The Bigger-is-Worse Effects of Model Size and Training Data of Large Language Model Surprisal on Human Reading Times

Byung-Doh Oh[1]

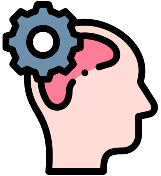
Department of Linguistics
The Ohio State University

25 April 2024
Universität des Saarlandes & SFB 1102

THE OHIO STATE UNIVERSITY

[1] Sep. 2024–: Center for Data Science, New York University

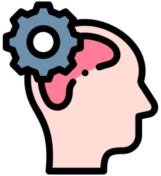
I landed in Frankfurt and took a



I landed in Frankfurt and took a

train

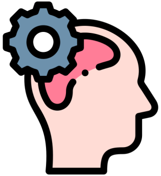
camel



I landed in Frankfurt and took a

train

camel



The more predictable train is easier to process than camel

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

I landed in Frankfurt and took a

train

camel

Human subjects

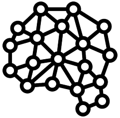
I landed in Frankfurt and took a

train

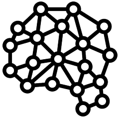
camel

Human subjects

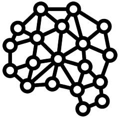
$\sim$



Model 1



Model 2



Model 3

Human subject data: Word-by-word reading times

Human
subjects

Human subject data: Word-by-word reading times

Human
subjects

I ====================================

Human subject data: Word-by-word reading times

Human subjects

```
= landed =============================
```

Human subject data: Word-by-word reading times

Human subjects

`======== in ==========================`

Human subject data: Word-by-word reading times

Human subjects

```
=========== Frankfurt ================
```

Human subjects

===================== and ============

Human subject data: Word-by-word reading times

Human subjects

`========================= took =======`

Human subject data: Word-by-word reading times

Human
subjects

`============================== a =====`

Human subject data: Word-by-word reading times

Human
subjects

`================================ camel`

Human subject data: Word-by-word reading times

Human subjects

`I landed in Frankfurt and took a camel`

Human subject data: Word-by-word reading times

Human subjects

`I landed in Frankfurt and took a camel`

Human subject data: Word-by-word reading times

Human subjects

`I landed in Frankfurt and took a camel`

Assumption: Processing difficulty causes delays in reading times!

Computational models: Large language models (LLMs)

Computational models: Large language models (LLMs)

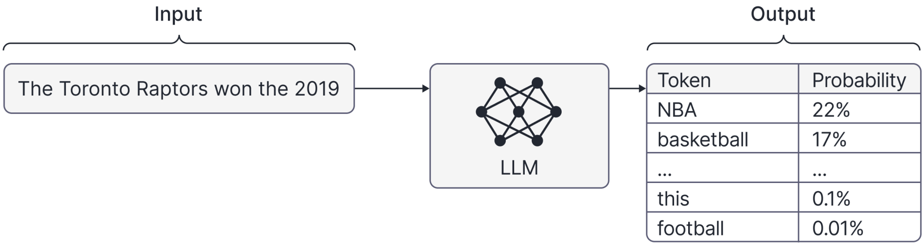
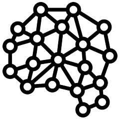


Figure from Borealis AI

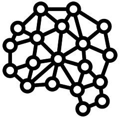
Link between human behavior and LLMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

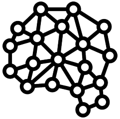
$\sim$



Model 1



Model 2

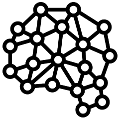


Model 3

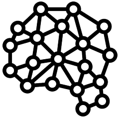
Link between human behavior and LLMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

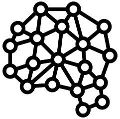
$\sim$



Model 1



Model 2



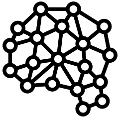
Model 3

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

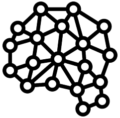
Link between human behavior and LLMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

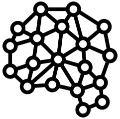
$\sim$



Model 1



Model 2



Model 3

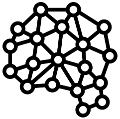
$$\mathrm{RT}(\text{train}) \propto -\log_2 \mathrm{P}(\text{train} \mid \text{I landed in Frankfurt and took a})$$

$$\mathrm{RT}(\text{camel}) \propto -\log_2 \mathrm{P}(\text{camel} \mid \text{I landed in Frankfurt and took a})$$

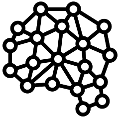
Link between human behavior and LLMs (Surprisal theory; Hale, 2001; Levy, 2008)

Human subjects

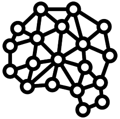
$\sim$



Model 1



Model 2



Model 3

Evaluation: How well does surprisal from Model n fit to human reading times?
(through regression modeling)

Roadmap

Roadmap

1. Phenomenon #1: The bigger-is-worse effect of model size (Oh & Schuler, 2023a)

Roadmap

1. Phenomenon #1: The bigger-is-worse effect of model size (Oh & Schuler, 2023a)
2. Phenomenon #2: The bigger-is-worse effect of training data (Oh & Schuler, 2023b)

Roadmap

1. Phenomenon #1: The bigger-is-worse effect of model size (Oh & Schuler, 2023a)
2. Phenomenon #2: The bigger-is-worse effect of training data (Oh & Schuler, 2023b)
3. Word frequency as a unified explanation (Oh, Yue, & Schuler, 2024)

Roadmap

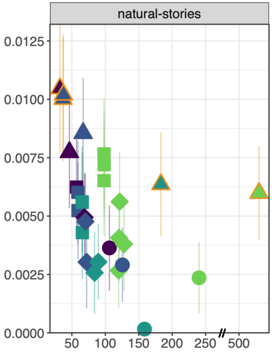
1. Phenomenon #1: The bigger-is-worse effect of model size (Oh & Schuler, 2023a)
2. Phenomenon #2: The bigger-is-worse effect of training data (Oh & Schuler, 2023b)
3. Word frequency as a unified explanation (Oh, Yue, & Schuler, 2024)
4. Conclusion

Phenomenon #1: The bigger-is-worse effect of model size

Oh and Schuler (2023a). Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *TACL*.

Better Fit

Poorer Fit



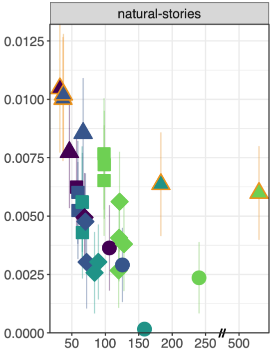
Wilcox et al. (2020)

More Accurate

Less Accurate

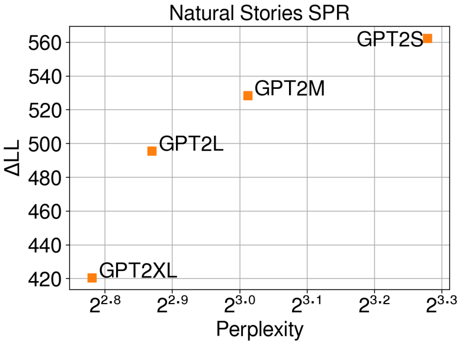
Better Fit

Poorer Fit



Wilcox et al. (2020)

More Accurate ⟷ Less Accurate



Oh, Clark, and Schuler (2022)

More Accurate, Larger ⟷ Less Accurate, Smaller

Replication with more LLM families

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

Replication with more LLM 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

Replication with more LLM 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: LLM surprisal

Model	#L	#H	d_{model}
GPT-2 Small	12	12	768
GPT-2 Medium	24	16	1024
GPT-2 Large	36	20	1280
GPT-2 XL	48	25	1600
GPT-Neo 125M	12	12	768
GPT-Neo 1.3B	24	16	2048
GPT-Neo 2.7B	32	20	2560
GPT-J 6B	28	16	4096
GPT-NeoX 20B	44	64	6144
OPT 125M	12	12	768
OPT 350M	24	16	1024
OPT 1.3B	24	32	2048
OPT 2.7B	32	32	2560
OPT 6.7B	32	32	4096
OPT 13B	40	40	5120
OPT 30B	48	56	7168
OPT 66B	64	72	9216

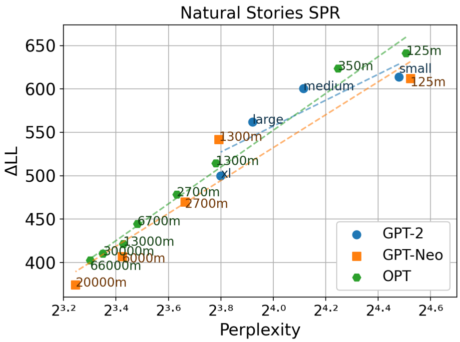
Replication with more LLM families

- 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: LLM surprisal
- Evaluation metric: Δlog-likelihood (ΔLL); how well does surprisal fit to RT?

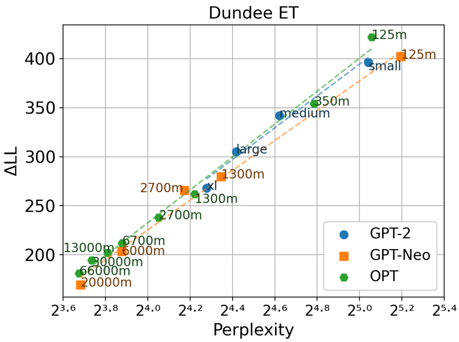
Model	#L	#H	d_{model}
GPT-2 Small	12	12	768
GPT-2 Medium	24	16	1024
GPT-2 Large	36	20	1280
GPT-2 XL	48	25	1600
GPT-Neo 125M	12	12	768
GPT-Neo 1.3B	24	16	2048
GPT-Neo 2.7B	32	20	2560
GPT-J 6B	28	16	4096
GPT-NeoX 20B	44	64	6144
OPT 125M	12	12	768
OPT 350M	24	16	1024
OPT 1.3B	24	32	2048
OPT 2.7B	32	32	2560
OPT 6.7B	32	32	4096
OPT 13B	40	40	5120
OPT 30B	48	56	7168
OPT 66B	64	72	9216

Better Fit

Poorer Fit



More Accurate, Larger ⟷ Less Accurate, Smaller



More Accurate, Larger ⟷ Less Accurate, Smaller

What linguistic factors drive this trend?

- Subsets defined by word-level and syntactic properties (Shain et al., 2018)

What linguistic factors drive this trend?

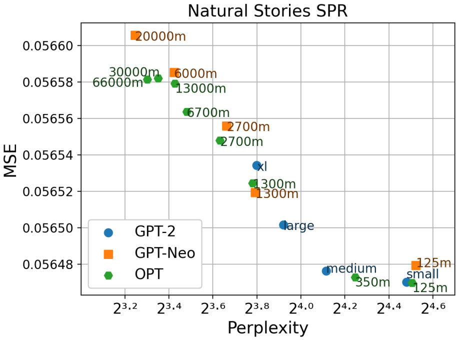
- Subsets defined by word-level and syntactic properties (Shain et al., 2018)
- Subsets with the largest differences in MSE between models identified

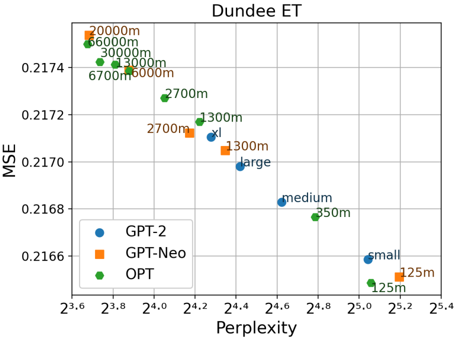
What linguistic factors drive this trend?

