Surprisal from Larger Transformer-based Language Models Predicts fMRI Data More Poorly

Yi-Chien Lin
The Ohio State University
Department of Linguistics
lin.4434@osu.edu

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

Abstract

There has been considerable interest in using surprisal from Transformer-based language models (LMs) as predictors of human sentence processing difficulty. Recent work has observed an inverse scaling relationship between Transformers' per-word estimated probability and the predictive power of their surprisal estimates on reading times, showing that LMs with more parameters and trained on more data are *less* predictive of human reading times. However, these studies focused on predicting latency-based measures. Tests on brain imaging data have not shown a trend in any direction when using a relatively small set of LMs, leaving open the possibility that the inverse scaling phenomenon is constrained to latency data. This study therefore conducted a more comprehensive evaluation using surprisal estimates from 17 pre-trained LMs across three different LM families on two functional magnetic resonance imaging (fMRI) datasets. Results show that the inverse scaling relationship between models' per-word estimated probability and model fit on both datasets still obtains, resolving the inconclusive results of previous work and indicating that this trend is not specific to latency-based measures.

1 Introduction

Expectation-based theories of syntactic comprehension (Hale, 2001; Levy, 2008) posit that comprehenders evaluate multiple hypotheses of a structure in parallel. Within this line of research, processing difficulty can be quantified using information-theoretic surprisal (Shannon, 1948) and the processing difficulty of a word is proportional to its surprisal estimate (i.e., negative log probability given the preceding context).

As Transformer-based models (Vaswani et al., 2017) become more widely incorporated into natural language processing tasks, there has been considerable interest in using surprisal from these models as predictors of processing difficulty (Merkx and Frank, 2021; Wilcox et al., 2020). Recent work has observed that surprisal from larger Transformer-based LMs, which have more parameters and lower model perplexity, is less predictive of reading times (Oh et al., 2022; Oh and Schuler, 2023; de Varda and Marelli, 2023; Shain et al., 2024; Oh et al., 2024). Larger LMs trained on larger amounts of data become more predictive of rare tokens throughout the course of training, resulting in lower surprisal estimates that diverge from human reading times (Oh et al., 2024). In addition to reading times, Oh et al. (2022) also use surprisal estimates from Transformers to predict brain imaging data, but observe no inverse or positive scaling when using a relatively small set of LMs. However, moving away from surprisal, recent studies observed a positive scaling on brain imaging data when using vectors directly from large language models as predictors (Schrimpf et al., 2021; Hosseini et al., 2024) so it may be tempting to think brain imaging data behaves differently from latency-based data (i.e., reading times).

This study presents evidence against this possibility by evaluating the scaling trend of the predictive power of surprisal estimates from a comprehensive set of 17 Transformer-based LMs across three LM families on two different brain imaging datasets: the original dataset (Shain et al., 2020), which was also evaluated in Oh et al. (2022), and a separate dataset (Pereira et al., 2018). Results show statistically significant inverse scalings on both datasets which resolve the inconclusive results from Oh et al. (2022) on Shain et al. (2020) when using surprisal estimates from a larger set of LMs and replicate inverse scaling on Pereira et al. (2018).

This work provides evidence that the inverse scaling of surprisal estimates is not specific to latency-based data. This piece of evidence provides a more complete picture of the relationship between the model sizes and predictive power of surprisal

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 179–186

estimates on psychometric data, which could potentially be helpful in providing researchers with insights into the appropriate use of LMs in understanding human sentence processing.

2 Related Work

Previous studies have examined the predictive power of surprisal from various types of LMs such as n-gram, Long Short-Term Memory network (Hochreiter and Schmidhuber, 1997), and recurrent neural network models on psychometric data (e.g., Goodkind and Bicknell, 2018; Wilcox et al., 2020). These studies observed a consistent trend in the positive relationship between a model's quality and its predictive power on psychometric data such as reading times: the better an LM's quality (or the higher an LM's per-word estimated probability), the better its fit to the psychometric data.

