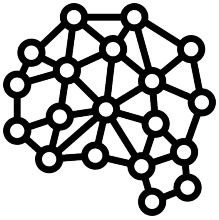
What can linguistic data tell us about the predictions of large language models?

Byung-Doh Oh

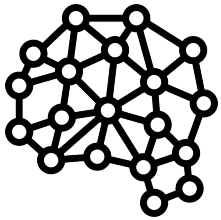
Center for Data Science
New York University

October 7, 2024
POSTECH CSE/AI



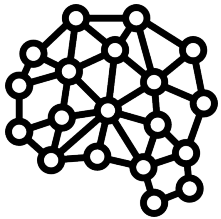
Language models

After this talk, I'll $\rightarrow$



Language models

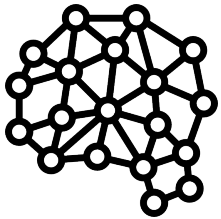
After this talk, I'll $\rightarrow$



Language models

$\rightarrow$ { *chat*, *meet*, *cry*, ...

After this talk, I'll $\rightarrow$

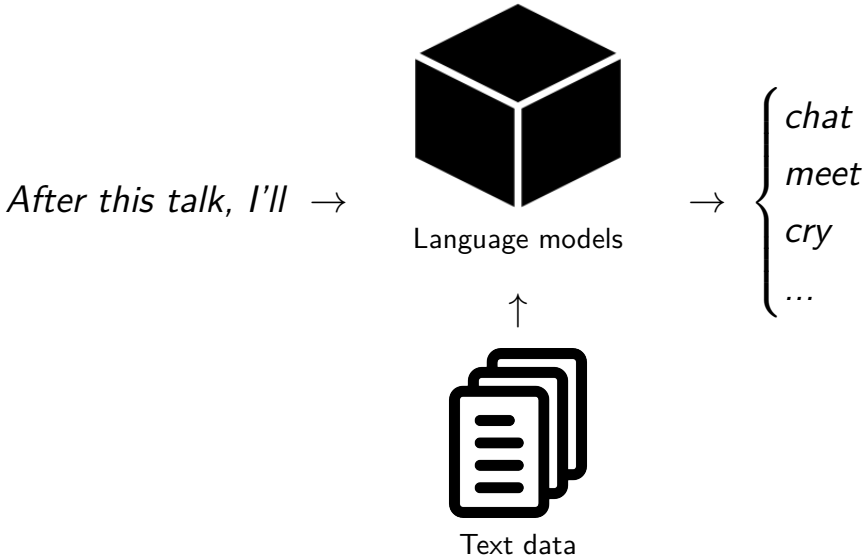


Language models

$\rightarrow$ { *chat*, *meet*, *cry*, ...

$\uparrow$

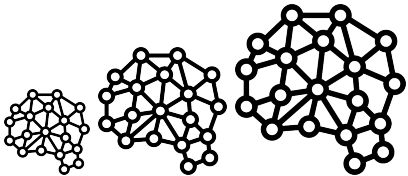
Text data


After this talk, I'll →
Language models
→
chat
meet
cry
...
Text data

Some open questions in the field

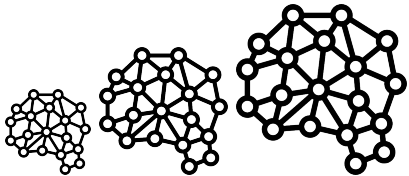
Some open questions in the field

Scaling behavior

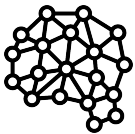


Some open questions in the field

Scaling behavior



Feature attribution

After this talk, I'll →  → *chat*

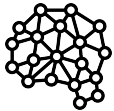
?

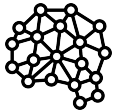
Outline

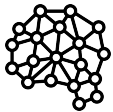
Outline

Reading time data

$\sim$



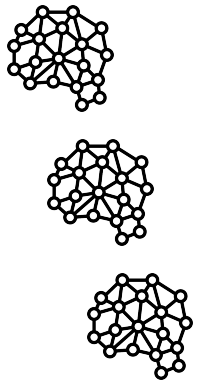




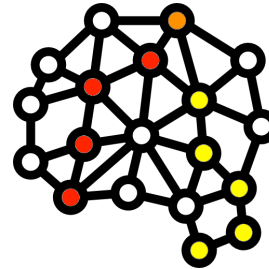
Outline

Reading time data

$\sim$



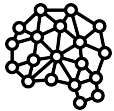
Linguistic annotations

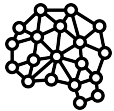


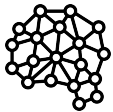
Outline

Reading time data

$\sim$







People make predictions too

After this talk, I'll

People make predictions too

After this talk, I'll

- chat
- cry

People make predictions too

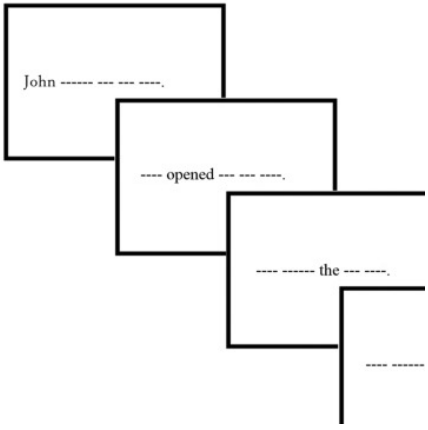
After this talk, I'll → chat / cry

The more predictable chat is easier to process than cry

(Balota et al., 1985; Ehrlich & Rayner, 1981; Kutas & Hillyard, 1980)

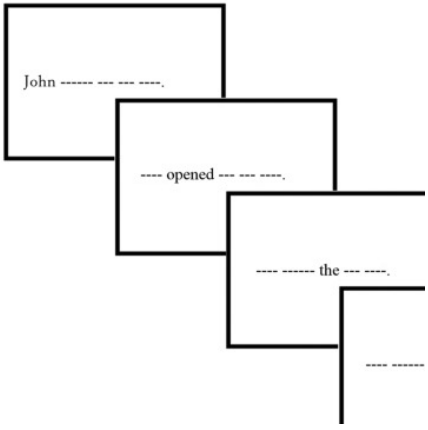
This shows up in reading time data

This shows up in reading time data



Self-paced reading

This shows up in reading time data



Self-paced reading

Eye-tracking

This shows up in reading time data

Text	After	this	talk,	I'll	chat

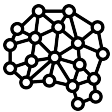
This shows up in reading time data

Text	After	this	talk,	I'll	chat
ReadingTime	718 ms	709 ms	847 ms	766 ms	886 ms

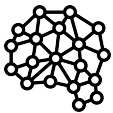
Link between reading time data and LMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

