Do Language Models Exhibit Human-like Structural Priming Effects?

Jaap Jumelet[1] **Willem Zuidema**[1] **Arabella Sinclair**[2]

[1]Institute for Logic, Language and Computation
University of Amsterdam

[2]School of Natural and Computing Sciences
University of Aberdeen

jumeletjaap@gmail.com w.h.zuidema@uva.nl arabella.sinclair@abdn.ac.uk

Abstract

We explore which linguistic factors—at the sentence *and* token level—play an important role in influencing language model predictions, and investigate whether these are reflective of results found in humans and human corpora (Gries and Kootstra, 2017). We make use of the structural priming paradigm, where recent exposure to a structure facilitates processing of the same structure. We don't only investigate whether, but also *where* priming effects occur, and what factors predict them. We show that these effects can be explained via the *inverse frequency effect*, known in human priming, where rarer elements within a prime increase priming effects, as well as lexical dependence between prime and target. Our results provide an important piece in the puzzle of understanding how properties within their context affect structural prediction in language models.

1 Introduction

Structural priming is the phenomenon where speakers are more likely to repeat a certain structure after being recently exposed to a sentence containing a congruent structure; in the following example a speaker is more likely to produce a Double Object (DO) construction (2a, the *target*) after having been exposed to a sentence with a congruent structure (1a, the *prime*) than after having been exposed to a sentence with an incongruent structure (1b, which illustrates the Prepositional Object (PO) dative construction):

(1) a. The girl gave [the boy]$_{NP}$ [the ball]$_{NP}$
b. The girl gave [the ball]$_{NP}$ [to the boy]$_{PP}$

(2) a. The baker gave [the lady]$_{NP}$ [the cake]$_{NP}$

Structural priming is well attested in humans, for both language production (Mahowald et al., 2016) and comprehension (Tooley, 2023). Interestingly, it has also been shown to occur in large language models (Prasad et al., 2019; Sinclair et al., 2022; Michaelov et al., 2023). Here, structural priming can be viewed as a simple form of 'in-context learning' (Dong et al., 2022), where the *task* is to generate a sentence (or compute its likelihood) with the target grammatical structure, influenced by the *demonstration* (the *prime* presented to the LLM before processing the target).

Priming effects in humans are typically stronger when there are shared words between prime and target, and when the prime is more unusual, or less frequent. This is the *inverse frequency* effect; it extends to other properties of structures themselves, and it is one of the main phenomena we focus on in this paper. To explain these effects without direct access to the underlying training data, we turn to factors known to predict priming effects from corpus linguistics (e.g. Gries, 2005; Jaeger and Snider, 2013), which highlight surprisal and structural preference as key factors, and demonstrate the importance of a more fine-grained method of measuring priming.

A second focus of this paper is examining the relationship between lexico-semantic overlap and the asymmetry of the priming effects observed. We examine priming at the token level, discovering that *where* priming takes place is important for understanding *how* lexico-semantic factors affect priming and for analysing the mechanisms underlying priming in models. Finally, we find that models' structural predictions are highly influenced by specific lexical items, and that they incorporate systematic properties of human production preferences learnt from the training data. We demonstrate that models, like humans, exhibit inverse frequency effects in terms of surprisal and verb preference, and that these are predictive of priming.

2 Structural Priming

Structural priming in humans is part of a rich literature on factors that impact human language processing, both in controlled experiments of production

Findings of the Association for Computational Linguistics: ACL 2024, pages 14727–14742

(Mahowald et al., 2016) and comprehension (Tooley, 2023), and analyses from corpus linguistics (e.g., Gries and Kootstra, 2017). We provide a brief theoretical background on structural priming in §2.1, and priming in language models in §2.2.

2.1 Properties of Priming

Production vs. Comprehension Structural priming has been shown to manifest in both language production and comprehension. Although recent work has shown that the underlying mechanisms for these two areas may not be as different as originally assumed (Segaert et al., 2013; Tooley and Bock, 2014) and are intricately related (Dell and Chang, 2014), numerous works have uncovered distinct differences in the factors that play a role for each modality (Ziegler and Snedeker, 2019). We therefore take the explicit stance in this paper that language models are likely to follow patterns found in human production, since they are exposed solely to human produced data, and for the factors we consider, we find this to be the case. While LMs are not necessarily expected to align directly with factors found in comprehension studies, arguably there may be similar acquisition mechanisms (e.g. error-based learning) that result in comprehension aligned behaviour. In this background, we focus on production- and corpus-based analyses of structural priming, unless unless explicitly mentioned otherwise.

Inverse Frequency One influential theory on the mechanism behind priming in humans is the implicit learning theory by Chang et al. (2006). This theory predicts that our expectation for a particular structure is proportional to degree of ***surprisal*** of having encountered this structure before. This effect —the *inverse frequency effect* (a *rarer* prime will boost priming more) — has indeed been confirmed experimentally to be a strong predictor of priming behaviour. Specifically, in language production in humans it has been found that highly surprising primes (as measured by language models) will have higher priming effects (Gries and Wulff, 2005; Jaeger and Snider, 2008, 2013; Fazekas et al., 2024).

Relatedly, ***structural preference***—which expresses within which structure a verb is most likely to occur—is another important factor when predicting priming behaviour: verbs that are strongly associated with one construction are more likely to be primed by that construction as well (Gries and Wulff, 2005; Gries et al., 2005; Bernolet and Hartsuiker, 2010). From this it then follows that priming effects are stronger when the prime sentence was of a less preferred structure: a prime containing the verb *gave*, for example, will prime subsequent targets more strongly when it is encountered in its *dispreferred* structure (PO) (Pickering and Branigan, 1998; Zhou and Frank, 2023). There exists an extensive line of work into determining the factors that govern this structural preference, which is driven by various complex syntax-semantic interactions (Green, 1974; Thompson and Koide, 1987; Gropen et al., 1989; Bresnan et al., 2007). Inspired by this literature, we find evidence in §6 of a verb-mediated inverse frequency effect in modern LLMs.

Lexical Dependence Many findings in production and corpus studies have shown that priming effects of sounds, words, meanings and structures interact: prime sentences and target sentences with shared words (***lexical overlap***), or words that share semantics (***semantic overlap***), boost structural priming (Hare and Goldberg, 1999; Jones et al., 2006; Hartsuiker et al., 2008; Snider, 2009; Gerard et al., 2010), and similar findings have been found in comprehension studies as well (Chiarello et al., 1990; Lucas, 2000; Traxler et al., 2014). A common explanation is that words in the prime that are identical or similar to words in the target already activate the relevant abstract syntactic frames. These frames, in turn, are most closely associated with verbs, or the syntactic head of the primed structure (Pickering and Branigan, 1998; Pickering and Ferreira, 2008; Reitter et al., 2011).

Lexical overlap effects in human experiments typically do not consider effect of preposition or determiner overlap, rather focusing on the content words. Findings have shown that structural priming does not depend on the repetition of function words, thus in humans there is a clear difference between content-word and function word repetition (Bock, 1989; Tree and Meijer, 1999; Pickering and Ferreira, 2008).

