Are Language Models Sensitive to Semantic Attraction? A Study on Surprisal

Yan Cong
Purdue University; Feinstein Institutes
yancong222@gmail.com

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Yu-Yin Hsu
The Hong Kong Polytechnic University
yu-yin.hsu@polyu.edu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Abstract

In psycholinguistics, *semantic attraction* is a sentence processing phenomenon in which a given argument violates the selectional requirements of a verb, but this violation is not perceived by comprehenders due to its attraction to another noun in the same sentence, which is syntactically unrelated but semantically sound.

In our study, we use autoregressive language models to compute the sentence-level and the target phrase-level Surprisal scores of a psycholinguistic dataset on semantic attraction.

Our results show that the models are sensitive to semantic attraction, leading to reduced Surprisal scores, although none of them perfectly matches the human behavioral patterns.

1 Introduction

Cases of similarity-based interference have always been at the center of interest for sentence processing studies, as they offer strong evidence for cue-based models of memory retrieval during language comprehension (Cunnings and Sturt, 2018). According to such accounts, interference emerges because an item with some cues has to be retrieved from memory, and because those cues are simultaneously matched by multiple items (Van Dyke, 2007; Lewis and Vasishth, 2013).

Consider the examples in (1) (Wagers et al., 2009):

(1) a. The key to the **cells** unsurprisingly were rusty.
b. The key to the **cell** unsurprisingly were rusty.

Compared to fully grammatical sentences, both elicit longer reading times in humans, but the effect is attenuated in 1a., where there is an **attractor** (*cells*) matching the number of the verb, causing an **illusion of grammaticality**. This phenomenon is known as **morphological attraction**.

Attraction has also been observed at the semantic level, as in the following example from the eye-tracking study by Cunnings and Sturt (2018):

(2) a. Julia saw the beer that the lady with the **meal** quite happily ate during an expensive night out.
b. Julia saw the beer that the lady with the **wine** quite happily ate during an expensive night out.

Again, both sentences are implausible, because *beer* violates the selectional restrictions of the verb *ate*, but the authors of the study observed that (2a) was processed faster than (2b), due to the presence of a semantically fitting noun (*meal*) that generates a **semantic illusion**. Both types of illusion are *facilitatory* interferences, as they attenuate the effects of anomalies leading to higher costs for the human language processing system. This is a case of **semantic attraction**.

The recent literature in Natural Language Processing (NLP), on the other hand, has shown an increasing interest in using the **Surprisal** scores (Hale, 2001; Levy, 2008) computed with Neural Language Models (NLMs) to account for sentence processing phenomena (Futrell et al., 2018; Van Schijndel and Linzen, 2018; Wilcox et al., 2018; Michaelov and Bergen, 2020, 2022a; Michaelov et al., 2023). This also includes investigations on interferences at the morphosyntactic level (Ryu and Lewis, 2021). To our knowledge, there have been no attempts to model semantic attraction with NLMs yet.

We aim at filling this gap by presenting a Surprisal-based analysis of a psycholinguistic dataset on semantic attraction with three autoregressive NLMs of different sizes. We found that NLMs are sensitive to both the plausibility of the sentences and semantic attraction effects. However, NLM Surprisal for a target phrase seems to be affected by attraction regardless of general sentence

*Proceedings of the The 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 141–148
July 13-14, 2023 ©2023 Association for Computational Linguistics

plausibility, differently from human reading behavior. On the other hand, sentence-level Surprisal is not affected by semantic attraction.

2 Related Work

2.1 Semantic Attraction in Implausible Sentences

The work by Cunnings and Sturt (2018) has recently brought evidence of the existence of semantic attraction in semantically implausible sentences. They collected eye-tracking fixations for sentences in four conditions, by crossing the factors of the plausibility of the sentence (the plausible or implausible arguments are in italic) and the plausibility of an attractor noun (in bold):

(3) a. Julia saw the *cake* that the lady with the **meal** quite happily ate during an expensive night out. (*plausible sentence*, *plausible attractor*)

b. Julia saw the *cake* that the lady with the **wine** quite happily ate during an expensive night out. (*plausible sentence*, *implausible attractor*)

c. Julia saw the *beer* that the lady with the **meal** quite happily ate during an expensive night out. (*implausible sentence*, *plausible attractor*)

d. Julia saw the *beer* that the lady with the **wine** quite happily ate during an expensive night out. (*implausible sentence*, *implausible attractor*)

The results showed that fixations were significantly longer in implausible sentences, but the effect was attenuated in presence of a plausible attractor (condition (3c)), while in plausible sentences the attractor did not have any significant effect. The authors explained the finding in terms of "verb-specific cues that may guide retrieval to grammatically illicit, but plausible, constituents during the resolution of filler-gap dependencies".