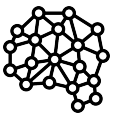
$\sim$



LM 1



LM 2

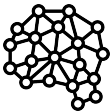


LM 3

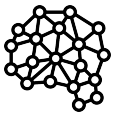
Link between reading time data and LMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

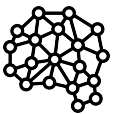
$\sim$



LM 1



LM 2



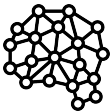
LM 3

Reading time of $w_t \propto \underbrace{-\log_2 \mathrm{P}(w_t \mid w_{1..t-1})}_{\text{surprisal}}$

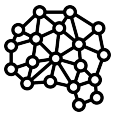
Link between reading time data and LMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

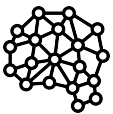
$\sim$



LM 1



LM 2



LM 3

Reading time of *chat* $\propto -\log_2 \mathrm{P}(\text{chat} \mid \text{After this talk, I'll})$

Reading time of *cry* $\propto -\log_2 \mathrm{P}(\text{cry} \mid \text{After this talk, I'll})$

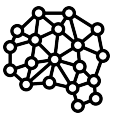
Link between reading time data and LMs (Surprisal theory; Hale, 2001; Levy, 2008)

Text	After	this	talk,	I'll	chat
ReadingTime	718 ms	709 ms	847 ms	766 ms	886 ms
$\text{Surprisal}_{\text{LM1}}$	4.95	1.32	6.40	3.04	3.39
$\text{Surprisal}_{\text{LM2}}$	3.53	0.69	5.73	1.14	1.37
$\text{Surprisal}_{\text{LM3}}$	3.50	0.59	5.13	0.63	0.79

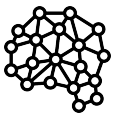
Human subjects

$\sim$

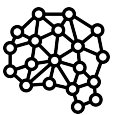
Regression modeling



LM 1



LM 2



LM 3

Methods: Evaluation of LM surprisal

Methods: Evaluation of LM surprisal

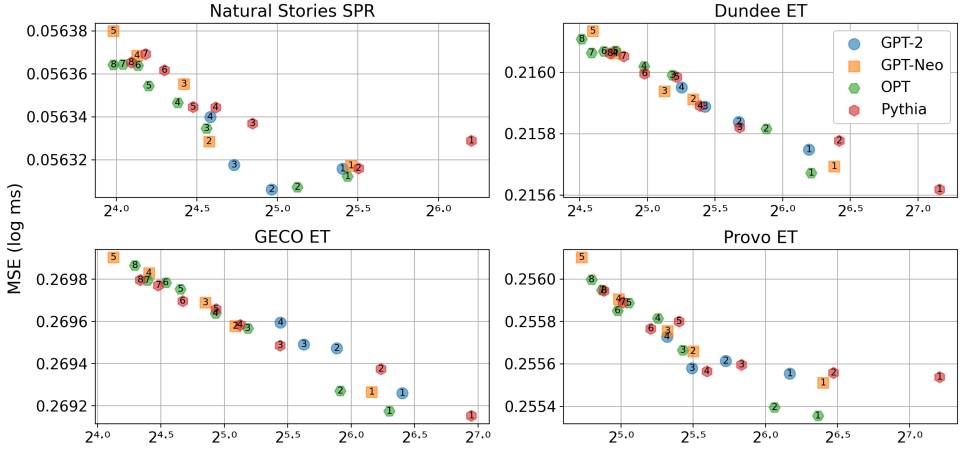
Transformer LMs	
GPT-2 Small	GPT-Neo 125M
GPT-2 Medium	GPT-Neo 1.3B
GPT-2 Large	GPT-Neo 2.7B
GPT-2 XL	GPT-J 6B
	GPT-NeoX 20B
OPT 125M	Pythia 70M
OPT 350M	Pythia 160M
OPT 1.3B	Pythia 410M
OPT 2.7B	Pythia 1B
OPT 6.7B	Pythia 1.4B
OPT 13B	Pythia 2.8B
OPT 30B	Pythia 6.9B
OPT 66B	Pythia 12B

Methods: Evaluation of LM surprisal

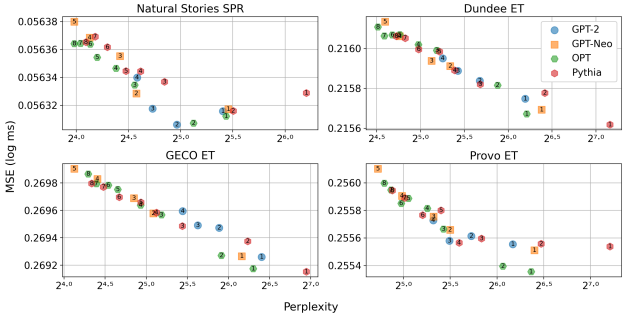
Transformer LMs	
GPT-2 Small	GPT-Neo 125M
GPT-2 Medium	GPT-Neo 1.3B
GPT-2 Large	GPT-Neo 2.7B
GPT-2 XL	GPT-J 6B
	GPT-NeoX 20B
OPT 125M	Pythia 70M
OPT 350M	Pythia 160M
OPT 1.3B	Pythia 410M
OPT 2.7B	Pythia 1B
OPT 6.7B	Pythia 1.4B
OPT 13B	Pythia 2.8B
OPT 30B	Pythia 6.9B
OPT 66B	Pythia 12B

Reading times	Genre	Data points
Natural Stories (SPR)	Wikipedia	384,905
Dundee (ET)	Newspaper	98,115
GECO (ET)	Novel	144,850
Provo (ET)	Short stories	52,960

