On Modeling Dative Alternations in Large Language Models

Nat Shi

University of Michigan - Ann Arbor

A thesis submitted in partial fulfillment

of the requirements for the degree of Bachelor of Science with Honors

in Cognitive Science

Advisor: Lisa Levinson, Ph.D

April 2024

Acknowledgement

I would like to thank my advisor Dr. Lisa Levinson for all her patience, enthusiasm, and support in guiding me throughout this process. She has been most forgiving towards my errant work habits, and I would not have arrived at this point without her.

Abstract

The capabilities of generative language models to produce human-like output have largely improved over the past several years, but the degree to which they can substitute for actual humans in language tasks are less certain. Studies have shown that language models exhibit productive behaviors and generalization to unseen concepts and contexts, but the extent to which these patterns mirror human behaviors as well as the mechanisms which bias models towards more human-like linguistic human production are unclear. Models may overgeneralize in ways that humans do not, so it is important to begin to understand where and when models are able to learn the same "exceptions" to grammatical rules that humans naturally acquire. Expanding off previous work in Leong and Linzen, 2023, this study assesses pretrained GPT-2 models in regards to their treatment towards dative alternations compared to native speaker judgments. To this end, we collect human acceptability judgments on pairs of sentences demonstrating dative alternation and evaluate the target models on the same dataset. From the differences between scores given to the two parallel dative constructions in question, we find that while the models are able to produce scores similar to human judgments on some level, their performance is weaker than anticipated by previous related works. In addition, we see that human judgments are seemingly affected by participants' familiarity with target verbs. In all, the results confirm some level of model ability to replicate human-like language processing on grammatical exceptions, but raise doubts about the extent and consistency.

Contents

Acknowledgement **ii**

Abstract **iii**

Contents **iv**

Introduction **1**

Generalization 2

Exceptions and Related Literature 3

Goal and Significance of this Study 4

Methods **6**

Data and Stimuli 6

Human Judgments 7

Model Scores 8

Results and Analysis **11**

Comparison with Language Models 11

On Frequency 12

On Sentence Length 14

Discussion and Conclusion **15**

Introduction

The improvement and proliferation of large language models (LLMs) has led to rapidly expanding usage in many fields. Appropriately, interest in the capabilities and potential uses of these models within linguistics, particularly those sub-fields that examine the mechanisms and patterns of natural language production and parsing, has been on the rise. The ability of LLMs to model natural human language seems to promise numerous benefits to neurolinguistics/ psycholinguistics research, from mapping aspects of language production to quantitative representations for computational analysis to supplementing human experiment data on unusual stimuli in studies. Thus, it is to our interest that language models behave similarly to humans in language tasks—not just in generating language that is vaguely qualitatively similar to natural human language, but also in reproducing specific behaviors and habits observed in human language processing, in both parsing and production tasks.

However, the modern LLM fails at some level to mechanistically model human language production. Popular high-performance LLMs from model families like BERT, GPT, etc., are all neural networks built on top of the transformer architecture (originally proposed by Vaswani et al., 2017) and trained on massive language text corpora; it is worth noting that though this branch of machine learning models originally stemmed from neurobiological inspirations (modeling the connections between neurons in the brain), output optimizations have long caused model architectures to veer sharply away from biological accuracy. These transformer neural networks take in text input, split sentences into tokens that are roughly equivalent to morphemes, encode them as numbers, and perform a series of linear and nonlinear mathematical transformations to convert these encoded tokens into numbers indicating the relative likelihood of each token being chosen based on previous or surrounding tokens. The likelihoods are derived from the underlying probability distribution that a model defines over its training input, so LLMs are probabilistic models of their enormous training corpora. Human language processing, on the other hand, has been understood to depend upon structure and hierarchy despite partial reliance on probabilistic/ frequency-correlated mechanisms—a contrast to the linear manner in which a

model moves through a sentence and judges the appropriateness of the next token (Everaert et al., 2015). Sequential probability is hardly the be-all and end-all of language processing for humans in the way that it is for models, so it is evident that language models do not actually model human language processing.

Furthermore, most working models are typically trained on an amount of language inputs far greater than that which the average human is exposed to during language acquisition. GPT-2—the pretrained model used in this study—was trained on WebText, a 40GB body of texts the OpenAI team assembled by scraping Internet forums. If the average English word is 4.7 characters in length, then adding an extra character for spaces and/ or punctuation allows us to approximate the number of words in WebText to be around 7.5 billion words. In contrast, Gilkerson et al., 2017 shows that children up to age three hear on average 11,000-16,000 words daily in a home environment; very roughly extrapolating under some loose assumptions implies that humans are exposed to around some 50-60+ million words by their teenaged years on average, and generally not exceeding 100 million. This is at least an order of magnitude below any reasonable approximation of the number of words in WebText.

Given these differences, LLMs provide us with a probabilistic model of language that humans have produced, and not a model of human language processing. A question then arises: how close can probabilistic predictions of language come to realities of human-produced language or human language parsing?