The follow-up study by Laurinavichyute and von der Malsburg (2022) instead used a forced choice completion judgement task to compare semantic and morphosyntactic attraction. First, they presented a target verb to the participants, and then they presented them with a sentence fragment, asking participants whether the verb could have been a fitting continuation for the sentence. In such a scenario, it is expected that violations will elicit negative answers, with attraction phenomena possibly increasing the error rates of the participants. Their stimuli contained violations either at the morphosyntactic or at the semantic level, and have either a morphosyntactic or a semantic attractor. The authors reported considerably higher error rates for the conditions with a violation and an attractor of the same type, supporting the idea that morphosyntactic and semantic attraction work similarly.

Our study on NLMs uses the stimuli from the datasets by Cunnings and Sturt (2018) to test whether they are sensitive to semantic attraction in sentence processing, which may be reflected by the Surprisal scores of the stimuli words. We also want to test whether semantic plausibility and attraction in NLMs interact like in humans, to what extent (cf. the claim in Cunnings and Sturt (2018) that semantic attraction has a facilitatory effect only when the sentence is not plausible) and if the effects are the same in NLMs of different sizes.

2.2 NLM Estimation of Word Surprisal

Transformer-based NLMs (Vaswani et al., 2017; Devlin et al., 2019; Radford et al., 2019) have become increasingly popular in NLP in recent years, and a number of studies designed tests to investigate their actual linguistic abilities (Tenney et al., 2019a; Jawahar et al., 2019; Tenney et al., 2019b). Some of these studies specifically analyzed the **Surprisal** scores computed by the models, to understand to what extent they are sensitive to linguistic phenomena that have been showed to affect human sentence processing. For example, Misra et al. (2020) investigated the predictions of BERT in a setting aimed at reproducing human semantic priming; they reported that BERT was indeed sensitive to "priming" and predicted a word with lower Surprisal values when the context included a related word as opposed to an unrelated one.Using a similar methodology, Cho et al. (2021) modeled the priming effect of verb aspect on the prediction of typical event locations, finding that BERT outputs lower surprisal scores for typical locations, but differently from humans, it does so regardless of verb aspect manipulations.

Michaelov and Bergen (2022a) investigated the issue of collateral facilitation, that is, when anomalous words in a sentence are processed more easily by humans because of the presence of semantically-related words in the context. They compared the Surprisal scores obtained with several Transformer NLMs and showed that most of them reproduce the same significant differences between conditions observed in humans. In Michaelov et al. (2023) the same authors used NLM Surprisal to replicate the effect of discourse context in reducing the N400 amplitude for anomalous words, using the Dutch stimuli of the experiments by Nieuwland and Van Berkum (2006).

Probably the closest relative to the topic of our study, Ryu and Lewis (2021) proved that the Surprisal values extracted with the GPT-2 language model predict the facilitatory effects of interference in ungrammatical sentences in which an attractor noun is matching in number with the verb or with a reflexive pronouns. However, they focused on morphosyntactic attraction, while we aim at modeling the facilitatory effects of semantic attraction.

3 Experimental Settings

3.1 Dataset

We derived our dataset from the Experiment 1 of the eye-tracking study by Cunnings and Sturt (2018). The authors employed a total of 32 items, each of them coming in four conditions, for a total of 128 stimuli. The stimuli were stories composed of an introduction sentence, a critical sentence and a wrap-up sentence. In our experiment, we just fed the NLMs with the critical sentence:

(4) Julia saw the *cake/beer* (plausible/implausible) that the lady with the **meal/wine** (plausible/implausible) quite happily ate during an expensive night out.