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

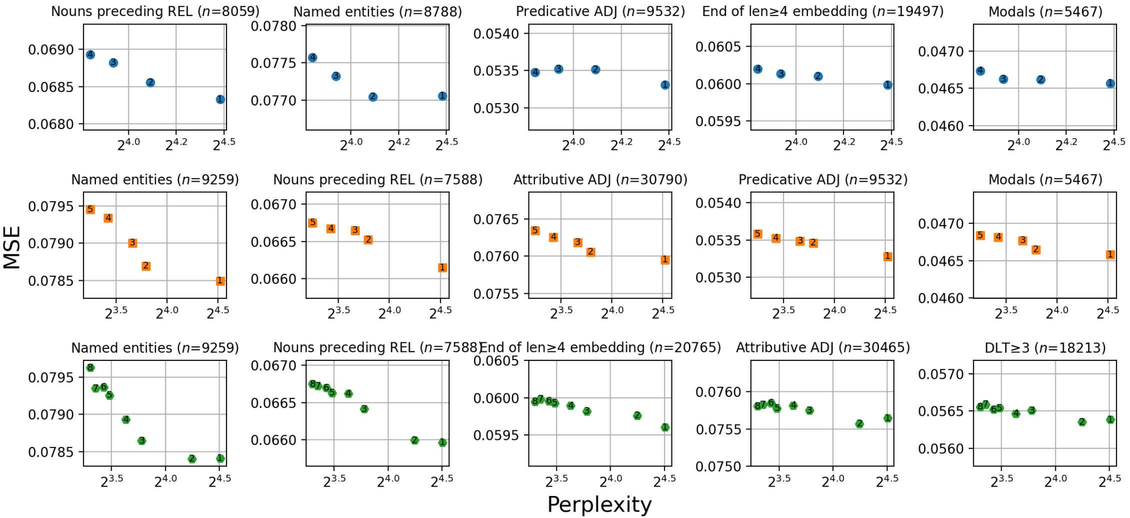
Poorer Fit

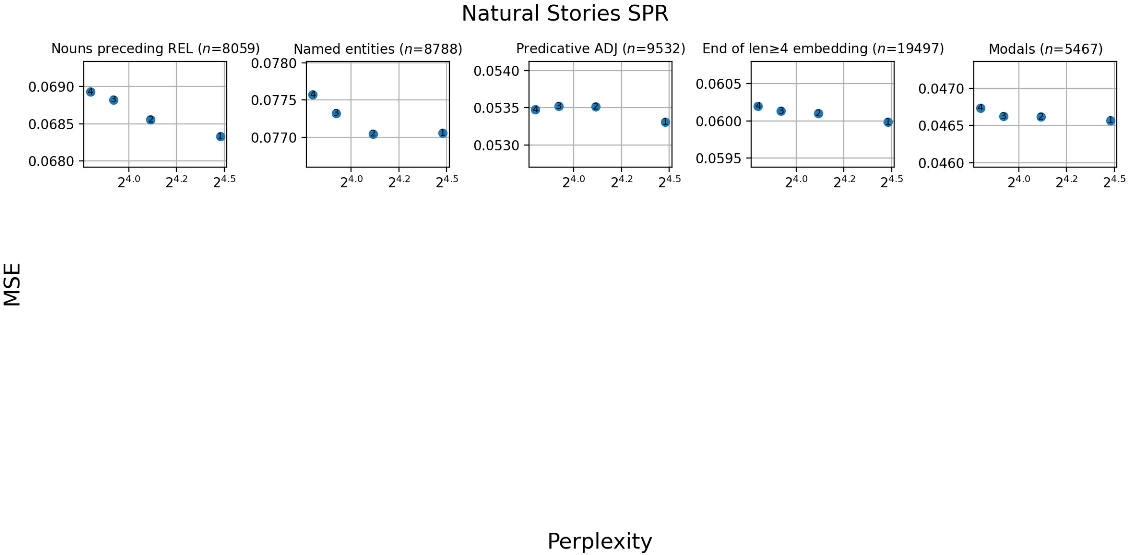
Better Fit



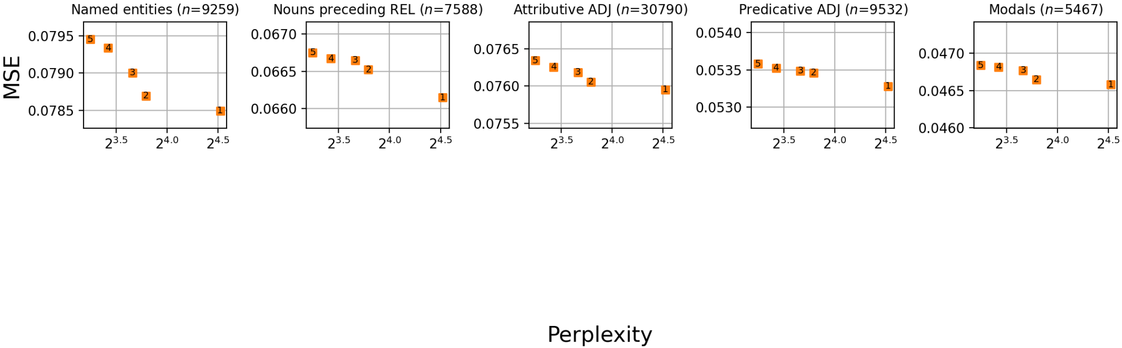


Natural Stories SPR

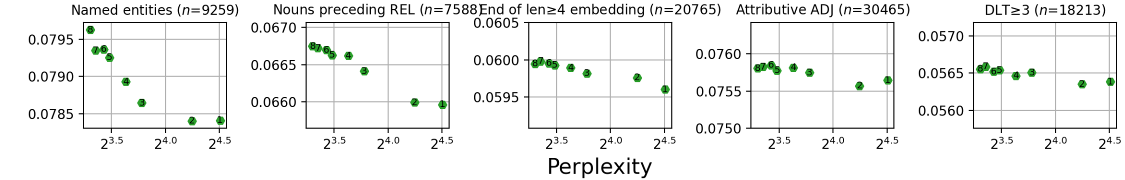



Natural Stories SPR
Nouns preceding REL (n=8059)
Named entities (n=8788)
Predicative ADJ (n=9532)
End of len≥4 embedding (n=19497)
Modals (n=5467)
MSE
Perplexity

