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

Byung-Doh Oh
Department of Linguistics
The Ohio State University, USA
oh.531@osu.edu

William Schuler
Department of Linguistics
The Ohio State University, USA
schuler.77@osu.edu

Abstract

This work presents a linguistic analysis into why larger Transformer-based pre-trained language models with more parameters and lower perplexity nonetheless yield surprisal estimates that are less predictive of human reading times. First, regression analyses show a strictly monotonic, positive log-linear relationship between perplexity and fit to reading times for the more recently released five GPT-Neo variants and eight OPT variants on two separate datasets, replicating earlier results limited to just GPT-2 (Oh et al., 2022). Subsequently, analysis of residual errors reveals a systematic deviation of the larger variants, such as underpredicting reading times of named entities and making compensatory overpredictions for reading times of function words such as modals and conjunctions. These results suggest that the propensity of larger Transformer-based models to 'memorize' sequences during training makes their surprisal estimates diverge from humanlike expectations, which warrants caution in using pre-trained language models to study human language processing.

1 Introduction

Expectation-based theories of sentence processing (Hale, 2001; Levy, 2008) postulate that processing difficulty is largely driven by how predictable upcoming linguistic material is given its context. In cognitive modeling, predictability operationalized by information-theoretic surprisal (Shannon, 1948) has been shown to be a strong predictor of behavioral and neural measures of processing difficulty (Demberg and Keller, 2008; Smith and Levy, 2013; Hale et al., 2018; Shain et al., 2020), providing empirical support for this position. As language models (LMs) directly define a conditional probability distribution of a word given its context required for surprisal calculation, they have frequently been evaluated as surprisal-based cognitive models of sentence processing.

Recently, it was observed that surprisal from larger variants of the pre-trained GPT-2 LM (Radford et al., 2019) that have more parameters and achieve lower perplexity is less predictive of self-paced reading times and eye-gaze durations collected during naturalistic reading (Oh et al., 2022). As the different variants of the pre-trained GPT-2 model share the primary architecture and training data, this offers an especially strong counterexample to previous work that showed a negative relationship between LM perplexity and predictive power of surprisal estimates (Goodkind and Bicknell, 2018; Hao et al., 2020; Wilcox et al., 2020). More broadly, this observation also contradicts the recent 'larger is better' trend of the NLP community, leaving open the question of why larger LMs perform worse. However, the Oh et al. (2022) results were part of a follow-up analysis in support of a separate claim about parser surprisal that only examined four model variants, so the results were not tested for statistical significance or extensively explored.

The current work fills that gap by conducting a detailed linguistic analysis of the positive relationship between LM perplexity and predictive power of surprisal estimates. First, the robustness of the trend observed in Oh et al. (2022) is examined by reproducing their results and additionally evaluating surprisal estimates from different families of Transformer-based LMs (GPT-Neo, OPT; Black et al., 2021, 2022; Wang and Komatsuzaki, 2021; Zhang et al., 2022) on their ability to predict human reading times. Results from regression analyses show a strictly monotonic, positive log-linear relationship between LM perplexity and fit to reading times for the five GPT-Neo variants and eight OPT variants on two separate datasets, which provides firm empirical support for this

Transactions of the Association for Computational Linguistics, vol. 11, pp. 336–350, 2023. https://doi.org/10.1162/tacl_a_00548
Action Editor: Marco Baroni. Submission batch: 10/2022; Revision batch: 12/2022; Published 3/2023.

trend. Subsequently, to provide an explanation for this positive relationship, residual errors from the regression models are analyzed with a focus on identifying linguistic phenomena that surprisal from larger variants accounts for less accurately compared to surprisal from their smaller counterparts. The results show that regression models with surprisal predictors from GPT-2, GPT-Neo, and OPT models generally underpredict reading times at nouns and adjectives, and that the degree of underprediction increases along with model size. This indicates that the poorer fit to human reading times achieved by surprisal estimates from larger Transformer-based LMs is primarily driven by their characteristic of assigning lower surprisal values to open-class words, which may be accurately predicted by extensive domain knowledge gleaned from large sets of training examples that are not available to humans. This suggests that as Transformer-based LMs get larger, they may be problematic for cognitive modeling because they are trained with non-human learning objectives and different inductive biases on vast quantities of Internet text.[1]

2 Related Work

In previous studies, surprisal estimates from several well-established types of LMs, including n-gram models, Simple Recurrent Networks (Elman, 1991), Gated Recurrent Unit networks (GRU; Cho et al., 2014), and Long Short-Term Memory networks (LSTM; Hochreiter and Schmidhuber, 1997), have been compared against behavioral measures of processing difficulty (e.g., Smith and Levy, 2013; Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019). Recently, as Transformer-based (Vaswani et al., 2017) models have dominated many NLP tasks, both large pre-trained and smaller 'trained-from-scratch' Transformer-based LMs have also been evaluated as models of processing difficulty (Wilcox et al., 2020; Hao et al., 2020; Merkx and Frank, 2021; Schrimpf et al., 2021).

A consistent finding that emerged out of these studies is that better LMs are also more predictive models of comprehension difficulty, or in other words, there is a negative correlation between LM perplexity and fit to human reading times. Goodkind and Bicknell (2018) compared surprisal estimates from a set of n-gram and LSTM LMs and observed a negative linear relationship between perplexity and regression model fit. Wilcox et al. (2020) evaluated n-gram, LSTM, Transformer, and RNNG (Dyer et al., 2016) models and replicated the negative relationship, although they note a more exponential relationship at certain intervals. Merkx and Frank (2021) provided further support for this trend using GRU and Transformer models with different numbers of layers.[2]

However, Oh et al. (2022) observed a directly contradictory relationship to this using surprisal estimates from pre-trained GPT-2 models (Radford et al., 2019). Using self-paced reading times from the Natural Stories Corpus (Futrell et al., 2021) and go-past durations from the Dundee corpus (Kennedy et al., 2003), the authors calculated the increase in log-likelihood (ΔLL) to a baseline linear-mixed effects (LME) model due to including a surprisal predictor. Their results showed that surprisal from the largest *XL* variant made the smallest contribution to regression model fit, followed by the smaller *Large*, *Medium*, and *Small* variants in that order, revealing a robust positive correlation between LM perplexity and predictive power of surprisal estimates. The same trend was replicated when unigram surprisal was included in the baseline, as well as when spillover effects were controlled for through the use of continuous-time deconvolutional regression (CDR; Shain and Schuler, 2021).