Generalization

One of the foremost concerns in regards to the natural language capabilities of LLMs is the question of how well models are able to demonstrate generalization. While the ability of language models to generalize patterns of language compositionally (that is, learning language primitives and assembling them in unfamiliar ways) is of general concern in the field of artificial language processing due to the benefits such capabilities bring to common AI tasks like text summarization, it is also an prerequisite for the ability to process language with results similar to those of humans, given the highly productive nature of human language (Baroni, 2020; Ontanon

et al., 2022). In fact, humans are known to generalize widely: even at young ages, children learn to extend grammatical patterns in certain lexical categories, to the extent of overgeneralization (for example, conjugating all verbs with the past tense ending -ed to the effect of forming incorrect words like “swimmed”) (Wonnacott, 2011). An early worry about transformer-based language models’ poor ability to perform such generalization quickly surfaced following the introduction of the architecture, and numerous studies have been conducted in order to systematically define, test, and improve such language models’ compositional generalization capabilities (Hupkes et al., 2020; Kim & Linzen, 2020; Lake & Baroni, 2017; Ontanon et al., 2022). Though initial results were not as optimistic, transformer neural nets have been shown in the past several years to be able to demonstrate compositional generalization on a number of benchmark tasks, especially following various architectural improvements (Ontanon et al., 2022). In this sense, the currently most widely used LLMs are able to reproduce at least some human-like language processing outcomes.

However, generalization in natural language consists not only of the ability to extend grammatical patterns to unknown words and contexts, but also to know where such patterns should not be extended as a matter of linguistic convention. Young children may overgeneralize certain language patterns, but they quickly learn where those patterns don’t apply; as such, the natural next step in probing models’ ability to perform compositional generalization is to test their ability to constrain that generalization to appropriate contexts just as people do (Wonnacott, 2011).

Exceptions and Related Literature

Following the confirmation of generalizing behaviors in transformer-based language models, a line of work exploring the question of how models perform on exceptions has emerged. One such branch of work seeks to assess model performance in comparison with human sentence judgments on specific productive grammatical patterns with known exceptions—that is, grammatical patterns that we tend to generalize to new contexts often but not always. Kann et al., 2019 and Hawkins et al., 2020 both assemble language datasets and tackle this through the lens of

verb frame alternations, linguistic phenomena in which there exist distinct parallel constructions with the same verb and objects describing the same events or ideas. The former work examines models to see if verb frames—or the syntactic contexts of target verbs—can be extracted from word and sentence embeddings, and establishes that models can partially decode fine-grained lexical information from these embeddings, but that this is not sufficient for models to parse alternations as effectively and consistently as desired (Kann et al., 2019). The latter specifically constructs a large dataset called DAIS consisting of sentence pairs based on the dative alternation; they then collect human preferences between paired parallel-structure sentences and model outputs on them to posit that recent language models (i.e. transformer-based neural net architectures) show promise in capturing human preferences, especially compared to older models (Hawkins et al., 2020).

Leong and Linzen, 2023 have recently explored this problem using the English passive voice: while the passive voice in English is highly productive, meaning that speakers very easily construct passive forms of unfamiliar active verbs, there are some English verbs that are broadly held to be "unpassivizable"—that is, they are exceptions. Using pretrained GPT-2 models of various sizes as well as a small GPT-2 model trained on a small corpus to emulate more human-realistic training for comparison purposes, they collected human acceptability judgments on sentence pairs and calculated judgment score differences termed "passive drops." The corresponding model "passive drops" were shown to correlate in some measure with the human "passive drops" within semantic groupings and as a whole, and the model trained on a small corpus outperformed the other pretrained models based on the metrics employed by the researchers, implying a nontrivial degree of agreement between model and human judgment on exceptions to grammatical generalizations.

Goal and Significance of this Study

This study is largely a replication and adaptation of Leong and Linzen's work, extending it to a different dataset and problem and interrogating additional and alternative features [1]. In that

[1] Data and code for analysis are available at https://github.com/UMWordLab/dative-exceptions

spirit, it is also an extension of Hawkins et al.'s work, as it chooses to adopt the dative alternation as its subject of investigation using the methods of Leong and Linzen. Altogether, the aim is to assess how well the results of both studies might be reflected in a combined replication as well as to what extent their results might fail to extend beyond the boundaries of the original studies, if at all.

The dative alternation is one of the exemplars of Baker's Paradox, or the problem where there exist seemingly arbitrary exceptions to productive patterns in language that are acquired by speakers without sufficient negative evidence. In other words, it is the problem of how people learn exceptions in language without being expressly exposed to them. While this study is not particularly concerned with delving into the theories surrounding how the paradox can be "solved" on the human side, it is interested in finding out if models can replicate human judgments on relevant tasks (it is simply enough to know that humans conform to the exceptions in question). In the case of the dative alternation, the exception lies in the fact that while datives in English can generally be constructed in one of two different ways, some verbs are not amenable to both constructions and only appear in one syntactic context or frame. As seen in the original example provided in Baker's original paper, only one of the following sentence pairs contains two grammatically sound sentences (Baker, 1979):

1. (a) John gave the book to Alice.

   (b) John gave Alice the book.

2. (a) George said something uncharitable to Maxine.

   (b) *George said Maxime something uncharitable.

