Large Language Models: Evaluation and Cognitive Applications

Machine Learning Summer School
Arequipa 2025

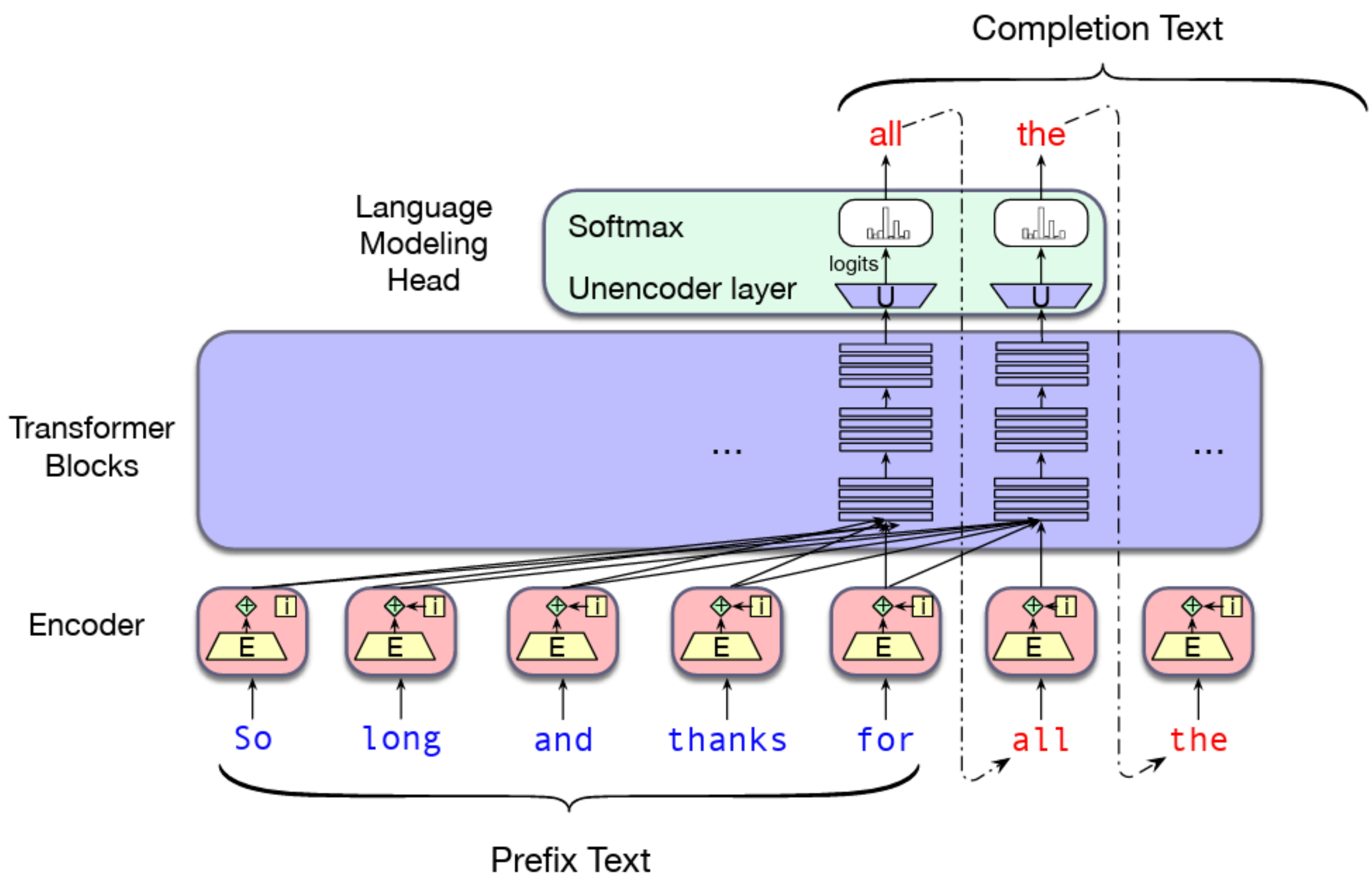
Tal Linzen
New York University and Google

Today

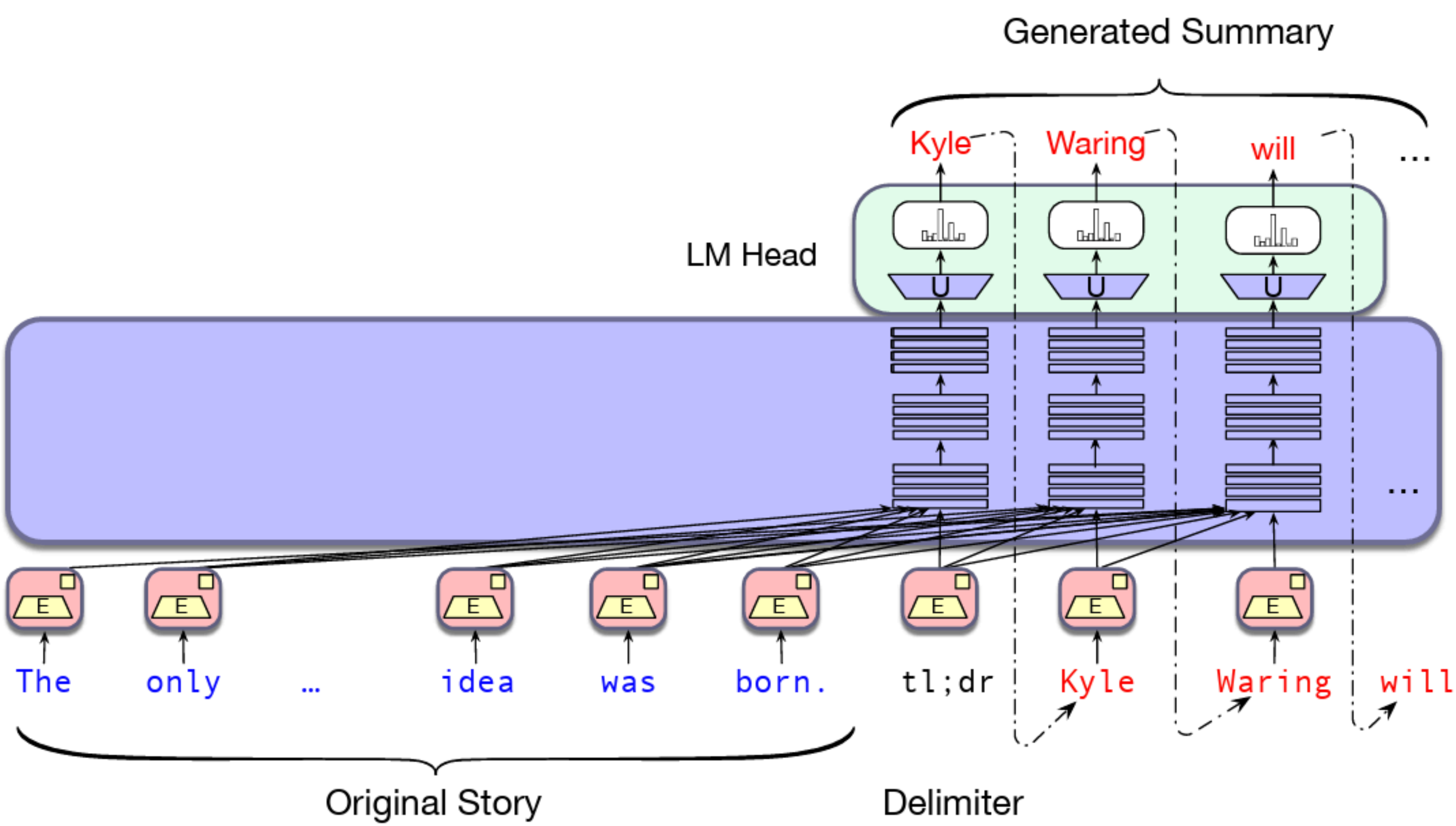
- Review: The LLM paradigm
- Review: The transformer architecture
- The potential of language models as cognitive models
- When do we want human-like language models?
- Improving data efficiency with formal language pretraining
- (Word prediction in LLMs and humans)

Review: the LLM paradigm for NLP (and beyond)

The LLM paradigm: pre-training



The LLM paradigm: aspirationally performing tasks zero-shot or few-shot



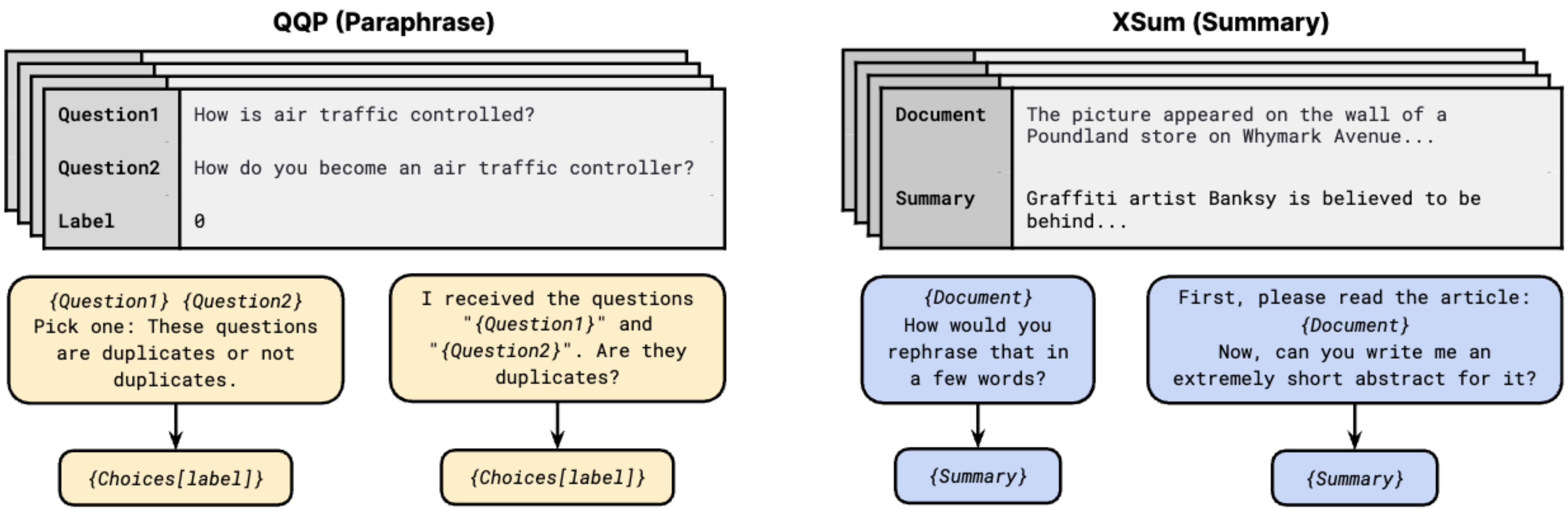
The prompting paradigm

- No formal separation between encoder and decoder: we provide the input through teacher forcing (this is called the “**prompt**”)

- E.g., to perform summarization, we append “TL;DR:” to the input (!), then generate (e.g. through greedy decoding; Radford et al 2019)

- “In-context learning”: we provide a few examples of the task in the prompt

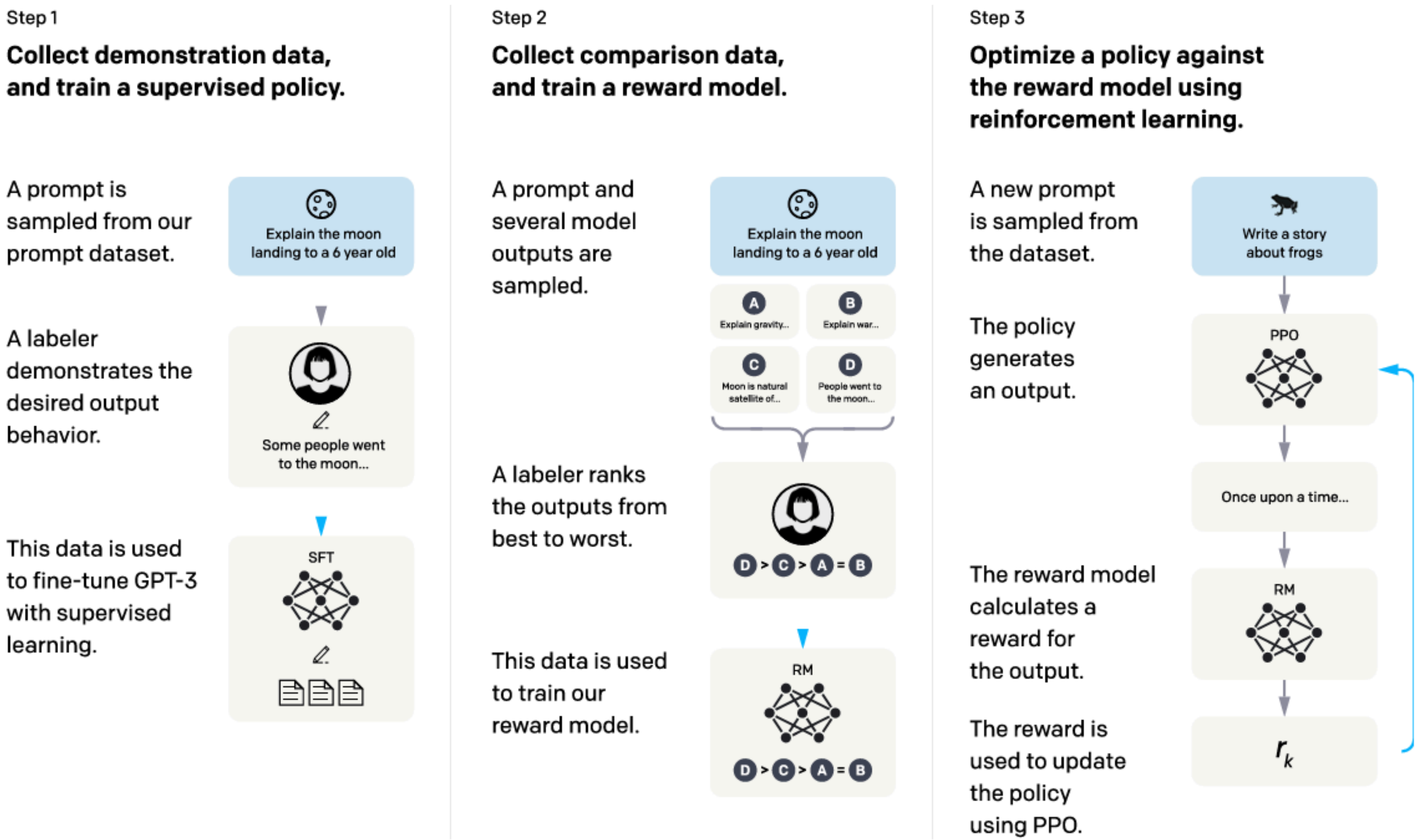
Post-training: Instruction tuning / supervised fine-tuning



- We continue training the LLM using the same objective (cross-entropy loss) on the expected outputs

(Sanh et al., 2022)

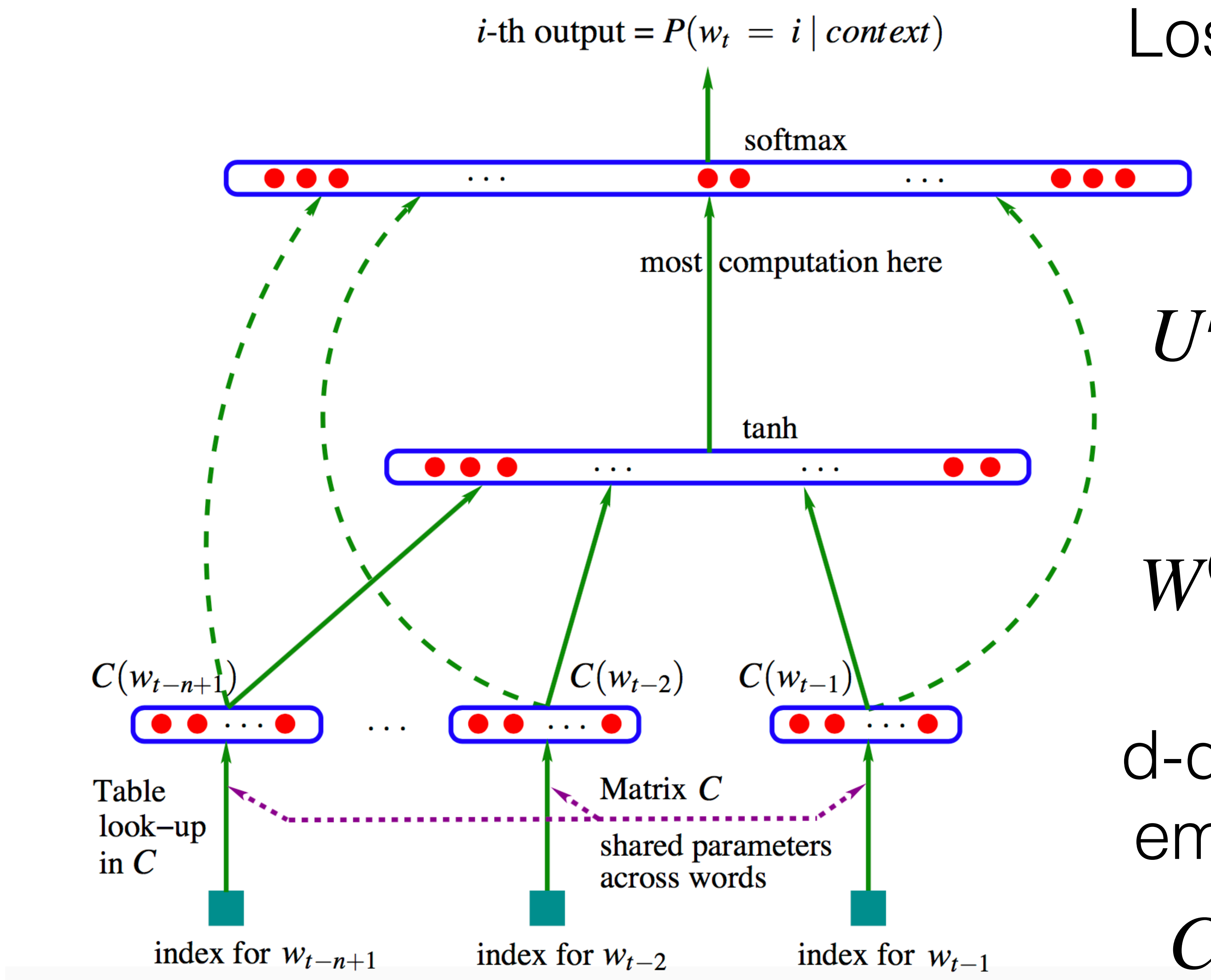
Post-training: Learning from preferences



(Ouyang et al., 2022)

Review: the transformer architecture

Background: Bengio et al. (2003)



Loss: $-log\hat{P}(w)$

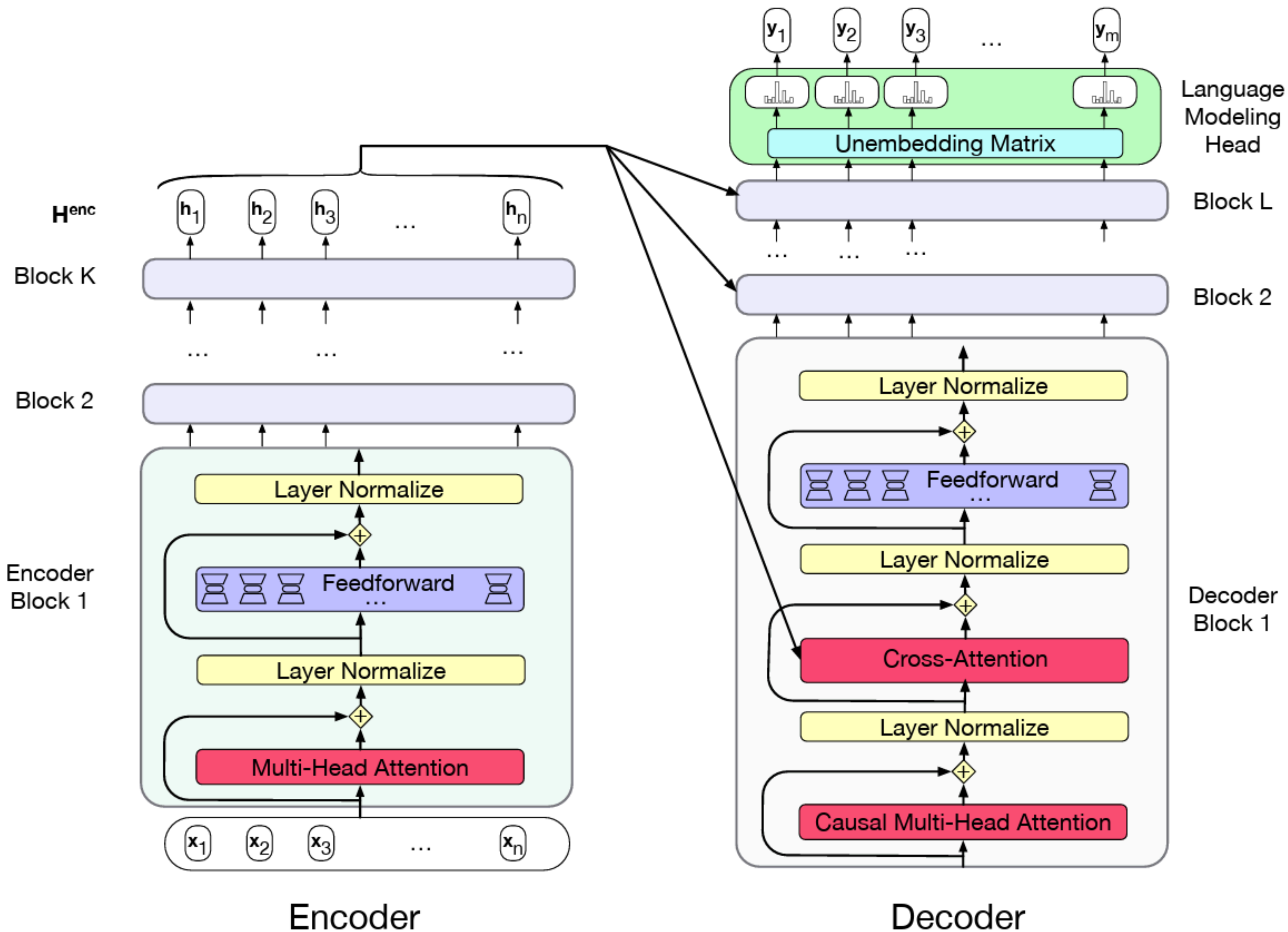
$U^{h \times |V|}$

$W^{(n-1)d \times h}$

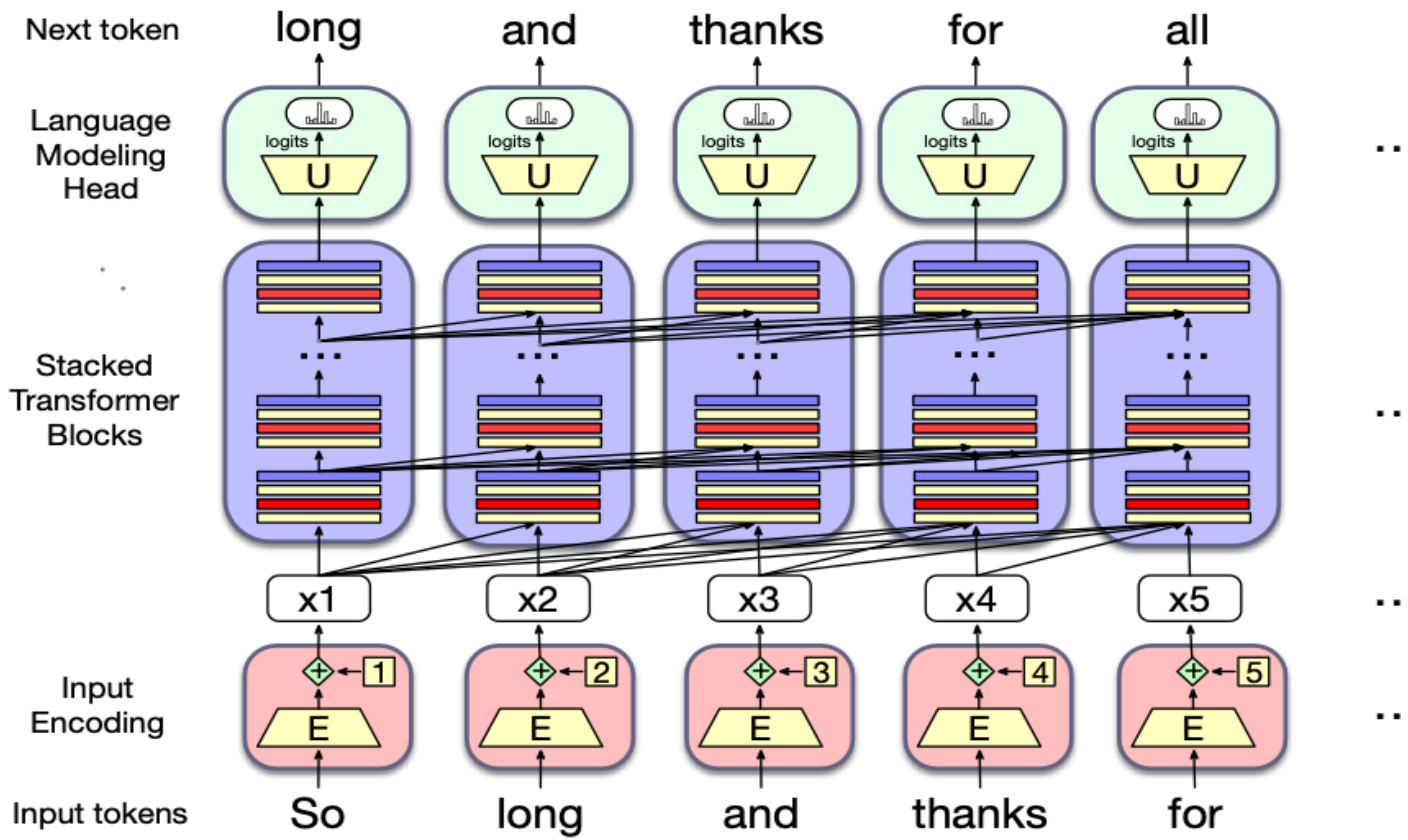
d-dimensional embeddings:

$C_1, \ldots, C_{|V|}$

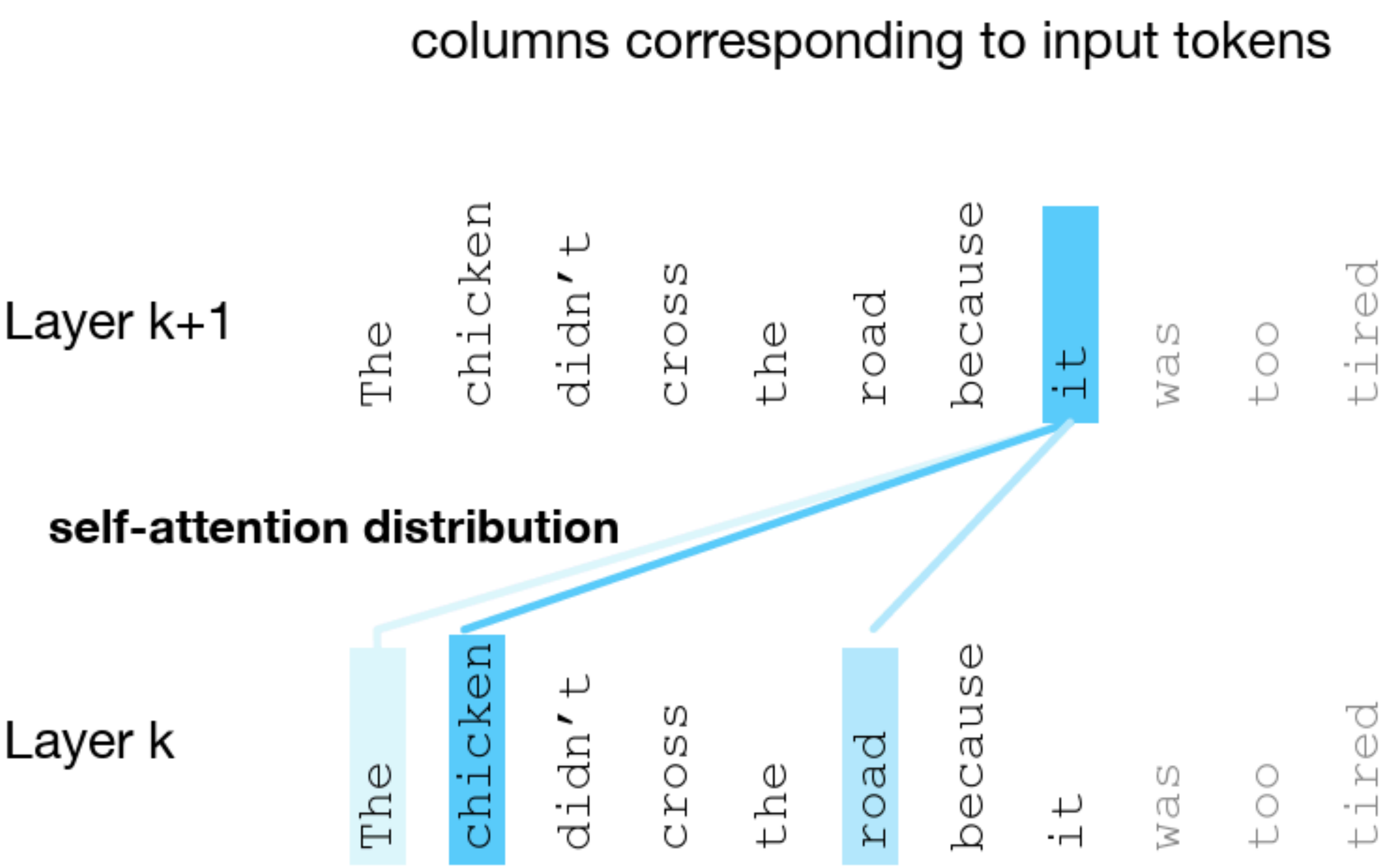
Transformer



Transformers as language models



Attention


columns corresponding to input tokens
Layer k+1
The
chicken
didn't
cross
the
road
because
it
was
too
tired
self-attention distribution
Layer k
The
chicken
didn't
cross
the
road
because
it
was
too
tired

Attention

Attention function

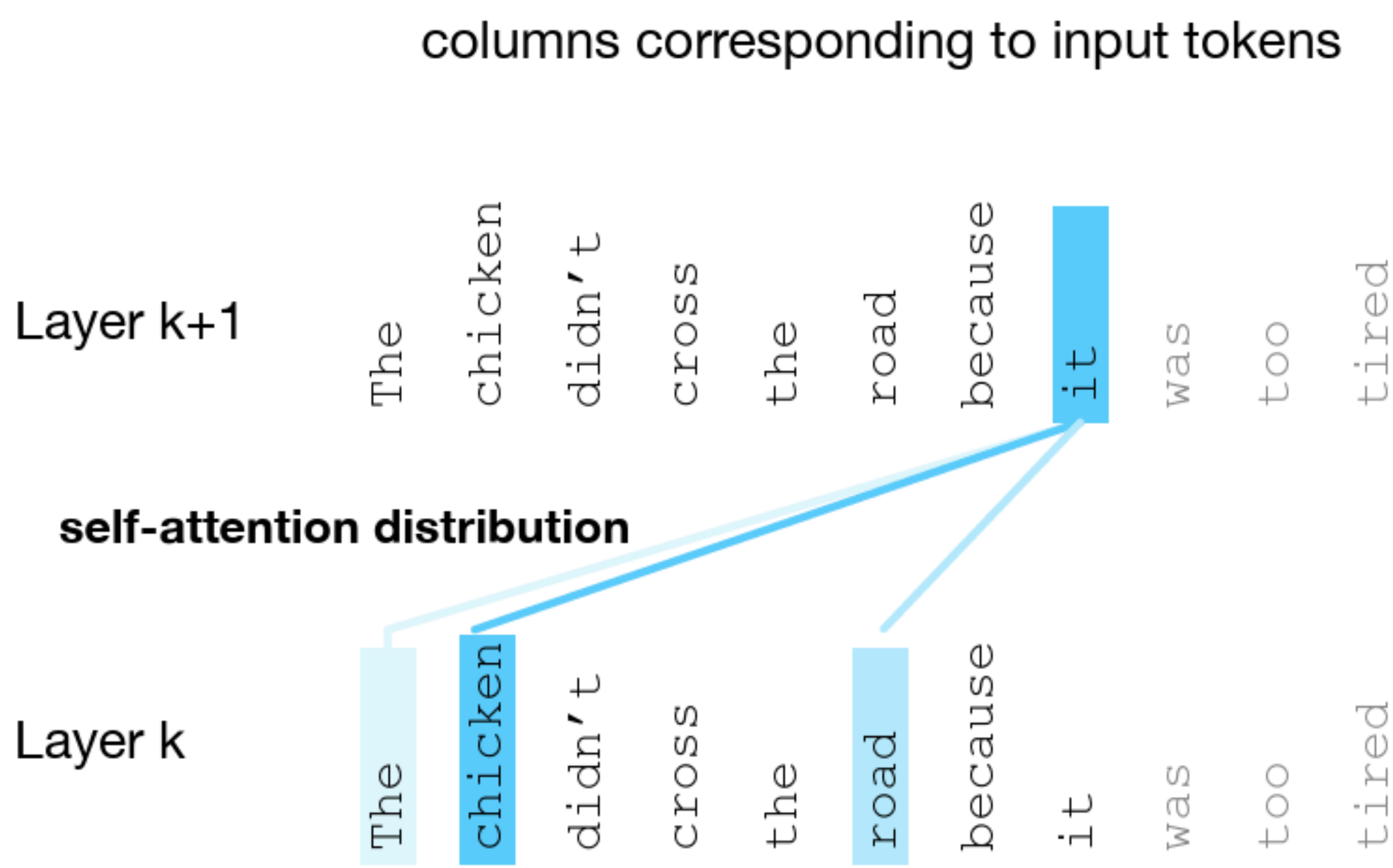
$$e_{ij} = a(s_{i-1}, h_j)$$

$$a(q,k) = \frac{q^T k}{\sqrt{|q|}}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_j \exp(e_{ij})}$$

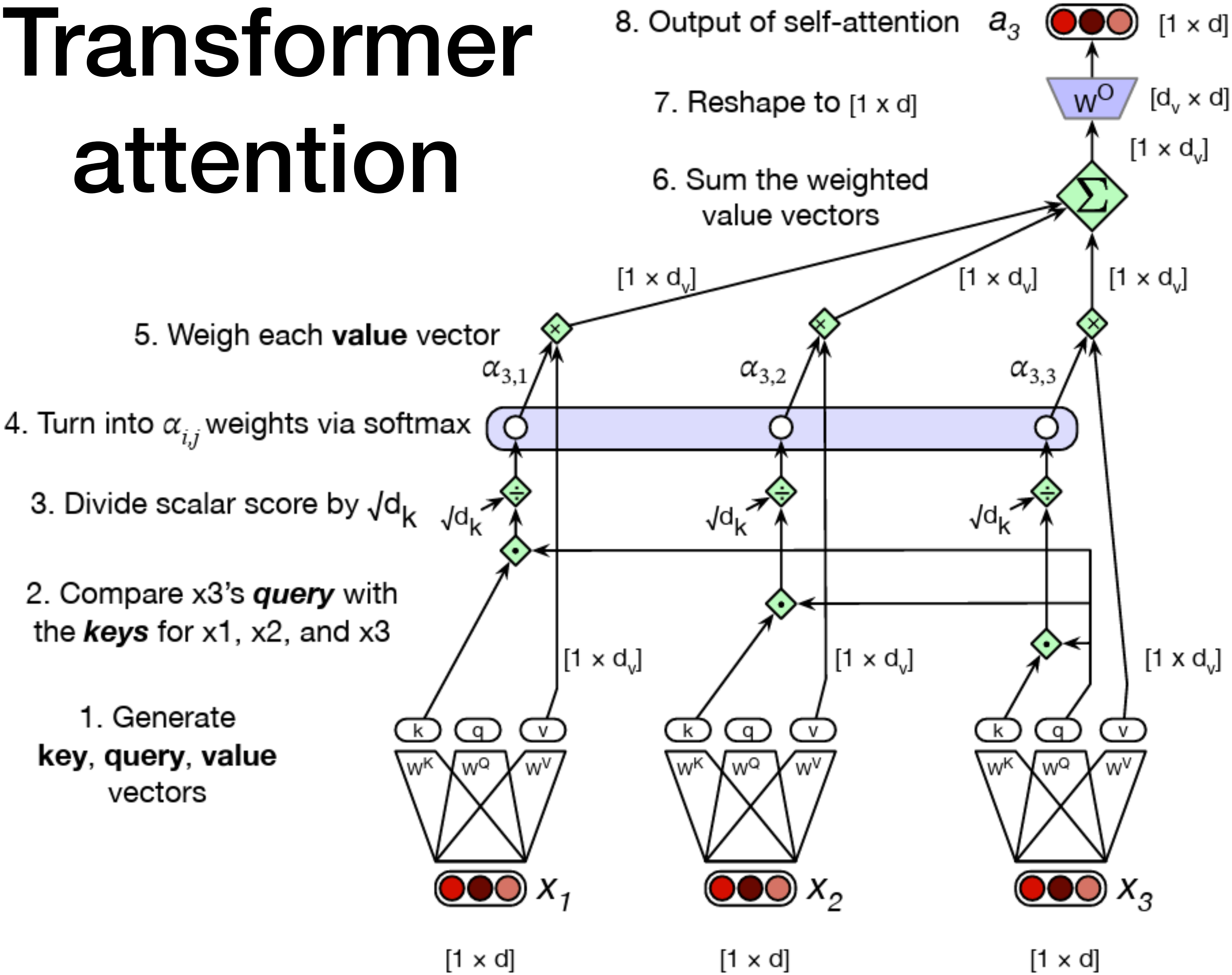
Softmax

Attention



columns corresponding to input tokens
Layer k+1
The
chicken
didn’t
cross
the
road
because
it
was
too
tired
self-attention distribution
Layer k
The
chicken
didn’t
cross
the
road
because
it
was
too
tired

Transformer attention



Multi-head attention

concatenation

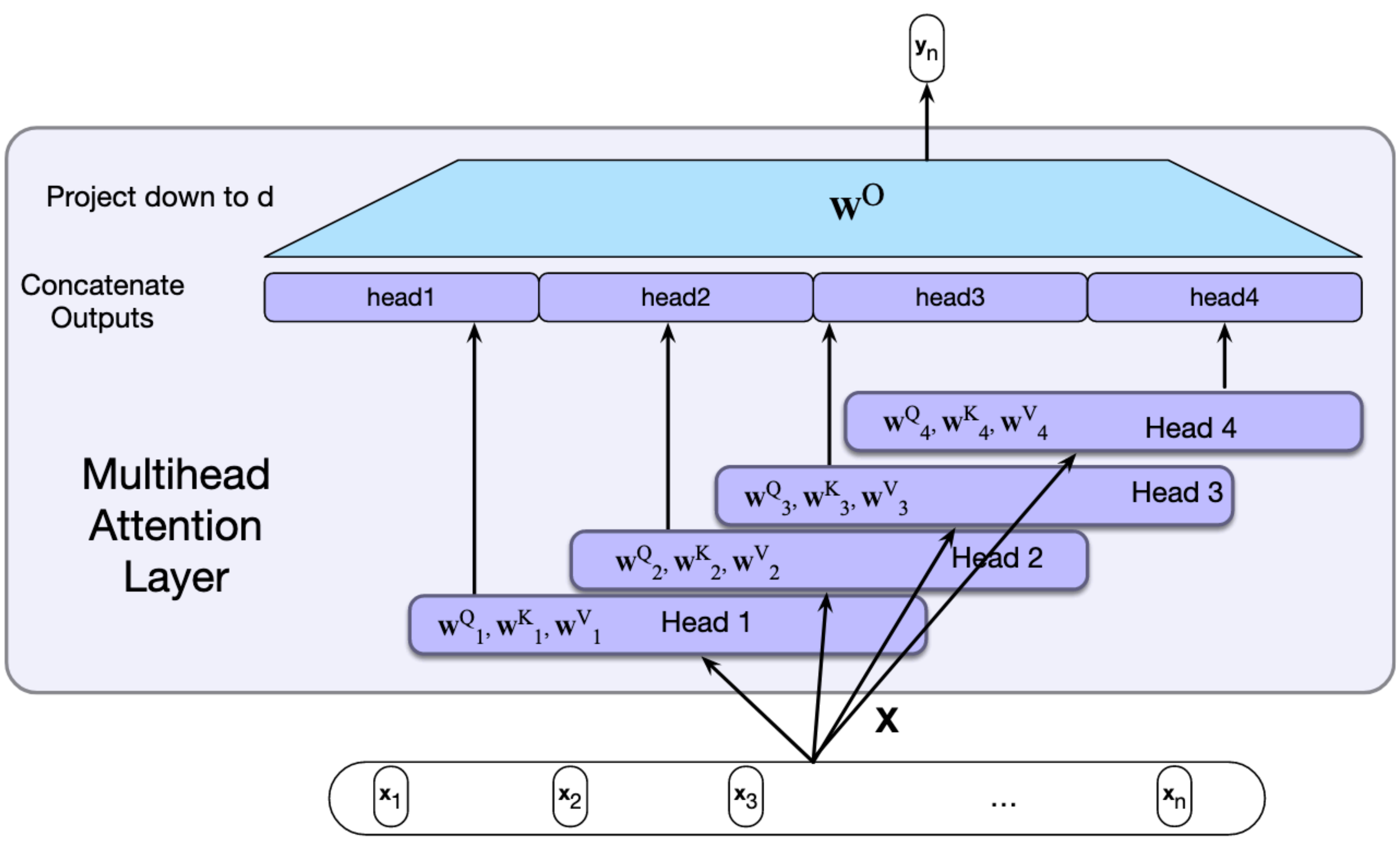
$$Y = [h_1; h_2; \ldots; h_k] W^O$$

$$h_i = \text{softmax}(Q_i, K_i, V_i)$$

$$Q_i = XW^{Q_i}$$

$$K_i = XW^{K_i}$$

$$V_i = XW^{V_i}$$



Language models as potential cognitive models

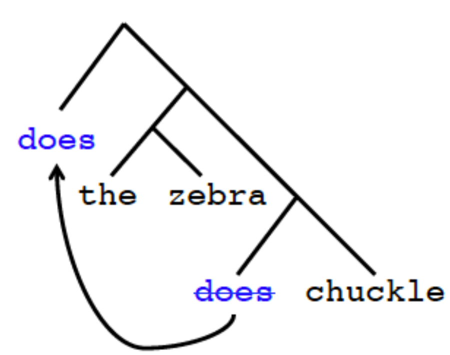
Deep learning and human language acquisition

- Modern neural networks are stronger learners than the cognitive models we had in the past—we can just unleash them on a corpus, without simplifying or annotating it
- We can ask: Which assumptions lead to the successful acquisition of linguistic generalizations? Do we need Universal Grammar? Perceptual grounding?
- And also: What representations emerge to support the network’s behavior?
- But we need to be able to control the assumptions: commercial “large” language models are not necessarily helpful

How difficult is it to acquire syntactic generalizations?

- **Input**: The zebra **does** chuckle.
- **Output:** **Does** the zebra chuckle?

MOVE-MAIN: Move the <u>main verb's</u> auxiliary to the front of the sentence.



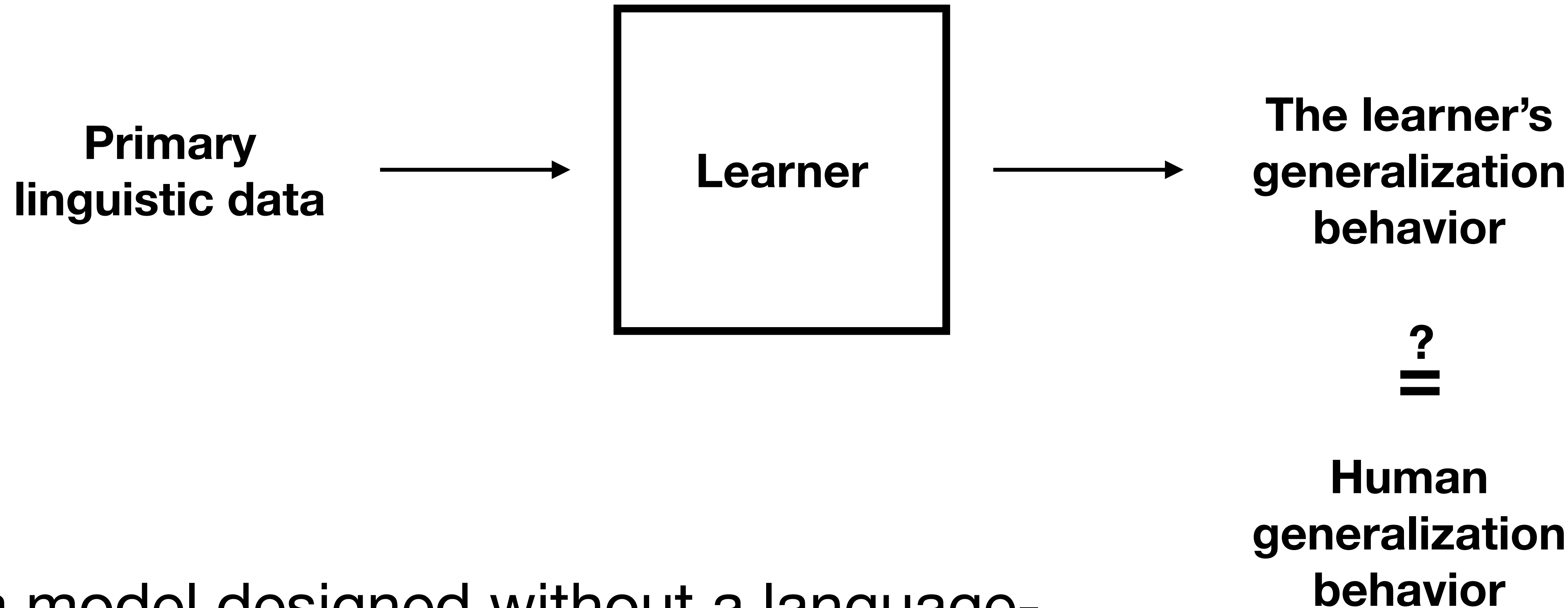
(Chomsky 1971; McCoy, Frank & Linzen, 2018, 2020)

How difficult is it to acquire syntactic generalizations?

First **Main**

- **Input**: My walrus that **will** eat **can** giggle.
  - MOVE-MAIN: **Can** my walrus that **will** eat _____ giggle?
  - MOVE-FIRST: **Will** my walrus that _____ eat **can** giggle?
- **Poverty of the stimulus argument:** children are not exposed to these cases, yet learn MOVE-MAIN
- Chomsky's solution to the learning problem: children only consider rules that are stated over a parse tree

Neural networks as computational infrastructure for cognitive modeling



- If a model designed without a language-specific bias generalizes like humans, innate inductive biases may not matter (data is enough)

Back to English question formation

First **Main**

- **Input**: My walrus that **will** eat **can** giggle.
  - MOVE-MAIN: **Can** my walrus that **will** eat ______ giggle?
  - MOVE-FIRST: **Will** my walrus that ______ eat **can** giggle?

Word prediction on CHILDES

Are you going to come with me or stay home?
No I hafta go now.
Don't you wanna go now Abe so that you'll be home in time to watch Charlie Brown?
Well I'm going.
How am I going to go if you're hanging on to me (.) without your jacket?
huh?
No I wanna be home to watch Charlie Brown.
Your new one?
I don't know .
Where'd you put it?
Abe.
Are you coming with me or not?