Moreover, recent work has shown that surprisal from neural LMs generally tends to underpredict human reading times of both targeted constructions and naturalistic text. For instance, van Schijndel and Linzen (2021) and Arehalli et al. (2022) observed that surprisal from neural LMs severely underpredicts the magnitude of garden-path effects demonstrated by human subjects. Additionally, Hahn et al. (2022) showed that surprisal from the pre-trained GPT-2 model fails to accurately predict the increase in reading times at the main verb of deeply embedded sentences. Kuribayashi et al. (2022) also demonstrated that neural LMs yield surprisal estimates that underpredict naturalistic reading times of English and Japanese text compared to those from neural LMs

[1] All code used in this work is available at: https://github.com/byungdoh/llm_surprisal.

[2] Although counterexamples to this trend have been noted, they were based on comparisons of LMs and incremental parsers that were trained on different data (Oh et al., 2021) or evaluation on Japanese, which has a different syntactic head-directionality than English (Kuribayashi et al., 2021).

that have a recency bias implemented as limited access to the previous context.

3 Main Experiment: Predictive Power of Language Model Surprisal Estimates

In order to examine whether the positive correlation observed by Oh et al. (2022) and others generalizes to larger Transformer-based models, surprisal predictors from different variants of the GPT-2, GPT-Neo, and OPT LMs were evaluated on self-paced reading times from the Natural Stories Corpus (Futrell et al., 2021) and go-past eye-gaze durations from the Dundee Corpus (Kennedy et al., 2003).

3.1 Response Data

The Natural Stories Corpus contains data from 181 subjects who read 10 naturalistic English stories that consist of a total of 10,245 tokens. The reading times were filtered to remove observations for sentence-initial and sentence-final words, observations from subjects who answered three or fewer comprehension questions correctly, and observations shorter than 100 ms or longer than 3000 ms, which resulted in a total of 770,102 observations. The Dundee Corpus contains data from 10 subjects who read 67 newspaper editorials that consist a total of 51,501 tokens. The durations were filtered to remove observations for unfixated words, words following saccades longer than four words, and words at sentence-, screen-, document-, and line-starts and ends. This resulted in a total of 195,507 observations.

Both datasets were subsequently partitioned into an exploratory set and a held-out set of roughly equivalent sizes.[3] This partitioning allows regression model selection (e.g., making decisions about random effects structure) and exploratory analyses to be conducted on the exploratory set and a single statistical significance test to be conducted on the held-out set, thereby obviating the need for multiple trials correction. This resulted in an exploratory set of 384,905 observations and a held-out set of 385,197 observations for the Natural Stories Corpus and an exploratory set of 98,115 observations and a held-out set of 97,392 observations for the Dundee Corpus. All observations were log-transformed prior to model fitting.

[3]This partitioning was conducted based on the sum of subject ID and sentence ID, resulting in each subject-by-sentence combination remaining intact in one partition.

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

Table 1: Model capacities of LM families whose surprisal estimates were examined in this work. #L, #H, and d_{model} refers to number of layers, number of attention heads per layer, and embedding size, respectively.

3.2 Predictors

Surprisal estimates calculated from four different variants of GPT-2 models (Radford et al., 2019) were used in Oh et al. (2022). In addition to GPT-2 surprisal, this experiment also evaluates surprisal estimates from five variants of GPT-Neo models (Black et al., 2021, 2022; Wang and Komatsuzaki, 2021)[4] and eight variants of OPT models (Zhang et al., 2022).[5] All of these LMs are decoder-only autoregressive Transformer-based models whose variants mainly differ in their capacity. The model capacities of the three LM families are summarized in Table 1.

Each story of the Natural Stories Corpus and each article of the Dundee Corpus was tokenized according to the three models' respective byte-pair encoding (BPE; Sennrich et al., 2016) tokenizer and was provided to each model variant to calculate surprisal estimates. In cases where

[4]Technically, the two largest variants are GPT-J and GPT-NeoX models, respectively, both of which have minor architectural differences from the GPT-Neo models. However, given that they share the same training data, they were considered to belong to the same family as the GPT-Neo models.

[5]The largest variant of the OPT model, which has about 175 billion parameters, was not used in this work due to constraints in computational resources.

each story or article did not fit into a single context window for the LMs, the second half of the previous context window served as the first half of a new context window to calculate surprisal estimates for the remaining tokens. In practice, most stories and articles fit completely within two context windows for the GPT-2 models that have a context size of 1,024 tokens, and within one context window for the GPT-Neo and OPT models that have a context size of 2,048 tokens. Additionally, when a single word w_t was tokenized into multiple subword tokens, negative log probabilities of subword tokens corresponding to w_t were added together to calculate $\mathsf{S}(w_t) = -\log \mathsf{P}(w_t \mid w_{1..t-1})$.

3.3 Regression Modeling

Subsequently, following the methods of Oh et al. (2022), a 'baseline' LME model that contains baseline predictors capturing low-level cognitive processing and seventeen 'full' LME models that contain the baseline predictors and each LM surprisal predictor were fit to the exploratory set of self-paced reading times and go-past durations using `lme4` (Bates et al., 2015). The baseline predictors include word length measured in characters and index of word position within each sentence (both self-paced reading and eye-tracking), as well as saccade length and whether or not the previous word was fixated (eye-tracking only).

All predictors were centered and scaled prior to model fitting, and the LME models included by-subject random slopes for all fixed effects as well as random intercepts for each subject and each word type. Additionally, for self-paced reading times collected from 181 subjects, a random intercept for each subject-sentence interaction was included. For eye-gaze durations collected from a much smaller number of 10 subjects, a random intercept for each sentence was included.

After the regression models were fit, the ΔLL values were first calculated for each regression model by subtracting the log-likelihood of the baseline model from that of a full regression model. Moreover, to examine the trend between LM perplexity and predictive power of surprisal estimates, the perplexity of each LM variant was calcuated on the two corpora.

