Unveiling the Language Processing of Humans and Machines

Byung-Doh Oh

Center for Data Science
New York University

25 February 2025
NTU Linguistics and Multilingual Studies

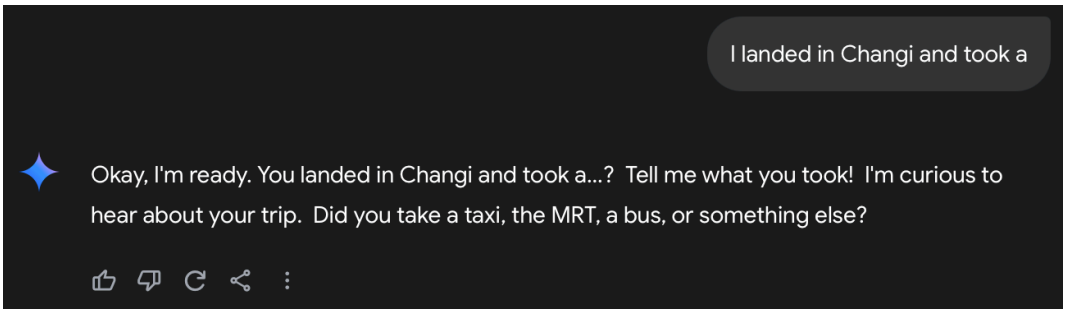
Human language processing rapid and efficient

Human language processing rapid and efficient

I landed in Changi and took a ...

Computational models are getting there

Computational models are getting there



Example from Gemini 2.0

The processing mechanisms are poorly understood in both

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

– Educational/clinical applications

The processing mechanisms are poorly understood in both

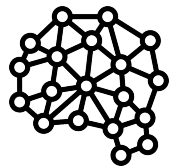
- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

- Educational/clinical applications
- More 'human-like' AI systems

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

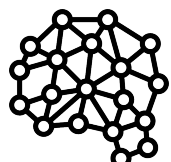
- Educational/clinical applications
- More 'human-like' AI systems



The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

- Educational/clinical applications
- More 'human-like' AI systems

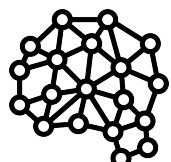


- AI safety/control

The processing mechanisms are poorly understood in both

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

- Educational/clinical applications
- More 'human-like' AI systems

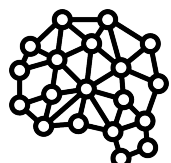


- AI safety/control
- Efficient training

The challenges

- What are the characteristics of the representation r that humans/models build?
- How is r integrated with bottom-up input w to form a new representation r'?

– Underlying computations are largely unobservable



– Computations are uninterpretable to the human researcher

The challenges

The challenges

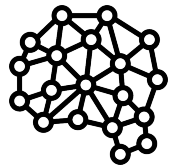
```
>>>
>>> model(torch.tensor(tokenizer("I landed in Changi").input_ids)).hidden_states[-1]
tensor([[-0.0796, -0.0654, -0.0842,  ..., -0.1442, -0.0456,  0.0143],
        [-0.0179,  0.7541, -0.6507,  ...,  0.1597,  0.2391, -0.0941],
        [ 0.1664,  0.3424,  0.5167,  ...,  0.0764,  0.1678,  0.0549],
        [ 0.6639,  0.6770,  0.3488,  ...,  0.7069, -0.5089, -0.0503],
        [-0.2492,  0.3408,  0.1225,  ...,  0.1948,  0.3410,  0.1411]],
       grad_fn=<ViewBackward0>)
>>> _
```

Example from GPT-2

My research program

- Underlying computations are largely unobservable

Psycholinguistics

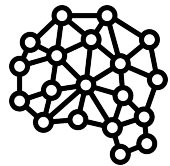


- Computations are uninterpretable to the human researcher

NLP/Interpretability

My research program

Computational modeling



– Underlying computations are largely unobservable

Psycholinguistics

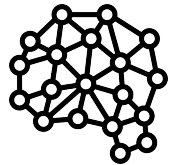
– Computations are uninterpretable to the human researcher

NLP/Interpretability

My research program

Computational modeling

Linguistic data



– Underlying computations are largely unobservable

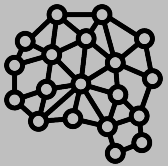
Psycholinguistics

– Computations are uninterpretable to the human researcher

NLP/Interpretability

Today's talk: Part #1

Computational modeling



- Underlying computations are largely unobservable

Psycholinguistics

- Computations are uninterpretable to the human researcher

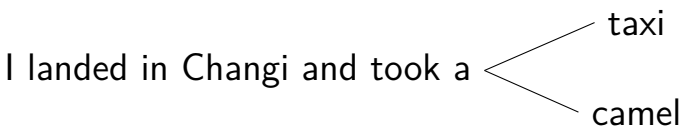
NLP/Interpretability

The shared principle of prediction (humans)

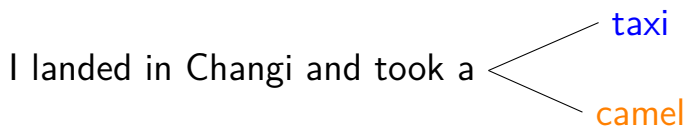
The shared principle of prediction (humans)

I landed in Changi and took a

The shared principle of prediction (humans)



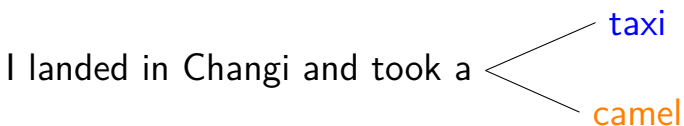
The shared principle of prediction (humans)



The more predictable taxi is easier to process than camel

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

The shared principle of prediction (humans)

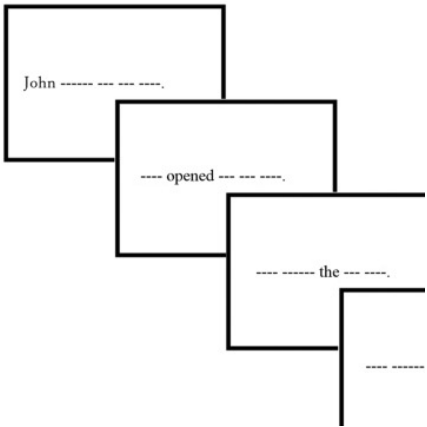


The more predictable taxi is easier to process than camel
(Balota et al., 1985; Ehrlich & Rayner, 1981; Kutas & Hillyard, 1980)

This shows up in reading time data collected from human subjects
(Schrimpf et al., 2021; Shain et al., 2024; Smith & Levy, 2013)

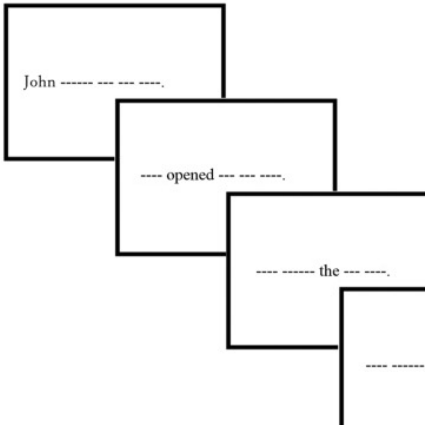
The shared principle of prediction (humans)

The shared principle of prediction (humans)



Self-paced reading

The shared principle of prediction (humans)



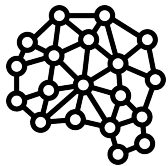
Self-paced reading

Eye-tracking

The shared principle of prediction (language models; LMs)

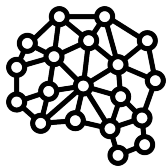
The shared principle of prediction (language models; LMs)

I landed in Changi and took a $\rightarrow$



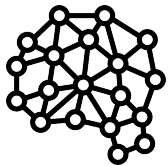
Language models

The shared principle of prediction (language models; LMs)

I landed in Changi and took a $\rightarrow$  $\rightarrow$ taxi

Language models

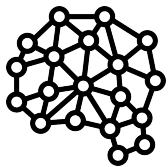
The shared principle of prediction (language models; LMs)

I landed in Changi and took a $\rightarrow$  $\rightarrow$

Language models

Word	Probability
taxi	0.1174
flight	0.0635
bus	0.0403
...	...
camel	3.8×10^{-5}

The shared principle of prediction (language models; LMs)

I landed in Changi and took a $\rightarrow$  $\rightarrow$

Language models

Word	Probability
taxi	0.1174
flight	0.0635
bus	0.0403
...	...
camel	3.8×10^{-5}

Models learn nontrivial linguistic structure by simply predicting the next word

(Futrell & Mahowald, 2025; Linzen & Baroni, 2021; Mahowald et al., 2024)

The shared principle of prediction (language models; LMs)

LMs are typically evaluated by how much probability they place on the ‘correct’ text

The shared principle of prediction (language models; LMs)

LMs are typically evaluated by how much probability they place on the 'correct' text

Perplexity: $\mathrm{P}(w_1, w_2, ..., w_N)^{-\frac{1}{N}}$, lower perplexity indicates higher probability

The shared principle of prediction (language models; LMs)

LMs are typically evaluated by how much probability they place on the 'correct' text

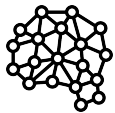
Perplexity: $\mathrm{P}(w_1, w_2, ..., w_N)^{-\frac{1}{N}}$, lower perplexity indicates higher probability

LMs with more free parameters (i.e. 'larger' models) achieve lower perplexity

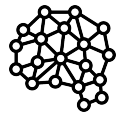
Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

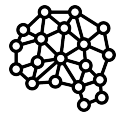
$\sim$



LM 1



LM 2

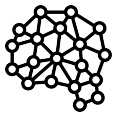


LM 3

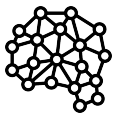
Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