While both constructions of the dative are acceptable for the sentences using the verb "give," this is not the case for the verb "say." Though the structural transformation that created the second sentence pair is identical to that in the first, "say" is not amenable to this structural change, or alternation—it is a non-alternating verb. Let us label the default prepositional construction the prepositional dative (PD) and the alternant the double-object (DO) construction. We expect that

an alternating verb could correctly appear in either construction, whereas a non-alternating verb would only support grammatically correct PD sentences. In this study, our goal is to determine if a model can agree with human acceptability judgments on the difference between scores for PD sentences and DO sentences constructed from alternating and non-alternating verbs. This will strengthen the line of inquiry about the capabilities and limitations of large language models while generating further points of exploration for the use of language models in relevant research.

Methods

Data and Stimuli

The stimuli for this experiment consist of a set of 200 sentence pairs extracted from the DAIS dataset, where each sentence pair contains the PD- and DO-constructed versions of a single sentence (Hawkins et al., 2020). The dataset originally contained 5000 such sentence pairs for 200 unique English verbs—100 of the most common alternating verbs and 100 of the most common non-alternating verbs from their source corpora. Each verb appeared in sentences of varying length with varying indirect object/ recipient arguments (a long sentence with an indefinite recipient, a long sentence with a definite recipient, a short sentence with an indefinite recipient, a short sentence with a definite recipient, and a sentence with a pronoun recipient); further variety was added to the dataset by including varying direct objects, with five options (definite or indefinite) for each verb to round out the total of 5000 sentence pairs.

This dataset size was rather unmanageable for the scope of this experiment, especially given the contrast with Leong and Linzen who used 28 distinct verbs, of which 18 were expected to be unpassivizable and 10 passivizable. Thus, it was necessary to subset the original dataset. The original experiments in Hawkins et al. seemed to demonstrate a trend in magnitudes of negative "DO preference," or preference for the prepositional construction, for both alternating and non-alternating verbs between the recipient and sentence length categories. In particular, participants showed strongest preference for the prepositional construction in long sentences, followed by those in short sentences, followed by those using pronoun recipients. Additionally, both alternating and non-alternating verbs demonstrated that the definite and indefinite categories

within the same sentence length were rather comparable or hardly different—this is despite remarks in the introduction regarding the effects of definiteness on alternation acceptability. As such, the definite sentences (given that indefinite objects sound less awkward in sentences presented without context) were the first to be removed, followed by the pronoun sentences, which had the smallest magnitudes and would be less straightforward to compare with the long and short indefinite sentences. With similar reasoning, the sentences with definite "themes," or direct object types, were next to be removed; from the remainder, every other sentence was chosen—a decision made in order to preserve a wider range of frequency ranks—alternating between indefinite theme options (note that this is a somewhat arbitrary meaning of alternating, as all verbs had the same number of definite and indefinite "themes" but not all of the actual words were the same). This reduced the dataset down to 100 verbs, each with a long indefinite and short indefinite sentence pair with an indefinite "theme" to form a set of 200 sentence pairs, or 400 sentences.

Human Judgments

We followed Leong and Linzen's use of a scale with a midpoint set at 50 (the midpoint could not be selected as a valid rating) to allow participants to rate sentence acceptability between 0 ("Completely Unacceptable") and 100 ("Completely Acceptable") by moving a slider. The 200 sentence pairs were split into eight groups Latin-square-style so that each group saw 50 out of the 100 verbs. Given the even split between alternating and non-alternating and DO/ PD constructions, each participant was shown only one of the four possible sentences per witnessed verb. The sentences were shuffled with 50 filler sentences and presented in a randomized order to participants, who were instructed to read 100 sentences and provide a rating for each. PD vs. DO scores for each verb were then calculated by subtracting average DO from average PD for each verb.

To accomplish this, we recruited and screened a random sample of 164 native English-speaking participants in the United States through the crowd-sourcing platform Prolific, with an original target of twenty unique participants per group. Consenting participants were given example attempts before starting the experiment and each took an average of thirteen

minutes to finish rating all 100 sentences. Every sentence received approximately twenty ratings that were considered in the final analysis. Participants were excluded if more than half of a pre-selected subset of ungrammatical filler sentences were assigned a score with a participant-specific z-score of over 0, or if more than half of a pre-selected subset of grammatically sensible filler sentences were assigned a score with a participant-specific z-score of under 0.

Overall, human judgments demonstrated a clear trend consistent with expectations of higher score differences for sentences containing non-alternating verbs (this accords with expectations because DO sentences are expected to be less notably than PD sentences for non-alternating verbs, resulting in a greater difference). A visualization of score ranges can be found in Figure 1.