3.4 Results

The results in Figure 1 show that surprisal from the smallest variant (i.e., GPT-2 Small, GPT-Neo 125M, and OPT 125M) made the biggest contribution to regression model fit on both self-paced reading times and eye-gaze durations for the three LM families. More notably, surprisal estimates from larger LM variants within each family yielded strictly poorer fits to reading times, robustly replicating the trend observed by Oh et al. (2022). Interestingly, the three LM families also seem to demonstrate a strong log-linear relationship between perplexity and ΔLL, as can be seen by the least-squares regression lines. All regression lines had a slope significantly greater than 0 at $p < 0.05$ level according to a one-tailed t-test, with the exception of the regression line for GPT-2 on Natural Stories ($p = 0.07$). This trend is highly significant overall by a binomial test (five results with $p < 0.05$ out of six trials), and directly contradicts the findings of recent studies that report a negative correlation between LM perplexity and predictive power of surprisal estimates.

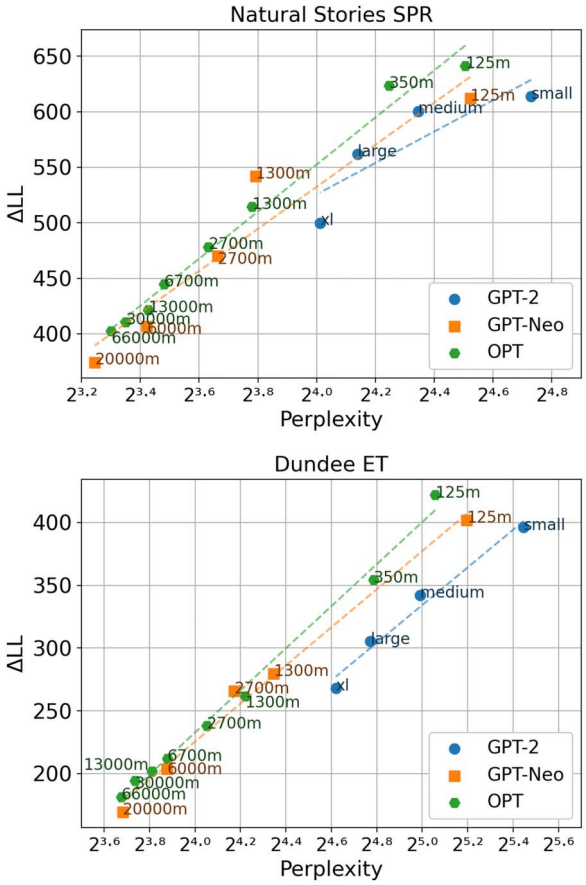

Figure 1: Perplexity measures from each LM variant, and improvements in regression model log-likelihood from including each surprisal estimate on the exploratory set of Natural Stories (top) and Dundee data (bottom). Dotted lines indicate the least-squares regression line for each LM family.

Additionally, comparison of the GPT-2 models and OPT models of similar model capacities (i.e., Small-125M, Medium-350M) shows that the OPT models generally both achieve lower perplexity and yield surprisal estimates that are more predictive of human reading times. Given the high similarity in model architecture between the two LMs, this trend seems to be due to the difference in the training data that were used. The most notable difference between the two training datasets is in their size, with the training set for GPT-2 estimated to be about 15B tokens and that for OPT estimated to be about 180B tokens (Thompson, 2022). However, the GPT-Neo models trained on about 247B tokens show no improvement over the OPT models, yielding a mixed picture. These results suggest that beyond a certain level, the quantity of training data may play a secondary role to the number of model parameters in capturing humanlike expectations.

4 Post-hoc Analysis: Linguistic Phenomena Underlying the Trend

In order to provide an explanation for the trend observed in Section 3, the residual errors from the regression models were analyzed to identify data points that surprisal from larger LM variants accounted for less accurately compared to surprisal from their smaller counterparts. For this analysis, a special emphasis was placed on identifying subsets of data points where surprisal from larger LM variants deviated more drastically from humanlike processing difficulty.

4.1 Calculation of Residual Errors

The seventeen LME models that contain each of the LM surprisal predictors described in Section 3.3 were used to generate predictions for all data points in the exploratory set of both self-paced reading times and go-past durations.[6] Subsequently, the predictions were subtracted from the target values to calculate the residual errors for each of the seventeen regression models.

However, a preliminary analysis of the LME models fitted to the Dundee Corpus revealed a discrepancy between model likelihood and mean squared error (MSE), where the regression models with higher likelihoods achieved similar MSEs to those with lower likelihoods. This is because the `lme4` package (Bates et al., 2015) minimizes the *penalized* residual sum-of-squares, which includes a Euclidean norm penalty on the spherical component of the random effects variables. In other words, an LME model can achieve higher likelihood than another if it can achieve similar MSE using less expressive random effects variables that have lower variance.

An inspection of the fitted random effects variables revealed that the by-word intercept was mostly responsible for the discrepancy between likelihood and MSE for the LME models fitted to the Dundee Corpus. More specifically, the LME models with surprisal estimates from larger LM variants had systematically higher variance for the by-word intercept, which allowed them to achieve similar MSEs at the cost of an increased penalty for the random effects variables. In order to control for this confound and bring model likelihood and MSE closer together, the seventeen LME models were fitted again to both corpora with the by-word random intercepts removed. Since the goal of this analysis was to identify data points that are responsible for the positive correlation between LM perplexity and fit to human reading times, it was thought that removing the by-word random intercepts would also yield a clearer picture with regard to words on which the surprisal estimates from larger LMs fall especially short.

The MSEs plotted in Figure 2, which generally replicate the inverse trend of ΔLLs in Figure 1, show that the removal of by-word random intercepts brought model likelihoods and MSEs closer. The residual errors from these newly fitted regression models were subsequently analyzed.

4.2 Annotation of Data Points

In order to guide the identification of linguistic phenomena underlying the trend observed in Section 3.4, each data point in both corpora was associated with various word- and sentence-level properties that are thought to influence real-time processing. These properties were derived from the manually annotated syntactic tree structures of both corpora from Shain et al. (2018).

Word-level properties reflect characteristics of the word that generally hold regardless of the surrounding context:

- Part-of-speech: the syntactic category of each word from a generalized categorial grammar annotation scheme (Nguyen et al., 2012; Shain

[6] The post-hoc analysis focused on the exploratory set, as the held-out set is reserved for statistical significance testing.