(CHILDES: MacWhinney 2000)

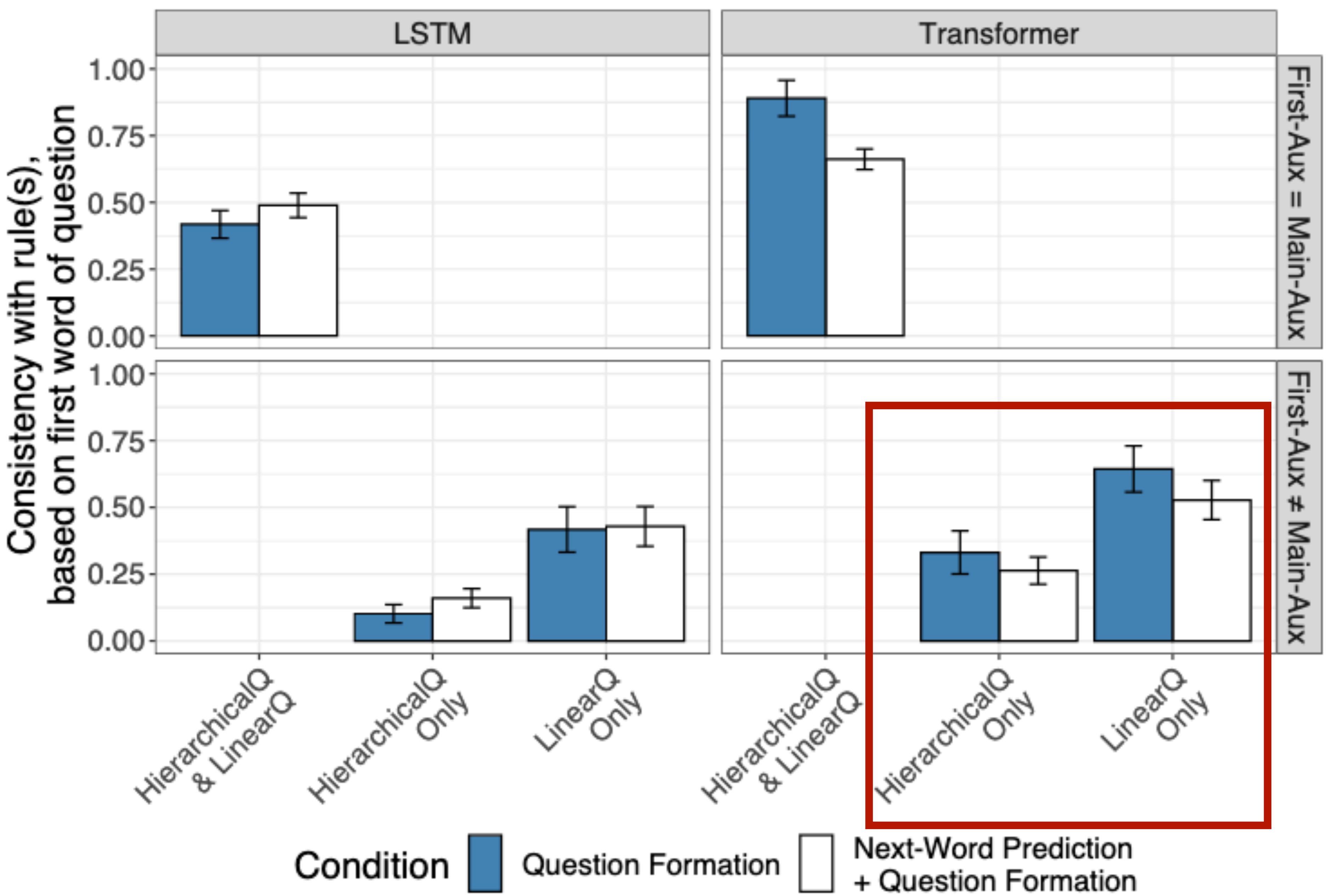
Child-directed speech: experimental setup

- Train ("pre-train") a language model on CHILDES
- Fine-tune it to form questions, based on the questions that actually occur in CHILDES:

```
(ROOT (SQ (VP (AUX does)
      (NP (PRP he))
      (VP (VB need) (NP (DT some) (NNS undies))))
      (. ?)))
```

`he needs some undies . does he need some undies ?`

Language models trained on child-directed speech learn the wrong generalization

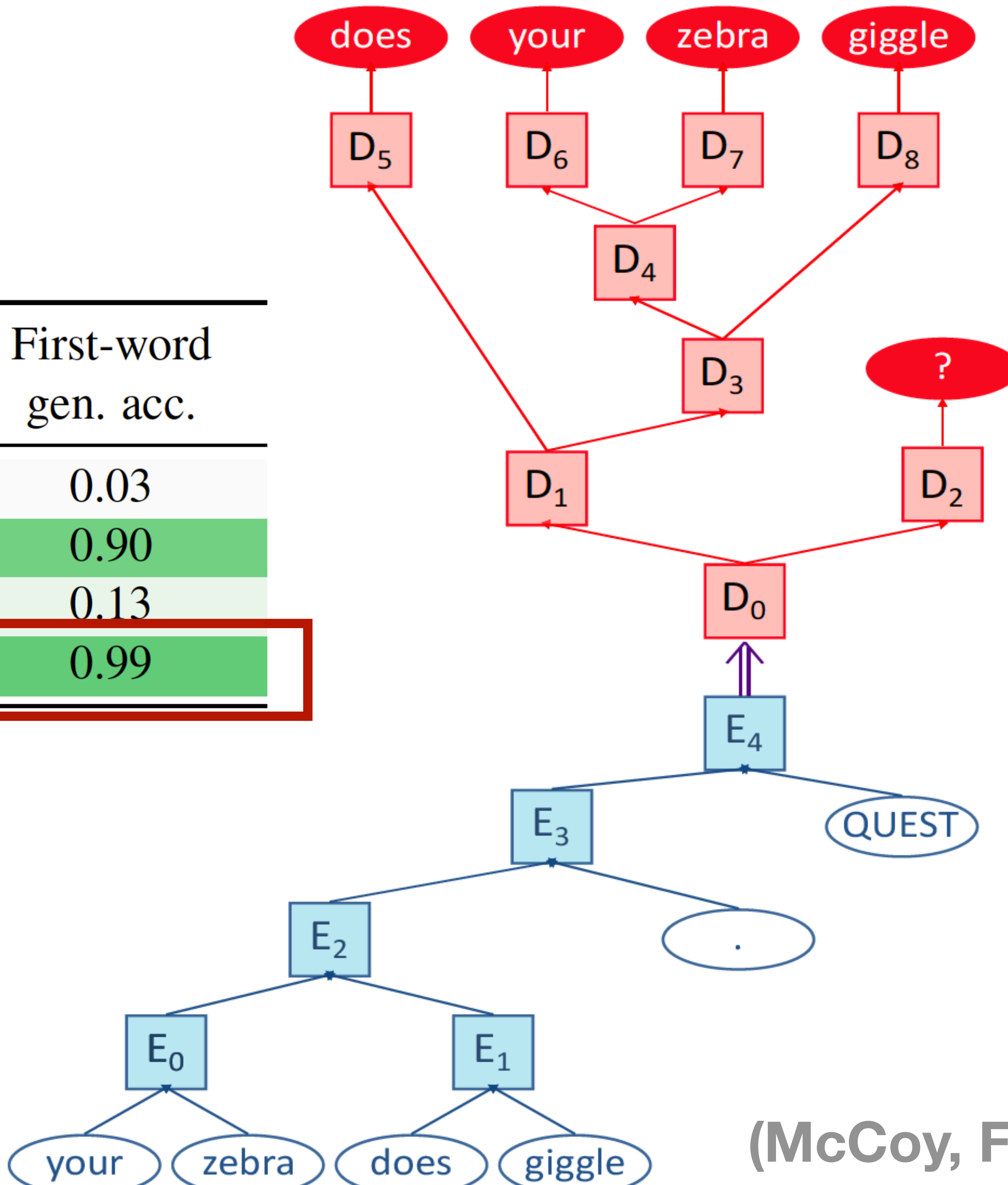


Most generated questions follow the (incorrect) linear generalization!

(Yedetore et al., 2023, ACL)

We can explicitly implement Chomsky's structure-sensitivity bias in a neural network!

Model	Full-sentence test acc.	First-word gen. acc.
Sequential/Sequential	0.88	0.03
Sequential/Tree	0.00	0.90
Tree/Sequential	0.96	0.13
Tree/Tree	0.96	0.99



(McCoy, Frank & Linzen, 2020)

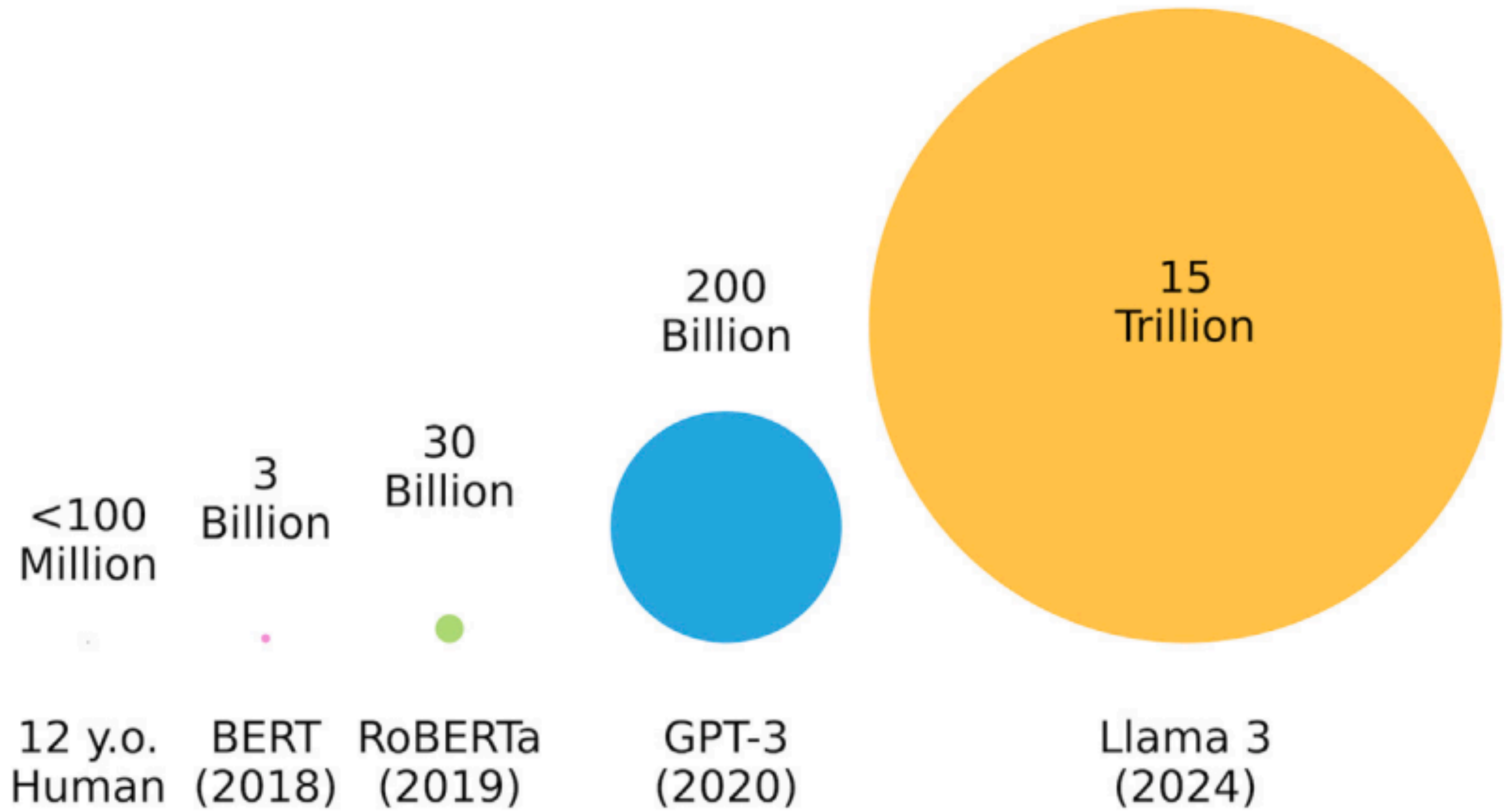
What about “LLMs”? You haven’t even mentioned ChatGPT in this section yet!

- Why do I think “large” language models are not very relevant?

- Size in and of itself is not a problem: you’d need a lot of parameters to describe all of the brain’s synapses, too

- The issue is that large deep learning models also require vastly more data than humans

Sample-efficiency and LLMs

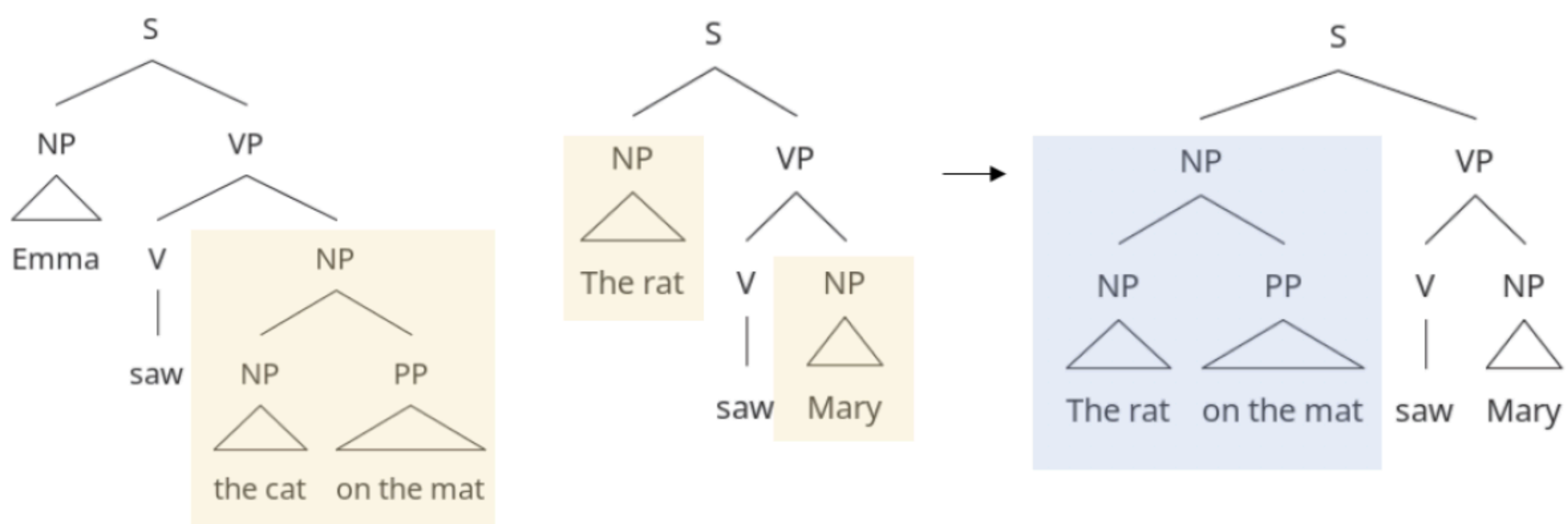


(Wilcox et al., 2025, Journal of Memory and Language)

Why size matters

- Most of the debates about language acquisition — and evaluations paradigms based on linguistics — have to do with **generalization**: how people learn to produce or understand structures they have never seen before

- If the model has observed every sentence structure known to mankind, it doesn't need to generalize!



(Kim & Linzen, 2020, EMNLP)

This is not just about the *amount* of data

- Most recent language models are trained on materials that are increasingly cognitively implausible:
  - Dozens of languages at the same time
  - Billions of lines of source code
  - ESL textbooks, dictionaries, linguistics articles…

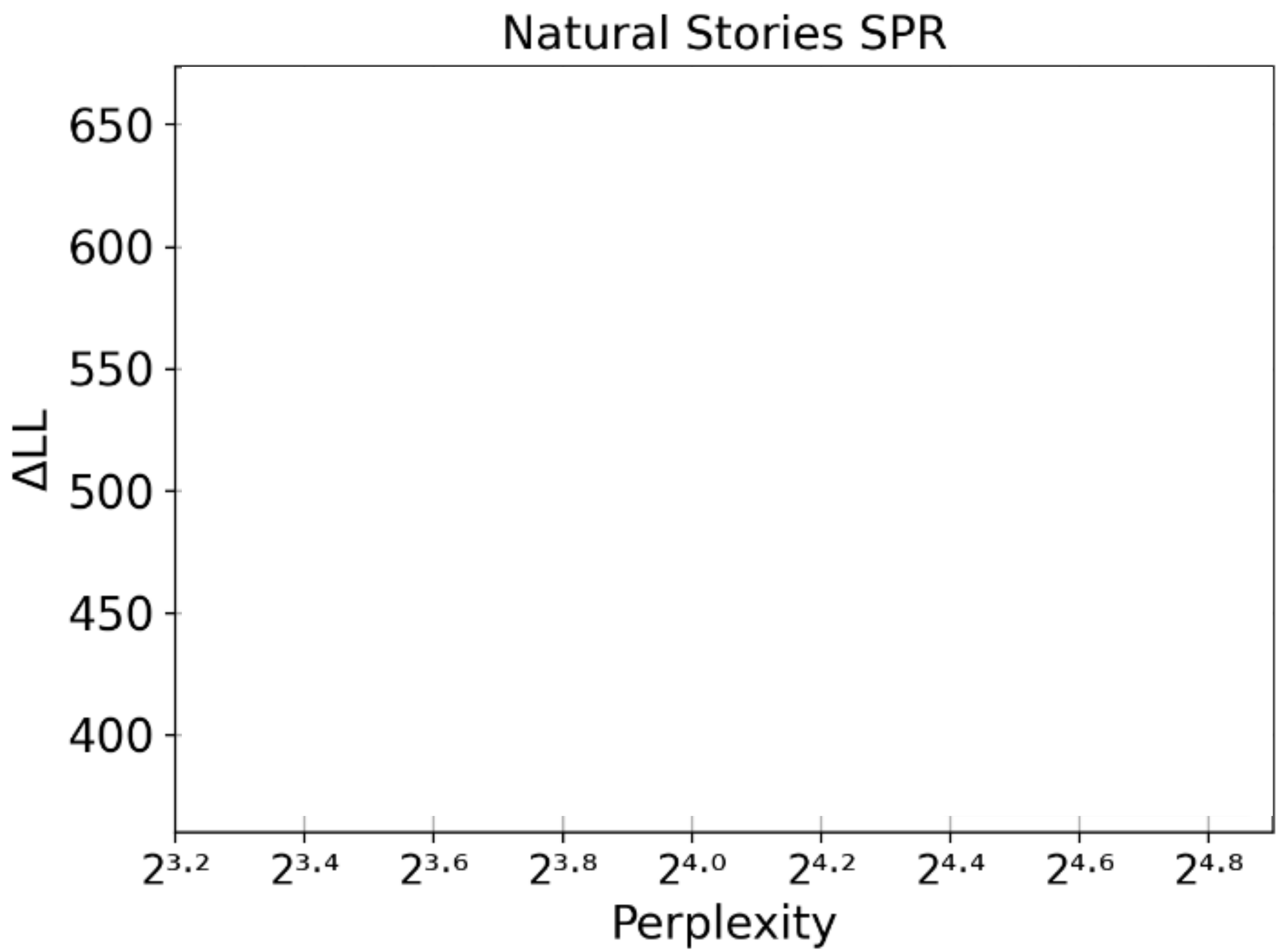
This is not just about the *amount* of data

- Interesting exception: human feedback (though OpenAI annotators' feedback is likely very different than parents')
- Even more recently development: we just don't know what's in the data! The training data is a trade secret

Predictions from larger language models are increasingly non-human-like

better fit to human reading times

worse fit to human reading times



more accurate word predictions

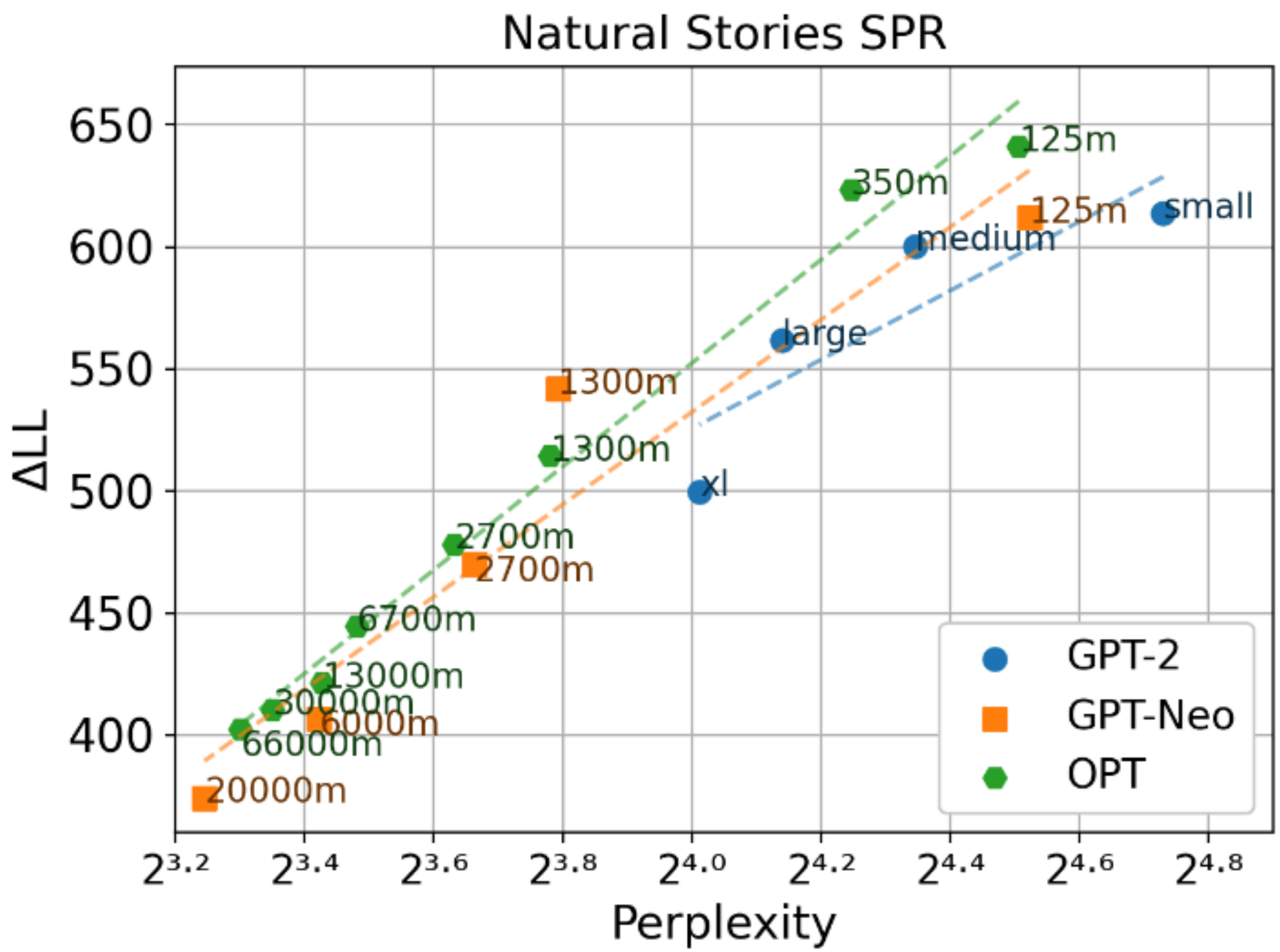
less accurate word predictions

(Oh & Schuler, 2022)

Predictions from larger language models are increasingly non-human-like

better fit to human reading times

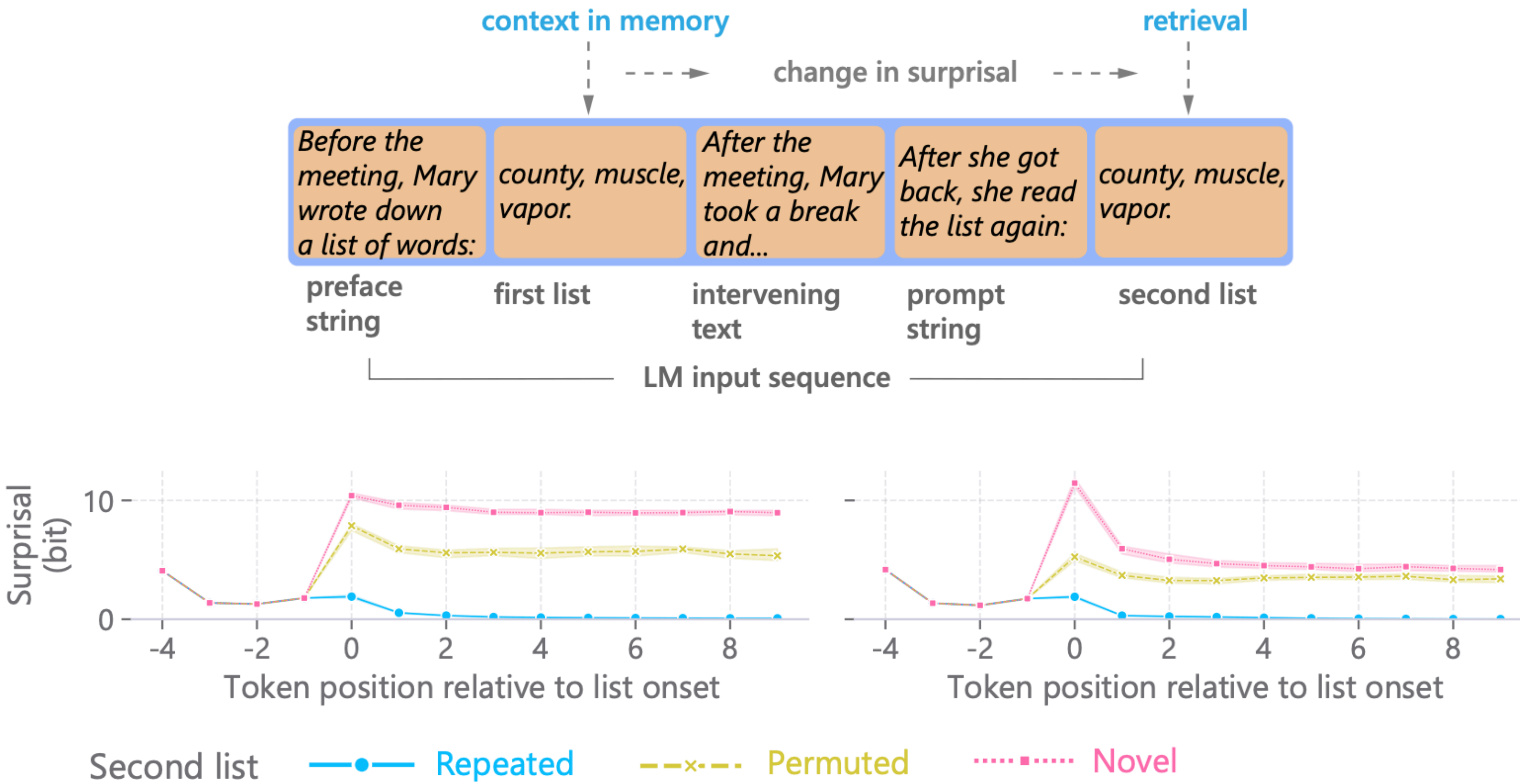
worse fit to human reading times



more accurate word predictions ⟷ **less accurate word predictions**

(Oh & Schuler, 2022)

Unlike humans, transformers have perfect memory recall for lists



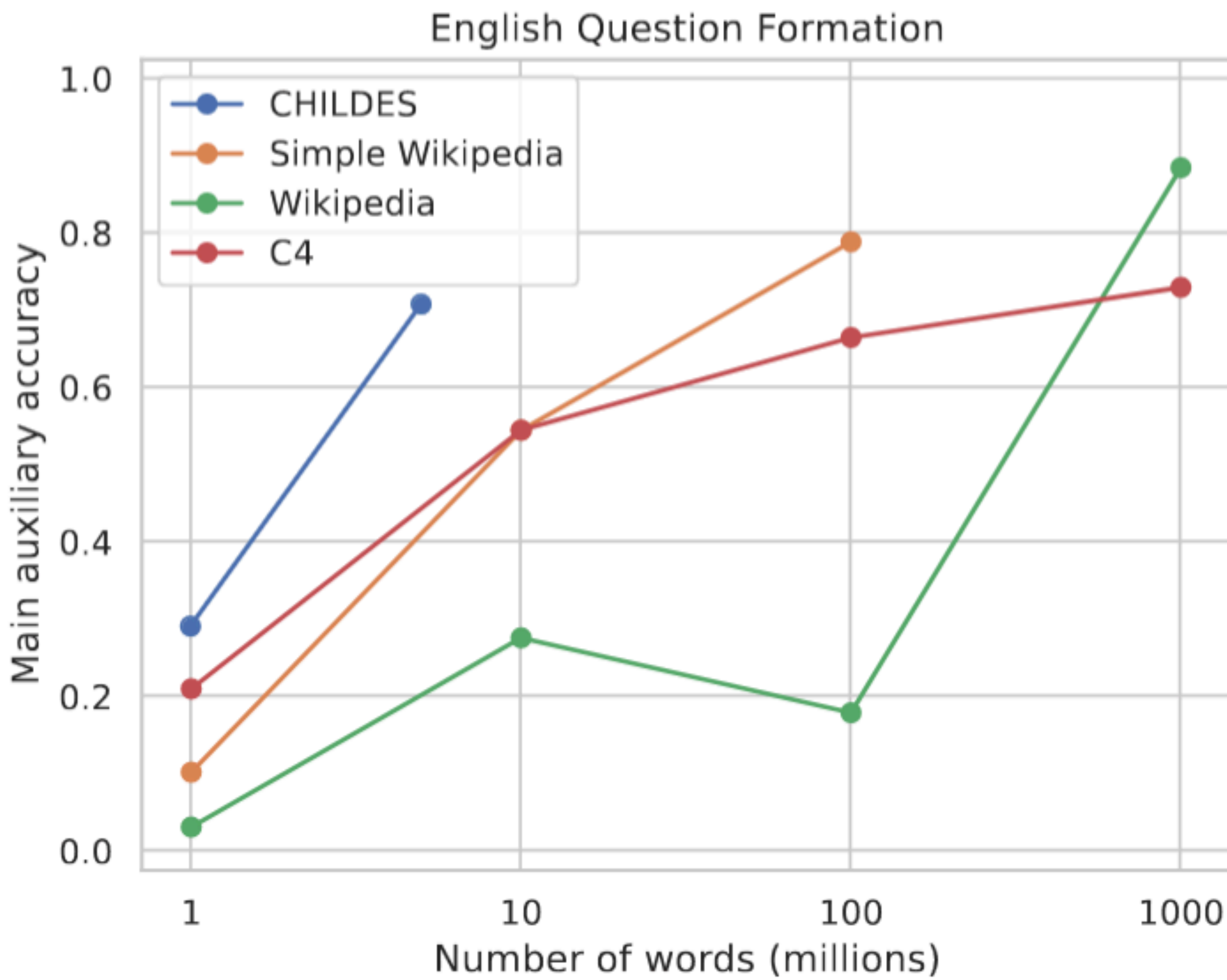
(Armeni, Honey & Linzen, 2022, CoNLL)

English question formation: reminder

First **Main**

- **Input**: My walrus that **will** eat **can** giggle.
  - MOVE-MAIN: **Can** my walrus that **will** eat _____ giggle?
  - MOVE-FIRST: **Will** my walrus that _____ eat **can** giggle?

