THE OHIO STATE UNIVERSITY

Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times?

Byung-Doh Oh, William Schuler

Introduction

- Expectation-based theories of sentence processing
  (Hale, 2001; Levy, 2008)

- Surprisal from LMs evaluated on measures of processing difficulty
  (e.g. Smith and Levy, 2013; Hale et al., 2018)

- Conflicting results regarding LM perplexity and fit to reading times
  (Goodkind and Bicknell 2018; Wilcox et al., 2020; Oh et al., 2022)

This Work

1. Evaluation of LLM surprisal on ability to predict human reading times

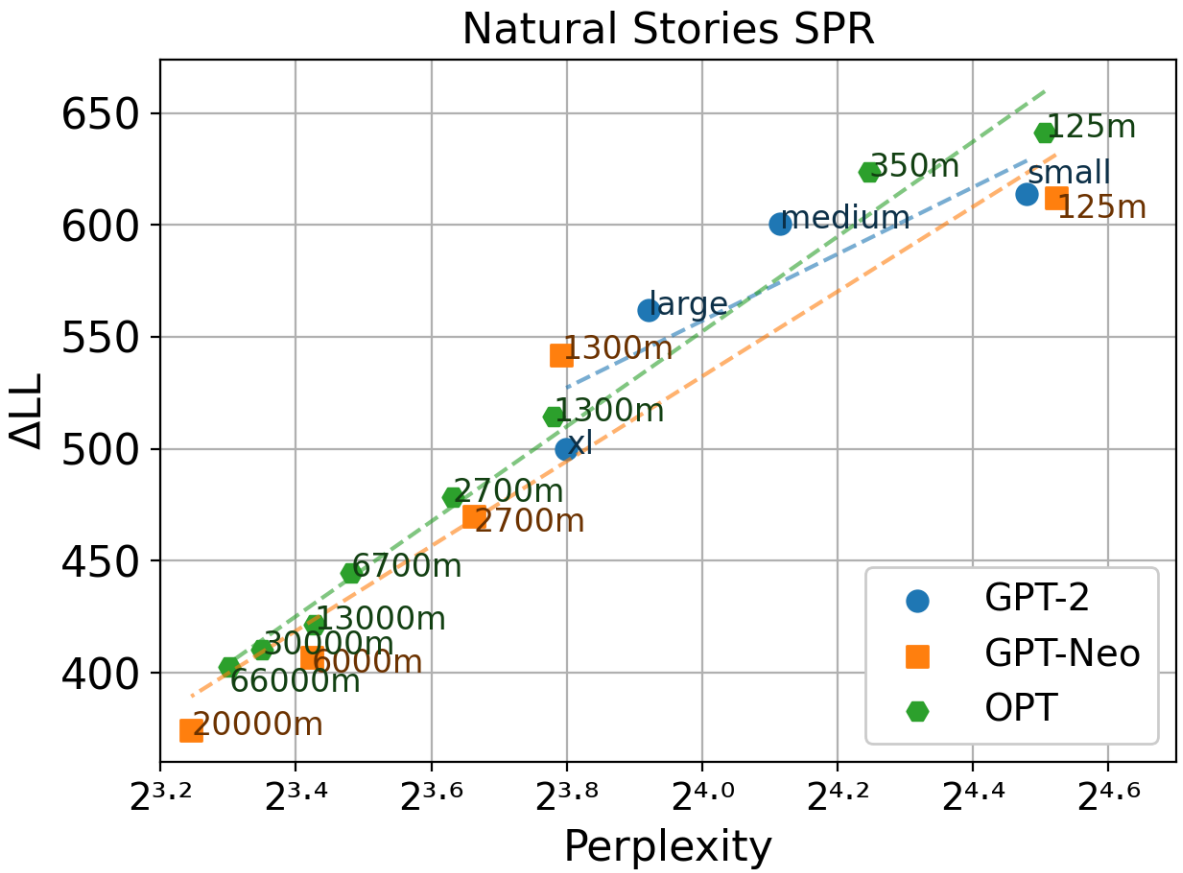
2. Identifying data points that drive the trend in fit to human reading times