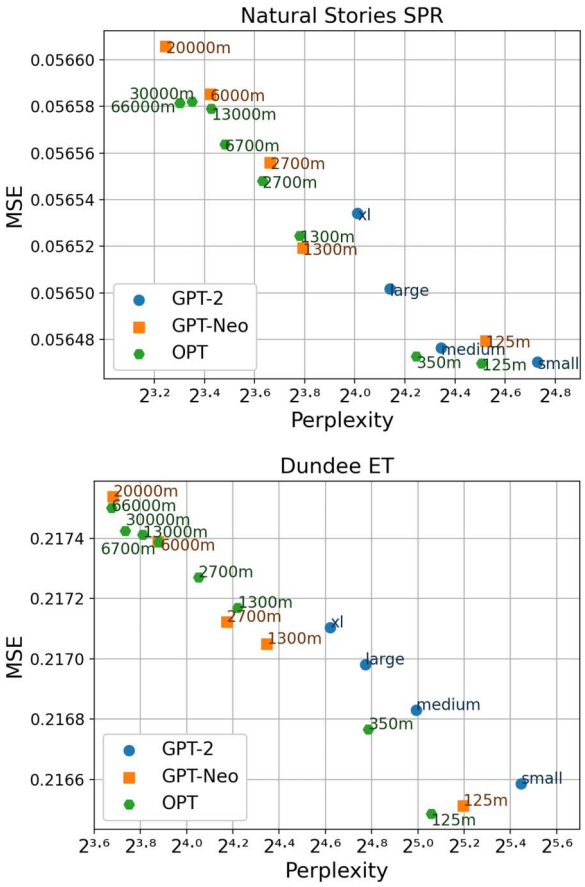

Figure 2: Perplexity measures from each LM variant, and mean squared errors of regression models that include each surprisal estimate on the exploratory set of Natural Stories (top) and Dundee data (bottom). Note that the larger values of MSE correspond to smaller values of log-likelihood in Figure 1.

et al., 2018). As these categories are defined in terms of primitive types (e.g., verbs and nouns) and type-combining operators (e.g., unsatisfied preceding and succeeding arguments), they make more fine-grained distinctions in terms of linguistic subcategorization.

- Named entities: a binary variable for whether or not the word is part of a proper name. Since words at the beginning of sentences were excluded from regression modeling, capitalization reliably identified such named entities.[7]

Sentence-level properties capture the syntactic structure of sentences, either in terms of dependencies or hierarchical phrases:

- Dependency Locality Theory (DLT; Gibson, 2000) cost: DLT posits that the construction of backward-looking dependencies between words (e.g., between a verb and its subject) incurs an 'integration' cost driven by memory retrieval operations. This cost is thought to be proportional to the length of the dependency in terms of the number of intervening discourse referents, which are operationalized as any noun or finite verb in this work.
- Left-corner parsing (Johnson-Laird, 1983): A left-corner parser incrementally derives phrasal structures from a series of lexical match and grammatical match decisions at every word.[8] These two decisions allow center-embedded constituents to be distinguished from non-embedded constituents. Additionally, the grammatical decision results in expectations about the upcoming syntactic category, which allows words before complete constituents (e.g., words before sentential clauses) to be identified.

This annotation allowed the data points in each corpus to be subsetted, which subsequently helped identify where surprisal from the larger LM variants deviated further from humanlike processing.

4.3 Iterative Slope-Based Analysis of Residual Errors

Subsequently, based on the properties annotated in Section 4.2, subsets of data points that strongly drive the trend in Figure 2 were identified. To this end, the linear relationship between log perplexity and MSEs was used; subsets of data points that drive the general trend should show larger differences in MSE between regression models, or in other words, have negative slopes that are steeper than the corpus-level slope.

Based on this idea, for every corpus-LM combination (i.e., {Natural Stories, Dundee} × {GPT-2, GPT-Neo, OPT}), a least-squares regression line was fitted between corpus-level log perplexity and MSEs of each subset defined by the properties outlined in Section 4.2. Subsequently, the subset with the steepest negative slope was identified. After excluding the identified subset, the above procedure was repeated to identify a new subset that showed the next strongest effect. For this analysis,

[7] Words like the pronoun *I* and names of fictional characters that appeared in the Natural Stories Corpus were manually excluded afterwards.

[8] See, e.g., Oh et al. (2022) for a more detailed definition of left-corner parsing models.