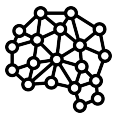
$\sim$



LM 1



LM 2



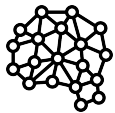
LM 3

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

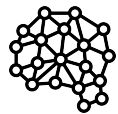
Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

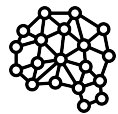
$\sim$



LM 1



LM 2



LM 3

Reading time of *taxi* $\propto -\log_2 \mathrm{P}(\text{taxi} \mid \text{I landed in Changi and took a})$

Reading time of *camel* $\propto -\log_2 \mathrm{P}(\text{camel} \mid \text{I landed in Changi and took a})$

Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

Text	I	landed	in	Changi
Reading Time	709 ms	847 ms	766 ms	886 ms
LM1 Surprisal	4.95	6.40	1.32	6.04
LM2 Surprisal	3.53	5.73	0.69	4.14
LM3 Surprisal	3.50	5.13	0.59	3.63

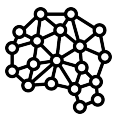
Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

Text	I	landed	in	Changi
Reading Time	709 ms	847 ms	766 ms	886 ms
LM1 Surprisal	4.95	6.40	1.32	6.04
LM2 Surprisal	3.53	5.73	0.69	4.14
LM3 Surprisal	3.50	5.13	0.59	3.63

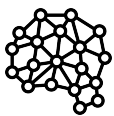
Human subjects

$\sim$

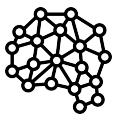
Regression modeling



LM 1



LM 2



LM 3

Theoretical link between humans and LMs (surprisal theory; Hale, 2001; Levy, 2008)

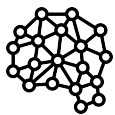
Text	I	landed	in	Changi
Reading Time	709 ms	847 ms	766 ms	886 ms
LM1 Surprisal	4.95	6.40	1.32	6.04
LM2 Surprisal	3.53	5.73	0.69	4.14
LM3 Surprisal	3.50	5.13	0.59	3.63

Long tradition in computational psycholinguistics (Demberg & Keller, 2008; Jurafsky, 1996; Oh et al., 2021)

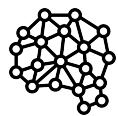
Human subjects

$\sim$

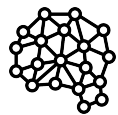
Regression modeling



LM 1



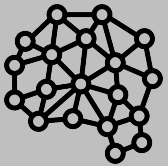
LM 2



LM 3

Today's talk: Part #1

Computational modeling



Psycholinguistics

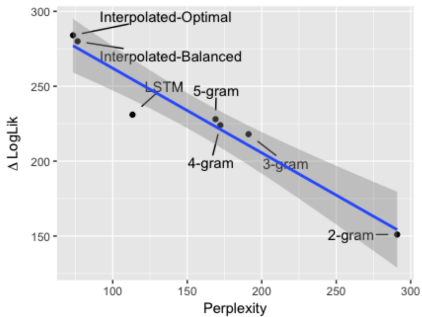
NLP/Interpretability

Oh and Schuler (2023a). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*.

People thought language processing is driven by accurate prediction

People thought language processing is driven by accurate prediction

Better Fit



Poorer Fit

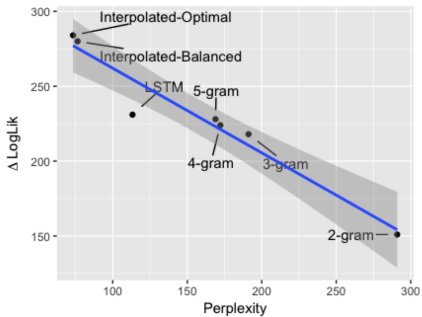
Goodkind and Bicknell (2018)

More Accurate ⟷ Less Accurate

People thought language processing is driven by accurate prediction

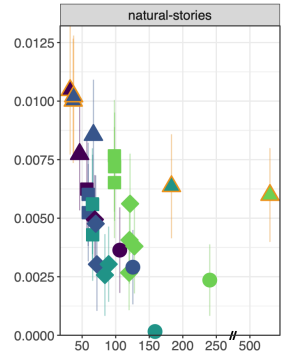
Better Fit

Poorer Fit



Goodkind and Bicknell (2018)

More Accurate ⟷ Less Accurate



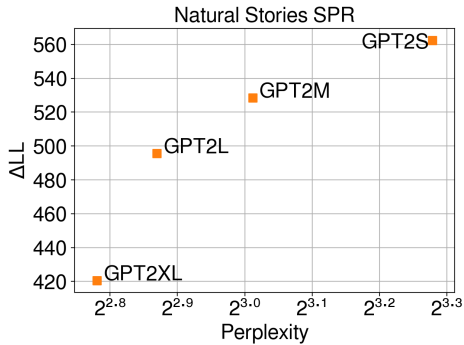
Wilcox et al. (2020)

More Accurate ⟷ Less Accurate

This relationship seems to break down with more contemporary LMs

This relationship seems to break down with more contemporary LMs

Better Fit

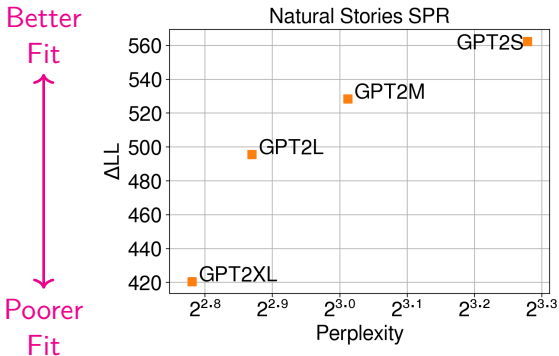


Poorer Fit

Oh et al. (2022)

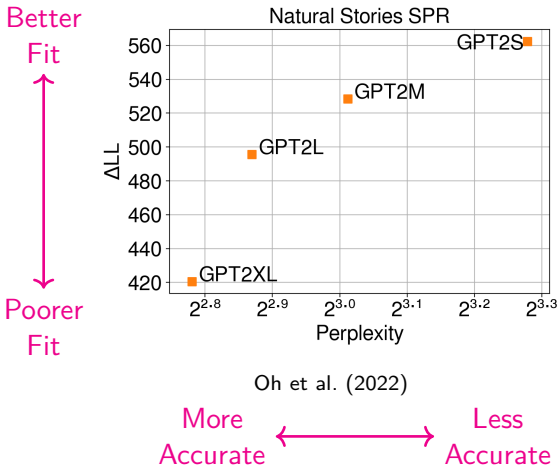
More Accurate ⟷ Less Accurate

This relationship seems to break down with more contemporary LMs



Does this trend replicate with other contemporary LMs?

This relationship seems to break down with more contemporary LMs



Does this trend replicate with other contemporary LMs?

If so, what drives this trend?

Methods: Replication with more LM families

- Regression models fit to reading times of Natural Stories and Dundee corpora (Futrell et al., 2021; Kennedy et al., 2003)

Methods: Replication with more LM families

- Regression models fit to reading times of Natural Stories and Dundee corpora (Futrell et al., 2021; Kennedy et al., 2003)

- Baseline predictors: word length/position, saccade length, previous word fixated

Methods: Replication with more LM families

- Regression models fit to reading times of Natural Stories and Dundee corpora (Futrell et al., 2021; Kennedy et al., 2003)
- Baseline predictors: word length/position, saccade length, previous word fixated
- Predictors of interest: LM surprisal

Model	Model size (#Parameters)
GPT-2 Small	∼117M
GPT-2 Medium	∼345M
GPT-2 Large	∼774M
GPT-2 XL	∼1.6B
GPT-Neo 125M	∼125M
GPT-Neo 1.3B	∼1.3B
GPT-Neo 2.7B	∼2.7B
GPT-J 6B	∼6B
GPT-NeoX 20B	∼20B
OPT 125M	∼125M
OPT 350M	∼350M
OPT 1.3B	∼1.3B
OPT 2.7B	∼2.7B
OPT 6.7B	∼6.7B
OPT 13B	∼13B
OPT 30B	∼30B
OPT 66B	∼66B

Methods: Replication with more LM families

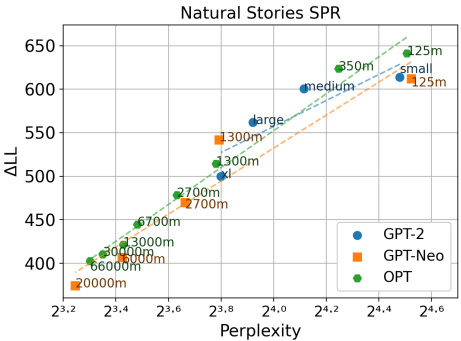
- Regression models fit to reading times of Natural Stories and Dundee corpora (Futrell et al., 2021; Kennedy et al., 2003)
- Baseline predictors: word length/position, saccade length, previous word fixated
- Predictors of interest: LM surprisal
- Evaluation: Δlog-likelihood (ΔLL); how well does surprisal fit to RT?

Model	Model size (#Parameters)
GPT-2 Small	∼117M
GPT-2 Medium	∼345M
GPT-2 Large	∼774M
GPT-2 XL	∼1.6B
GPT-Neo 125M	∼125M
GPT-Neo 1.3B	∼1.3B
GPT-Neo 2.7B	∼2.7B
GPT-J 6B	∼6B
GPT-NeoX 20B	∼20B
OPT 125M	∼125M
OPT 350M	∼350M
OPT 1.3B	∼1.3B
OPT 2.7B	∼2.7B
OPT 6.7B	∼6.7B
OPT 13B	∼13B
OPT 30B	∼30B
OPT 66B	∼66B

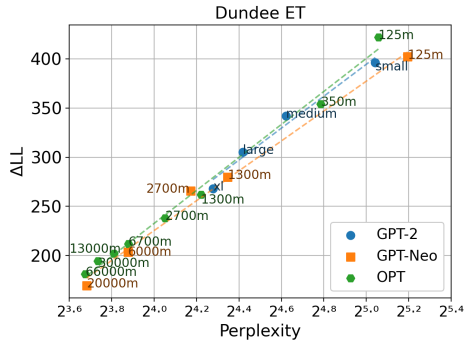
Strictly monotonic, positive relationship

Better Fit

Poorer Fit



More Accurate, Larger ⟷ Less Accurate, Smaller



More Accurate, Larger ⟷ Less Accurate, Smaller

What linguistic factors drive this trend?