Larger LMs are more accurate, but less human-like

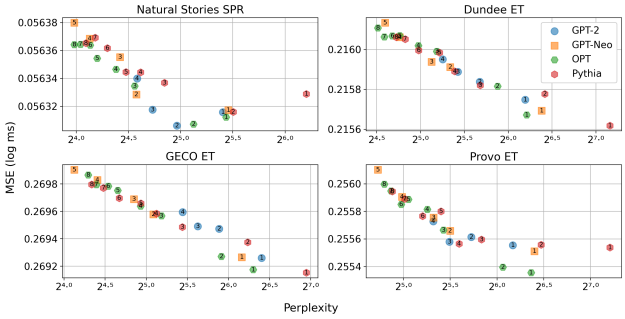


Larger LMs are more accurate, but less human-like



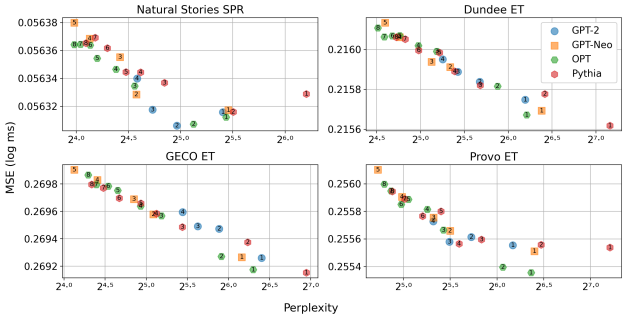
Larger LMs are more accurate, but less human-like

- Systematic divergence due to model size



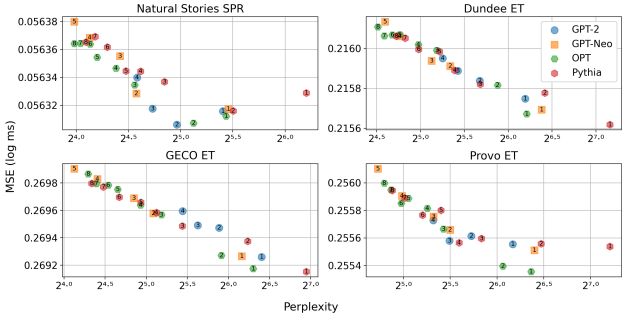
Larger LMs are more accurate, but less human-like

- Systematic divergence due to model size
- What can this tell us about the predictions LMs learn?



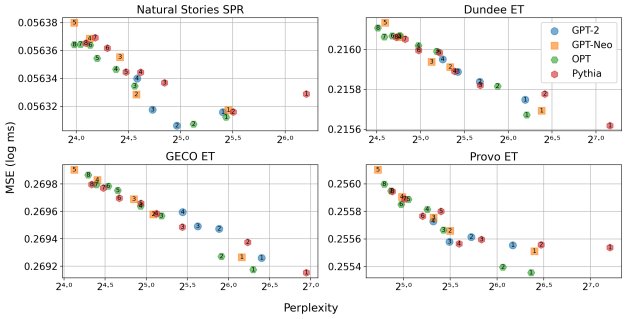
Larger LMs are more accurate, but less human-like

- Models: Larger models can learn with fewer gradient updates (Tirumala et al., 2022)



Larger LMs are more accurate, but less human-like

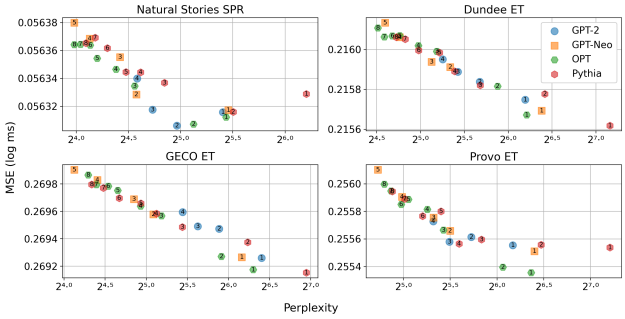
- Models: Larger models can learn with fewer gradient updates (Tirumala et al., 2022)
- Humans: Rare words are difficult to process (Shain, 2024)

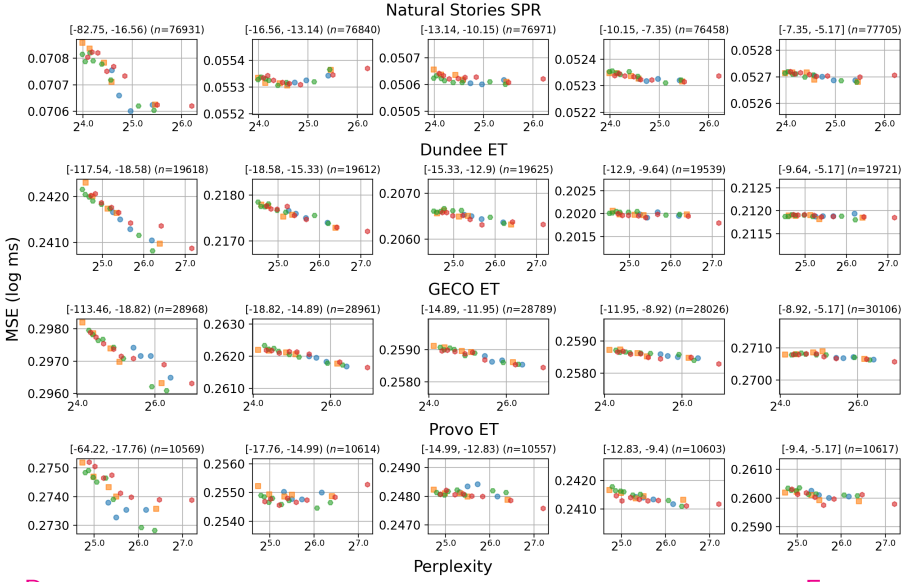


Larger LMs are more accurate, but less human-like

- Models: Larger models can learn with fewer gradient updates (Tirumala et al., 2022)
- Humans: Rare words are difficult to process (Shain, 2024)

→ *Suggests the strong role of word frequency*




Natural Stories SPR
Dundee ET
GECO ET
Provo ET
[-82.75, -16.56) (n=76931)
[-16.56, -13.14) (n=76840)
[-13.14, -10.15) (n=76971)
[-10.15, -7.35) (n=76458)
[-7.35, -5.17] (n=77705)
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
[-113.46, -18.82) (n=28968)
[-18.82, -14.89) (n=28961)
[-14.89, -11.95) (n=28789)
[-11.95, -8.92) (n=28026)
[-8.92, -5.17] (n=30106)
[-64.22, -17.76) (n=10569)
[-17.76, -14.99) (n=10614)
[-14.99, -12.83) (n=10557)
[-12.83, -9.4) (n=10603)
[-9.4, -5.17] (n=10617)
MSE (log ms)
Perplexity
Rare Words
Frequent Words

What gives larger LMs the advantage on rare words?