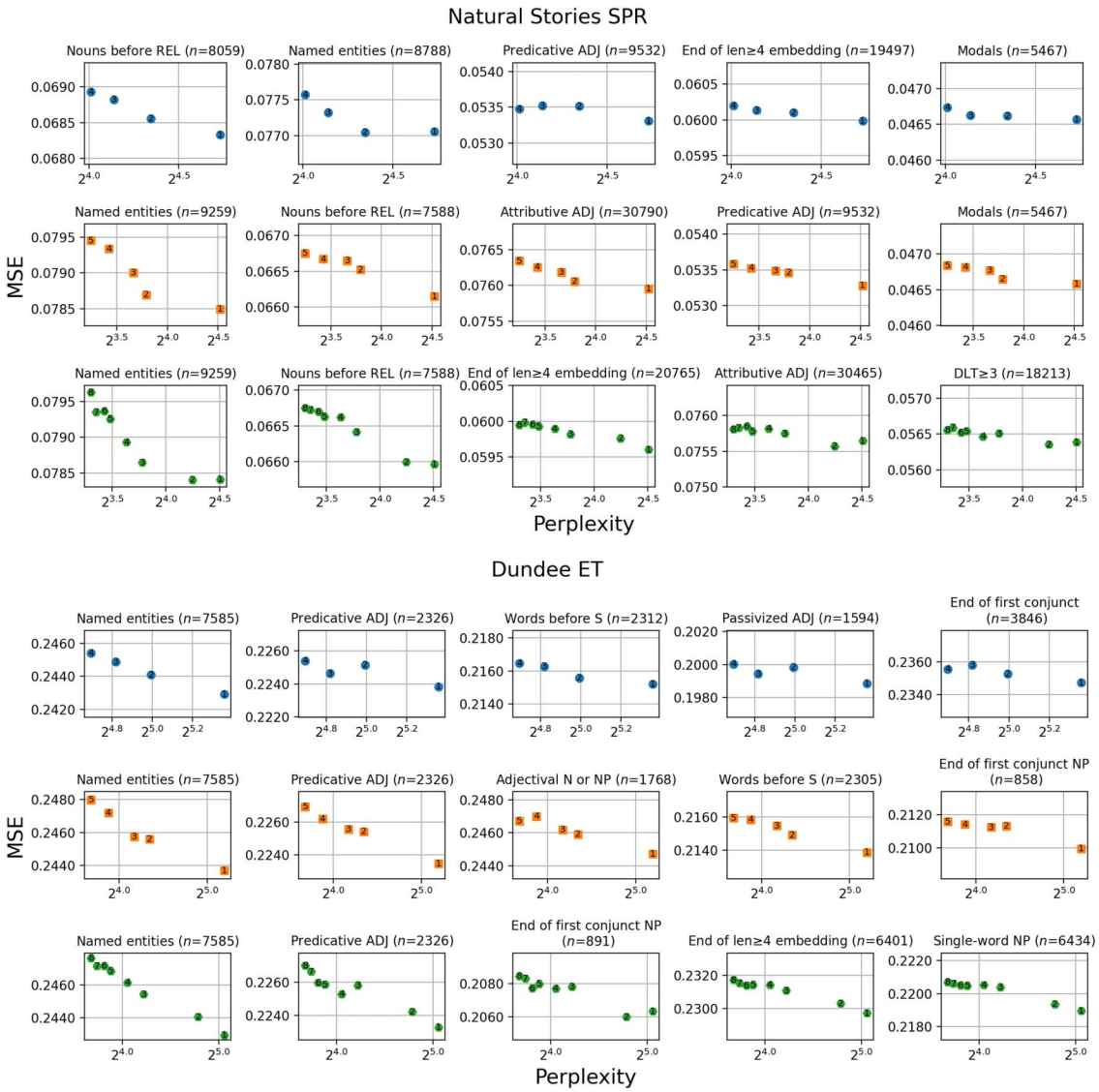

Figure 3: Corpus-level perplexity measures from each GPT-2, GPT-Neo, and OPT model variant (top, middle, and bottom rows, respectively), and mean squared errors of regression models that include each surprisal estimate on the top five subsets (columns ordered from left to right) of Natural Stories self-paced reading data (top panel) and Dundee eye-tracking data (bottom panel). The ordered labels represent LM variants of different sizes, with '1' representing the smallest variant. ADJ: adjective, N: noun, NP: noun phrase, REL: relativizer, S: sentential clause.

only subsets that contained more than 1% of the data points in each corpus were considered at each iteration.[9]

Additionally, once the subsets of interest were identified, the data points in each subset were further separated according to whether the regression model underpredicted or overpredicted the target reading times. In order to identify whether the trend of MSEs is driven primarily by systematic underprediction or overprediction of reading times, the average surprisal from each LM variant and the sum of squared errors (SSE) were calculated for each subset. SSEs instead of MSEs were analyzed because different regression models had different numbers of underpredicted vs. overpredicted data points, and because points close to 0 can distort the MSEs and obscure the overall trend of mispredictions.

4.4 Results

The results in Figure 3 show that on each corpus, similar subsets were identified as driving the trend of MSEs across different LM families. On

[9]This criterion amounts to >3,849 data points for Natural Stories and >981 data points for Dundee at the first iteration. Although this may seem like a lenient criterion, this was necessary to examine phenomena that lie at the long tail of the Zipfian distribution of word frequencies.

the Natural Stories Corpus, these subsets were primarily determined by the word's syntactic category, such as named entity nouns, nouns before relativizers, attributive and predicative adjectives, and modals. The top subsets of the Dundee Corpus were similarly determined by syntactic category, such as named entity nouns, predicative and passivized adjectives, and single-word noun phrases (e.g., pronouns). Subsets defined by the syntactic structure of sentences were less commonly identified from the Natural Stories Corpus, with ends of center-embedded constituents spanning four or more words and words with high DLT costs emerging. From the Dundee Corpus, ends of center-embedded constituents, ends of first conjunct constituents (both overall and noun phrases specifically), and beginnings of adjectival noun phrases (e.g., *family* in *a family size pack*) were identified. Additionally, words preceding a sentential clause were identified, which corresponded to conjunctions and ends of adjuncts. On most of these subsets, the MSEs of each regression model were higher than those on the entire corpus, which indicates that the misprediction of reading times that pre-trained LM surprisal already has difficulty modeling is exacerbated as the models get larger. Subsets such as modals of Natural Stories and first conjunct NP endings of Dundee are counterexamples to this general trend.

The average surprisal values[10] and SSEs from underpredicted and overpredicted data points in Figure 4 shed more light on the direction and magnitude of mispredictions from each regression model. For example, on the subset of named entities, which emerged as the top two subsets across all corpus-by-LM combinations, the larger LM variants show systematically higher SSEs due to underprediction. This strong discrepancy highlights a mismatch between human sentence processing and language modeling; named entity terms (e.g., *Elvis Presley*) have been shown to incur increased processing times compared to their common noun counterparts (e.g., *a singer*) due to various semantic associations that are retrieved (Proverbio et al., 2001; Wang et al., 2013). In contrast, for language models, named entity terms that typically consist of multiple tokens have high mutual information, making it easy for them to accurately predict subsequent tokens given the first (e.g., *Presley* given *Elvis*), resulting in especially lower surprisal estimates for larger LM variants.

Similarly, across the two corpora and three LM families, the trend of MSEs for other nouns as well as adjectives appears to be consistently driven by more severe underpredictions from regression models containing surprisal estimates from larger LM variants. On these subsets, the difference in average surprisal values between the smallest and largest LM variants was typically above 2 bits, which is larger than the difference in log perplexity (i.e., corpus-level average surprisal, Figure 2) between these variants. This indicates that these subsets represent words that the larger LM variants predict especially accurately, which results in low surprisal estimates that deviate from human reading times.

