Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown the emergent capability of in-context learning (ICL). One line of research has claimed that ICL is functionally equivalent to gradient descent, a type of error-driven learning mechanism. In this paper, we introduce a new way of diagnosing whether ICL is functionally performing error-driven learning. Our approach is based on the *inverse frequency effect* (IFE)—a phenomenon in which an agent's behavior is influenced to a greater degree when presented with improbable examples as compared to more likely ones. The IFE has previously been identified in psycholinguistics where humans exhibit the IFE in the context of structural priming (the tendency for people to produce sentence structures they have encountered recently). In that context, the IFE has been used as evidence that human structural priming must involve error-driven learning mechanisms. In our experiments, we simulated structural priming with ICL and found that LLMs indeed display the IFE, with the effect being stronger in larger models. We conclude that at least in the case we studied, ICL is indeed a type of error-driven learning, supporting the hypothesis that an error signal is implicitly computed in the forward pass during ICL. Our results suggest that both humans and LLMs make use of error-driven processing mechanisms in on-line processing.

1 Introduction

To what extent do humans and language models use similar processing mechanisms? On the one hand, language processing by language models and human learners appears to be substantially different. Humans, unlike LLMs, display the ability to flexibly adapt to new contexts. And language models typically require massive amounts of training data and a large number of parameters to exhibit human-like performance. Yet recent pre-trained large language models (LLMs) have shown the capacity for in-context learning (ICL): they flexibly adapt to novel tasks with only a small number of demonstrations provided as prompts in the context window without any parameter updates (Brown et al., 2020). This emergent capability could provide a way to bridge the divide between language models and humans: perhaps ICL is a processing mechanism that can flexibly adapt to novel contexts without the need for substantial training.

Within the body of research about why ICL arises and how it operates, one line of work has aimed to deepen our theoretical understanding of ICL by offering *functional* interpretations of ICL as a kind of implicit gradient descent during inference. Von Oswald et al. (2023) demonstrate that Transformer models, with appropriate choices of parameters, *can* process in-context demonstrations in a way that is functionally equivalent to performing gradient updates on the same demonstration examples. Garg et al. (2022), Zhang et al. (2024), and Ahn et al. (2024) show that standard Transformers (Vaswani et al., 2017) can be trained to implement learning algorithms for linear regressions under ICL-based training objectives. Dai et al. (2023) provide a mathematical construction showing a dual form between Transformer attention and gradient descent and interpreted ICL as a meta-optimization process that performs implicit fine-tuning. However, Shen et al. (2023) argue that the importance of these theoretical demonstrations is limited in that they do not take ICL to be an emergent property, instead assuming a training objective that optimizes for ICL. These demonstrations therefore deviate from actual LLMs pre-trained with natural data with a language modeling training objective. Indeed, Shen et al. find inconsistencies between ICL and gradient descent in real models, and therefore leave the equivalence between ICL and gradient descent as an open question.

In this paper, we aim to better characterize *what kind of learning mechanism ICL is* by drawing a connection between ICL and human learning mech-

anisms, using a case study that allows us to evaluate off-the-shelf LLMs using natural language data. Specifically, we investigate **whether ICL is a type of *error-driven learning* such that an error signal is implicitly computed during the forward pass**. By focusing on this question, we take some first steps toward understanding whether ICL involves implicit gradient descent because gradient descent is one type of error-driven learning. To approach this question, we treat ICL as a processing mechanism of LLMs and borrow insights from methods of studying processing mechanisms in humans: we examine to what extent LLMs show the *inverse frequency effect* (IFE), a phenomenon in human structural priming (Branigan and Pickering, 2017) that has been argued to require an error-driven processing mechanism in humans, *implicit learning* (e.g., Chang et al., 2006). We focus our attention on the widely studied linguistic phenomenon of the dative alternation where the IFE has been robustly attested in humans. We demonstrate that LLMs not only show a robust IFE under standard fine-tuning (which involves explicit gradient descent, a type of error-driven learning), but also varying degrees of the IFE under the ICL setting, with larger models showing a stronger IFE. Given the correlation between ICL ability and model size, we use our results to conclude that, at least in the cases we study, ICL is indeed an error-driven learning mechanism.