What gives larger LMs the advantage on rare words?

We know that:

What gives larger LMs the advantage on rare words?

We know that:

- Transformers can repeat input tokens (cf. RNNs): $B \ldots B$

What gives larger LMs the advantage on rare words?

We know that:

- Transformers can repeat input tokens (cf. RNNs): $B \ldots B$
- Transformers can induce bigram patterns (Elhage et al., 2021): $AB \ldots AB$

What gives larger LMs the advantage on rare words?

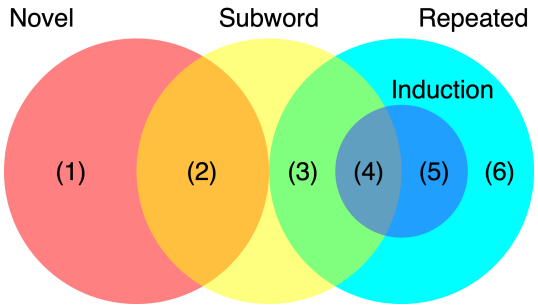
We know that:

- Transformers can repeat input tokens (cf. RNNs): $B \ldots B$
- Transformers can induce bigram patterns (Elhage et al., 2021): $AB \ldots AB$
- Language modeling typically uses subword tokenization: *miller*

What gives larger LMs the advantage on rare words?

We know that:

- Transformers can repeat input tokens (cf. RNNs): $B \ldots B$
- Transformers can induce bigram patterns (Elhage et al., 2021): $AB \ldots AB$
- Language modeling typically uses subword tokenization: *miller*



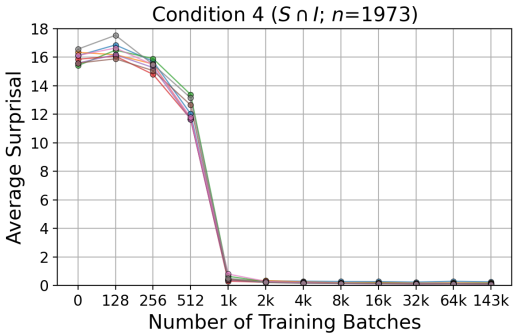
What gives larger LMs the advantage on rare words?

Surprisal from Pythia LMs at various points during training

What gives larger LMs the advantage on rare words?

Surprisal from Pythia LMs at various points during training

$AB \ldots AB$, where AB is in same word

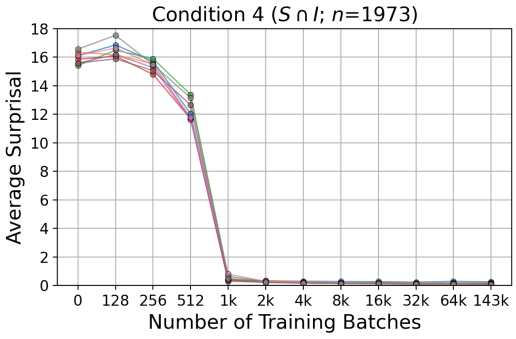


70M 160M 410M 1B 1.4B 2.8B 6.9B 12B

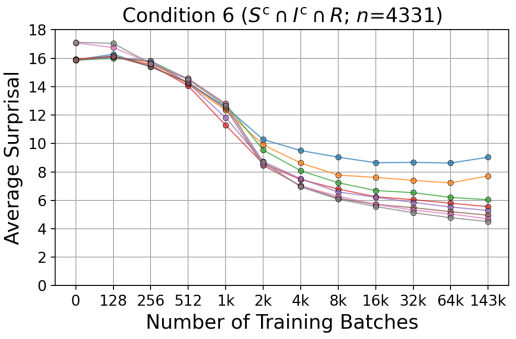
What gives larger LMs the advantage on rare words?

Surprisal from Pythia LMs at various points during training

AB ... *AB*, where *AB* is in same word



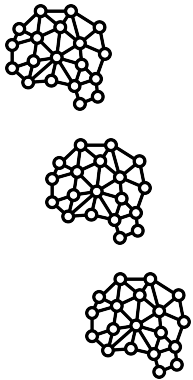
AB ... *CB*, where *CB* is *not* in same word



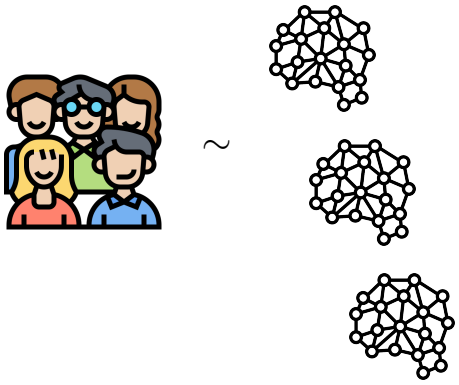
70M 160M 410M 1B 1.4B 2.8B 6.9B 12B

Reading time data

$\sim$



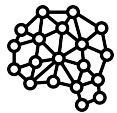
Reading time data

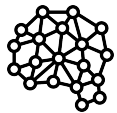


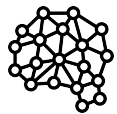
- Larger LMs predict the next word better, but are less human-like

Reading time data

$\sim$



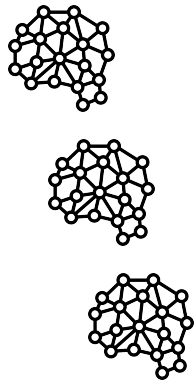




- Larger LMs predict the next word better, but are less human-like
- Strong influence of word frequency on LMs' probabilities and their divergence

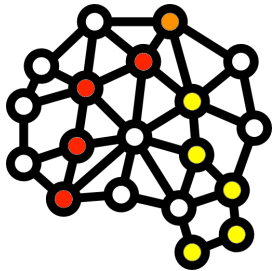
Reading time data

$\sim$



- Larger LMs predict the next word better, but are less human-like
- Strong influence of word frequency on LMs' probabilities and their divergence
- Larger models make more accurate predictions in novel local contexts