In contrast, the subset of modals on the Natural Stories Corpus identified for the GPT-2 and GPT-Neo models shows a completely opposite trend in which more severe overpredictions drive the overall trend of MSEs. This seems to be more due to the difference in the estimated regression coefficients rather than the difference in the LM surprisal estimates themselves. The average surprisal values on this subset show that the difference between their smallest and largest variants is less than 1 bit, which indicates that the LM variants are making more similar predictions about modals. However, since surprisal predictors from larger LM variants are generally smaller in magnitude, the regression models assign them higher coefficients in order to predict reading times, resulting in a systematic overprediction given surprisal predictors of similar values. This also explains the trend observed for words preceding a sentential clause on the Dundee Corpus, which mainly consisted of conjunctions.

Finally, while they were less common, subsets based on syntactic complexity were also identified as driving the differential fit to reading times. On the Natural Stories Corpus, a systematic underprediction of regression models with OPT surprisal was observed on words with a DLT cost of greater than or equal to three. These words mainly consist of nouns and finite verbs that complete long-distance dependencies. While finite verbs in general were not identified as subsets that showed a strong effect, it is likely that the increased reading times caused by the construction

[10]Since perplexity is equivalent to exponentiated average surprisal, the average surprisal values for each subset are roughly comparable to LM perplexity of, e.g., Figure 2. However, caution is warranted as these values are calculated over data points of reading times instead of tokens.

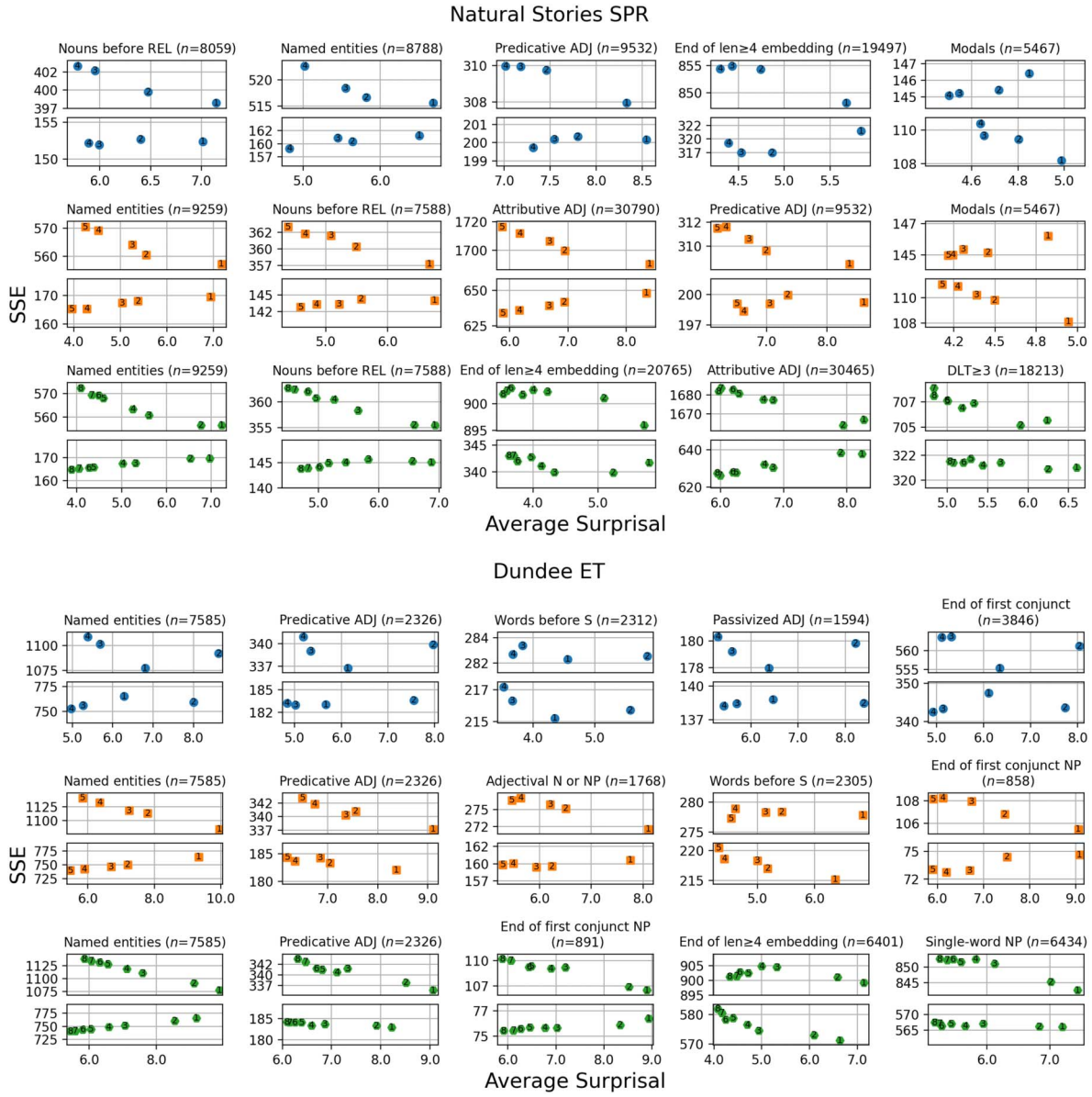

Figure 4: Average surprisal from each GPT-2, GPT-Neo, and OPT model variant, and sum of squared errors of regression models that include each surprisal estimate on the top five subsets of Natural Stories self-paced reading data and Dundee eye-tracking data. The top and bottom subplots of each row represent values from underpredicted and overpredicted data points, respectively.

of long-distance dependencies made the underpredictions more salient. Ends of center-embedded constituents of length greater than or equal to four were also identified, which typically corresponded to nouns and adjectives. On the Natural Stories Corpus, the trend of more severe underpredictions driving the effect is consistent with other noun and adjective subsets. However, on the Dundee Corpus, overpredictions seem to be responsible for the overall trend in this subset, which may hint at subtle differences in how syntactic complexity is manifested in self-paced reading times and eye-gaze durations.

Taken together, these results indicate that the poorer fit to human reading times achieved by surprisal estimates from larger Transformer-based language models is primarily driven by their characteristic of assigning lower surprisal values to open-class words like nouns and adjectives, which may be accurately predicted by extensive domain knowledge gleaned from large sets of training examples that are not available to humans. In other words, the extra parameters of the larger LM variants may be improving predictions of such words in a way that is beyond human ability.