The sentences in the four conditions, as shown in Example (4), were differing for i) a fitting or a selectional preference-violating direct object (in *italic*) for the verb in the subordinate clause (underlined), which would determine the plausibility of the sentence; ii) a plausible or an implausible attractor noun (in **bold**), not syntactically related with the verb but with a high degree of thematic fit with it.[1] The authors reported main effects of both sentence plausibility (implausible sentences induce longer fixations) and attractor plausibility (a plausible attractor has a facilitatory effect) in the total viewing times.[2] They also reported a significant interaction between the two: total viewing times for implausible sentences were shorter when the attractor was plausible compared to implausible, while no significant difference was observed in plausible sentences as a result of attractor plausibility.

3.2 Language Models

For the models in this paper, we use the implementation of Minicons (Misra, 2022)[3], an open source library that provides a standard API for behavioral and representational analyses of NLMs. We make the code and the test data available for additional testing.[4] We experiment with three variants of autoregressive LMs of different sizes: the original GPT-2 Base, with 124 million parameters (Radford et al., 2019); DistilGPT-2 with 82 million parameters (Sanh et al., 2019), trained as a student network with the supervision of GPT-2; and GPT-Neo that, with 1.3 billion parameters (Gao et al., 2020; Black et al., 2021), is close to the size of the smallest models of the GPT-3 family.

Using autoregressive NLMs, we computed the Surprisal scores at the target in the stimuli (the verb in the subordinate clause), and also at the level of the entire sentence. When the NLMs tokenizer splits the target in more than one token, we take the average of the Surprisal scores of its subtokens.

More formally, the Surprisal of the target T in the context C (**Surp**) was computed as:

$$Surp(T|C) = \frac{\sum_{t \in T} -logP(t|C)}{count(t)} \tag{1}$$

where $P(t|C)$ is the probability of each subtoken $t \in T$ given the previous context C, while $count(t)$ is the number of subtokens in the target phrase T.

The Surprisal of the sentence S (**SentSurp**) instead is simply the sum of the Surprisals of each token T normalized by the length of the sentence:

$$SentSurp(S) = \frac{\sum_{T \in S} Surp(T)}{count(T)} \tag{2}$$

where $count(T)$ is the total number of tokens in the sentence S.[5]

[1] We refer to the notion of *thematic fit* as the degree of compatibility between a predicate and a noun filling one of its semantic roles (McRae and Matsuki, 2009; Sayeed et al., 2016; Santus et al., 2017).

[2] To address a remark by Reviewer 1, we checked the logarithmic frequencies of the attractor nouns (the target nouns were the same in all conditions), which were not mentioned in the original study (see the materials in the Appendix). We have not found any significant difference between noun frequencies across conditions.

[3] https://github.com/kanishkamisra/minicons-experiments

[4] https://github.com/yancong222/transformers-semantic-attraction-surprisal

[5] Notice that the sentences may differ in the number of tokens, in the cases when the object and/or the attractor nouns are splitted by the tokenizer. This is why we did not use the

	GPT-2			DistilGPT-2			GPTNeo		
	B	SE	*p*	*B*	SE	*p*	*B*	SE	*p*
Intercept	9.72	0.54	<.001	9.62	0.54	<0.001	9.92	0.49	<0.001
SentPlaus	3.40	0.28	<0.001	2.17	0.28	<0.001	4.39	0.31	<0.001
AttrPlaus	0.84	0.28	0.003	1.01	0.28	<0.001	0.84	0.31	.008
Length	0.13	0.20	0.08	0.22	0.19	0.09	0.05	0.18	0.78
SentPlaus:AttrPlaus			0.29			0.19			0.11
S1-A0 : S0-A0	-3.69	0.39	< 0.001	-2.45	0.31	< 0.001	-4.87	0.43	< 0.001
S0-A1 : S0-A0	-1.12	0.39	0.021	-1.38	0.31	< 0.001	-1.32	0.43	0.013
S1-A1 : S0-A0	-4.24	0.39	< 0.001	-3.26	0.31	< 0.001	-5.22	0.43	< 0.001
S0-A1 : S1-A0	2.56	0.39	< 0.001	1.07	0.31	0.003	3.55	0.43	< 0.001
S1-A1 : S1-A0	-0.56	0.39	0.48	-0.82	0.31	0.039	-0.36	0.43	0.84
S1-A1 : S0-A1	-3.12	0.39	< .001	-1.89	0.31	< 0.001	-3.91	0.43	< 0.001

Table 1: Summary for the results of predictors of **Surp**, and of the interaction between *SentPlaus* and *AttrPlaus*. In the pairwise comparisons *cond1:cond2*, the reference level is *cond2* (meaning, if the estimate *B* is negative, the Surprisal of *cond1* is lower than *Cond2*, otherwise it is higher).

