"Is Bigger Always Better?": Comparing The Surprisals Of LLMs Against Humans For Sentence Comprehension

Anonymous ACL submission

Abstract

This paper investigates the similarities and differences between human and machine language processing by comparing human and machine surprisals from two self-paced-reading corpora. The study examines how the frequency distribution of surprisals changes with increasing context length and presents evidence that with greater context, both humans and machine language models can better predict upcoming words, resulting in narrow surprisal values. The study also analyzes how machine surprisals behave differently from human surprisals across parts of speech tags, and shows that increasing context size leads to better correlation with human processing effort. The findings also suggest that with increasing model complexity, machine language models may capture a wider range of cognitive and neural processes, potentially providing a more accurate representation of human language processing.

1 Introduction

Sentence Comprehension is a topic is of interest for those studying human language processing as well as those studying algorithmic Natural Language Processing. It is a complicated process that involves integrating many levels of linguistic information. And so, conducting a comparative analysis between humans and machine models (capable of natural language understanding and generation) represents an intriguing research direction.

Sentence Comprehension is thought to be a process that combines the ability to predict and eventually integrate the upcoming words into the working memory. And in order to understand this phenomenon, researchers have employed various methods over the years. One such method is surprisal. The concept of surprisal (Hale, 2016) originating in information theory, has been utilized to measure the level of unexpectedness of a word based on the preceding context. It has also found its place in psycholinguistics, where it been formulated as a function of metrics such as reading time(Monsalve et al., 2012). Previous studies have shown that higher levels of surprisal[1] predict longer reading times (Lowder et al., 2018; Monsalve et al., 2012; Staub, 2015). Additionally, surprisal has been utilized to evaluate the ability of neural language models (eg. Davis and Van Schijndel (2020)) to learn "human-like" language structures. In the present study, we employ surprisal to investigate large pretrained language models (based on the Transformer architecture) and how are they related with notions of human surprisal (estimated from reading times).

Recent advances in deep neural networks have led to the development of large language models (based on the Transformer architecture), such as the GPT family of auto-regressive models, that are capable of generating high-quality text. The effectiveness of these models has opened up new avenues for investigating the extent to which language models capture human cognitive processes (Michaelov et al., 2021; Binz and Schulz, 2023; Kuribayashi et al., 2022).

This study aims to connect machine surprisal (calculated from the logits returned from GPT2 language models) with human reading times. In this paper, we consider reaction times from self-paced reading[2] experiments as the indicator of reading times. We are interested in understanding how increasing the parameter size of the models affect the relationship between machine surprisal and human reading times. Specifically, we explore if larger models can be considered to be better models of human cognition. In doing this, we also look at how context length affects the surprisals in both humans and machines.

This paper adopts the stand that the human language processing system and the GPT2 models represent two different types of "language models". And thus, machine surprisal (obtained from GPT2 models), and reading times (collected from humans), reflect the processing mechanisms of these two distinct "language models"[3]. And so our thesis is that, as a model becomes more "cognitively plausible", the predictability of a

[1] These studies use eye-tracking data and Recurrent Neural Networks Language Models to calculate surprisal

[2] http://www.intro2psycholing.net/resources/experiments/selfpaced.php

[3] This can be seen as a way to think about extrinsic evaluation of the two very different 'language models'

word as reflected in the machine surprisal would be close to the predictability (as defined in Bianchi et al. (2020)) of a word as per a human participant. And if the progress in language modelling using the Transformer architecture over the last few years is taken into consideration, models with larger parameter sizes have out-performed smaller mdoels. And so the question is: are models with larger parameter sizes also more "cognitively plausible"?

To do this, we use reading time data from two existing datasets and surprisals calculated from four different GPT2 language models with varying parameter sizes. Our study contributes to the growing body of research that is exploring the extent to which deep neural networks can serve as credible models of human language processing.

2 Related Work

Studies on the cognitive plausibility of language models have highlighted both their strengths and weaknesses in modeling human language processing. On one hand, multiple studies have shown that language models can predict a range of language processing phenomena like reading times, word recognition, and syntactic processing with a high degree of accuracy (Smith and Levy, 2013; Frank and Bod, 2011; Demberg and Keller, 2008). These studies suggest that language models may be able to capture some aspects of the cognitive processes underlying language comprehension. However, it should be mentioned that in most psycholinguistic accounts of sentence processing, reading times often imply data collected from physiological data like eye-tracking or fMRI data among others. But in this paper, we primarily look at the much coarser reaction time data.