2.2 Structural Priming in Language Models

Structural priming has been used to investigate abstract language representations in language models. A number of (early) papers used fine-tuning on a small sample of items of a particular structure, and measured its impact on related items (van Schijndel and Linzen, 2018; Prasad et al., 2019). Sin-

clair et al. (2022) measure the impact of congruent and incongruent prime sentences on a subsequent target, paralleling approaches in psycholinguistics that view priming as resulting from *residual activations* (Branigan et al., 1999). Using this approach, LMs are shown to exhibit priming effects that are cumulative, susceptible to recency effects, boosted by lexico-semantic overlap, and persisting in cross-lingual settings (Michaelov et al., 2023; Xiao et al., 2024).

One key finding of Sinclair et al. (2022) is that priming effects are often asymmetric: when comparing alternative structures in the dative and transitive data, they remark that some of these structures are more susceptible to priming than their alternatives. In §4 we confirm this observation for the dative; the strength and direction of the asymmetry are a surprising result, given priming effects are typically higher for the *opposite* alternation in humans (Bock, 1989; Kaschak et al., 2011; Reitter et al., 2011). We show that this finding extends to a wide range of state-of-the-art LLMs, and is predictable via other inverse frequency effects.

3 Measures, Data & Models

3.1 Sentence-level Priming Effect

To measure the priming effect, we make use of the measure of Sinclair et al. (2022), which has recently also been adapted by Sinha et al. (2023) and Michaelov et al. (2023). The Priming Effect (PE) is defined as the difference in log probability of a target sentence T^X when preceded by a prime P^X that has the same congruent structure X (PO/DO), and the log probability of the same target T^X that is preceded by a prime P^Y of incongruent structure Y (to contrast this measure with the measure from §3.2, we will refer to it as the *sentence*-level Priming Effect, s-PE):

$$s\text{-}PE(\text{X}) = \log P(\text{T}^\text{X}|\text{P}^\text{X}) - \log P(\text{T}^\text{X}|\text{P}^\text{Y}) \quad (1)$$

The conditional probability of $\log P(\text{T}^\text{X}|\text{P}^\text{X})$ is computed as the sum of log probabilities of all tokens in the target sentence:

$$\log P(\text{T}^\text{X}|\text{P}^\text{X}) = \sum_i \log P_{LM}(\text{T}^\text{X}_i|\text{P}^\text{X}, \text{T}^\text{X}_{<i}) \quad (2)$$

3.2 Token-level Priming Effect

The Priming Effect metric of Eq. 1 shows whether a target sentence is primed by structural congruence as a whole, but does not provide insight into *which* tokens within the target were most responsible for such an effect. To investigate this, we introduce the **token-level priming effect** metric (w-PE), which expresses priming effects for each individual target token T^X_i:

$$\begin{aligned} w\text{-}PE(\text{X}, i) = &\log P(\text{T}^\text{X}_i|\text{P}^\text{X}, \text{T}^\text{X}_{<i}) \\ &- \log P(\text{T}^\text{X}_i|\text{P}^\text{Y}, \text{T}^\text{X}_{<i}) \end{aligned} \quad (3)$$

Note that the sentence-level PE decomposes into a sum of w-PE scores; as such w-PE expresses the relative contribution of each target token to s-PE:

$$s\text{-}PE(\text{X}) = \sum_i w\text{-}PE(\text{X}, i)$$

3.3 The Prime-LM Corpus

We use the dative constructions from the Prime-LM corpus of Sinclair et al. (2022), similar to examples (1) and (2) in §1. This subset of sentences is convenient for our purposes, because we can select both prime-target pairs with *no* lexical overlap and minimal semantic similarity between nouns and verbs, as well as pairs with varying degrees of overlap and varying degrees of semantic similarity. The datives thus allow us to not only measure structural priming, but also inspect the role of lexical overlap in more detail. We briefly explain the subsets we select for our experiments (each containing 15.000 prime/target pairs), as well as two additional sub-conditions we introduce to the lexical overlap category.

Core contains a) no lexical overlap exists between prime and target sentences, not even between function words, and b) no semantic association exists between prime and target exists in the USF free association norms dataset (Nelson et al., 2004). In our experiments, we use the Core condition as a baseline.

Semantic Similarity contains explicit pairwise semantic similarity between prime and target, where similarity is assessed by a non-zero human association from the USF dataset or a minimum cosine similarity of at least 0.4 based on GPT2-large embeddings. We consider three conditions : i) all nouns are semantically similar, ii) the verbs are similar, iii) all nouns *and* verbs are similar.

Lexical Overlap ensures lexical items are shared across prime and target. We consider three

such conditions : i) all nouns overlap, ii) determiners and prepositions overlap, iii) verbs overlap. We create two additional conditions, iv) determiner overlap and v) preposition overlap. This allows us to separately measure the impact of determiners and prepositions, since *verb overlap* necessitates preposition overlap.

3.4 Models

We consider the following (auto-regressive) LLMs. For models with an * we also test their *aligned* versions. PE scores are computed using the `diagnnose` library (Jumelet, 2020).

GPT2-large (Radford et al., 2019): This is the impactful 2019 model from OpenAI, with 774M parameters, trained only on a (causal) language modelling objective.

***Llama-2-7b** (Touvron et al., 2023): We consider both the 7B base model and the RLHF/PPO aligned *chat* model (Ouyang et al., 2022).

***Falcon-7b** (Almazrouei et al., 2023): We consider both the 7B base, and the *instruction-tuned* variant fine-tuned on dialogue data taken from ChatGPT (OpenAI, 2023).

***Mistral-7b** (Jiang et al., 2023): This 7B model is the current state-of-the-art in this size bracket. We also consider the instruction-tuned variant, trained on similar data to Falcon-7b.

***Zephyr** (Tunstall et al., 2023): An aligned version of Mistral-7b using Direct Preference Optimization (Rafailov et al., 2023).

4 Exp1: Measuring Structural Priming

We aim to better understand the asymmetrical priming effects observed in Sinclair et al. (2022), to gain a more detailed picture of how lexical overlap affects this asymmetry. We start our experimental setup with their sentence-level approach, considering a wider range of large, contemporary LLMs. In the next section we then examine priming effects at a more fine-grained level.

Priming Effects are skewed and correlated We compute the s-PE scores for the models of §3.4 on the Core condition of Prime-LM. We observe there exists a strong negative correlation between the PE scores of the prepositional object and the double object. In Figure 1A we plot those results as a scatter plot in the space formed by PE score for one construction against PE score for the alternative construction. This representation highlights that, for Llama-2-Chat and all the other models we consider, the s-PE of PO constructions is *negatively* correlated with that of DO constructions (ρ: -0.72 to -0.77), and that only for a fraction of sentences there exists a positive priming effect in both directions (26 to 38%).

This correlation and skew towards one of the two constructions were already observed by Sinclair et al. (2022) for GPT2-large and other relatively small LMs. Interestingly, correlation and skew do also exist in the newest, large LMs, and, moreover, are even more pronounced. Figure 1C shows the mean s-PE in the same *PE space* for all the LLMs we considered. GPT2-large shows the least skew, Llama-2-chat the most (for completeness, the plot also shows the strength of the correlation between the PE(PO) and PE(DO) scores, as well as the spread of the distribution.) Note that this observed behaviour of *large* LMs is less consistent and far more asymmetric than results in the human literature, where priming effects, while typically asymmetric (Bock, 1989) , are generally observed to be positive for both structures.