	GPT-2			DistilGPT-2			GPTNeo		
	B	SE	*p*	*B*	SE	*p*	*B*	SE	*p*
Intercept	7.20	0.53	<0.001	7.84	0.56	<0.001	7.94	0.51	<0.001
SentPlaus	0.10	0.02	<0.001	0.06	0.02	0.011	0.16	0.02	<0.001
AttrPlaus	0.02	0.02	0.382	0.03	0.02	0.06	0.01	0.02	0.829
Length	-0.07	0.01	<.001	-0.08	0.02	<0.001	-0.10	0.01	<0.001
SentPlaus:AttrPlaus			0.33			0.038			0.84
S1-A0 : S0-A0	-0.10	0.03	< 0.001	-0.06	0.03	0.078	-0.172	0.03	< 0.001
S0-A1 : S0-A0	-0.01	0.03	0.96	-0.01	0.03	0.967	-0.02	0.03	0.89
S1-A1 : S0-A0	-0.12	0.03	< 0.001	-0.11	0.03	< 0.001	-0.17	0.03	< 0.001
S0-A1 : S1-A0	0.09	0.03	0.003	0.05	0.03	0.214	0.15	0.03	< 0.001
S1-A1 : S1-A0	-0.02	0.03	0.77	-0.05	0.03	0.249	0.01	0.03	0.992
S1-A1 : S0-A1	-0.11	0.03	< 0.001	-0.10	0.03	< 0.001	-0.15	0.03	< 0.001

Table 2: Summary for the results of predictors of **SentSurp**, and of the interaction between *SentPlaus* and *AttrPlaus*. In the pairwise comparisons *cond1:cond2*, the reference level is *cond2* (meaning, if the estimate *B* is negative, the Surprisal of *cond1* is lower than *Cond2*, otherwise it is higher).

For each NLM, we fitted a linear mixed-effects model using **Surp** or **SentSurp** as the dependent variable, which was estimated for each of the experimental stimuli. The independent variables were: the plausibility of the sentence *SentPlaus* (plausible vs. implausible; plausible as the base of comparison), the plausibility of the attractor *AttrPlaus* (plausible vs. implausible; plausible as the base of comparison), their interactions, and the token length of the stimulus *length*. We included items as a random intercept in our models. We use the LME4 package (Bates et al., 2014) for model fitting and results; the pairwise comparisons with Tukey adjustment were carried out by the EMMEANS package (Lenth, 2019) in R.

4 Results

The findings of the experiments are summarized in Tables 1 and 2.

Considering the main effects, we found that all models were able to distinguish plausible from implausible items at the sentence level (see SentPlaus in Tables 1 and 2), with significantly higher Surprisal scores for the latter.

As shown in Table 1, the models based on **Surp** were also sensitive to the attractor plausibility, and marginally to the token length of the stimuli. No significant main effect of interaction between sentence and attractor plausibility was found. The models based on **SentSurp** (Table 2) were sensitive to token length, but not to the attractor plausibility, with the only exception of a marginal significance for DistilGPT2. The SentSurp model based on DistilGPT2 is the only one showing (at least marginally) significant effects for the plausibility of both sentence (p = 0.011) and attractor (p = 0.06) and for their interaction (p = 0.038) (see Table 2 and Figure 1), while no interaction was found in any of the other models. The fact that this behavior was found in the smallest model may represent another case of what has been called “inverse scaling” in the NLM literature, that is, the performance decreases at the increase of model size (Wei et al., 2022; Jang et al., 2023), or in the

sum of the Surprisal scores, as per Reviewer 3’s comment.

case of psycholinguistic modeling, the behavior becomes less human-like (Michaelov and Bergen, 2022b; Oh and Schuler, 2022).

