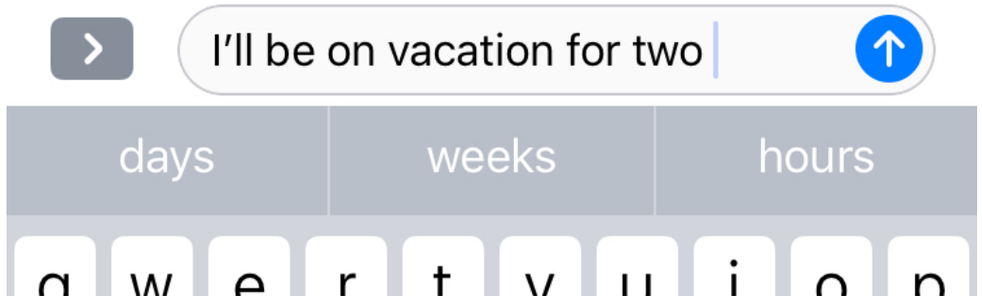
# Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context

Urvashi Khandelwal, He He, Peng Qi, Dan Jurafsky

# Language Modeling

# Language Modeling

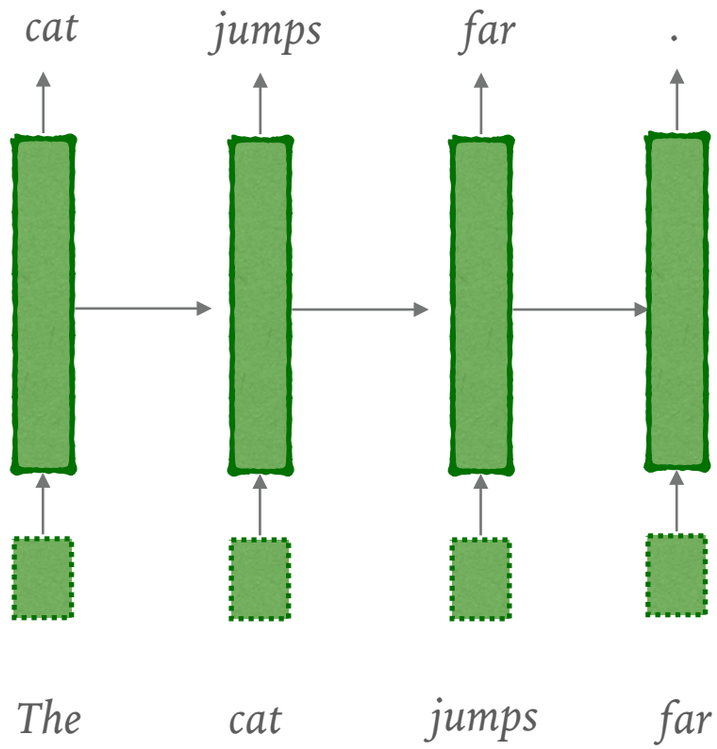



Goal: estimate the distribution of next possible words

$P(w_n \mid w_1 \dots w_{n-1})$

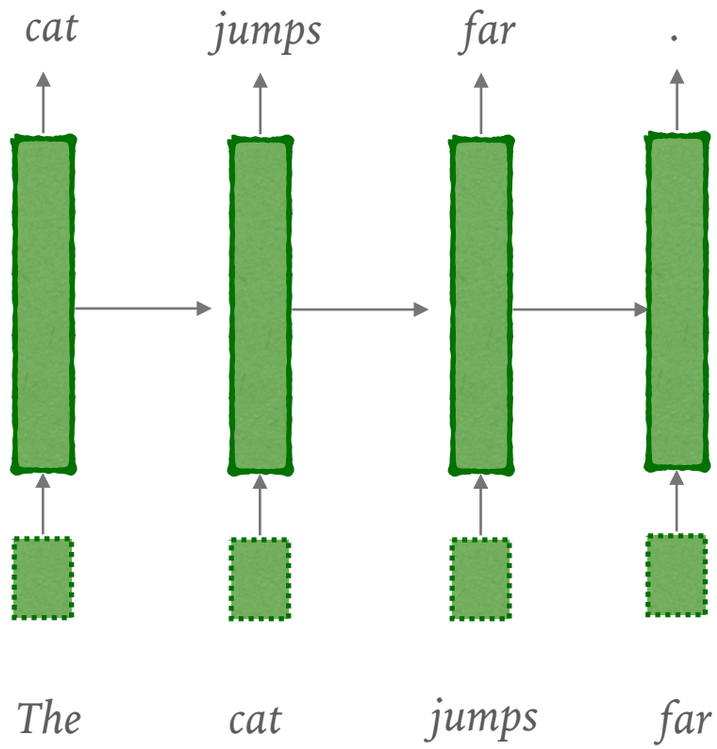
- Can be used to **generate** sentences word by word
- **Recurrent** architectures have been very successful at modeling long sequences

# Recurrent Neural Network




- Words are processed sequentially and information is passed from one state to the next
- Context unlimited (but training can have vanishing gradient issues)

# Recurrent Neural Network

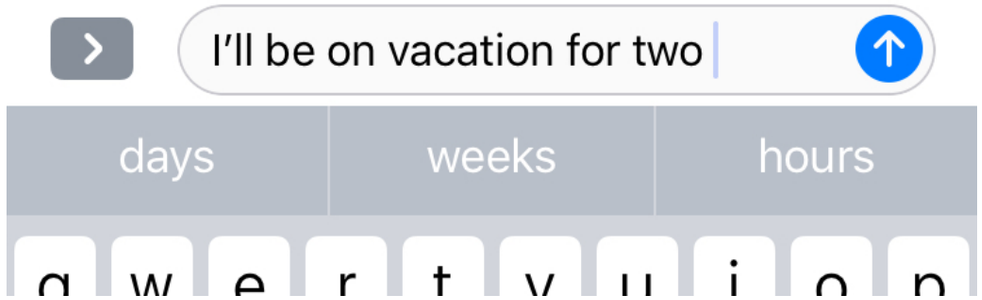



- Words are processed sequentially and information is passed from one state to the next
- Context unlimited (but training can have vanishing gradient issues)

Significant improvements over n-gram language models

Why?

# Language Modeling Evaluation




Goal: estimate the distribution of next possible words

$P(w_n \mid w_1 \dots w_{n-1})$

- Loss <-> Perplexity

# How do Language Models use Context?

# Experimental Setting

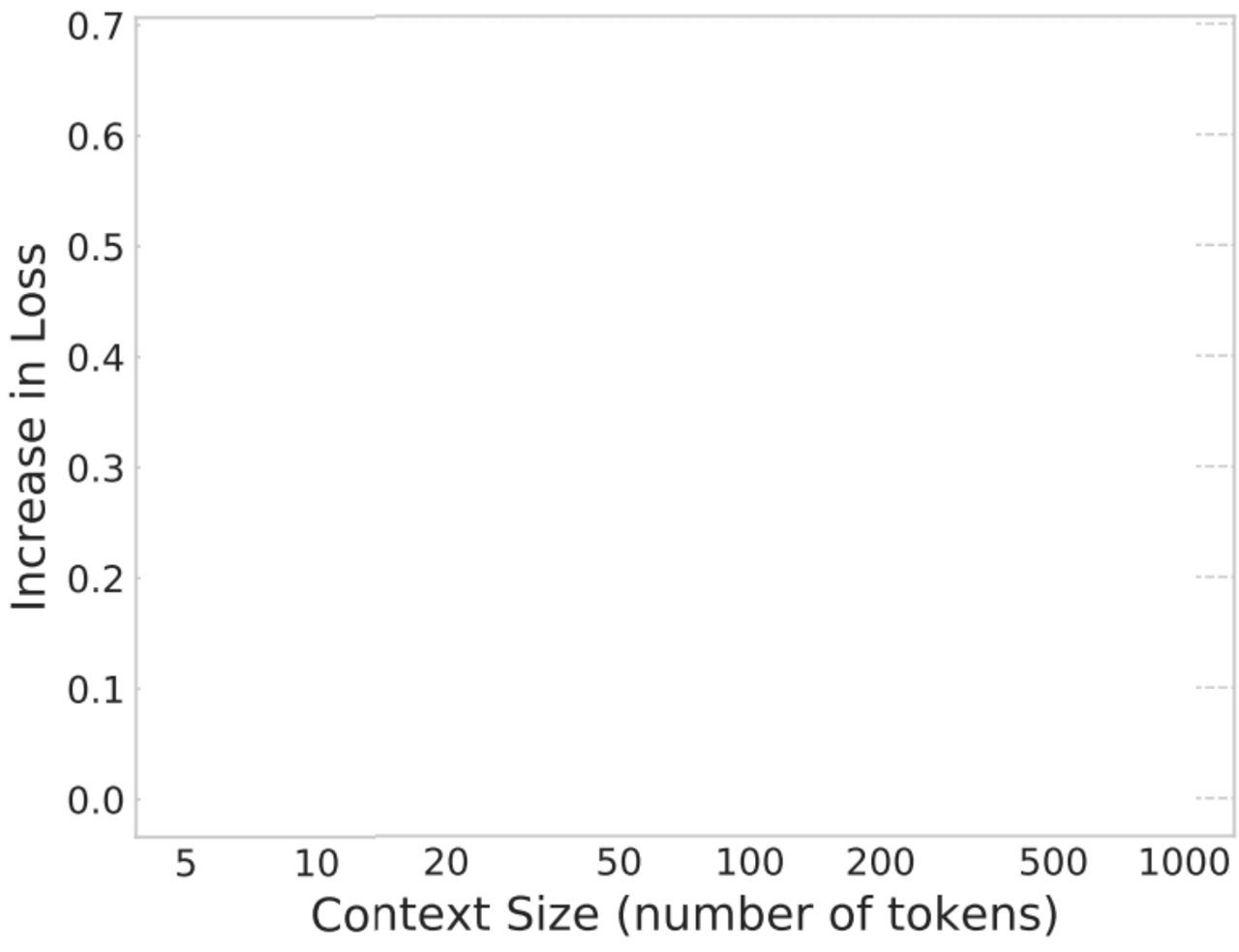
- Pretrained Model on 2 datasets
- Prior Context is changed at **test time only**

# Question 1: How much Context is used by LM?

- Motivation:
  - In RNNs, context is theoretically infinite
  - How much context is necessary for best performance?

# Question 1: How much Context is used by LM?

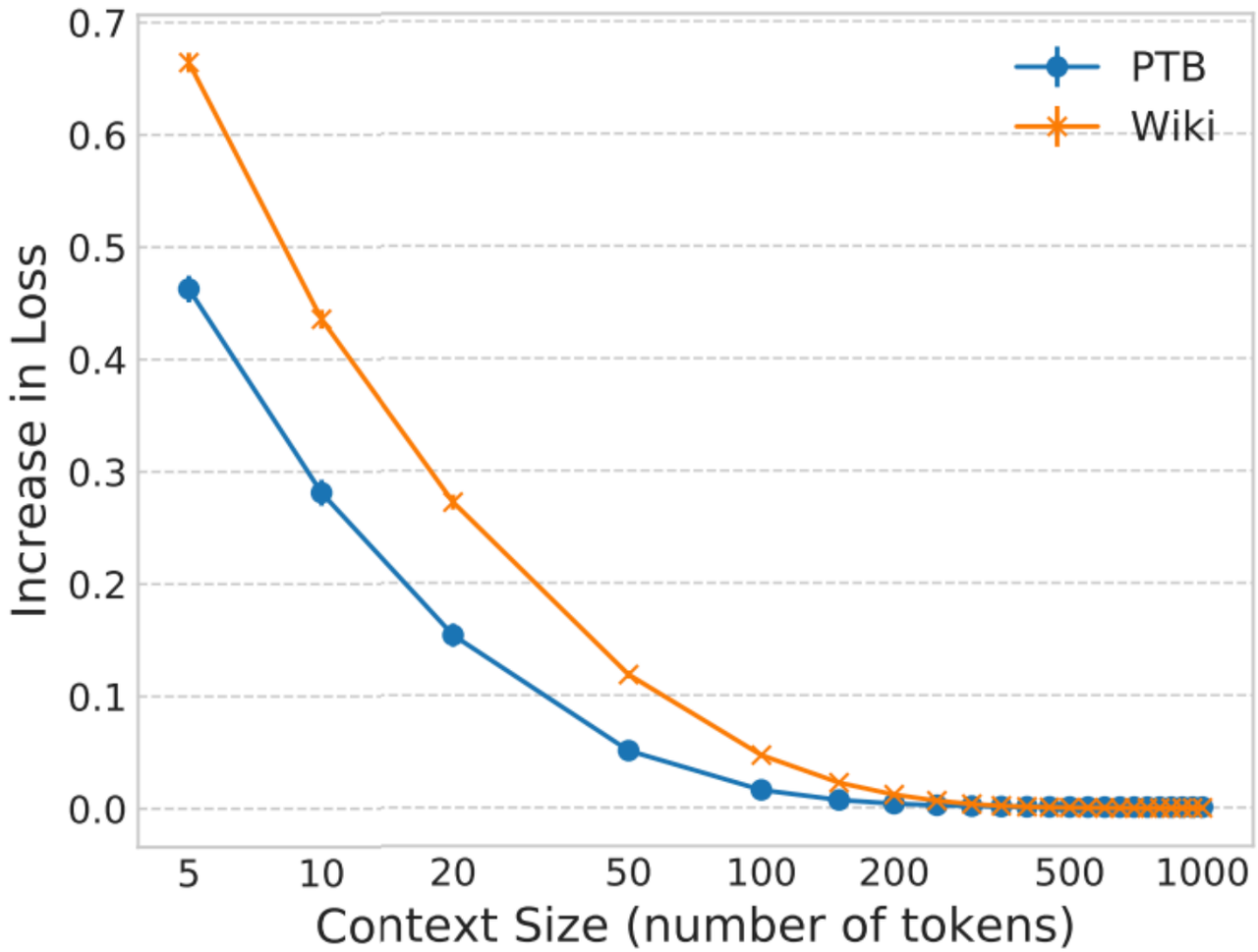
- Experiment: test time feed **most recent n tokens only**




(a) Varying context size.