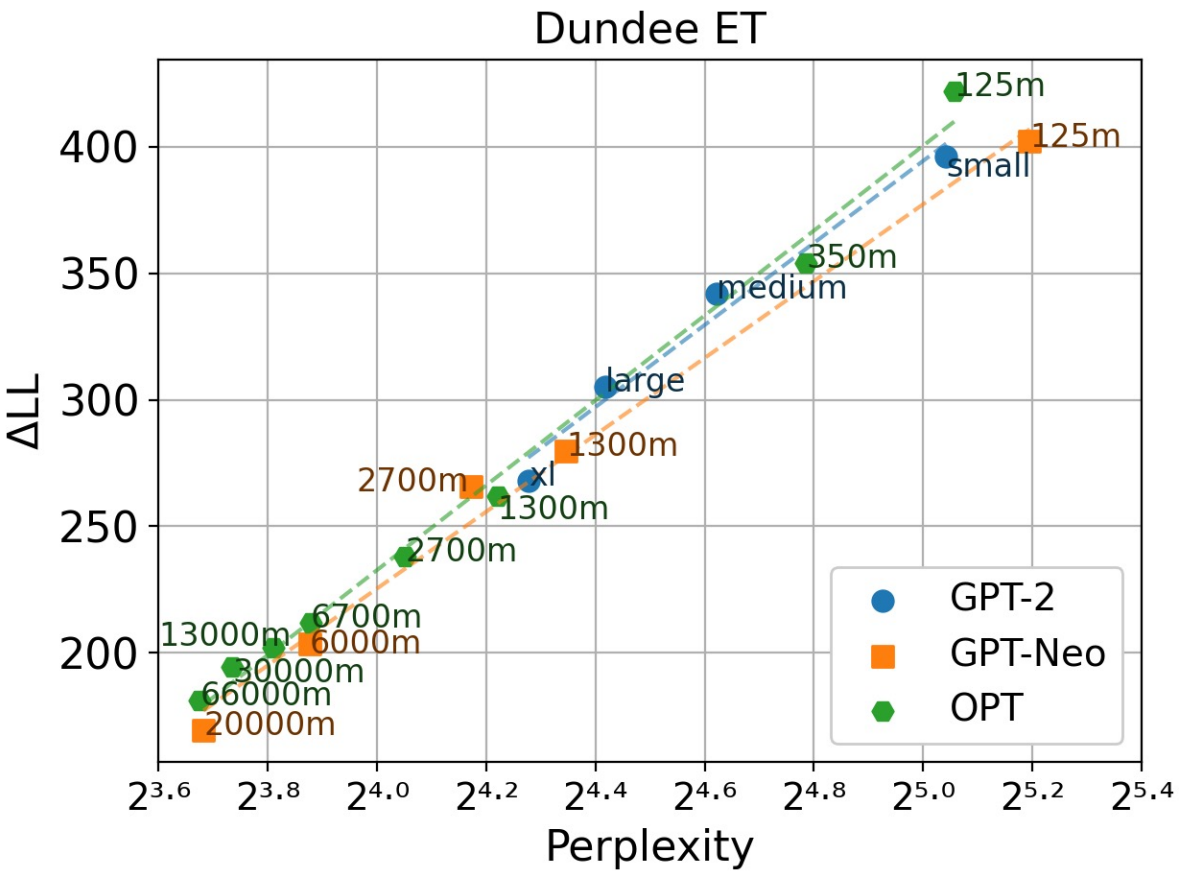
Replication Study: Evaluation on Reading Times

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

Model	#L	#H	d_{model}	Parameters
GPT-2 Small	12	12	768	∼124M
GPT-2 Medium	24	16	1024	∼355M
GPT-2 Large	36	20	1280	∼774M
GPT-2 XL	48	25	1600	∼1558M
GPT-Neo 125M	12	12	768	∼125M
GPT-Neo 1300M	24	16	2048	∼1300M
GPT-Neo 2700M	32	20	2560	∼2700M
GPT-J 6B	28	16	4096	∼6000M
GPT-NeoX 20B	44	64	6144	∼20000M
OPT 125M	12	12	768	∼125M
OPT 350M	24	16	1024	∼350M
OPT 1.3B	24	32	2048	∼1300M
OPT 2.7B	32	32	2560	∼2700M
OPT 6.7B	32	32	4096	∼6700M
OPT 13B	40	40	5120	∼13000M
OPT 30B	48	56	7168	∼30000M
OPT 66B	64	72	9216	∼66000M