The post hoc analyses of the pairwise comparison showed some interesting contrasts. We noticed that significant differences were found between the plausible sentences with plausible attractors and the two implausible conditions (i.e. in Figure 1, a vs. c and a vs. d, with $ps < 0.001$). Differently from human total viewing times, no significant differences and no consistent facilitatory effects are observed between c and d in the **SentSurp** models (notice also in Figure 1 that the median of condition d. is actually slightly lower than c., and the medians for c. and d. tend to be close in all the **SentSurp** models, cf. the boxplots in the Appendix, right column), while facilitation is found for all the **Surp** models.

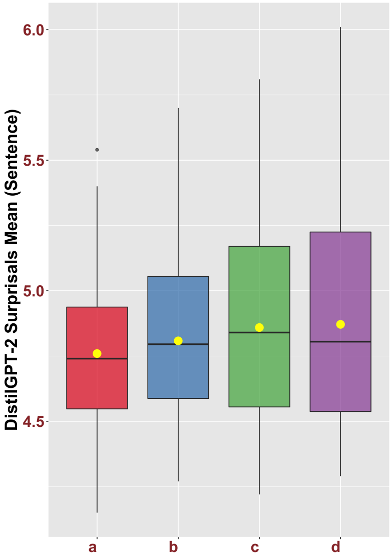

Figure 1: Sentence Surprisal scores from DistilGPT-2 (means in yellow). Conditions are the same of Ex. 3.

It is also noticeable that all models show no sensitivity to plausible attractors with the sentence-level Surprisal metrics, but the Surprisal at the target word with implausible attractors is always significantly higher. However, since no significant main effect of interactions was found for **Surp** models, we conclude that semantic attraction seems to to have a general facilitation effect on its own, regardless of sentence plausibility.

It would be interesting, in the future, to analyze how the attractors concretely affect the predictions, for example using techniques like contrastive explanations (Yin and Neubig, 2022) that can shed light on which tokens contribute to the prediction of the target verb rather than a plausible alternative word (in our case, this could be a verb in a thematic fit relation with the implausible attractor noun, e.g. *drank* for *wine* in examples 2. *b-d*).

5 Conclusions

In this work, we presented a study on Surprisal to investigate whether NLMs predictions are sensitive to semantic attraction. Our results on the data of the eye-tracking experiment by Cunnings and Sturt (2018) reveal that all models are sensitive to the general plausibility of the sentence, and that semantically-plausible attractors decrease the Surprisal at the target phrase, although this effect generally does not interact with sentence plausibility as in humans.

At the sentence level, no effects of attractor plausibility were observed, with the only, partial exception of a marginal significance with DistilGPT2. Interestingly, the most human-like pattern -including the interaction- has been observed with this model, the smallest one, although the specific contrasts between conditions pattern differently from human total viewing times.

Acknowledgments

We would like to thank the three anonymous reviewers for their insightful feedback.

This research was made possible by the start-up research fund (1-BD8S) at the Hong Kong Polytechnic University.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting Linear Mixed-effects Models Using lme4. *arXiv preprint arXiv:1406.5823*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. *zenodo.org*.

Won Ik Cho, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Modeling the Influence of Verb Aspect on the Activation of Typical Event Locations with BERT. In *Findings of ACL-IJCNLP*.

Ian Cunnings and Patrick Sturt. 2018. Retrieval Interference and Semantic Interpretation. *Journal of Memory and Language*, 102:16–27.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. 2018. RNNs as Psycholinguistic Subjects: Syntactic State and Grammatical Dependency. *arXiv preprint arXiv:1809.01329*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv preprint arXiv:2101.00027*.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2023. Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts. In *Transfer Learning for Natural Language Processing Workshop*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What Does BERT Learn about the Structure of Language? In *Proceedings of ACL*.

Anna Laurinavichyute and Titus von der Malsburg. 2022. Semantic Attraction in Sentence Comprehension. *Cognitive Science*, 46(2):e13086.

Russell Lenth. 2019. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R Package Version 1.4.2.