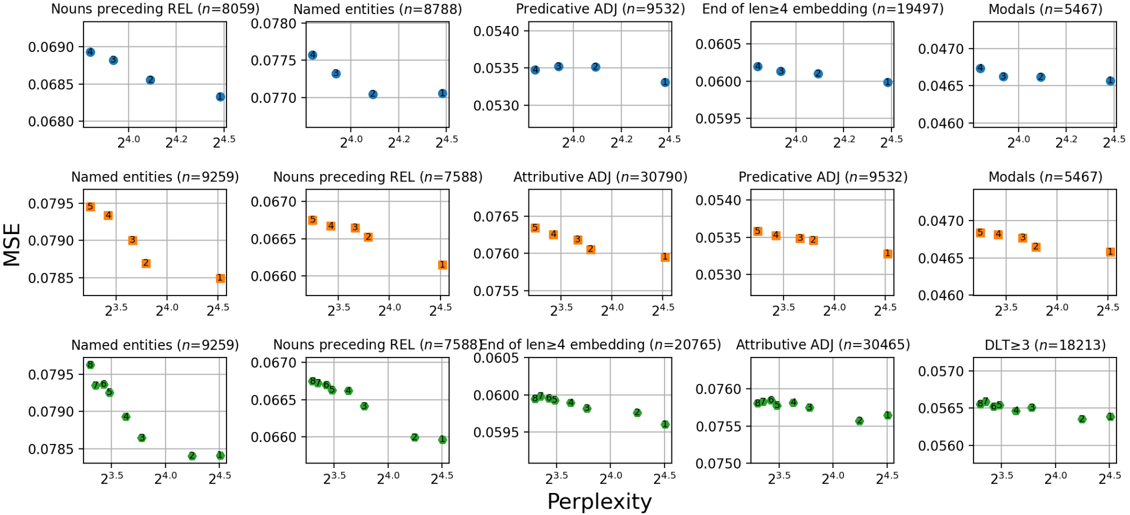
Natural Stories SPR



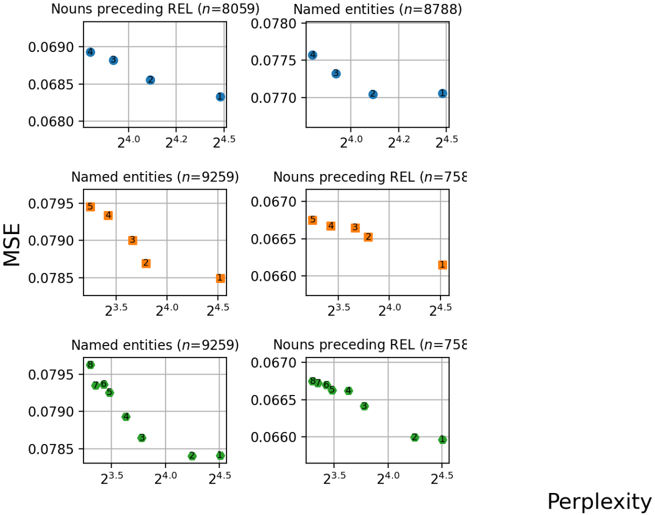
Natural Stories SPR



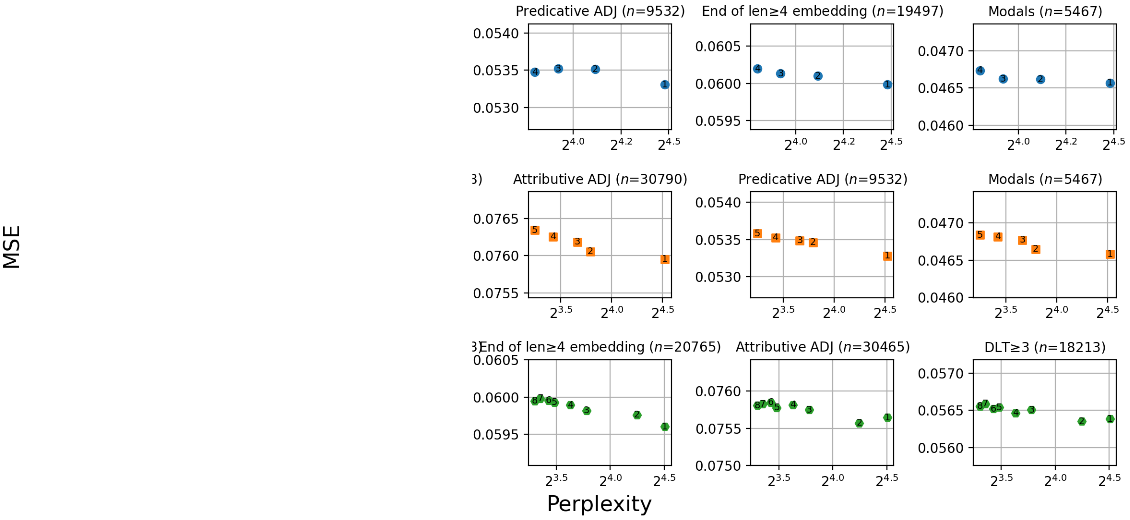
Natural Stories SPR



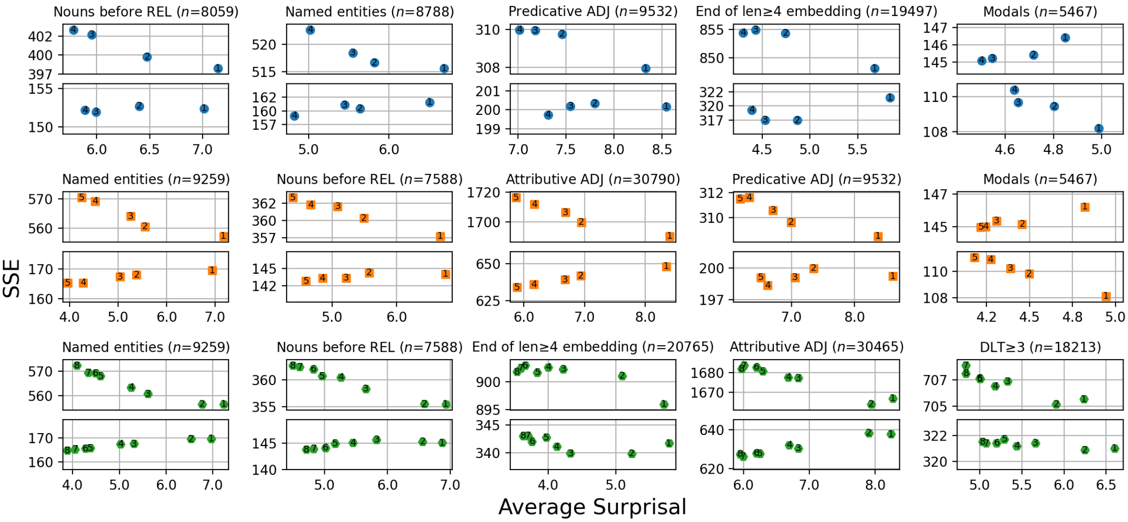
Natural Stories SPR

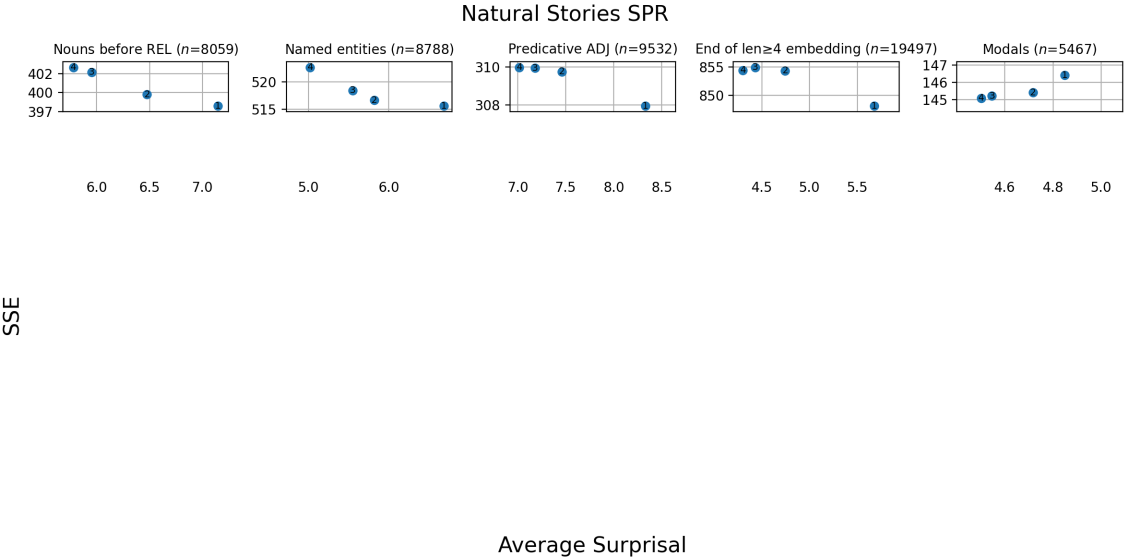


Natural Stories SPR

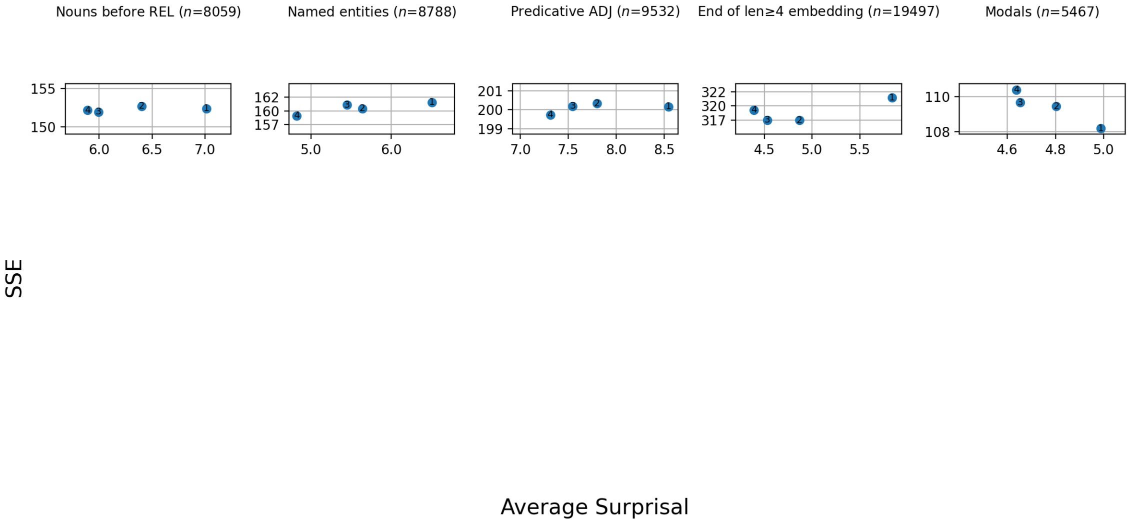


Natural Stories SPR

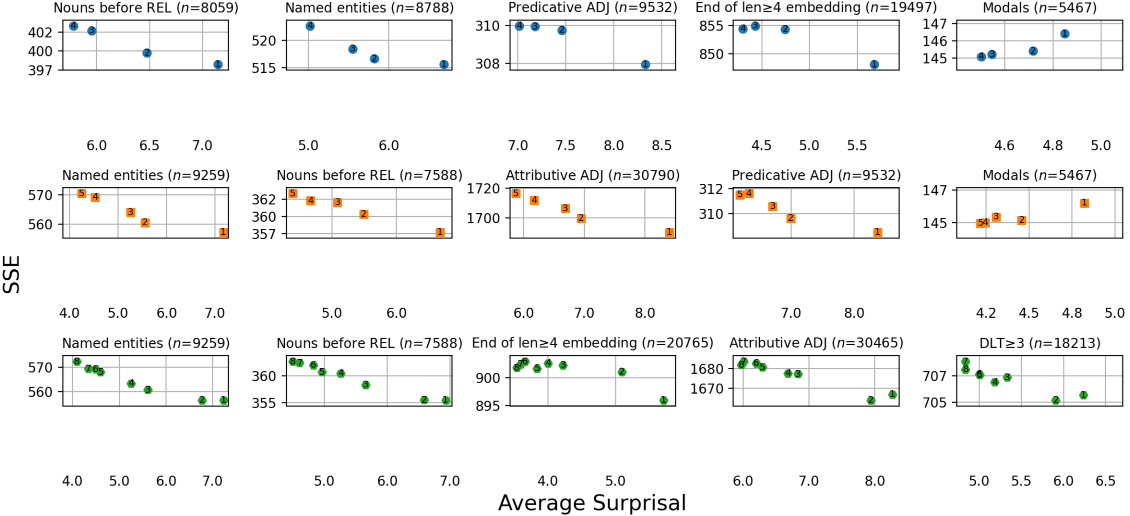



Natural Stories SPR
Nouns before REL (n=8059)
Named entities (n=8788)
Predicative ADJ (n=9532)
End of len≥4 embedding (n=19497)
Modals (n=5467)
402
400
397
520
515
310
308
855
850
147
146
145
6.0
6.5
7.0
5.0
6.0
7.0
7.5
8.0
8.5
4.5
5.0
5.5
4.6
4.8
5.0
SSE
Average Surprisal

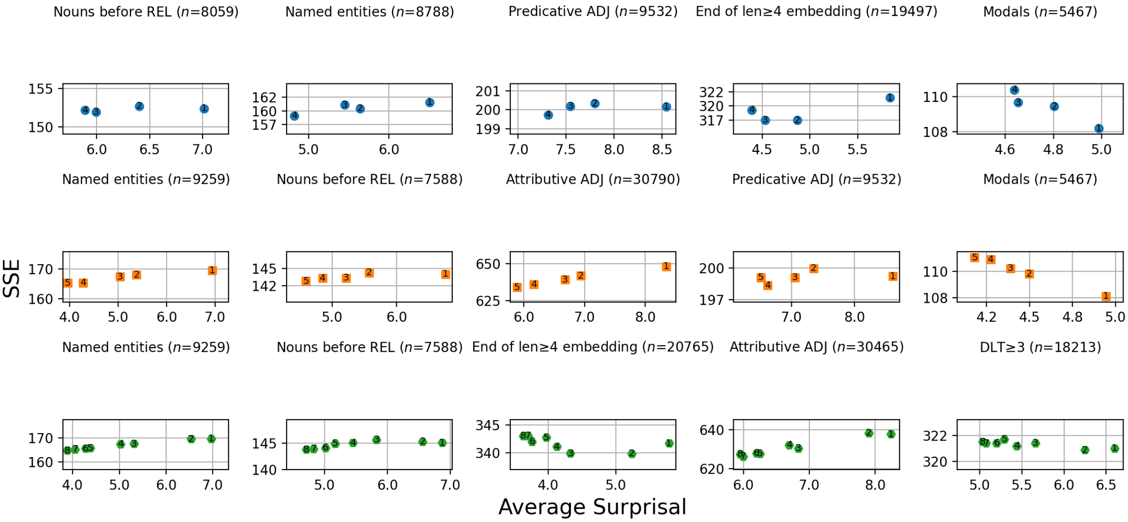
Natural Stories SPR



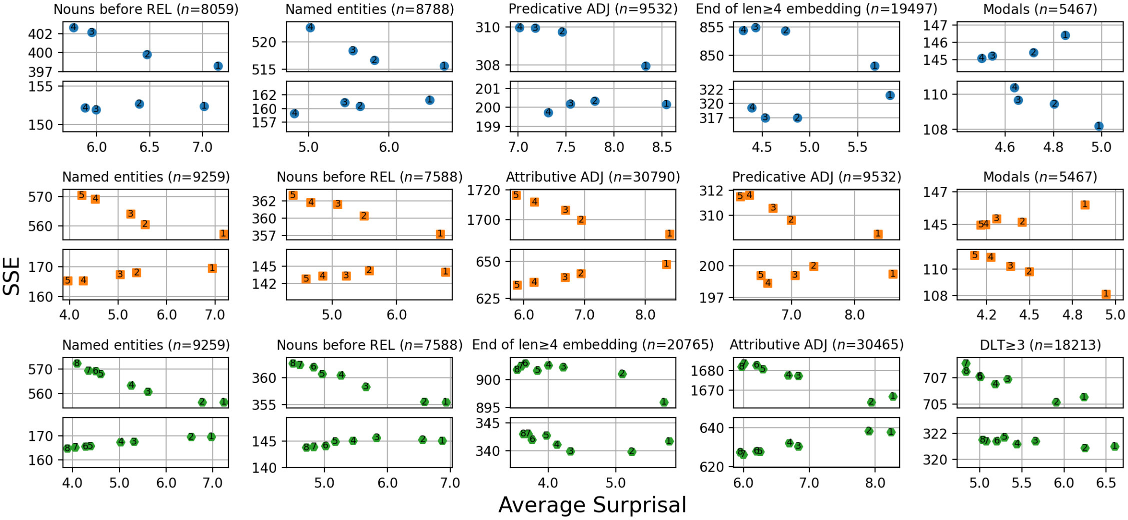
Natural Stories SPR



Natural Stories SPR



Natural Stories SPR



Summary: Bigger-is-worse effect of model size

- Surprisal from larger models show strictly poorer fits to human reading times

Summary: Bigger-is-worse effect of model size

- Surprisal from larger models 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: Bigger-is-worse effect of model size