The cognitive plausibility of the corpus can have surprising effects



(Mueller & Linzen, 2023)

BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

- Yearly shared task (started in 2023) with a standardized evaluation pipeline

- 100-million-word cognitively plausible corpus: child-directed speech, transcribed speech, children's books…

Conclusions: Language models as potential cognitive models

- Deep learning can be a useful **infrastructure** for studying how learning outcomes are affected by theoretical assumptions about learners' input and inductive biases

- To study those assumptions, we usually need human-size models trained on human-appropriate data

- We may also need models that are resource-limited in human-like ways (e.g. in terms of their memory capacity)

- These may not be the models that corporations find lucrative: we need to train different models for cognitive modeling

When do we want human-like language models?

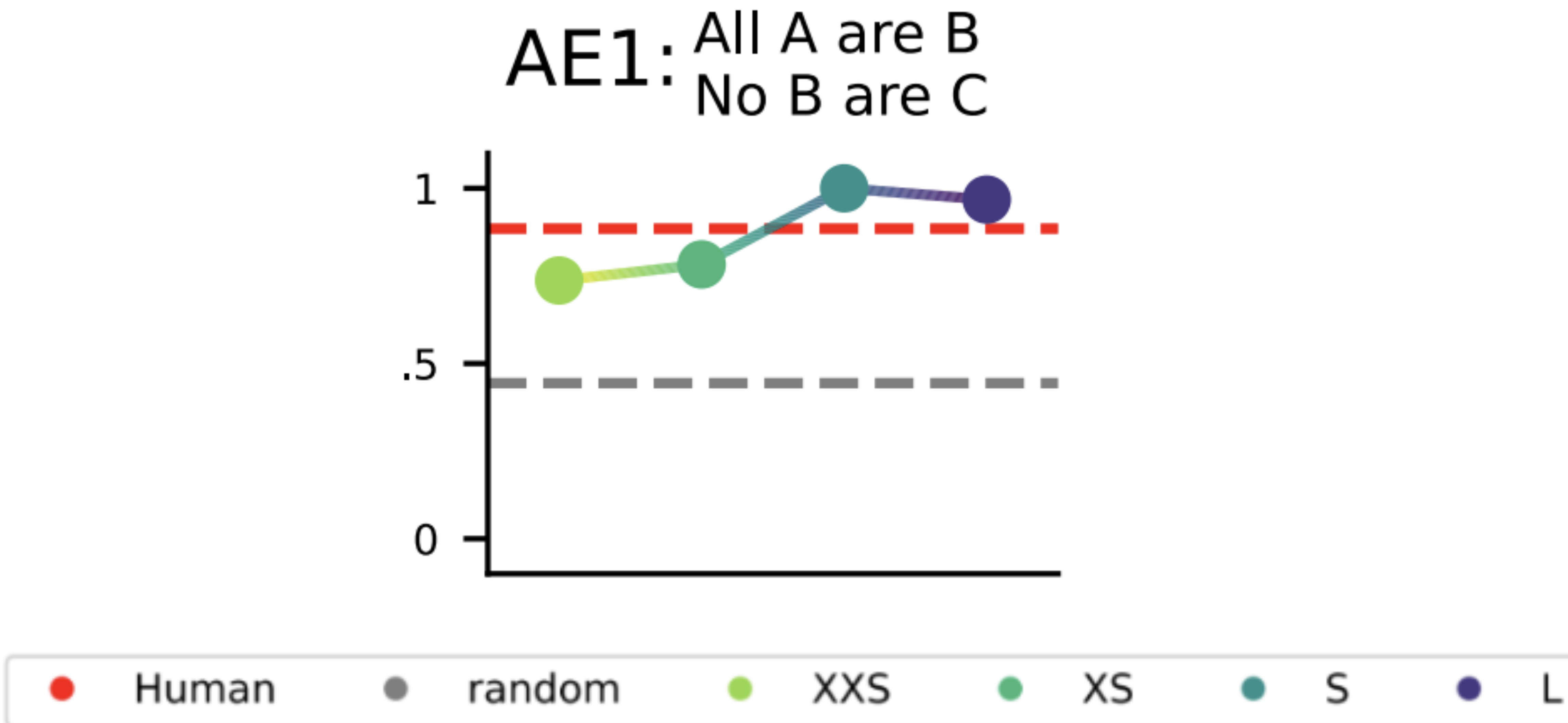
Syllogistic reasoning

- *All bakers are artists.*
- *All artists are chemists.*
- What follows?
- The correct answer: all artists are not chemists

Syllogistic reasoning

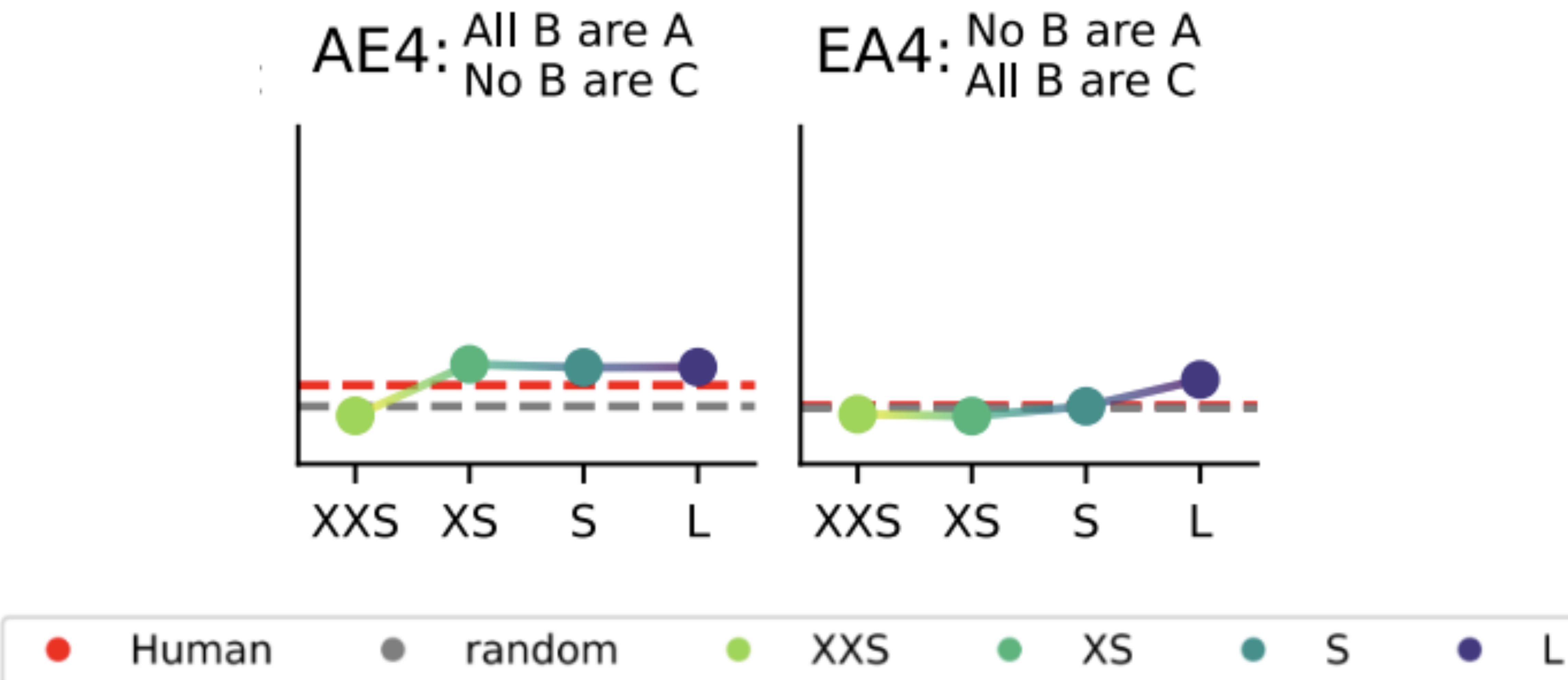
- *All bakers are artists.*
- *No chemists are bakers.*
- What follows?
- The correct answer: some artists are not chemists
- Almost all human participants reason incorrectly here!
- Do we want to use this as a benchmark for LLM reasoning?

Models do well on some of the same syllogisms as people



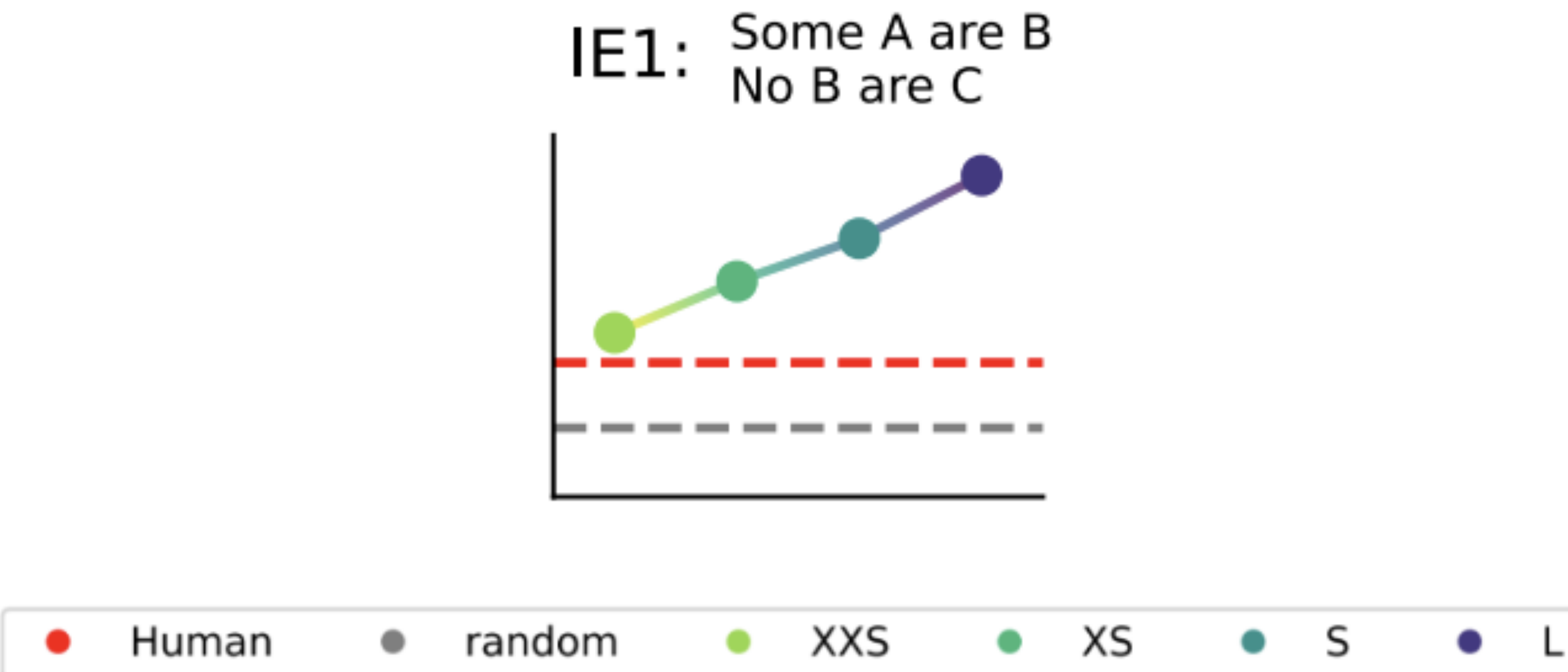
(Eisape et al., 2024)

Models do poorly on some of the same syllogisms as people



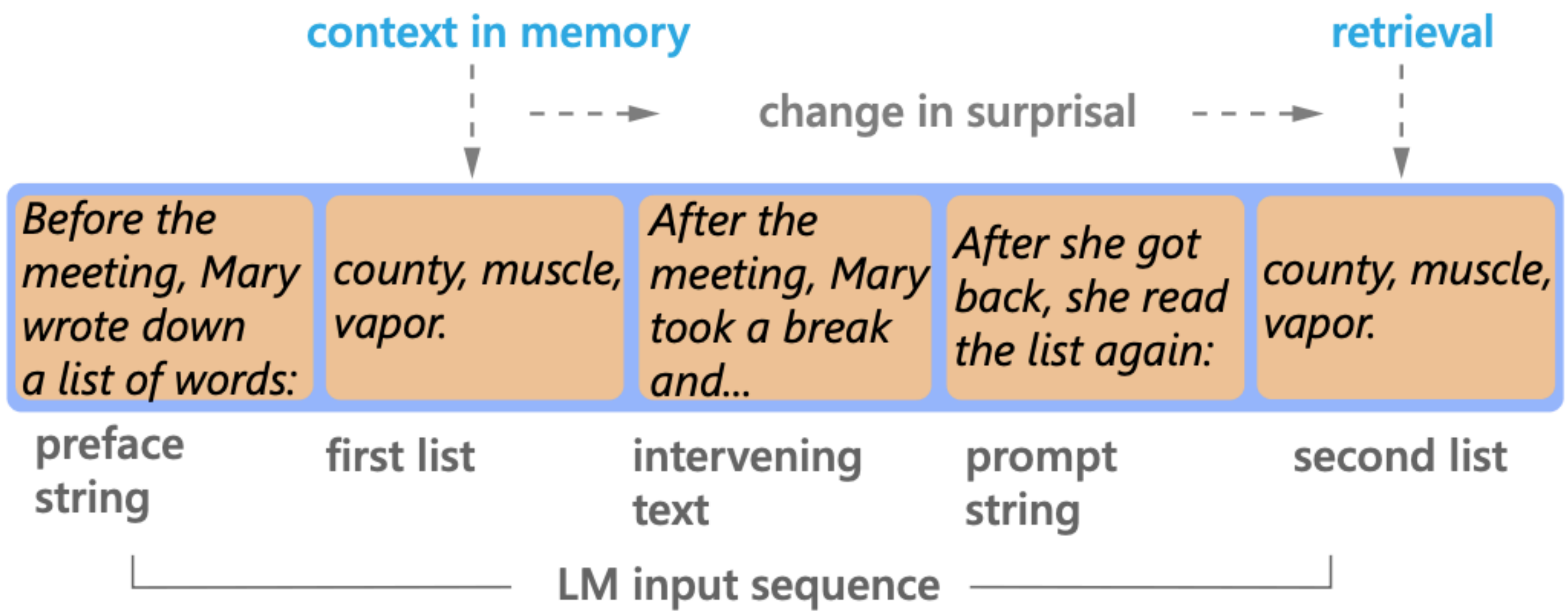
(Eisape et al., 2024)

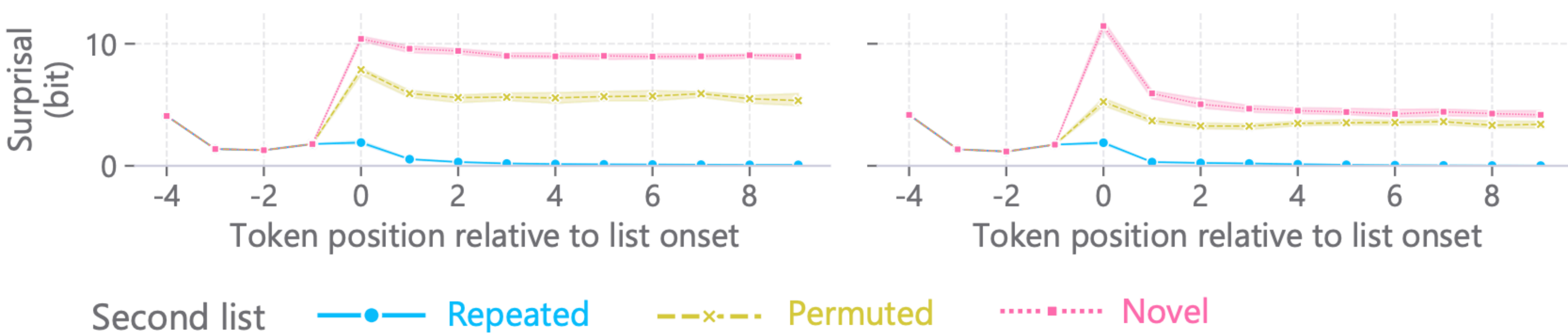
Models, especially larger ones, do better than people on some syllogisms



(Eisape et al., 2024)

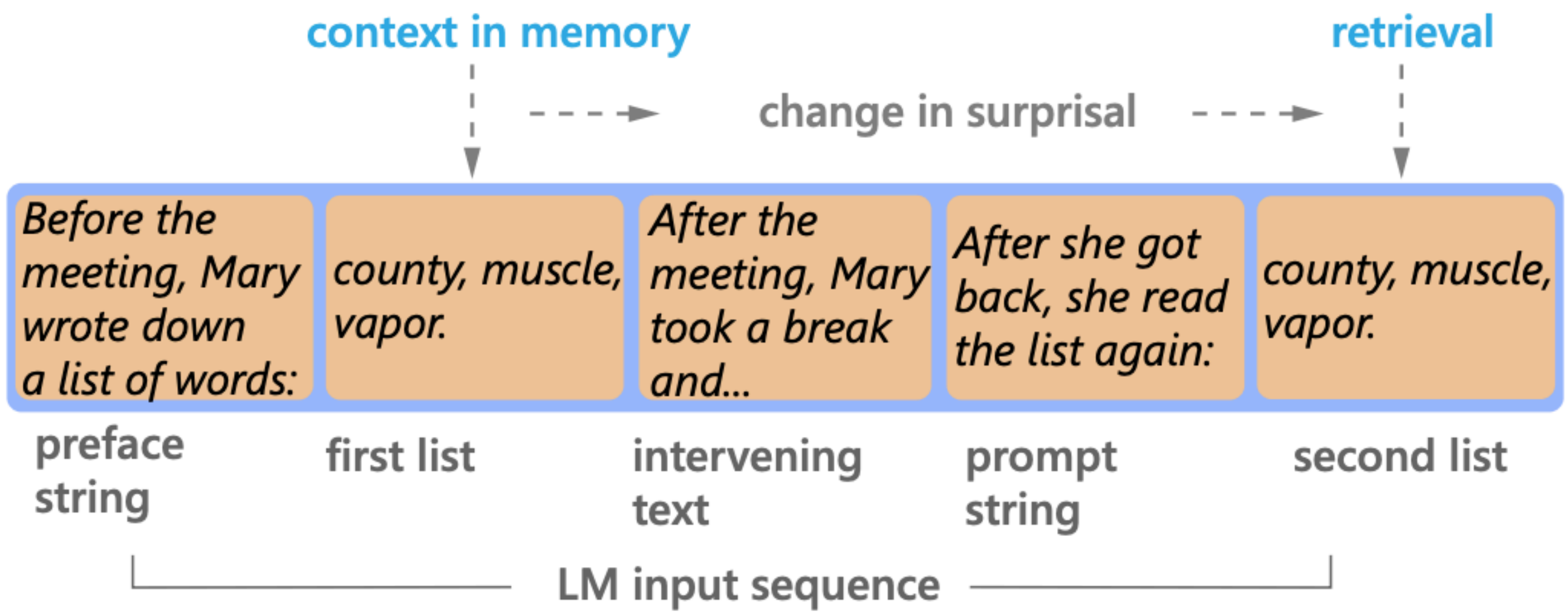
Another example: unlike humans, transformers have perfect memory recall for lists

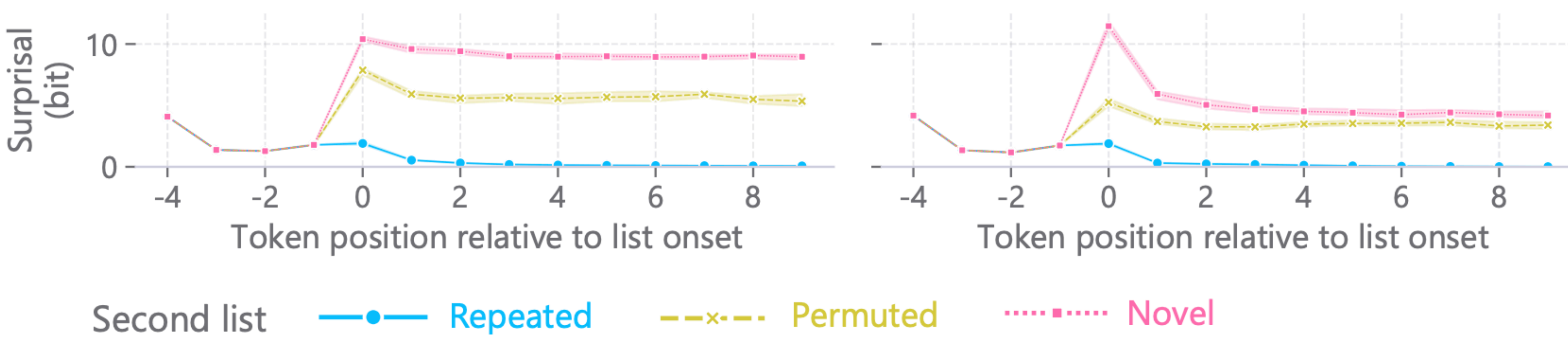




(Armeni, Honey & Linzen, 2022, CoNLL)

For most applications, great working memory is a good thing!





(Armeni, Honey & Linzen, 2022, CoNLL)

Conclusions: When do we want human-like language models?