5 Discussion and Conclusion

This work presents results using multiple large pre-trained LMs showing that larger variants with more parameters and better next-word prediction performance (i.e., lower perplexity) nonetheless yield surprisal estimates that are less predictive of human reading times (i.e., smaller contribution to regression model fit), corroborating and expanding upon earlier results based on the GPT-2 LM (Oh et al., 2022).

First, in order to examine the generalizability of this trend, surprisal estimates from five variants of the GPT-Neo LM and eight variants of the OPT LM were evaluated in terms of their ability to predict self-paced reading times and eye-gaze durations. The regression analysis revealed a strictly monotonic, positive log-linear relationship between perplexity and fit to reading times for five GPT-Neo variants and eight OPT variants, providing robust empirical support for this trend. Additionally, the different data used to train each LM family seem to influence the quality of surprisal estimates, although more pre-training data did not necessarily result in surprisal estimates that are more predictive of reading times.

Subsequently, to identify the data points that are responsible for the positive relationship, a post-hoc analysis of the residual errors from each regression model was conducted. The results showed that the difference in MSEs between regression models containing surprisal predictors from different LM variants was especially large on nouns and adjectives, such as named entity terms and predicative adjectives. A further inspection of their predictions showed that the trend of MSEs on these words was driven mainly by underpredictions of reading time delays, which were exacerbated as the larger LM variants predicted the words more accurately and assigned lower surprisal values. This tendency also led to higher regression coefficients for surprisal estimates from larger LM variants, which resulted in a systematic overprediction at function words like conjunctions and modals that had similar surprisal estimates across LM variants.

The 'more and more superhuman' predictions of larger LM variants observed in this work are consistent with findings from recent analyses of Transformer-based LMs. For example, a mathematical analysis of Transformers (Elhage et al., 2021) showed that a layer of self-attention essentially functions as a lookup table that keeps track of bigram statistics of the input data. Given this observation, it may be the case that the larger LM variants with more attention heads at their disposal have the capability to learn stronger local associations between tokens. This possibility was empirically supported from the perspective of memorization by Carlini et al. (2022), who found that larger variants of the GPT-Neo model returned more sequences verbatim from the pre-training data during greedy decoding. This behavior may explain why nouns and adjectives showed the strongest effect in the post-hoc analysis; since adjectives and nouns typically have higher type-frequency than verbs or function words, it may be the case that nouns and adjectives that are rarely seen during training are predicted much more faithfully by the larger LM variants with higher model capacity. Additionally, this also suggests that as these pre-trained LMs continue to get bigger, they will continue to degrade as models of humanlike language comprehension.

The 'trained-from-scratch' LMs studied in earlier psycholinguistic modeling work (e.g., Goodkind and Bicknell, 2018; Wilcox et al., 2020) show a negative relationship between perplexity and fit to reading times. However, based on regression results following the same protocols as Section 3, surprisal estimates from LMs trained in Wilcox et al. (2020) generally seem to be less predictive of human reading times than those from pre-trained LMs examined in this work. Given the especially large discrepancy in model size between newly trained LMs and pre-trained LMs, it may be the case that they capture two distinct regimes in terms of the relationship between LM performance and predictive power of surprisal estimates. While the results of the current study clearly show that surprisal estimates from smaller pre-trained LM variants are more predictive of reading times, it remains to be seen how much smaller LMs can become before the predictive power of surprisal estimates starts to decrease. With recently increasing effort in developing efficient NLP models, future work could explore the extent to which, for example, knowledge distillation techniques (Sanh et al., 2019) can result in LMs that are more predictive of humanlike processing difficulty.

Additionally, the importance of being 'adequately surprised' at nouns like named entity

terms that was identified in the current study may also explain similar recent counterexamples to the trend observed between model perplexity and fit to reading times (Oh et al., 2021; Kuribayashi et al., 2021). Oh et al. (2021) showed that incorporating a character model to estimate word generation probabilities within an incremental left-corner parser resulted in more predictive surprisal estimates compared to those from a baseline parser that treats words as symbols, although at a cost of higher test perplexity. The character model may be effectively assigning higher surprisal values to these rare words, thereby achieving better fit to human reading times. The reading times of Japanese text studied in Kuribayashi et al. (2021) were measured in larger units (i.e., *bunsetsu*; roughly equivalent to phrases) than typical English words. Therefore, the Japanese LMs analyzed in that study are likely to have been trained on and have made predictions on text that has been tokenized into 'sub-bunsetsu' tokens, which may have made a different picture emerge from results based on purely word-based LMs of earlier work.

In general, the tendency of pre-trained LM surprisal to underpredict reading times observed in this work is consistent with recent empirical shortcomings of neural LM surprisal. For example, van Schijndel and Linzen (2021) and Arehalli et al. (2022) found that surprisal from neural LMs severely underpredicts the magnitude of garden-path effects demonstrated by human subjects. Similarly, Hahn et al. (2022) showed that surprisal from GPT-2 fails to accurately predict the increase in reading times at the main verb of deeply embedded sentences. Kuribayashi et al. (2022) also demonstrated that implementing a recency bias by deterministically truncating the context window of neural LMs leads to surprisal estimates that alleviate the underpredictions of full neural LM surprisal on naturalistic reading times of English and Japanese text. Taken together, these results suggest that neural LMs do not make abstract, linguistic generalizations like people do.

Moreover, there are also efforts to evaluate other memory- and attention-based predictors calculated from Transformer-based LM representations on their ability to predict human behavior. For instance, Ryu and Lewis (2021) drew connections between the self-attention mechanism of Transformers and cue-based retrieval models of sentence comprehension (e.g., Lewis et al., 2006). Their proposed attention entropy, which quantifies the diffuseness of attention weights over previous tokens, was found to show profiles that are consistent with similarity-based interference observed during the processing of subject-verb agreement. Oh and Schuler (2022) expanded upon this idea and showed that the entropy of attention weights at a given timestep as well as the shift in attention weights across consecutive timesteps are robust predictors of naturalistic reading times over GPT-2 surprisal. Hollenstein and Beinborn (2021) calculated the norm of the gradient of each input token on two eye-tracking corpora using BERT (Devlin et al., 2019) as a metric of saliency, which showed higher correlations to fixation durations compared to raw attention weights.