On the other hand, some researchers have raised concerns about the limitations of language models in capturing the full complexity of human language processing. For example, Bender and Koller (2020) argue that language models may be limited by the assumptions and biases present in the training data, and may not be able to capture certain aspects of linguistic knowledge, such as pragmatic reasoning or world knowledge.

Nonetheless, the use of language models in cognitive science research has opened up new avenues for investigating the cognitive processes underlying language comprehension. For example, recent studies have used language models to investigate how syntactic and semantic factors interact during language processing (Linzen et al., 2016; Hupkes et al., 2018), and how individual differences in working memory and attention influence language processing ((Schwering and MacDonald, 2020) Other studies have explored the relationship between language model surprisal and brain activity during language processing (Hale et al., 2018), further supporting the idea that language models capture some aspects of human language processing. Moreover, recent research has also shown that the correlation between language model surprisal and reading times may depend on the size of the language model (Futrell et al., 2019; Gulordava et al., 2018).

The present paper contributes to this line of research by investigating how the parameter size of language models and context size affect the relationship between reading time and language model surprisal.

3 Datasets

We use two different datasets to study how language models were similar/dissimilar to humans in terms of the information theoretic definition of surprisal. Both corpora were curated using text from existing novels and each corpus represented a different aspect of context. Since both corpora have previously been used to test theories in psycholinguistics, they seemed to be good candidates to test the hypothesis that language models were valid computational models of human language processing.

3.1 Natural Stories Corpus

The Natural Stories Corpus (Futrell et al., 2017) was created as an attempt to include low-frequency syntactic constructions into sentences so that different processing theories could be tested. The corpus consists of 10 stories with a total of 10,245 words and 485 sentences. As part of the dataset, the authors also include parse trees generated using the Stanford parser (followed by hand correcting of the results) and self-paced reading data. For the purpose of this paper we use the actual stories from the corpus and the self-paced reading data to make comparisons with the machine results.

3.2 UCL corpus

The UCL corpus (Frank et al., 2013) was designed as a standard dataset for the evaluation of computational psycholinguistic models. The dataset consists of 361 stimuli sentences collected from 3 different novels. As part of the dataset, the authors release the reading times and eye-tracking data from a psycholinguistic task of self-paced reading (of 43 subjects) of the sentences in the dataset.

Overall, the stimuli from the Natural Stories corpus were longer in comparison with the stimuli from the UCL corpora. Now, we know that human reading times are affected by a multitude of factors. And hence, the disparate nature of the corpora could help pin down the common similarities in the humans and machines by comparing the reading time and machine surprisal behavior across the two corpora.

4 Concepts Used

Reading time is a measure commonly used in psycholinguistics to study how language is processed in the human brain. It refers to the amount of time a person takes to read a specific piece of text. Reading time is affected by various factors, including complexity of the content, the reader's level of expertise and their cognitive and linguistic abilities. Studying reading time can provide insights into the cognitive processes involved in reading.

Surprisal or self-information is a concept in information theory that measures the degree of unexpectedness of an event or a message. In the context of language processing (Hale, 2001; Levy, 2008), surprisal refers to the level of uncertainty or unpredictability associated with a given word or sequence of words in a text. It is typically calculated using probabilistic language models that estimate the probability of a word given its preceding context. Words with low probability or high surprisal are more difficult to process and can lead to slower reading times and increased cognitive processing effort. Processing Effort was earlier related with uncertainty by Frank (2010).

5 Methodology

We calculate surprisal from four different models[4] of the GPT2 family by using the usual information-theoretic formulation of surprisal. The surprisal of observing a particular word w_i given its preceding context $(w_0, w_1, ..., w_{i-1})$ can be calculated using the following equation:

$$S(w_i|w_0...w_{i-1}) = -\log_2 P(w_i|w_0...w_{i-1}) \quad (1)$$

where $P(w_i|w_0...w_{i-1})$ is the probability of observing word w_i given its preceding context. The logarithm base 2 is used to convert the probability into bits, which represents the amount of information conveyed by the occurrence of the word w_i in the context. The surprisal value $S(w_i|C_i)$ is high when the observed word is unexpected given its preceding context and low when the word is highly predictable.