Replication Study: Evaluation on Reading Times



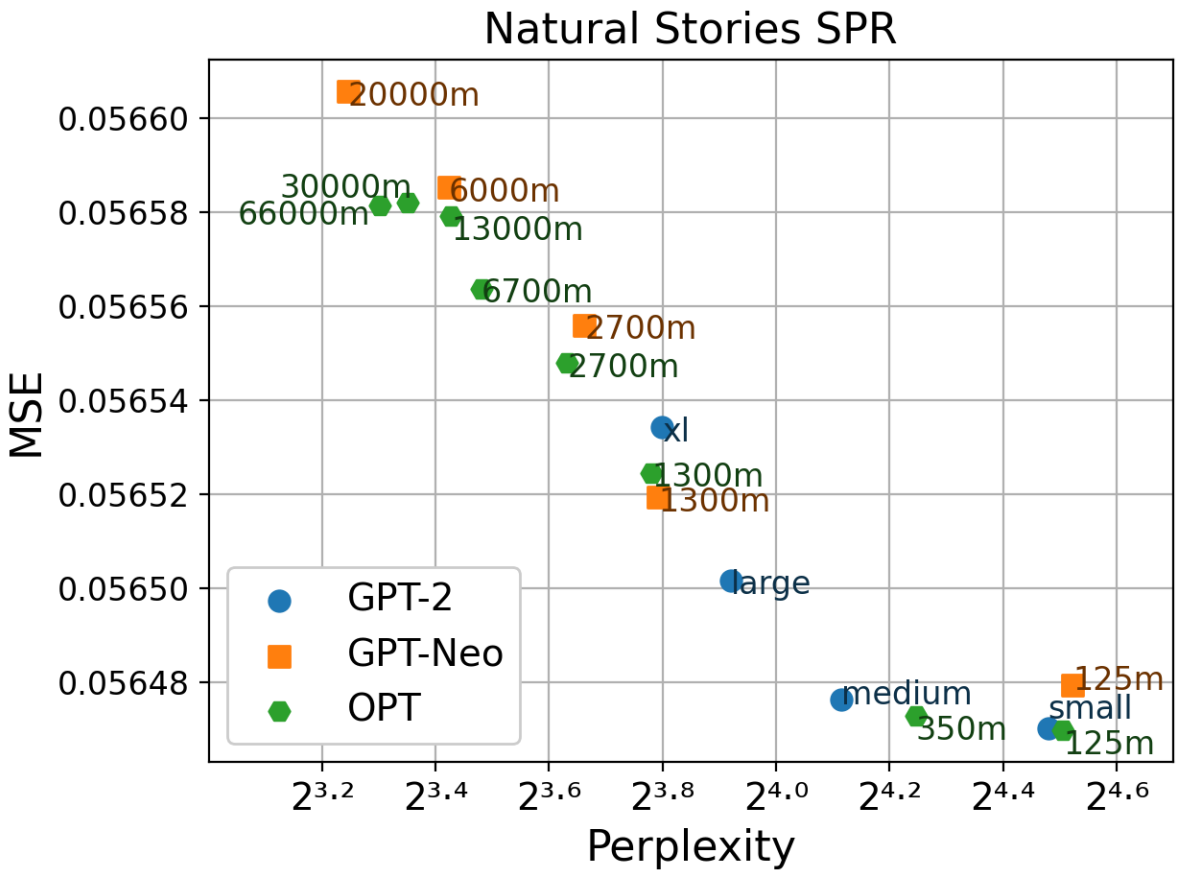


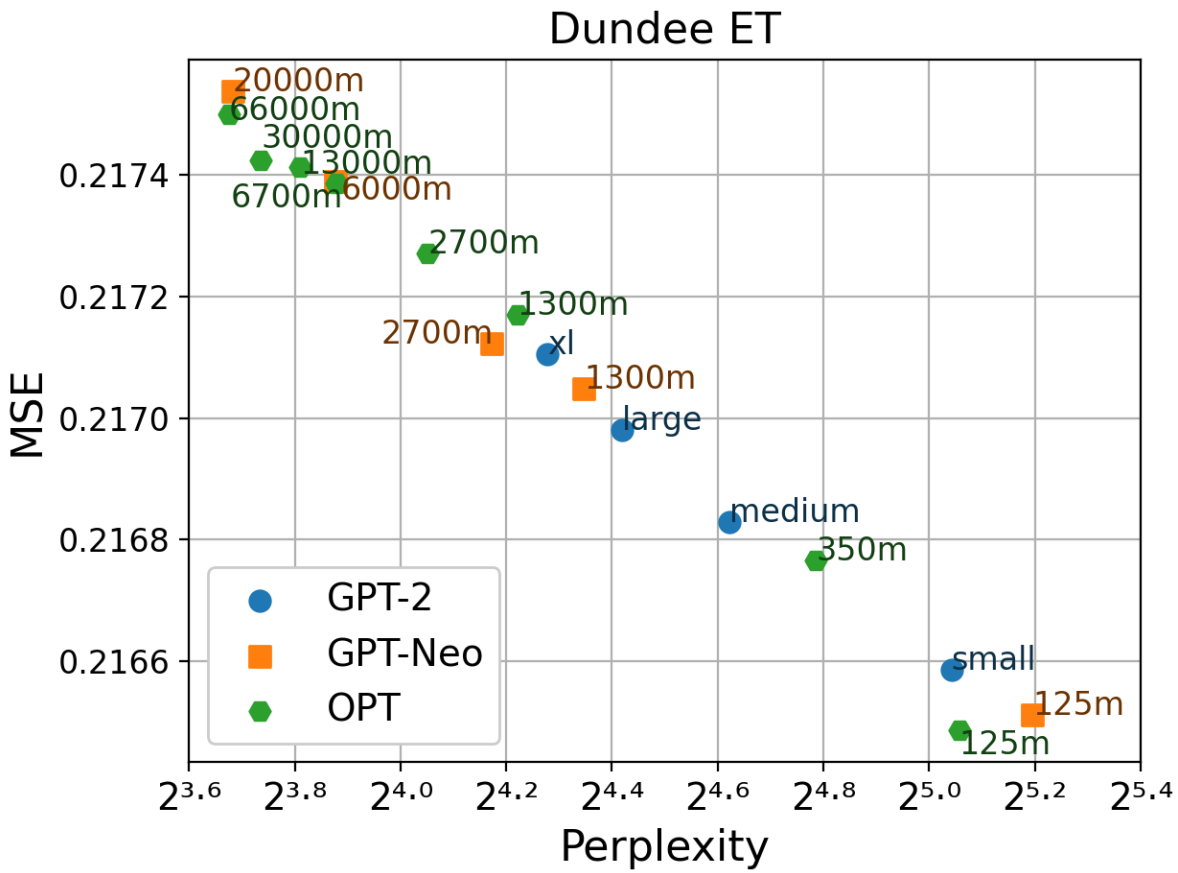
Post-hoc Analysis:
Linguistic Phenomena Underlying the Trend

- Data points associated with word-level and syntactic properties
  (Shain et al., 2018)