# Question 1: How much Context is used by LM?

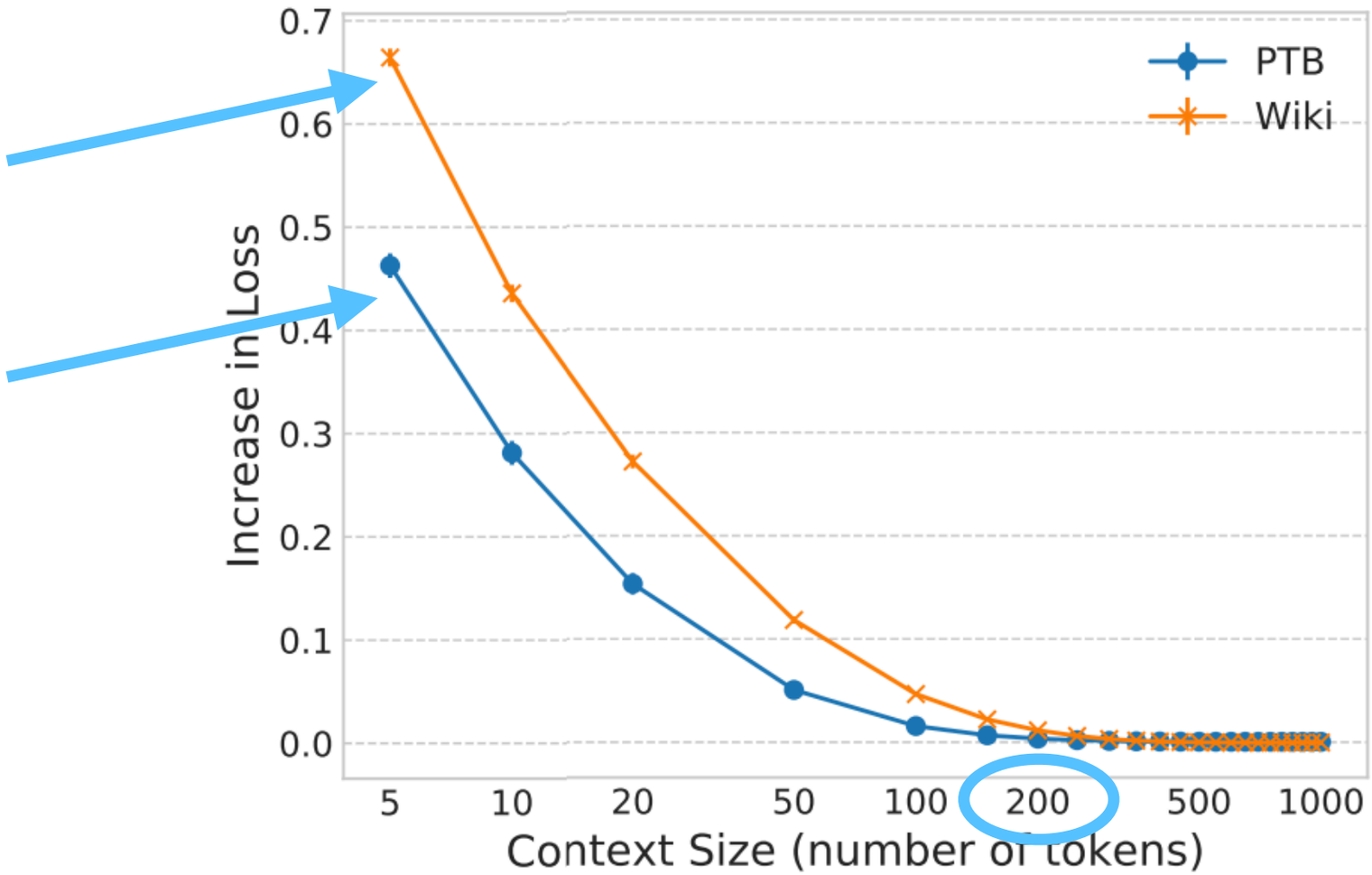
- Experiment: test time feed **most recent n tokens only**




(a) Varying context size.

# Question 1: How much Context is used by LM?

- Experiment: test time feed **most recent n tokens only**



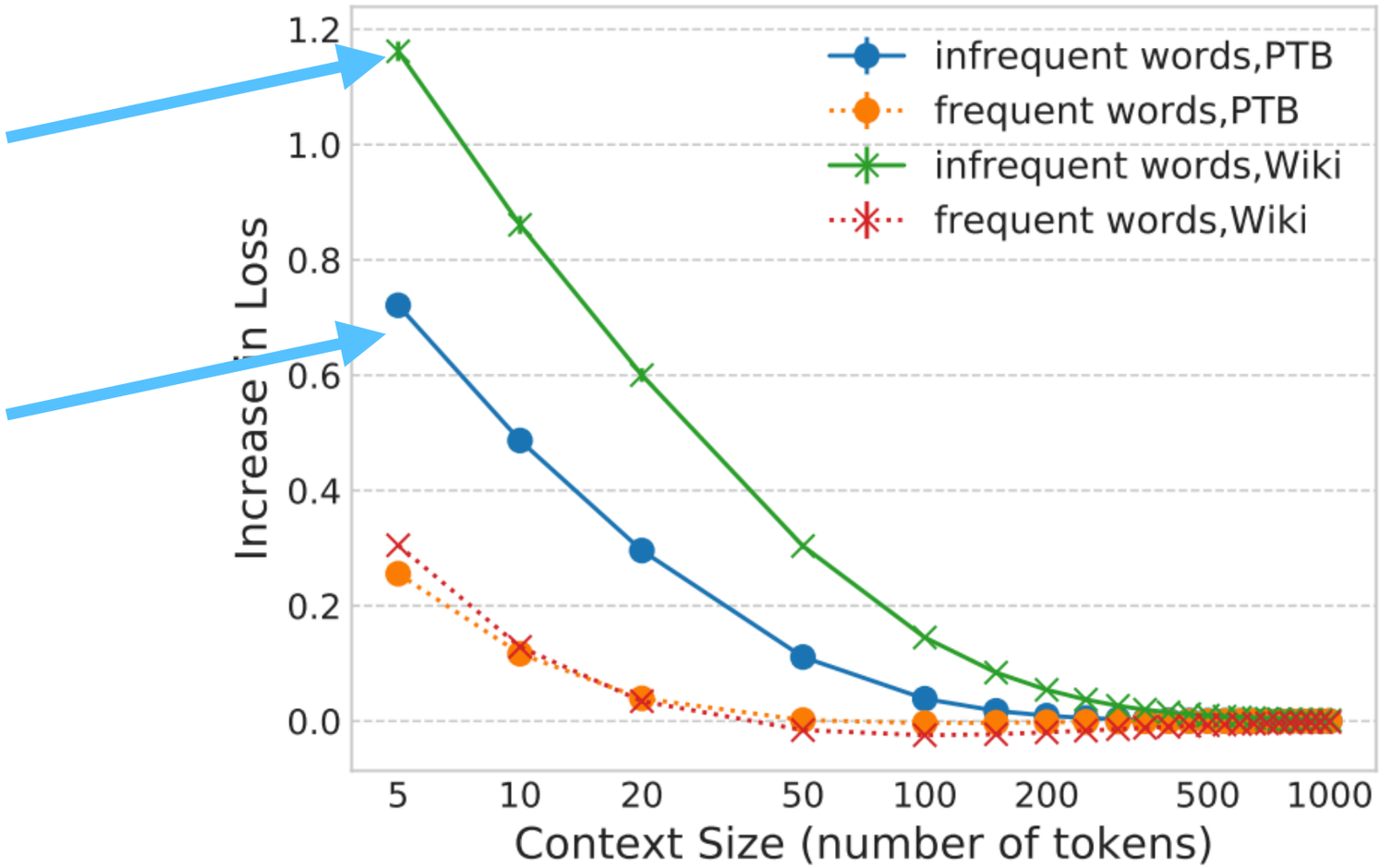

(a) Varying context size.

# Question 2: Do different words need different context?

- Motivation:
  - words such as ”the” or ”.” probably need less context to predict
  - function words v. content words
  - frequent words v. rare words

# Question 2: Do different words need different context?

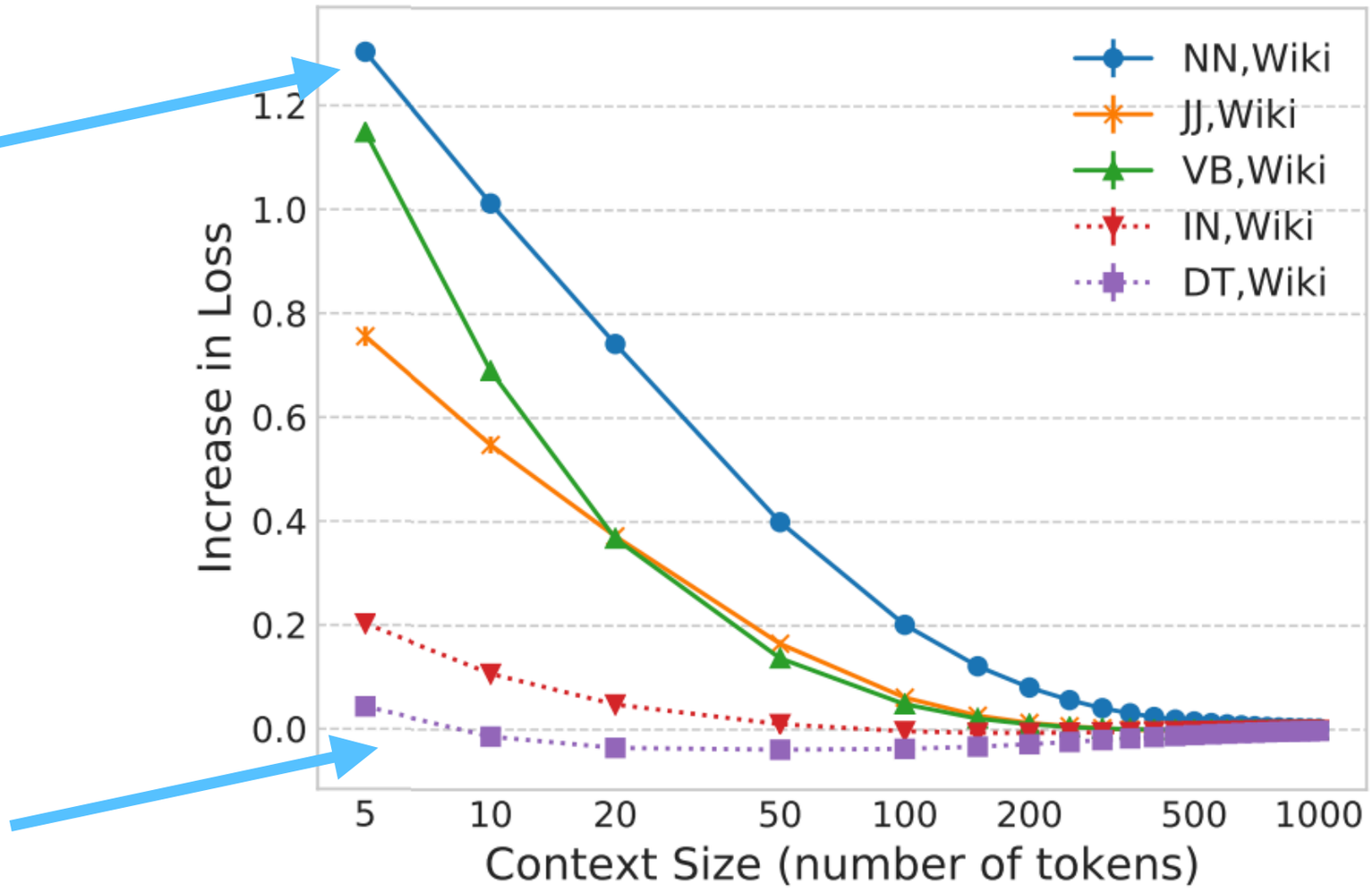
- Experiment: test time feed **most recent n tokens only**




(c) Frequent vs. infrequent words.