Finally, it is becoming more common in psycholinguistic modeling to use surprisal from pre-trained LMs as a baseline predictor to study various effects in naturalistic sentence processing (e.g., Ryu and Lewis, 2022; Clark and Schuler, 2022). The broader implication of the current study is that researchers should not select the largest pre-trained LM available based on the widely held 'larger is better' assumption of the NLP community. As a general practice, surprisal estimates from smaller pre-trained LM variants should be incorporated to form a more rigorous baseline, which will guard against drawing unwarranted scientific conclusions.

Acknowledgments

We thank our TACL action editor and the reviewers for their helpful comments. This work was supported by the National Science Foundation grant #1816891. All views expressed are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning*, pages 301–313.

Christoph Aurnhammer and Stefan L. Frank. 2019. Comparing gated and simple recurrent

neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, pages 112–118.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136. https://doi.org/10.18653/v1/2022.bigscience-1.9

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large scale autoregressive language modeling with Mesh-Tensorflow. *Zenodo*. https://doi.org/10.5281/zenodo.5297715

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv preprint*, arXiv:2202.07646v2. https://doi.org/10.48550/arXiv.2202.07646

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734. https://doi.org/10.3115/v1/D14-1179

Christian Clark and William Schuler. 2022. Evidence for composition operations in broad-coverage sentence processing. In *35th Annual Conference on Human Sentence Processing*.

Vera Demberg and Frank Keller. 2008. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 109(2):193–210. https://doi.org/10.1016/j.cognition.2008.07.008, PubMed: 18930455

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 199–209. https://doi.org/10.18653/v1/N16-1024

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for Transformer circuits.

Jeffrey L. Elman. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225. https://doi.org/10.1007/978-1-4615-4008-3_5

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. 2021. The Natural Stories corpus: A reading-time corpus of English texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55(1):63–77. https://doi.org/10.1007/s10579-020-09503-7, PubMed: 34720781

Edward Gibson. 2000. The Dependency Locality Theory: A distance-based theory of linguistic complexity. In *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, pages 95–126, Cambridge, MA. MIT Press.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*, pages 10–18. https://doi.org/10.18653/v1/W18-0102

Michael Hahn, Richard Futrell, Edward Gibson, and Roger P. Levy. 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, PubMed: 36260742

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, pages 1–8. https://doi.org/10.3115/1073336.1073357

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. 2018. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2727–2736. https://doi.org/10.18653/v1/P18-1254

Yiding Hao, Simon Mendelsohn, Rachel Sterneck, Randi Martinez, and Robert Frank. 2020. Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–86. https://doi.org/10.18653/v1/2020.cmcl-1.10

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 141–150. https://doi.org/10.18653/v1/2021.acl-short.19

Philip N. Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA.

Alan Kennedy, Robin Hill, and Joël Pynte. 2003. The Dundee Corpus. In *Proceedings of the 12th European Conference on Eye Movement*.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5203–5217. https://doi.org/10.18653/v1/2021.acl-long.405

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006, PubMed: 17662975

Richard L. Lewis, Shravan Vasishth, and Julie A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10(10):447–454. https://doi.org/10.1016/j.tics.2006.08.007, PubMed: 16949330

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. https://doi.org/10.18653/v1/2021.cmcl-1.2

Luan Nguyen, Marten van Schijndel, and William Schuler. 2012. Accurate unbounded dependency recovery using generalized categorial grammars. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 2125–2140.

Byung-Doh Oh, Christian Clark, and William Schuler. 2021. Surprisal estimators for human reading times need character models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3746–3757.

Byung-Doh Oh, Christian Clark, and William Schuler. 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, PubMed: 35310956

Byung-Doh Oh and William Schuler. 2022. Entropy- and distance-based predictors from GPT-2 attention patterns predict reading times over and above GPT-2 surprisal. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9324–9334.

Alice Mado Proverbio, Stefania Lilli, Carlo Semenza, and Alberto Zani. 2001. ERP indexes of functional differences in brain activation during proper and common names retrieval. *Neuropsychologia*, 39(8):815–827. https://doi.org/10.1016/S0028-3932(01)00003-3, PubMed: 11369405

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Soo Hyun Ryu and Richard L. Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with Transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 61–71. https://doi.org/10.18653/v1/2021.cmcl-1.6

Soo Hyun Ryu and Richard L. Lewis. 2022. Using Transformer language model to integrate surprisal, entropy, and working memory retrieval accounts of sentence processing. In *35th Annual Conference on Human Sentence Processing*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Marten van Schijndel and Tal Linzen. 2021. Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, 45(6):e12988. https://doi.org/10.1111/cogs.12988, PubMed: 34170031

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45). https://doi.org/10.1073/pnas.2105646118, PubMed: 34737231

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. https://doi.org/10.18653/v1/P16-1162

Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. 2020. fMRI reveals language-specific predictive coding during naturalistic sentence comprehension. *Neuropsychologia*, 138:107307. https://doi.org/10.1016/j.neuropsychologia.2019.107307, PubMed: 31874149

Cory Shain and William Schuler. 2021. Continuous-time deconvolutional regression for psycholinguistic modeling. *Cognition*, 215:104735. https://doi.org/10.1016/j.cognition.2021.104735, PubMed: 34303182

Cory Shain, Marten van Schijndel, and William Schuler. 2018. Deep syntactic annotations for broad-coverage psycholinguistic modeling. In *Workshop on Linguistic and Neuro-Cognitive Resources*.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading

time is logarithmic. *Cognition*, 128:302–319. https://doi.org/10.1016/j.cognition.2013.02.013, PubMed: 23747651

Alan D. Thompson. 2022. What's in my AI? A comprehensive analysis of datasets used to train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. *Life-Architect.ai Report*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 billion parameter autoregressive language model. https://github.com/kingoflolz/mesh-transformer-jax

Lin Wang, Zude Zhu, Marcel Bastiaansen, Peter Hagoort, and Yufang Yang. 2013. Recognizing the emotional valence of names: An ERP study. *Brain and Language*, 125(1):118–127. https://doi.org/10.1016/j.bandl.2013.01.006, PubMed: 23467262

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained Transformer language models. *arXiv preprint*, arXiv:2205.01068v4. https://doi.org/10.48550/arXiv.2205.01068