- Chomsky distinguished “competence” (a system’s in-principle capabilities) from “performance” (how it deploys those capabilities in practice): working memory is a clear case where the two diverge

- For commercial applications, it’s unclear if ever want to match human “performance” (including errors and limitations)

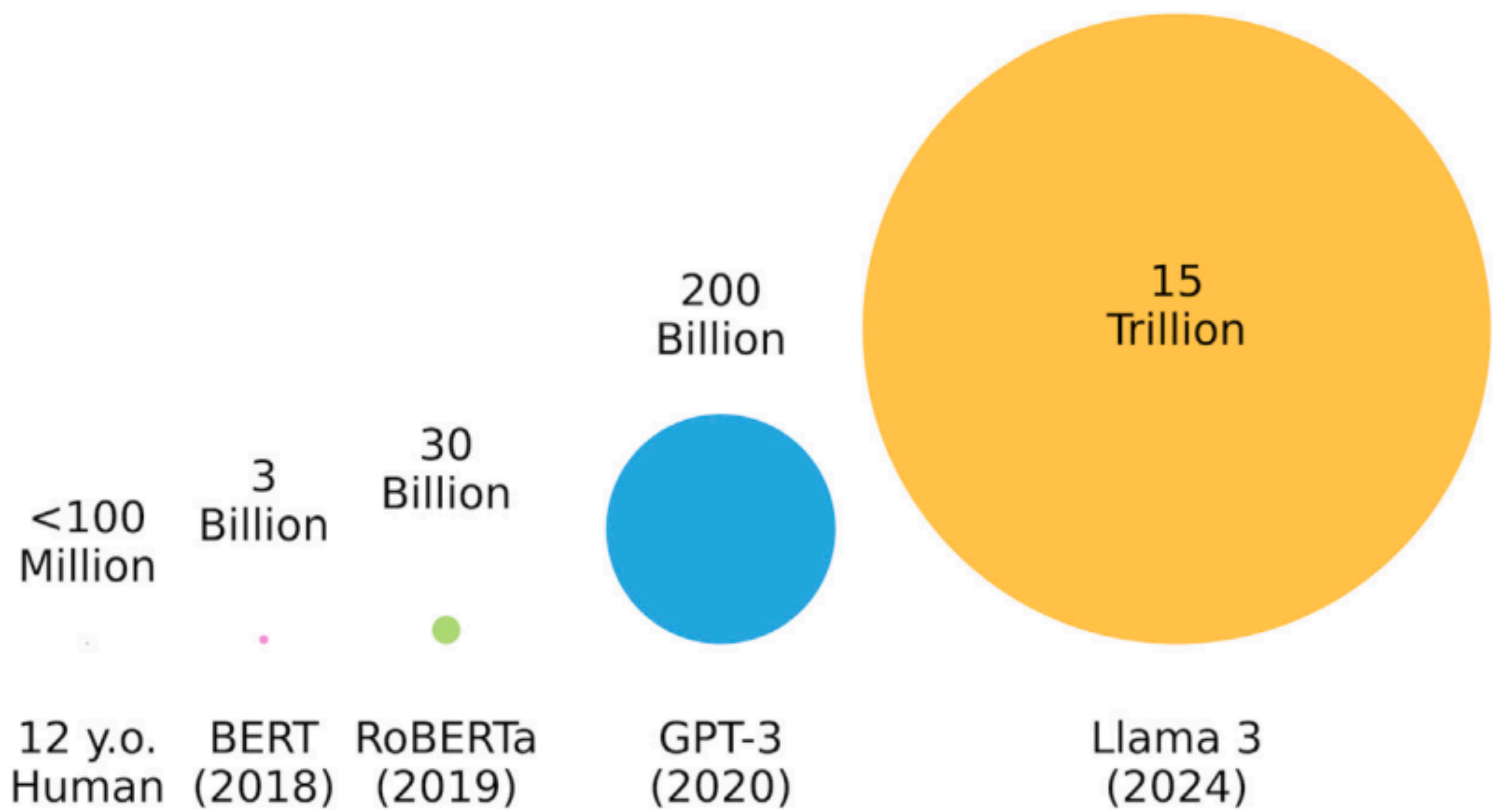
- If we do want to match it — for cognitive modeling — we will need models that are resource-limited in human-like ways, e.g. in terms of their memory capacity (unlike transformers)

Improving data efficiency with formal language pretraining

Motivation

- (Untrained) neural networks, e.g. transformers, have weak inductive biases
    - With enough data, transformers can learn to model not just language, but also protein structure, images, etc

Sample-efficiency and LLMs

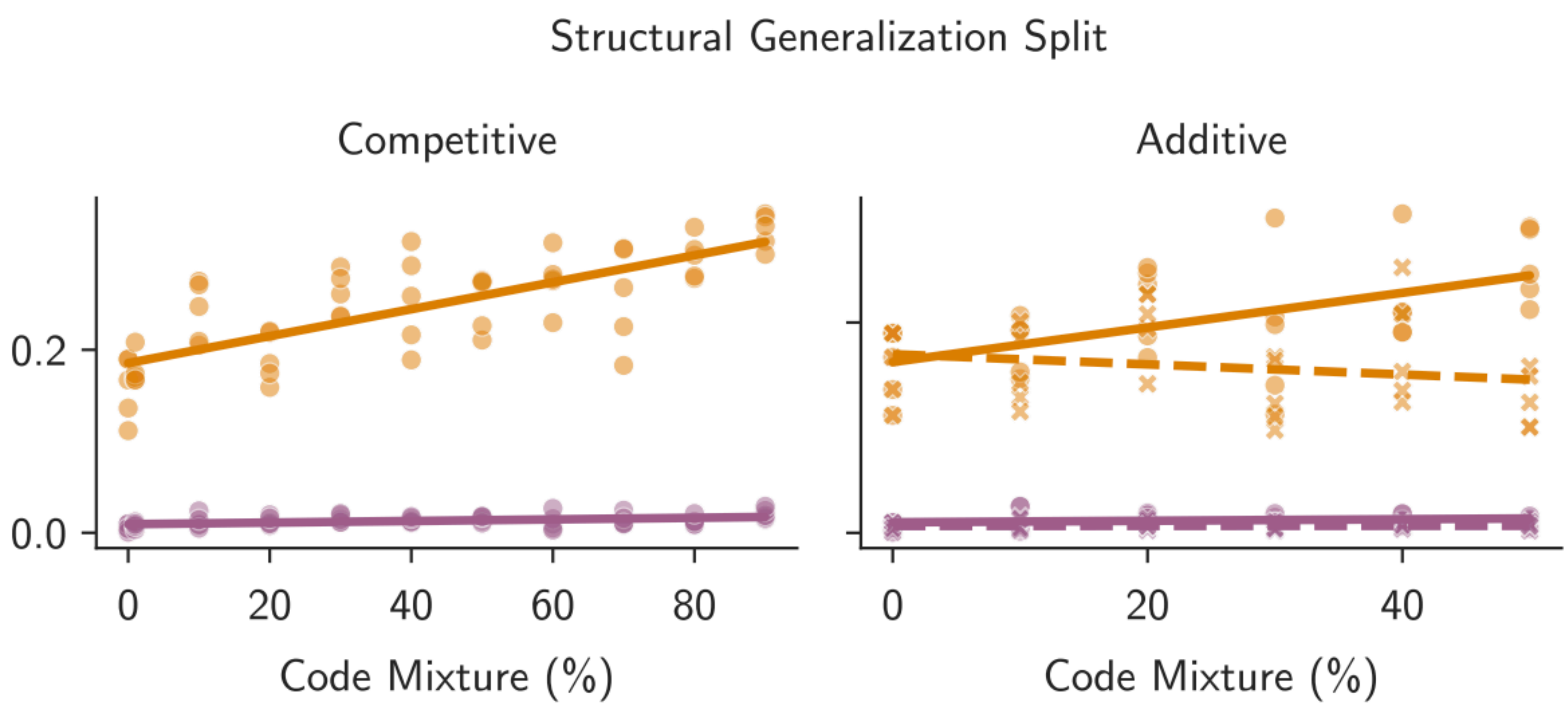


(Wilcox et al., 2025, Journal of Memory and Language)

Pre-pretraining

- Hypothesis 1: by pre-pretraining a transformer on a formal language, we can
    - Increase sample efficiency (counted by natural language tokens)
    - Increase compute efficiency (counted by total number of tokens)
    - Improve generalization in natural language

Indirect evidence for the hypothesis: pretraining on source code improves performance on natural language tasks



(Petty, van Steenkiste and Linzen, TMLR, 2025)

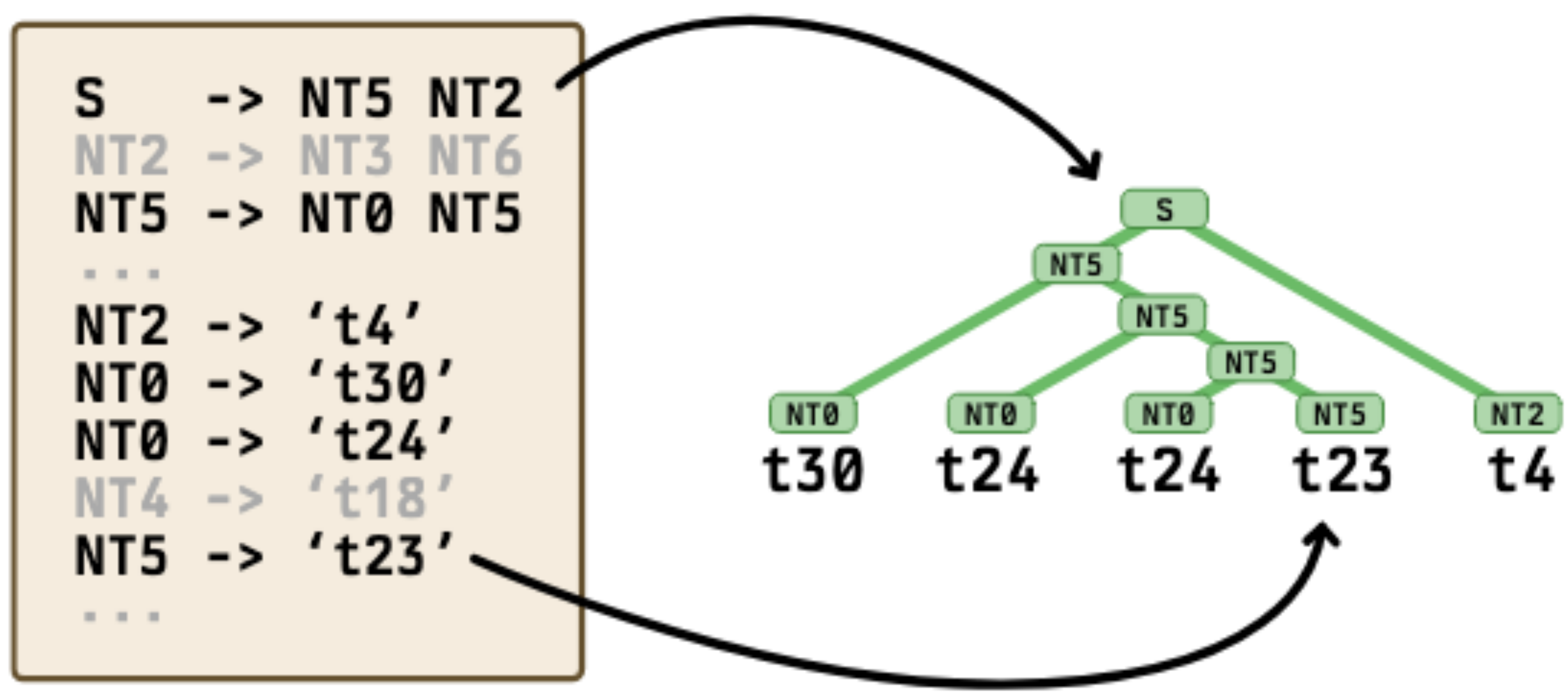
Pre-pretraining

- Hypothesis 2: the formal languages that would work best are those that
  - Contain structure that mimics the structures found in natural language
  - Are a good fit to the model's computational architecture (can be learned effectively)

The languages we use

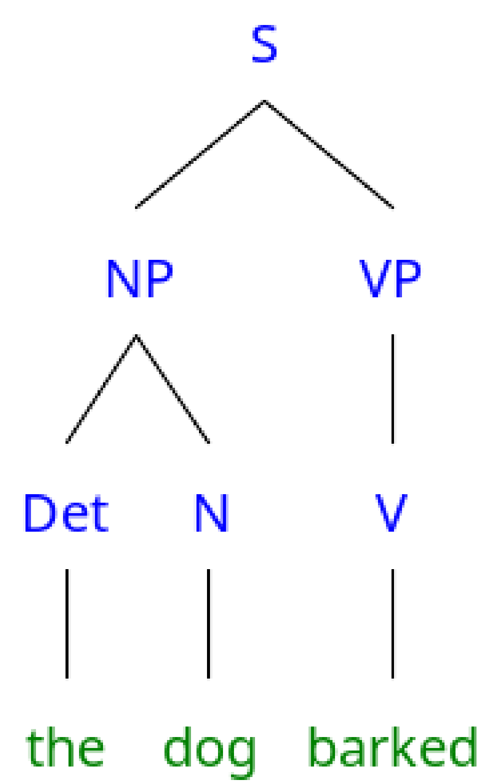
Language	Example
1-Dyck	((( )))
k-Dyck	([{ }])
k-Shuffle Dyck	([{]) }
ww	1 2 3 1 2 3

Background: the Chomsky hierarchy



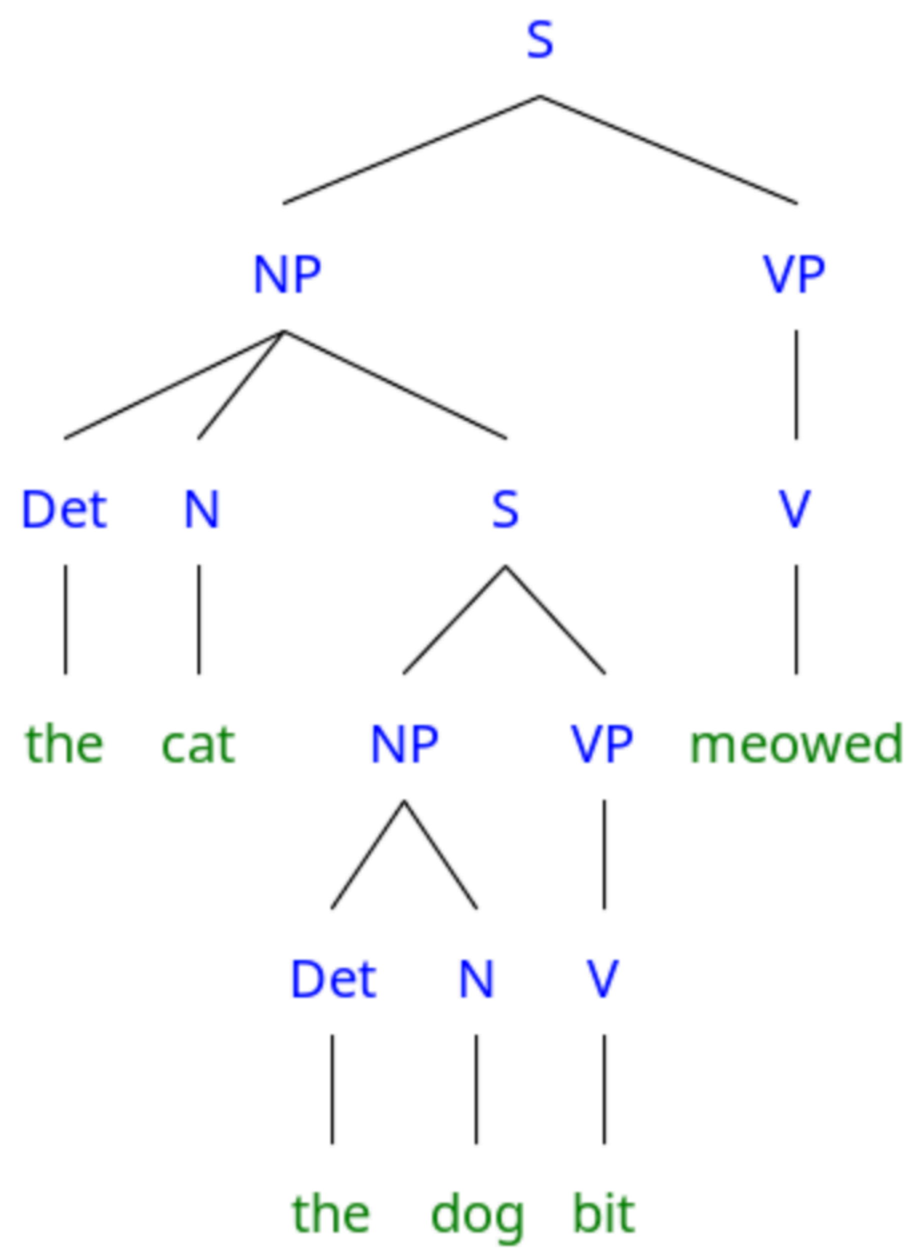
Context-free languages can be generated from a context-free grammar: many phenomena in natural language syntax are context-free, but some are context-sensitive

Context-free languages as a model of natural languages



Context-free languages as a model of natural languages

NP → Det N S



Background: transformer complexity

- First-order logic with majority (FO(M)) is an upper bound on transformer expressivity: any language a transformer can implement is in FO(M) (Merrill and Sanharwal, 2023)
- C-RASP is a lower bound on transformer expressivity: if a language is definable with a C-RASP program, there exists a transformer that recognizes it (Yang and Chiang, 2024)

The languages

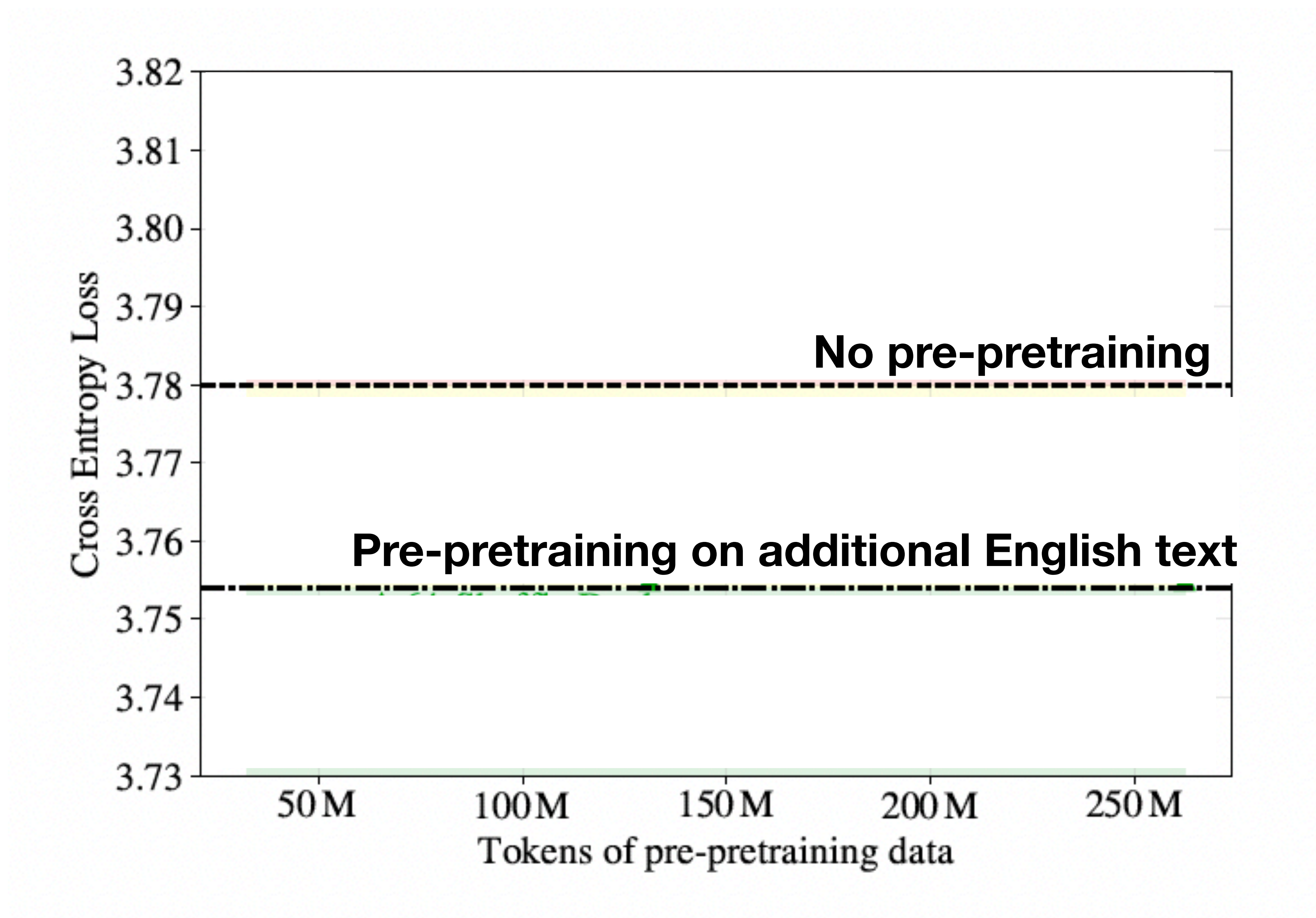
Language	**Example**
1-Dyck	((( )))
k-Dyck	([{ }])
k-Shuffle Dyck	([{]) }
ww	1 2 3 1 2 3

	Context-free	Context-sensitive
C-RASP	1-Dyck	k-Shuffle Dyck
FO(M)	k-Dyck	ww

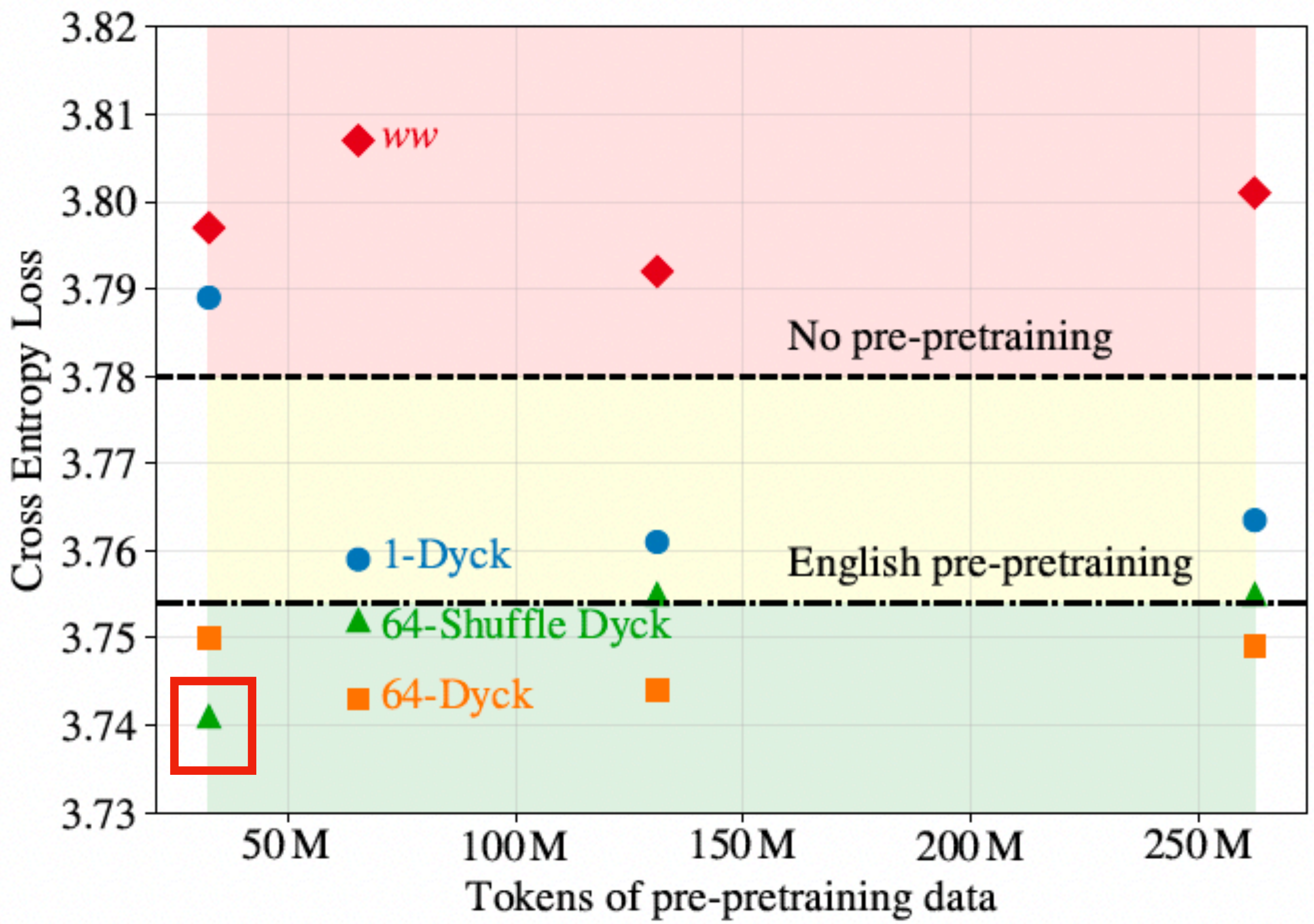
Experimental setup

- Architecture based on Pythia 160M transformer LMs (Biderman et al 2023)
- We use the C4 natural language corpus
- We train for 10000 steps, or 600M tokens of natural language
- Preceded by pre-pretraining on a formal language