Lexical overlap balances Priming Effects Next, we investigate the priming effects of the LMs where either the semantic similarity or lexical overlap between prime and target is increased. The results for lexical overlap are shown in Figure 1C–E (additional plots regarding semantic similarity are in Appendix B). The plots show that an increase in lexical overlap of any type moves all models more solidly into the upper-right quadrant of the PE-space. That is, it pushes all LM priming behaviours to become both stronger and more balanced (less skewed towards one or the other construction). This is especially prevalent for the overlap in verbs and function words. We will explore the impact of these factors in more detail in the next section.

5 Exp2: Locating Structural Priming

Sentence-level analysis does not allow us to investigate individual token-level predictions, making it impossible to examine *where* in the target sentence priming effects are at their strongest. To better understand the s-PE results of §4 we thus compute the token-level w-PE scores for the same conditions.

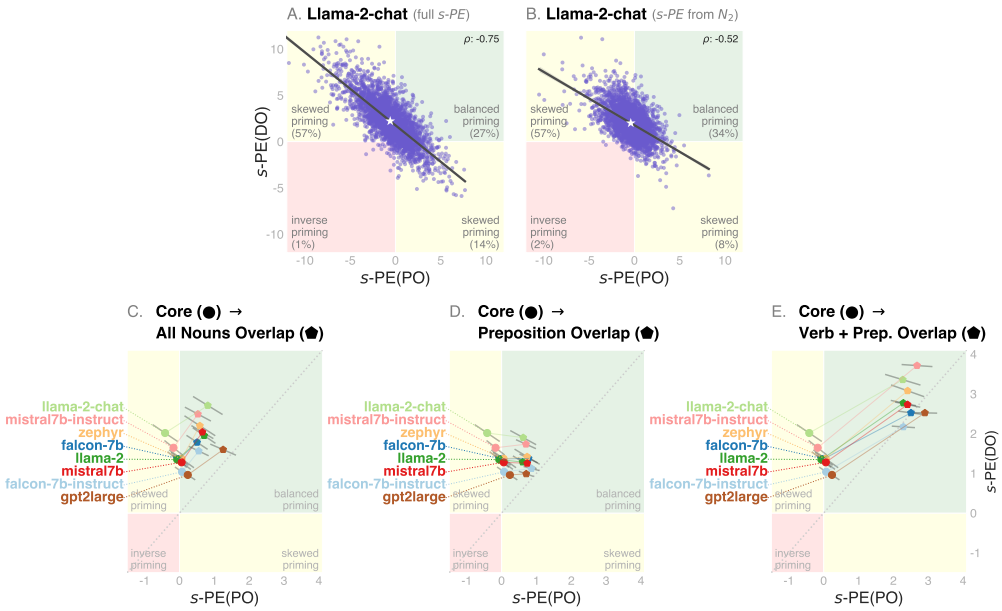

Figure 1: We plot PE results against one another. The four quadrants in this '*PE space*': *balanced priming* where the PE is positive in both directions, *skewed priming* where it's only positive in one, and *inverse priming* when the PE is negative in both directions. There exists a strong negative correlation between priming effects of opposite structures (A). Only a small portion of the data is primed in both directions for Core. Priming becomes more balanced when measured from the point of divergence in the target (B, §5), or when lexical overlap is increased (C–E).

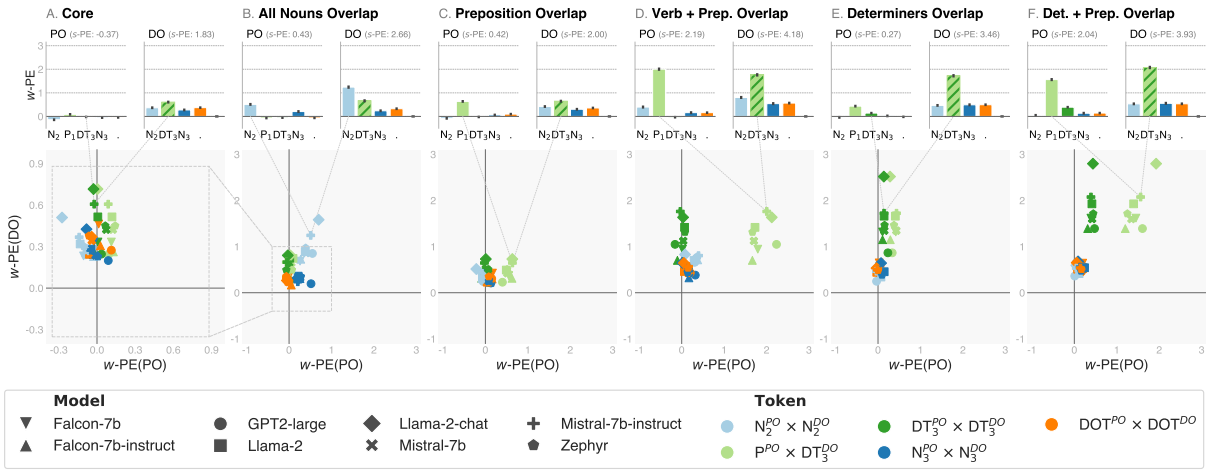

Figure 2: The w-PE scores for the *Core* and *Lexical Overlap* conditions. Scores are grouped by token (based on colour) and model (based on shape). To exemplify how these *Priming Space* coordinates map to a bar chart, we show the Mistral-7b-instruct scores at the top of each plot. Note that the *Core* results are plotted at a different scale than the other conditions. PO: *The girl gave the* **ball** to the **boy** . DO: *The girl gave the* **boy** the **ball** .

Structural Divergence Figure 3 shows the average w-PE scores for Llama-2 on the *Core* condition, which exhibits much higher sentence-level priming effects for the DO sentence than for the PO sentence. The *token-level* scores show that the treatment of the two sentences starts to diverge from the position of the *second* noun onwards (the second noun is is the patient for PO, whereas it is the recipient for DO). Prior to that, the target sentence is, in fact, the same for both PO/DO alternations and as such will be inversely proportional to each other: scores up till this point merely show a target's bias towards a prime of a particular structure, regardless of structural congruence.

This provides a partial explanation for the strong negative correlation between s-PE scores that was observed in §4: since (roughly) half of s-PE score is made up of w-PE scores that have a perfectly

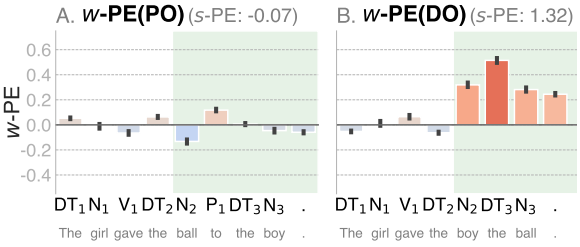

Figure 3: The token-level Priming Effect reveals which token predictions in the target sentence contributed the most to the overall sentence-level Priming Effect, here averaged for Llama-2 over the *Core*. It is inversely correlated up to the point of divergence between the two structures, at the position of the second noun (N_2). Only after that point can the congruence between prime and target start playing a role.

negative correlation of -1, the overall correlation of s-PE scores is strongly affected by this. Based on this insight we compute a modified s-PE score that is only measured from the point of target sentence divergence (s_δ):

$$s_\delta\text{-}PE(\text{X}) = \sum_{i>4} w\text{-}PE(\text{X}, i) \quad (4)$$

This allows us to confirm that the negative correlation between PO and DO decreases with s_δ-PE (ρ: -0.76 to -0.52 for Llama-2-chat; Figure 1B).

