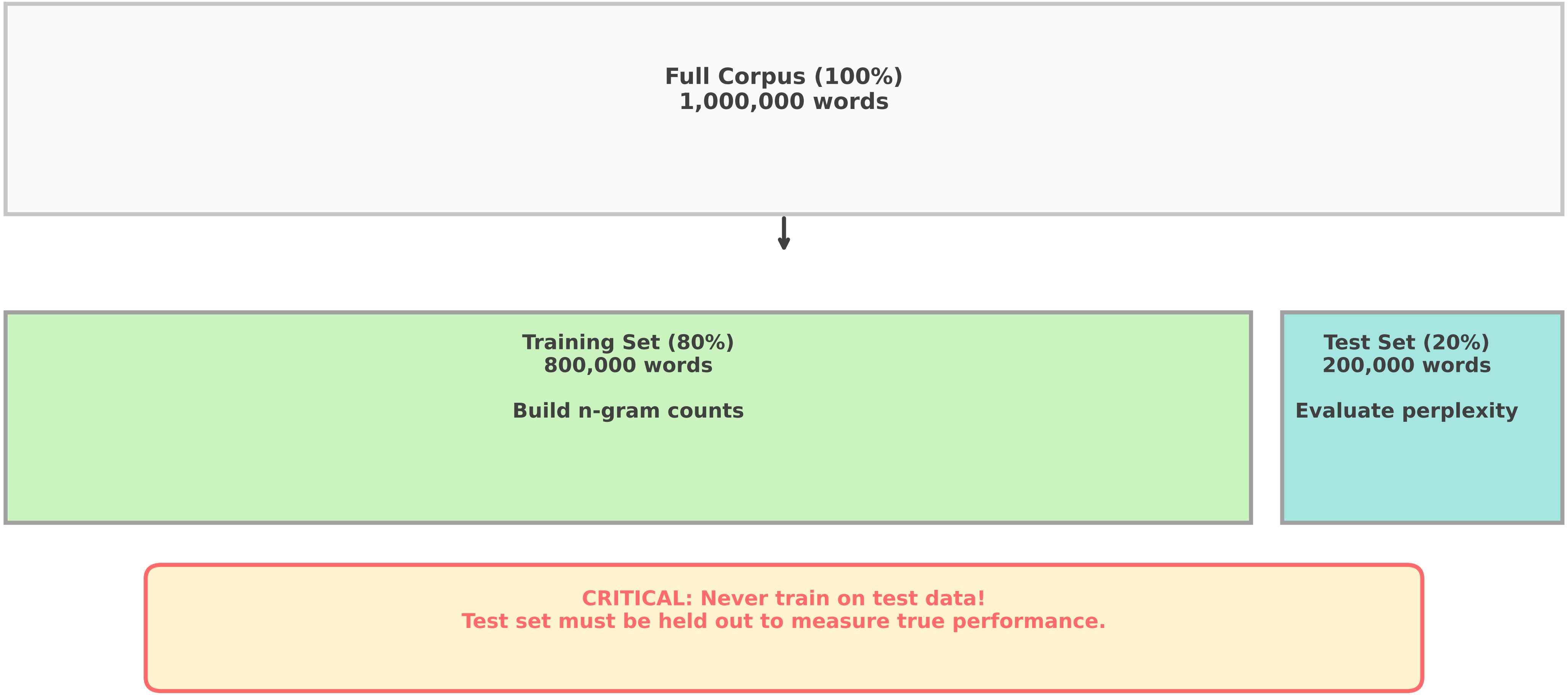
Train/Test Split: Proper Evaluation Setup



Full Corpus (100%)
1,000,000 words

Training Set (80%)
800,000 words

Build n-gram counts

Test Set (20%)
200,000 words

Evaluate perplexity

CRITICAL: Never train on test data!
Test set must be held out to measure true performance.