As mentioned earlier, we consider reading time to be proportional to the processing effort (P). More specifically:

$$RT \propto P$$

Specifically, we define the reading time (RT) as a function of different cognitive processes (φ), length of the word (L) and the effort required to read the word (E). And hence, we envisage the functional form of RT to be given as:

$$RT_w = f(\varphi, L_w, E_w) \quad (2)$$

[4] gpt2=124M parameters; gpt2-medium=355M parameters, gpt2-large:774M parameters; gpt-xl: 1.5B parameters

Hence we hypothesize that the ratio of RT to the length of a word would be proportional to the 'actual effort' taken to read the word. In other words:

$$\frac{RT_w}{L_w} \propto E_w$$

And so the question is, do bigger models lead to better correlation with the observed human data?

We assume that both GPT style models and human language models incorporate incremental processing mechanisms. There is also some evidence that human brains think in terms of 'sub-words' (Solomyak and Marantz, 2009; Nieuwland, 2019). And hence, we assume that even for humans, if a word is represented as k sub-words in the language processing system, then:

$$p_{word} = p_1 \times ... \times p_k$$

Also, "processing effort" to read the can be written as:

$$P = f(p_1) + ... + f(p_k)$$

which implies that:

$$P = \theta \log_2 (p_{word})$$

where θ is a scaling parameter. Now, given that we assume RT_{word} to be an indicator of p_{word}, we rewrite "processing effort" as:

$$PE_w = \log_2 (E_w) \quad (3)$$

We also assume that for models with greater degree of "cognitive plausibility", this ratio would have a positive correlation with the machine values of surprisal. And so the question is, are bigger models more "cognitively plausible"?

6 Observations

We start our analysis by looking at the nature of frequency distribution of the normalized reading times and machine surprisals from both the corpora.

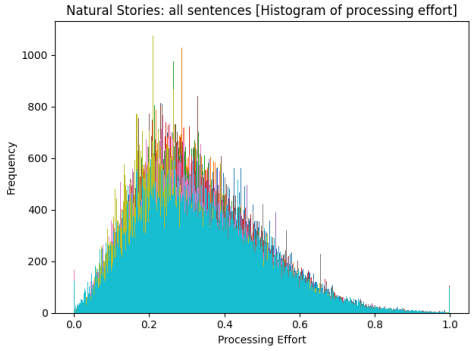

Figure 1: Natural Stories: Histogram of Human "processing effort" across all stimuli

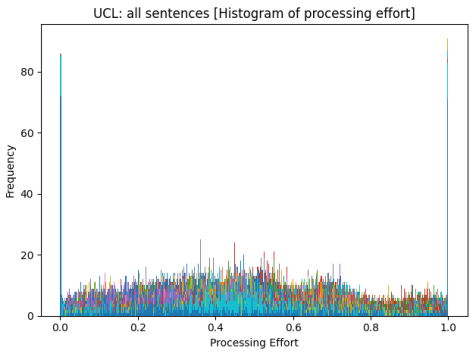

Figure 2: UCL: Histogram of Human "processing effort" across all stimuli

We perform normalization on both the Processing Effort and Machine Surprisal to compare the human and machine surprisals on a comparable scale (0 to 1). It seems that the frequency distribution of both the Processing Effort and Machine Surprisal has a single mode for the Natural Stories corpus but has multiple modes for the UCL corpus. To confirm the nature of modality of the distributions, we proceed to perform the Hartigan Dip-test of Unimodality (Hartigan and Hartigan, 1985) implemented using Python[5]. The results for the test for both corpora are shown in Tables 1 and 3.

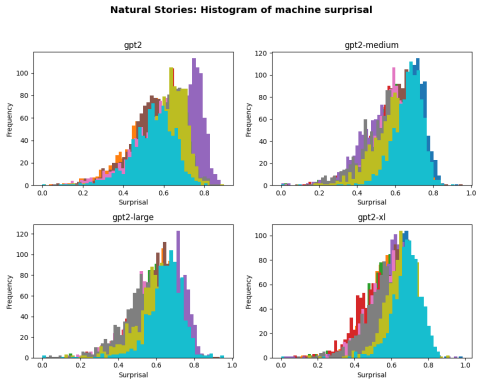

Figure 3: Natural Stories: Histogram of frequencies of surprisals of GPT2 models with different parameter sizes.