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

Richard L Lewis and Shravan Vasishth. 2013. An Activation-based Model of Sentence Processing as Skilled Memory Retrieval. In *Cognitive Science*, pages 375–419. Routledge.

Ken McRae and Kazunaga Matsuki. 2009. People Use their Knowledge of Common Events to Understand Language, and Do So as Quickly as Possible. *Language and Linguistics Compass*, 3(6):1417–1429.

James A Michaelov and Benjamin K Bergen. 2020. How Well Does Surprisal Explain N400 Amplitude under Different Experimental Conditions? In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022a. Collateral Facilitation in Humans and Language Models. In *Proceedings of CONLL*.

James A Michaelov and Benjamin K Bergen. 2022b. 'Rarely'a Problem? Language Models Exhibit Inverse Scaling in their Predictions Following 'Few'-type Quantifiers. *arXiv preprint arXiv:2212.08700*.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. Can Peanuts Fall in Love with Distributional Semantics? *arXiv preprint arXiv:2301.08731*.

Kanishka Misra. 2022. minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models. *arXiv preprint arXiv:2203.13112*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of EMNLP*.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Byung-Doh Oh and William Schuler. 2022. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? In *Transactions of the Association for Computational Linguistics*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. In *Open-AI Blog*.

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 NAACL Workshop on Cognitive Modeling and Computational Linguistics*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Proceeding of the NeurIPS EMC*$^2Workshop$.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of ACL*.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019b. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. *arXiv preprint arXiv:1905.06316*.

Julie A Van Dyke. 2007. Interference Effects from Grammatically Unavailable Constituents During Sentence Processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(2):407.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. SUBTLEX-UK: A New and Improved Word Frequency Database for British English. *Quarterly Journal of Experimental Psychology*, 67(6):1176–1190.

Marten Van Schijndel and Tal Linzen. 2018. Modeling Garden Path Effects without Explicit Hierarchical Syntax. In *Proceedings of CogSci*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Matthew W Wagers, Ellen F Lau, and Colin Phillips. 2009. Agreement Attraction in Comprehension: Representations and Processes. *Journal of Memory and Language*, 61(2):206–237.

Jason Wei, Yi Tay, and Quoc V Le. 2022. Inverse Scaling Can Become U-shaped. *arXiv preprint arXiv:2211.02011*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What Do RNN Language Models Learn about Filler-gap Dependencies? *arXiv preprint arXiv:1809.00042*.

Kayo Yin and Graham Neubig. 2022. Interpreting Language Models with Contrastive Explanations. In *Proceedings of EMNLP*.

Appendix

Descriptive statistics

The statistics for the Surprisal scores can be seen in Table 3 and 4, while the logarithmic frequencies of attractor and target nouns are in Table 5 and 6 (notice that the target nouns were the same in all the experimental conditions).

Boxplots

The boxplots for the Surprisal scores for all the metrics and models are shown in Figure 2.

	Sentence			
Models	Min	Max	Mean	Std
GPT-2	3.88	5.76	4.525	0.319
DistilGPT-2	4.150	6.010	4.824	0.399
GPT-Neo	3.400	5.460	4.268	0.391

Table 3: Cunnings dataset Surprisal mean descriptive statistics (sentence).

	Target			
Models	Min	Max	Mean	Std
GPT-2	0.74	17.35	7.597	3.759
DistilGPT-2	0.67	19.66	7.984	3.039
GPT-Neo	1.40	18.09	7.308	3.819

Table 4: Cunnings dataset Surprisal mean descriptive statistics (target phrase).

Cond.	Min	Max	Mean	Std
a,c	0.000002	0.000513	0.000085	0.000123
b,d	0.000001	0.000513	0.000077	0.000122

Table 5: Log-transformed frequency statistics for the attractor nouns across conditions in the Cunnings dataset. The frequencies were extracted with the Wordfreq library (Speer, 2022), which relies on the SUBTLEX database (Van Heuven et al., 2014).

Cond.	Min	Max	Mean	Std
a,b,c,d	0.000001	0.000525	0.000053	0.000109

Table 6: Log-transformed frequency frequency statistics for the target nouns in the Cunnings dataset. The frequencies were extracted with the Wordfreq library (Speer, 2022), which relies on the SUBTLEX database (Van Heuven et al., 2014).

