Language Model Performance on Penn Treebank
(40K vocabulary, 1M tokens)

Perplexity (lower is better)

Model Type

Unigram	Bigram	Trigram	5-gram (Kneser-Ney)	LSTM (2018)	Transformer (GPT-2)
962	170	148	141	78	35

Adding context helps a lot! -82%

Neural models break through! -45%

Attention is all you need! -55%

Perplexity ≈ Average branching factor (choices per word)