Linguistic annotations

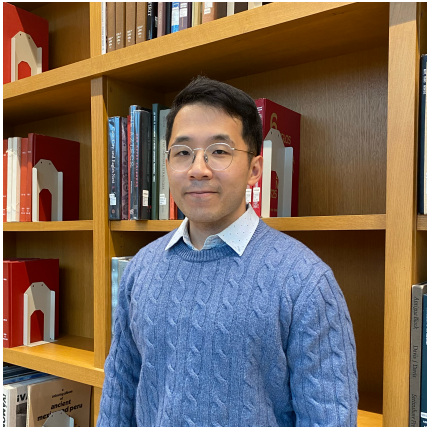


We want to understand a model's predictions

Consider a computer vision model that detects emotions, given an image:

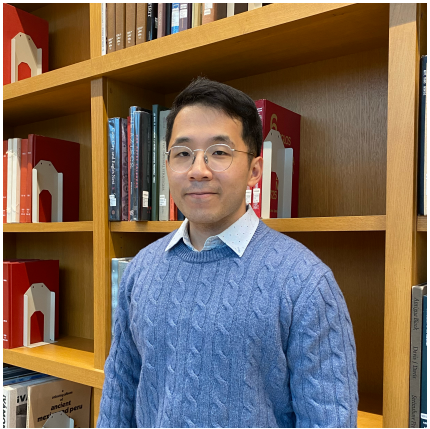
We want to understand a model's predictions

Consider a computer vision model that detects emotions, given an image:



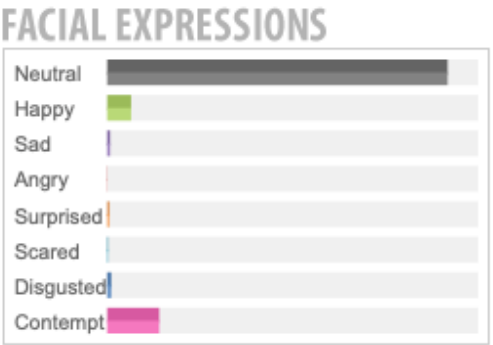
We want to understand a model's predictions

Consider a computer vision model that detects emotions, given an image:



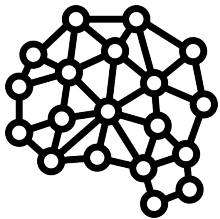
→

Contempt?



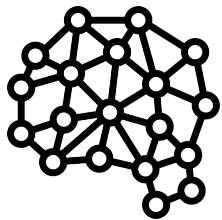
Results from http://noldus.com

Feature attribution methods

I have questions. I'll $\rightarrow$  $\rightarrow$ *ask*

Language model

Feature attribution methods

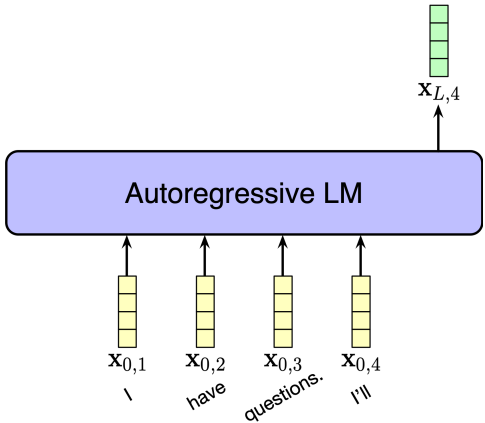
I have questions. I'll $\rightarrow$  $\rightarrow$ *ask*

Language model

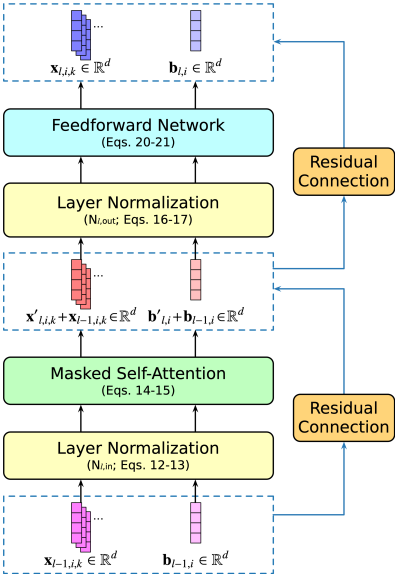
- What in the input sequence led to the prediction of *ask*?

The challenge

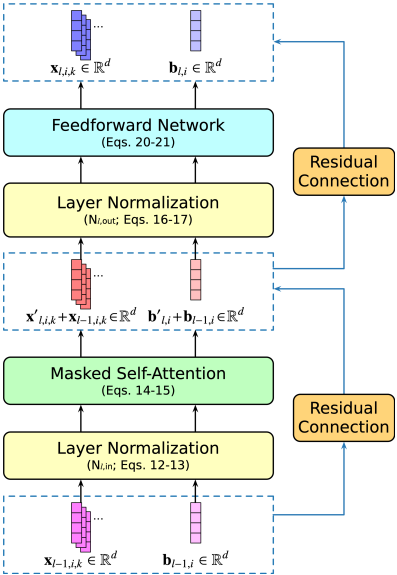
Transformers mix representations in non-linear ways