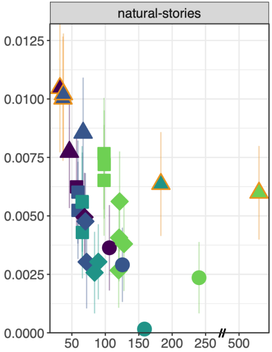
- Surprisal from larger models show strictly poorer fits to human reading times
- Effect mostly driven by underprediction of reading times by LLM surprisal
  (see e.g. Arehalli et al., 2022; Hahn et al., 2022; van Schijndel & Linzen, 2021)
- Likely due to extensive domain knowledge from massive amounts of training examples

Phenomenon #2: The bigger-is-worse effect of training data

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

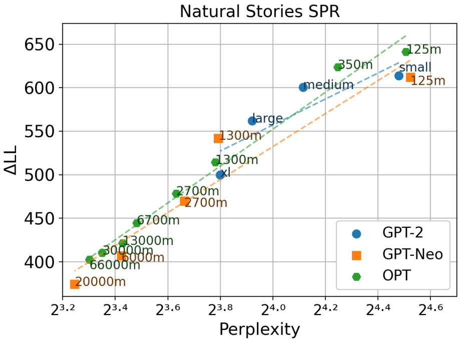
Better Fit

Poorer Fit



Wilcox et al. (2020)

More Accurate, Larger ⟷ Less Accurate, Smaller



Oh and Schuler (2023a)

More Accurate, Larger ⟷ Less Accurate, Smaller

Evaluating LLMs trained on less data

- Regression models fit and ΔLL calculated

Evaluating LLMs trained on less data

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

Model	#L	#H	d_{model}
Pythia 70M	6	8	512
Pythia 160M	12	12	768
Pythia 410M	24	16	1024
Pythia 1B	16	8	2048
Pythia 1.4B	24	16	2048
Pythia 2.8B	32	32	2560
Pythia 6.9B	32	32	4096
Pythia 12B	36	40	5120

Evaluating LLMs trained on less data

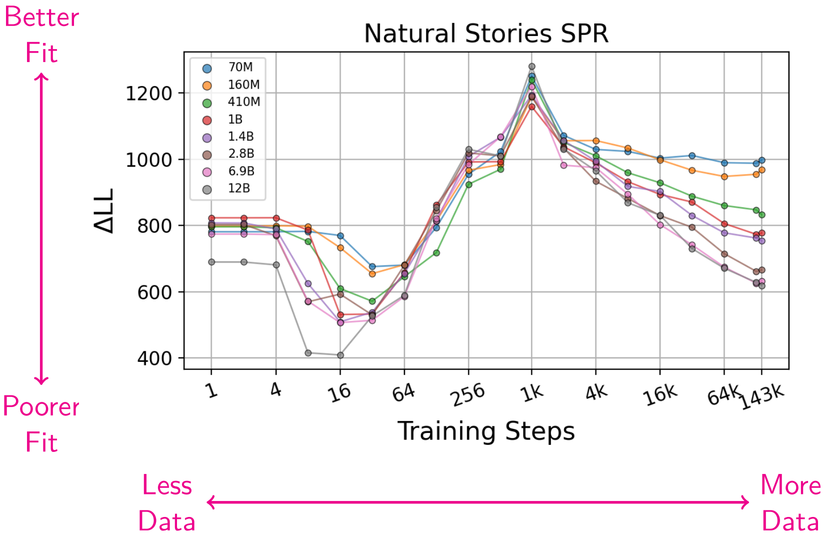
- Regression models fit and ΔLL calculated
- Predictors of interest: LLM surprisal
- Trained on identical batches of 1024×2048 ($\sim$2 million) tokens

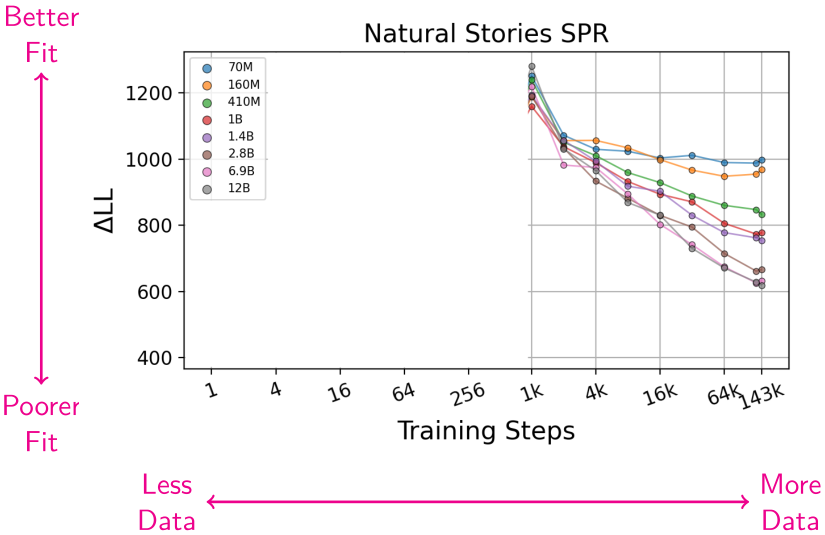
Model	#L	#H	d_{model}
Pythia 70M	6	8	512
Pythia 160M	12	12	768
Pythia 410M	24	16	1024
Pythia 1B	16	8	2048
Pythia 1.4B	24	16	2048
Pythia 2.8B	32	32	2560
Pythia 6.9B	32	32	4096
Pythia 12B	36	40	5120

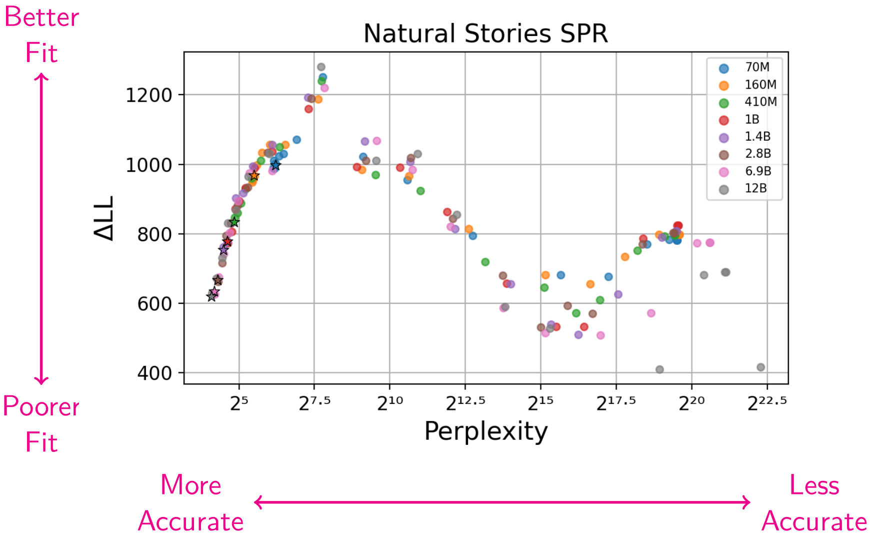
Evaluating LLMs trained on less data

- Regression models fit and ΔLL calculated
- Predictors of interest: LLM surprisal
- Trained on identical batches of 1024×2048 ($\sim$2 million) tokens
- Checkpoints available after {1, 2, 4, ..., 512, 1000, 2000, ..., 143000} batches

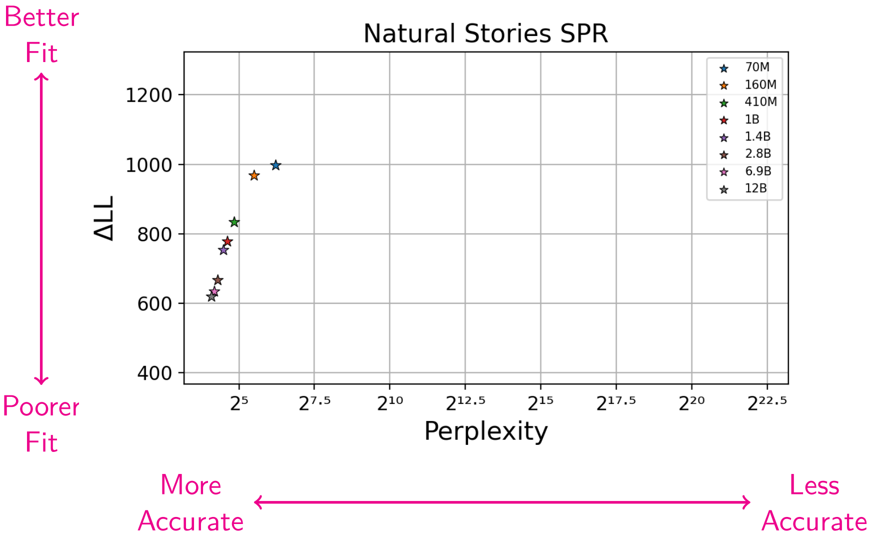
Model	#L	#H	d_{model}
Pythia 70M	6	8	512
Pythia 160M	12	12	768
Pythia 410M	24	16	1024
Pythia 1B	16	8	2048
Pythia 1.4B	24	16	2048
Pythia 2.8B	32	32	2560
Pythia 6.9B	32	32	4096
Pythia 12B	36	40	5120

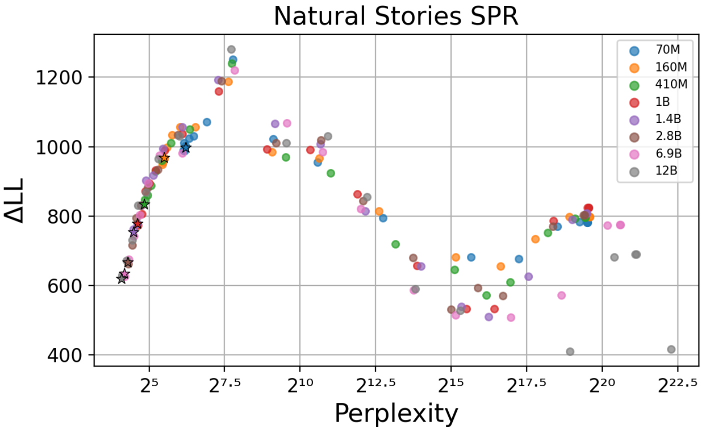

Better
Fit
Poorer
Fit
Natural Stories SPR
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
ΔLL
1200
1000
800
600
400
1
4
16
64
256
1k
4k
16k
64k
143k
Training Steps
Less
Data
More
Data


Better
Fit
Poorer
Fit
Natural Stories SPR
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
1200
1000
800
600
400
ΔLL
1
4
16
64
256
1k
4k
16k
64k
143k
Training Steps
Less
Data
More
Data



Better
Fit
Poorer
Fit
Natural Stories SPR
ΔLL
1200
1000
800
600
400
2^5
2^7.5
2^10
2^12.5
2^15
2^17.5
2^20
2^22.5
Perplexity
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
More
Accurate
Less
Accurate


Better Fit
Poorer Fit
Natural Stories SPR
ΔLL
1200
1000
800
600
400
2^5
$2^{7.5}$
2^{10}
$2^{12.5}$
2^{15}
$2^{17.5}$
2^{20}
$2^{22.5}$
Perplexity
70M
160M
410M
1B
1.4B
2.8B
6.9B
12B
More Accurate
Less Accurate



Better Fit

Poorer Fit

More Accurate

Less Accurate

How small can we go?