# Question 2: Do different words need different context?

- Experiment: test time feed **most recent n tokens only**



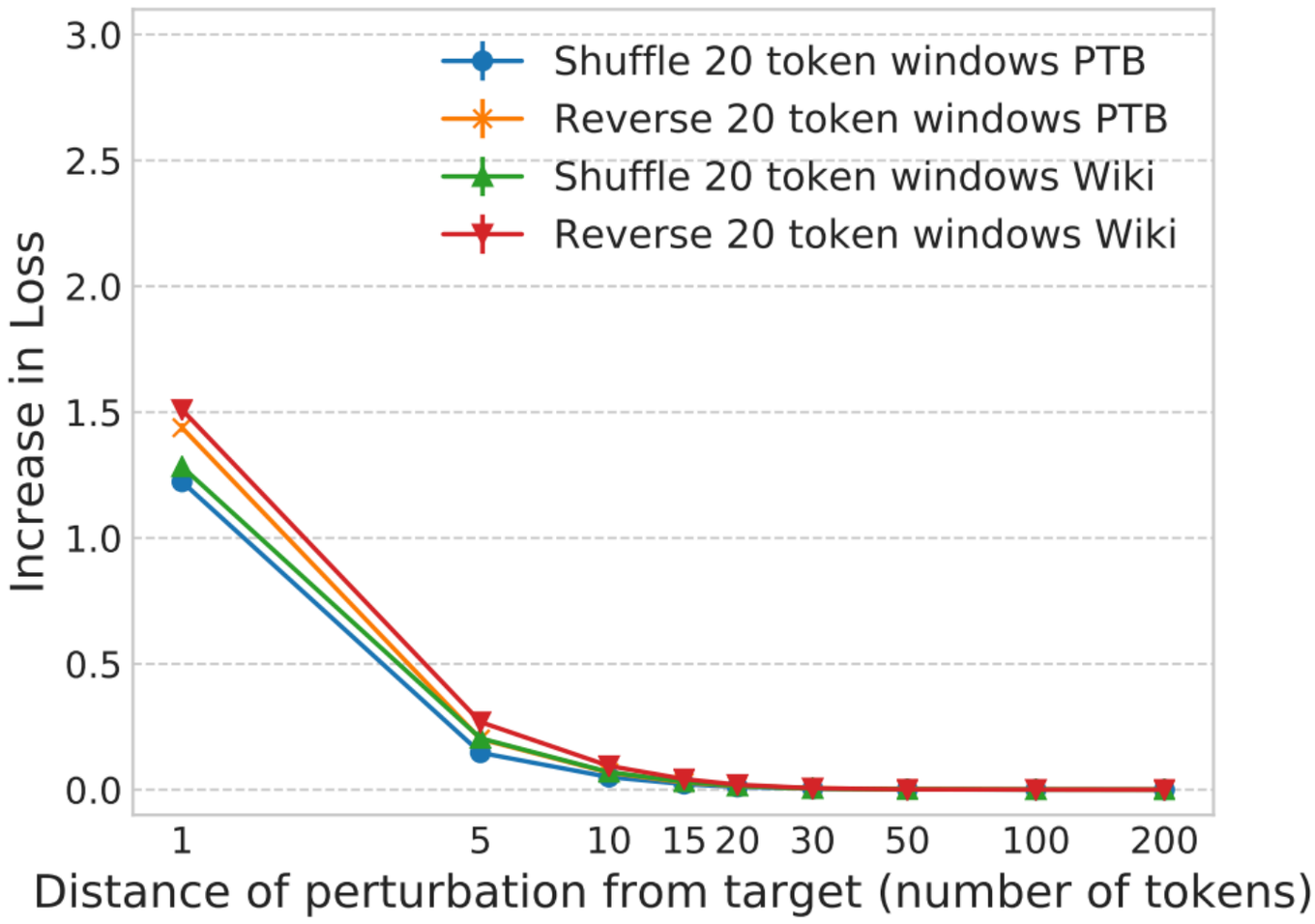

(d) Different parts-of-speech.

# Question 3: Does word order matter?

- Motivation:
  - Previous experiments showed LM required context size of around 200 tokens
  - But does it matter what the context looks like?

# Question 3: Does word order matter?

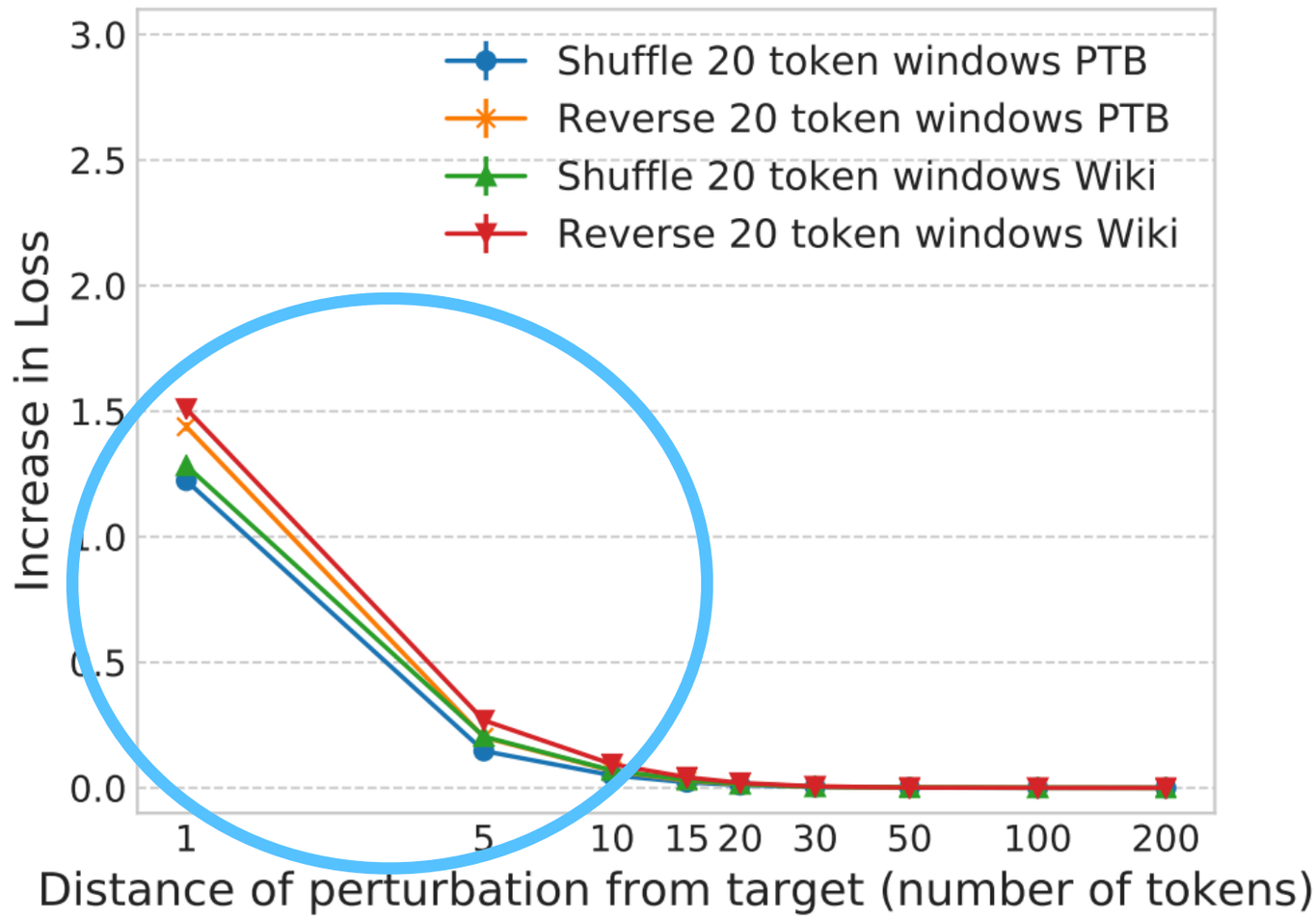
- Experiment: test time **permute substrings** (shuffle, reverse)




(a) Perturb order locally, within 20 tokens of each point.

# Question 3: Does word order matter?

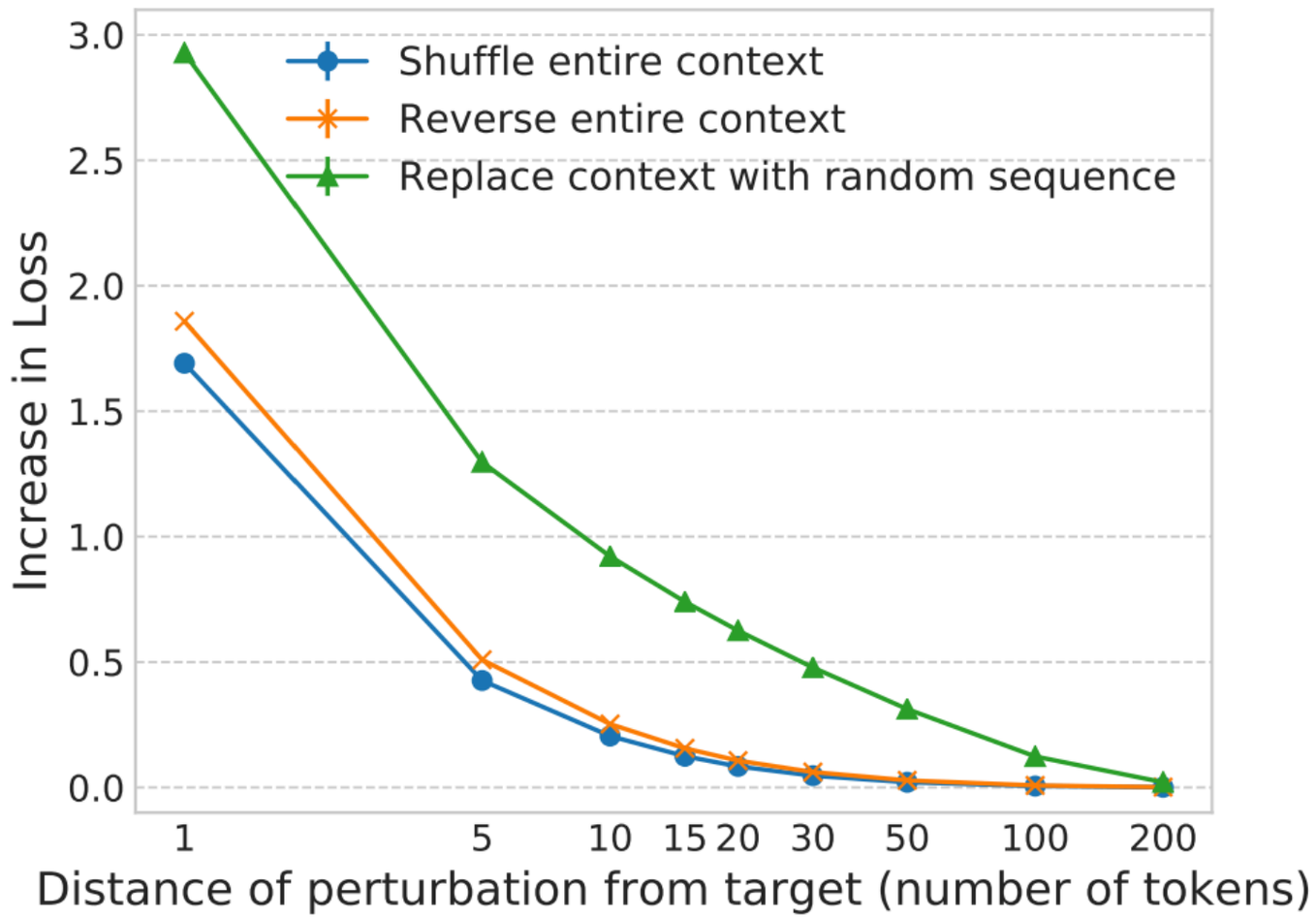
- Experiment: test time **permute substrings** (shuffle, reverse)




(a) Perturb order locally, within 20 tokens of each point.