Model Scores

Model data were collected using the same set of stimuli and evaluated on three of the original WebText-pretrained GPT-2 models (I downloaded GPT-2-small, GPT-2-medium, and GPT-2-large; attempts to run GPT-2-XL were met with hardware limitations) from Hugging Face and five smaller GPT-2 models pretrained on a smaller dataset. The three Hugging Face models were trained on the same WebText corpus but have varying parameter counts (117M, 345M, and 774M, respectively, according to the Hugging Face site). The five smaller GPT-2 models (called GPT-2_100M in the context of this study) were versions of the small-size GPT-2 pretrained by Leong and Linzen on a small English text corpus similar in quality to WebText but limited to 100 million words in order to mimic a more realistically human-like level of exposure to language, given the disproportion between word quantities in model training corpora and human language encounters during development. Having downloaded the models from Leong and Linzen's study, we then extracted scores for each sentence and calculated a difference by subtracting sentence scores for DO sentences from sentence scores for PD sentences for each model, averaging across the five 100M models for a 100M model average in the same manner as Leong and Linzen. Due

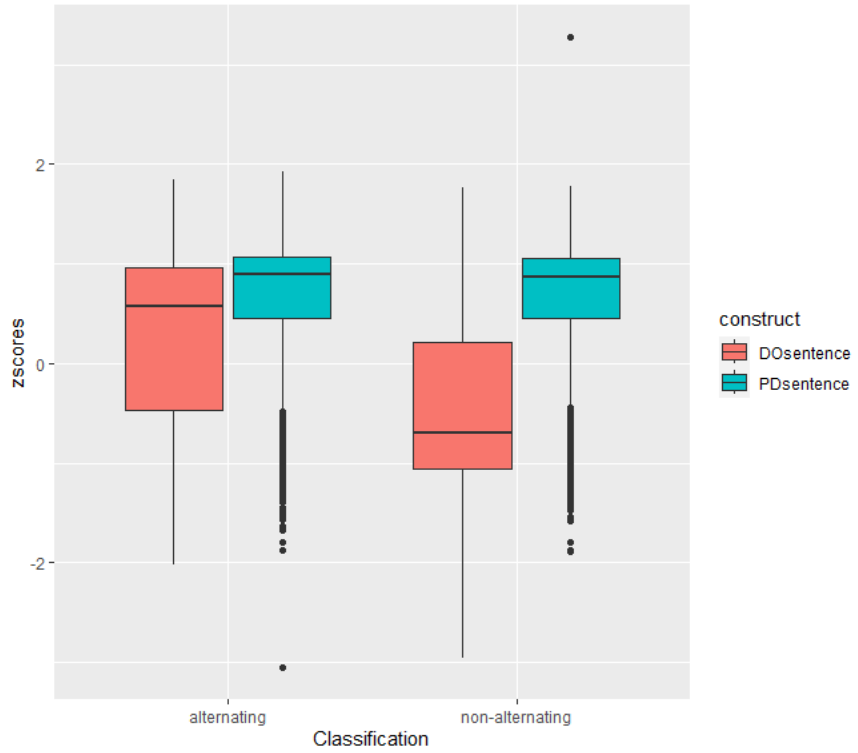

Figure 1

Z-scores of participants' acceptability scores, divided by verb classification and sentence construct. It can be seen that most participants assigned lower scores to DO sentences with non-alternating verbs than to PD sentences with non-alternating verbs, and that the median score given to DO sentences with non-alternating verbs is lower than a large majority of scores given to DO sentences with alternating verbs.

to some issues in the code given in the original repository, we used the Minicons library[2] to perform sequence scoring of sentences consistent with the syntactic evaluation paradigm referenced in Leong and Linzen. The paradigm in question, defined as "Mean LP" in Lau et al., 2017, provides a reasonable method to score sentences in acceptability judgment experiments, where a sentence is scored by first summing the log-probabilities assigned to each token by the model and then normalizing through division by sentence length (that is, token count) to mitigate effects of variable sentence length on the score:

[2] https://github.com/kanishkamisra/minicons

$$\text{Mean LP} = \frac{\log p_m(\xi)}{|\xi|}$$

where $p_m(\xi)$ describes the probability the model assigns to the sentence.

The normalization is significant given that sentences with the PD construction will have more words than their counterpart sentences with the DO construction, and different sentence pairs also have different numbers of words, which would cause longer sentences to seem more disproportionately "probable." It may be worth noting that this accords with the description given by Leong and Linzen in regards to their model scoring method, but not their code implementation.

The results are ambiguous at first glance: while the model scores seem to share the result of assigning lower scores to DO sentences constructed with non-alternating verbs similar to the human acceptability judgments as seen in Figure 2, they also seem to output comparatively higher scores for the PD sentences in general, an effect which is less marked in the human acceptability judgments and affects corresponding score difference calculations.