What linguistic factors drive this trend?

- Text annotated with word-level and syntactic properties (Shain et al., 2018)

What linguistic factors drive this trend?

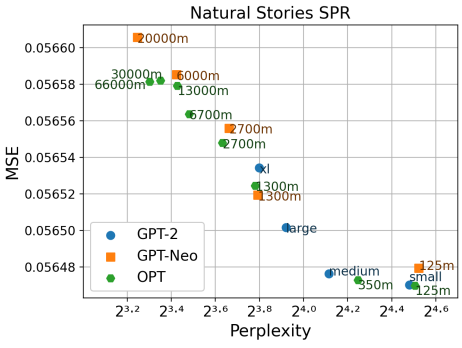
- Text annotated with word-level and syntactic properties (Shain et al., 2018)
- Subsets with the largest differences in MSE between models identified

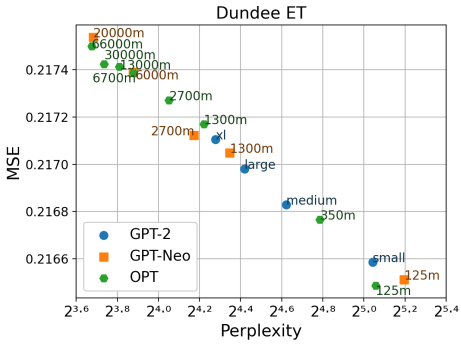
What linguistic factors drive this trend?

- Text annotated with word-level and syntactic properties (Shain et al., 2018)
- Subsets with the largest differences in MSE between models identified

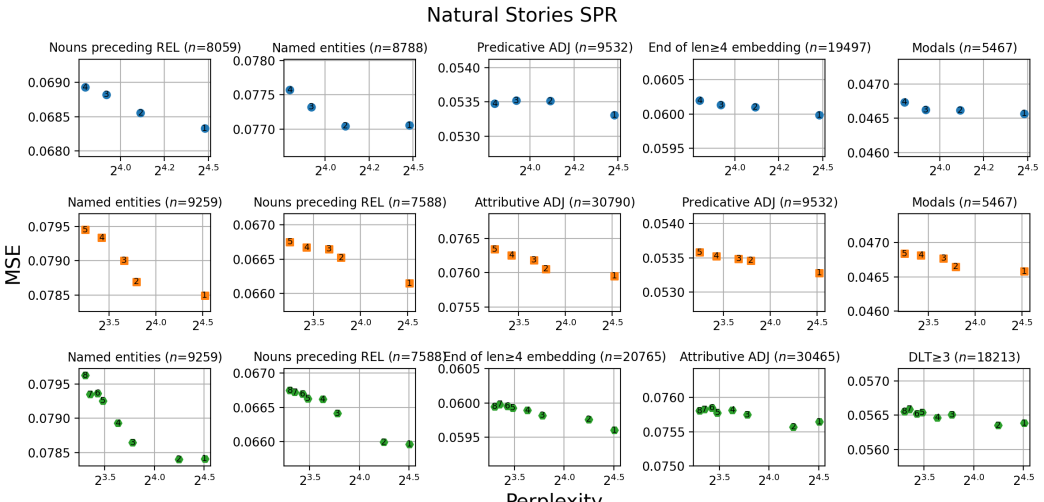
Poorer Fit

Better Fit

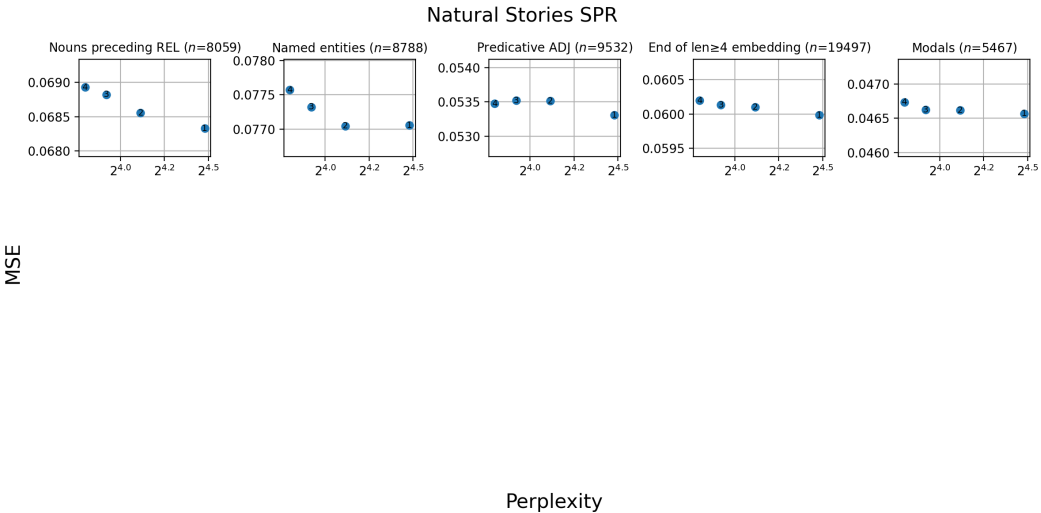




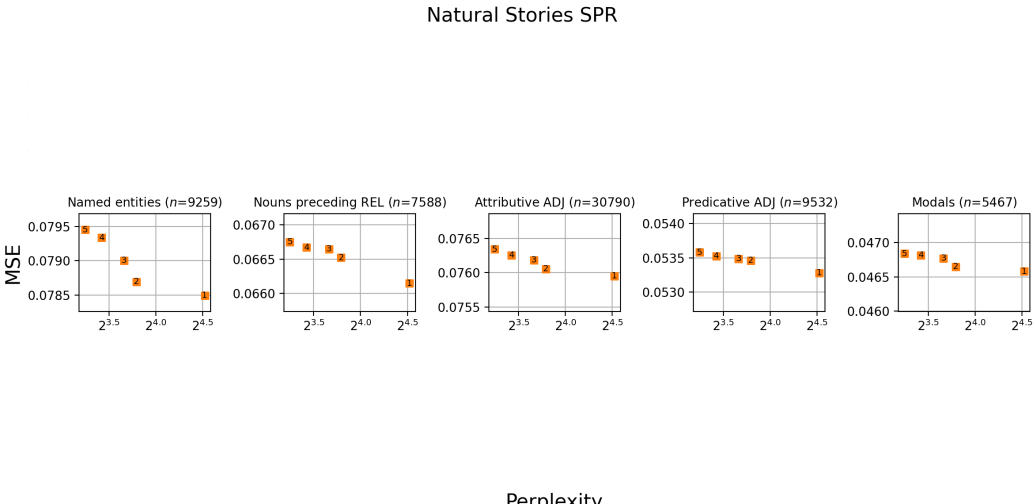
Driven by open-class words like named entities



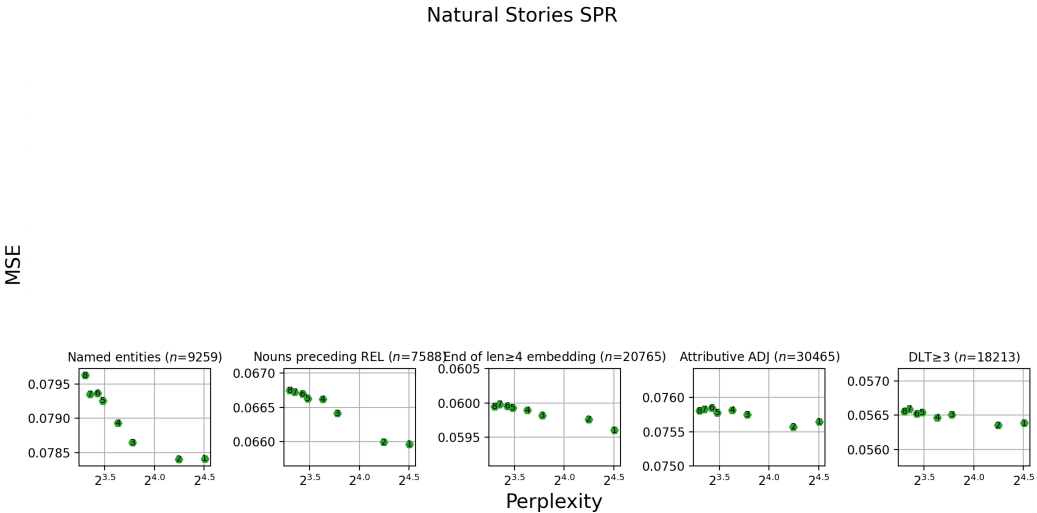
Driven by open-class words like named entities



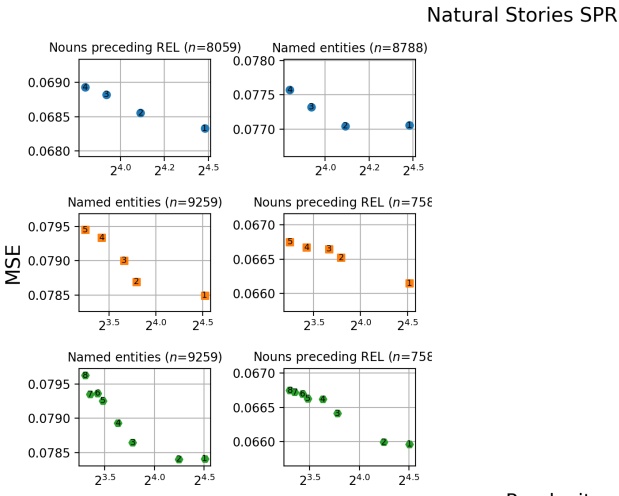
Driven by open-class words like named entities