# Question 3: Does word order matter?

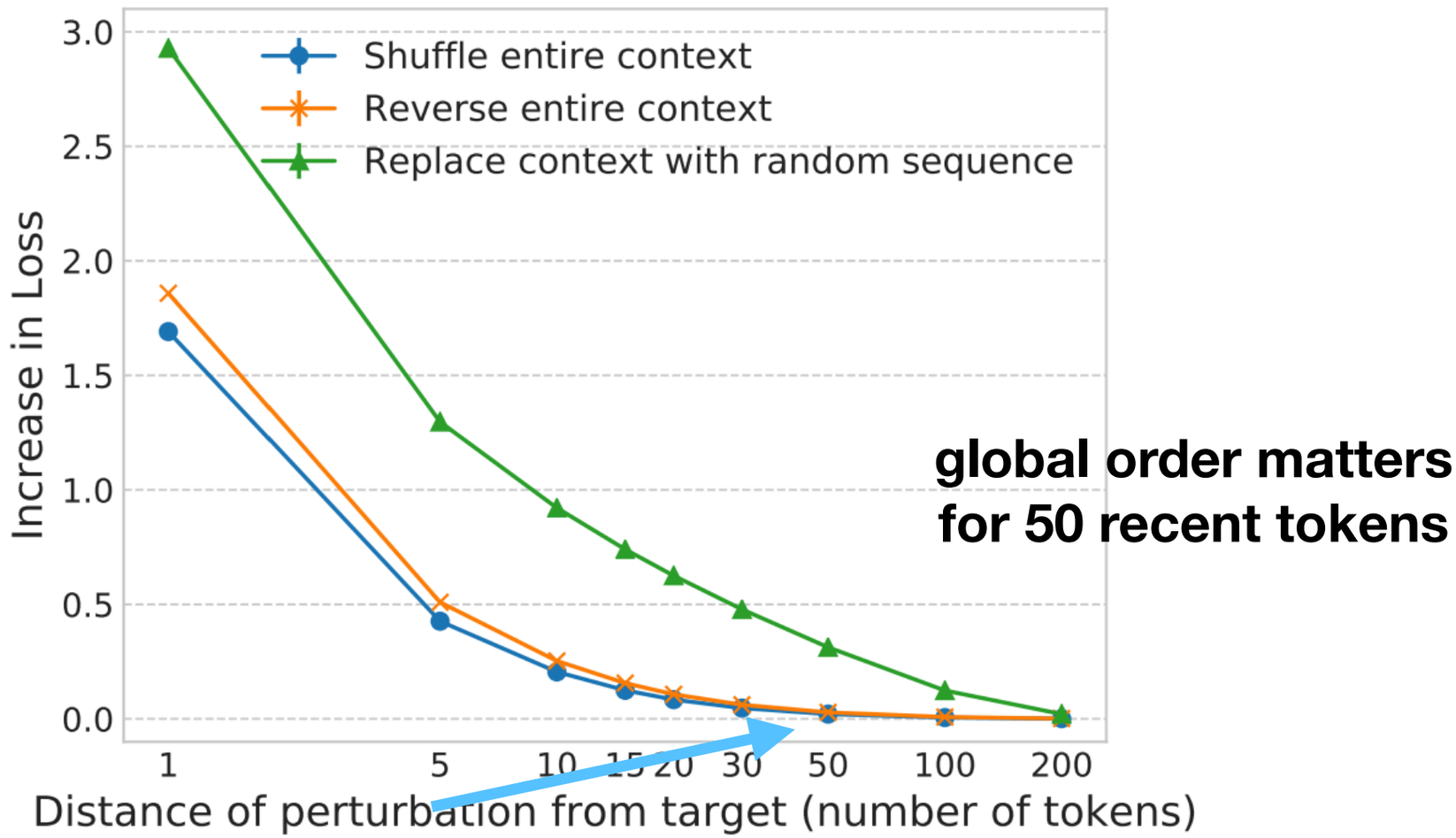
- Experiment: test time **permute substrings** (shuffle, reverse)




(b) Perturb global order, i.e. all tokens in the context before a given point, in Wiki.

# Question 3: Does word order matter?

- Experiment: test time **permute substrings** (shuffle, reverse)



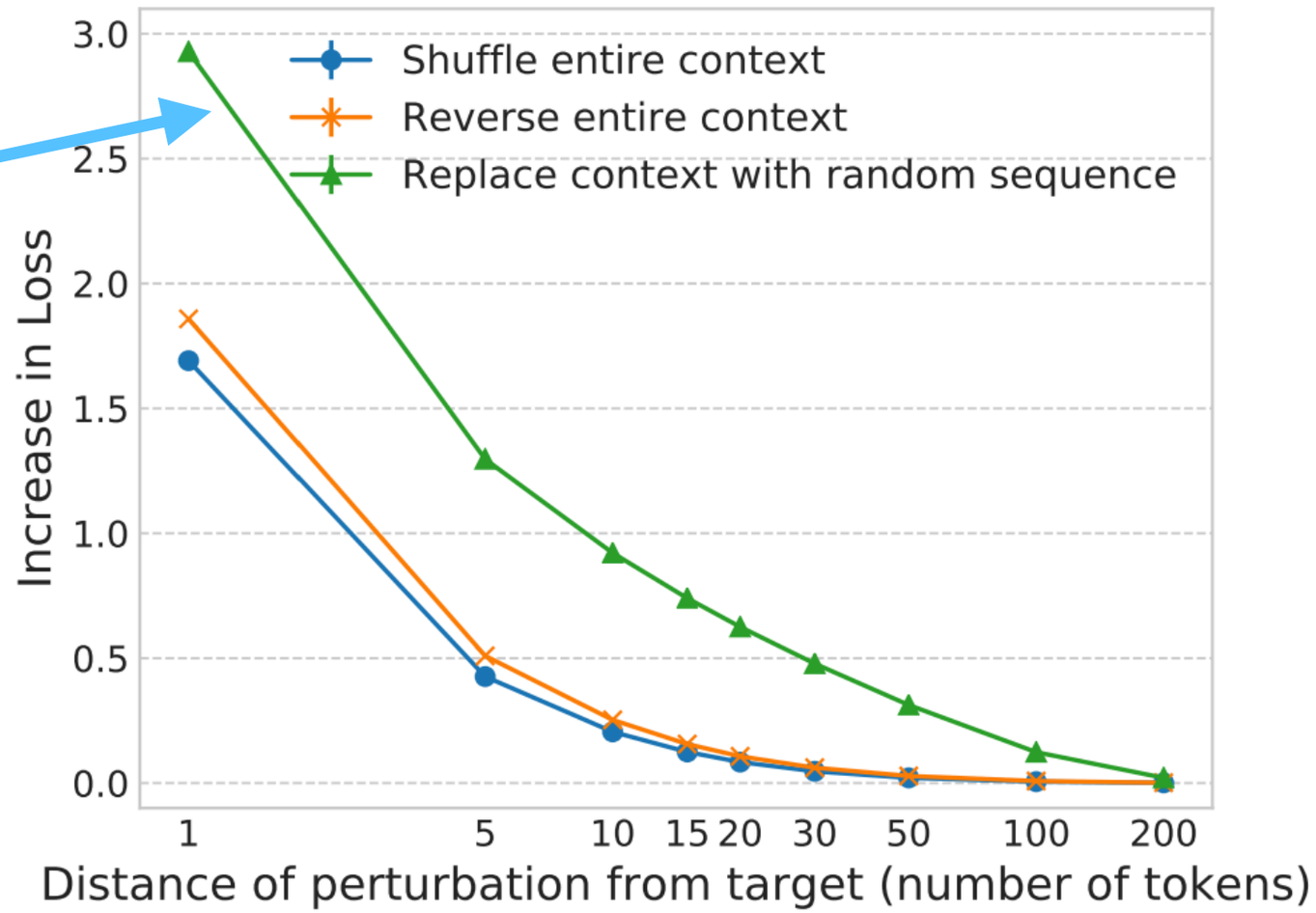

(b) Perturb global order, i.e. all tokens in the context before a given point, in Wiki.

# Question 3: Does word order matter?

- Experiment: test time **permute substrings** (shuffle, reverse)

**far away tokens can't be random**

**order doesn't matter**



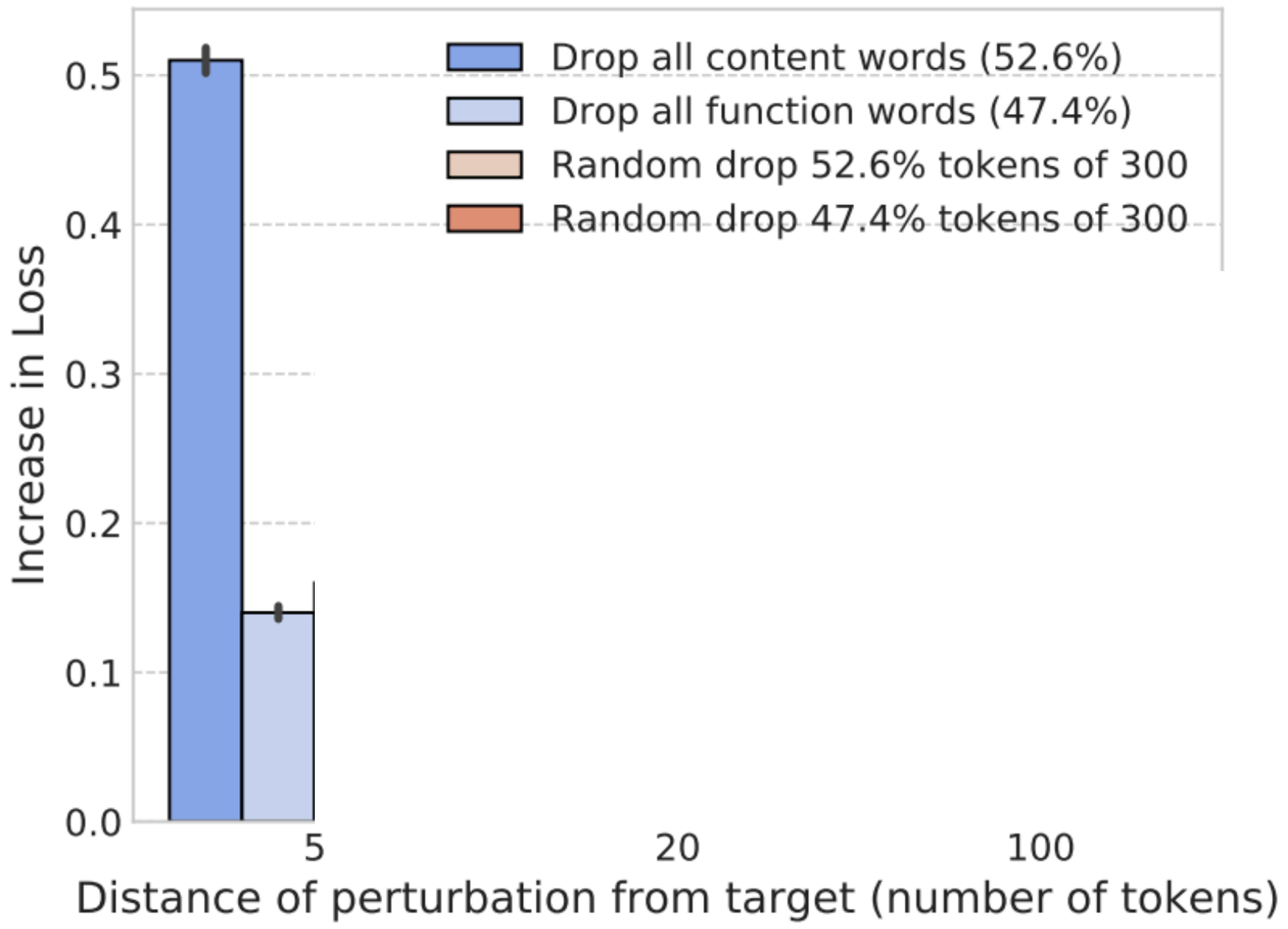

(b) Perturb global order, i.e. all tokens in the context before a given point, in Wiki.

# Question 4: What kind of words are important?

- Motivation:
  - 200 words is a lot to pay attention to
  - Are all words in the context equally important?

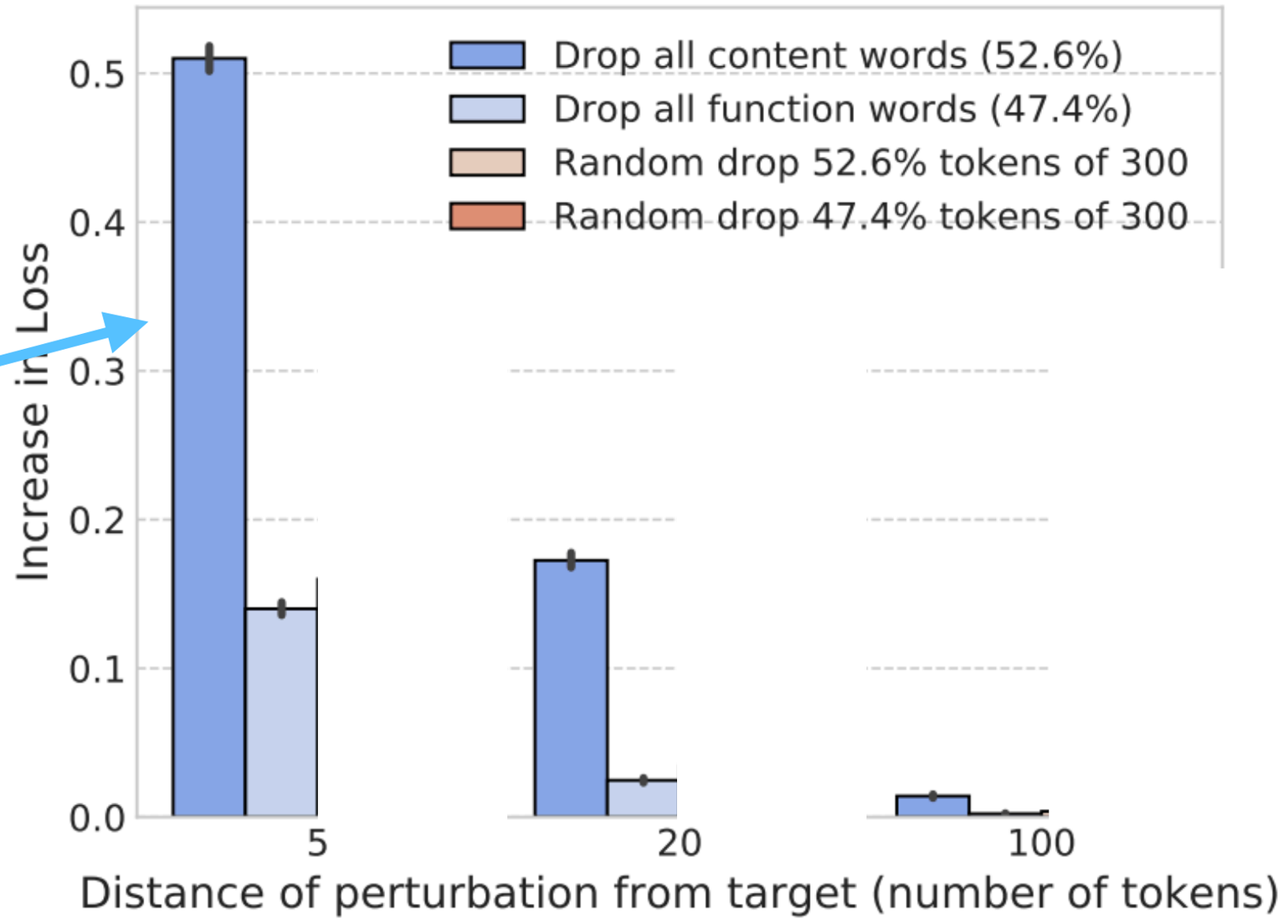
# Question 4: What kind of words are important?