Category	d	p-value
human-RT	0.00066	0.311688
GPT2	0.00183	1.0
GPT2-medium	0.00197	1.0
GPT2-large	0.00189	1.0
GPT2-xl	0.00226	0.995005

Table 1: p-values for Dip-Test results for Natural Stories corpus

[5]https://github.com/BenjaminDoran/unidip

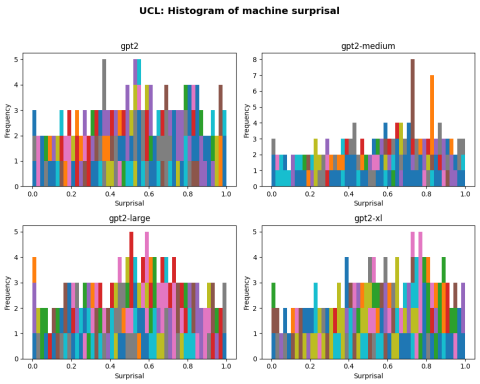

Figure 4: UCL: Histogram of frequencies of surprisals of GPT2 models with different parameter sizes.

Based on the p-values in Table 3, we conclude that the frequency distribution of human processing effort and machine surprisal are indeed multimodal ($p<0.05$) for the UCL corpus. The frequency distribution of processing effort and surprisal for the Natural Stories corpus Table 1 on the other hand seems to unimodal as suspected earlier. However, to ascertain if the nature of this distribution changes with increasing context length, we repeat the Dip-test after splitting the stimuli in the Natural Stories Corpus into three parts based on their lengths. Hence, the first part of the split contained the first $\frac{1}{3}$ of the stimuli and so on. The results from this Dip-test are shown in Table 2.

Model	Half 1	Half 2	Half 3
human-RT	0.087912	0.771229	0.742258
GPT2	1.0	1.0	0.996003
GPT2-medium	1.0	0.986014	1.0
GPT2-large	0.825174	1.0	1.0
GPT2-xl	1.0	0.941059	0.989011

Table 2: p-values for Dip-Test of three halves results for Natural Stories corpus

Even when looking at the individual halves in Table 2, we see that the distributions remain unimodal. But it should be kept in mind that the length of each half was almost 80 to 90 times that of the average sentence length of the UCL corpus. And so, we investigate if this property of modality of distributions was some kind of statistical artefact caused due to the the length of the Natural Stories corpus.

Thus it appears that the nature of distribution of both human and machine surprisals change with increasing context length. We suspect that this is an effect of working memory (Baddeley, 1992) and integration of words in action. We hope to delve into more details about it in the future.

We further investigate this property of the effect of

Item	d	p-value
human-RT	0.03691	0.000999
GPT2	0.03649	0.000999
GPT2-medium	0.03647	0.000999
GPT2-large	0.03641	0.000999
GPT2-xl	0.03641	0.000999

Table 3: p-values for Dip-Test results for UCL corpus

context length on the distribution of surprisals using an "artificial corpus" where we can easily control the length of sentences and automatically generate a corpus to perform our analysis. In the next section, we describe and then run the machine models on our corpus of "artificial sentences" generated using an open-source generator to repeat the analysis methods that we used in this section.

6.1 Artificial Sentence Corpus

There has been a phenomenal growth in the quality of output of Natural Language Generation systems in the last few years. They have thus emerged as an interesting way to generate test data for the line of research that we are advocating in this paper. To study how the frequency distribution of surprisal corresponding to different GPT2 models differ with differing context sizes, we create an "Artificial Sentence Corpus" using the state-of-the-art ChatGPT[6] system. ChatGPT has lately captured the public imagination on account of its ability to generate "coherent responses to various questions" (Shahriar and Hayawi, 2023). To create the corpus, we prompt ChatGPT to construct sentences of different lengths[7]. In this way, we generate 200 'artificial stories'. The sentences were then fed into all four of the GPT2 models for obtaining the values of surprisal. For the purpose of analysis, we club the sentences into four separate length bins: 0-15 (50 sentences), 15-50 (50 sentences), 50-100 (50 sentences), 100-300 (50 sentences).

For the analysis of the surprisals obtained from the models, we perform a dip-test on the frequency distributions of the 4 length bins for the four different models of the GPT2 family. As mentioned earlier, p-value of less than 0.05 in the test indicates that the distribution is multimodal. And conversely, p-values more than 0.05 indicate that the distribution is unimodal. Figure 5 shows that, in terms of the results from the dip-test, for all models, sentences of different context lengths exhibit different patterns of distributional modality in their surprisals. For all models, the surprisals are multimodal till bin2. But bins 3 and 4 correspond to a p-value way more than 0.05 for the dip-test. In other words, for sentences with more than 50 words, the GPT2 models starts being

