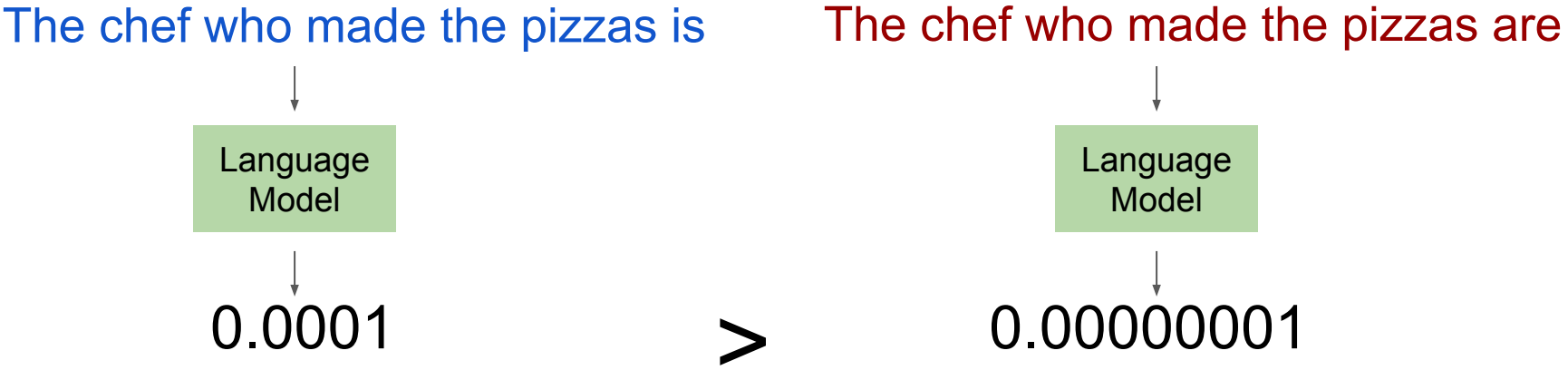
Neural networks as linguistic test subjects

How do we understand language behavior in **language models**?

One method: *minimal pairs. Is the acceptable sentence higher-probability?*



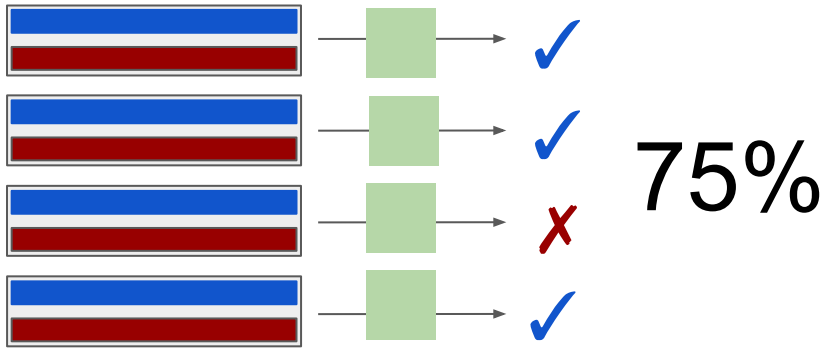
Premise: A language model should assign higher probability to the acceptable sentence in any minimal pair.

Linzen et al., 2016

Neural networks as linguistic test subjects

Steps to conduct a *minimal pairs* test on a language model:

1. Gather or construct a test set of minimal pairs which require specific aspects of understanding to distinguish.
2. Run your language model on the pairs, and report percent of pairs the model predicts as desired.



Neural networks as linguistic test subjects

Example: Do LMs show Subject-Verb number agreement across attractors?

The chef who made the pizzas and talked to the customers is

subject *attractor* *attractor* *verb*

	n=0	n=1	n=2	n=3	n=4
Random	50.0	50.0	50.0	50.0	50.0
Majority	32.0	32.0	32.0	32.0	32.0
LSTM, H=50†	6.8	32.6	≈50	≈65	≈70
Our LSTM, H=50	2.4	8.0	15.7	26.1	34.65
Our LSTM, H=150	1.5	4.5	9.0	14.3	17.6
Our LSTM, H=250	1.4	3.3	5.9	**9.7**	13.9
Our LSTM, H=350	**1.3**	**3.0**	**5.7**	**9.7**	**13.8**
1B Word LSTM (repl)	2.8	8.0	14.0	21.8	20.0
Char LSTM	**1.2**	5.5	11.8	20.4	27.8