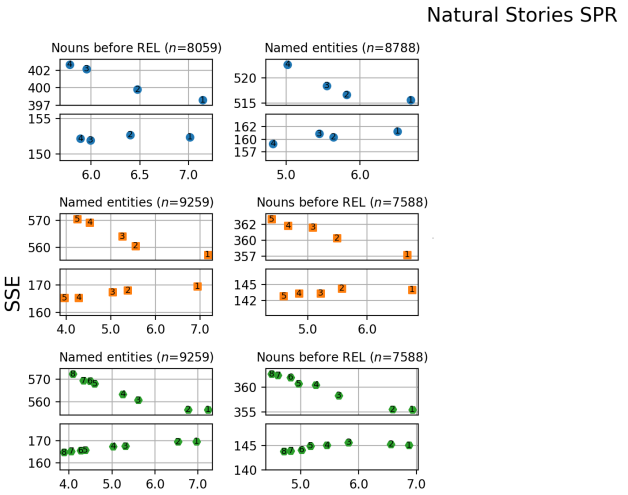
Driven by open-class words like named entities



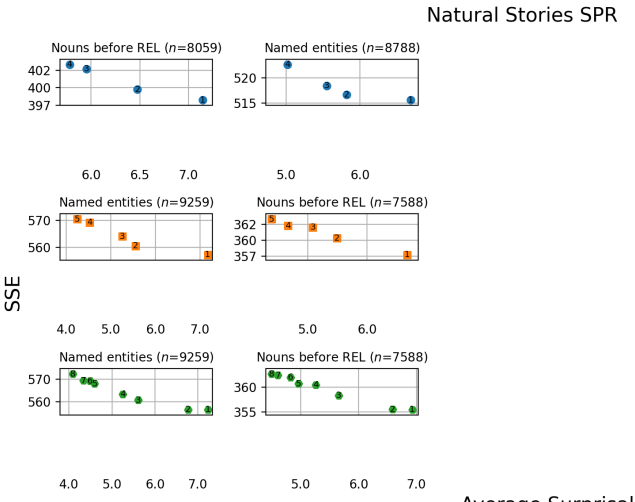
Driven by open-class words like named entities



Reading times are underpredicted by larger LMs



Reading times are underpredicted by larger LMs



Some examples

(In a passage about the Roswell UFO incident)

... In January nineteen ninety-seven, Karl _____, one of the more prominent pro-UFO researchers, ...

Some examples

(In a passage about the Roswell UFO incident)
... In January nineteen ninety-seven, Karl _____, one of the more prominent pro-UFO researchers, ...

(In a passage about the Tulip mania)
... At one point twelve acres of land were offered for a Semper _____ bulb. ...

Some examples

(In a passage about the Roswell UFO incident)
... In January nineteen ninety-seven, Karl _____, one of the more prominent pro-UFO researchers, ...

(In a passage about the Tulip mania)
... At one point twelve acres of land were offered for a Semper _____ bulb. ...

Larger LMs appear to make increasingly 'superhuman' predictions of such words

Summary: The bigger-is-worse effect of LM size

Surprisal from larger LMs show strictly poorer fits to human reading times

Summary: The bigger-is-worse effect of LM size

Surprisal from larger LMs show strictly poorer fits to human reading times

Effect mostly driven by underprediction of reading times by LLM surprisal
(see e.g. Arehalli et al., 2022; Hahn et al., 2022; van Schijndel & Linzen, 2021)

Summary: The bigger-is-worse effect of LM size

Surprisal from larger LMs show strictly poorer fits to human reading times

Effect mostly driven by underprediction of reading times by LLM surprisal
(see e.g. Arehalli et al., 2022; Hahn et al., 2022; van Schijndel & Linzen, 2021)

Likely due to extensive domain knowledge from massive amounts of training examples

Summary: The bigger-is-worse effect of LM size

Surprisal from larger LMs show strictly poorer fits to human reading times

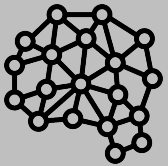
Effect mostly driven by underprediction of reading times by LLM surprisal
(see e.g. Arehalli et al., 2022; Hahn et al., 2022; van Schijndel & Linzen, 2021)

Likely due to extensive domain knowledge from massive amounts of training examples

Suggests typical language processing is not informed by accurate factual knowledge

Today's talk: Part #1

Computational modeling



Psycholinguistics

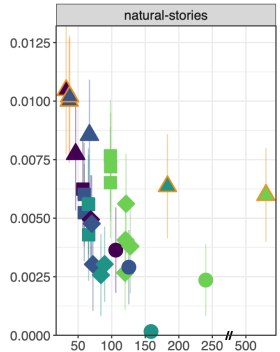
NLP/Interpretability

Oh and Schuler (2023b). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

(Still) conflicting results about next-word prediction accuracy

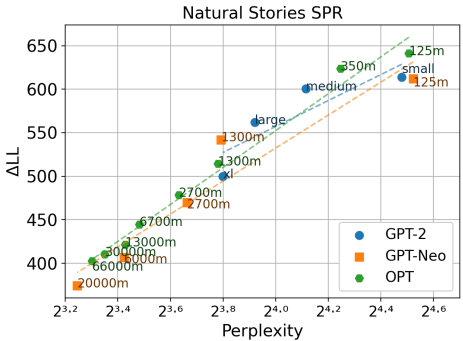
Better Fit

Poorer Fit



Wilcox et al. (2020)

More Accurate ⟷ Less Accurate



Oh and Schuler (2023a)

More Accurate ⟷ Less Accurate

Covering the middle ground (training data)

- Regression models fit and ΔLL calculated

Covering the middle ground (training data)

- Regression models fit and ΔLL calculated
- Predictors of interest: LM surprisal

Model	Model size (#Parameters)
Pythia 70M	∼70M
Pythia 160M	∼160M
Pythia 410M	∼410M
Pythia 1B	∼1B
Pythia 1.4B	∼1.4B
Pythia 2.8B	∼2.8B
Pythia 6.9B	∼6.9B
Pythia 12B	∼12B

Covering the middle ground (training data)

- Regression models fit and ΔLL calculated
- Predictors of interest: LM surprisal
- Trained on identical batches of 1024×2048 ($\sim$2 million) tokens

Model	Model size (#Parameters)
Pythia 70M	$\sim$70M
Pythia 160M	$\sim$160M
Pythia 410M	$\sim$410M
Pythia 1B	$\sim$1B
Pythia 1.4B	$\sim$1.4B
Pythia 2.8B	$\sim$2.8B
Pythia 6.9B	$\sim$6.9B
Pythia 12B	$\sim$12B

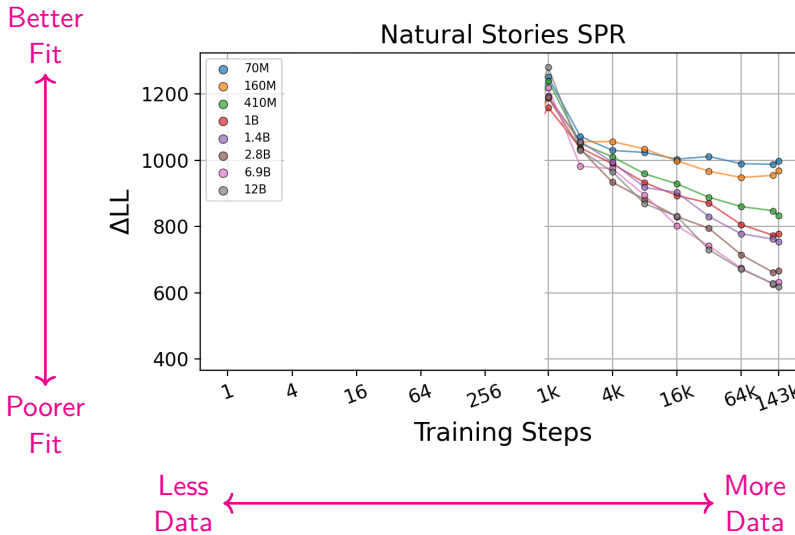
Covering the middle ground (training data)

- Regression models fit and ΔLL calculated
- Predictors of interest: LM surprisal
- Trained on identical batches of 1024×2048 ($\sim$2 million) tokens
- Intermediate checkpoints throughout training evaluated

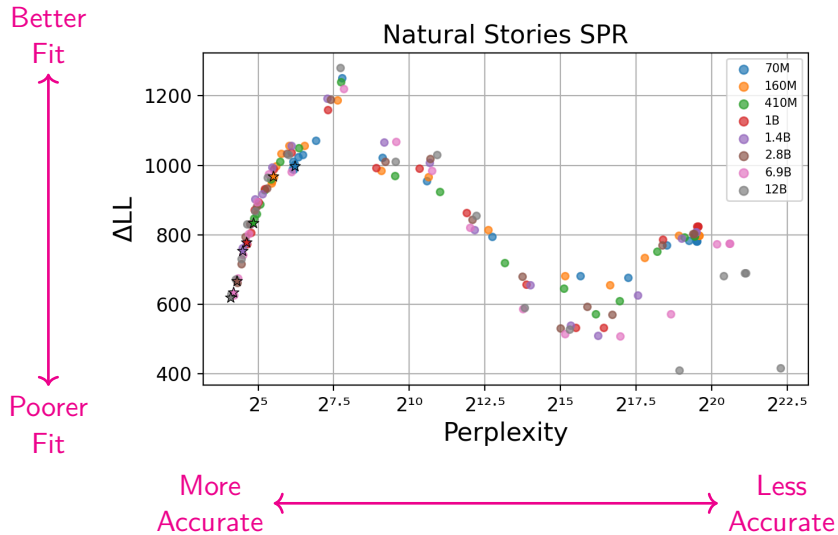
Model	Model size (#Parameters)
Pythia 70M	$\sim$70M
Pythia 160M	$\sim$160M
Pythia 410M	$\sim$410M
Pythia 1B	$\sim$1B
Pythia 1.4B	$\sim$1.4B
Pythia 2.8B	$\sim$2.8B
Pythia 6.9B	$\sim$6.9B
Pythia 12B	$\sim$12B