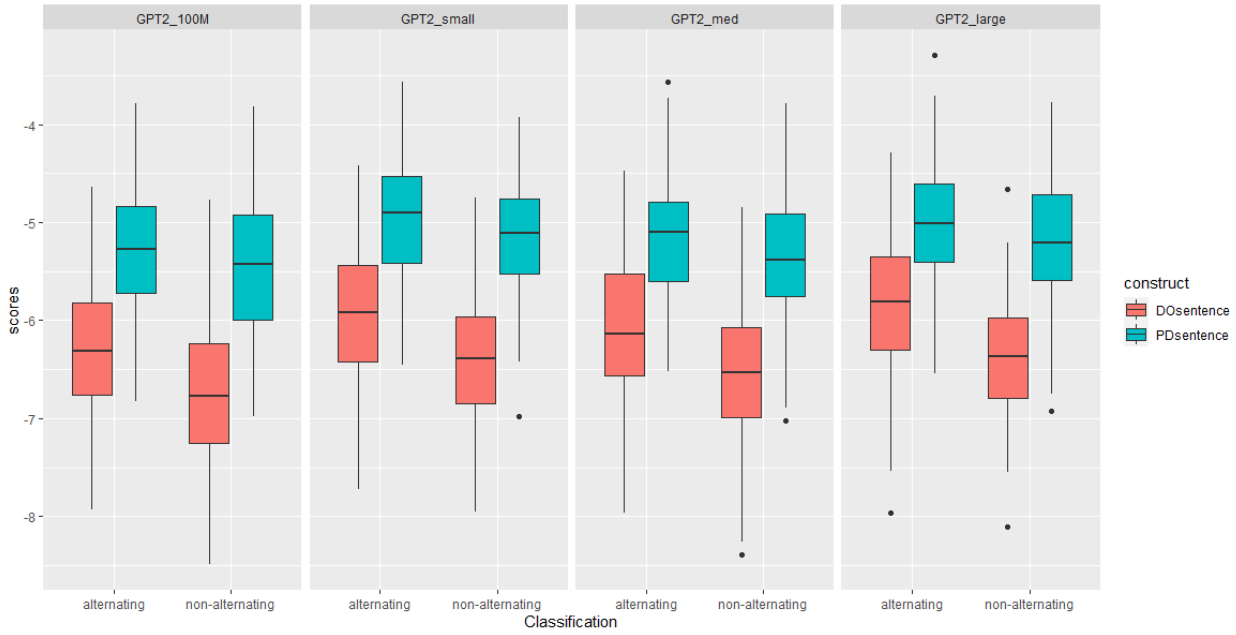

Figure 2

Scores extracted from models. Judging from quartiles, scores assigned to DO sentences with non-alternating verbs are generally lower than those assigned to DO sentences with alternating verbs; however, all DO scores are also notably lower as a whole than PD scores.

Results and Analysis

Comparison with Language Models

Following the steps taken in Leong and Linzen, we graphed the differences between scores for PD sentences and DO sentences (referring to them as score differences from hereon) from the models against the score differences from the human participants, as seen in Figure 3. In comparison to the results from Leong and Linzen, the correlation coefficients for all models are rather low (see Table 1), and the score differences produced from the models are not as clearly separable by verb classification. At most, there is a weak positive relationship between the score differences from human judgments and model outputs. Neither do we see the pattern from Leong and Linzen where the model trained on a smaller corpus outperforms the others; in fact, the averaged score differences of the 100M models trained by Leong and Linzen seem to underperform in comparison to the other models. Such a result would have implied that even data similar to the amount that a human might see is sufficient for a model to learn human-like preferences. Instead, the model which performs best is GPT-2-large, implying that model size may have a nontrivial effect on the ability of the model to judge these exceptions based on distributional training data. Reducing the training dataset size may have even had a negative effect on model results, as the same-sized GPT-2-small, trained on the full WebText corpus, outperforms GPT-2_100M on producing scores more similar to human judgments.

Table 1

Models and correlation coefficients

Model	r_s
GPT-2_100M	0.148
GPT-2-small	0.2135
GPT-2-medium	0.2271
GPT-2-large	0.3082

While the cause is unclear, the models' tendency to prefer PD sentence constructions for both alternating and non-alternating verbs may be contributing to the relative lack of clear separability between score differences for sentences with alternating and non-alternating verbs. Reasons for this may merit further future study; if WebText is ever made public, one could attempt to perform analysis on the training material for the Hugging Face models to see if prevalence of

either construction in the original data might be correlated to this pattern. It would also be interesting to see if model predictions might change if sentences were not presented in isolation.