- Smaller LMs trained following the procedures of the Pythia LM

How small can we go?

- Smaller LMs trained following the procedures of the Pythia LM

Model	#L	#H	d_{model}	#Parameters
Repro 1-1-64	1	1	64	∼6M
Repro 1-2-128	1	2	128	∼13M
Repro 2-2-128	2	2	128	∼13M
Repro 2-3-192	2	3	192	∼20M
Repro 2-4-256	2	4	256	∼27M
Repro 3-4-256	3	4	256	∼28M
Repro 4-6-384	4	6	384	∼46M
Repro 6-8-512	6	8	512	∼70M

How small can we go?

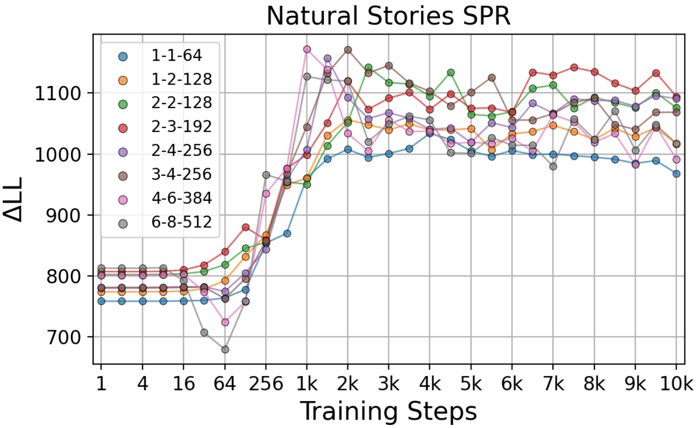
- Smaller LMs trained following the procedures of the Pythia LM

Model	#L	#H	d_{model}	#Parameters
Repro 1-1-64	1	1	64	$\sim$6M
Repro 1-2-128	1	2	128	$\sim$13M
Repro 2-2-128	2	2	128	$\sim$13M
Repro 2-3-192	2	3	192	$\sim$20M
Repro 2-4-256	2	4	256	$\sim$27M
Repro 3-4-256	3	4	256	$\sim$28M
Repro 4-6-384	4	6	384	$\sim$46M
Repro 6-8-512	6	8	512	$\sim$70M

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

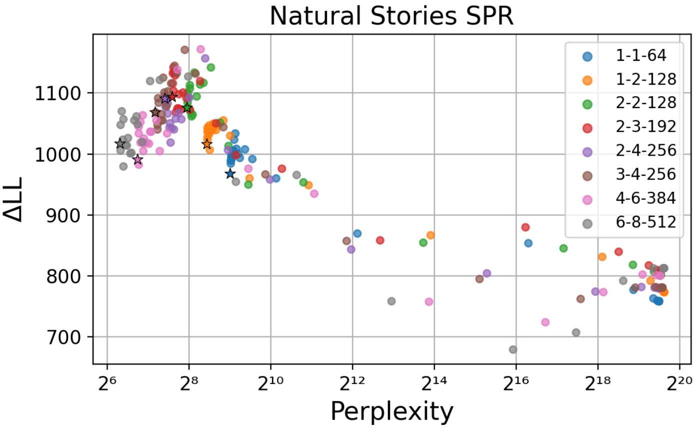
Better Fit

Poorer Fit



Less Data

More Data

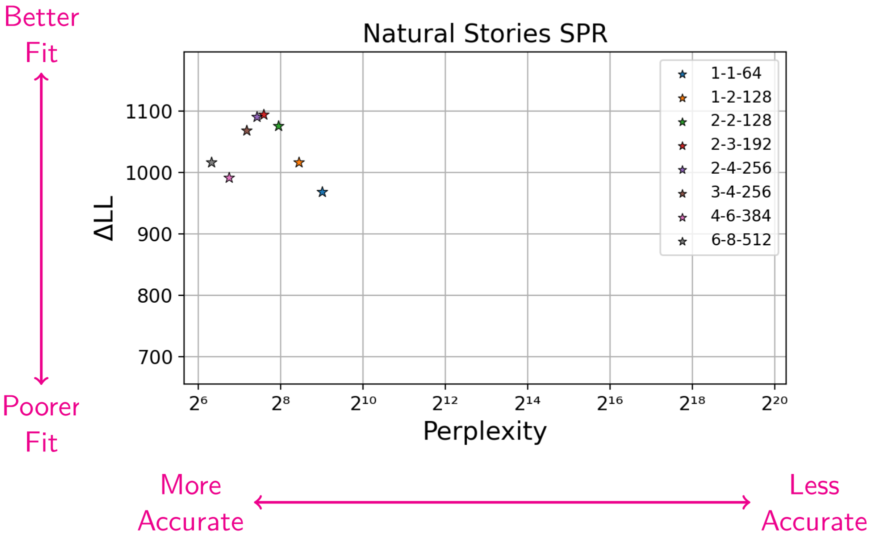


Better Fit

Poorer Fit

More Accurate

Less Accurate


Better
Fit
Poorer
Fit
Natural Stories SPR
ΔLL
1100
1000
900
800
700
2^6
2^8
2^{10}
2^{12}
2^{14}
2^{16}
2^{18}
2^{20}
Perplexity
1-1-64
1-2-128
2-2-128
2-3-192
2-4-256
3-4-256
4-6-384
6-8-512
More
Accurate
Less
Accurate

Summary: Bigger-is-worse effect of training data

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

Summary: 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: Bigger-is-worse effect of training data

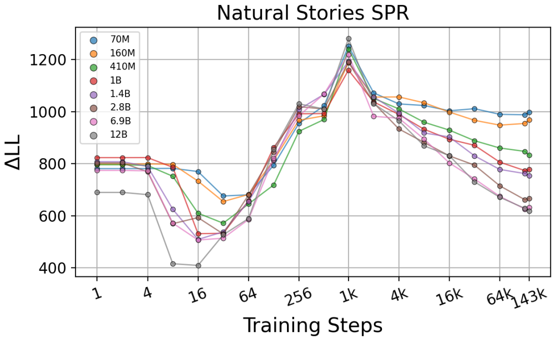
- Fit to reading times starts to degrade after about two billion tokens of training data
- Strong interaction between model size and training data amount after the peak
- Consolidates conflicting results about LM perplexity and fit to reading times

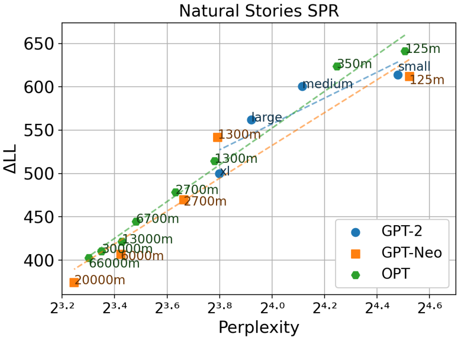
Word frequency as a unified explanation

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. *Proceedings of EACL*.

Better Fit

Poorer Fit





Less Data ⟷ More Data

Larger Models ⟷ Smaller Models

Insights from the scaling behavior of LLMs

- Larger models 'learn faster' given the same amount of exposure (Tirumala et al., 2022)

Insights from the scaling behavior of LLMs

- Larger models ‘learn faster’ given the same amount of exposure (Tirumala et al., 2022)
- Early in training, all models similarly learn to predict frequent function words (Xia et al., 2023)

Insights from the scaling behavior of LLMs

- Larger models ‘learn faster’ given the same amount of exposure (Tirumala et al., 2022)
- Early in training, all models similarly learn to predict frequent function words (Xia et al., 2023)

Word frequency modulates the difference in surprisal estimates as a function of model size and training data amount, which drives their adverse effects on fit to human reading times.

Revisiting the bigger-is-worse effect of model size

- LME models fit to reading times of Natural Stories, Dundee, Ghent, and Provo corpora (Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018)

Revisiting the bigger-is-worse effect of model size

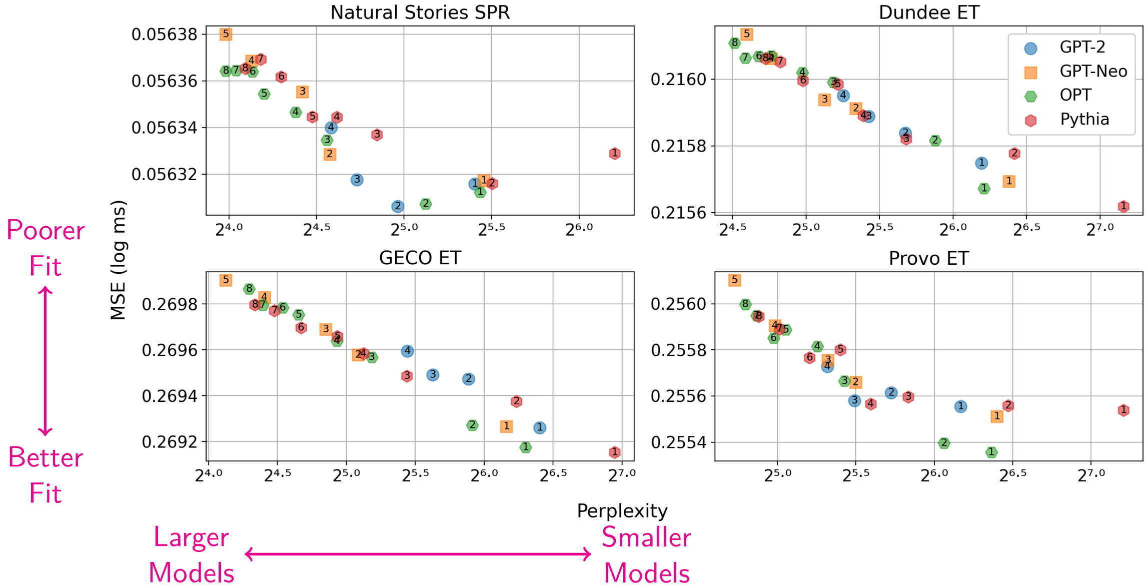
- LME models fit to reading times of Natural Stories, Dundee, Ghent, and Provo corpora (Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018)
- Baseline predictors: Word length/position, unigram surprisal (The Pile; Gao et al., 2020), saccade length, previous word fixated