Results: language modeling loss

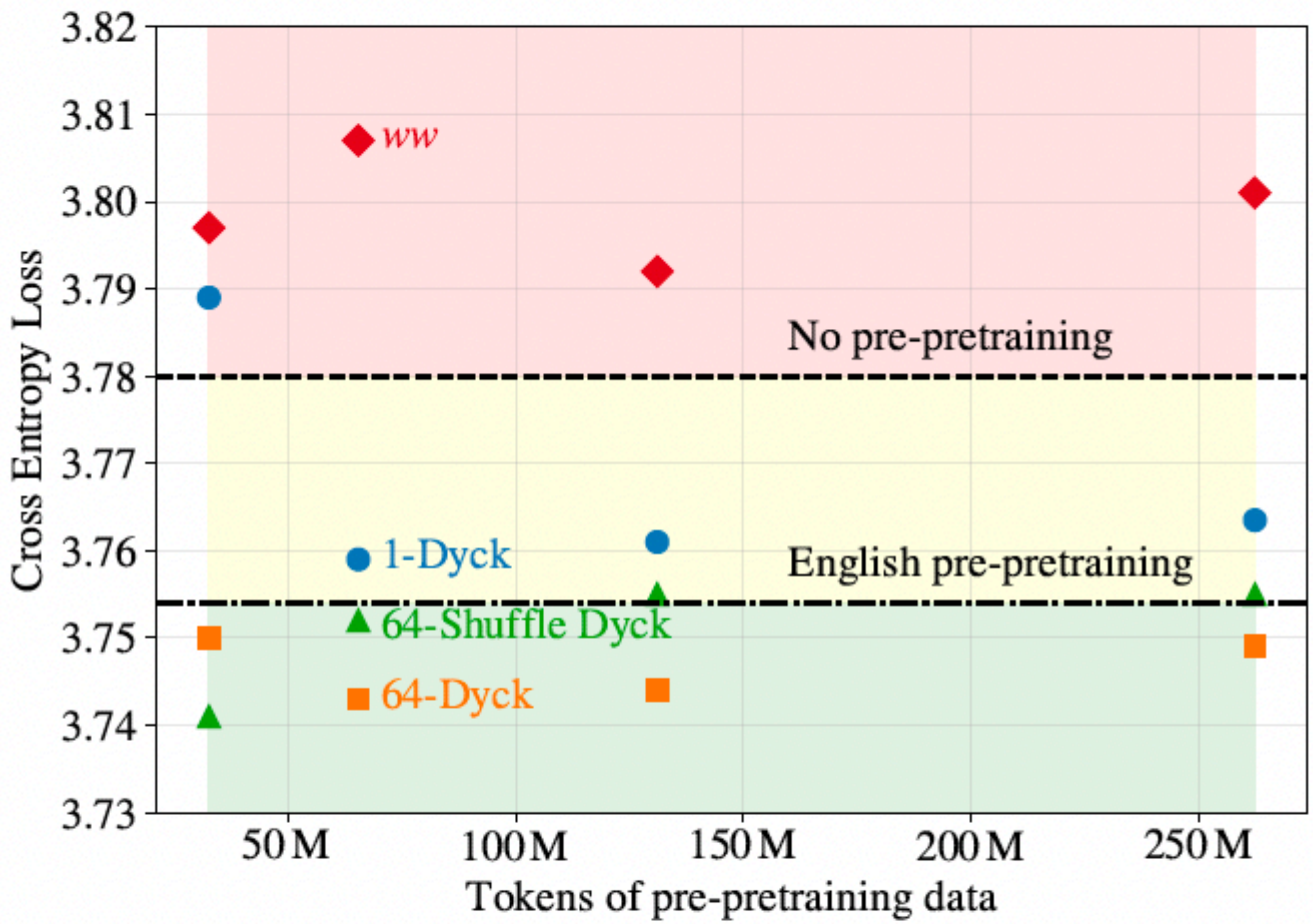


Results: language modeling loss



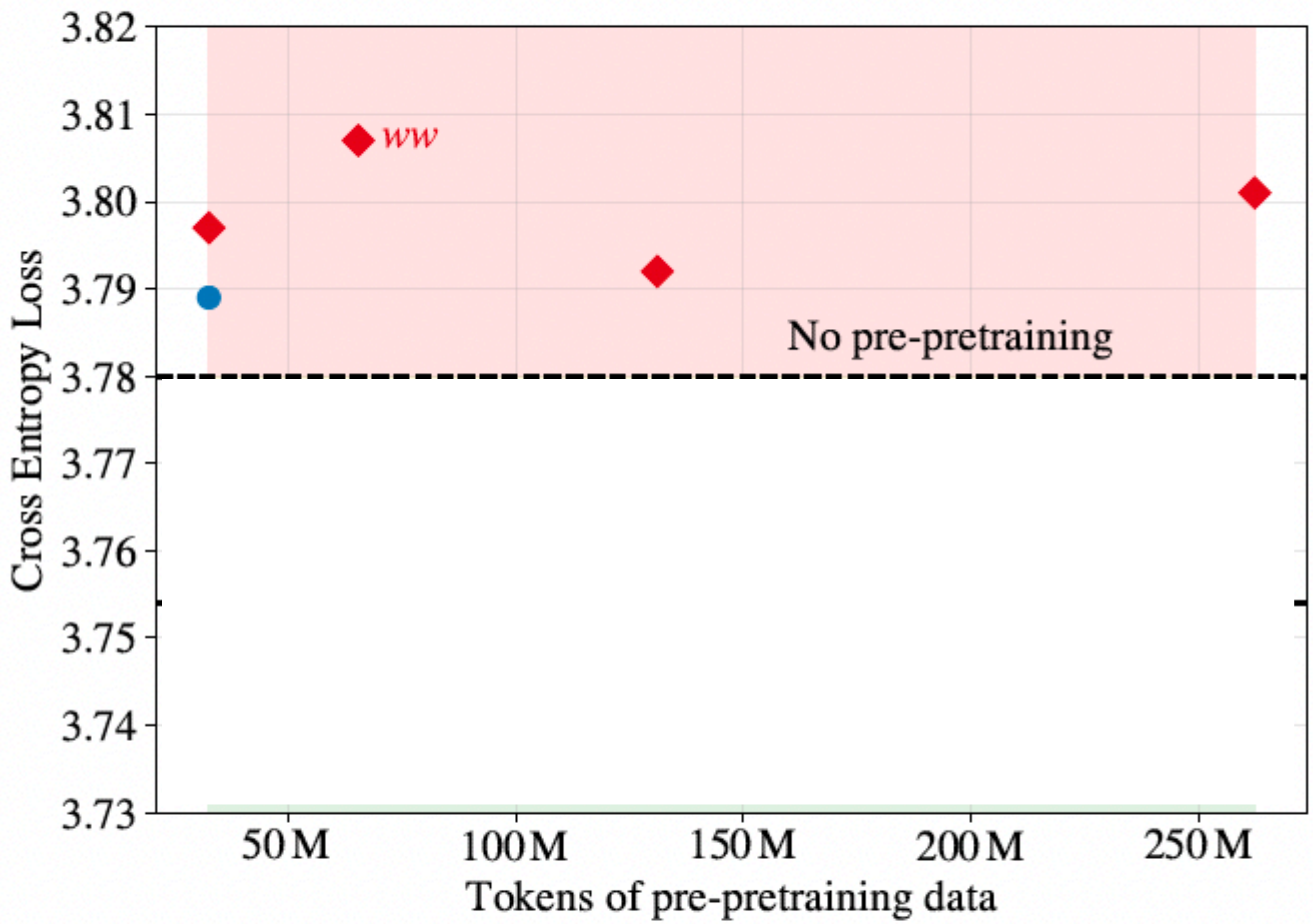
It is best to pre-preretrain for fewer tokens on Shuffle Dyck

Results: language modeling loss



Past a certain point, additional prepretraining no longer helps and may hurt

Results: language modeling loss



Prepretraining on the copy language ww is always harmful

Targeted syntactic evaluation with minimal pairs (BLiMP)

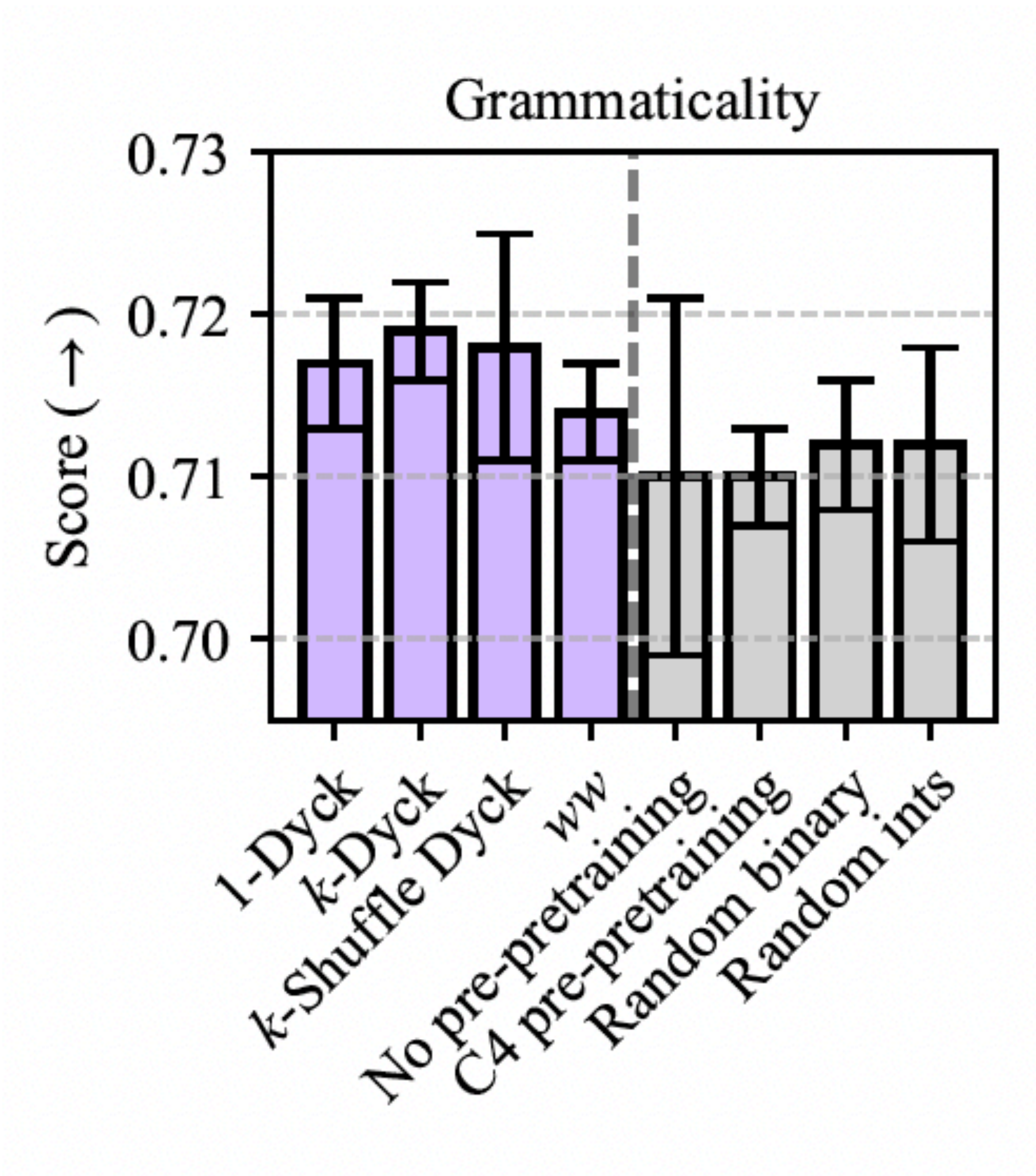
Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
BINDING	7	*Carlos said that Lori helped him.*	*Carlos said that Lori helped himself.*
CONTROL/RAISING	5	*There was bound to be a fish escaping.*	*There was unable to be a fish escaping.*

Is the probability of the acceptable example higher than the probability of the unacceptable one?

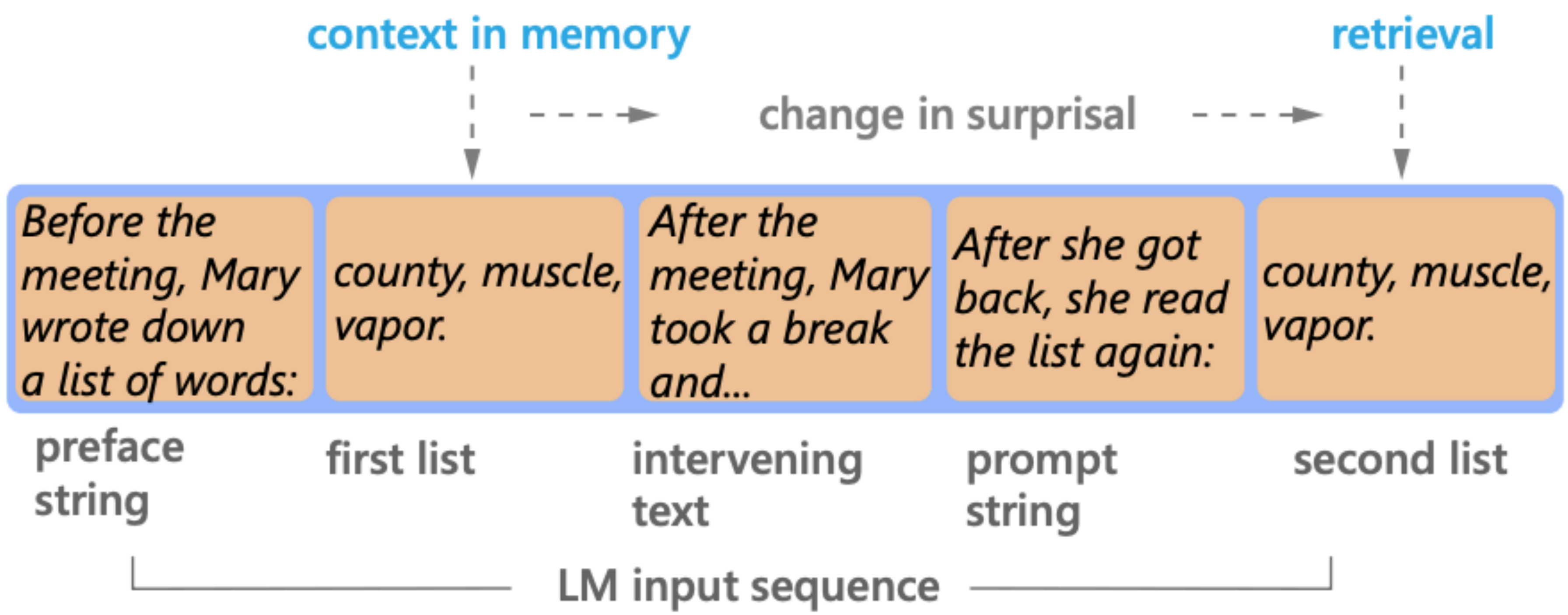
(Marvin & Linzen 2018, Warstadt et al 2019)

Targeted syntactic evaluation with minimal pairs (BLiMP)

Comparing models at the optimal amount of prepretraining for each setup:



Targeted evaluation: retrieval

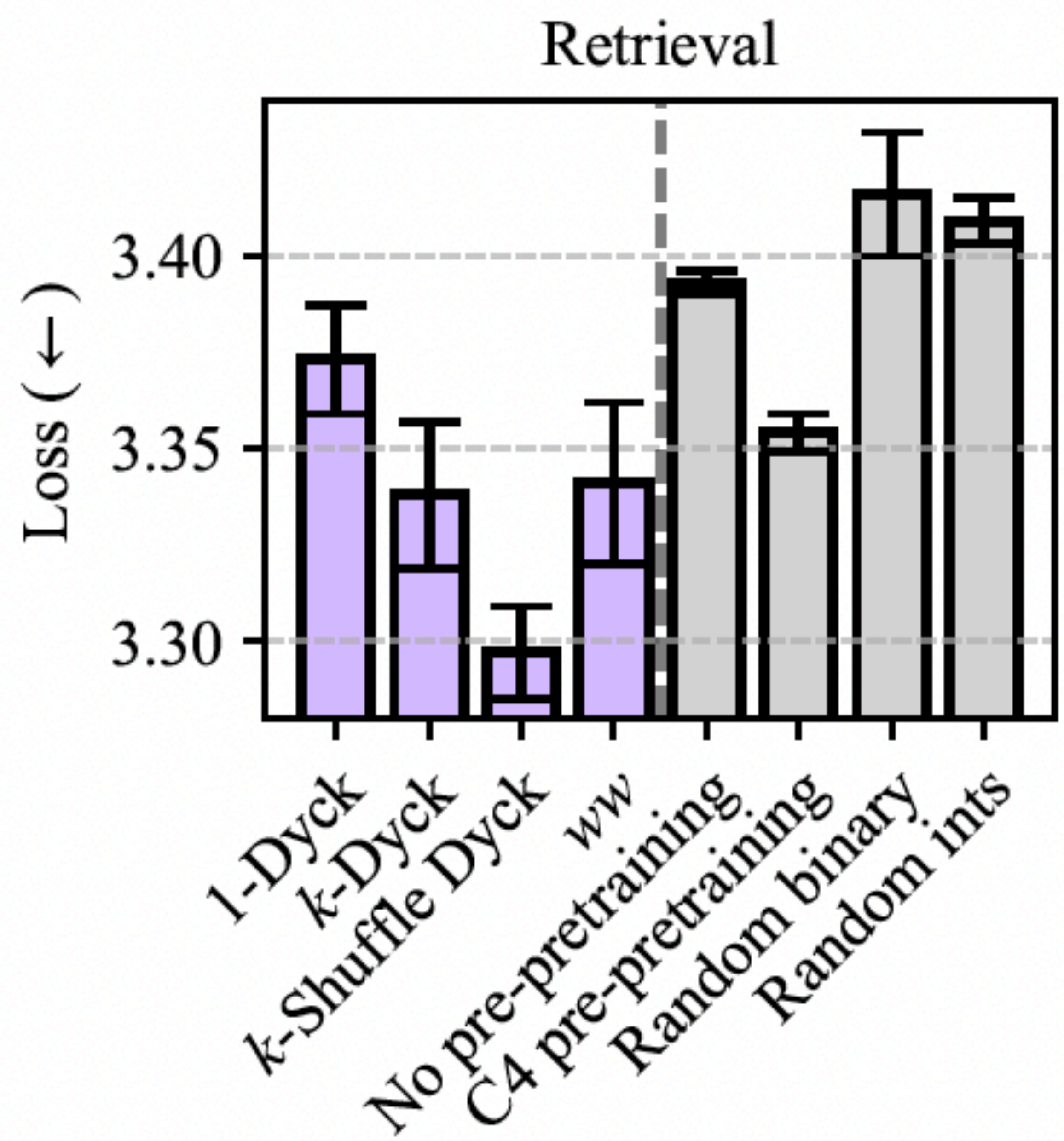


We expect the loss on the second repetition of the list to be lower

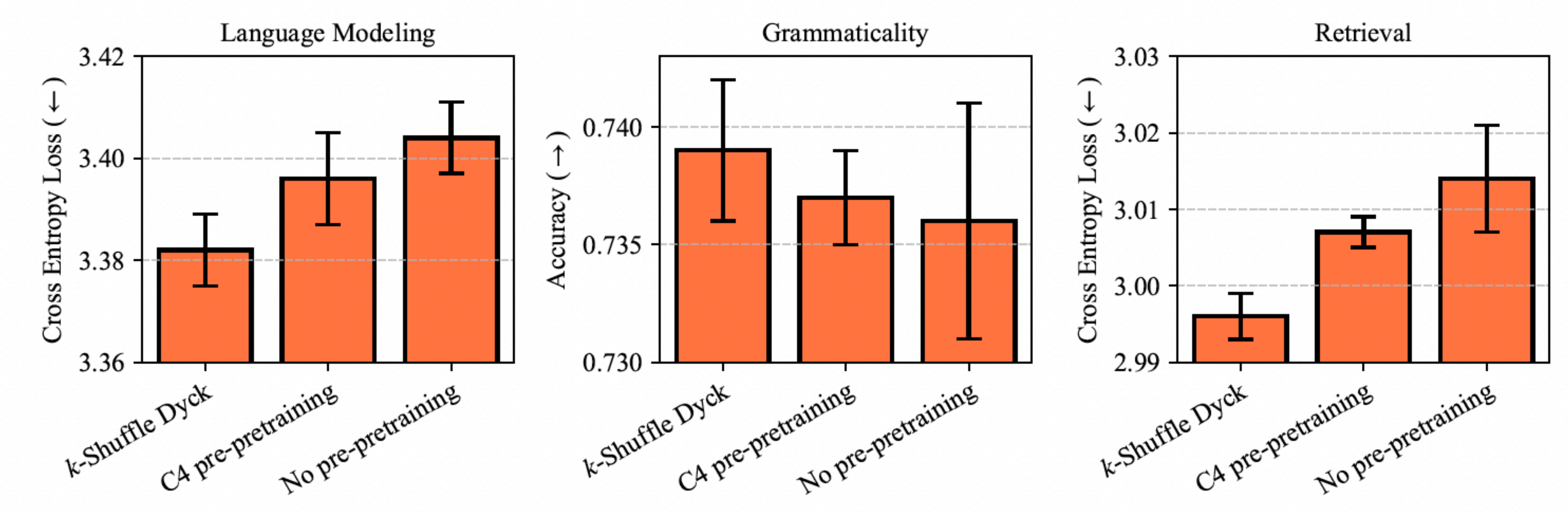
(Armeni, Honey & Linzen 2022)

Targeted evaluation: retrieval

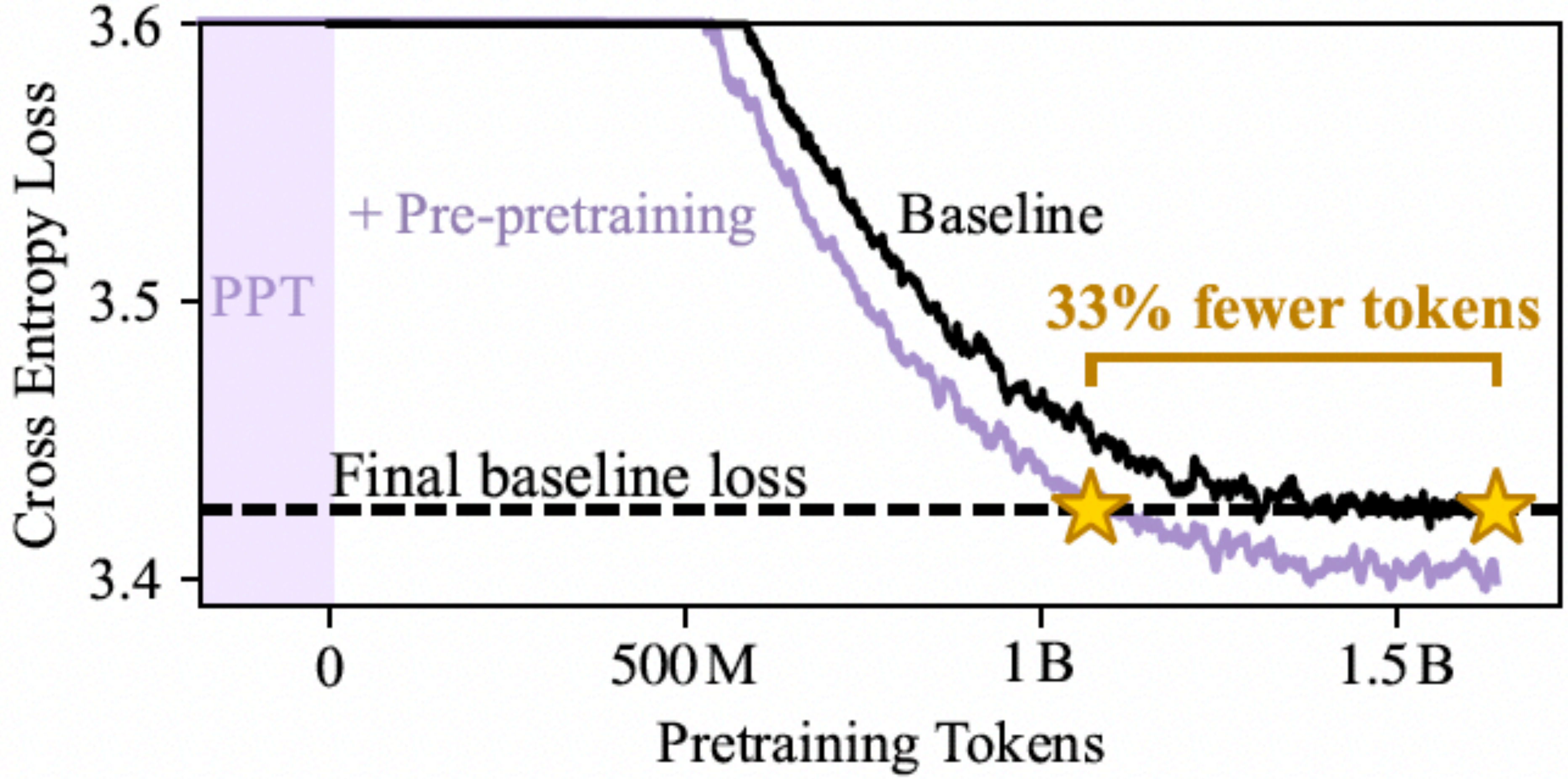
Comparing models at the optimal amount of pre-pretraining for each language



Scaling up to 1B parameters



Scaling up to 1B parameters



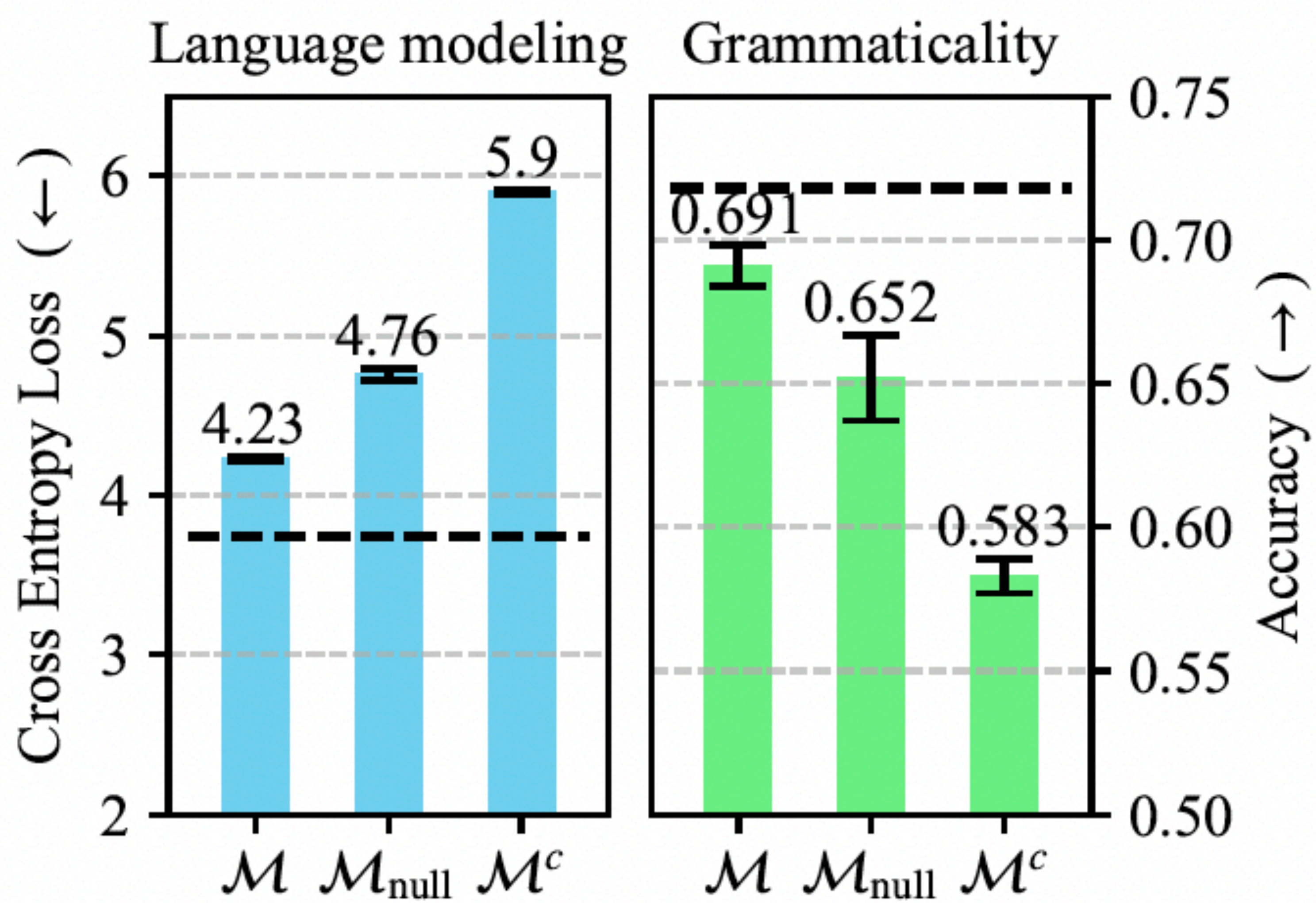
Mechanistic analysis

- Hypothesis: subnetworks learned through pre-pretraining are then reused to process natural language