Sweet spot at around two billion tokens

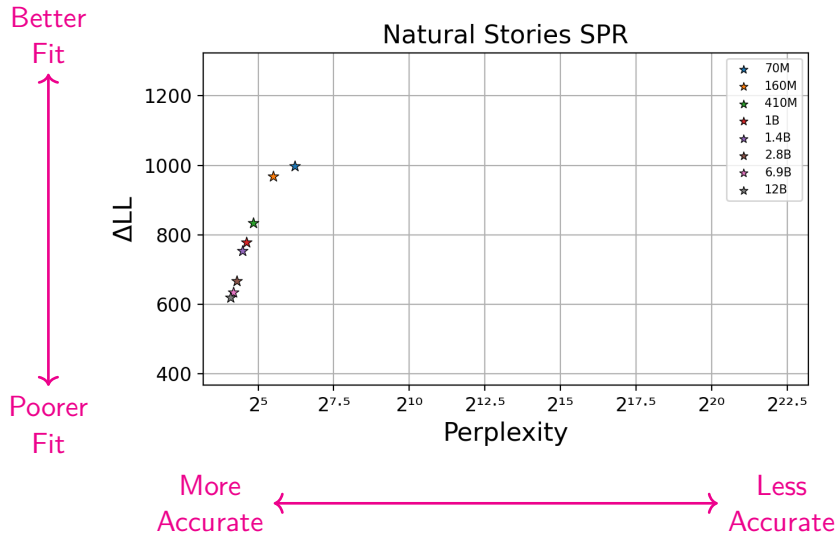
Sweet spot at around two billion tokens



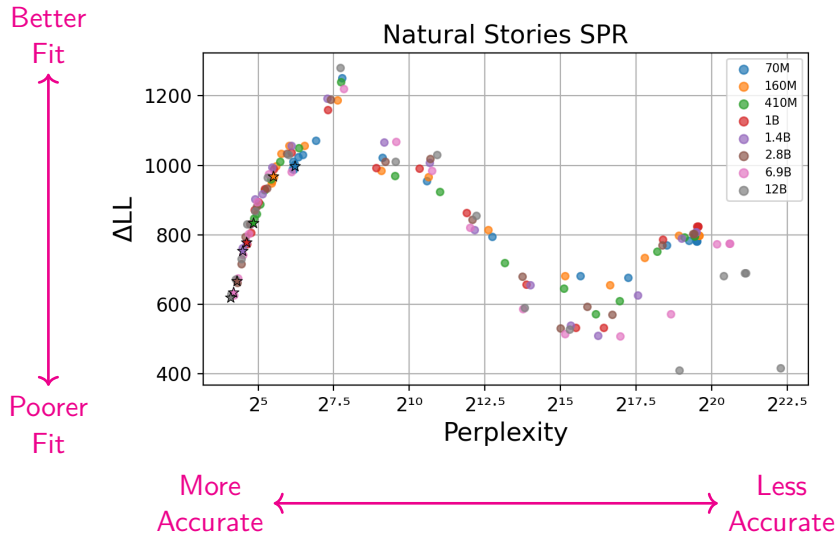
There are two different regimes



There are two different regimes



There are two different regimes



Covering the middle ground (model size)

- Smaller LMs trained following the procedures of the Pythia LM

Covering the middle ground (model size)

- Smaller LMs trained following the procedures of the Pythia LM

Model	Model size (#Parameters)
Repro 1-1-64	∼6M
Repro 1-2-128	∼13M
Repro 2-2-128	∼13M
Repro 2-3-192	∼20M
Repro 2-4-256	∼27M
Repro 3-4-256	∼28M
Repro 4-6-384	∼46M
Repro 6-8-512	∼70M

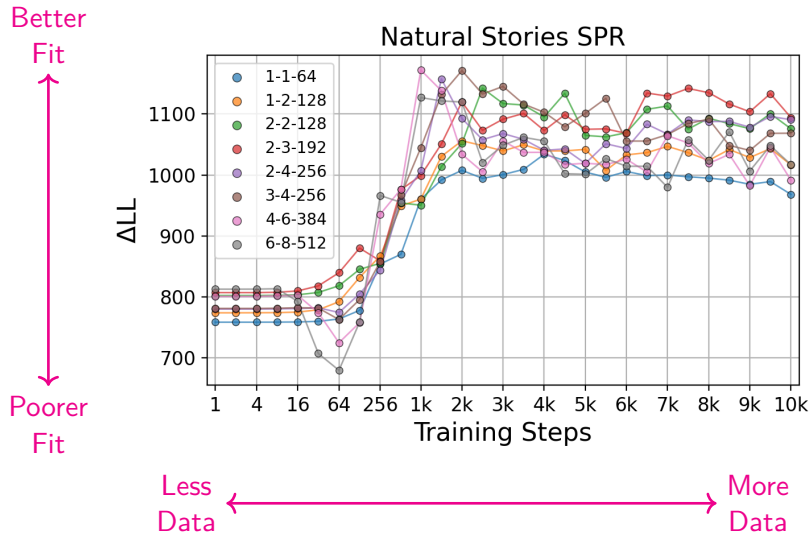
Covering the middle ground (model size)

- Smaller LMs trained following the procedures of the Pythia LM

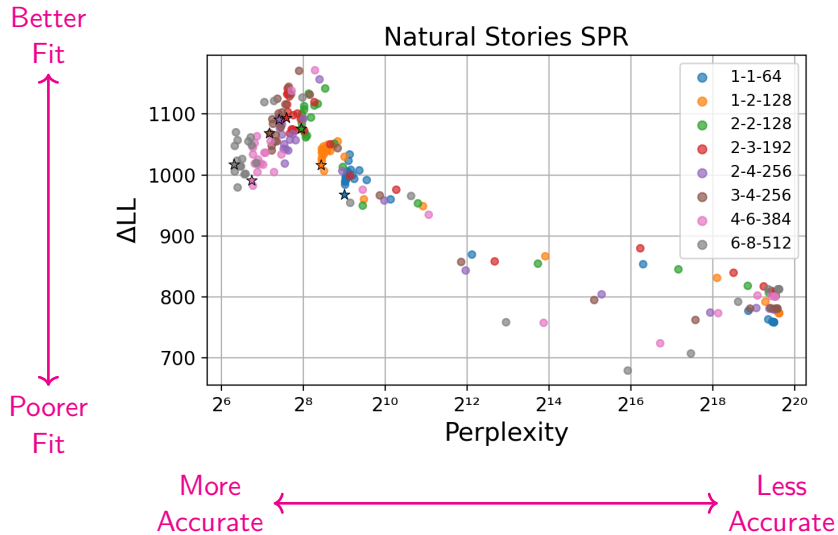
Model	Model size (#Parameters)
Repro 1-1-64	∼6M
Repro 1-2-128	∼13M
Repro 2-2-128	∼13M
Repro 2-3-192	∼20M
Repro 2-4-256	∼27M
Repro 3-4-256	∼28M
Repro 4-6-384	∼46M
Repro 6-8-512	∼70M

- LMs evaluated after {1, 2, 4, ..., 512, 1000, 1500, ..., 10000} training steps

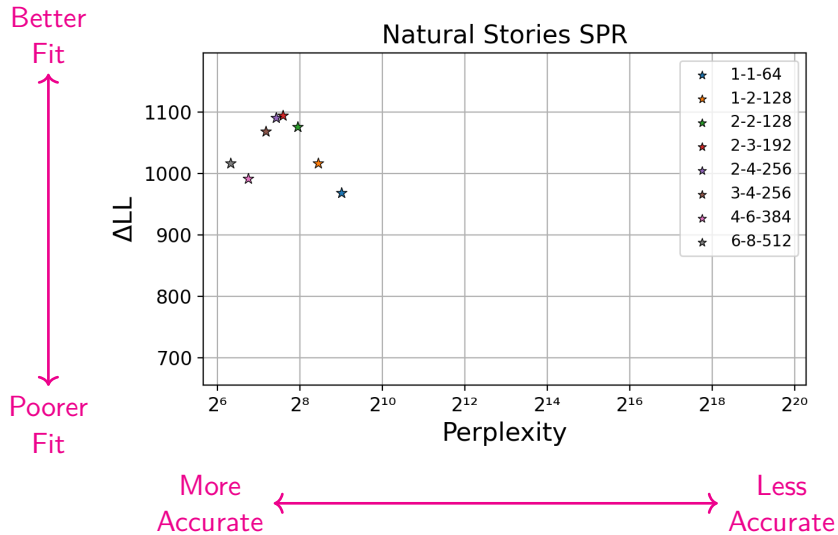
Smaller LMs converge earlier



The two different regimes, again



The two different regimes, again



Summary: The bigger-is-worse effect of training data

Fit to reading times starts to degrade after about two billion tokens of training data

Summary: The bigger-is-worse effect of training data

Fit to reading times starts to degrade after about two billion tokens of training data

Strong interaction between model size and training data amount after the peak

Summary: The bigger-is-worse effect of training data

Fit to reading times starts to degrade after about two billion tokens of training data

Strong interaction between model size and training data amount after the peak

Consolidates conflicting results about LM perplexity and fit to reading times

Summary: The bigger-is-worse effect of training data

Fit to reading times starts to degrade after about two billion tokens of training data