Recently, there has been increasing interest in comparing surprisal estimates from Transformer-based models against behavioral and neural measures of processing difficulty. Differing from previous findings, Oh et al. (2022) observed a completely opposite *inverse* scaling relationship between LMs' per-word estimated probability and models' fit to reading times when using surprisal estimates from a set of pre-trained GPT-2 models (Radford et al., 2019). In other words, larger LMs (with higher per-word estimated probability) are less predictive of reading times. This observation was further replicated using surprisal estimates from a larger set of Transformer-based LMs (Oh and Schuler, 2023; Oh et al., 2024).

Moreover, recent work in this line of study found that surprisal estimates from neural LMs have a tendency to underpredict reading times. Van Schijndel and Linzen (2021) and Arehalli et al. (2022) showed that neural LM surprisal successfully predicts human precessing difficulty in garden-path constructions but consistently underestimates its magnitude. Kuribayashi et al. (2022) also observed that surprisal estimates from neural LMs without context limitations underpredict reading times of English and Japanese naturalistic text. Furthermore, Oh and Schuler (2023) conducted an analysis of the inverse scaling relationship between Transformer-based LMs' per-word estimated probability and the predictive power of surprisal estimates, observing that the poorer fit to reading times of larger LMs is mainly driven by the lower surprisal values those LMs assign to open-class words such as nouns and adjectives. Oh et al. (2024) followed up on this observation and showed that word frequency explains the inverse correlation between the size of Transformer-based LMs, the amount of training data, and the predictive power of surprisal estimates. They found that larger LMs get better at predicting rare words, resulting in lower surprisal which diverges from human reading times.

3 Experiment 1: Predictive Power of Surprisal on Shain et al. (2020)

Experiment 1 revisited Oh et al. (2022) and evaluated the predictive power of surprisal estimates from a larger set of Transformer-based LMs on Shain et al. (2020) (hereafter '**Natural Stories fMRI**'). We collected surprisal estimates from 17 pre-trained autoregressive Transformer-based LMs and used them to predict the blood oxygenation level-dependent (BOLD) signals from Natural Stories fMRI. The experiment setup followed that of Shain et al. (2020).[1]

3.1 Response Data

Natural Stories fMRI (Shain et al., 2020) was collected at a two-second fixed time interval from 78 subjects while they listen to a recording of the Natural Stories Corpus (Futrell et al., 2021) which consists of naturalistic English stories. This dataset includes the time series of BOLD signals of several functional regions of interest (fROIs) in the language network (Fedorenko et al., 2011), which were identified with a localizer task. For each fROI, the BOLD signals were averaged across all voxels within that fROI.

Following Shain et al. (2020), Natural Stories fMRI was partitioned into fit, exploratory, and held-out sets which contain around 50%, 25%, and 25% of the data points, respectively. Specifically, we grouped the BOLD signals into 30-second chunks and distributed those chunks into the sets to ensure the time series data in the sets are sequential and to avoid having data which is temporally too close to each other in each set. This resulted in 100,084, 50,818, and 51,393 data points in the fit, exploratory, and held-out sets. The regression models were fit using the fit set and the results reported in this study are on the held-out set.

[1] Code for replicating the experiments in this study is available at https://github.com/modelblocks/modelblocks-release.

3.2 Predictors

We collected surprisal estimates from model variants of three autoregressive Transformer-based LM families which were used in Oh et al. (2022) and Oh and Schuler (2023): GPT-2 (Radford et al., 2019), GPT-Neo (Black et al., 2021, 2022; Wang and Komatsuzaki, 2021), and OPT (Zhang et al., 2022)[2] families. The hyperparameters of each LM are detailed in Table 1, Appendix A.