- Experiment: test time **drop certain words**



# Question 4: What kind of words are important?

- Experiment: test time **drop certain words**

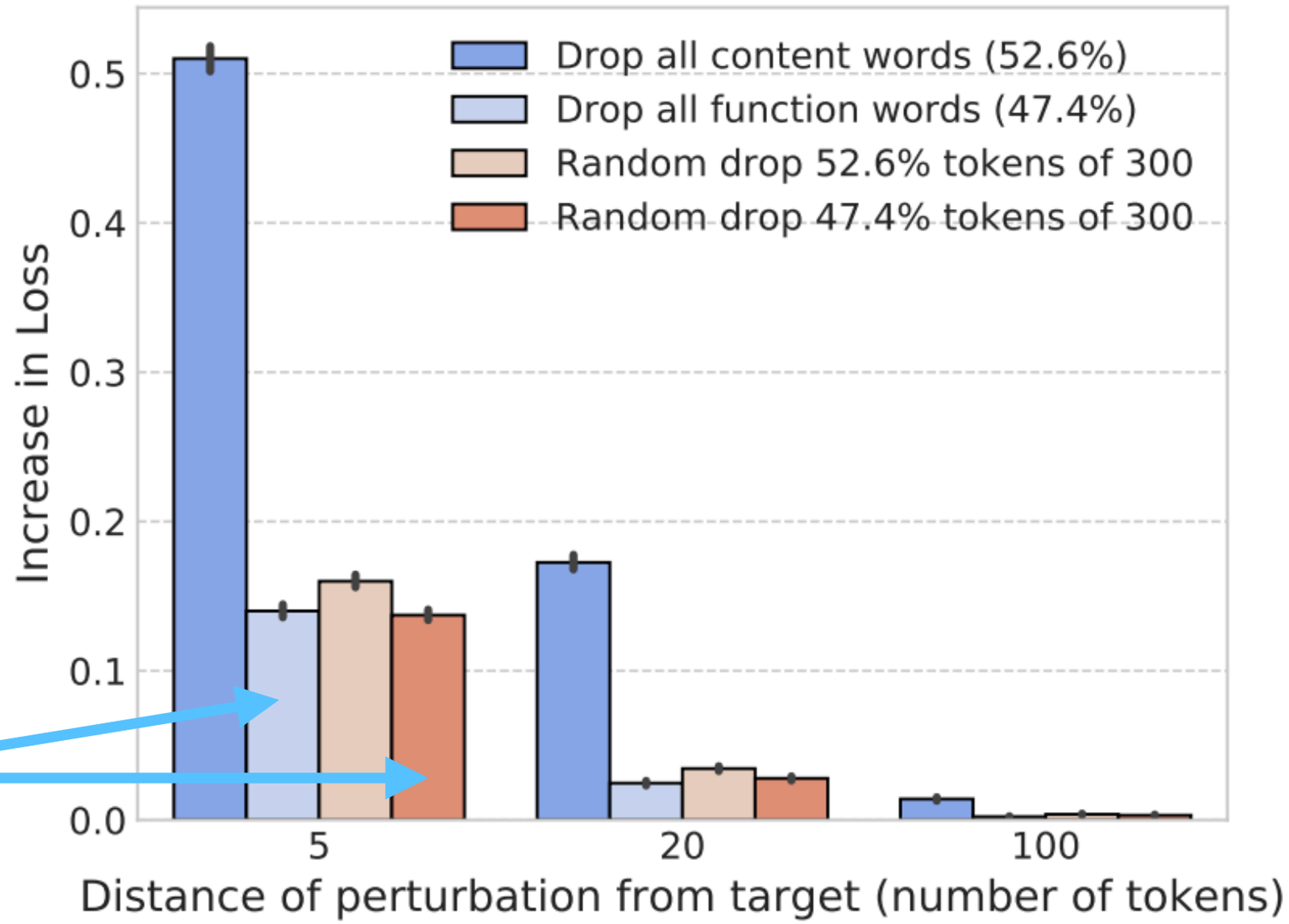



**content words matter more**

**particularly local**

# Question 4: What kind of words are important?

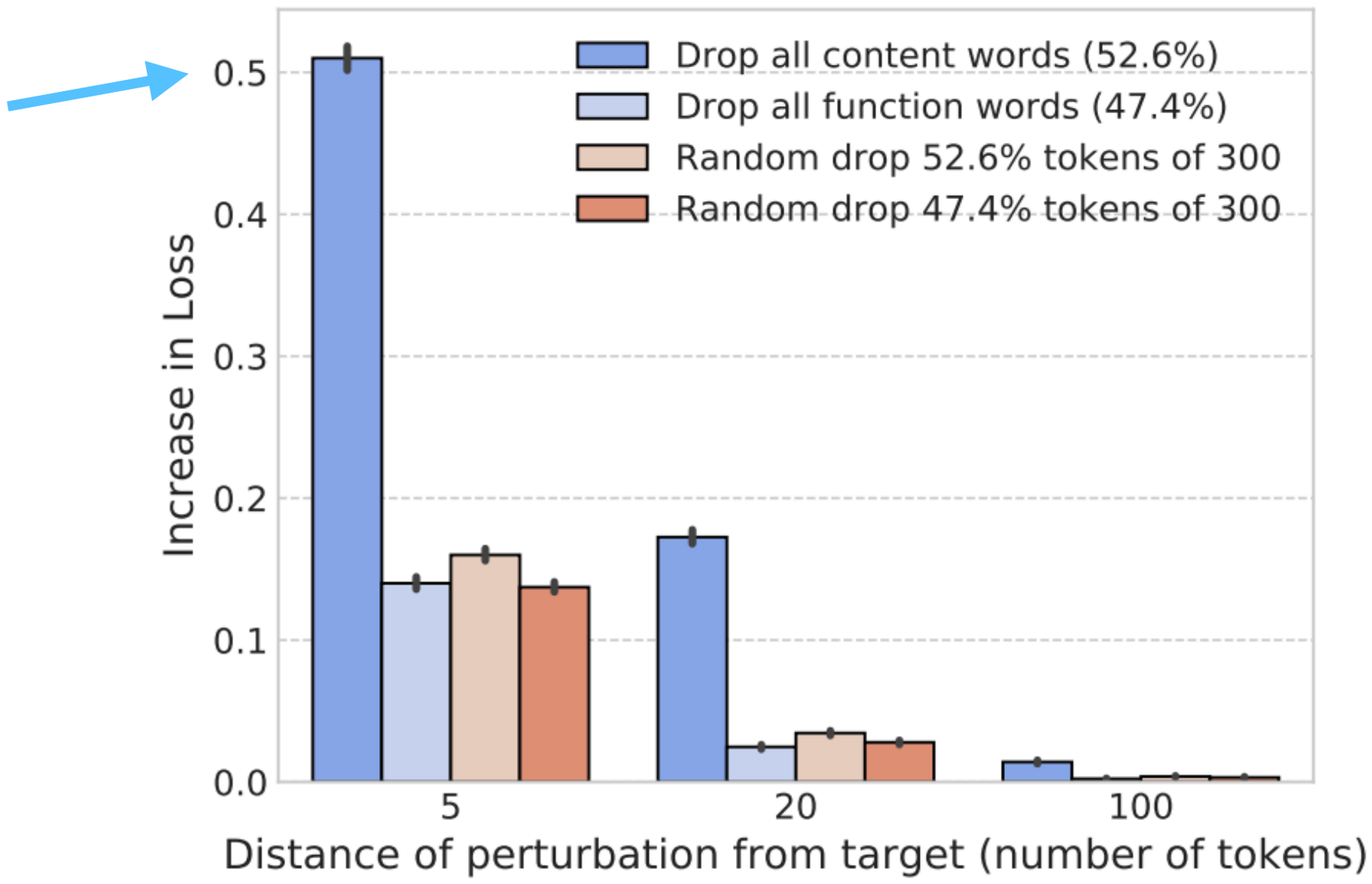
- Experiment: test time **drop certain words**




**function words == random**

# Question 4: What kind of words are important?

- Experiment: test time **drop certain words**



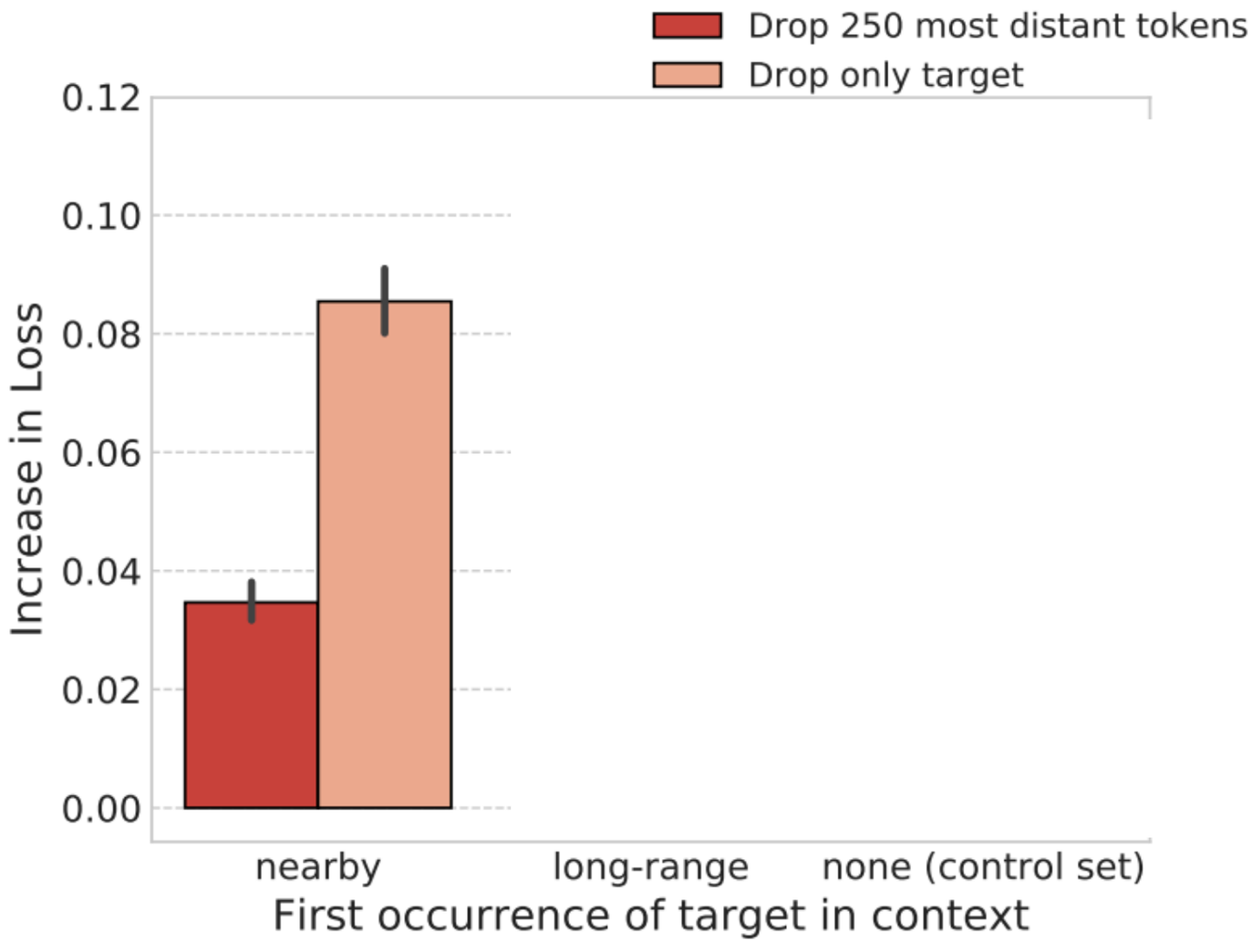
# Question 5: Is cache/copy important?

- Motivation:
  - Information in very distant context is useful, but can be hard to recall/remember/use
  - Copy mechanisms like caching and attention work well
  - Can LSTMs learn to copy words?