Linear approximation of Transformers

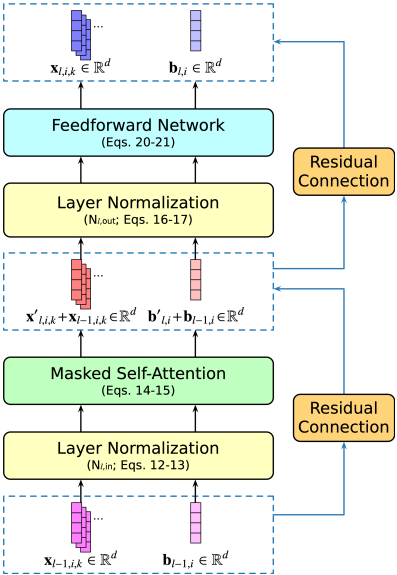


Linear approximation of Transformers



1. Attention weights treated as fixed scalars

Linear approximation of Transformers

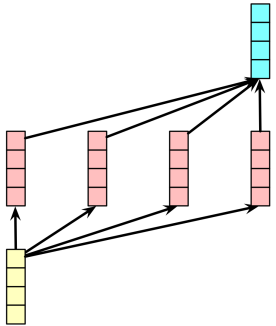


1. Attention weights treated as fixed scalars

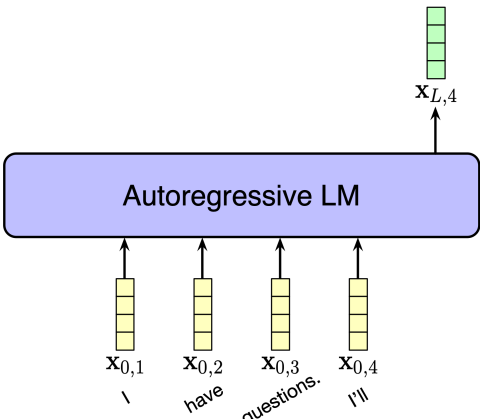
2. FFN approximated with:

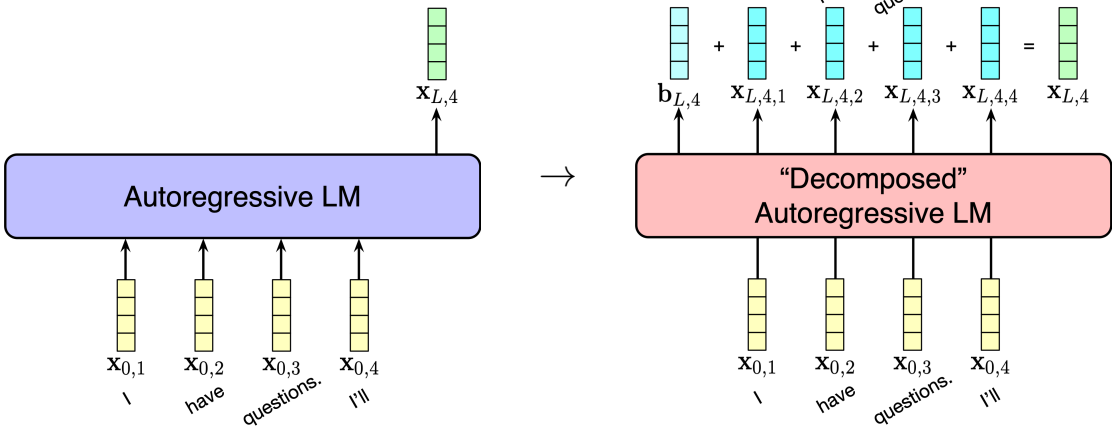
$$\begin{aligned}\mathrm{FF}_l(\mathbf{y}) &= \mathbf{F}_{l,2}\,\sigma(\mathbf{F}_{l,1}\,\mathbf{y} + \mathbf{f}_{l,1}) + \mathbf{f}_{l,2} \\ &\approx \mathbf{F}_{l,2}(\mathbf{s} \odot (\mathbf{F}_{l,1}\,\mathbf{y} + \mathbf{f}_{l,1}) + \mathbf{i}) + \mathbf{f}_{l,2}\end{aligned}$$

Linear approximation allows representations to be separated



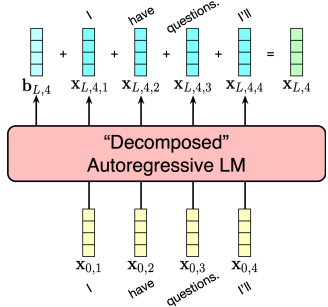
Representation at timestep 4
attributable to input at timestep 1


$\mathbf{x}_{L,4}$
Autoregressive LM
$\mathbf{x}_{0,1}$
I
$\mathbf{x}_{0,2}$
have
$\mathbf{x}_{0,3}$
questions.
$\mathbf{x}_{0,4}$
I'll


$\mathbf{x}_{L,4}$
Autoregressive LM
$\mathbf{x}_{0,1}$
I
$\mathbf{x}_{0,2}$
have
$\mathbf{x}_{0,3}$
questions.
$\mathbf{x}_{0,4}$
I'll
I
have
questions.
I'll
$\mathbf{b}_{L,4}$
+
$\mathbf{x}_{L,4,1}$
+
$\mathbf{x}_{L,4,2}$
+
$\mathbf{x}_{L,4,3}$
+
$\mathbf{x}_{L,4,4}$
=
$\mathbf{x}_{L,4}$
"Decomposed"
Autoregressive LM
$\mathbf{x}_{0,1}$
I
$\mathbf{x}_{0,2}$
have
$\mathbf{x}_{0,3}$
questions.
$\mathbf{x}_{0,4}$
I'll

Methods: Characterizing the most important words

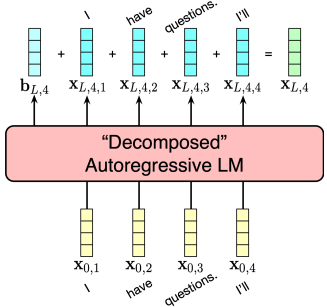
I have a questions. I'll $\rightarrow$



$\rightarrow$ *ask*

Methods: Characterizing the most important words

I have a questions. I'll $\rightarrow$

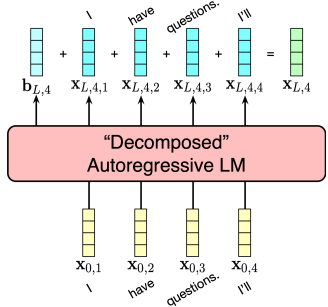


$\rightarrow$ *ask*

1. Ablate each vector and find what causes the largest drop in P(*ask* | ...)

Methods: Characterizing the most important words

I have a questions. I'll $\rightarrow$



$\rightarrow$ *ask*

1. Ablate each vector and find what causes the largest drop in P(*ask* | ...)
2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

Methods: Characterizing the most important words

2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

Methods: Characterizing the most important words

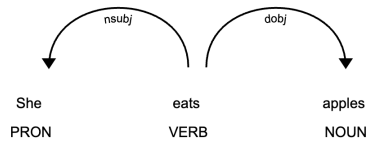
2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