Each story of the Natural Stories Corpus was first tokenized using the byte-pair encoding (Sennrich et al., 2016) tokenizer corresponding to each model variant. The tokenized texts were then input to each LM to calculate token-wise surprisal estimates. In cases where a tokenized story exceeds a single context window of the LMs, to calculate the surprisal estimates for the remaining tokens, the latter half of the previous context window was used as the first half of the context window. If a word consists of multiple subword tokens, we aggregate the token-level surprisal estimates to get the word-level surprisal estimate.[3]

Recent studies (Oh and Schuler, 2024; Pimentel and Meister, 2024) have shown that, for languages which use whitespace characters in their orthography, the leading whitespaces (i.e., the whitespace characters which immediately precede the tokens) added by LMs' tokenizers can result in a sum over all word probabilities that is greater than one. This study therefore incorporated whitespace-trailing decoding into surprisal calculation. For each word, the probability of the leading whitespace was reallocated to its preceding word when calculating word-level probabilities.

Since the response data of Natural Stories fMRI is the time series of BOLD signals, we first collected the surprisal estimates of the words and then applied a hemodynamic response function (HRF; Boynton et al. 1996) to convolve those surprisal estimates so that the convolved surprisal estimates align with the response data. A canonical HRF convolves predictors into the shape of the BOLD signals when a subject is presented with a stimulus. The HRF we used for this experiment is shown in the following equation, giving magnitude $f(x)$ for offset time x in seconds:

$$f(x) = \frac{5.2^{5.4} x^{5.4-1} e^{-5.2x}}{\Gamma(5.4)} - 0.35 \frac{7.35^{10.8} x^{10.8-1} e^{-7.35x}}{\Gamma(10.8)}$$

[2]Due to the constraints in computational resources available to us, we did not include the largest OPT model variant, which has around 175 billion parameters.

[3]Although we tokenized the text at token-level instead of word-level, the relationship between the predictive power of the surprisal and the LM quality is unlikely to make much difference, as Oh and Schuler (2025) found that there is a relatively small effect of token granularity on the relationship.

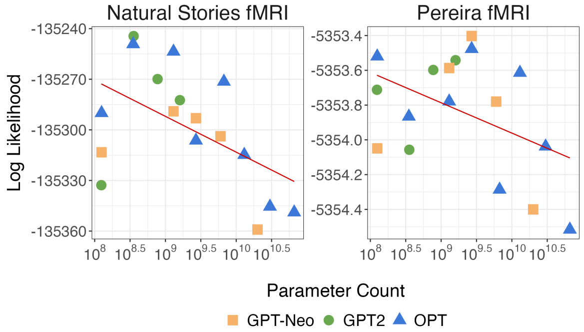

Figure 1: Predictive power of surprisal estimates from pretrained LMs on BOLD signals of Natural Stories fMRI (left) and Pereira fMRI (right).

3.3 Regression Modeling

Once the surprisal estimates were collected, we fit a set of linear mixed-effect (LME) models to BOLD signals with `lme4` (Bates et al., 2015) using the fit set. During regression modeling, in addition to surprisal estimates, we also included a set of baseline predictors, including the index of the word position in the sentence and the word length in characters. All models were fit using raw BOLD values along with the above baseline predictors. The predictors were centered, scaled, and convolved with the HRF before being used to fit the LME models. Following Barr et al. (2013), during model fitting, we started with the maximal LME models and eliminated one random effect at a time until the LME models converge. The maximal converging formula for Natural Stories fMRI was:

```
BOLD ~ z.(sentpos) + z.(wlen) + z.(surp) +
       (1 + z.(wlen) + z.(surp) | subject) +
       (1 + z.(surp) | storyid)
```

where `sentpos` is the word position in the sentence, `wlen` is word length, `surp` is surprisal estimate, and `storyid` is the index of a story.[4] We show the predictive power with the model log likelihood calculated on the held-out set.