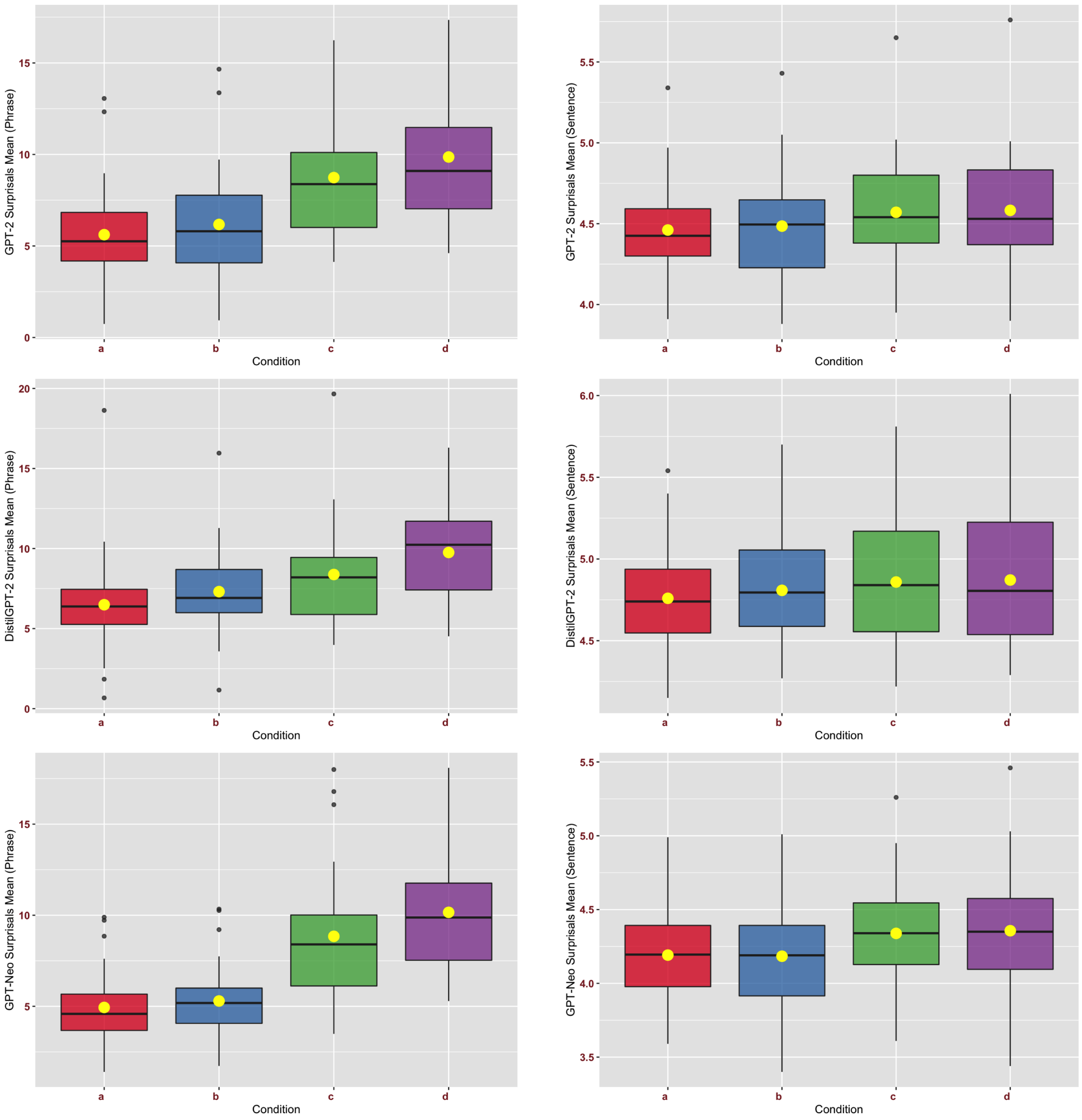

Figure 2: Boxplots of the Surprisal for all the metrics-model combinations: target Surprisal scores on the left, sentence Surprisal on the right; GPT-2 in the top row, DistilGPT-2 in the middle row, GPTNeo at the bottom.