- Subsets with the largest differences in SE between models identified

Post-hoc Analysis:
Linguistic Phenomena Underlying the Trend



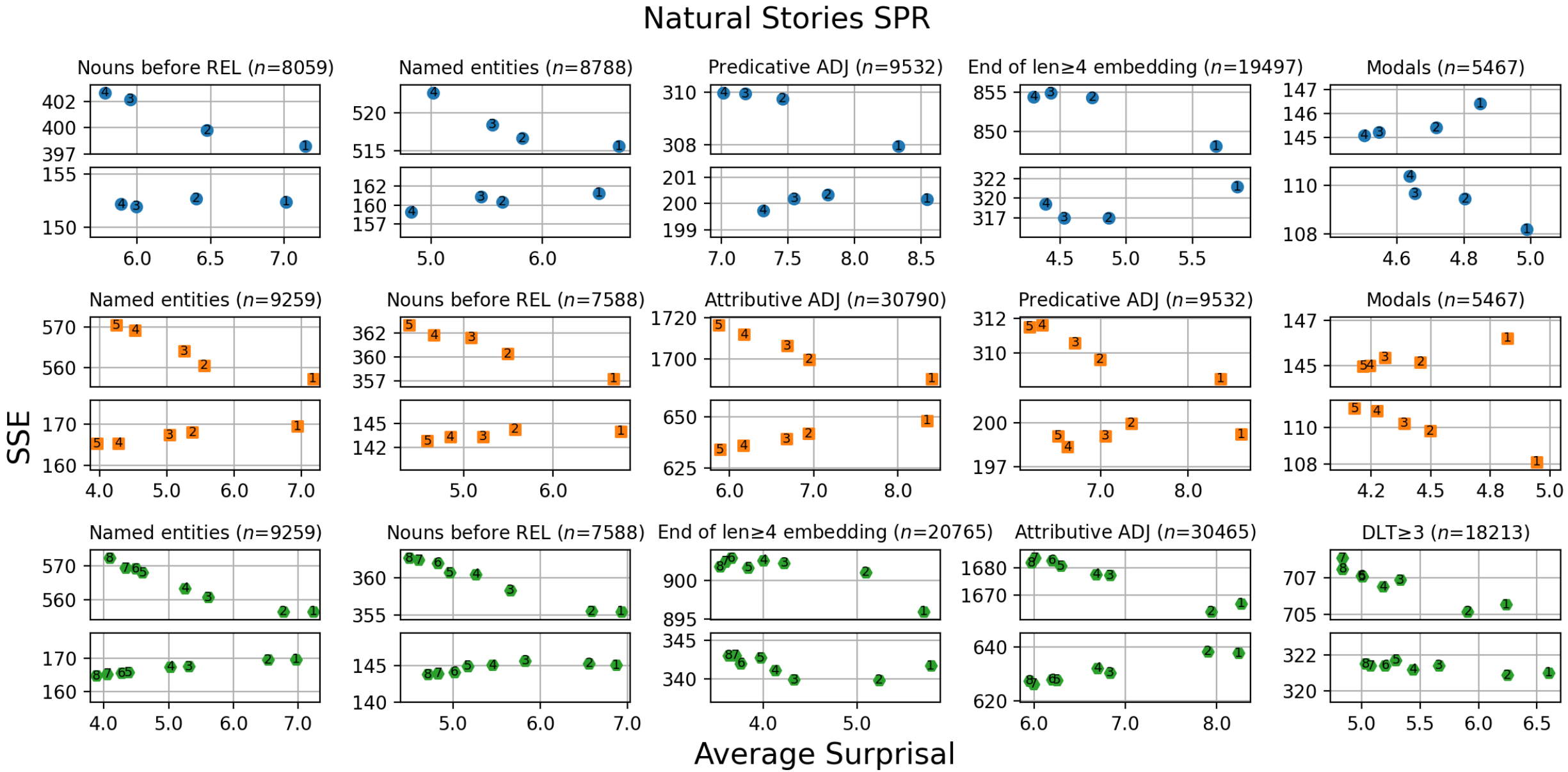


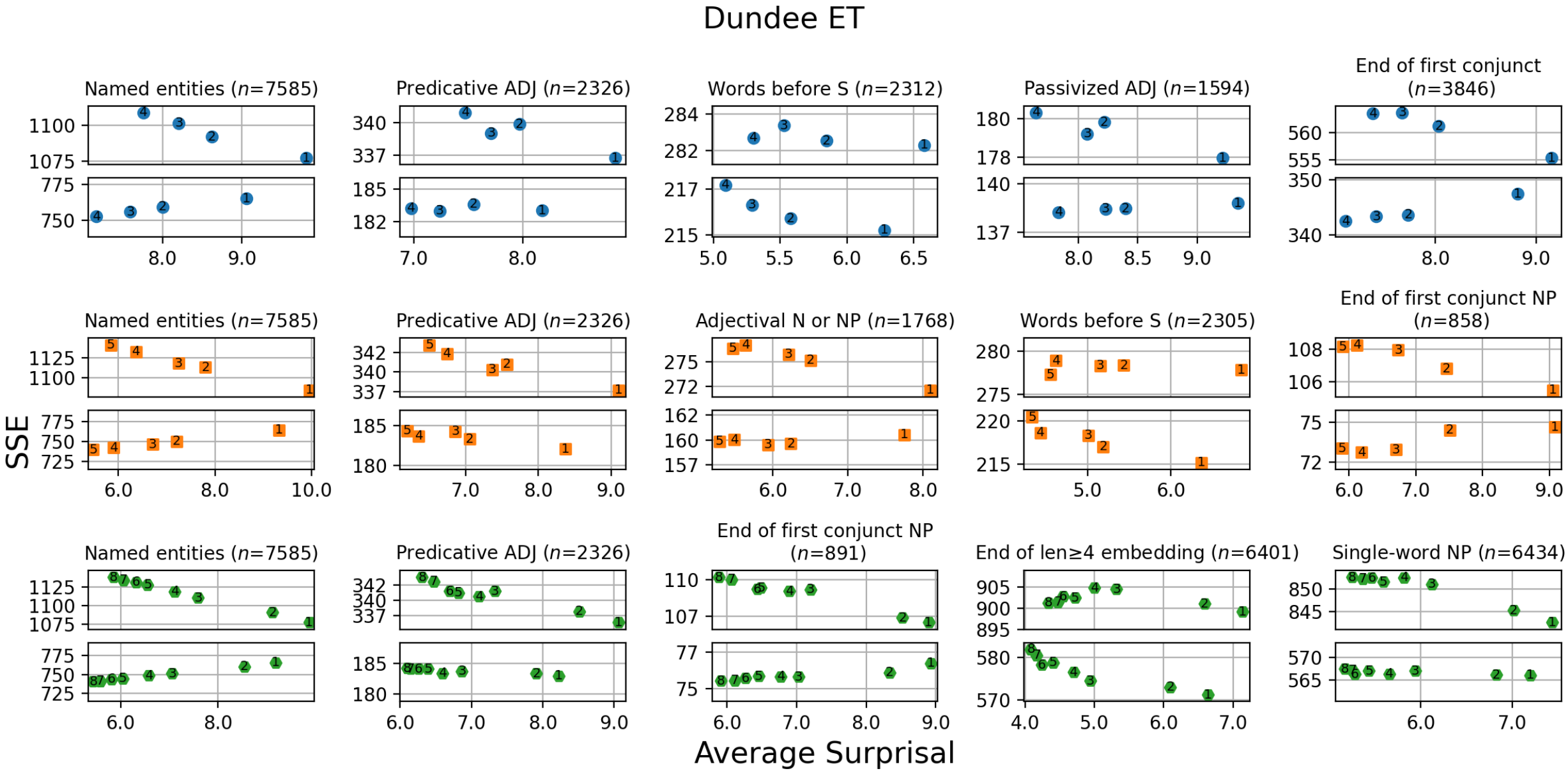
Post-hoc Analysis:
Linguistic Phenomena Underlying the Trend

- Data points associated with word-level and syntactic properties
  (Shain et al., 2018)

- Subsets with the largest differences in SE between models identified

- Data points further categorized as underpredictions or overpredictions


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)
Named entities (n=9259)
Nouns before REL (n=7588)
Attributive ADJ (n=30790)
Predicative ADJ (n=9532)
Modals (n=5467)
Named entities (n=9259)
Nouns before REL (n=7588)
End of len≥4 embedding (n=20765)
Attributive ADJ (n=30465)
DLT≥3 (n=18213)
SSE
Average Surprisal


Dundee ET
Named entities (n=7585)
Predicative ADJ (n=2326)
Words before S (n=2312)
Passivized ADJ (n=1594)
End of first conjunct (n=3846)
Named entities (n=7585)
Predicative ADJ (n=2326)
Adjectival N or NP (n=1768)
Words before S (n=2305)
End of first conjunct NP (n=858)
SSE
Named entities (n=7585)
Predicative ADJ (n=2326)
End of first conjunct NP (n=891)
End of len≥4 embedding (n=6401)
Single-word NP (n=6434)
Average Surprisal

Conclusion

- ‘Bigger-is-worse’ effect of LM surprisal robustly replicated
  (Oh et al., 2022)