Strong interaction between model size and training data amount after the peak

Consolidates conflicting results about LM perplexity and fit to reading times

Suggests the need to refine surprisal theory in terms of the quantity of language input

More recently within this line of work

Identifying word frequency as an explanation of these two effects (Oh et al., 2024)

More recently within this line of work

Identifying word frequency as an explanation of these two effects (Oh et al., 2024)

Correcting word probability calculation protocols from LMs (Oh & Schuler, 2024a)

More recently within this line of work

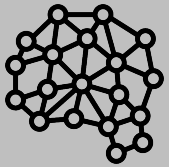
Identifying word frequency as an explanation of these two effects (Oh et al., 2024)

Correcting word probability calculation protocols from LMs (Oh & Schuler, 2024a)

Evaluating whether RT data has been leaked in LM training data (Oh et al., under review)

Today's talk: Part #2

Linguistic data



- Underlying computations are largely unobservable

Psycholinguistics

- Computations are uninterpretable to the human researcher

NLP/Interpretability

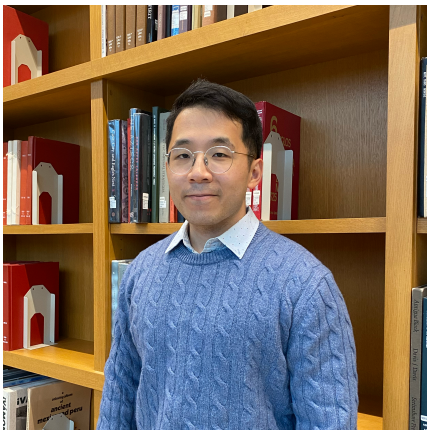
We want to understand a model's predictions

We want to understand a model's predictions

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

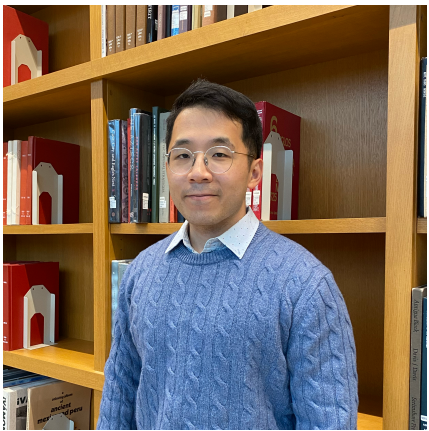
We want to understand a model's predictions

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



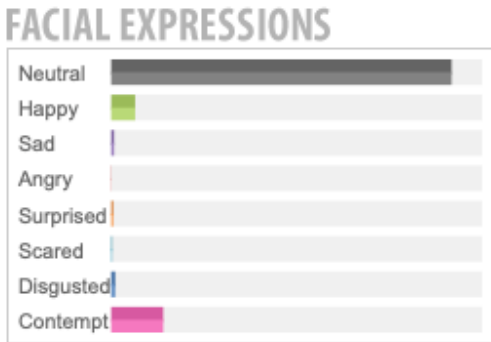
We want to understand a model's predictions

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



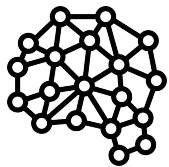
→

Contempt?



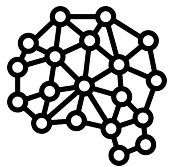
Results from http://noldus.com

This applies to language models as well

I landed in Changi $\rightarrow$  $\rightarrow$ Airport

Language models

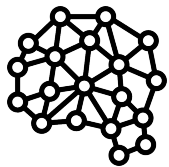
This applies to language models as well

I landed in Changi $\rightarrow$  $\rightarrow$ Airport

Language models

What in the input sequence led to the prediction of Airport?

This applies to language models as well

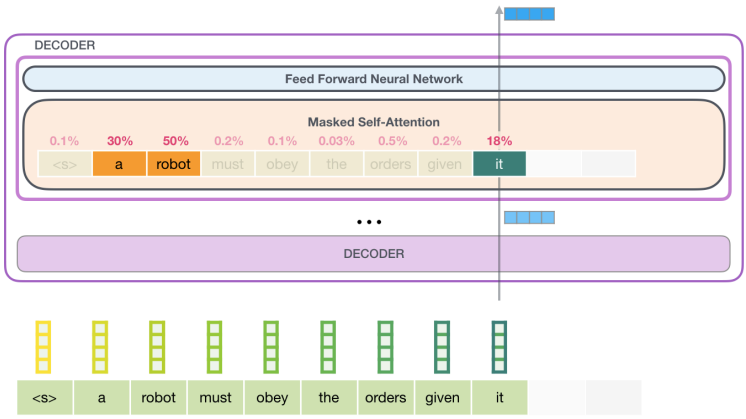
I landed in Changi $\rightarrow$  $\rightarrow$ Airport

Language models

What in the input sequence led to the prediction of Airport?

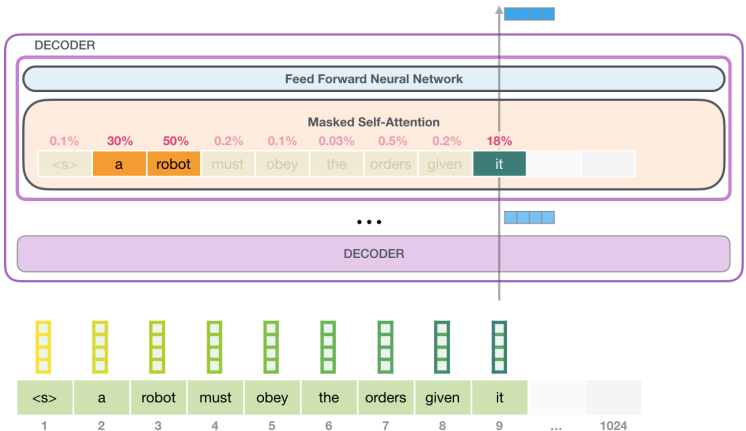
Feature attribution methods aim to attribute the prediction to specific input features

Many contemporary LMs are based on Transformers (Vaswani et al., 2017)



From https://jalammar.github.io/illustrated-gpt2/

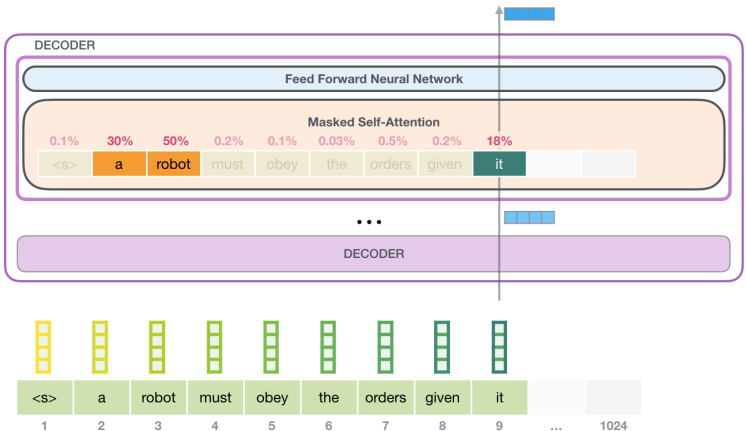
Many contemporary LMs are based on Transformers (Vaswani et al., 2017)



From https://jalammar.github.io/illustrated-gpt2/

Self-attention:
Weighted average of all representations

Many contemporary LMs are based on Transformers (Vaswani et al., 2017)



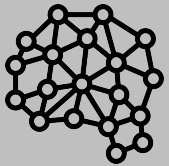
From https://jalammar.github.io/illustrated-gpt2/

Self-attention:
Weighted average of all representations

Feedforward NN:
Non-linear transform of representation

Today's talk: Part #2

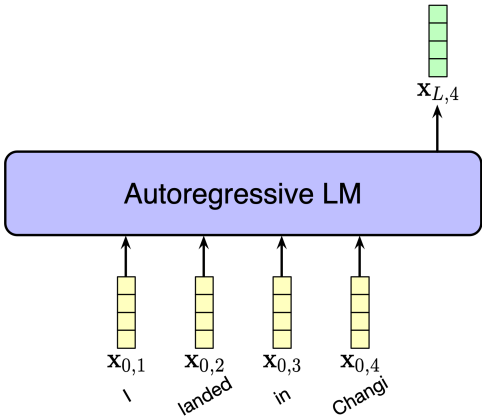
Linguistic data



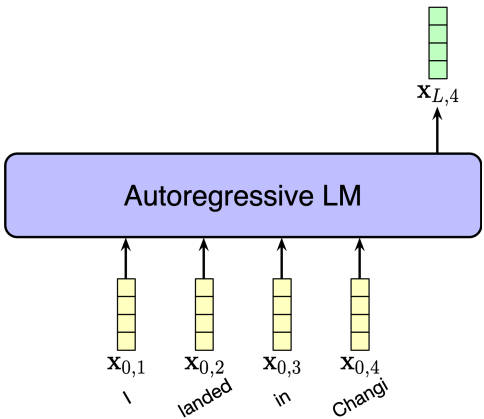
Psycholinguistics *NLP/Interpretability*

Oh and Schuler (2023c). 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*.