Lexical Dependence We focus our analysis on *lexical overlap* (Figure 2), which showed the strongest balancing effects in §4.[1] Priming behaviour could be distributed in two ways across the target: either uniformly or peaked at a particular token, and either balanced or skewed towards one structure. From the point of structural divergence, we have the noun (N_2) of the first noun phrase, followed by the function word (P_1:PO, DT_3:DO) that marks the start of the second noun phrase of the construction. Priming effects for N_2 stem from cues with respect to the *semantic role* (e.g. gave the ball_{PO} | boy_{DO}). Numerous works in production have shown priming to already take place at this location (Pickering and Branigan, 1998; Cleland and Pickering, 2003). We would therefore expect to find some evidence of *balanced* priming from the N_2 within the core condition. However, the *skewed* priming we observe in *Core* (2A) suggests that the semantic role of the noun does not play as important a part in structural prediction for models.

[1]Results for semantic similarity are provided in Appendix C.

Indeed, we observe the most consistent and balanced priming effects from the start of the second NP (w-PE(PO, P_1), w-PE(DO, DT_3)), suggesting that models only narrow their structural predictions later on within a sentence.

Next, we observe the local impact of lexical overlap between prime and target. For overlapping nouns, we can see that the w-PE for both N_2 and N_3 has increased significantly for both PO and DO. The other tokens, on the other hand, are not impacted by this overlap at all: the priming boost manifests itself solely at the position of the overlapping token. For verb overlap, we show that the increased $s-$PE scores here stem from the verb as well as the prepositional overlap (necessary when sharing the same verb and preserving semantics) resulting in significantly larger w-PE(PO, P) scores (Figures 2C and D). Interestingly, verb overlap also leads to a boost in N_2 and N_3, compared to the *Core*. This shows that, under this condition, the model *is* aware of the expected semantic role in the N_2 position: the verb overlap has primed the model in the DO case to expect an animate entity here (and inanimate for PO). Unlike findings in the human literature for both production (Bock, 1989) and comprehension (Traxler, 2008), we observe prepositional overlap strongly boosting priming effects in the language models we investigate.

6 Exp3: Explaining Structural Priming

We now take a closer look at the factors that impact Priming Effects by conducting a regression analysis inspired by factors from corpus linguistics and production studies (Gries, 2005; Jaeger and Snider, 2013). Following Gries (2011), we make use of linear mixed effects models to determine salient word and sentence level factors that predict priming, to discover whether models display consistent behaviour with respect to one another and to human patterns of priming in production they may have learnt. We first describe the factors we use in our regression analysis in §6.1, and present the results in §6.2.

6.1 Priming Factors

We investigate the two broad categories of factors discussed in §2: *lexical dependence*, making use of the various conditions of the Prime-LM corpus (§3.3), and *inverse frequency*, choosing to focus on sentence-level surprisal and the structural preference of the verbs used.

Lexical Dependence We include pairwise token-level ***semantic similarity*** across prime and target content words, measured as the cosine similarity of the word embeddings taken from GPT2-large. We also include sentence-level similarity, based on the sentence embeddings of MPNet (Song et al., 2020), a high-performance sentence encoder. Here, we compute the cosine similarity between the PO prime and target embeddings. We add ***lexical overlap*** as a binary factor per token to our analysis. This allows us to separate overlap effects in conditions where multiple tokens overlap, which is not possible in corpus-level experiments.

Surprisal We include the ***surprisal*** of the congruent and incongruent prime and target, based on the negative log likelihood of the language model itself. Surprisal gives us a measure of how predictable or expected the sentences are as a whole, encompassing within-sentence collocation frequency effects.

Structural Preference Whereas corpus-based analysis of preferences is often based on normalised frequency statistics (Gries and Stefanowitsch, 2004), we base preference on the average probability difference of a verb in two alternating structures:

$$\text{PO-}pref(v) = \frac{1}{|\mathcal{V}|} \sum_{s \in \mathcal{V}} \log P(s^{\text{PO}}) - \log P(s^{\text{DO}}) \tag{5}$$

where $\mathcal{V}$ is the set of sentences containing verb v. This score expresses a verb's preference towards a particular structure on a scale from DO to PO. Hawkins et al. (2020) and Veenboer and Bloem (2023) provide a similar methodology for measuring structural preferences in LMs. For computing these scores we make use of the prime sentences from the *Core* condition of PrimeLM.

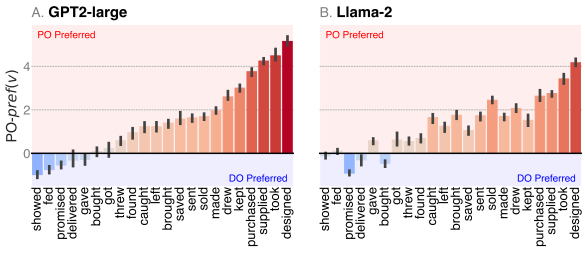

Figure 4: Structural preferences for GPT2-large and Llama-2, expressing the preference of a ditransitive verb with respect to a prepositional object versus a double object construction. The verb order of Llama-2 is based on the sorted order of GPT2-large.

The majority of verbs have a preference towards PO structure (as an example, Figure 4 contains the preferences of GPT2-large and Llama-2). This is not in line with dative usage preferences found in English, although it varies across vernaculars: some production and corpus studies suggest American English has a 2:1 preference towards DO constructions Bock and Griffin (2000); Grimm and Bresnan (2009), whereas Australian English has a PO preference (Bresnan and Ford, 2010). Preference towards PO in LMs may be confounded by transitive verb phrases followed by a prepositional modifier.

We also compute the Spearman correlations between the preference orders of all LMs and humans (Gries and Stefanowitsch, 2004), which reveals that there exists a high degree of variance across models and low correlation across models and human preference order (full figure in Appendix D). We leave a more thorough investigation of these differences open for future work that can take inspiration from established linguistic findings (Gropen et al., 1989; Arnold et al., 2000).

6.2 Modeling Priming Effects

Linear Mixed Model We fit a linear mixed model (LMM) using the factors of §6.1 that are added as fixed effects (Baayen et al., 2008). We fit two LMMs: one for predicting s_δ-PE(PO) and one for s_δ-PE(DO), which will provide insights whether different factors predict these effects. Fitting is done based on 30.000 items: 15.000 items are sampled for the *Core* condition, and 15.000 items are sample from the *Semantic Similarity* and *Lexical Overlap* conditions. This provides a balanced dataset of the *Core* and conditions that diverge from this baseline. We add a by-LM random intercept to account for individual model biases, akin to by-speaker random effects in human priming studies (Gries, 2011; Jaeger and Snider, 2013). All factors are centred and scaled to unit variance.

Results We report full LMM with coefficients in Figure 5, next to the reported effects found in the literature on human priming (z-scores and standard errors in Appendix E). The LMM reaches an R^2 of 0.257 (PO) and 0.227 (DO), which indicates that a large fraction of PE could still be predicted based on other factors and more complex interactions. We leave a more extensive exploratory analysis for future work, and focus on confirmatory hypothesis testing for now (Tukey, 1980; Barr et al., 2013).