Overall, our results suggest that error-driven learning is an aspect of processing shared by humans and ICL-using LLMs. Our study has implications for both NLP/ML (1, 2, & 3) and linguistically-motivated analysis of LLMs (4 & 5):

(1) We find evidence that ICL in off-the-shelf LLMs can be viewed as a form of error-driven learning in at least some cases.

(2) We show how the IFE can serve as a diagnostic for error-driven learning, paving the way for future work that can use this diagnostic to further investigate the circumstances under which ICL functions like error-driven learning.

(3) We generalize the notion of ICL beyond the standardly assumed prompt format of input-output pairs, establishing a connection between priming and prompting.

(4) We show that LLMs qualitatively display an important property of human language processing: the IFE in structural priming.

(5) While most human-LLM comparisons focus on representations, our experiments go one step further by analyzing the processing mechanisms used by LLMs.

2 Background and Related Work

In this section, we lay out the building blocks necessary for motivating why we use the IFE to diagnose the error-driven nature of ICL. Our experimental approach is formally described in Section 3.1.

2.1 Structural Priming in Psycholinguistics

Encountering a syntactic structure can predispose speakers to repeat that structure in the near future, a phenomenon known as syntactic priming (Bock, 1986). For example, after encountering a double object (DO) sentence (e.g., *Alice gave Bob a book*), speakers tend to produce another DO structure (e.g., *The student sent the professor a letter*) rather than a semantically equivalent prepositional dative (PD) structure (e.g., *The student sent a letter to the professor*). Structural priming has been interpreted as an *adaptation* mechanism, where speakers adapt lexical and syntactic predictions to the current context (Jaeger and Snider, 2013), similar to the way in which LLMs adapt their outputs on the basis of prompts. Examples of priming typically involve two sentences; we will refer to them as the *prime sentence* and the *target sentence*, where the prime sentence comes first and influences the form of the target sentence.

One important aspect of structural priming is the *inverse frequency effect* (Jaeger and Snider, 2008; Bernolet and Hartsuiker, 2010; Kaschak et al., 2011): less-preferred syntactic alternatives cause stronger overall priming than more-preferred structures, where the degree to which a structure is preferred is operationalized as its relative frequency in the speaker's experience. This can be seen by looking at sentences involving verbs with gradient structural preferences, or *verb biases* (or alternation biases; see Hawkins et al., 2020 for a systematic investigation of verb biases in neural models). The verb *give* allows both prepositional dative (PD) (*The doctor gave a book to the judge*) and double object (DO) (*The doctor gave the judge a book*) structures, but is biased toward (i.e., occurs more often with) DO in English. Under the IFE, a "prime" sentence involving *give* in its (less preferred) PD structure will cause a greater priming effect than a DO sentence, i.e., a PD prime will

increase the probability of a PD occurring more than a DO prime will increase the probability of a DO occurring. The strength of PD priming (i.e., the increase in the probability of a PD target given a PD prime) inversely correlates with the expectation of a PD prime, determined by its verb bias (Bernolet and Hartsuiker, 2010).

Two theories have been proposed to account for structural priming. *Transient activation theory* (Pickering and Branigan, 1998) claims that the activation of structural representations from the prime persists for a short time (in working memory) so that it is easier to reactivate the same structure at the next relevant opportunity. This form of transient activation theory does not however account for the IFE because the amount of residual activation is independent from verb biases and does not involve any error-driven mechanism. In contrast, *implicit learning theory* (Chang et al., 2006) claims that humans are continuously and dynamically adjusting their probabilistic knowledge concerning the occurrence of grammatical structures (including verb biases) on the basis of their experience and use such information to predict the form of linguistic input. Crucially, under standard theories of learning, the update performed by the learner is error-driven, such that a larger update is performed in situations where the learner's predictions are farther from the truth. In the context of priming, this would mean that priming strength is determined by the difference between the learner's predictions and the actual prime sentence: the less the learner expects the observed prime structure, the larger the resulting gradient is that updates their expectations for that structure in the future, yielding a larger priming effect. Therefore, implicit learning, unlike transient activation, predicts the IFE. The two theories are not mutually exclusive and can co-exist to account for priming, a framing known as the dual mechanism account (Tooley and Traxler, 2010).