- We prune 50% of the attention heads so as to minimize the impact on language modeling loss **on the formal language**

- Then we test for transfer to natural language

Mechanistic analysis



$\mathscr{M}$: the sparse subnetwork we find

$\mathscr{M}_{null}$: a random subnetwork of the same size

$\mathscr{M}^C$: the complement of $\mathscr{M}$

Pre-pretraining: conclusions

- Pre-pretraining on formal languages improves sample efficiency and generalization
- It is more efficient to pre-pretrain on formal language than on more natural language!
- The formal language needs to be match the structural complexity of natural language
- Some support for the computational compatibility hypothesis

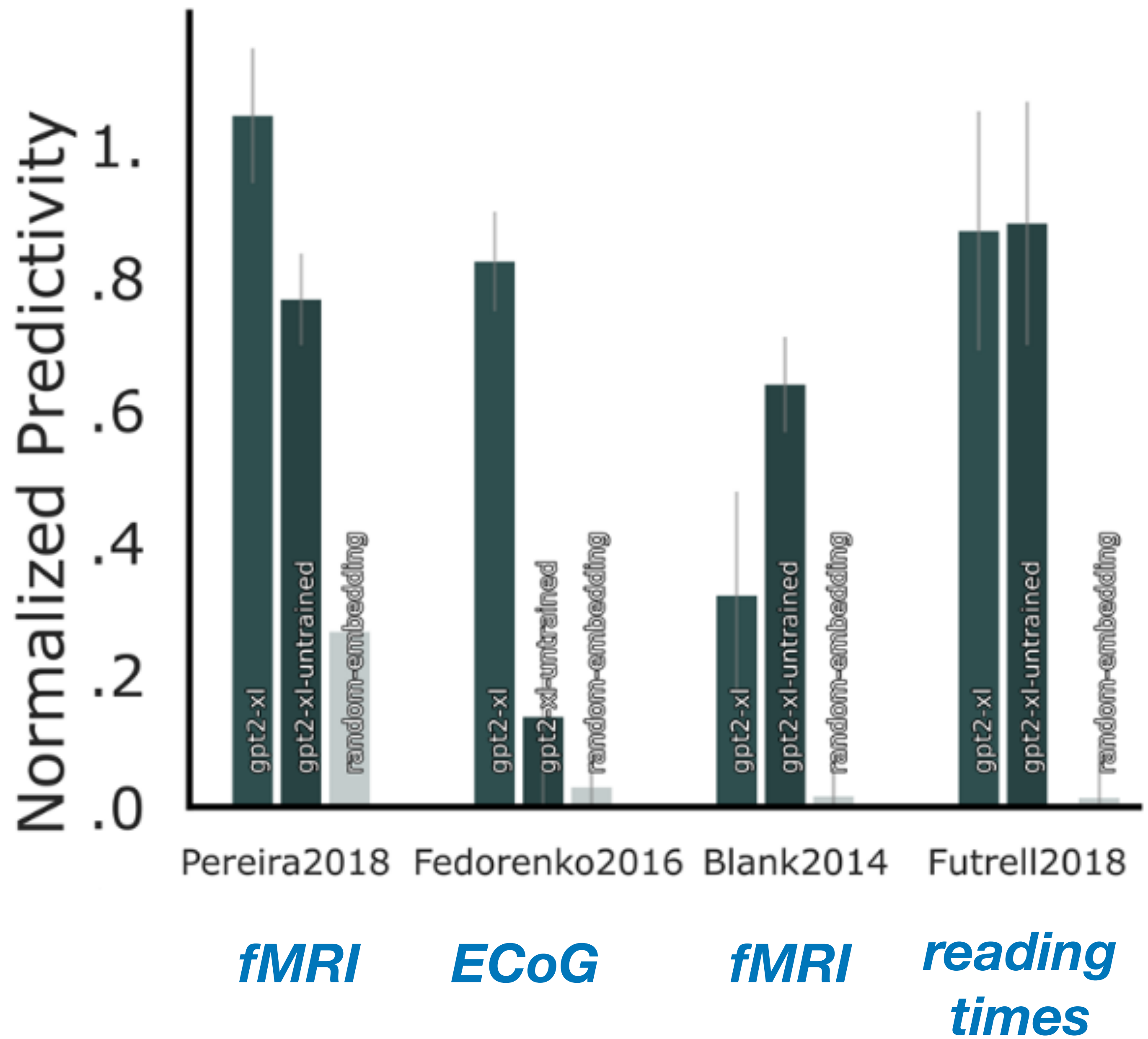
Future work

	Context-free	Context-sensitive
C-RASP	1-Dyck	k-Shuffle Dyck
FO(M)	k-Dyck	ww

- We need more languages in each cell of the table to draw clearer conclusions
- The computational compatibility hypothesis leads us to expect other architectures to show different patterns: e.g., we can extend to RNNs

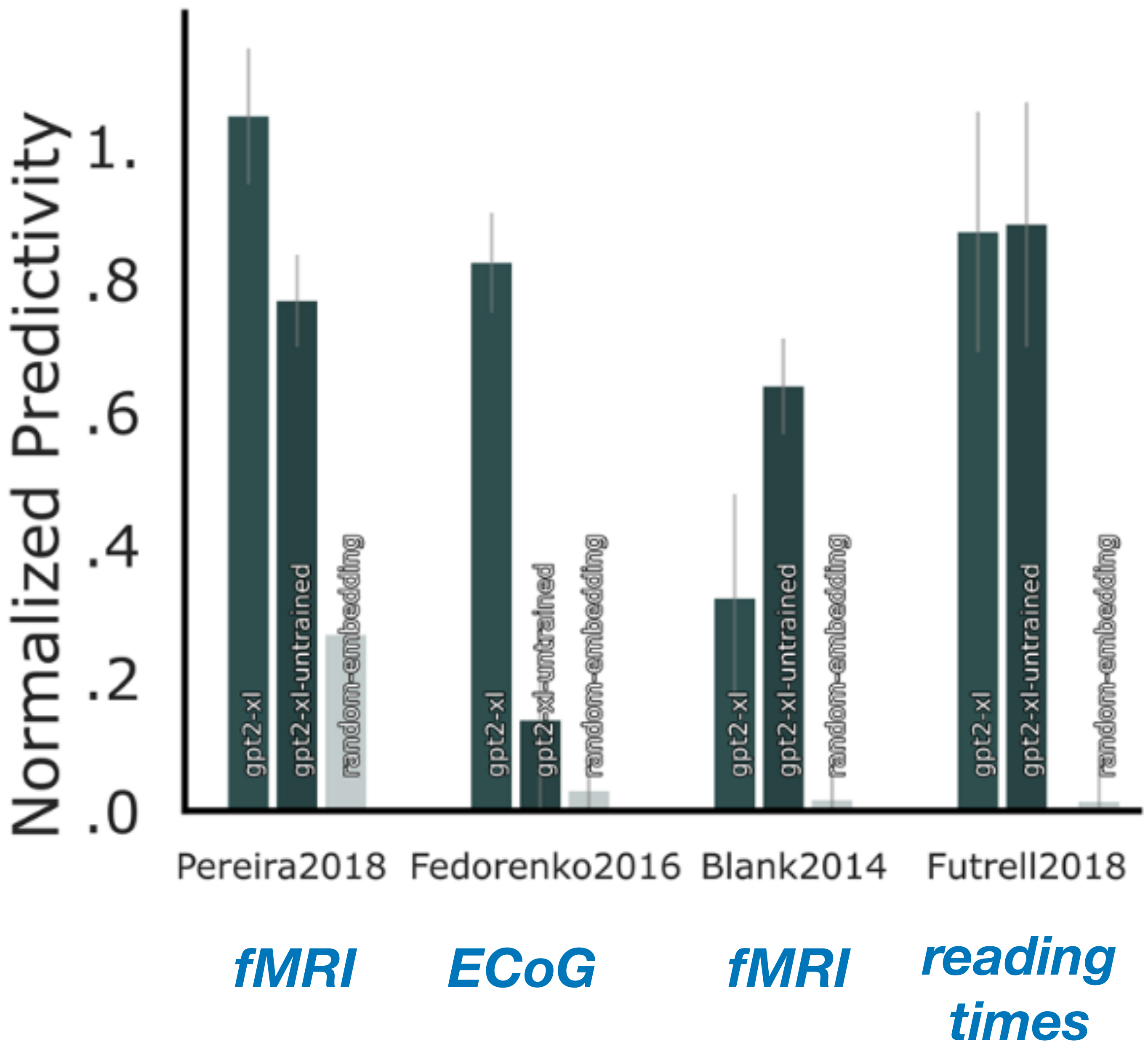
Word prediction in LLMs and humans

Deep learning next-word prediction models explain most of the explainable variance in brain activity and reading times!



(Schrimpf et al., 2021, PNAS)

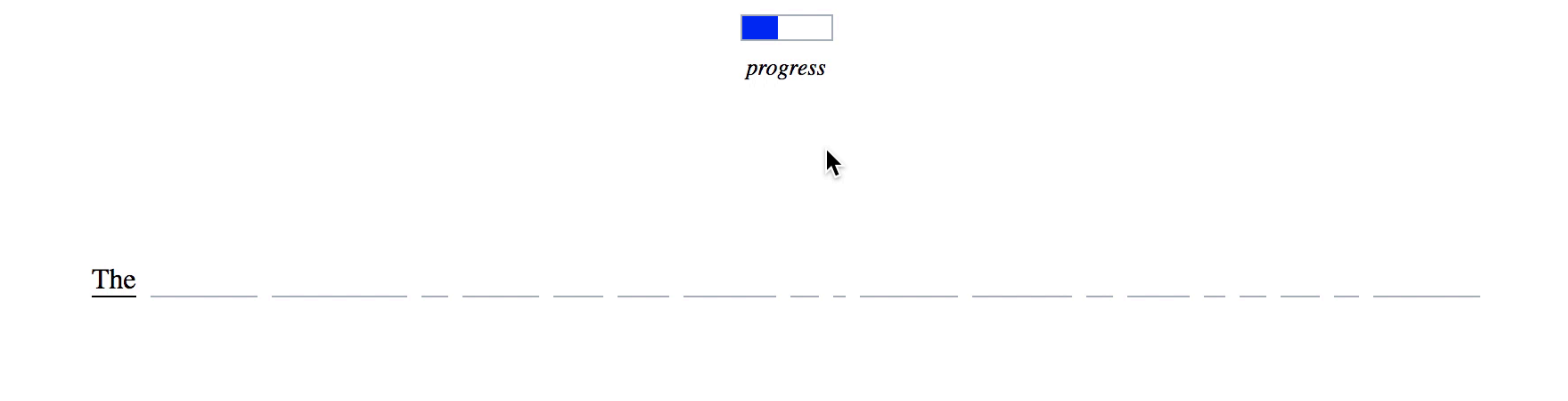
Can we use them to explain how people process syntactically complex sentences?



(Schrimpf et al., 2021, PNAS)

Before the woman visited the famous doctor had been drinking.

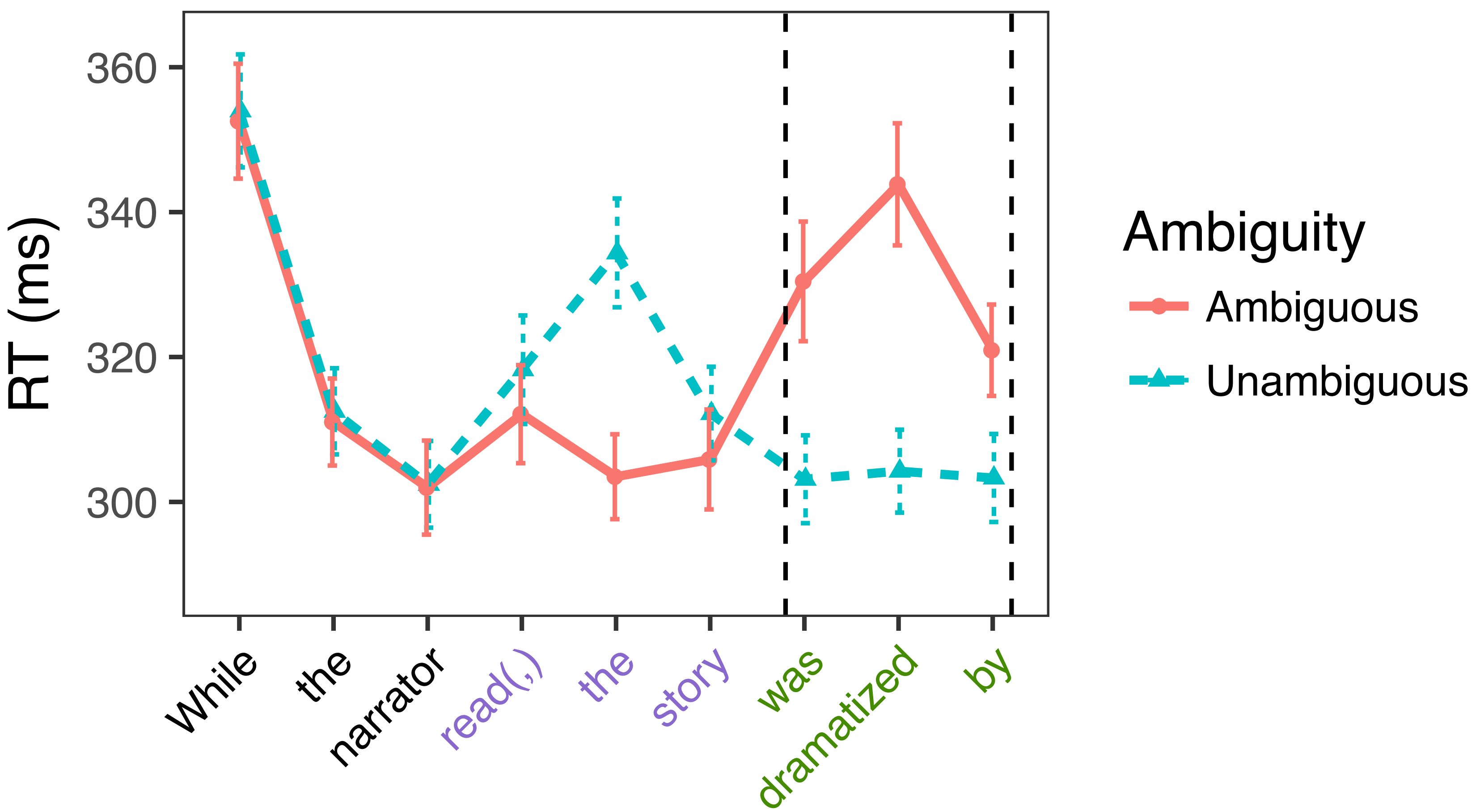
Self-paced reading with a moving window



Did the journalist meet an actress?

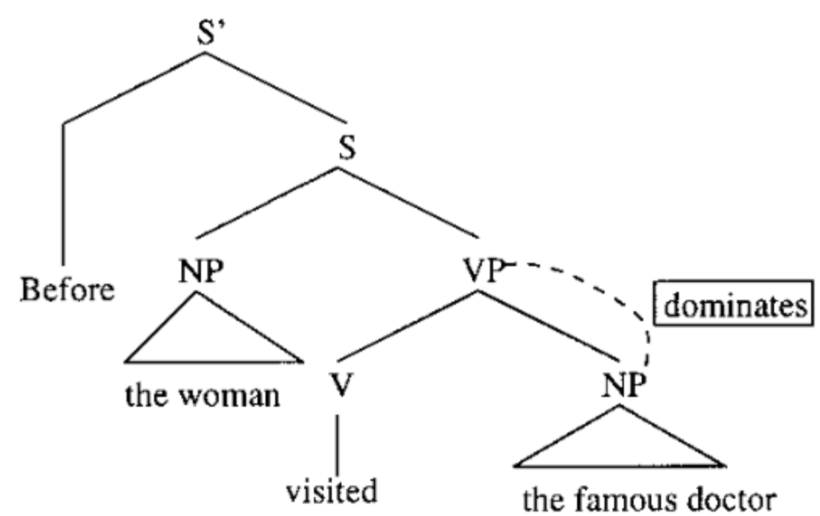
While the narrator read the story **was dramatized by** the actors.

While the narrator read, the story **was dramatized by** the actors.



Reanalysis in a serial parser

Before the woman visited the famous doctor had been drinking.



(Frazier, 1987;
Sturt et al. , 1999)

Alternative account: predictability

weaker neural responses

read faster

The children went outside to play.

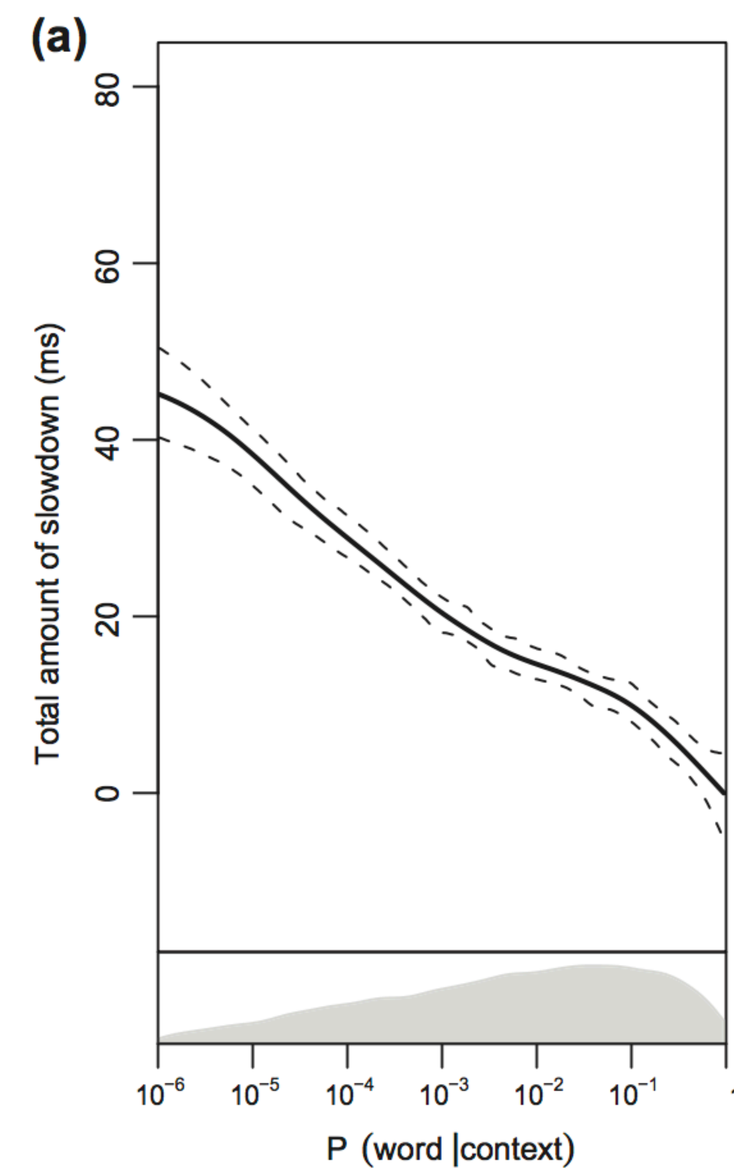
The professor went home to play.

read more slowly

stronger neural responses

(Ehrlich and Rayner, 1981; Kutas and Hillyard, 1984)

Surprisal



(Smith & Levy, 2013)

The surprisal account of garden path sentences

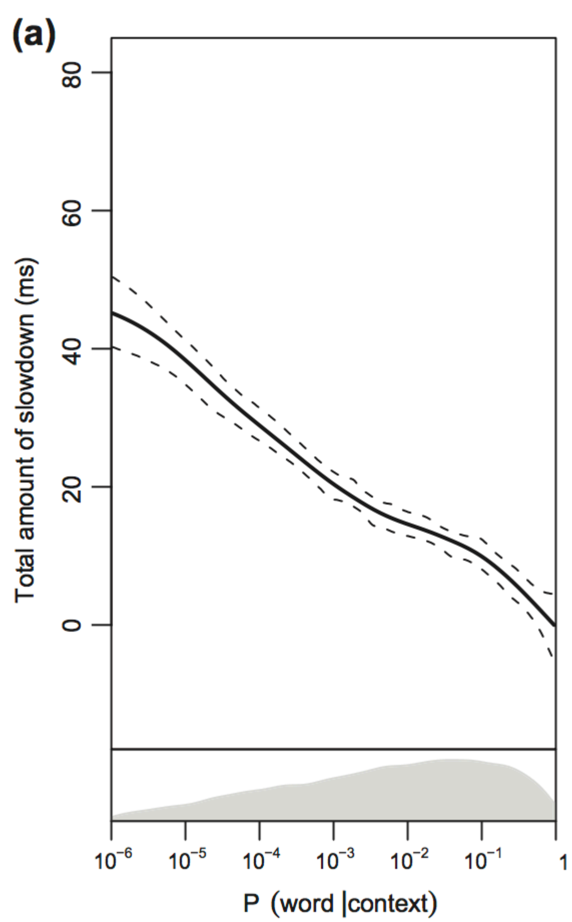
Even though the girl phoned the instructor
was very upset with her for missing a lesson.

Unpredictable!
(Hale, 2001; Levy, 2013)

Parsimonious explanation!

Quantitative test of surprisal

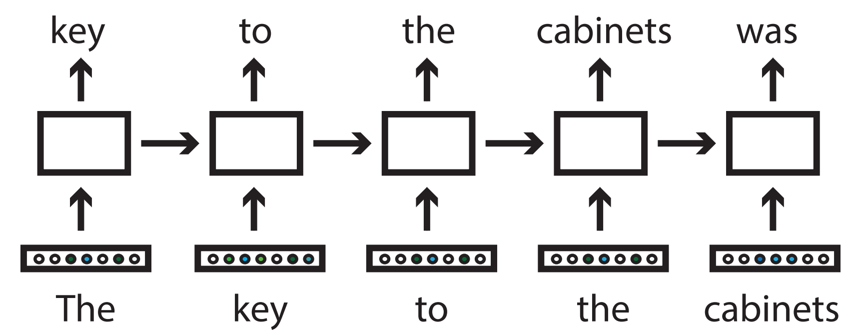
1. Using filler items from a self-paced reading study, estimate the slowdown δ that readers experience for each unit (bit) of surprisal



(van Schijndel & Linzen, 2021, Cognitive Science)

2. Estimate surprisal from a word prediction model trained on a text corpus:

$$-log_2\hat{P}(w_n = w^k \,|\, w_1, \ldots, w_{n-1})$$



3. The predicted magnitude of the garden path effect is δ times the difference in surprisal across contexts:

The employees understood the contract **would be changed** very soon.

The employees understood **that** the contract **would be changed** very soon.

The Syntactic Ambiguity Processing (SAP) Benchmark

- 2000 self-paced reading subjects and ~350 eye tracking subjects, each reading:
  - Garden path constructions: MV/RR, NP/S, and NP/Z
  - Subject-gap vs. object-gap relative clauses (Staub, 2010)
  - Relative clause attachment ambiguities (Dillon et al., 2019)
  - Outright agreement errors

(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, JML)

Comparing matched regions in ambiguous and unambiguous sentences

The girl fed the lamb ***remained relatively calm*** ...

The girl who was fed the lamb ***remained relatively calm*** ...

The girl found the lamb ***remained relatively calm*** ...