The challenge

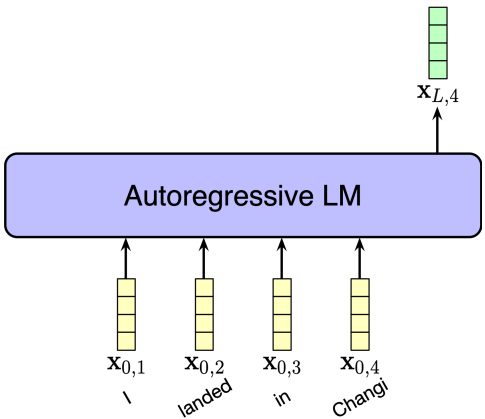


The challenge



Transformers mix representations with non-linear functions

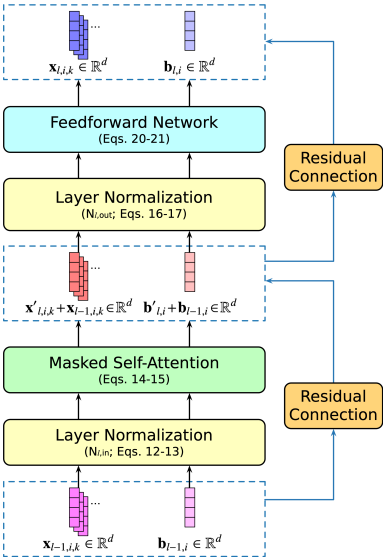
The challenge



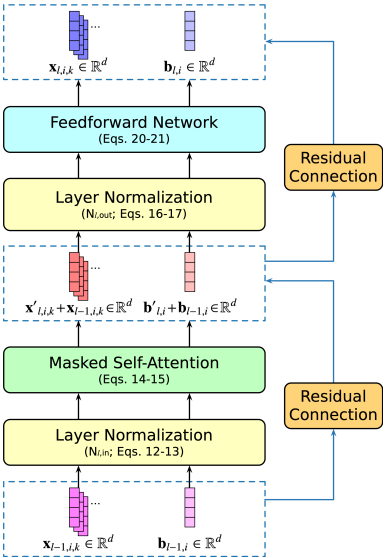
Transformers mix representations with non-linear functions

This means the output cannot be decomposed into the sum of input

This work presents a linear approximation of Transformers

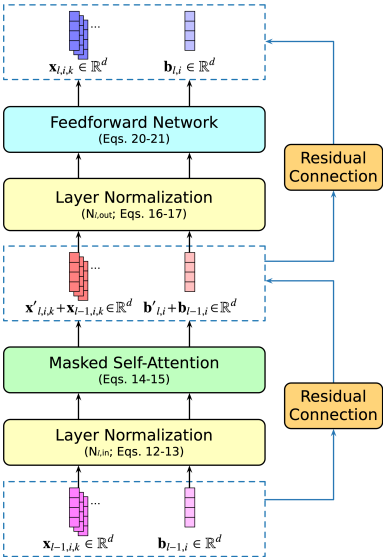


This work presents a linear approximation of Transformers



- Self-attention:
  Attention weights treated as fixed scalars

This work presents a linear approximation of Transformers

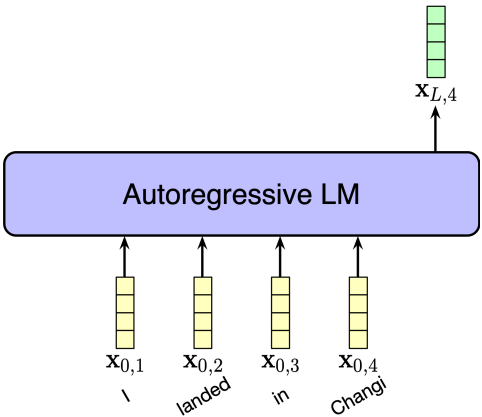


- Self-attention:
  Attention weights treated as fixed scalars

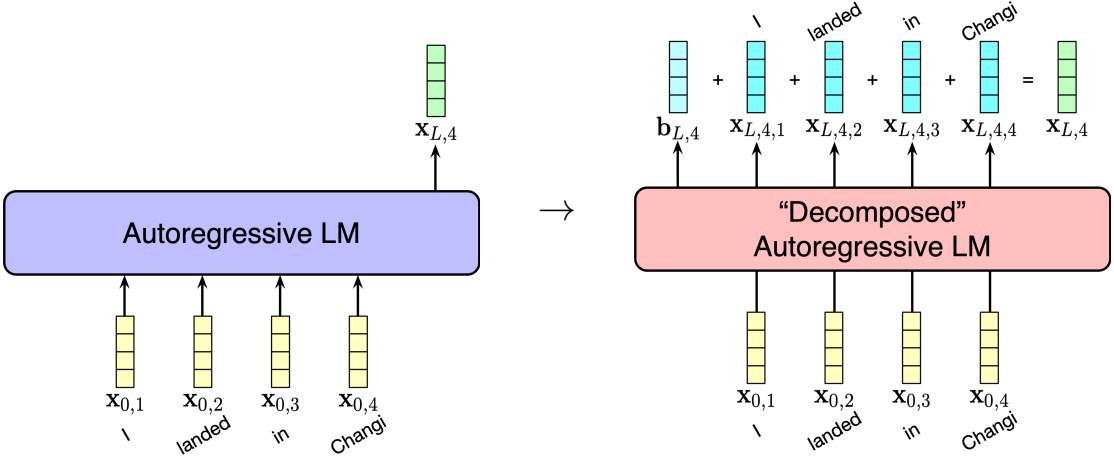
- Feedforward NN is approximated by:

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

This allows the output to be decomposed

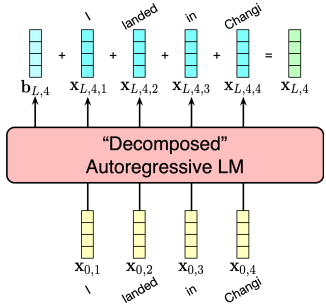


This allows the output to be decomposed



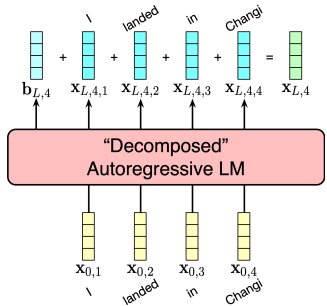
Methods: Characterizing the most important words

I landed in Changi $\rightarrow$



$\rightarrow$ *Airport*

Methods: Characterizing the most important words

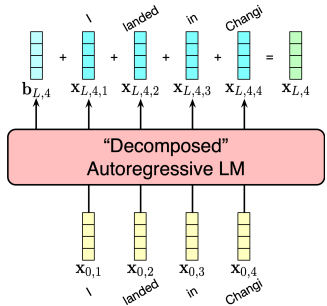
I landed in Changi $\rightarrow$ 

$\rightarrow$ *Airport*

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

Methods: Characterizing the most important words

I landed in Changi $\rightarrow$



$\rightarrow$ *Airport*

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

Methods: Characterizing the most important words

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

Methods: Characterizing the most important words

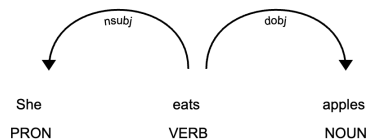
2. Annotate (*Changi*, *Airport*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference

- Pointwise mutual information (PMI): $\log_2 \frac{P(x,y)}{P(x)P(y)}$, where x, y are words

Methods: Characterizing the most important words

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

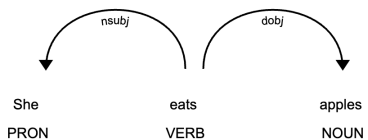
- Pointwise mutual information (PMI): $\log_2 \frac{P(x,y)}{P(x)P(y)}$, where x, y are words
- Dependency:



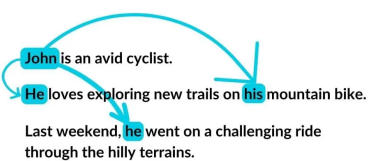
Methods: Characterizing the most important words

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

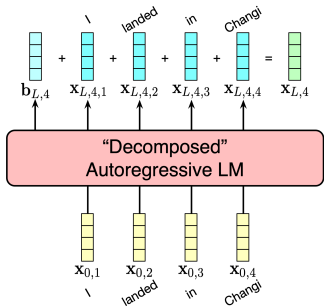
- Pointwise mutual information (PMI): $\log_2 \frac{P(x,y)}{P(x)P(y)}$, where x, y are words
- Dependency:



- Coreference:



Methods: Characterizing the most important words

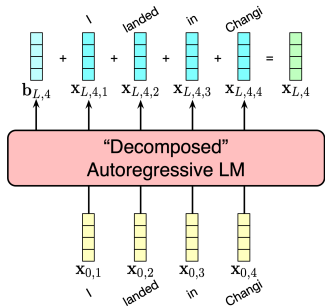
I landed in Changi $\rightarrow$ 

$\rightarrow$ *Airport*

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

Methods: Characterizing the most important words

I landed in Changi $\rightarrow$



$\rightarrow$ *Airport*

1. Ablate each blue vector and find what causes the largest drop in P(*Airport* | ...)
2. Annotate (*Changi*, *Airport*) with $\text{PMI}_{\text{bigram}}$, $\text{PMI}_{\text{document}}$, dependency, coreference
3. Fit stepwise linear regression to the drop in P(*Airport* | ...)

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

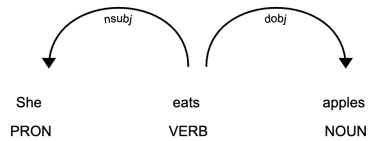
Results from OPT-125M LM on CoNLL-2012 corpus

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

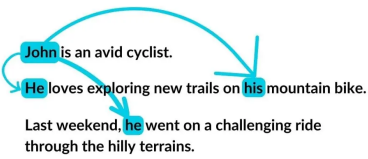
Predictor	Rank	Increase in LogLik
$\text{PMI}_{\text{bigram}}$	1	6151.262*
$\text{PMI}_{\text{document}}$	2	3194.815*
Dependency	3	1981.778*
Coreference	4	25.883*

Follow-up experiments

- Dependency:

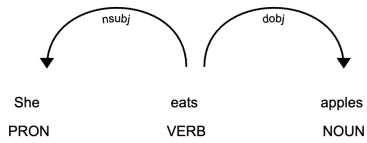


- Coreference:

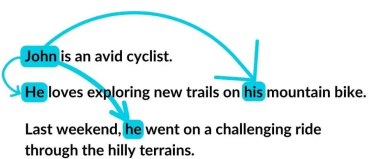


Follow-up experiments

- Dependency:



- Coreference:



These relationships seem less important for LMs; is this true for all subtypes?