[6]https://chat.openai.com

[7]Example prompt: Generate 10 stories each having 110 words

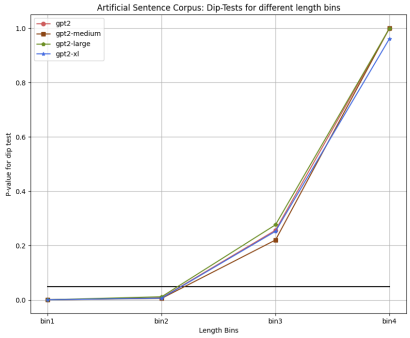

Figure 5: P-values for Dip-Test for different length bins across different models of the GPT2 family

surprised in a specific range about most words and word classes. We find this behavior very intriguing because we saw a similar feature while studying the frequency distribution of human processing effort for the Natural Stories corpus in Table 2.

Model	Bin 1	Bin2 2	Bin 3	Bin 4
GPT2	0.0009	0.0009	0.2498	1.0000
GPT2-m	0.0009	0.0009	0.2418	0.9980
GPT2-l	0.0009	0.0009	0.2428	1.0
GPT2-x	0.0009	0.0009	0.1648	0.9960

Table 4: p-values for Dip-Test of four bins for Artificial Stories corpus

6.2 Surprisal vs PE: A Part-Of-Speech (POS) perspective

Our next step involves analyzing the human processing effort and GPT2 surprisal in terms of parts-of-speech. To do this, we extract the POS tags for every sentence in the two corpora and gather the processing effort and surprisal scores for each sentence, from human observations and machine predictions respectively. After calculating the average scores for both processing effort and surprisal across all categories, we visualize the data in a plot.

Looking at the UCL corpus for human participants (Figure 6), we observe that nouns and verbs (content words) require slightly less processing effort compared to other Part-Of-Speech categories. A similar trend can be seen in the Natural Stories corpus (Figure 7). Previous research on whether function words or content words take longer to process during reading has not yielded a clear consensus. Some studies suggest that function words are processed faster (Staub and Clifton Jr, 2006; Schmauder et al., 2000), while others indicate that content words are processed more quickly (Rayner et al., 1986, 2000). Moreover, most studies

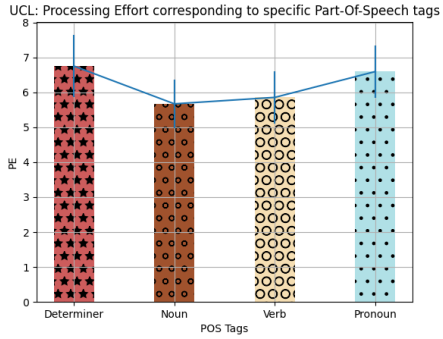

Figure 6: UCL: POS wise average PE for human participants

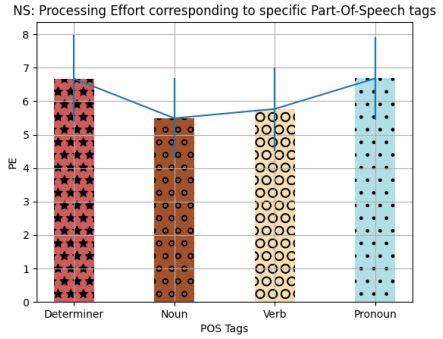

Figure 7: Natural Stories: POS wise average PE for human participants

on reading time rely on fixation times and other physiological data, whereas we used reaction times from self-paced reading experiments for our observations.

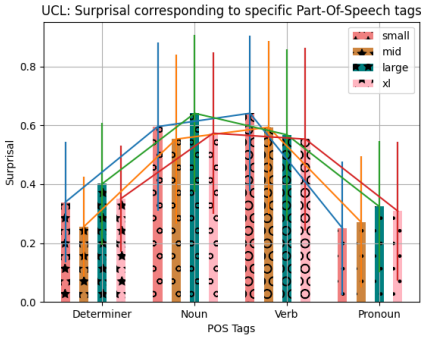

Figure 8: UCL: POS-wise average surprisal for machines

When analyzing the machines' performance (Figures 8 to 10), we notice that categories like Determiners and Pronouns (function words) result in lower surprisal scores than categories like Verbs and Nouns (content words). Additionally, as the context size increases (UCL vs. Natural Stories), the difference in surprisal between the content words and that of function words decrease.