3.4 Results

Figure 1 (left) presents the relationship between the regression model log likelihood and log-transformed parameter count for Natural Stories

[4]The original design of the LME model for Natural Stories fMRI is shown in Appendix B.

fMRI. The results show a significant decrease in predictive power when parameter count (model size) increases ($p < 0.01$ by a permutation test with 1,000 permutations), resolving the inconclusive results in Oh et al. (2022), which only evaluate the predictive power of surprisal estimates from a relatively small set of LMs on this same dataset.[5]

4 Experiment 2: Predictive Power of Surprisal on Pereira et al. (2018)

Experiment 1 resolves the inconclusive results on Natural Stories fMRI from Oh et al. (2022) by showing an inverse scaling phenomenon with a larger set of LMs. To examine whether this phenomenon is generalizable across other brain imaging datasets, we additionally evaluated the predictive power of the same set of LMs on a different fMRI dataset, Pereira et al. (2018) (hereafter '**Pereira fMRI**').

4.1 Response Data

Pereira fMRI contains data from Experiments II and III in Pereira et al. (2018). Eight subjects participated in Experiment II and read 96 English passages, which includes 384 sentences in total; six subjects participated in Experiment III and read 72 English passages, which consists of 243 sentences in total. Each passage consists of three to four sentences and sentences were presented to subjects one at a time. Like Natural Stories fMRI, Pereira fMRI also includes BOLD signals from the voxels in the language network. The responses of this dataset are per-sentence BOLD signals, which were obtained by aggregating activation across regions. Following Oh et al. (2022), Pereira fMRI was partitioned into fit, exploratory, and held-out sets, resulting in the sizes of 2,268, 1,132, and 1,130 data points respectively in each set.

4.2 Procedure

In general, the procedure of Experiment 2 followed that of Experiment 1 except that since the response data for Pereira fMRI is per-sentence BOLD signals, we followed recent work on the same dataset (Hosseini et al., 2024) and used the surprisal estimate of the sentence-final word as the predictor for each response data point.[6] The motivation of using the sentence-final word surprisal is that its preceding context has already been taken into consideration when calculating the surprisal estimates of the last words of the sentences.

When fitting the LME models, we used sentence length in words and index of the passage containing the sentence as baseline predictors. Similar to the procedure in Experiment 1, following Barr et al. (2013), we started with the maximal LME models and eliminated the random effects one by one until the models no longer have serious convergence issues. The maximal converging formula was:

```
BOLD ~ z.(sentlen) + z.(passpos) +
    z.(lastwordsurp) + (1 + z.(sentlen) +
    z.(passpos) + z.(lastwordsurp) | subject) +
    (1 + z.(sentlen) + z.(lastwordsurp) | sentid)
```

where `sentlen` is sentence length in words, `passpos` is the sentence position within passage, `lastwordsurp` is last-word surprisal estimate, and `sentid` is the index of each sentence.[7]

4.3 Results

Figure 1 (right) shows the results of Experiment 2. Similar to the results of Experiment 1, the results on Pereira fMRI also show that as parameter count (model size) increases, the regression model log likelihood decreases significantly ($p < 0.05$). The replication of the inverse scaling phenomenon on a different dataset suggests that this phenomenon is not dataset-specific.[8]

[5]The choice of model families was based on previous work which observed inverse scaling on reading times (Oh et al., 2022; Oh and Schuler, 2023). Based on the inverse scaling trend observed for Pythia and the similarity to inverse scaling results for larger off-the-shelf LLMs found in Oh and Schuler (2023); Oh et al. (2024), we expect more recent, larger models to yield substantially worse fits.

[6]Hosseini et al. (2024) use the model representations of the sentence-final words and not the surprisal estimates as the representation of the entire sentence.

[7]The original design of the LME model for Pereira fMRI is shown in Appendix B.

[8]For both corpora, in addition to running experiments using LME models, per reviewers' request, we also ran linear regression for both datasets (1) using one data point per sentence and (2) averaging across participants.

For Pereira fMRI, since the response data is already per-sentence BOLD signals, for each sentence, we averaged across the BOLD signals across all participants. Similar to our method described in Experiment 2, which was based on the methodology in described Hosseini et al. (2024), we used the last-word surprisal estimate as the representation of each data point. We then performed linear regression modeling using six-fold cross-validation. The test fold sizes range from 87 to 122 data points and the training fold sizes range from 505 to 540 data points.