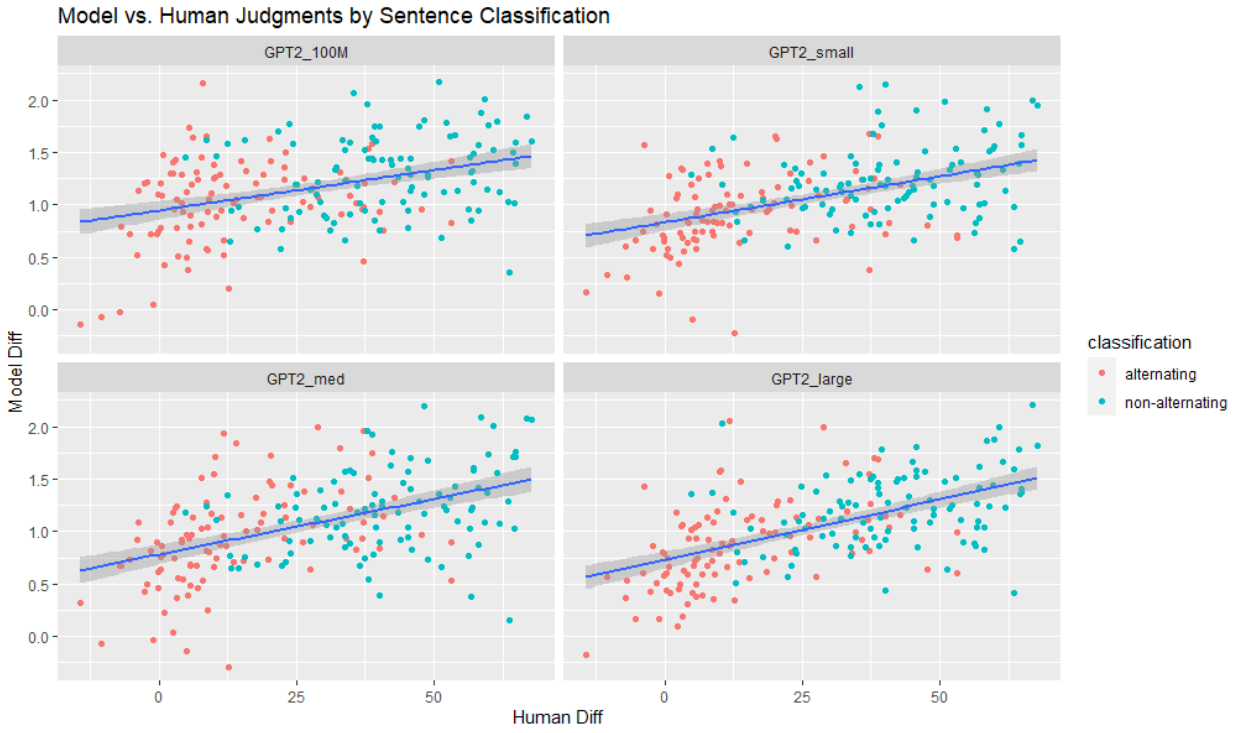

Figure 3

Score differences from models vs. average score differences from humans for each sentence pair. While there is some degree of correlation, it is weak. Note that even visually, score differences from human judgments exhibit a degree of separability between sentences with alternating and non-alternating verbs.

On Frequency

Meanwhile, further inspection of the data reveals an interesting trend in judgments in relation to verb frequency. While these verb frequencies (taken from the DAIS dataset and based on the original corpus from which the verbs were drawn) are relative rankings and not absolute, there is a rather compelling relationship between the magnitude of differences in sentence score differences from the human judgments, as seen in Figure 4. Note that this pattern, while still present to some degree, does not exhibit itself quite as strongly in the model data.

We additionally fitted a mixed linear model to predict human and model scores from frequency rank and verb classification with sentence construct type using verb ID, sentence pair ID, and participant ID as random intercepts. The results relevant to frequency rank in particular are shown in Table 2. Estimated slopes are small and p-values high enough to lack statistical significance ($p > 0.05$) for all the models, further substantiating the relative lack of effect of frequency rank on model outputs in comparison to human outputs.

Table 2

Frequency Rank Effects

Judgment Source	Slope Estimate	p-value
Human	-0.17307	0.000306
GPT2_100M	-2.524e-03	0.217
GPT2-small	-2.672e-04	0.891
GPT2-medium	-2.312e-03	0.23
GPT2-large	-1.970e-03	0.298

On the one hand, these results imply that human acceptability as a metric or target of study may be biased by participants' familiarity with various words. Not only were score differences for non-alternating verbs higher, but score differences for alternating verbs were lower for more familiar verbs; in comparison, the score differences for the two groups was much closer for the less familiar verbs. That the models were far less impacted by frequency rank calls into question why this effect was not observed—while the training corpora for the models might not be exactly the same as the sources for the DAIS dataset, the relative frequency rankings of verbs should not change drastically (though one could perform further analysis on Leong and Linzen's training data or WebText-like corpora to estimate the extent of any such differences). Perhaps it is possible that the large amounts of data models are trained on results in less biased treatment of words that human participants would have trouble easily retrieving grammatical contexts for, but if it was only a matter of training corpus, we might expect the 100M model to show some difference. Thus, any cause for this might have more to do with how models are processing input and what biases or lack thereof they demonstrate. Perhaps the same mechanisms that allow language models to generate outputs on input stimuli for which sufficient human judgments are difficult to

gather is what is causing the models to fail to exhibit a pattern more similar to the human data here, but this is conjecture that requires further research.