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

- Smaller pre-trained LMs should be used to study sentence processing

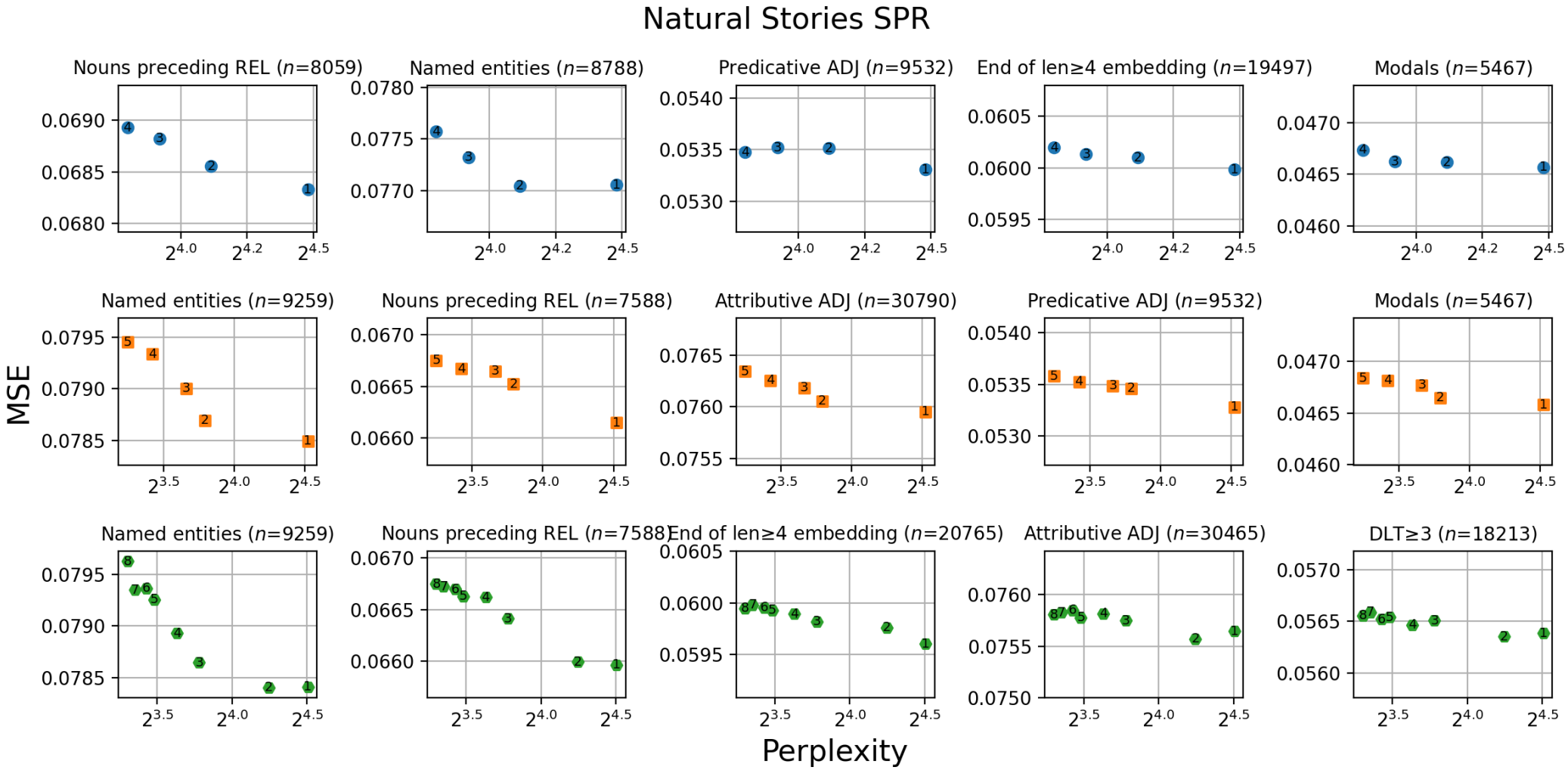
THE OHIO STATE UNIVERSITY

Thank you!