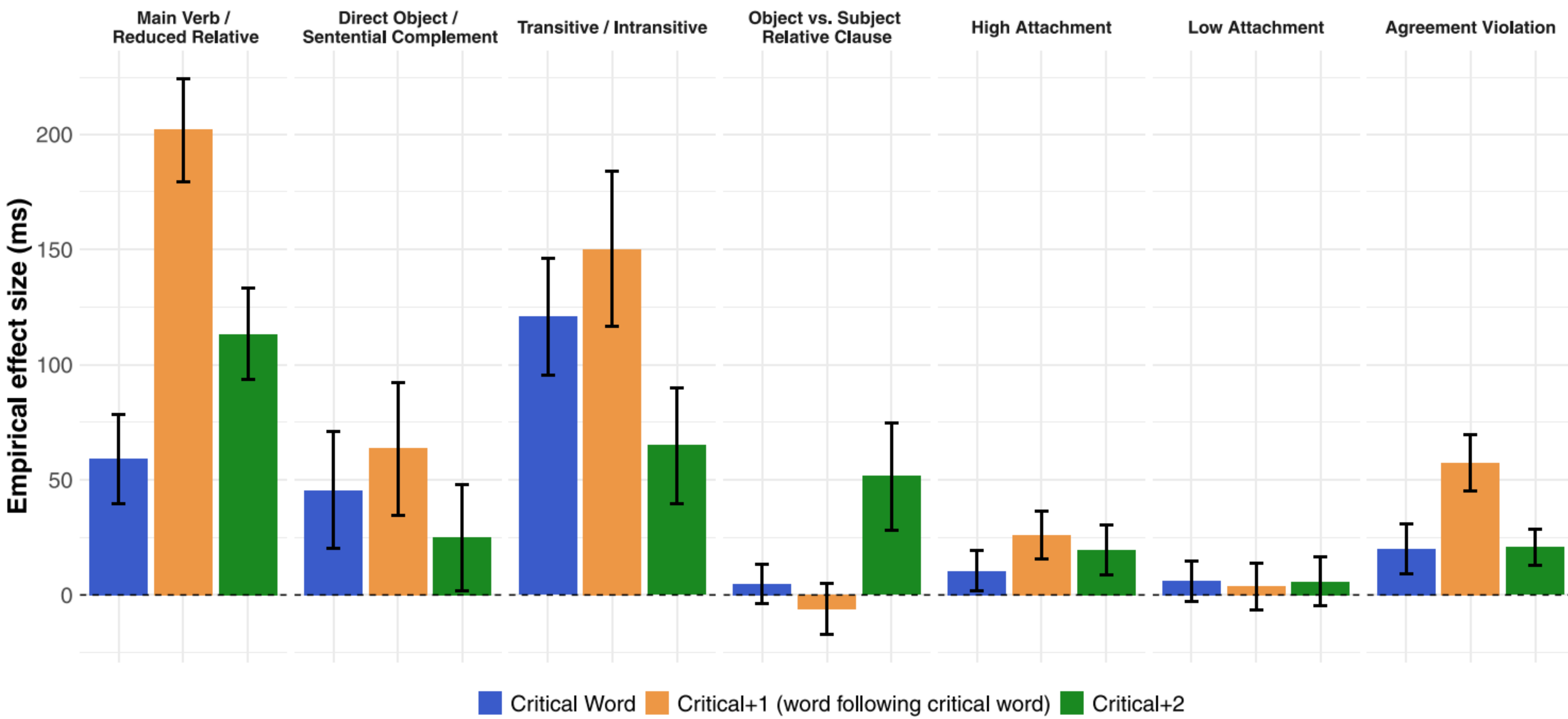
The girl found that the lamb ***remained relatively calm*** ...

When the girl attacked the lamb ***remained relatively calm*** ...

When the girl attacked, the lamb ***remained relatively calm*** ...

(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, JML)

Results: human reading times



(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, JML)

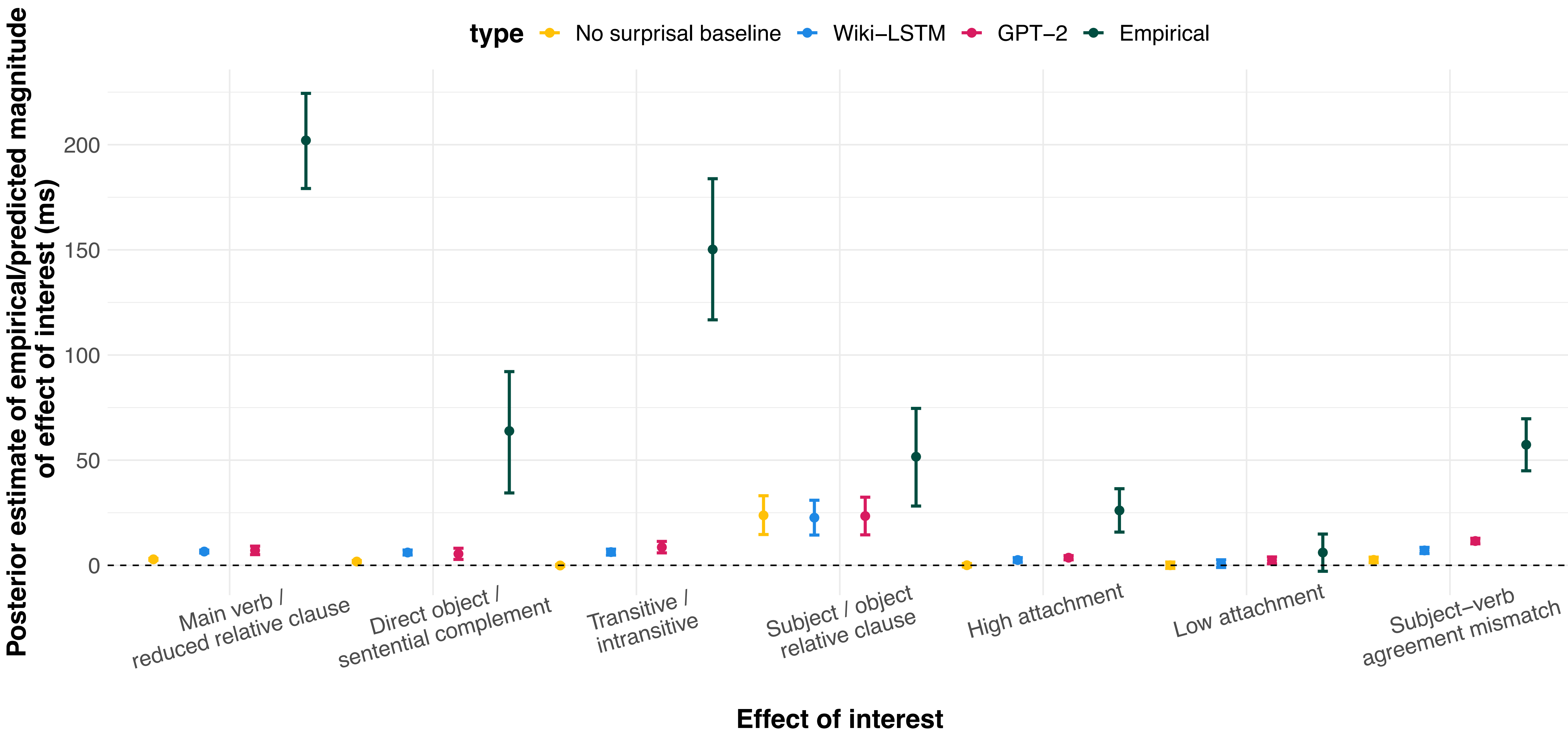
Language models

- GPT-2 small: a transformer LM trained by OpenAI on 40 GB of web data (~5-10 billion words)

- LSTM: trained by Gulordava et al. (2018) on a Wikipedia corpus of around 80M words

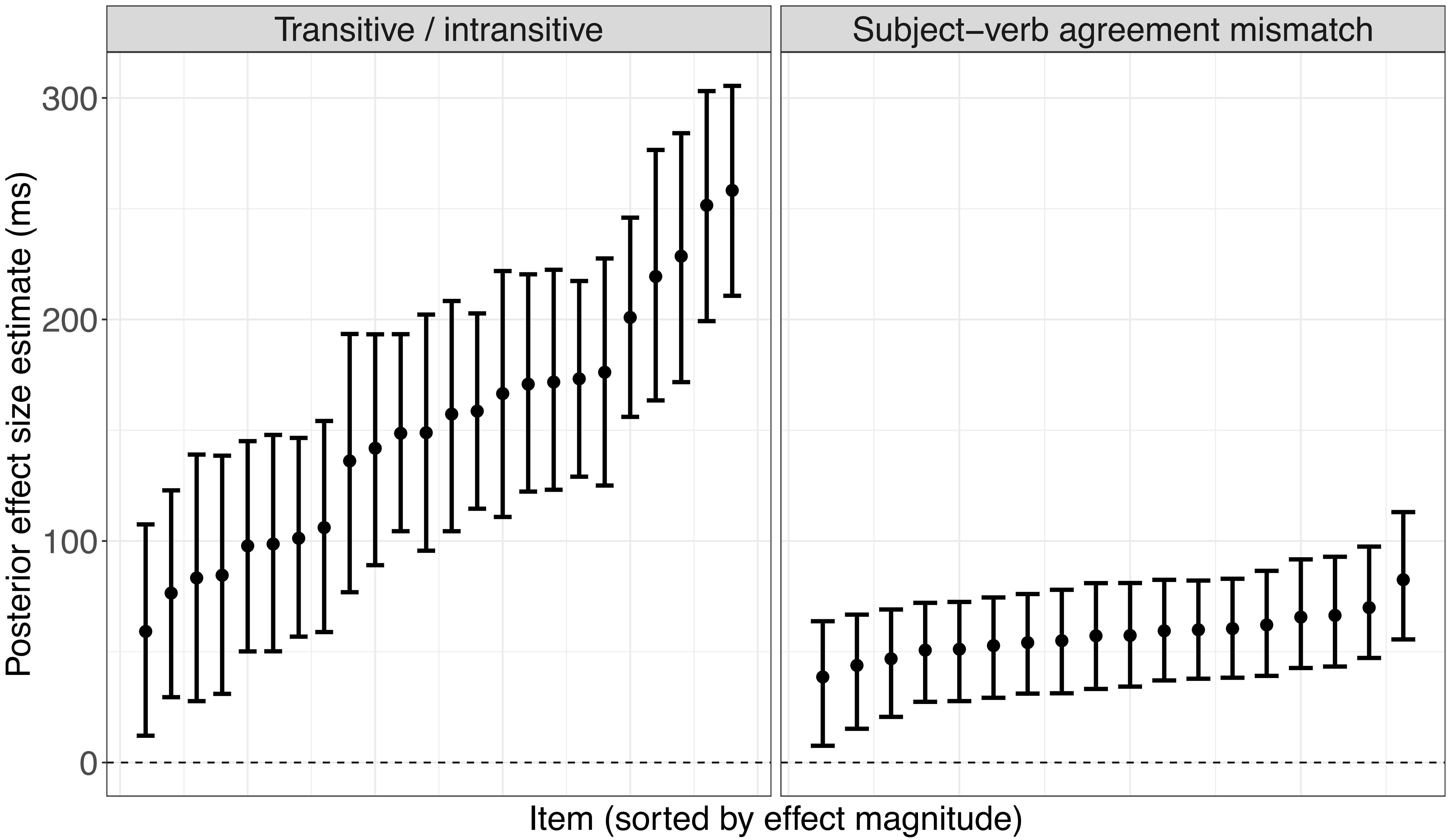
(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, JML)

Condition mean estimates: comparison to language model predictions

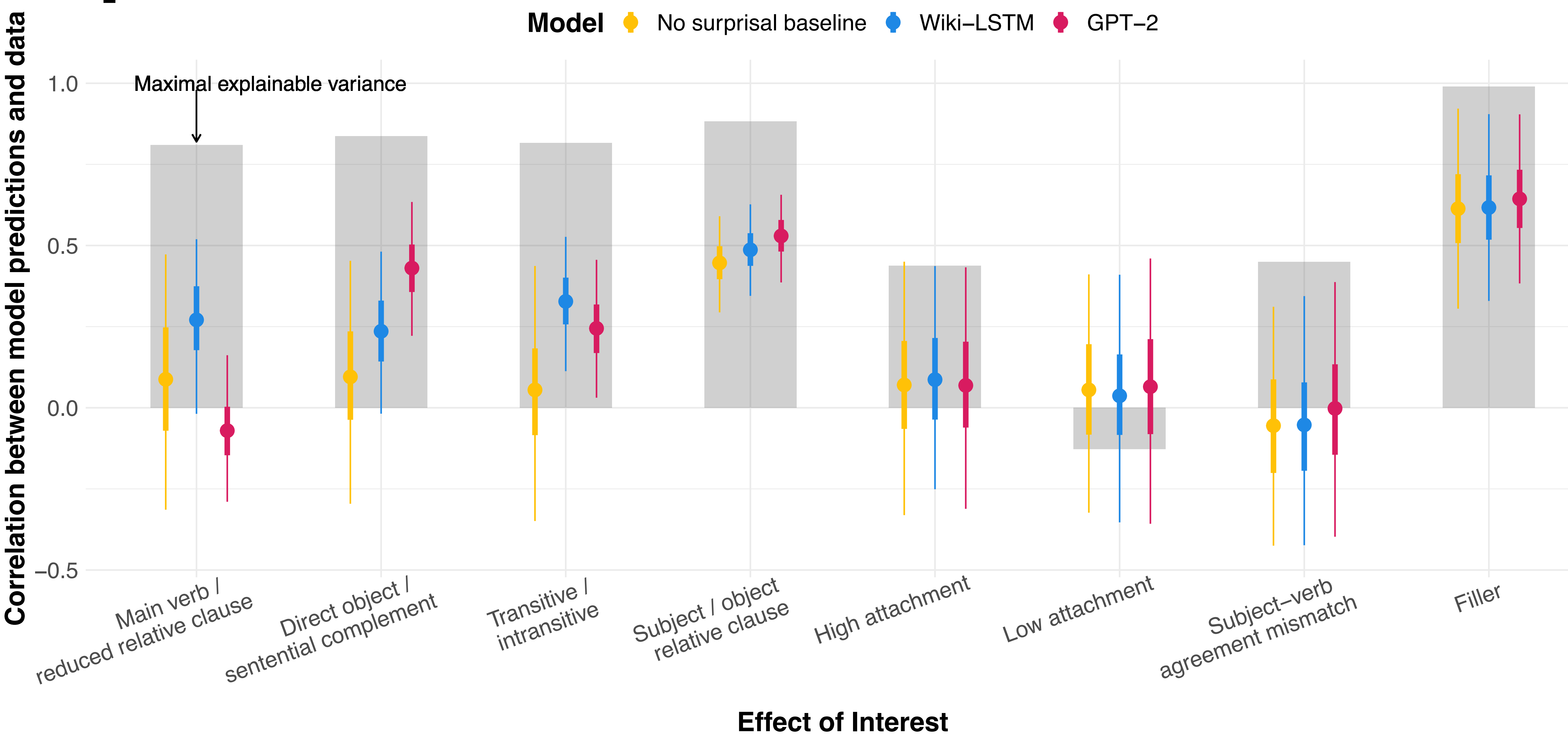


(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, Journal of Memory and Language)

We have enough data to compute meaningful item-level estimates



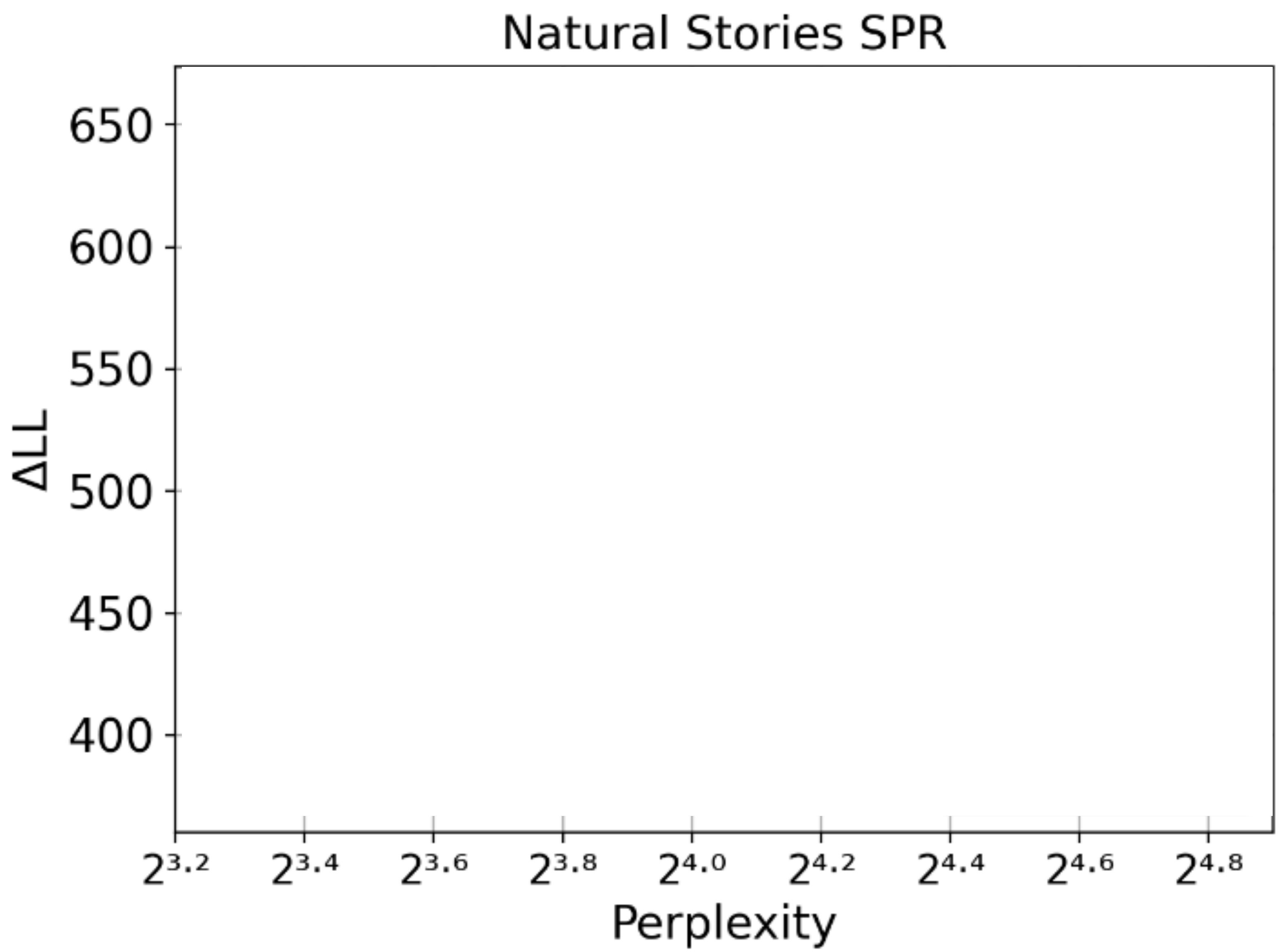
Can our language models predict item-wise variation?



Could language models that make more accurate word predictions help?

better fit to
human reading
times

worse fit to
human reading
times



more accurate
word predictions

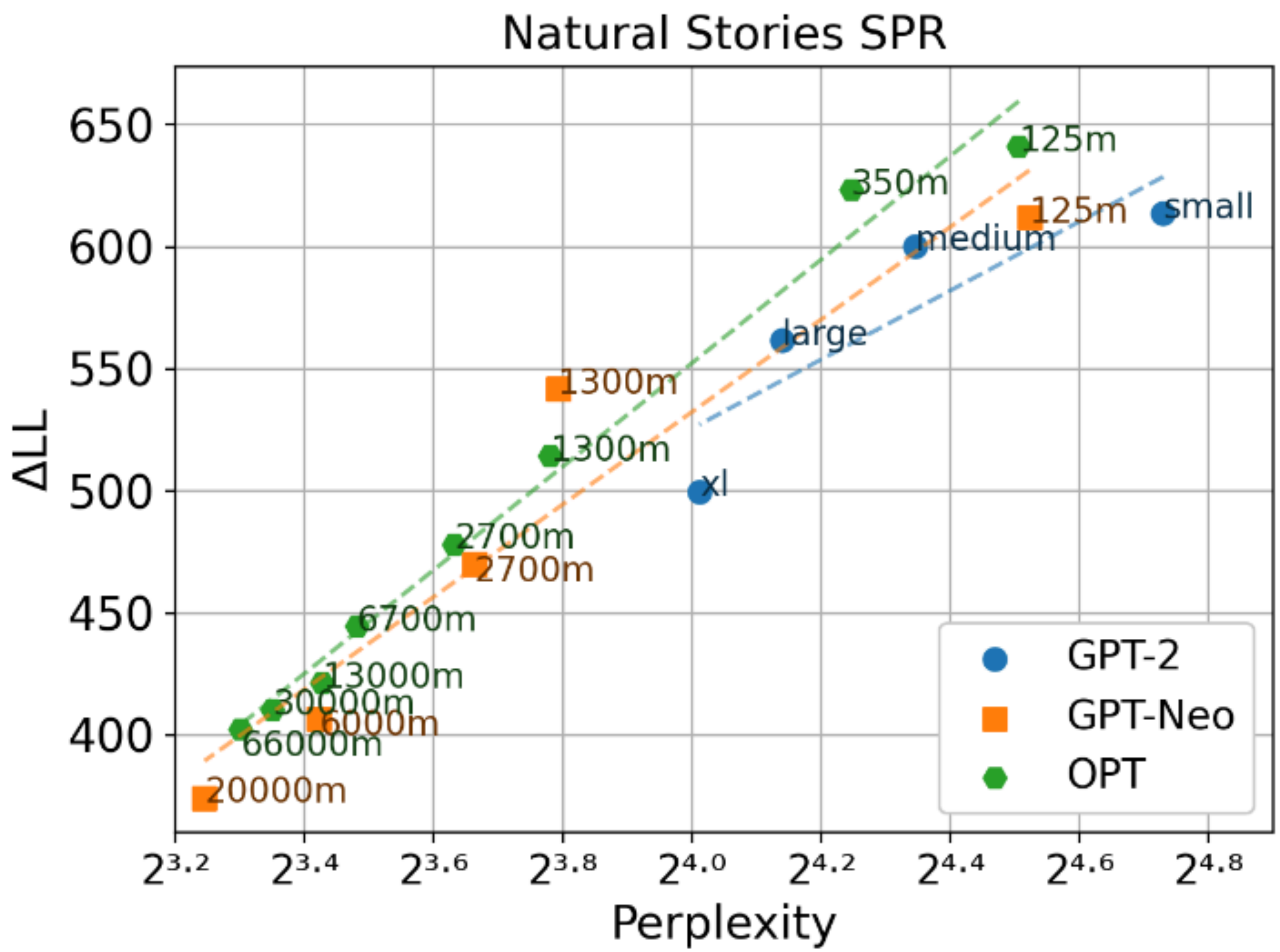
less accurate
word predictions

(Oh & Schuler, 2022)

Could language models that make more accurate word predictions help?

better fit to human reading times

worse fit to human reading times



more accurate word predictions

less accurate word predictions

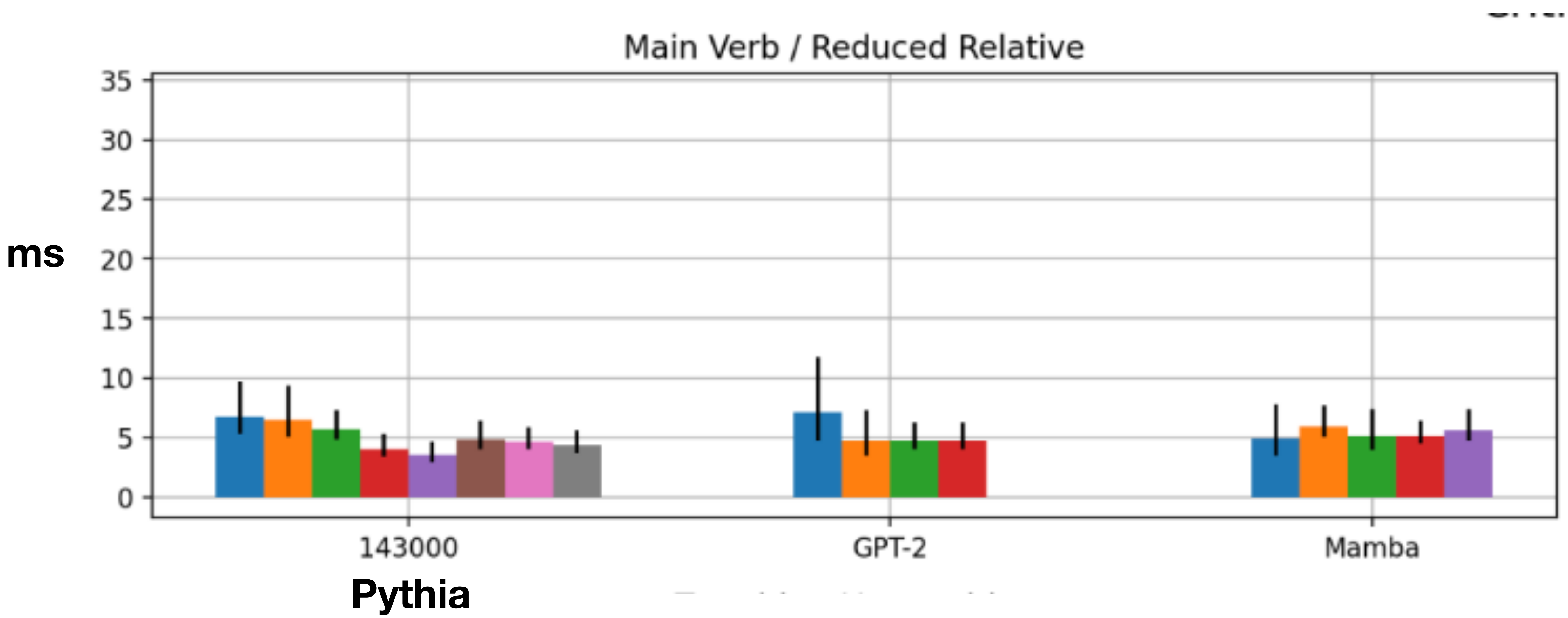
(Oh & Schuler, 2022)

Could language models that make more accurate word predictions help? Testing bigger LMs

- GPT-2 small, medium, large and extra-large (transformer)
- Pythia models of sizes 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B and 12B (transformers as well)
- Mamba of sizes 130M, 370M, 790M, 1.4B, 2.8B (state space models)

(Huang, Arehalli, Kugemoto, Muxica, Prasad, Dillon & Linzen, 2024, JML)

Could language models that make more accurate word predictions help? Testing bigger LMs



Pythia: 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B
GPT-2 S, M, L, XL
Mamba 130M, 370M, 790M, 1.4B, 2.8B

(Oh, Dillon and Linzen, in prep)

Conclusions: Word prediction and cognitive modeling

- To more closely mimic human processing, we will likely need models that are resource-limited in human-like ways (Timkey and Linzen 2023, Warstadt et al. 2023):
    - Trained on a cognitively-plausible corpus
    - Only consider a small number of interpretations concurrently
    - Have limited working memory
- These are not necessarily going to be the models developed by OpenAI etc: cognitive scientists need to train models ourselves