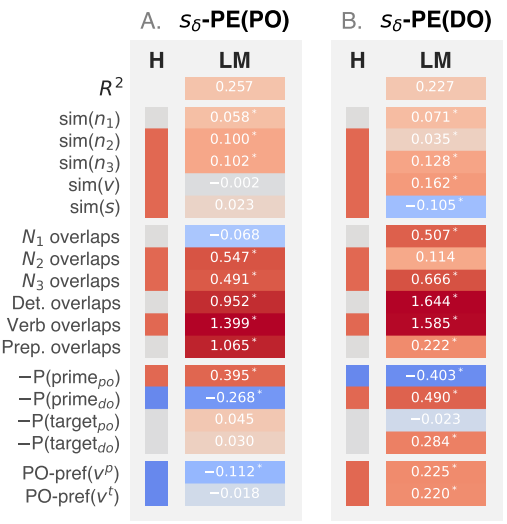

Figure 5: LMM coefficients for (A) predicting s_δ-PE(PO) and (B) s_δ-PE(DO), shown side-by-side with reported effects for predicting human priming in production- and corpus-based studies. Significant LLM coefficients ($p < 10^{-3}$) are denoted by an asterisk.

Semantic Similarity As in human production and corpus linguistics findings, we observe priming effects are predicted by semantic similarity between prime and target words (Snider, 2009), with the most consistent effects across structures for noun similarity (Cleland and Pickering, 2003), although the effect is relatively weak compared to other factors. Sentence similarity however, was not a predictive factor. In part this could be due to the sentence encoder not being sensitive to the structurally similar PO items that we use for computing sentence-level similarity. Incorporating the feature-based Grower similarity employed by Snider (2009) could be an alternative to explore the relation between sentence-level semantic similarity and priming.

Lexical Overlap We observe that lexical overlap is the strongest predictor of priming behaviour, in particular for primes sharing the same verbs, prepositions and determiners as the targets. This is in line with human findings; the meta analysis of Mahowald et al. (2016) shows lexical overlap is 'the most consistent moderator of syntactic priming'. We also observe that shared determiner overlap is consistently of high importance when predicting model PE, something observed but given far less attention in the human literature. Contrary to human findings in both production (e.g. Bock, 1989) and comprehension (e.g. Traxler, 2008), prepositional overlap is one of the strongest priming predictors. This indicates that priming in LMs is strongly driven by lexical cues, tying in with our observation in §5 that priming effects are highly influened by single token prediction, and this is driven more strongly by function than content words.

Surprisal Similar to Jaeger and Snider (2013) and Fazekas et al. (2024), we find that priming effect is predicted by prime surprisal, in both directions for PO and DO. This is evidence for an inverse frequency effect: a less frequent/plausible prime leads to an *increase* in priming effect. Target surprisal is less significant: only DO surprisal is a significant predictor.

Structural Preference We find that verb preference plays a highly predictive role, which again provides evidence for inverse frequency effects. A verb that has a structural preference for PO will lead to a higher DO priming effect, and vice versa, in line with results observed in human data (Gries and Wulff, 2005; Gries et al., 2005). This provides further explanation for the DO skewed priming effects that models display: for most of DO targets, their primes will *not* be in the preferred structure, thus boosting priming effects.

7 Discussion & Conclusion

In this paper, we seek to better understand the mechanisms that may underlie structural priming behaviour in LLMs. Borrowing insights from empirical and theoretical work on priming in humans, we investigate how, where and why a range of modern LLMs assign higher or lower probabilities to target sentences depending on preceding context, allowing us to investigate the extent to which language models are influenced by structure and semantics when making upcoming predictions.

Do models demonstrate structural priming? We find, in line with Sinclair et al. (2022), that models exhibit asymmetrical priming effects, and that this is even more pronounced in newer, larger LMs. By introducing a token-level priming effect we are able to locate more precisely potential sources of this asymmetry. We observe the direction of the asymmetry in PE is consistently *inverse* to priming effects in humans: where humans consistently display higher PE for the PO alternation, rather than DO, which we observe in models. We speculate that the verb preference effects we find in §6, which are predictive of PE as in humans, may play a role in this. Finally, through observing priming at the

token level, we observe that balanced priming in models is only visible from later on within the target, later than we may expect the same effects to be observed in humans.

How do humans and models compare? In predicting model PE using human factors known to predict priming in humans and in human corpora, we observe that lexico-semantic and frequency-related predictors of priming in humans also predict priming in LLMs. However, although we find lexico-semantic overlap of content words to be a reliable predictor of priming, we find that *function word overlap* plays a surprisingly predictive role, which has not been shown to be the case for humans with dative constructions (Bock, 1989; Pickering and Ferreira, 2008). Similar to human findings, we observe—consistently across models—inverse frequency effects of prime surprisal and verb preference (e.g. Gries and Wulff, 2005; Jaeger and Snider, 2008). This demonstrates that models are able to pick up on highly abstract factors influencing language predictions in humans from corpora, and highlights how influential seemingly small properties of the context are when it comes to upcoming model predictions.

What are the implications of implicit learning? We showed that priming is driven by similar inverse frequency effects observed in human priming. From a broader perspective, this is a striking finding. Inverse frequency effects have been argued to stem from an error-based implicit learning procedure (Chang et al., 2006): we adapt future predictions proportionally to recent predictive errors. This cognitive mechanism then leads to detectable patterns in human-produced corpus data (Jaeger and Snider, 2013), on which LLMs are trained. LLMs are thus able to pick up on this highly abstract pattern, which shows that their priming behaviour is far more intricate than a simple repetition-based mechanism. An interesting endeavour for future work would be to test this finding in a setting with control over data distribution (e.g. Jumelet and Zuidema (2023b)), to ensure that inverse frequency effects do not stem from some other indirect effect of language learning.

Comprehension and production in LLMs We build on Sinclair et al. (2022), who design the priming effect metric to measure *comprehension* in LLMs forcing its prediction on a fixed target without allowing for open-ended production. It is important to remember, however, that through the way LLMs are trained, their predictions will be driven by human *production* patterns in the training data, thus motivating our choice to base our predictions on findings from corpus and production studies. Although the mechanisms for comprehension and production in LLMs are highly linked—they rely on the same output distribution—it would be interesting to investigate priming in LLMs in a generation-based setting as well. A thorough investigation in this direction may provide deeper insights into the relation between LLM behaviour and cognitive theories of human language processing (e.g., Dell and Chang, 2014).

Outlook Language models as cognitive models can potentially aid in discovering important properties of human linguistic behaviour (Futrell et al., 2019; Linzen and Baroni, 2021; Wilcox et al., 2023); we thus view our results as a contribution to defining the border where human patterns are replicated in models. Looking outwards from this detailed analysis, future studies could investigate the extent to which these priming effects influence structural repetition patterns in generation, complementing existing work finding priming-like lexical repetition effects in LLM generation (Molnar et al., 2023). Furthermore, a more detailed investigation in the exact nature of the (potentially) hierarchical representations that underlie priming behaviour could take inspiration from parsing-based theories of priming (Prasad and Linzen, 2024), or deploy techniques from interpretability research to uncover hierarchical structure (Murty et al., 2023; Jumelet and Zuidema, 2023a). Not only will such investigations provide deeper insights into the cognitive plausibility of LLMs (Beinborn and Hollenstein, 2023), but it may also yield a better understanding of the mechanisms underlying in-context learning (Min et al., 2022; Han et al., 2023).