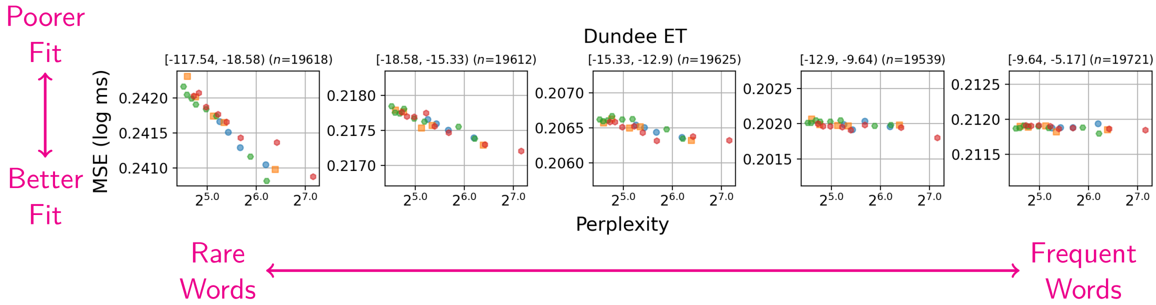
Revisiting the bigger-is-worse effect of model size

- LME models fit to reading times of Natural Stories, Dundee, Ghent, and Provo corpora (Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018)
- Baseline predictors: Word length/position, unigram surprisal (The Pile; Gao et al., 2020), saccade length, previous word fixated
- Predictors of interest: LLM surprisal

Revisiting the bigger-is-worse effect of model size

- LME models fit to reading times of Natural Stories, Dundee, Ghent, and Provo corpora (Cop et al., 2017; Futrell et al., 2021; Kennedy et al., 2003; Luke & Christianson, 2018)
- Baseline predictors: Word length/position, unigram surprisal (The Pile; Gao et al., 2020), saccade length, previous word fixated
- Predictors of interest: LLM surprisal
- Mean squared errors calculated on each quintile defined by unigram log-probability


Natural Stories SPR
Dundee ET
GECO ET
Provo ET
GPT-2
GPT-Neo
OPT
Pythia
MSE (log ms)
Perplexity
Poorer Fit
Better Fit
Larger Models
Smaller Models



Poorer Fit
Better Fit
Dundee ET
[-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)
MSE (log ms)
0.2420
0.2415
0.2410
0.2180
0.2175
0.2170
0.2070
0.2065
0.2060
0.2025
0.2020
0.2015
0.2125
0.2120
0.2115
$2^{5.0}$
$2^{6.0}$
$2^{7.0}$
Perplexity
Rare Words
Frequent Words

Revisiting the bigger-is-worse effect of training data amount

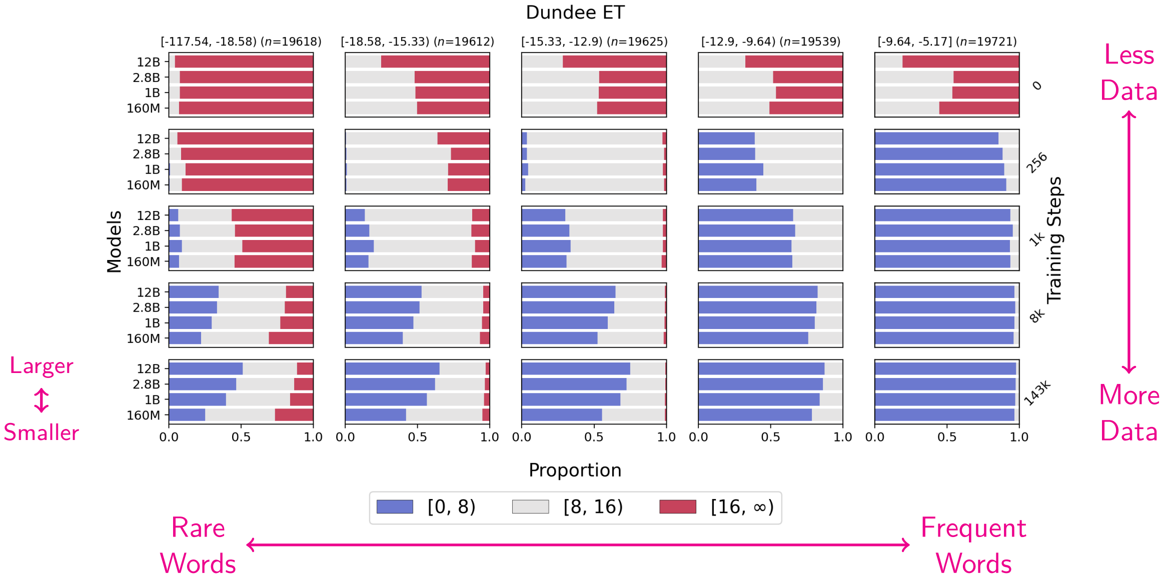
- Similar regression modeling procedures as Experiment 1

Revisiting the bigger-is-worse effect of training data amount

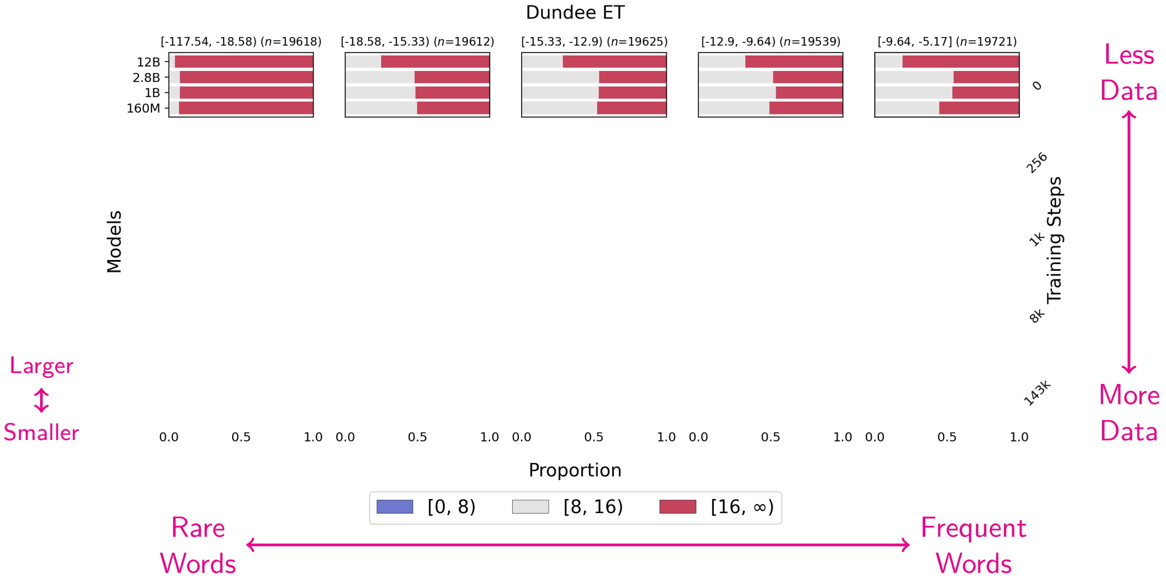
- Similar regression modeling procedures as Experiment 1
- Pythia surprisal after $\{0, 128, 256, 512, 1k, 2k, 4k, 8k, 143k\}$ training steps

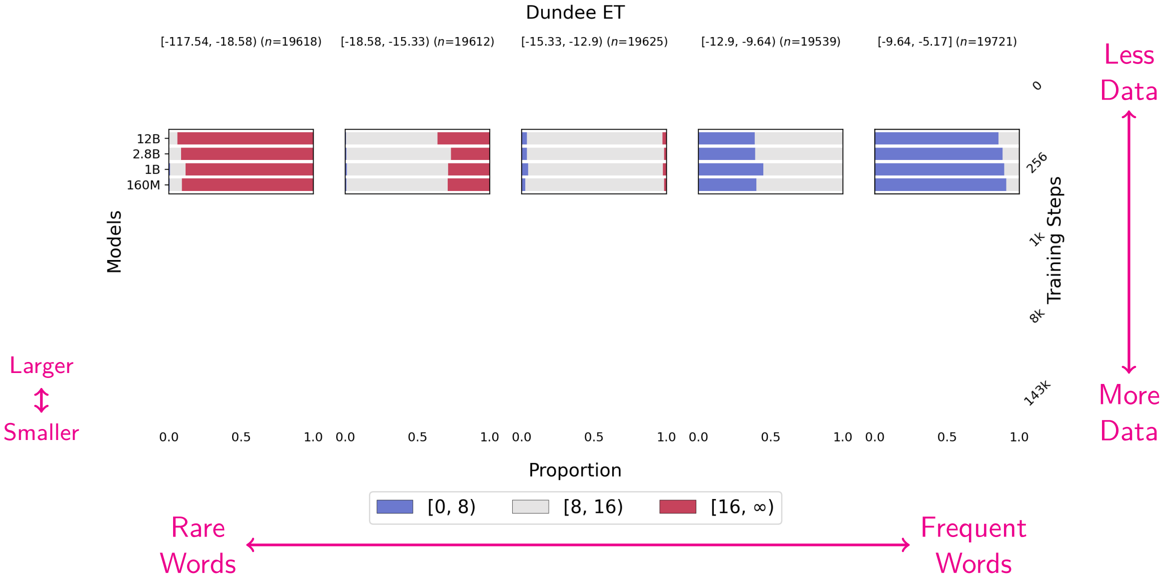
Revisiting the bigger-is-worse effect of training data amount

- Similar regression modeling procedures as Experiment 1
- Pythia surprisal after $\{0, 128, 256, 512, 1k, 2k, 4k, 8k, 143k\}$ training steps
- Surprisal values and MSEs analyzed by quintile defined by unigram log-probability

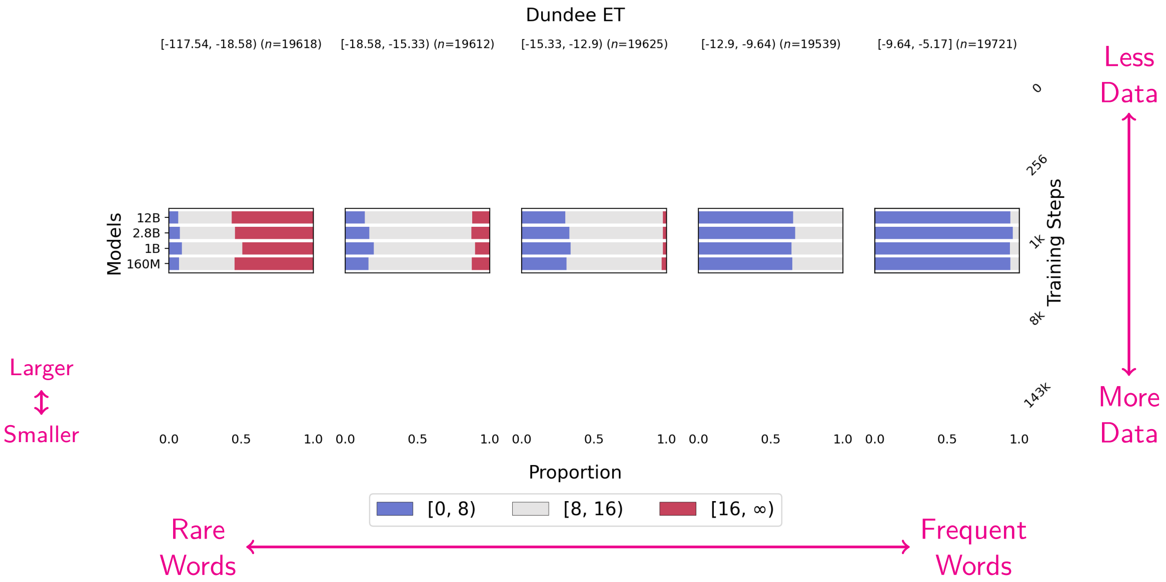


Dundee ET
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
12B
2.8B
1B
160M
Models
0
256
1k
8k
143k
Training Steps
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words