# of attractors between subject and verb

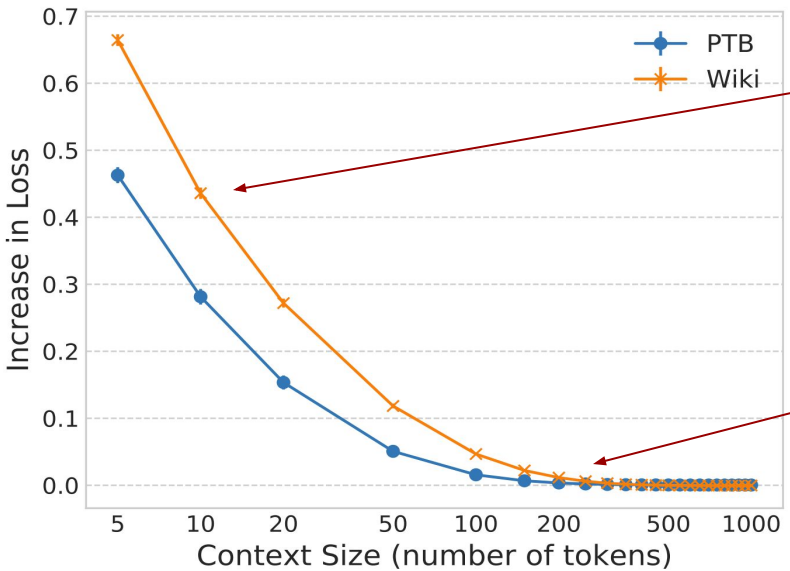
Error rate on a large corpus of minimal pairs

LMs do *really* well!?

[Kuncoro et al., 2018]

Neural networks as linguistic test subjects

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.



Only giving the LM 10 words of context at test time makes the test error go up.

Only giving the LM 250 words of context *doesn't change its loss*, so it's not using contexts longer than 250 words much.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

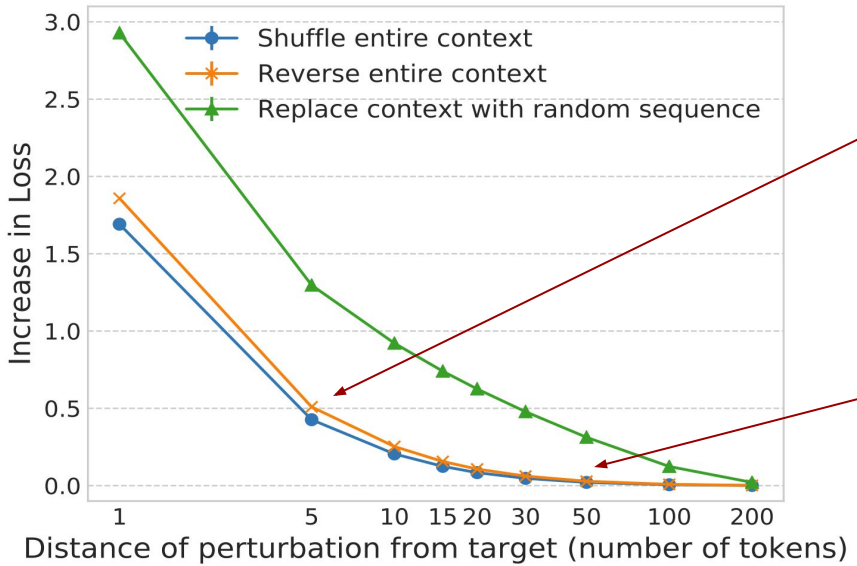
Question: How does an LSTM language model use its long-distance contexts?

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.



Shuffling the order of the context further than 5 words away increases loss, so the LM cares about word order past 5 words.

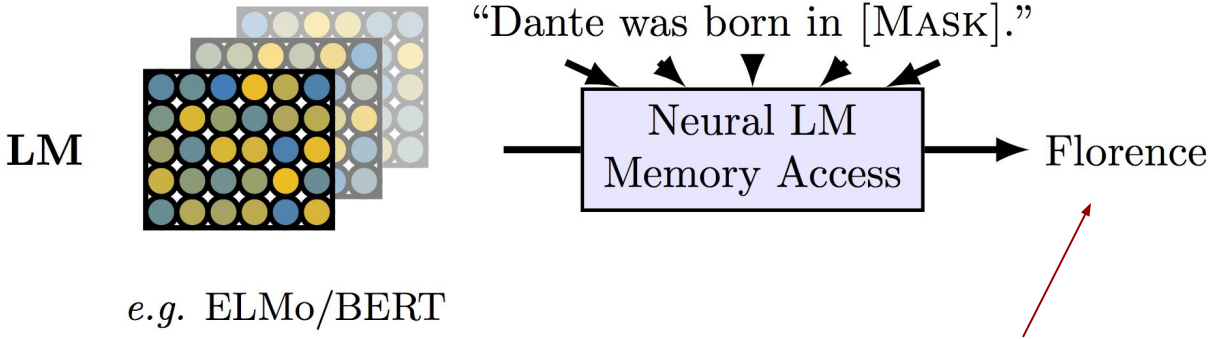
Shuffling the word order of the context further than 50 words away *doesn't* increase loss, so the LM treats words 50-250 effectively as a bag-of-words.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

Question: Do LMs memorize factual relations?

Method:



Check if most likely word under the LM is a correct answer.

Eval: % of these relations for which this holds.

[Petroni et al., 2019]

Neural networks as linguistic test subjects

Question: Do LMs memorize factual relations?

Evaluation:

Baseline: Return word that shows up most with the subject (Dante) and the relation (born in)

BERT-base and BERT-large: memorize a surprising number of facts

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**

[Khandelwal et al., 2018]