Limitations

One clear limitation of our work stems from the specific nature of our analyses. More work remains to be done to investigate whether these results generalise across other constructions within English or further extend to other languages. It also remains an open question whether the constraints within the dataset we use influence our outcomes: do these results generalise to a wider range of vocabulary, or a more complex set of sentences. Additionally, even within the scope of our analyses, the effects of

model size on the results we observe are interesting, but would need to be tested more systematically to draw firm conclusions from them. Likewise, while we purposefully include a selection of base models and their alignment tuned variants to investigate whether there are any striking differences, the sample is too small to make any meaningful inference.

Acknowledgements

We would like to thank our three anonymous reviewers and meta-reviewer for their thoughtful comments and discussion, which helped refine this paper. We would like to thank Raquel Fernández for her fundamental role in the paper upon which this paper builds. AS would like to thank colleagues Anastasia Klimovich-Gray and Agnieszka Konopka from the department of Psychology for their useful discussions and psycholinguistic perspective. JJ would like to thank Jakub Szymanik, Roberto Zamparelli, and Alexey Koshevoy for insightful discussions during his Trento visit.

References

Liz Allen, Alison O'Connell, and Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? how the contributor role taxonomy (credit) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1):71–74.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *CoRR*, abs/2311.16867.

Jennifer E. Arnold, Anthony Losongco, Thomas Wasow, and Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, 76(1):28–55.

R.H. Baayen, D.J. Davidson, and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412. Special Issue: Emerging Data Analysis.

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.

Lisa Beinborn and Nora Hollenstein. 2023. *Cognitive Plausibility in Natural Language Processing*. Springer Nature.

Sarah Bernolet and Robert J. Hartsuiker. 2010. Does verb bias modulate syntactic priming? *Cognition*, 114(3):455–461.

Kathryn Bock. 1989. Closed-class immanence in sentence production. *Cognition*, 31(2):163–186.

Kathryn Bock and Zenzi M Griffin. 2000. The persistence of structural priming: Transient activation or implicit learning? *Journal of experimental psychology: General*, 129(2):177.

Holly P. Branigan, Martin J. Pickering, and Alexandra A. Cleland. 1999. Syntactic priming in written production: Evidence for rapid decay. *Psychonomic Bulletin & Review*, 6(4):635–640.

Joan Bresnan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen. 2007. Predicting the dative alternation. In *Cognitive foundations of interpretation*, pages 69–94. KNAW.

Joan Bresnan and Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in american and australian varieties of english. *Language*, 86(1):168–213.

Franklin Chang, Gary S. Dell, and Kathryn Bock. 2006. Becoming syntactic. *Psychological review*, 113 2:234–72.

Christine Chiarello, Curt Burgess, Lorie Richards, and Alma Pollock. 1990. Semantic and associative priming in the cerebral hemispheres: Some words do, some words don't... sometimes, some places. *Brain and language*, 38(1):75–104.

Alexandra A Cleland and Martin J Pickering. 2003. The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, 49(2):214–230.

Gary Dell and Franklin Chang. 2014. The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 369:20120394.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Judit Fazekas, Giovanni Sala, and Julian Pine. 2024. Prime surprisal as a tool for assessing error-based learning theories: A systematic review. *Languages*, 9(4).

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Jeffrey Gerard, Frank Keller, and Themis Palpanas. 2010. Corpus evidence for age effects on priming in child language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Georgia M. Green. 1974. Semantics and syntactic regularity. Bloomington: Indiana University Press.

Stefan Gries. 2011. Studying syntactic priming in corpora. *Converging evidence: Methodological and theoretical issues for linguistic research*, pages 143–165.

Stefan Gries and Anatol Stefanowitsch. 2004. Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics*, 9:97–129.

Stefan Gries and Stefanie Wulff. 2005. Do foreign language learners also have constructions? *Annual Review of Cognitive Linguistics*, 3:182–200.

Stefan Th. Gries. 2005. Syntactic priming: A corpus-based approach. *Journal of psycholinguistic research*, 34(4):365–399.

Stefan Th. Gries, Beate Hampe, and Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, 16(4):635–676.

Stefan Th. Gries and Gerrit Jan Kootstra. 2017. Structural priming within and across languages: a corpus-based perspective. *Bilingualism: Language and Cognition*, 20(2):235–250.

Scott Grimm and Joan Bresnan. 2009. Spatiotemporal variation in the dative alternation: A study of four corpora of british and american english. In *Third International Conference Grammar and Corpora*, pages 22–24.

Jess Gropen, Steven Pinker, Michelle Hollander, Richard Goldberg, and Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language*, 65(2):203–257.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics.

Mary L Hare and Adele E Goldberg. 1999. Structural priming: Purely syntactic? In *Proceedings of the twenty first annual conference of the Cognitive Science Society*, pages 208–211. Psychology Press.

Robert J. Hartsuiker, Sarah Bernolet, Sofie Schoonbaert, Sara Speybroeck, and Dieter Vanderelst. 2008. Syntactic priming persists while the lexical boost decays: Evidence from written and spoken dialogue. *Journal of Memory and Language*, 58(2):214–238.

Robert Hawkins, Takateru Yamakoshi, Thomas Griffiths, and Adele Goldberg. 2020. Investigating representations of verb bias in neural language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4653–4663, Online. Association for Computational Linguistics.

T. Florian Jaeger and Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *The 30th Annual Meeting of the Cognitive Science Society (CogSci08)*, pages 1061–1066, Washington, D.C.

T. Florian Jaeger and Neal E. Snider. 2013. Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, 127(1):57–83.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Michael N Jones, Walter Kintsch, and Douglas JK Mewhort. 2006. High-dimensional semantic space accounts of priming. *Journal of memory and language*, 55(4):534–552.

Jaap Jumelet. 2020. diagNNose: A library for neural activation analysis. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 342–350, Online. Association for Computational Linguistics.

Jaap Jumelet and Willem Zuidema. 2023a. Feature interactions reveal linguistic structure in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8697–8712, Toronto, Canada. Association for Computational Linguistics.

Jaap Jumelet and Willem Zuidema. 2023b. Transparency at the source: Evaluating and interpreting language models with access to the true distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4354–4369, Singapore. Association for Computational Linguistics.

Michael P Kaschak, Timothy J Kutta, and John L Jones. 2011. Structural priming as implicit learning: Cumulative priming effects and individual differences. *Psychonomic bulletin & review*, 18:1133–1139.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212.

Margery Lucas. 2000. Semantic priming without association: A meta-analytic review. *Psychonomic bulletin & review*, 7:618–630.

Kyle Mahowald, Ariel James, Richard Futrell, and Edward Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91:5–27.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Aron Molnar, Jaap Jumelet, Mario Giulianelli, and Arabella Sinclair. 2023. Attribution and alignment: Effects of local context repetition on utterance production and comprehension in dialogue. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 254–273.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D. Manning. 2023. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Martin J Pickering and Holly P Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and language*, 39(4):633–651.

Martin J. Pickering and Victor S. Ferreira. 2008. Structural priming: A critical review. *Psychological Bulletin*, 134(3):427.

Grusha Prasad and Tal Linzen. 2024. Spawning structural priming predictions from a cognitively motivated parser. *arXiv preprint arXiv:2403.07202*.

Grusha Prasad, Marten van Schijndel, and Tal Linzen. 2019. Using priming to uncover the organization of syntactic representations in neural language models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 66–76.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

David Reitter, Frank Keller, and Johanna D Moore. 2011. A computational cognitive model of syntactic priming. *Cognitive science*, 35(4):587–637.