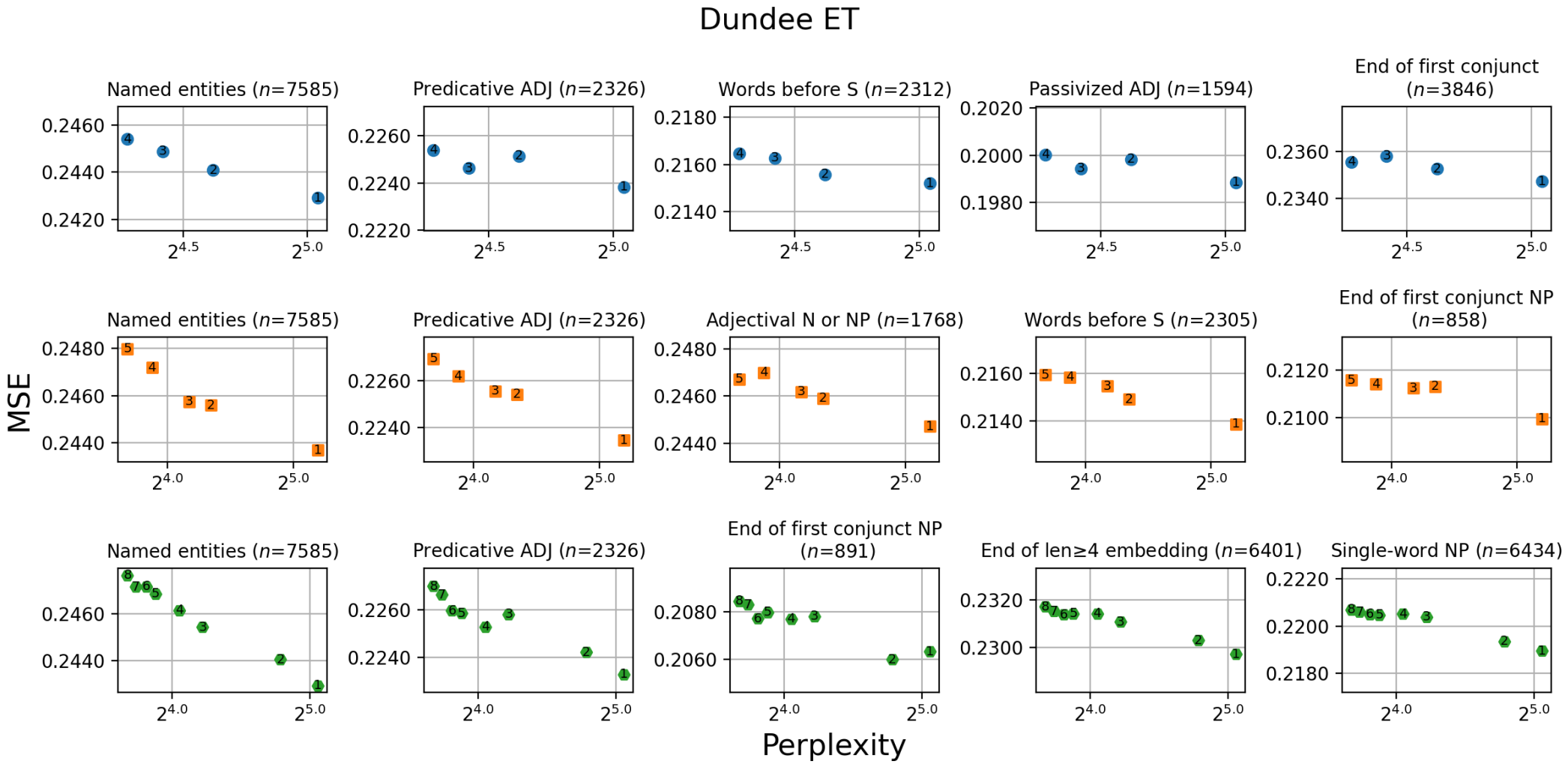
This work was supported by NSF Grant #1816891.

Code for this work is publicly available at https://github.com/byungdoh/llm_surprisal.

Supplementary: Top-5 Subsets (Natural Stories)



Supplementary: Top-5 Subsets (Dundee)



Supplementary: Regression Modeling

- Filtering criteria
  - SPR: initial/final words, <100 ms, >3000 ms, <4 correct answers
  - ET: initial/final words, unfixated words, after >4 word saccades

- By-subject random slopes for all main effects
  - (1 | subject:sentence) for Natural Stories
  - (1 | sentence) for Dundee