# Question 5: Is cache/copy important?

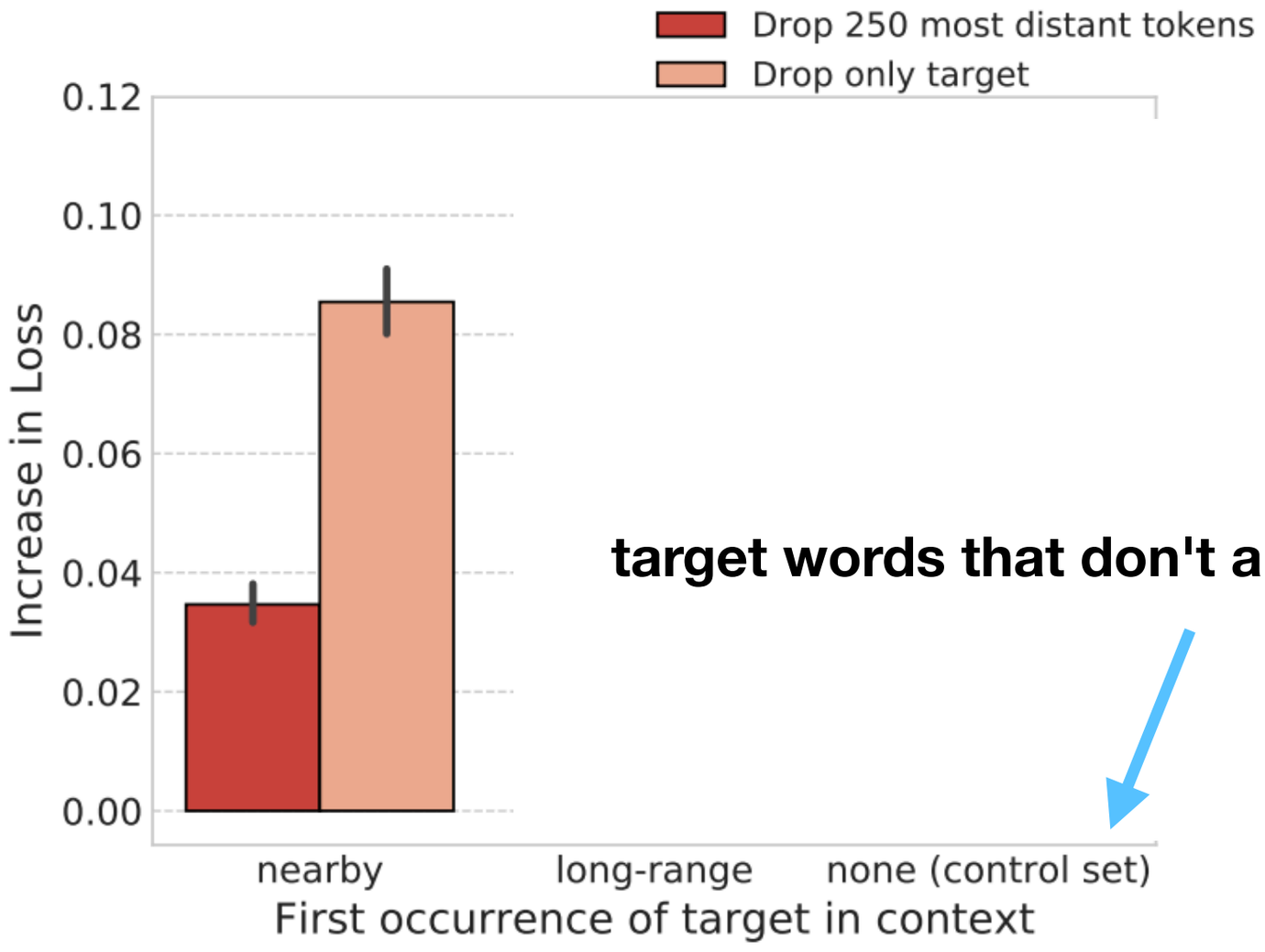
- Experimental Setup:
  - 2 types of context:
    - **nearby**: within 50 most recent tokens
    - **long range**: beyond 50 tokens
  - 3 types of target words:
    - **nearby copy**
    - **far copy**
    - **no copy**

# Question 5: Is cache/copy important?



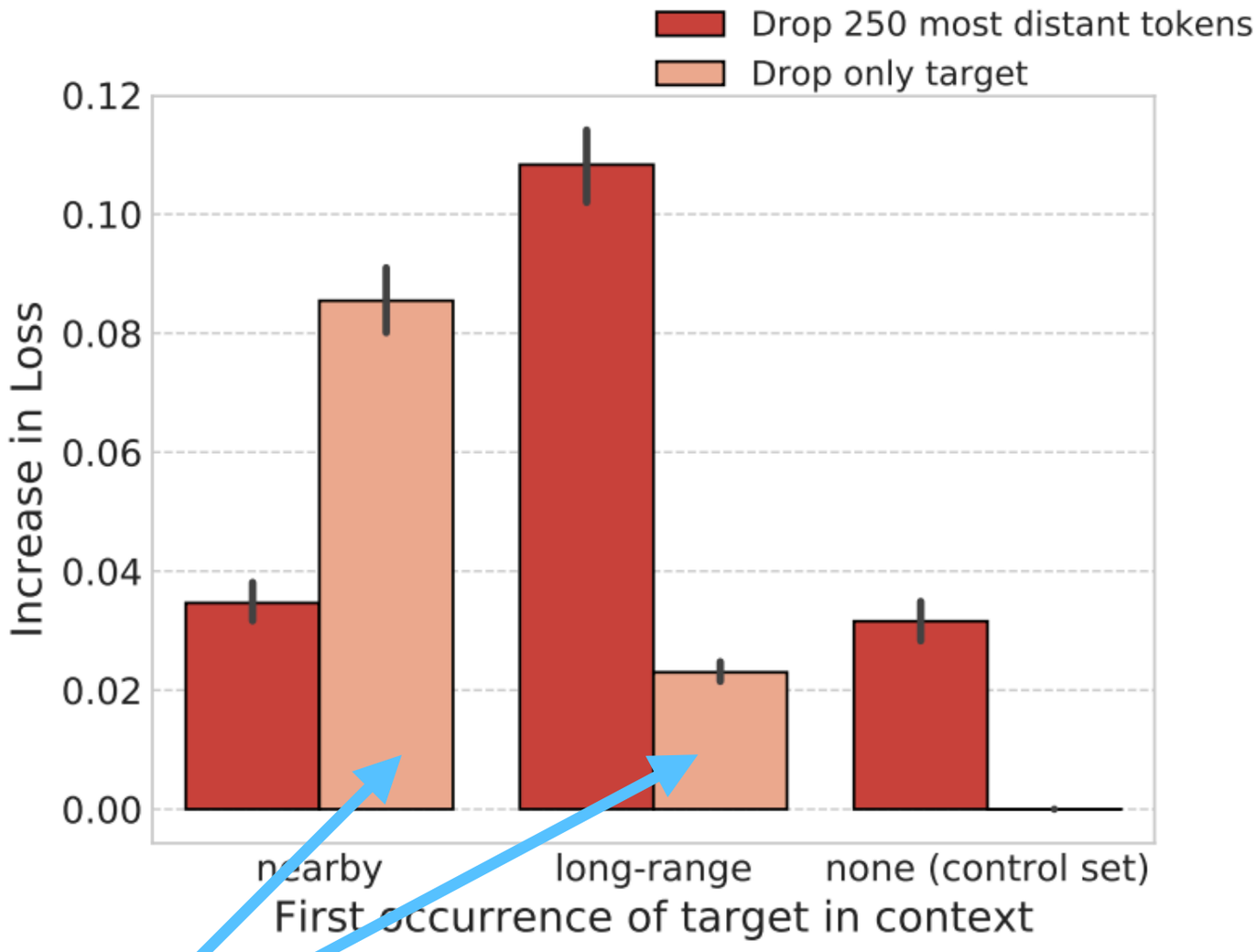

(a) Dropping tokens

# Question 5: Is cache/copy important?




(a) Dropping tokens

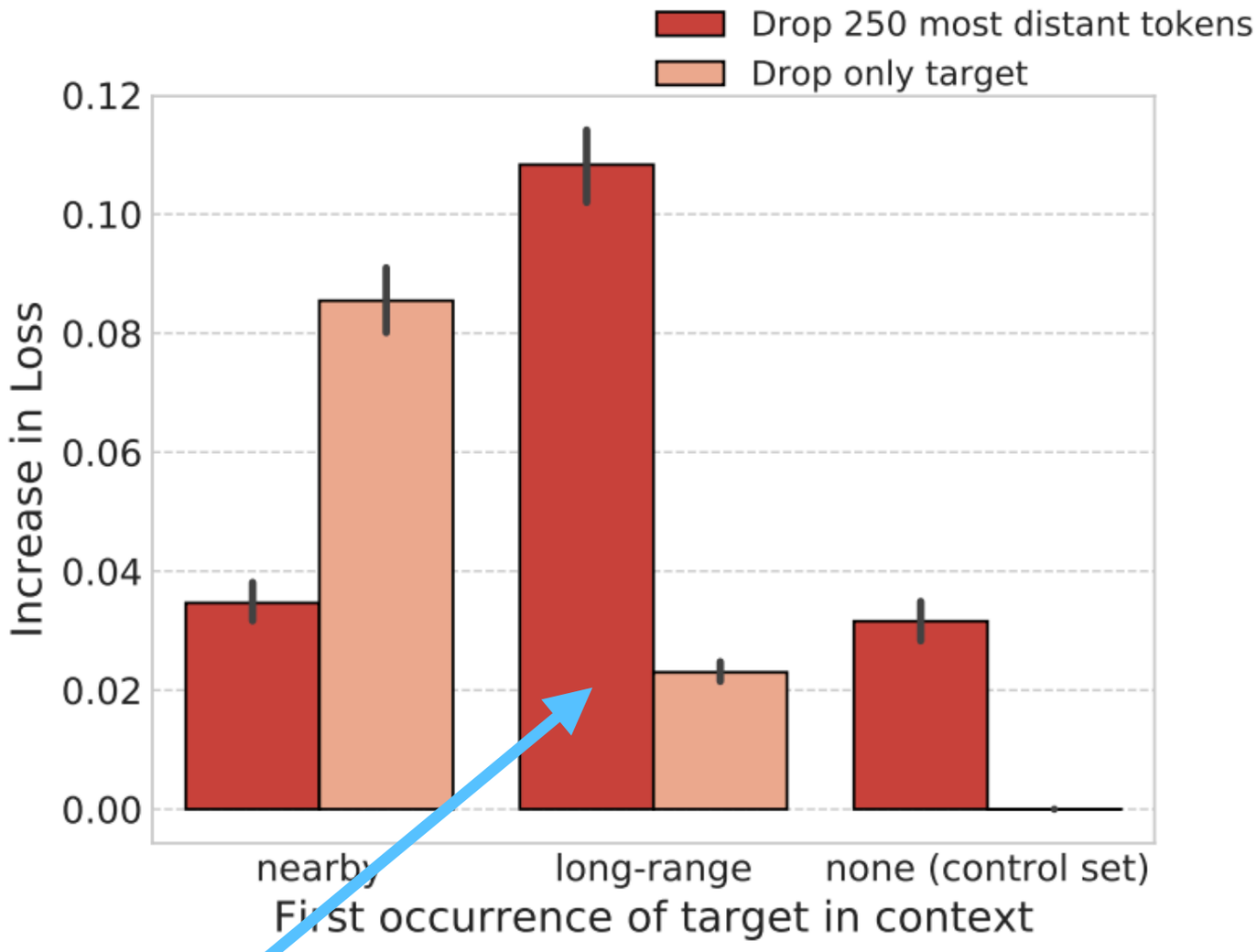
# Question 5: Is cache/copy important?