To determine which models are the most similar to

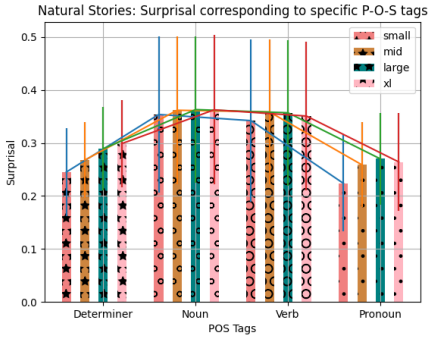

Figure 9: Natural Stories: POS-wise average PE for machines

humans (based purely on POS categories), we compute the Wasserstein distance between humans and machines for both datasets.

Model	Determiners	Noun	Verb	Pronoun
GPT2-s	6.09	5.27	5.50	5.84
GPT2-m	6.01	5.22	5.45	5.87
GPT2-l	6.15	5.31	5.42	5.92
GPT2-x	6.10	5.24	5.40	5.90

Table 5: Wasserstein distance (humans and machines) across POS for UCL corpus

The Wasserstein distance measures how much 'effort' it would take to transform one distribution to another. In this case, lower distances imply greater similarity among surprisal and processing effort for a particular POS category. Tables 5 and 6 shows that the Wasserstein distances were relatively lower for the Natural Stories Corpus than the UCL corpus. But there doesn't seem to be any evidence suggesting that the similarity between patterns of human processing effort and machine surprisal increased with increased parameter sizes.

Model	Determiners	Noun	Verb	Pronoun
GPT2-s	5.92	4.85	5.11	5.91
GPT2-m	5.94	4.85	5.13	5.94
GPT2-l	5.96	4.85	5.13	5.95
GPT2-x	5.97	4.85	5.12	5.95

Table 6: Wasserstein distance (humans and machines) across POS for Natural Stories corpus

Hence, our analysis suggests that the patterns of surprisal and processing effort for Part-Of-Speech tags are vastly different between machines (pretrained GPT2 systems) and humans, regardless of the parameter size of the GPT2 models. However, there is weak evidence that with longer context size, the models exhibit more "human-like" behavior (based on the Wasserstein dis-

tance).

We propose that the discrepancy in behavior for POS tags may be due to the way GPT2 models are trained. Humans do not learn a language by processing terabytes of text data, and this likely results in the representation and processing of word-classes in the GPT-style neural models differing significantly from those in the internal language models of humans. Additionally, as we noted in our previous analysis of the distribution of surprisal and processing effort, the distributions become more unimodal with increasing context length. Thus, the relatively smaller Wasserstein distance for the Natural Stories corpus may be explained by this phenomenon.

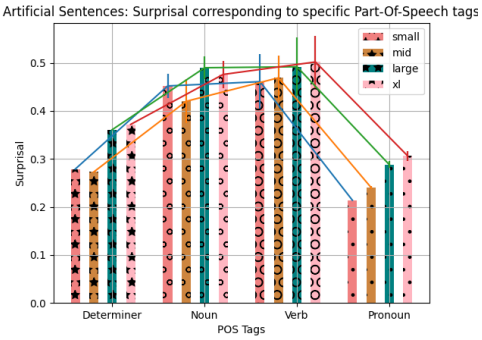

Figure 10: Artificial Sentences: POS wise average surprisal for machines

Sobieszek and Price (2022) explored why the statistical capabilities of GPT3 might allow it to 'play tricks' that make its responses seem more plausible than truthful. The comparatively less surprisal associated with function in this case might be a sign of those tricks and needs further exploration.

6.3 Comparison of complexities

We will now introduce a metric for comparing the "effort" required to process sentences in a corpus by both the GPT2 systems and human participants. This metric is partly inspired by the work of Frank (2010) and is based on a perplexity-like measure. Specifically, we define our metric as the mean of Processing Effort, which itself is a logarithmic function of surprisals. This measures the amount of unexpected information contained in a sentence. Our metric is based on a fundamental principle of information theory and is therefore applicable to measures of machine surprisal (and behaves like perplexity).

For a given sentence, we define sentence complexity as follows:

$$C = \frac{\sum_{i=1}^{n} PE_i}{n} = \frac{\sum_{i=1}^{n} S_i}{n} \quad (4)$$

In Equation (4), 'PE' and 'S' refer to Processing Effort and Surprisal respectively as defined in the previous sections. Mathematically speaking, the summation of the surprisal terms translate to the product of probabilities. In the following sections we use this metric to compare the performance of the GPT2 models (of different parameter sizes) with averages of human performance effort complexities.