- Pointwise mutual information (PMI): $\frac{P(x,y)}{P(x)P(y)}$, where x, y are words

Methods: Characterizing the most important words

2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

- Pointwise mutual information (PMI): $\frac{P(x,y)}{P(x)P(y)}$, where x, y are words
- Dependency:

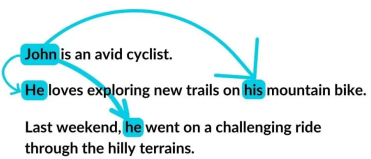


Methods: Characterizing the most important words

2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

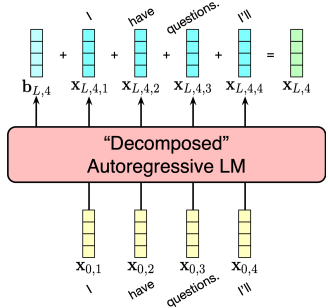
- Pointwise mutual information (PMI): $\frac{P(x,y)}{P(x)P(y)}$, where x, y are words
- Dependency:

- Coreference:



Methods: Characterizing the most important words

I have a questions. I'll $\rightarrow$

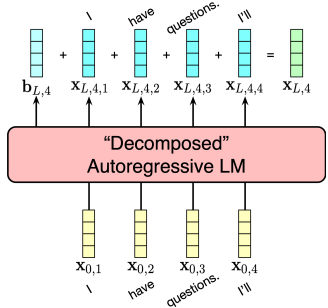


$\rightarrow$ *ask*

1. Ablate each vector and find what causes the largest decrease in P(*ask* | ...)
2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

Methods: Characterizing the most important words

I have a questions. I'll $\rightarrow$



$\rightarrow$ *ask*

1. Ablate each vector and find what causes the largest decrease in P(*ask* | ...)
2. Annotate (*questions*, *ask*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference
3. Fit stepwise regression using four variables to the decrease in P(*ask* | ...)

LMs rely on high-PMI words to make next-word predictions

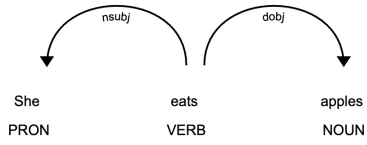
Results from OPT-125M model on CoNLL-2012 corpus

(Pradhan et al., 2012; Zhang et al., 2022)

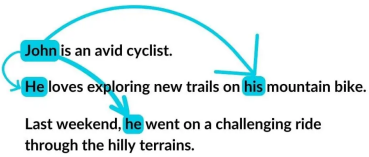
Predictor	Rank	Increase in LogLik
$\text{PMI}_{\text{bigram}}$	1	6151.262*
$\text{PMI}_{\text{document}}$	2	3194.815*
Dependency	3	1981.778*
Coreference	4	25.883*

Follow-up experiments

- Dependency:

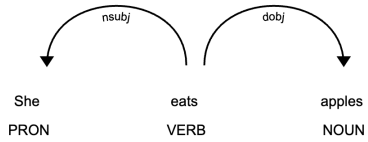


- Coreference:

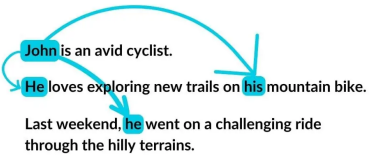


Follow-up experiments

- Dependency:



- Coreference:



Are earlier words in these relationships the most important for predicting later words?

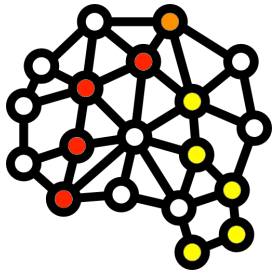
Dependency: When they are high-PMI

Relation	Precision	PMI_{bigram}	$PMI_{document}$
...	...	...	...
Compound	80.44	4.97	2.93
Adjectival modifier	82.55	4.36	2.17
...	...	...	...
Microaverage	56.20	1.11	1.58

Coreference: When the same word is repeated

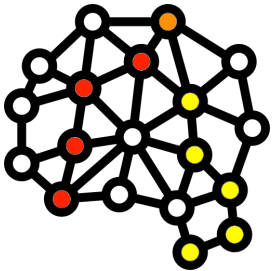
Part-of-speech	Precision	Repeated %
...	...	...
Proper noun (singular)	61.21	68.80
Proper noun (plural)	70.67	68.00
...	...	...
Microaverage	38.21	43.26

Linguistic annotations



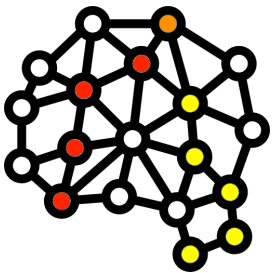
- Framework for decomposing representations in Transformers

Linguistic annotations



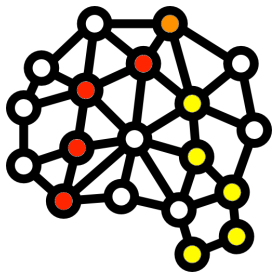
- Framework for decomposing representations in Transformers
- Models seem to rely on collocational associations and repetitions

Linguistic annotations

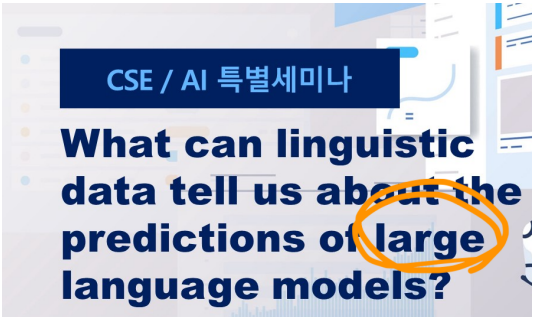


- Framework for decomposing representations in Transformers
- Models seem to rely on collocational associations and repetitions
- But these do overlap with (some) dependency/coreference relationships

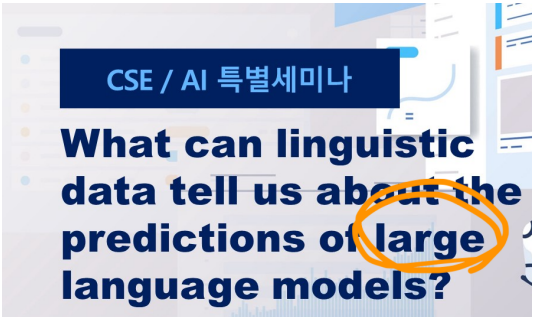
Linguistic annotations



Some closing remarks



Some closing remarks

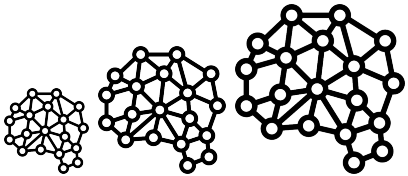


1) Foundation for “scaling laws”

Studying the behavior of smaller models lets us generalize to larger models.