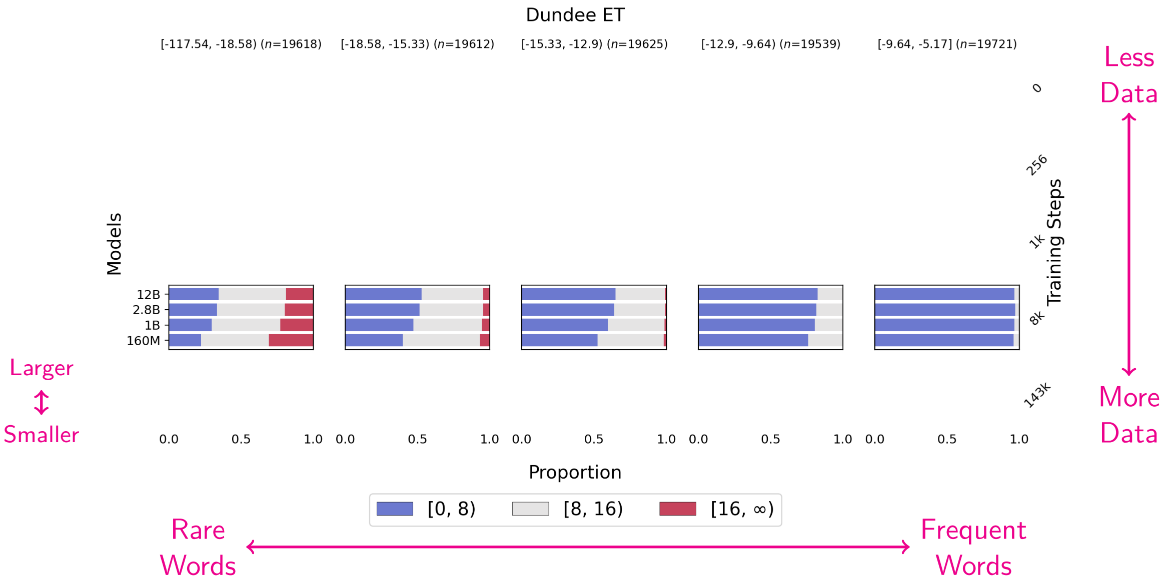

Dundee ET
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
12B
2.8B
1B
160M
Models
Training Steps
0
256
1k
8k
143k
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words



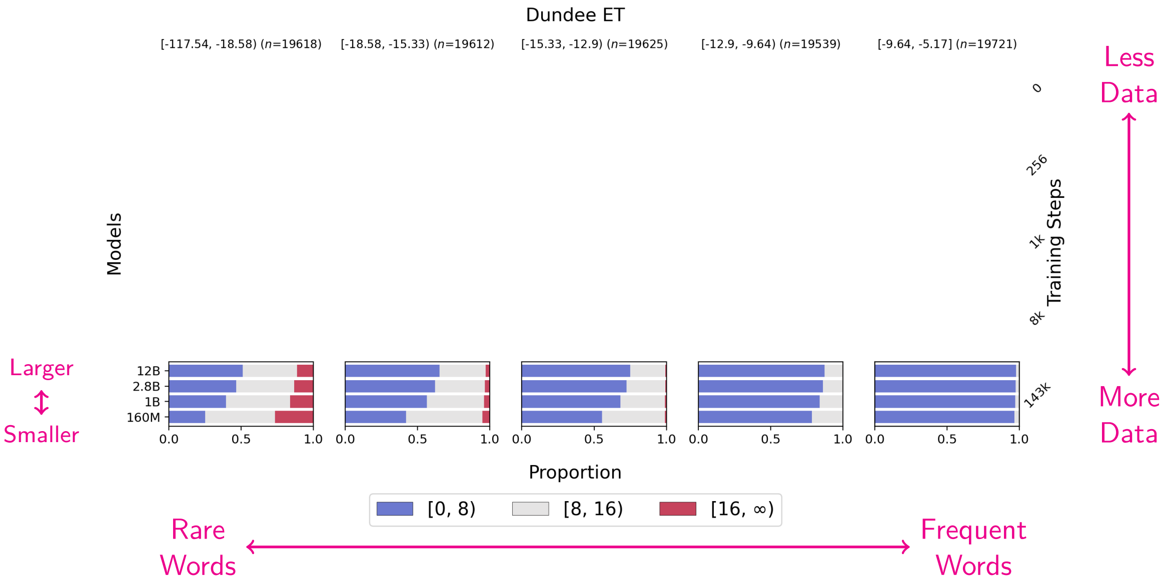
Dundee ET
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
12B
2.8B
1B
160M
Models
0
256
1k
8k
143k
Training Steps
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words



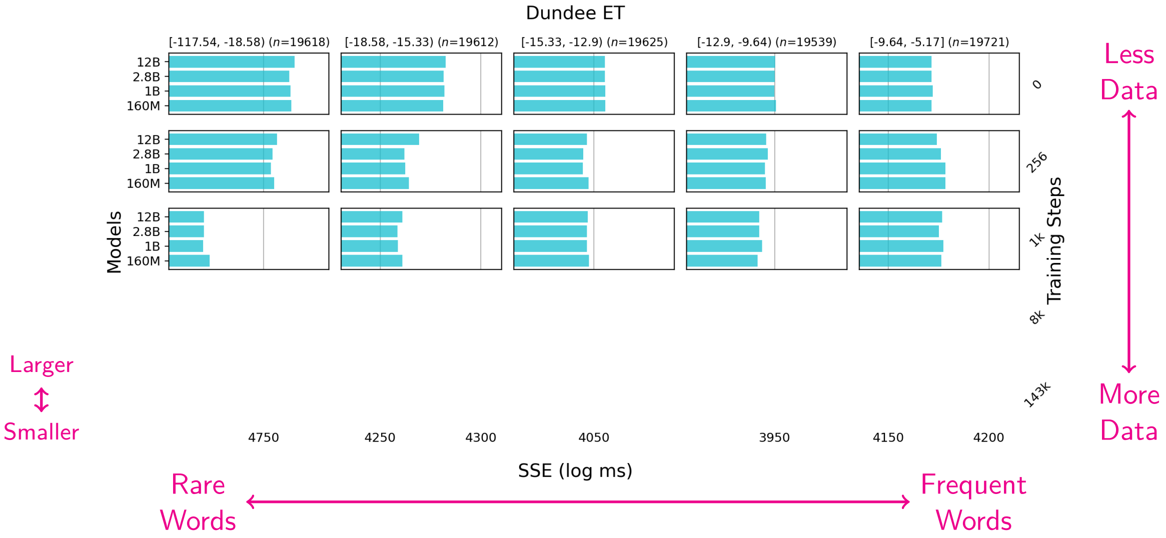
Dundee ET
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
Models
12B
2.8B
1B
160M
Training Steps
0
256
1k
8k
143k
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words



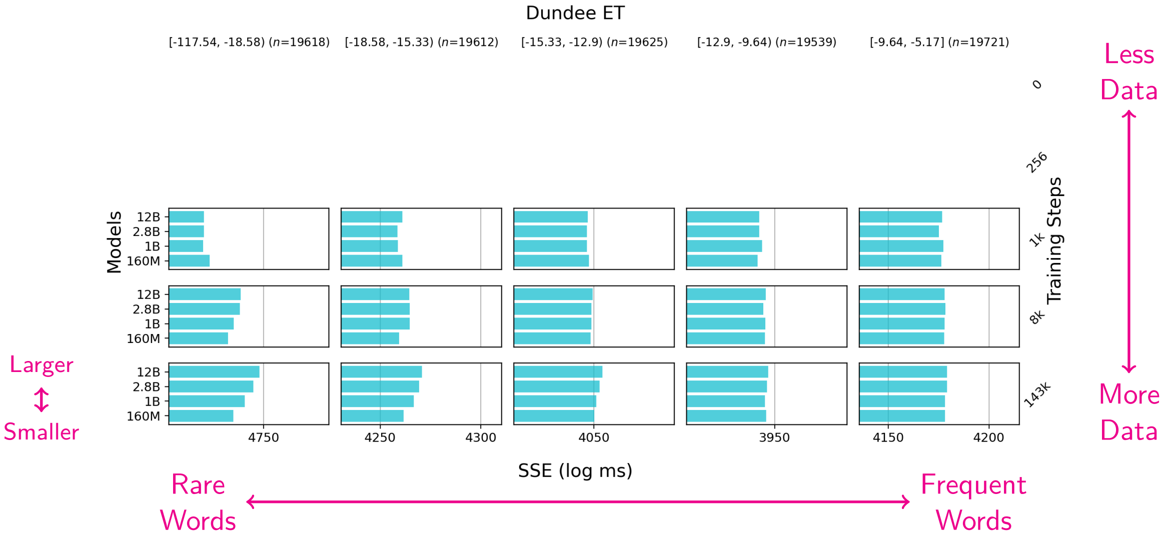
Dundee ET
[-117.54, -18.58) (n=19618)
[-18.58, -15.33) (n=19612)
[-15.33, -12.9) (n=19625)
[-12.9, -9.64) (n=19539)
[-9.64, -5.17] (n=19721)
Models
12B
2.8B
1B
160M
Training Steps
0
256
1k
8k
143k
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words


Dundee ET
[-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)
Models
12B
2.8B
1B
160M
Training Steps
0
256
1k
8k
143k
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Larger
Smaller
Less Data
More Data
Rare Words
Frequent Words

Dundee ET
[-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)
12B
2.8B
1B
160M
Models
0
256
1k
8k
143k
Training Steps
4750
4250
4300
4050
3950
4150
4200
SSE (log ms)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words



Dundee ET
[-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)
12B
2.8B
1B
160M
Models
0
256
1k
8k
143k
Training Steps
4750
4250
4300
4050
3950
4150
4200
SSE (log ms)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words



Dundee ET
[-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)
Models
12B
2.8B
1B
160M
Training Steps
0
256
1k
8k
143k
4750
4250
4300
4050
3950
4150
4200
SSE (log ms)
Less Data
More Data
Larger
Smaller
Rare Words
Frequent Words

What enables larger models to predict rare words?

- One possibility is that larger models have a longer ‘effective’ context window

What enables larger models to predict rare words?

- One possibility is that larger models have a longer ‘effective’ context window
- Method: Limiting the context to the most recent $\{49, 24, 9\}$ tokens (Kuribayashi et al., 2022)

What enables larger models to predict rare words?