For a sentence ψ, To make the comparison between the GPT2 models and the human participants, we observe the difference defined as:

$$D_i = C(PE_\psi) - C(S_\psi) \quad (5)$$

In other words, for a sentence i, Equation (5) yields the difference in complexities of processing it by humans and by machines. Hence, if a sentence was easier (in terms of surprisal values) for humans to read in comparison to the humans, then the term D_ψ would be <0 and vice versa.

Model	Average Difference
GPT2	-0.0950
GPT2-medium	-0.1347
GPT2-large	-0.1063
GPT2-xl	-0.1245

Table 7: Average of difference of complexities for UCL corpus

From Tables 7 and 8, it seems that larger models indeed behave more "human-like" in terms of this metric with large contexts.

Model	Average Difference
GPT2	-0.6357
GPT2-medium	-0.6158
GPT2-large	-0.6157
GPT2-xl	-0.6135

Table 8: Average of difference of complexities for NS corpus

6.4 Do bigger models 'look' like humans?

Finally, we investigate the relationship between reading behavior and processing effort by conducting a joint analysis (as done for example with different data by Reichle et al. (1998)) of Reaction Times and Eye-Tracking data from the UCL corpus. Our goal was to eventually compare these measures with values obtained from machine surprisal, which is a computational measure of information processing difficulty.

Many (if not most) papers exploring similar research questions use eye-tracking metrics for estimating reading time. However we use reaction times as our metric for reading time in this paper. And hence, we first tested

the the Pearson correlation between the Processing effort measured by Reading Time and Fixation Duration. We calculated the Processing effort for a sentence by taking the means of both the reading times and total fixation duration for each word. The results (Table 9) showed a strong positive correlation between Reaction Times and Fixation Times, indicating that these measures are both sensitive indicators of processing effort during reading.

correlation coefficient	p-value
0.5592	$<< 0.05$

Table 9: Result of Pearson correlation between Fixation Duration and Reaction time for sentences in UCL corpus

The scientific literature commonly considers eye-tracking metrics, such as fixation times, as a standard measure for estimating cognitive load. This is often the norm in psycholinguistics since this metric reflects the amount of attention and processing resources required to read a given text (Kliegl et al., 2004). Therefore, we conduct a correlation analysis to examine whether Processing Effort, as measured by both mean Reaction Times and mean Fixation duration, is correlated with machine surprisal.

Model	Correlation Coefficient	P-value
GPT2	0.0910	$<< 0.05$
GPT2-m	0.1597	$<< 0.05$
GPT2-l	0.1521	$<< 0.05$
GPT2-x	0.1721	$<< 0.05$

Table 10: Result of Pearson correlation between Reading Time and Machine Surprisal for sentences in UCL corpus

Model	Correlation Coefficient	P-value
GPT2	0.1053	$<< 0.05$
GPT2-m	0.1920	$<< 0.05$
GPT2-l	0.1871	$<< 0.05$
GPT2-x	0.2170	$<< 0.05$

Table 11: Result of Pearson correlation between Fixation Duration and Machine Surprisal for sentences in UCL corpus

From Tables 10 and 11, we see that as the number of parameters in the models increased, the correlation coefficient statistic also increased. This suggests that surprisals generated from models with greater parameter size have a stronger positive correlation with human physiological data collected in the form of Reading Time and Fixation Duration.

This observation suggests that more complex models can perhaps capture a wider range of cognitive and neural processes, and thus provide a more accurate representation of human reading behavior.

7 Discussion

The development of neural networks was initially inspired by the functioning of human neurons, but practical applications have since driven their engineering. However, language is a uniquely human trait, and it is therefore crucial to investigate whether Language Model (LM) training approaches that do not perfectly mimic human language learning are able to learn the same aspects of language. This paper contributes to this line of inquiry by comparing human and machine surprisals from two self-paced-reading corpora. The study's findings, including the change in frequency distribution of surprisals and POS analysis, suggest that with greater context, both humans and machine language models can better predict upcoming words, resulting in narrow surprisal values. However the nature of representation of syntax in Transformer-based models (in comparison to humans) requires further investigation. Also, the observation that increasing model parameter size leads to better correlation with human processing effort, emphasizes the benefits of scaling up language models and the need for more research into emergent capabilities of larger models.

8 Conclusion

This paper compared human and machine surprisals from two existing self-paced-reading corpora. The estimation of human surprisal was done using reaction times. Analysis showed that the nature of frequency distribution of the surprisals for both humans and pretrained LMs changed with increasing context length. We observe that the distribution starts with being multimodal, but it quickly becomes unimodal with increasing context length. Additionally, we found that for long context lengths, in both GPT2 models and humans, surprisal peaked at a specific range for most words.