Dependency: Important when they are high-PMI

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

Coreference: Important 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

Summary: Decomposition of LMs

Framework for decomposing representations in Transformers

Summary: Decomposition of LMs

Framework for decomposing representations in Transformers

LMs seem to rely on collocational associations and repetitions

Summary: Decomposition of LMs

Framework for decomposing representations in Transformers

LMs seem to rely on collocational associations and repetitions

But these do overlap with some dependency/coreference relationships

Summary: Decomposition of LMs

Framework for decomposing representations in Transformers

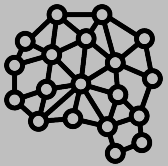
LMs seem to rely on collocational associations and repetitions

But these do overlap with some dependency/coreference relationships

Sheds light on the predictive mechanism underlying 'blackbox' LMs

Today's talk: Part #2

Linguistic data



Psycholinguistics *NLP/Interpretability*

Oh and Schuler (in revision). The attenuation of frequency effects in large language models. *Journal of Memory and Language*.

How do these processing mechanisms interact with model/data size?

How do these processing mechanisms interact with model/data size?

We know from previous work that:

How do these processing mechanisms interact with model/data size?

We know from previous work that:

- Transformers can easily repeat input tokens: $B \ldots B$

How do these processing mechanisms interact with model/data size?

We know from previous work that:

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

How do these processing mechanisms interact with model/data size?

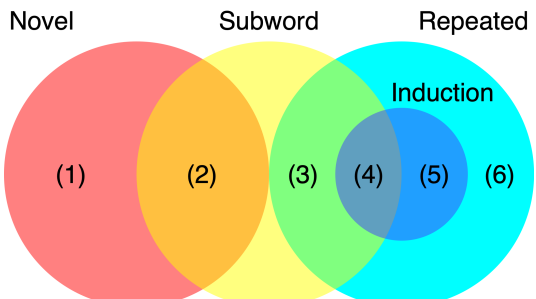
We know from previous work that:

- Transformers can easily repeat input tokens: $B \ldots B$
- Transformers can induce bigram patterns (Elhage et al., 2021): $AB \ldots AB$
- LMs typically use subword tokenization: *miller*

How do these processing mechanisms interact with model/data size?

We know from previous work that:

- Transformers can easily repeat input tokens: $B \ldots B$
- Transformers can induce bigram patterns (Elhage et al., 2021): $AB \ldots AB$
- LMs typically use subword tokenization: *mill*er



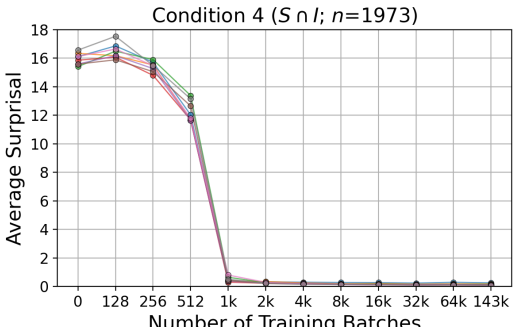
How do these processing mechanisms interact with model/data size?

Surprisal from Pythia LMs at various points during training

How do these processing mechanisms interact with model/data size?

Surprisal from Pythia LMs at various points during training

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

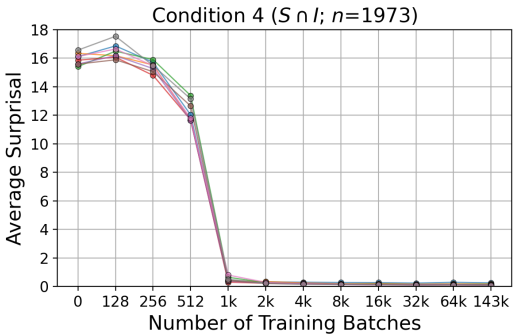


70M 160M 410M 1B 1.4B 2.8B 6.9B 12B

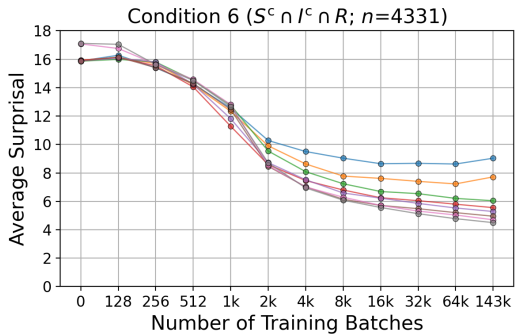
How do these processing mechanisms interact with model/data size?

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

More recently within this line of work

Manipulating the granularity of LM's predictions (from characters to words; Oh & Schuler, 2024b)

More recently within this line of work

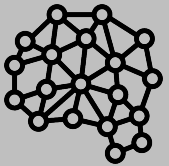
Manipulating the granularity of LM's predictions (from characters to words; Oh & Schuler, 2024b)

Studying the 'processing mechanisms' of state space models (e.g. Dao & Gu, 2024)

Conclusion

Computational modeling

Linguistic data



Psycholinguistics

NLP/Interpretability

Future directions for the Oh Research Group (name tentative)

1) Advancing computational models of human language processing

Addressing the limitations of LMs (e.g. perfect memory of previous words; Clark, Oh, & Schuler, 2025)

1) Advancing computational models of human language processing

Addressing the limitations of LMs (e.g. perfect memory of previous words; Clark, Oh, & Schuler, 2025)

Conducting more fine-grained analyses of eye movements during reading

1) Advancing computational models of human language processing

Addressing the limitations of LMs (e.g. perfect memory of previous words; Clark, Oh, & Schuler, 2025)

Conducting more fine-grained analyses of eye movements during reading

Modeling individual variation in language processing

2) Studying the representations of NLP models

Developing methods for identifying syntactic/semantic interpretations within LMs

2) Studying the representations of NLP models

Developing methods for identifying syntactic/semantic interpretations within LMs

- Garden path sentences: *The boy fed the chicken stayed ...*
- Semantic anomalies: *The manager forgot which waitress the customer served ...*

2) Studying the representations of NLP models

Developing methods for identifying syntactic/semantic interpretations within LMs

- Garden path sentences: *The boy fed the chicken stayed ...*
- Semantic anomalies: *The manager forgot which waitress the customer served ...*

Intervening on the models' representations for controlled text generation

2) Studying the representations of NLP models

Developing methods for identifying syntactic/semantic interpretations within LMs

- Garden path sentences: *The boy fed the chicken stayed ...*
- Semantic anomalies: *The manager forgot which waitress the customer served ...*

Intervening on the models' representations for controlled text generation

Leveraging linguistic structure for efficient model training

3) Supporting ongoing research at NTU LMS

Developing computational models of SLA, code-switching, language contact, ...

3) Supporting ongoing research at NTU LMS

Developing computational models of SLA, code-switching, language contact, ...

Applying NLP/automated tools to conversation/discourse analysis

3) Supporting ongoing research at NTU LMS

Developing computational models of SLA, code-switching, language contact, ...

Applying NLP/automated tools to conversation/discourse analysis

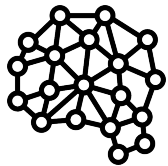
Supporting documentation efforts for under-resourced languages

Thank you for listening!

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

Computational modeling

Linguistic data



Psycholinguistics *NLP/Interpretability*

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

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. *Proceedings of the 26th Conference on Computational Natural Language Learning*, 301–313.

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.

Clark, C., Oh, B.-D., & Schuler, W. (2025). Linear recency bias during training improves transformers' fit to reading times. *Proceedings of the 31st International Conference on Computational Linguistics*, 7735–7747.

Dao, T., & Gu, A. (2024). Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 10041–10071.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

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*.

References II

Futrell, R., Gibson, E., Tily, H. J., Blank, I., Vishnevetsky, A., Piantadosi, S., & Fedorenko, E. (2021). The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, *55*(1), 63–77.

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *arXiv preprint*, *arXiv:2501.17047*.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, 10–18.

Hahn, M., Futrell, R., Gibson, E., & Levy, R. P. (2022). A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, *119*(43), e2122602119.

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*.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Kennedy, A., Hill, R., & Pynte, J. (2003). The Dundee Corpus. *Proceedings of the 12th European Conference on Eye Movement*.

References III

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, *28*(6), 517–540.

Oh, B.-D., Clark, C., & Schuler, W. (2021). Surprisal estimators for human reading times need character models. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 3746–3757.

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963.

Oh, B.-D., & Schuler, W. (2023a). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Oh, B.-D., & Schuler, W. (2023b). Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1915–1921.

References IV

Oh, B.-D., & Schuler, W. (2023c). 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.

Oh, B.-D., & Schuler, W. (2024a). Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 3464–3472.

Oh, B.-D., & Schuler, W. (2024b). The impact of token granularity on the predictive power of language model surprisal. *arXiv preprint, arXiv:2412.11940*.

Oh, B.-D., & Schuler, W. (in revision). The attenuation of frequency effects in large language models. *Journal of Memory and Language*.

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.

Oh, B.-D., Zhu, H., & Schuler, W. (under review). Assessing the leakage of naturalistic reading time corpora in language model pre-training datasets. *ACL Rolling Review*.

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.

References V

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.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45), e2105646118.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121.

Shain, C., van Schijndel, M., & Schuler, W. (2018). Deep syntactic annotations for broad-coverage psycholinguistic modeling. *Workshop on Linguistic and Neuro-Cognitive Resources*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 6000–6010.

References VI

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 1707–1713.

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*.

Research Seminar

Unveiling the Language Processing of Humans and Machines
By Dr Byung-Doh Oh

Please fill in and submit your feedback here:

https://forms.office.com/r/2nP10gTMpv