In this study, we assume the correctness of the arguments from psycholinguistics that error-driven learning mechanisms are necessary to explain the IFE. Therefore, by examining whether LLMs show the IFE in the ICL setting, we can infer whether there is an implicit error signal computed in ICL.

2.2 Structural Priming in Neural Language Models

Structural priming has been used as a tool for probing the mental representations of structure in humans, under the assumption that these representations define the notion of similarity underlying priming (Branigan and Pickering, 2017). Previous works have adopted this paradigm for probing learned linguistic representations in neural networks. van Schijndel and Linzen (2018) and Prasad et al. (2019) show that LSTMs (Gulordava et al., 2018) are capable of adapting to syntactic structures under fine-tuning. Specifically, these researchers **fine-tune** model weights on prime sentences and test target sentence probabilities in the updated model. This is a direct analog of the implicit learning account of structural priming as it *involves weight updates*. Recently, Sinclair et al. (2022) have shown that the GPT2 family (Radford et al., 2019) shows robust structural priming without the use of explicit weight updates; when a prime sentence is **concatenated** to its corresponding target sentence, the model assigns a higher probability to the target sentence than it does when the target sentence appears without a prime. Other works have demonstrated crosslingual structural priming in large language models (Michaelov et al., 2023), suggesting that structural priming is robustly detected in LLMs. In concurrent work, Jumelet et al. (2024) study in considerable detail the linguistic factors the give rise to structural priming in current state-of-the-art LLMs, including presenting evidence that LLMs display the IFE; our work differs from theirs in that we study the IFE as a diagnostic for error-driven learning, while Jumelet et al.'s goal is to compare LLM priming with human priming. In sum, there is a large body of work demonstrating that LLMs, like humans, show structural priming. This sets the stage for our current study of investigating the processing mechanisms underlying priming.

3 Current Study

3.1 Overview of Our Approach

We first clarify our conceptualization of ICL. In most of the literature, ICL is conceptualized as involving explicit demonstration-answer pairs. However, Chen et al. (2024) find that parallel structures in pre-training data substantially contribute to ICL capabilities in LLMs, where parallel structures are defined as pairs of phrases that appear in the same context window and that follow similar templates. Although these parallel structures are not explicitly framed as input-output pairs for a task, they can be conceptualized as in-context examples of less-structured implicit "tasks" such as copy-

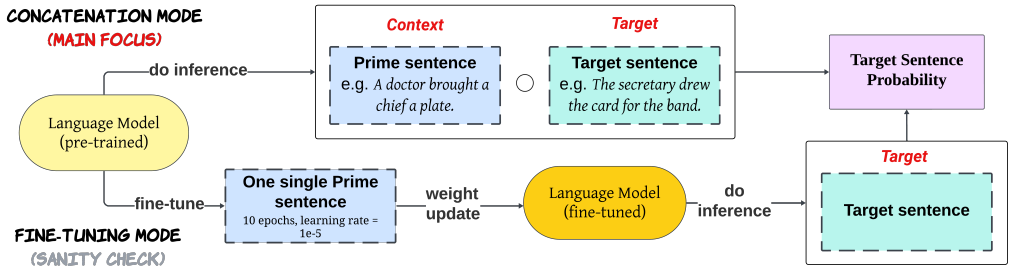

Figure 1: An overview of our experimental design.

ing n-grams, repeating syntactic constructions, and producing related types of world knowledge. Inspired by this perspective, we conceptualize ICL as a more generalized notion that involves *a sensitivity to parallelism* through generic next-token prediction: any text in the context window will influence the model's logits toward preferring similar text. This generalized notion of ICL, namely, the contextual influence of prompts in the context window, is analogous to priming in humans. The fact that Chen et al. (2024) showed that parallel structures in pre-training substantially contribute to the emergence of ICL supports the view that we can frame ICL in these more general, parallelism-focused terms. Under this generalized framing, ICL can produce structural priming: a model that observes a sentence with a particular syntactic structure (say, the DO structure) can use ICL to replicate that syntactic structure, even though the sentence cannot be cleanly viewed as a demonstration-answer pair. That is, the less structured, implicitly defined "task" encoded by the DO sentence used as the prompt could be interpreted as "producing another sentence following the DO structural template exemplified in the prompt."