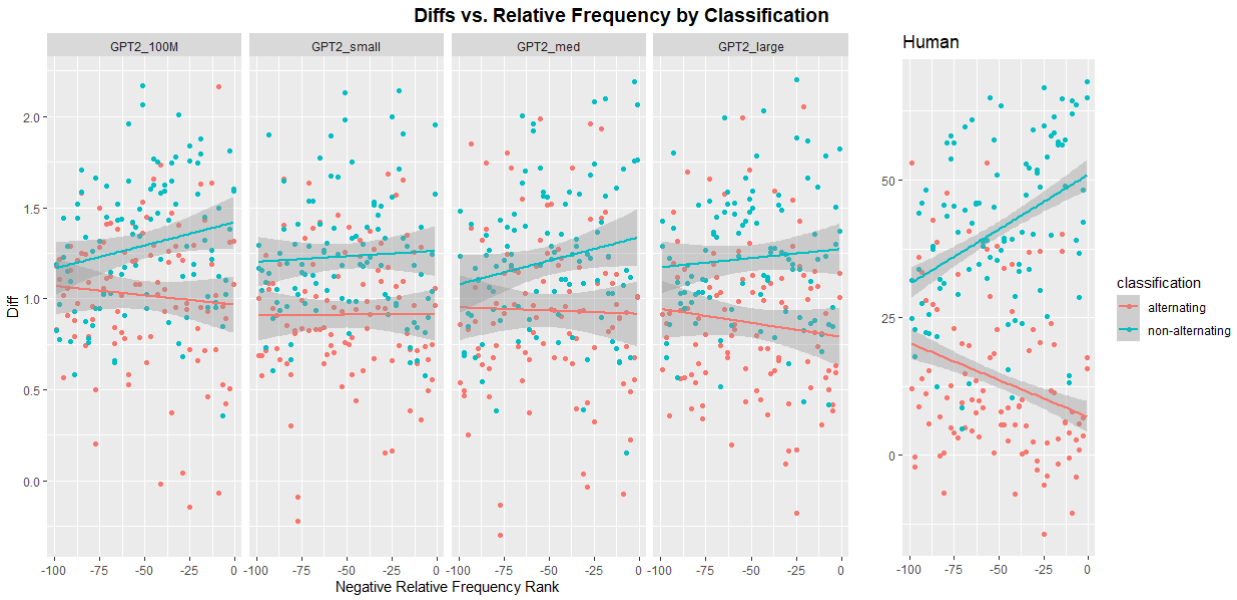

Figure 4

Score differences for each verb for models and human judgments compared to relative frequency ranks, with the lowest (near -100) ranks being most infrequently occurring.

On Sentence Length

Another feature worthy of some note was the seeming effect of sentence length on results, particularly the effect of sentence length on human judgment results. Distributions of score differences for alternating-verb long sentences vs. short sentences look quite different in the human data, as seen in Figure 5; this same effect is far less pronounced in the model data. For both alternating and non-alternating verbs, human sentence score differences are generally lower for the short sentences, with narrower distributions. In some of the models, however, this effect is not only not observed for non-alternating verbs, but even reversed. While there are questions and further examinations to be pursued on the pronounced possible relationship between sentence length and human acceptability judgments, in the context of this study, it can only indicate that this is another aspect along which human and language model sentence judgments may differ. Despite the fact that for some tasks a lack of bias due to varying sentence length may be

beneficial, our primary concern here is merely whether or not this behavior is human-like, in which case it appears not to be so.

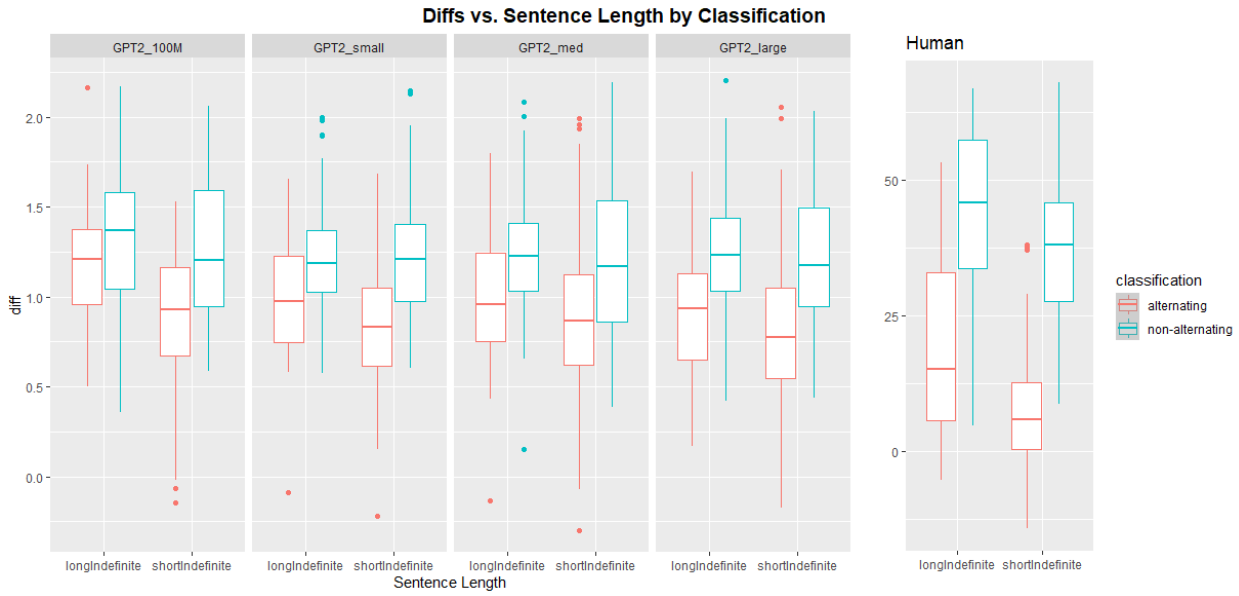

Figure 5

Score differences for each verb grouped by sentence length; human judgments seem to be more heavily affected.