Katrien Segaert, Gerard Kempen, Karl Magnus Petersson, and Peter Hagoort. 2013. Syntactic priming and the lexical boost effect during sentence production and sentence comprehension: An fmri study. *Brain and Language*, 124(2):174–183.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2023. Language model acceptability judgements are not always robust to context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6043–6063, Toronto, Canada. Association for Computational Linguistics.

Neal Snider. 2009. Similarity and structural priming. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31, pages 925–937.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sandra A. Thompson and Yuka Koide. 1987. Iconicity and 'indirect objects' in english. *Journal of Pragmatics*, 11(3):399–406.

Kristen M Tooley. 2023. Structural priming during comprehension: A pattern from many pieces. *Psychonomic Bulletin & Review*, 30(3):882–896.

Kristen M. Tooley and Kathryn Bock. 2014. On the parity of structural persistence in language production and comprehension. *Cognition*, 132(2):101–136.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Matthew J Traxler. 2008. Structural priming among prepositional phrases: Evidence from eye movements. *Memory & Cognition*, 36(3):659–674.

Matthew J Traxler, Kristen M Tooley, and Martin J Pickering. 2014. Syntactic priming during sentence comprehension: evidence for the lexical boost. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40(4):905.

Jean E Fox Tree and Paul JA Meijer. 1999. Building syntactic structure in speaking. *Journal of Psycholinguistic Research*, 28:71–90.

John W. Tukey. 1980. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of LM alignment. *CoRR*, abs/2310.16944.

Marten van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710, Brussels, Belgium. Association for Computational Linguistics.

Tim Veenboer and Jelke Bloem. 2023. Using collostructional analysis to evaluate BERT's representation of linguistic constructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12937–12951, Toronto, Canada. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

Bushi Xiao, Chao Gao, and Demi Zhang. 2024. Modeling bilingual sentence processing: Evaluating rnn and transformer architectures for cross-language structural priming. *arXiv preprint arXiv:2405.09508*.

Zhenghao Zhou and Robert Frank. 2023. What affects priming strength? simulating structural priming effect with pips. *Proceedings of the Society for Computation in Linguistics*, 6(1):413–417.

Jayden Ziegler and Jesse Snedeker. 2019. The use of syntax and information structure during language comprehension: Evidence from structural priming. *Language, Cognition and Neuroscience*, 34(3):365–384.

A Author Contributions

Following the Contributor Role Taxonomy (CRediT; Allen et al. 2019):

	JJ	WZ	AS
Conceptualisation	✓	✓	✓
Methodology	✓		✓
Software	✓		
Data Curation	✓		✓
Investigation	✓		
Visualisation	✓		
Analysis	✓		✓
Writing - Original Draft	✓		✓
Writing - Review & Editing	✓	✓	✓
Supervision		✓	✓

B Sentence-level asymmetry effects

The increased semantic similarity, in Figure 6, can be seen to primarily have an effect on boosting the PE score of the dominant structure (DO), while leaving the PE score of the opposite structure unaffected (PO). Furthermore, it can be seen that the the effect of increasing the similarity of nouns and verbs is not *linearly additive*: increasing similarity of both nouns and verbs has a far greater impact than the individual conditions combined.

C Token-level PE for Increased Semantic Similarity

The token-level PE scores for the 3 conditions with increased semantic similarity are shown in Figure 7.

D Preference Order Correlation

The Spearman correlation between LMs and human structural preference order are shown in Figure 9. Surprisingly, there exists a high degree of variation in preference order, both within models and across

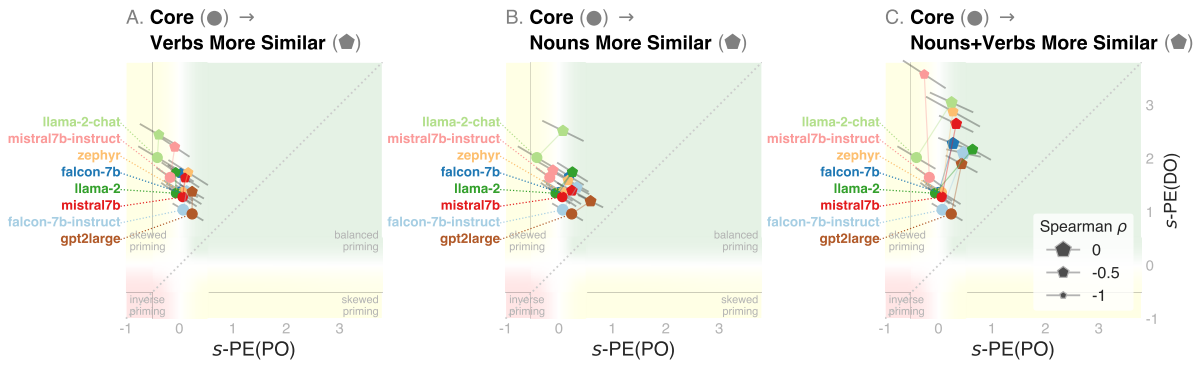

Figure 6: Sentence-level Priming Effects for conditions with an increased semantic similarity across nouns and verbs.

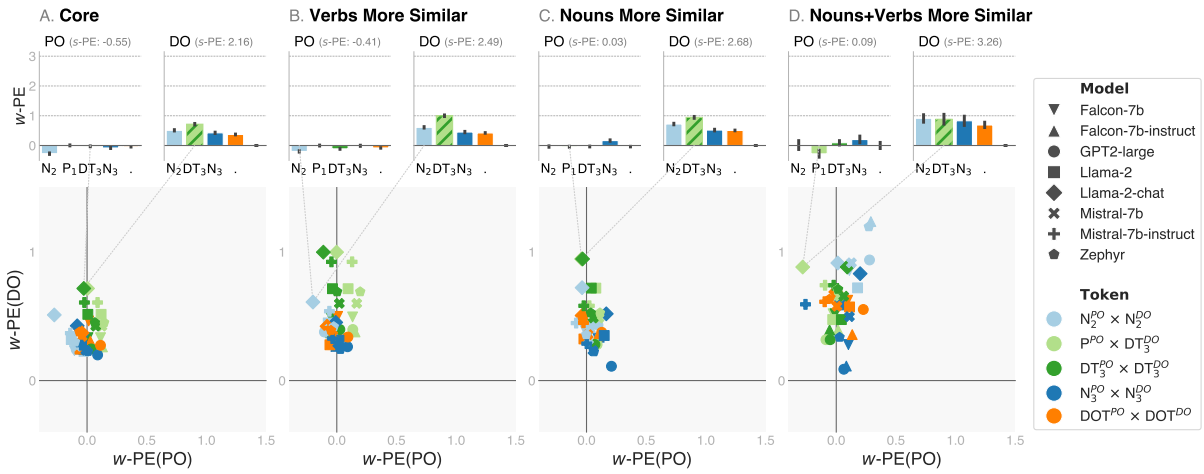

Figure 7: w-PE scores for increased semantic similarity between prime and target.