For Natural Stories fMRI, since the response data is the time series of BOLD signals, we do not have direct information about sentence segmentation. In order to obtain an estimate of the sentence-level BOLD signals and surprisal estimates, in general, we followed Feghhi et al. (2024) which used an fMRI dataset collected at a two-second time interval (Blank et al., 2014), similar to Natural Stories fMRI. We mapped each response data point to the word (along with its surprisal

5 Conclusion

Recent studies (Oh et al., 2022; Oh and Schuler, 2023; Oh et al., 2024) have observed a robust negative relationship between Transformer-based LMs' quality and the predictive power of surprisal estimates from those LMs on latency data. This work generalizes that result by examining the predictive power of surprisal estimates from Transformer-based models on brain imaging data. The predictive power of surprisal estimates from 17 pre-trained LMs across three LM families was evaluated on two fMRI datasets. Results show that the inverse scaling between model size and model fit still obtains on brain imaging data. The inverse relationship between the predictive power of surprisal estimates from those LMs and parameter count (model size) on brain imaging data is not as strict as that on latency-based data, but this may be due to the fact that brain imaging data is much noisier than latency-based data. Nevertheless, the observation of the inverse scaling phenomenon on two distinct brain imaging datasets indicates that this trend is not specific to latency-based measures, suggesting that brain imaging data do not behave substantially differently from latency data. We speculate the deviation of larger LMs from brain imaging data is due to the same reason for reading times (Oh and Schuler, 2023). This additional piece of evidence reinforces the observation of inverse scaling phenomenon on latency data (Oh et al., 2022; Oh and Schuler, 2023), suggesting that smaller models could potentially be helpful in providing researchers with insights into the appropriate use of LMs in understanding sentence processing.

estimate) that is closest to four seconds before that data point. Note that for this analysis, we did not apply HRF convolution to the predictors. This allowed us to obtain an estimate of sentence segmentation points since we have the sentence id information of each word. To get the sentence-level BOLD signals, we averaged all BOLD signals that were mapped to the same sentence id. As for sentence-level surprisal estimates, it was not possible to use last-word surprisal since the BOLD signals are not sentence-aligned, so for each sentence, we averaged all surprisal estimates that were within that sentence and were mapped to some BOLD signals.

Once we obtained the sentence-level BOLD signals for each participant and surprisal estimates, we averaged across participants' responses. Since the number of data points were reduced to a small amount after obtaining the sentence-level responses/surprisal estimates and averaging across subjects, we performed linear regression modeling using eight-fold cross-validation. For each iteration, we used the data of one story as the test data and the remaining seven stories as the training data. The test fold sizes range from 210 to 390 data points and training fold sizes range from 1992 to 2172 data points.

For both corpora in this analysis, we report the mean squared errors averaged across each iteration. In this analyses, Natural Stories fMRI still shows a significant inverse scaling ($p < 0.05$) while Pereira fMRI shows numerically but not statistically significant inverse scaling. We speculate that the reason why statistical significance was not observed for inverse scaling on Pereira fMRI under this experimental setting might be because the BOLD signals used for regression modeling were averaged across subjects, which could potentially make the data somewhat sparser and therefore underpowered.

Limitations

This study attempts to examine the predictive power of surprisal estimates from Transformer-based models on fMRI data. Large language models evaluated in this work are trained on English text and the datasets evaluated are collected from English speakers. Therefore, these findings may or may not be replicated in other languages.

Ethics Statement

The fMRI datasets used in this study are from published work (Shain et al., 2020; Pereira et al., 2018). Details regarding the data collection, validation and other relevant procedures are described in respective publications. Since this work focuses on evaluating the predictive power of large language models on fMRI data, the potential risks and harmful impacts posed by this study on the society appear to be minimal.

Acknowledgments

We thank the anonymous reviewers for their helpful feedback. This work was supported by the National Science Foundation grant #2313140. 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 (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dale J. Barr, Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

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

Sidney 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, virtual+Dublin. Association for Computational Linguistics.