Discussion and Conclusion

The goal of this study was to examine the ability of pretrained LLMs to produce judgments similar to human participants on a well-known case of a grammatical exception in English. To do so, we compared model scores to human acceptability judgments on sentences demonstrating the dative alternation, subtracting scores for double-object constructions from scores for prepositional constructions. Overall, the results provide mixed support for the ability of models to demonstrate human-like exceptionality despite past work suggesting that models can efficiently acquire such patterns for both dative alternations and other exceptions. While there was a degree of correlation between sentence score differences produced by models and score differences from human acceptability judgments, it was not strong enough to make a compelling case for the ability of models to produce significantly human-like judgments on dative constructions. This differs to some extent from the results of Hawkins et al., 2020 (though some

details are consistent, such as overall trends and better performance of larger models), and part of the difference may be due to differences in how acceptability was assessed, as Hawkins et al. frames human judgment as a function of binary preference as opposed to differences in acceptability. In such a setup, users would have been exposed to both constructions simultaneously and thus may have displayed somewhat different biases. However, the failure of models to produce clearly different results between alternating and non-alternating verbs is a point of concern; there is more to be examined in understanding why models had such strong preferences for specific sentence constructions regardless of whether verbs were amenable to the double-object construction, and whether or not this could be partially explained by patterns in the models' original training corpora.

Further examination of this specific problem could also explore the same task with more varied examples of English grammatical exceptions using different-sized models to see which exception patterns result in better performance from larger or smaller models, and whether or not these are consistent between replications as well as between different model families. If the passives exceptions case explored in Leong and Linzen consistently generates better results on smaller models and the dative alternation exceptions case explored in this study as well as in Hawkins et al. consistently finds better performance from larger models, there may be some implications for how language models learn exceptions, as well as what types of exceptions are more easily acquired from training data.

Additional considerations from this study include the effects of factors like sentence length and verb frequency rank. While an analysis of the prevalence of prepositional vs. double-object constructions in the original training corpora of the models was not performed and may constitute a worthy future exploration, we did examine the effects of relative frequency rank, which was proposed by Leong and Linzen, 2023 as an extension in their work, and found that more frequently-appearing verbs elicited more exaggerated differences in human acceptability judgments between acceptable and unacceptable sentences. However, as this effect was not shown as clearly in the model scores, there is some exploration to be done in ascertaining the cause for

this. Note that this study did not perform direct frequency analysis on training corpora, which may have affected the results; likewise, this study does not perform any analysis on absolute frequencies of verbs. A thorough exploration of frequency effects might require study of large models pre-trained on public datasets (as opposed to GPT-2, as WebText is still unreleased) or the training of models such as GPT-2 on accessible corpora for experimentation. Similarly, human biases related to sentence length were not reproduced by the models; while one could test for sentence length effects on other exceptions, it would also be worthwhile to determine the extent to which sentence lengths can bias human acceptability judgments to begin with, and whether or not these are correlated with some other grammatical feature that was unaccounted for in this study.

Finally, the original DAIS dataset and related datasets contain other parameters (e.g. manipulations of the direct object, use of pronouns, etc.) that were neglected in this study but might prove fruitful directions of inquiry in future investigations.

Altogether, the results of this study illustrate that while language models are able to produce judgments on English grammatical exceptions that are correlated with human judgments, this correlation is weaker than may have been implied by previous related work. On the one hand, our results imply that there is further work to be done both in understanding human acceptability biases and model biases towards specific grammatical constructs. On the other hand, our results suggest that language models struggle to reproduce certain important aspects of human language and should thus be used with due caution in tasks which may be affected by such features. While distributional information may have some role in helping language models acquire a level of human-like treatment of exceptions, there is a large gap between the current weakly correlated results and the level of human-likeness we might prefer for practical detailed uses.

References

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, *10*(4), 533–581. Retrieved April 8, 2024, from http://www.jstor.org/stable/4178133

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190307. https://doi.org/10.1098/rstb.2019.0307

Everaert, M. B., Huybregts, M. A., Chomsky, N., Berwick, R. C., & Bolhuis, J. J. (2015). Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in Cognitive Sciences*, *19*(12), 729–743. https://doi.org/https://doi.org/10.1016/j.tics.2015.09.008

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265. https://doi.org/10.1044/2016_ajslp-15-0169

Hawkins, R. X. D., Yamakoshi, T., Griffiths, T. L., & Goldberg, A. E. (2020). Investigating representations of verb bias in neural language models. *CoRR*, *abs/2010.02375*. https://arxiv.org/abs/2010.02375

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020, July). Compositionality decomposed: How do neural networks generalise? (extended abstract) [Journal track]. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 5065–5069). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2020/708

Kann, K., Wastadt, A., Williams, A., & Bowman, S. R. Verb argument structure alternations in word and sentence embeddings. In: *2*. University of Massachusetts Amherst Libraries, 2019, January, 287–297. https://doi.org/10.7275/q5js-4y86

Kim, N., & Linzen, T. (2020, November). COGS: A compositional generalization challenge based on semantic interpretation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing*

(emnlp) (pp. 9087–9105). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.731

Lake, B. M., & Baroni, M. (2017). Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:46761158

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241. https://doi.org/https://doi.org/10.1111/cogs.12414

Leong, C. S.-Y., & Linzen, T. (2023, June). Language models can learn exceptions to syntactic rules. In T. Hunter & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics 2023* (pp. 133–144). Association for Computational Linguistics. https://aclanthology.org/2023.scil-1.11

Ontanon, S., Ainslie, J., Fisher, Z., & Cvicek, V. (2022, May). Making transformers solve compositional tasks. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 3591–3607). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.251

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wonnacott, E. (2011). Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1–14. https://doi.org/https://doi.org/10.1016/j.jml.2011.03.001