We also find that the machine surprisals behave very differently than human surprisals across parts of speech tags. We find that GPT2 models are way less surprised by more "predictable" POS tags like determiners than humans. We also present rudimentary evidence that with increasing context length, GPT2 models (irrespective of size) might be more similar to humans in terms of being surprised by specific word-classes.

Finally, we show that increasing the parameter sizes seems to make models perform more "human-like" for sentence-level metrics. We also find that surprisal from eye-tracking metrics seem to correlate better with GPT2 surprisals than surprisal from reading times.

References

Alan Baddeley. 1992. Working memory. *Science*, 255(5044):556–559.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Bruno Bianchi, Gastón Bengolea Monzón, Luciana Ferrer, Diego Fernández Slezak, Diego E Shalom, and Juan E Kamienkowski. 2020. Human and computer estimations of predictability of words in written language. *Scientific reports*, 10(1):4396.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Forrest Davis and Marten Van Schijndel. 2020. Recurrent neural network language models always learn english-like relative clause attachment. *arXiv preprint arXiv:2005.00165*.

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

Stefan L Frank. 2010. Uncertainty reduction as a measure of cognitive processing effort. In *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, pages 81–89.

Stefan L Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological science*, 22(6):829–834.

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45:1182–1190.

Richard Futrell, Edward Gibson, Hal Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. 2017. The natural stories corpus. *arXiv preprint arXiv:1708.05763*.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9):397–412.

John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. 2018. Finding syntax in human encephalography with beam search. *arXiv preprint arXiv:1806.04127*.

John A Hartigan and Pamela M Hartigan. 1985. The dip test of unimodality. *The annals of Statistics*, pages 70–84.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. 2004. Length, frequency, and predictability effects of words on eye movements in reading. *European journal of cognitive psychology*, 16(1-2):262–284.

Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. 2022. Context limitations make neural language models more human-like. *arXiv preprint arXiv:2205.11463*.

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

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Matthew W Lowder, Wonil Choi, Fernanda Ferreira, and John M Henderson. 2018. Lexical predictability during natural reading: Effects of surprisal and entropy reduction. *Cognitive science*, 42:1166–1183.

James A Michaelov, Megan D Bardolph, Seana Coulson, and Benjamin K Bergen. 2021. Different kinds of cognitive plausibility: why are transformers better than rnns at predicting n400 amplitude? *arXiv preprint arXiv:2107.09648*.

Irene Fernandez Monsalve, Stefan L Frank, and Gabriella Vigliocco. 2012. Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 398–408.

Mante S Nieuwland. 2019. Do 'early'brain responses reveal word form prediction during language comprehension? a critical review. *Neuroscience & Biobehavioral Reviews*, 96:367–400.

Keith Rayner, David A Balota, and Alexander Pollatsek. 1986. Against parafoveal semantic preprocessing during eye fixations in reading. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 40(4):473.

Keith Rayner, Gretchen Kambe, and Susan A Duffy. 2000. The effect of clause wrap-up on eye movements during reading. *The Quarterly Journal of Experimental Psychology: Section A*, 53(4):1061–1080.

Erik D Reichle, Alexander Pollatsek, Donald L Fisher, and Keith Rayner. 1998. Toward a model of eye movement control in reading. *Psychological review*, 105(1):125.

A René Schmauder, Robin K Morris, and David V Poynor. 2000. Lexical processing and text integration of function and content words: Evidence from priming and eye fixations. *Memory & Cognition*, 28:1098–1108.

Steven C Schwering and Maryellen C MacDonald. 2020. Verbal working memory as emergent from language comprehension and production. *Frontiers in human neuroscience*, 14:68.

Sakib Shahriar and Kadhim Hayawi. 2023. Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations. *arXiv preprint arXiv:2302.13817*.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Adam Sobieszek and Tadeusz Price. 2022. Playing games with ais: The limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364.

Olla Solomyak and Alec Marantz. 2009. Lexical access in early stages of visual word processing: A single-trial correlational meg study of heteronym recognition. *Brain and language*, 108(3):191–196.

Adrian Staub. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, 9(8):311–327.

Adrian Staub and Charles Clifton Jr. 2006. Syntactic prediction in language comprehension: evidence from either... or. *Journal of experimental psychology: Learning, memory, and cognition*, 32(2):425.