Idan Blank, Nancy Kanwisher, and Evelina Fedorenko. 2014. A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations. *Journal of neurophysiology*, 112(5):1105–1118.

Geoffrey M. Boynton, Stephen A. Engel, Gary H. Glover, and David J. Heeger. 1996. Linear systems analysis of functional magnetic resonance imaging in human v1. *Journal of Neuroscience*, 16(13):4207–4221.

Andrea de Varda and Marco Marelli. 2023. Scaling in cognitive modelling: a multilingual approach to human reading times. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 139–149, Toronto, Canada. Association for Computational Linguistics.

Evelina Fedorenko, Michael K. Behr, and Nancy Kanwisher. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433.

Ebrahim Feghhi, Nima Hadidi, Bryan Song, Idan A Blank, and Jonathan C Kao. 2024. What are large language models mapping to in the brain? a case against over-reliance on brain scores. *arXiv preprint arXiv:2406.01538*.

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelina Fedorenko. 2021. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resources and Evaluation*, 55:63–77.

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 (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Eghbal A. Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

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*, pages 10421–10436, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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.

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.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger Transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Byung-Doh Oh and William Schuler. 2024. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3464–3472, Miami, Florida, USA. Association for Computational Linguistics.

Byung-Doh Oh and William Schuler. 2025. The impact of token granularity on the predictive power of language model surprisal. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.

Byung-Doh Oh, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2644–2663, St. Julian's, Malta. Association for Computational Linguistics.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9(1):963.

Tiago Pimentel and Clara Meister. 2024. How to compute the probability of a word. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18358–18375, Miami, Florida, USA. Association for Computational Linguistics.

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

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):e2105646118.

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 (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Levy. 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, 121(10):e2307876121.

Claude Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656.

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.

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. Curran Associates, Inc.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

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. *Preprint*, arXiv:2205.01068.

A Hyperparameters of Models Examined in this Study

Model Variant	#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
GPT2- XL	48	25	1600	~1.6B
GPT-Neo 125M	12	12	768	~125M
GPT-Neo 1.3B	24	16	2048	~1.3B
GPT-Neo 2.7B	32	20	2560	~2.7B
GPT-J 6B	28	16	4096	~6B
GPT-NeoX 20B	44	64	6144	~20B
OPT 125M	12	12	768	~125M
OPT 350M	24	16	1024	~350M
OPT 1.3B	24	32	2048	~1.3B
OPT 2.7B	32	32	2560	~2.7B
OPT 6.7B	32	32	4096	~6.7B
OPT 13B	40	40	5120	~13B
OPT 30B	48	56	7168	~30B
OPT 66B	64	72	9216	~66B

Table 1: Hyperparameters of the models examined in this study. #L refers to the number of layers of that model; #H refers to the number of attention heads each layer has; d_{model} refers to the embedding size of the model.

B Maximal LME models

During LME model fitting, we followed Barr et al. (2013) and started with the maximal LME models. We eliminated one random effect at a time until the LME models converge. To ensure that the regression includes both by-subject and by-item random effects, we removed random effects iteratively from the by-subject and by-item effects. That is, if the LME model does not converge, we first removed a random effect from the by-item group; if the LME model still does not converge, we then move on to removing a random effect from the by-subject group, and so on.

Following is the original design of the LME model for Natural Stories fMRI, in which the eliminated random effects are bolded.

```
BOLD ~ z.(sentpos) + z.(wlen) + z.(surp) +
    (1 + z.(sentpos) + z.(wlen) +
    z.(surp) | subject) + (1 + z.(sentpos) +
    z.(wlen) + z.(surp) | storyid)
```

As for Pereira fMRI, the original design of the LME model is the following, with the eliminated random effect bolded:

```
BOLD ~ z.(sentlen) + z.(passpos) +
    z.(lastwordsurp) + (1 + z.(sentlen) +
    z.(passpos) + z.(lastwordsurp) | subject) +
    (1 + z.(sentlen) + z.(passpos) +
    z.(lastwordsurp) | sentid)
```