Our research question is **whether an error signal is implicitly computed during the forward pass of processing the prompt in the generalized ICL setting**. We investigate this question by testing whether LLMs show the IFE in the ICL setting. Given that (i) it has been argued by psycholinguists that only error-driven learning mechanisms can give rise to the IFE; (ii) processing the prime sentence in the context window conditions the probability of the target sentence in our generalized notion of ICL; and (iii) standard structural priming in the ICL setting has been robustly observed, we hypothesize that the strength of the IFE will positively correlate with the strength of the ICL capability of LLMs: the stronger the ICL capability is, the better the error signal will be computed in the forward pass, leading to a stronger IFE.

We simulate structural priming across LLMs of various sizes with the two previously explored methods mentioned in Section 2.2 (see Figure 1). The **`Fine-Tuning`** mode updates the LLM's parameters on the basis of a single prime sentence, and the updated model is used to infer the probability of the target sentence. The **`Concatenation`** mode instead uses (our generalized notion of) ICL, where the prime sentence is concatenated with the target sentence, so that the prime constitutes the prompt/context at the point in which the probability of the target sentence is measured. The `Fine-Tuning` mode serves as a sanity check that LLMs are able to show the IFE with **explicit** error-driven learning. Because of the nature of gradient descent, changes to model weights during fine-tuning are a function of the degree to which a prime sentence is correctly predicted. A more unexpected prime will therefore give rise to larger gradients and updates, as well as a larger priming effect. Once we have demonstrated this result empirically, the stage will then be set for our main focus: using the `Concatenation` mode to diagnose whether ICL involves **implicit** error-driven learning.

3.2 Corpus

We adapted the *Core Dative* PRIME-LM Corpus from Sinclair et al. (2022) to create our dataset. We briefly introduce the relevant properties and refer the readers to the original paper for details. The dative corpus consists of sentences in two forms:

(6) **DO**: $\texttt{DP}_{subj}$ `V` $\texttt{DP}_{iobj}$ $\texttt{DP}_{dobj}$
e.g., *A girl bought a guy a coffee*.

(7) **PD**: $\texttt{DP}_{subj}$ `V` $\texttt{DP}_{dobj}$ `Prep` $\texttt{DP}_{iobj}$
e.g., *A girl bought a coffee for a guy*.

Each `DP` is a determiner with a common noun (120 distinct nouns in total). The corpus was constructed in a way that ensures semantic plausibility and

controls for the degree of semantic association and lexical overlap between prime and target sentences.

Since our goal is to study the IFE, we need to ensure variation in the occurrence of verbs with different verb biases in the the prime and target sentences. To do this, for each prime sentence involving one of 22 possible verbs, we sampled 50 target sentences with no lexical overlap. This means that the verbs in prime and target sentences are always distinct. Each prime-target pair can be instantiated structurally in 4 different ways, depending on whether the target and prime sentences are PD or DO: $t_{\text{PD}}|p_{\text{PD}}$, $t_{\text{PD}}|p_{\text{DO}}$, $t_{\text{DO}}|p_{\text{PD}}$, $t_{\text{DO}}|p_{\text{DO}}$ (i.e., target sentence t conditioned on prime p). This resulted in 92400 prime-target pairs.[1] An example of $t_{\text{PD}}|p_{\text{DO}}$ is *"A doctor brought a chief a plate. The secretary drew the card for the band."*

We also created an alternative dataset of the same size by replacing all indirect object DPs with a pronoun. This was motivated by a corpus study (see Appendix A) that showed that the most common indirect objects in DO sentences are animate pronouns, suggesting that it may be important to consider pronouns in analyses of DO vs. PD sentences. The importance of animacy has also previously been noted by Bresnan et al. (2007). The presence and absence of pronouns lead to different verb biases for LLMs, which affect their IFE behaviors. We return to this point in the discussion.