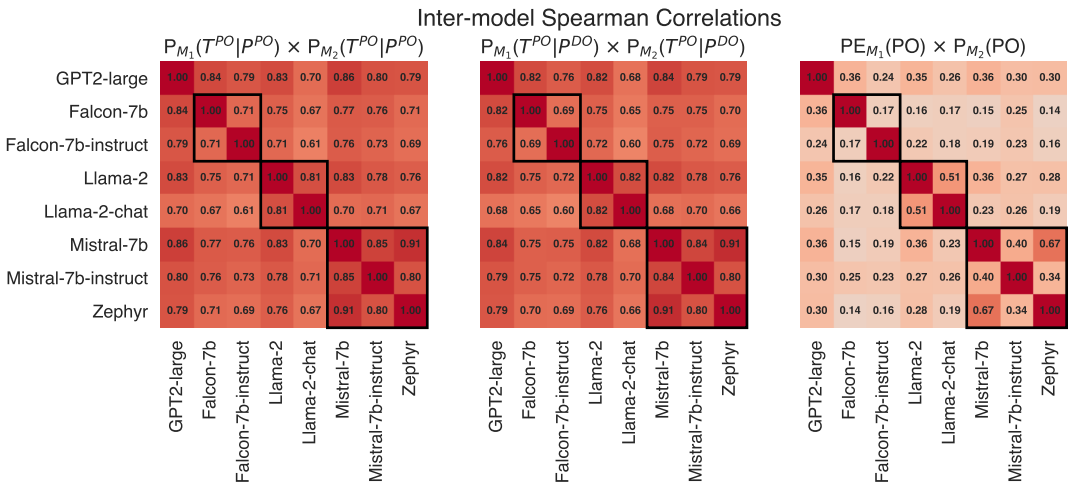

Figure 8: Correlations between LMs.

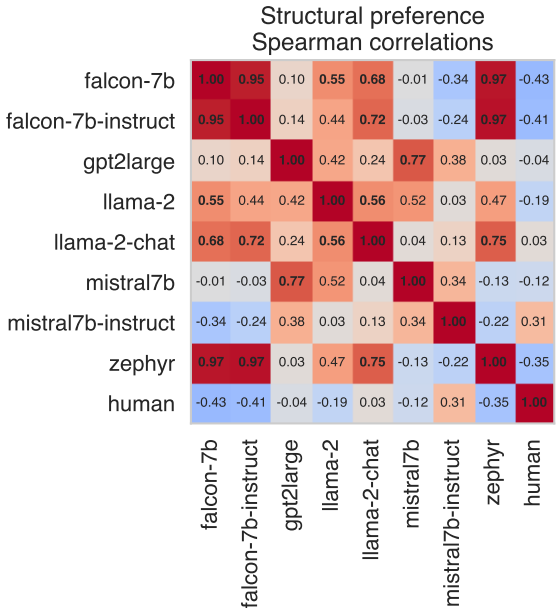

Figure 9: The Spearman correlation between LMs and human structural preference order.

models and human preference. Only *Falcon-7b* and its instruction-tuned variant retain a high preference overlap; all the other aligned LMs diverge quite strongly from their base model. None of the LMs have a significant correlation with respect to the reported human preference order, which is in contrast to Hawkins et al. (2020)'s positive findings of strong correlations between model and human preferences. We leave a more thorough investigation of these differences open for future work.

E Linear Mixed-effect Model Summary

The LMM results including coefficients, standard error, z-score and p-values are shown in Table 1 and 2.

F Model Size and Alignment

Interestingly, although the sample is too small to make broad generalisations, in the models we test, we observe larger models exhibit more *skewed* priming behaviour in the core, and higher susceptibility to lexical boosting than the smaller GPT2. We also observe no strong patterns to distinguish alignment tuning from base models, in fact, one surprising finding is the degree of difference in PE for a given prime target pair that a base and alignment model from the same base will have (See Figure 8, which has the models grouped by base model).

0	Model:	MixedLM	Dependent Variable:	s_δ-PE(PO)
1	No. Observations:	30000	Method:	REML
2	No. Groups:	8	Scale:	1.8203
3	Min. group size:	3655	Log-Likelihood:	-51620.4121
4	Max. group size:	3858	Converged:	No
5	Mean group size:	3750.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-0.072	0.111	-0.652	0.515	-0.289	0.145
sim(n_1)	0.041	0.010	4.260	0.000	0.022	0.060
sim(n_2)	0.097	0.009	10.381	0.000	0.079	0.115
sim(n_3)	0.109	0.009	11.652	0.000	0.091	0.127
sim(v)	0.010	0.009	1.092	0.275	-0.008	0.027
sim(s)	-0.000	0.017	-0.008	0.994	-0.034	0.034
N_1 overlaps	-0.148	0.052	-2.850	0.004	-0.250	-0.046
N_2 overlaps	0.696	0.056	12.510	0.000	0.587	0.805
N_3 overlaps	0.473	0.050	9.405	0.000	0.374	0.571
Det. overlaps	1.010	0.026	38.149	0.000	0.958	1.062
Verb overlaps	1.491	0.046	32.578	0.000	1.401	1.581
Prep. overlaps	1.025	0.027	38.574	0.000	0.973	1.077
$-$P(prime$_{po}$)	0.390	0.023	16.788	0.000	0.345	0.436
$-$P(prime$_{do}$)	-0.259	0.025	-10.504	0.000	-0.307	-0.210
$-$P(target$_{po}$)	0.044	0.023	1.885	0.059	-0.002	0.089
$-$P(target$_{do}$)	0.055	0.025	2.223	0.026	0.006	0.103
PO-pref(v^p)	-0.129	0.010	-13.377	0.000	-0.148	-0.110
PO-pref(v^t)	-0.029	0.010	-3.041	0.002	-0.048	-0.010
Group Var	0.097	0.061				

Table 1: Raw LMM results for predicting s_δ-PE(PO).

0	Model:	MixedLM	Dependent Variable:	s_δ-PE(DO)
1	No. Observations:	30000	Method:	REML
2	No. Groups:	8	Scale:	2.2337
3	Min. group size:	3655	Log-Likelihood:	-54688.2103
4	Max. group size:	3858	Converged:	No
5	Mean group size:	3750.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	1.339	0.111	12.047	0.000	1.121	1.557
sim(n_1)	0.071	0.011	6.701	0.000	0.050	0.092
sim(n_2)	0.021	0.010	2.018	0.044	0.001	0.041
sim(n_3)	0.122	0.010	11.826	0.000	0.102	0.142
sim(v)	0.171	0.010	17.197	0.000	0.151	0.190
sim(s)	-0.044	0.019	-2.331	0.020	-0.082	-0.007
N_1 overlaps	0.456	0.057	7.966	0.000	0.344	0.568
N_2 overlaps	-0.146	0.061	-2.384	0.017	-0.266	-0.026
N_3 overlaps	0.717	0.055	12.947	0.000	0.608	0.826
Det. overlaps	1.671	0.029	57.289	0.000	1.614	1.728
Verb overlaps	1.535	0.050	30.433	0.000	1.436	1.634
Prep. overlaps	0.244	0.029	8.349	0.000	0.187	0.302
$-$P(prime$_{po}$)	-0.333	0.026	-12.989	0.000	-0.383	-0.283
$-$P(prime$_{do}$)	0.416	0.027	15.332	0.000	0.363	0.469
$-$P(target$_{po}$)	-0.045	0.026	-1.759	0.079	-0.095	0.005
$-$P(target$_{do}$)	0.311	0.027	11.452	0.000	0.258	0.365
PO-pref(v^p)	0.231	0.011	21.761	0.000	0.210	0.252
PO-pref(v^t)	0.215	0.011	20.262	0.000	0.194	0.236
Group Var	0.097	0.031				

Table 2: Raw LMM results for predicting s_δ-PE(DO).