- One possibility is that larger models have a longer 'effective' context window
- Method: Limiting the context to the most recent {49, 24, 9} tokens (Kuribayashi et al., 2022)
  I landed in Frankfurt and took a _____

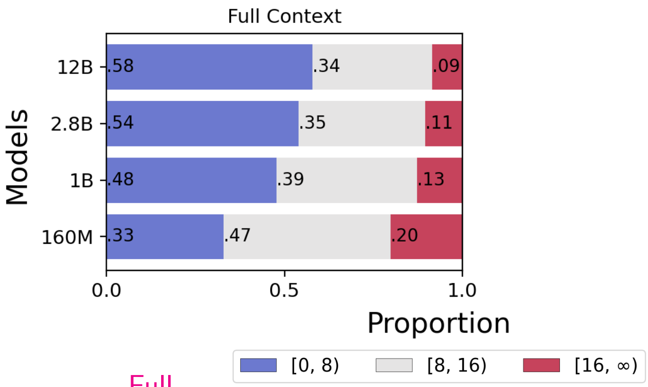
What enables larger models to predict rare words?

- One possibility is that larger models have a longer 'effective' context window
- Method: Limiting the context to the most recent $\{49, 24, 9\}$ tokens (Kuribayashi et al., 2022)
  I landed in Frankfurt and took a _____ $\rightarrow$ *and took a* _____

What enables larger models to predict rare words?

- One possibility is that larger models have a longer ‘effective’ context window
- Method: Limiting the context to the most recent {49, 24, 9} tokens (Kuribayashi et al., 2022)
  I landed in Frankfurt and took a _____ $\rightarrow$ *and took a _____*
- Change in Pythia surprisal values analyzed on the quintile of the rarest words

Dundee ET

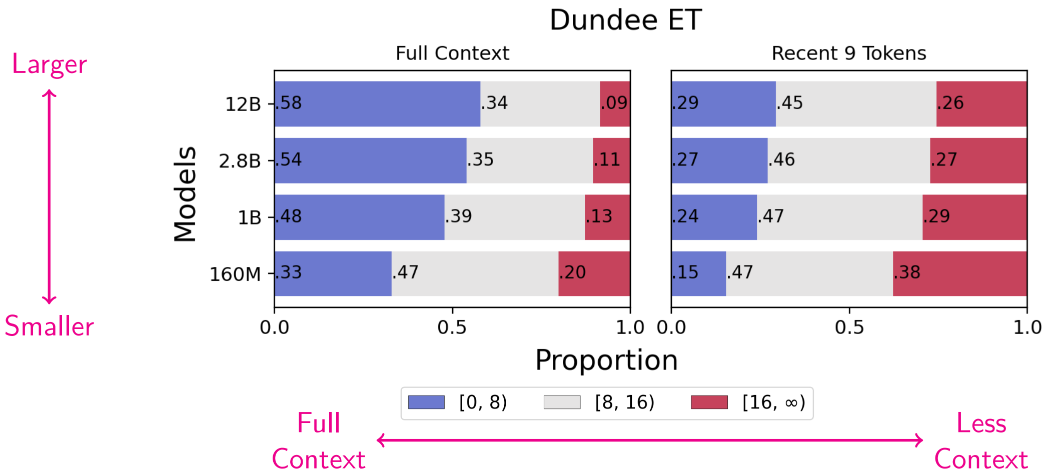


Larger

Smaller

Full Context

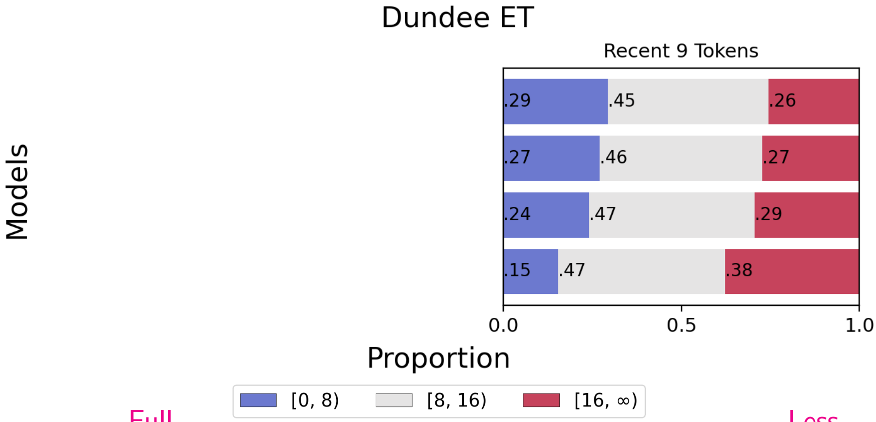
Less Context



Larger
Smaller
Dundee ET
Full Context
Recent 9 Tokens
Models
12B
2.8B
1B
160M
.58
.34
.09
.54
.35
.11
.48
.39
.13
.33
.47
.20
.29
.45
.26
.27
.46
.27
.24
.47
.29
.15
.47
.38
0.0
0.5
1.0
Proportion
[0, 8)
[8, 16)
[16, ∞)
Full Context
Less Context

Larger

Smaller



Full Context

Less Context

Summary: Word frequency as a unified explanation

- Word frequency explains the adverse effects of model size and training data amount

Summary: Word frequency as a unified explanation

- Word frequency explains the adverse effects of model size and training data amount
- Larger model and training data sizes contribute to accurate predictions of rare words

Summary: Word frequency as a unified explanation

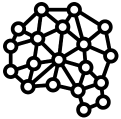
- Word frequency explains the adverse effects of model size and training data amount
- Larger model and training data sizes contribute to accurate predictions of rare words
- The associations that give larger models an advantage are widespread

Conclusion

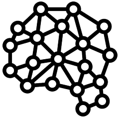
Takeaways

Human subjects

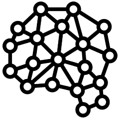
$\sim$



Model 1



Model 2

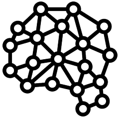


Model 3

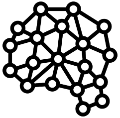
Takeaways

Human subjects

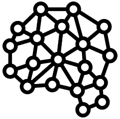
$\sim$



Model 1



Model 2



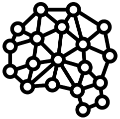
Model 3

1. Which models are closer to human behavior among Models $1..n$?

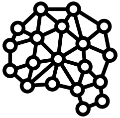
Takeaways

Human subjects

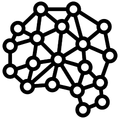
$\sim$



Model 1



Model 2



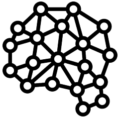
Model 3

1. Which models are closer to human behavior among Models $1..n$?
   Smaller LLMs trained on less data

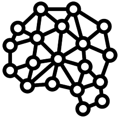
Takeaways

Human subjects

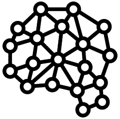
$\sim$



Model 1



Model 2



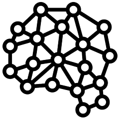
Model 3

1. Which models are closer to human behavior among Models $1..n$?
   Smaller LLMs trained on less data
2. Why is Model i less human-like than Model j?

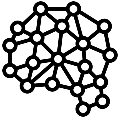
Takeaways

Human subjects

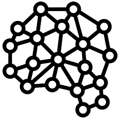
$\sim$



Model 1



Model 2



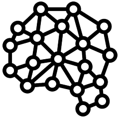
Model 3

1. Which models are closer to human behavior among Models $1..n$?
   Smaller LLMs trained on less data
2. Why is Model i less human-like than Model j?
   Accurate predictions of rare words

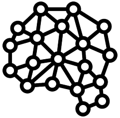
Implications

Human subjects

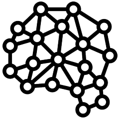
$\sim$



Model 1



Model 2

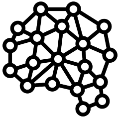


Model 3

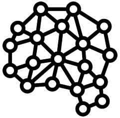
Implications

Human subjects

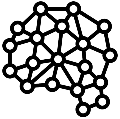
$\sim$



Model 1



Model 2

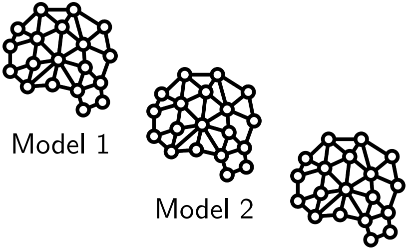


Model 3

1. Surprisal theory could be refined to assume a realistic amount of data

Implications

$\sim$

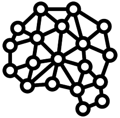


1. Surprisal theory could be refined to assume a realistic amount of data
2. Caution for using LLM surprisal to study other psycholinguistic questions!
   (e.g. Hoover et al., 2023; Shain, 2023)

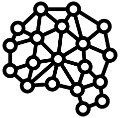
Future directions

Human subjects

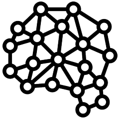
$\sim$



Model 1



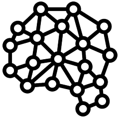
Model 2



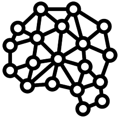
Model 3

Human subjects

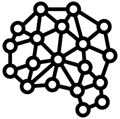
$\sim$



Model 1



Model 2



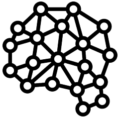
Model 3

1. What drives the predictions of Model k?

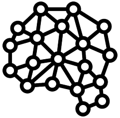
Future directions

Human
subjects

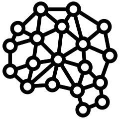
$\sim$



Model 1



Model 2



Model 3

1. What drives the predictions of Model k?
2. Do these results generalize to other constructions or languages?

Thank you for listening!

oh.531@osu.edu byungdoh.github.io
byungdoh/{llm_surprisal,slm_surprisal}

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. https://arxiv.org/abs/2210.12187

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.

Cop, U., Dirix, N., Drieghe, D., & Duyck, W. (2017). Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading. *Behavior Research Methods*, *49*(2), 602–615. https://doi.org/10.3758/s13428-016-0734-0

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.

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*, 63–77. https://doi.org/10.1007/s10579-020-09503-7

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint, arXiv:2101.00027*. https://arxiv.org/abs/2101.00027

References II

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. https://doi.org/10.1073/pnas.2122602119

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://www.aclweb.org/anthology/N01-1021/

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, *7*, 350–391. https://doi.org/10.1162/opmi_a_00086

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

Kuribayashi, T., Oseki, Y., Brassard, A., & Inui, K. (2022). Context limitations make neural language models more human-like. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 10421–10436. https://aclanthology.org/2022.emnlp-main.712

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

References III

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, *50*(2), 826–833. https://doi.org/10.3758/s13428-017-0908-4

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. https://doi.org/10.3389/frai.2022.777963

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. https://doi.org/10.1162/tacl_a_00548

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. https://aclanthology.org/2023.findings-emnlp.128/

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/

Shain, C. (2023). Word frequency and predictability dissociate in naturalistic reading. *PsyArXiv preprint*. https://osf.io/preprints/psyarxiv/9zdfw/

References IV

Shain, C., van Schijndel, M., & Schuler, W. (2018). Deep syntactic annotations for broad-coverage psycholinguistic modeling. *Workshop on Linguistic and Neuro-Cognitive Resources.* http://lrec-conf.org/workshops/lrec2018/W9/pdf/9_W9.pdf

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

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*, e12988. https://doi.org/10.1111/cogs.12988

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. https://cognitivesciencesociety.org/cogsci20/papers/0375

Xia, M., Artetxe, M., Zhou, C., Lin, X. V., Pasunuru, R., Chen, D., Zettlemoyer, L., & Stoyanov, V. (2023). Training trajectories of language models across scales. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 13711–13738. https://aclanthology.org/2023.acl-long.767