3.3 Language Models

We study a set of Transformer models that have been claimed to show ICL capabilities to varying extents (Lee et al., 2024). First, we used **GPT2** (Radford et al., 2019) in three of its sizes (SMALL, MEDIUM, LARGE), with 85M, 302M, and 708M parameters, respectively. All versions were loaded from the package `transformerLens` (Nanda and Bloom, 2022). Second, we used **Llama2** (Touvron et al., 2023) in three versions: 7B (6.5B parameters), 7B-CHAT (6.5B parameters), and 13B (13B parameters). All versions were loaded using the Huggingface `transformers` library (Wolf et al., 2020). Finally, we used **GPT3-base** (Brown et al., 2020) with the DAVINCI-002 version (175B parameters), accessed via the OpenAI API. The models vary in size, and correspondingly, in their ICL capabilities. We predict that there will be a stronger IFE as size increases.[2]

[1] We use t and p for individual target and prime sentences, T and P for sets of targets and primes, and $\mathcal{P}$ for probability.

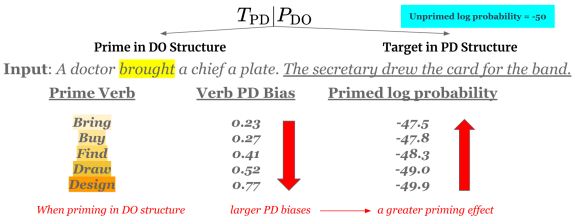

Figure 2: A demonstration of the IFE: a stronger priming effect of a DO prime is predicted as PD-bias increases. The numerical values of the primed log probabilities are hypothetical, for illustration purposes only.

3.4 Quantifying Verb Biases

The verb bias for a specific verb is the likelihood of producing structure X compared to the alternative structure Y. In human experiments, baseline verb biases are estimated as the ratio of the number of times one structure occurs over the sum of both structures; such occurrences can be counted either in natural production experiments or in corpus searches (Zhou and Frank, 2023). Here, we compute a verb bias for each verb analogously as the ratio of the model probability of one structure over the sum of the probabilities of both structures. The probability of a sentence s is the product of probabilities assigned by LLMs to each token w_i: $\mathcal{P}(s) = \prod_i \mathcal{P}(w_i|\overline{w}_{<i})$.This measures how likely it is for the model to see or produce this sentence. Then, given a set of sentences $\mathcal{S}_V$ with ditransitive verb V, where each sentence t_{PD} with structure PD always has its counterpart t_{DO} in the opposite structure of DO (see 6 and 7), the **PD-bias of verb V** is the mean normalized probability of sentences in structure PD:

$$bias(V, \text{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\text{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\text{PD}})}{\mathcal{P}(t_{\text{PD}}) + \mathcal{P}(t_{\text{DO}})} \quad (1)$$

The DO-bias of verb V, i.e., $bias(V, \text{DO})$, is defined analogously. We compute verb bias separately for each model (see Appendix C for GPT3), and use a model's biases in our assessment of whether it shows the IFE.

3.5 Simulating Structural Priming

As stated in Section 3.1, we use two modes to simulate structural priming. Following van Schijndel

[2] We also tested the LSTM models from Gulordava et al. (2018) with the `Concatenation` mode and found no evidence of structural priming.

and Linzen (2018), for the `Fine-Tuning` mode, we update the parameters by fine-tuning the model on a single prime sentence with learning rate $1e^{-5}$ for 10 epochs (see the full fine-tuning details in Appendix B), and we use the updated model to do inference on the target sentence. Following Sinclair et al. (2022), for the `Concatenation` mode, we condition a target sentence on a prime sentence through directly concatenating them, separated by a period, without any weight updates.

The probability of the target sentence after priming is the product of probabilities assigned to its tokens: $\mathcal{P}(t_X|p_X) = \prod_i \mathcal{P}(t_{X_i}|p_X, t_{X_{<i}})$. Under standard priming, the probability of the target sentence t_X should be greater after being primed by a ditransitive prime sentence of either structure: $\mathcal{P}(t_X|p_X) > \mathcal{P}(t_X)$ and $\mathcal{P}(t_X|p_Y) > \mathcal{P}(t_X)$. Being primed by the same structure should induce a larger probability increase than being primed by the opposite structure: $\mathcal{P}(t_X|p_X) > \mathcal{P}(t_X|p_Y)$.