**copy nearby words**

(a) Dropping tokens

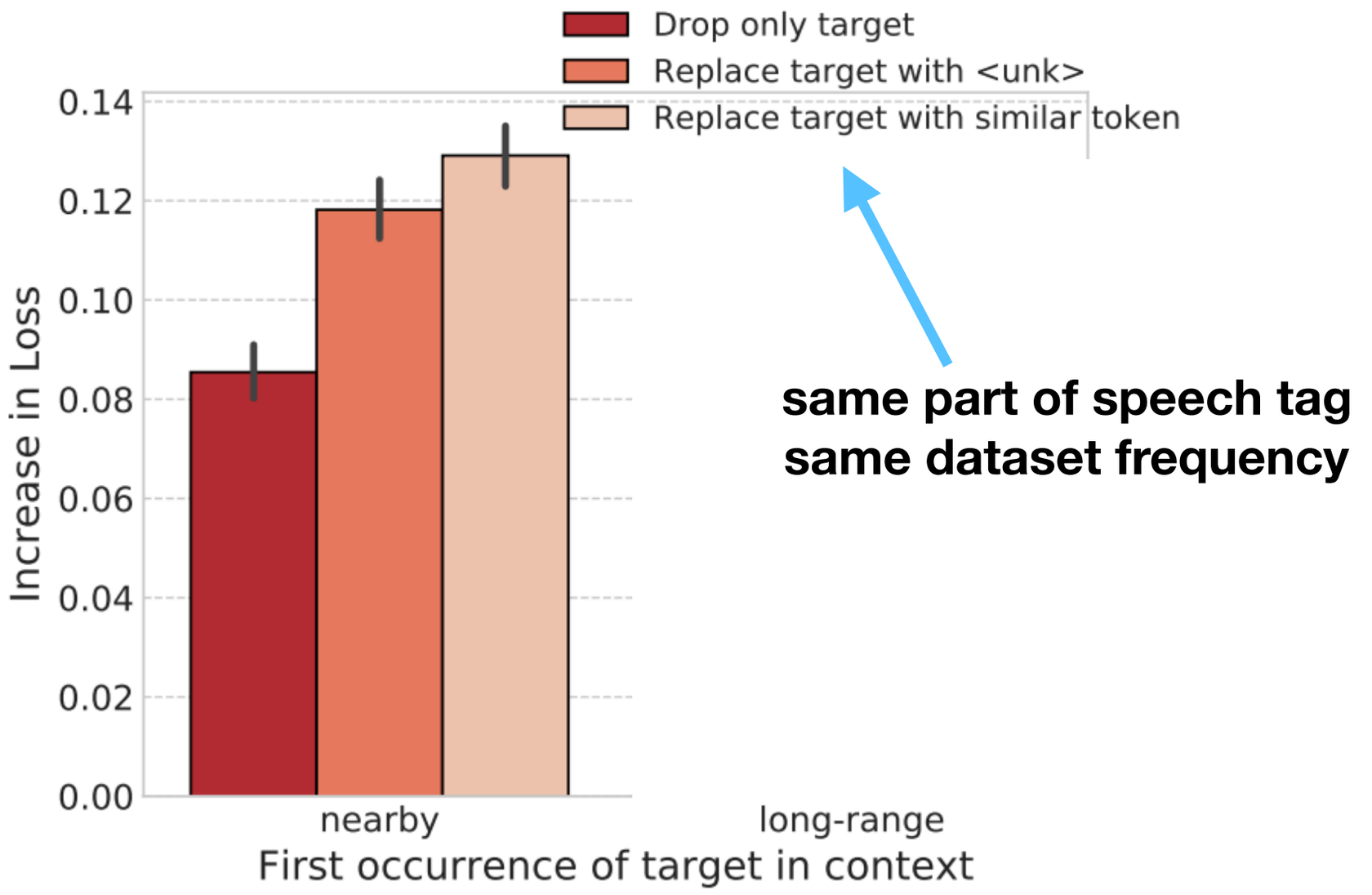
# Question 5: Is cache/copy important?




(a) Dropping tokens

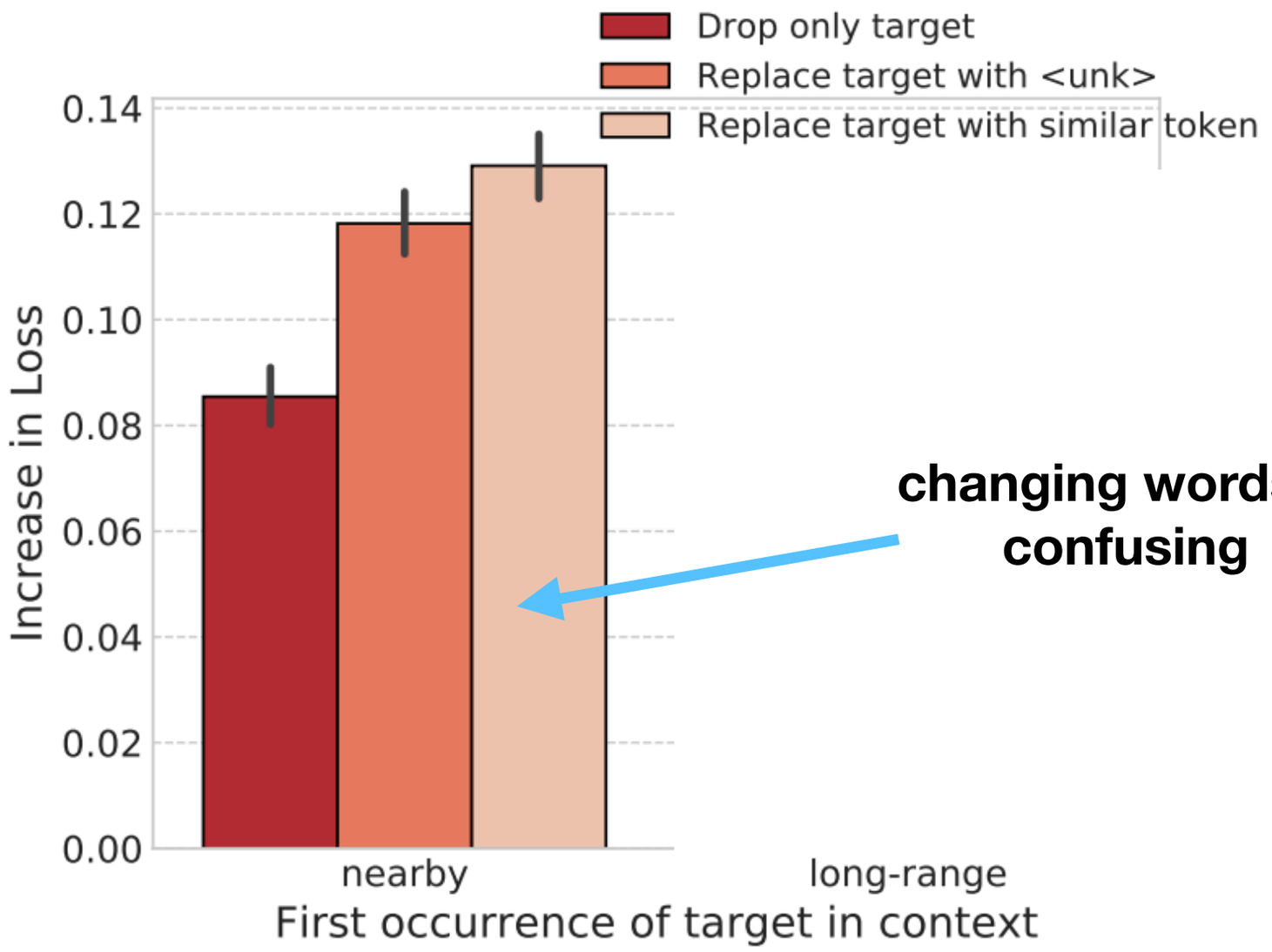
**long range targets need distant context**

# Question 5: Is cache/copy important?



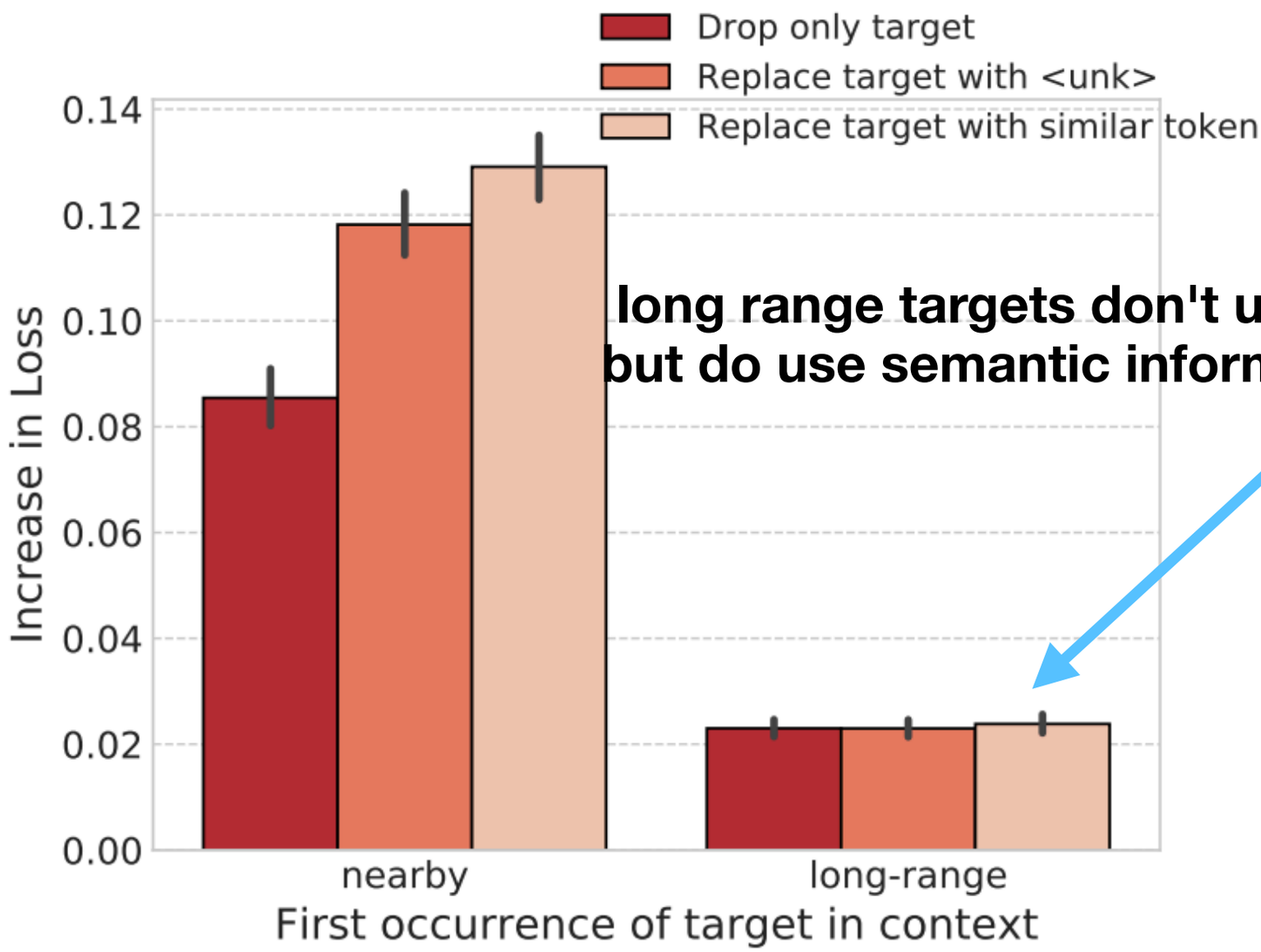

(b) Perturbing occurrences of target word in context.

# Question 5: Is cache/copy important?




(b) Perturbing occurrences of target word in context.

# Question 5: Is cache/copy important?




(b) Perturbing occurrences of target word in context.

# Caching

- How does Cache work?
  - records hidden state for a tilmestep *t*
  - compute distribution over states in the cache

$$P_{\text{cache}}(w_t|w_{t-1},\ldots,w_1;h_t,\ldots,h_1)$$

# Caching

- How does Cache work?
  - records hidden state for a tilmestep *t*
  - compute distribution over states in the cache
  - interpolate with standard vocabulary softmax
  - cache can upweight certain words in the past

# Caching

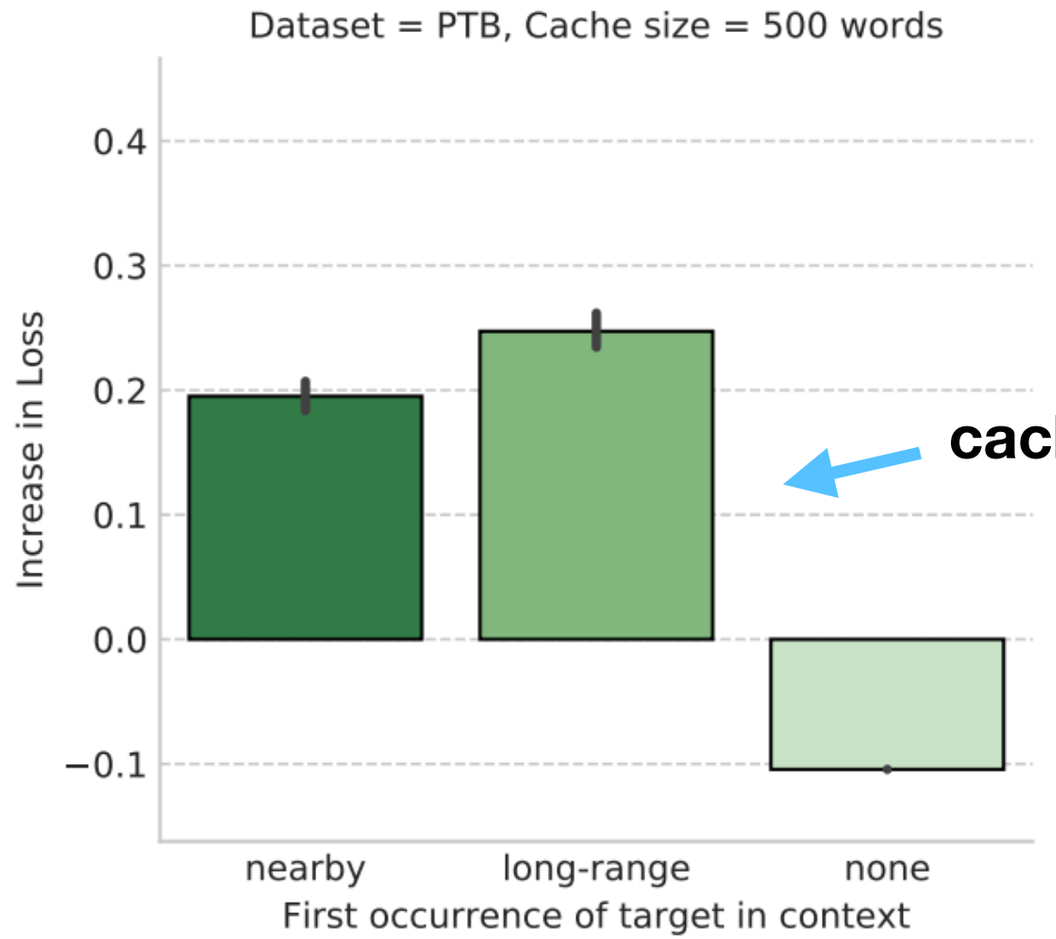
- How does Cache work?
  - records hidden state for a tilmestep *t*
  - compute distribution over states in the cache
  - interpolate with standard vocabulary softmax
  - cache can upweight certain words in the past