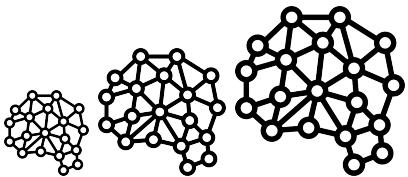
1) Foundation for “scaling laws”

Studying the behavior of smaller models lets us generalize to larger models.



1) Foundation for “scaling laws”

Studying the behavior of smaller models lets us generalize to larger models.



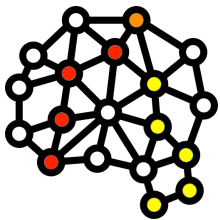
...

2) We can't use LLMs for everything

Smaller models are necessary when data is limited, and they give us control.

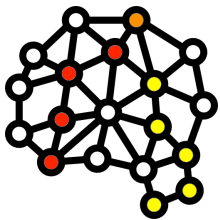
2) We can't use LLMs for everything

Smaller models are necessary when data is limited, and they give us control.

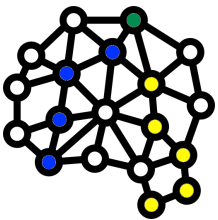


2) We can't use LLMs for everything

Smaller models are necessary when data is limited, and they give us control.



"Steering"



3) Are people good NLP models?

3) Are people good NLP models?

Probably not.

3) Are people good NLP models?

Probably not.

But:

3) Are people good NLP models?

Probably not.

But:

- People have robust representations that generalize

3) Are people good NLP models?

Probably not.

But:

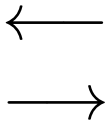
- People have robust representations that generalize
- People are efficient learners

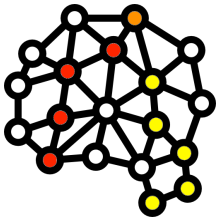
3) Are people good NLP models?

Probably not.

But:

- People have robust representations that generalize
- People are efficient learners
- People ultimately determine the "quality" of language

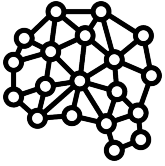


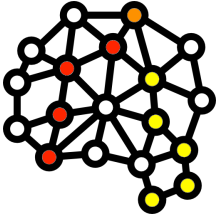


Thank you for listening!

oh.b@nyu.edu byungdoh.github.io

$\sim$





Oh, Yue, and Schuler (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *Proc. EACL*.
+ Some material under review

Oh and Schuler (2023). Token-wise decomposition of autoregressive language model hidden states for analyzing model predictions. *Proc. ACL*.

Image credits

https://www.flaticon.com
https://thenounproject.com
https://www.bitbrain.com
https://spotintelligence.com

Patterson and Nicklin (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*(1), 100045.

References I

Balota, D. A., Pollatsek, A., & Rayner, K. (1985). The interaction of contextual constraints and parafoveal visual information in reading. *Cognitive Psychology*, *17*(3), 364–390.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … Olah, C. (2021). A mathematical framework for Transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://aclanthology.org/N01-1021/

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Oh, B.-D., & Schuler, W. (2023). Token-wise decomposition of autoregressive language model hidden states for analyzing model predictions. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 10105–10117. https://aclanthology.org/2023.acl-long.562

References II

Oh, B.-D., Yue, S., & Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, 2644–2663. https://aclanthology.org/2024.eacl-long.162/

Patterson, A. S., & Nicklin, C. (2023). L2 self-paced reading data collection across three contexts: In-person, online, and crowdsourcing. *Research Methods in Applied Linguistics*, *2*(1), 100045. https://doi.org/10.1016/j.rmal.2023.100045

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., & Zhang, Y. (2012). CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. *Joint Conference on EMNLP and CoNLL - Shared Task*, 1–40. https://aclanthology.org/W12-4501

Shain, C. (2024). Word frequency and predictability dissociate in naturalistic reading. *Open Mind*, *8*, 177–201.

Tirumala, K., Markosyan, A., Zettlemoyer, L., & Aghajanyan, A. (2022). Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, *35*, 38274–38290. https://proceedings.neurips.cc/paper_files/paper/2022/file/fa0509f4dab6807e2cb465715bf2d249-Paper-Conference.pdf

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: Open pre-trained Transformer language models. *arXiv preprint, arXiv:2205.01068v4*. https://arxiv.org/abs/2205.01068

LMs rely on high-PMI words (full)

Predictor	β	t-value	Increase in LogLik
Word index	0.034	1.919	-
Distance	1.126	62.755	-
Log probability	-0.083	-5.350	-
$\text{PMI}_{\text{bigram}}$	1.220	70.857	6151.262*
$\text{PMI}_{\text{document}}$	1.286	73.952	3194.815*
Dependency	1.055	63.720	1981.778*
Coreference	0.123	7.195	25.883*

Dependency: When they are high-PMI

Relation	Precision	PMI_{bigram}	$PMI_{document}$
Nominal subject	61.15	1.38	1.44
Direct object	70.43	0.91	1.57
Oblique	52.54	-0.68	1.54
Compound	80.44	4.97	2.93
Nominal modifier	53.84	-0.41	1.84
Adjectival modifier	82.55	4.36	2.17
Determiner	52.03	1.51	1.08
Case marker	52.38	-0.29	1.08
Microaverage	56.20	1.11	1.58

Coreference: When the same word is repeated

Mention head part-of-speech	Precision	Repeated %
Personal pronoun	26.55	30.92
Possessive pronoun	23.29	30.59
Proper noun (singular)	61.21	68.80
Proper noun (plural)	70.67	68.00
Common noun (singular)	43.39	48.75
Common noun (plural)	47.01	55.03
Possessive ending	46.28	40.91
Microaverage	38.21	43.26