# Caching

- (as an aside, caching tends to cache many many things)

# Question 6: How does Cache Affect Copy?

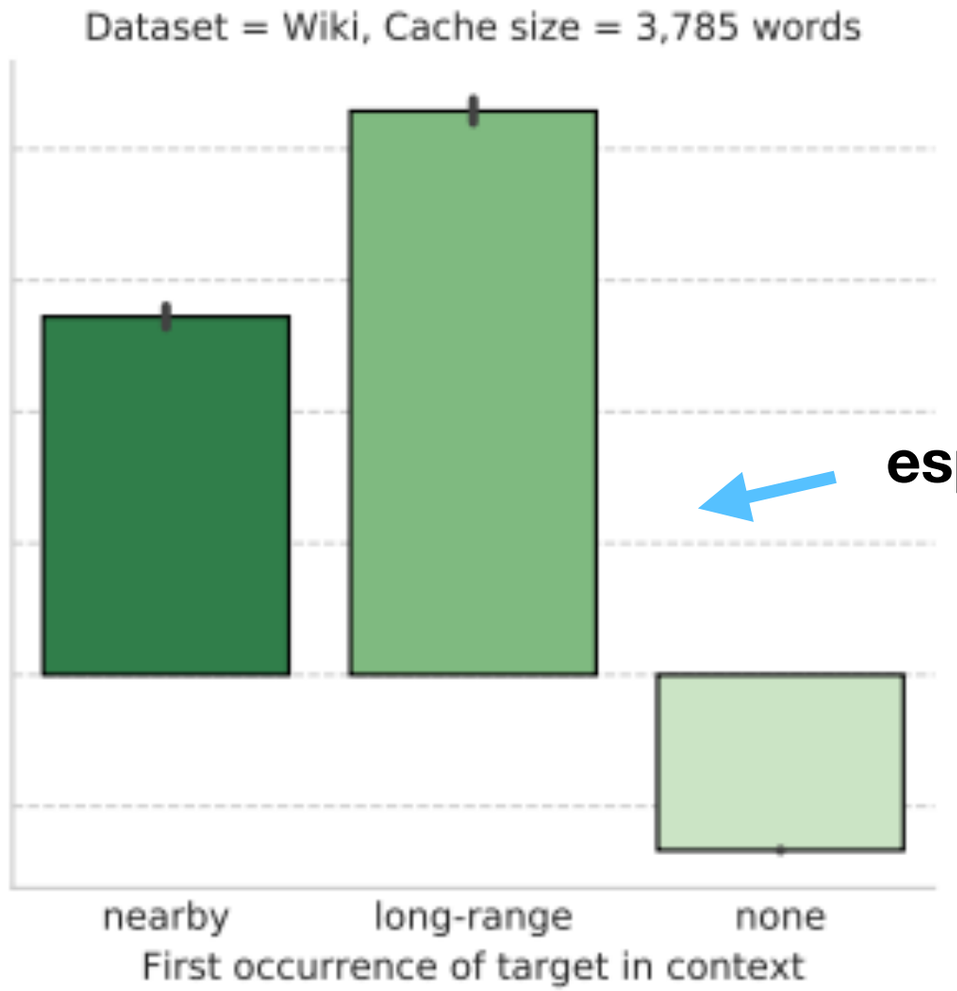
- Experiment: evaluate LM with and without cache, measure perplexity difference for copy near, copy far, copy none




**cache helps copy far words**

# Question 6: How does Cache Affect Copy?

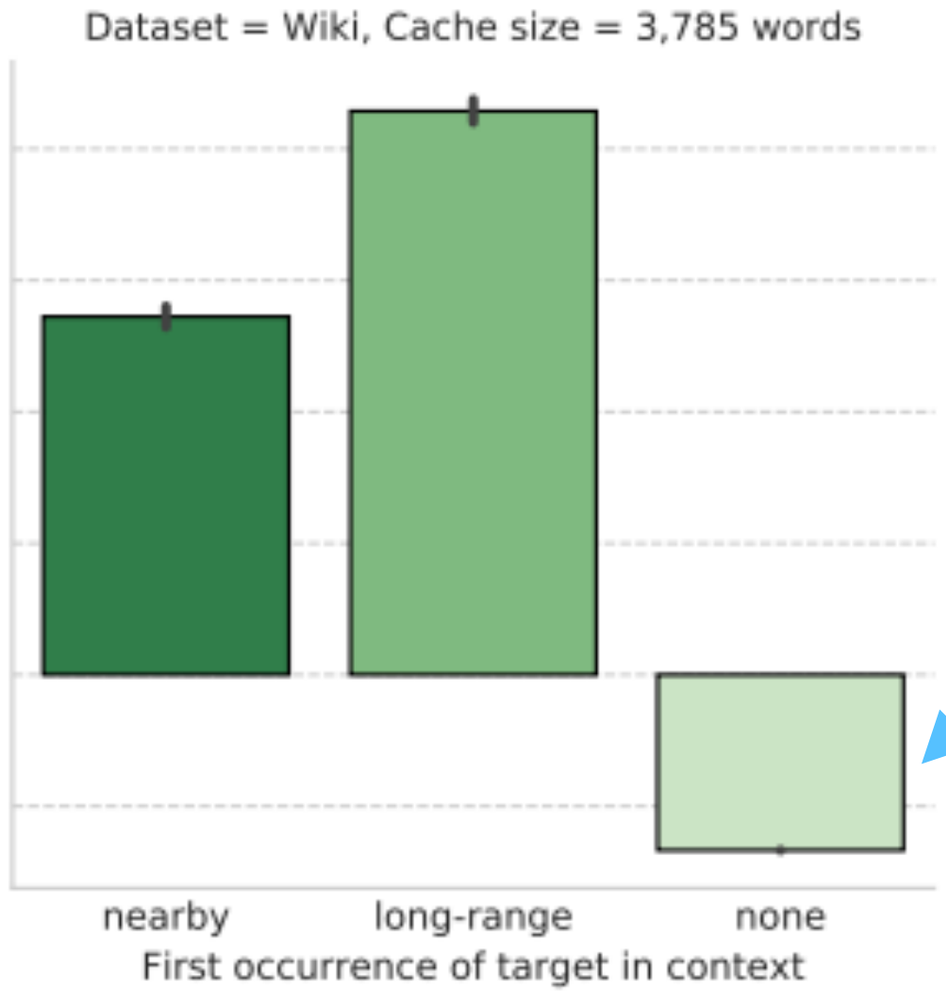
- Experiment: evaluate LM with and without cache, measure perplexity difference for copy near, copy far, copy none




**especially for wikipedia**

# Question 6: How does Cache Affect Copy?

- Experiment: evaluate LM with and without cache, measure perplexity difference for copy near, copy far, copy none



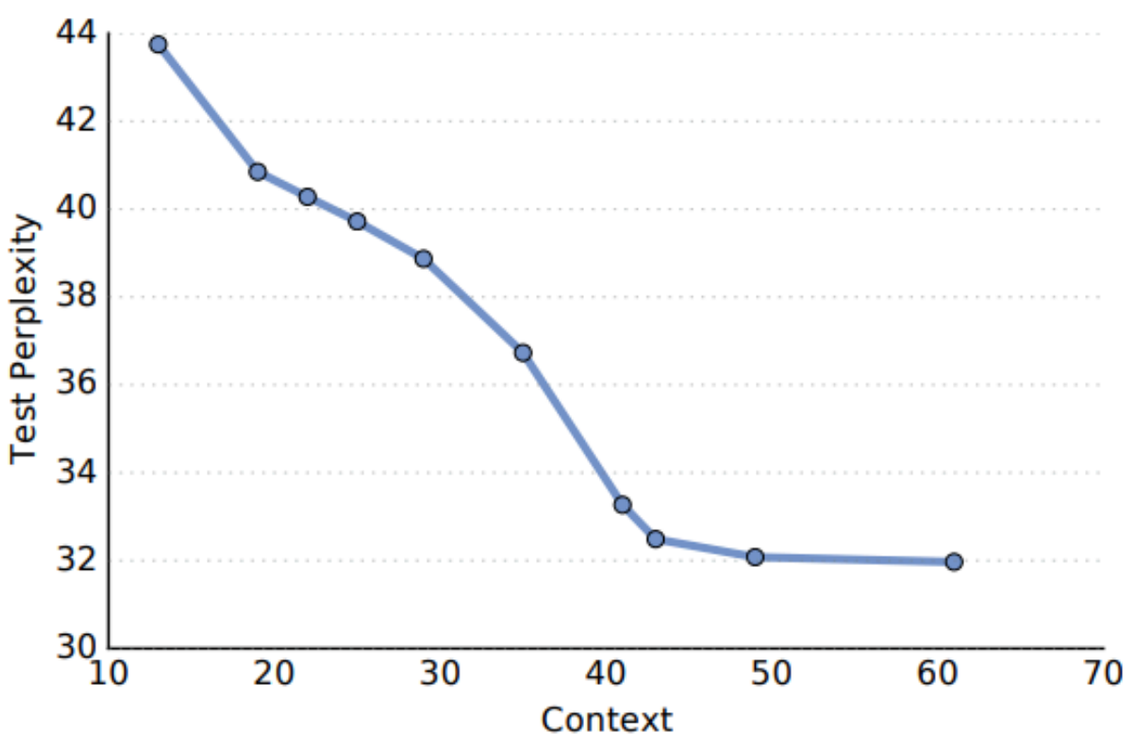

**cache also hurts**

# Why does the Cache Hurt performance for none?

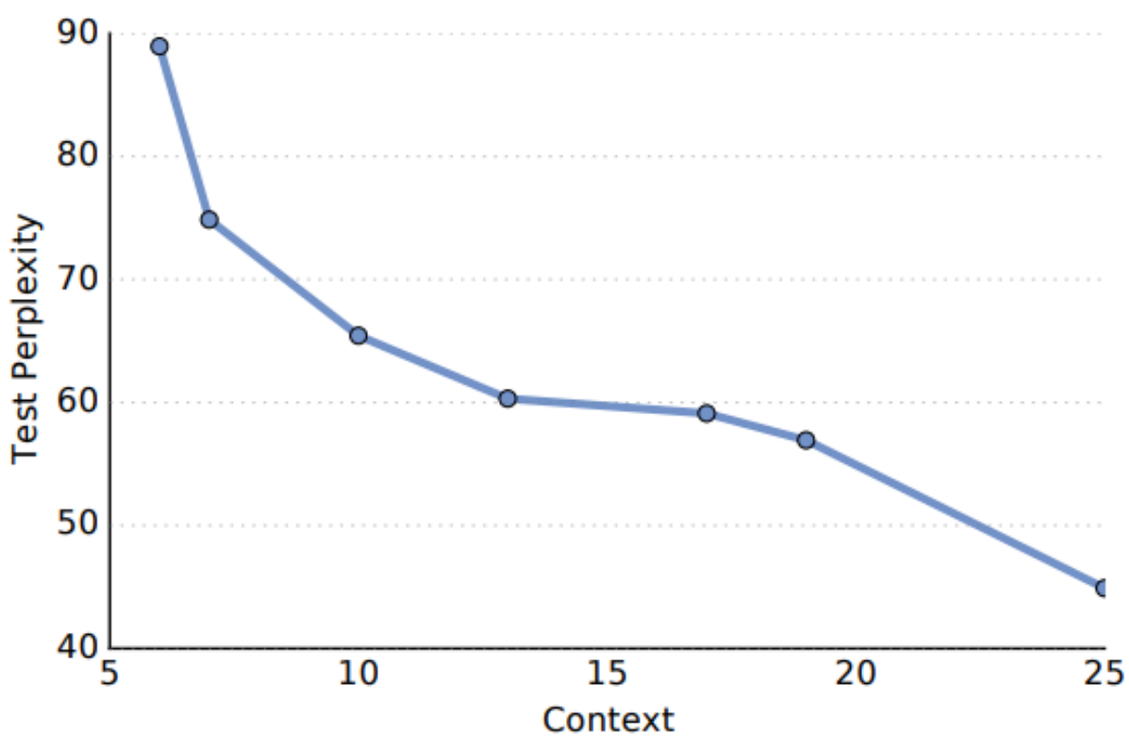
- If there is no information in the cache (e.g. cannot copy):
  - cache probability distribution is flat
  - when interpolate with model probability, flattens it as well

# Some Questions

- Impact of Dataset, Model?
- Context v. Capacity?




google billion words




wikitext 103