3.6 Predictions about the Inverse Frequency Effect

Recall that, under the IFE, the priming strength of structure X inversely correlates with the prime verb's X-bias. That is, the degree to which the target production's probability deviates from the baseline is determined by the prime verb. Consider, for example, the case in which the prime sentence has a DO structure and the target sentence has a PD structure (other combinations of structures work analogously). For each prime verb V, we computed the PrimeBias for the PD target structure given a DO prime sentence, where PrimeBias is defined as the normalized target probability primed by this verb over a set of target sentences:

$$PrimeBias(\text{PD}|\text{DO}, V) = \frac{1}{|T_{\text{PD}}| \cdot |P^V_{\text{DO}}|} \sum_{t_{\text{PD}} \in T_{\text{PD}}} \sum_{p^V_{\text{DO}} \in P^V_{\text{DO}}} \frac{\mathcal{P}(t_{\text{PD}}|p^V_{\text{DO}})}{\mathcal{P}(t_{\text{DO}}|p^V_{\text{DO}}) + \mathcal{P}(t_{\text{PD}}|p^V_{\text{DO}})} \quad (2)$$

As shown in Figure 2, the IFE predicts that with a PD target and DO prime sentence, as the prime verb V's PD-bias increases, the prime sentence is less expected, resulting in *a larger priming strength towards the DO direction* in the target production, i.e., a smaller $PrimeBias(\text{PD}|\text{DO}, V)$ value. Similarly, as PD-bias increases, a PD prime sentence will result in *a smaller priming strength towards the PD direction* in the target production,

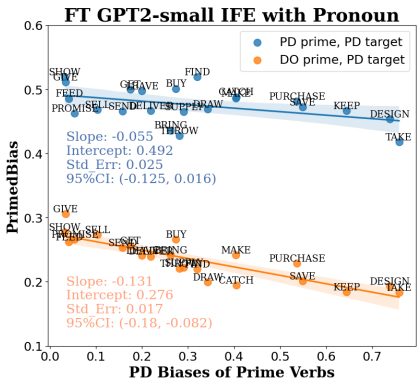

Figure 3: GPT2-SMALL shows a robust IFE under the `Fine-Tuning` mode, shown by the two negative slope.

i.e., again a smaller $PrimeBias(\text{PD}|\text{PD}, V)$ value. Therefore, when plotting $PrimeBias(\text{PD}|\text{DO}, V)$ and $PrimeBias(\text{PD}|\text{PD}, V)$ against increasing PD verb biases and fitting a line with linear regression, the IFE predicts **negative slopes for both plots**. Moreover, standard priming predicts that $PrimeBias(\text{PD}|\text{PD}, V)$ should have a higher intercept than $PrimeBias(\text{PD}|\text{DO}, V)$ since the former increases $\mathcal{P}(T_{\text{PD}})$ more than the latter.[3]

4 Results and Analysis

For each model and for each prime verb, we plotted $PrimeBias(\text{PD}|\text{PD}, V)$ and $PrimeBias(\text{PD}|\text{DO}, V)$ against increasing verb biases and used linear regression to find the relationship between priming strength and verb biases. We report the 95% confidence intervals of the fitted slopes to assess the significance of the fitted lines, where intervals below 0 suggest that the slopes are more likely to be negative.

4.1 `Fine-Tuning` Mode

We applied the `Fine-Tuning` mode to GPT2-SMALL. As is shown in Figure 3, the $T_{\text{PD}}|P_{\text{PD}}$ condition has a larger intercept than the $T_{\text{PD}}|P_{\text{DO}}$ condition, suggesting that the `Fine-Tuning` mode is able to capture standard structural priming. Further, we observe negative slopes in both lines, suggesting that the `Fine-Tuning` mode is able to capture the IFE. The 95% confidence interval for the $T_{\text{PD}}|P_{\text{DO}}$ condition is below 0, demonstrating a stronger IFE than the $T_{\text{PD}}|P_{\text{PD}}$ condition.

[3] The other two conditions, namely $T_{\text{DO}}|P_{\text{PD}}$ and $T_{\text{DO}}|P_{\text{DO}}$, are mathematically guaranteed to have slopes that are exactly opposite from $T_{\text{PD}}|P_{\text{PD}}$ and $T_{\text{PD}}|P_{\text{DO}}$, respectively, and the intercept should be 1 minus the intercept of its counterpart.

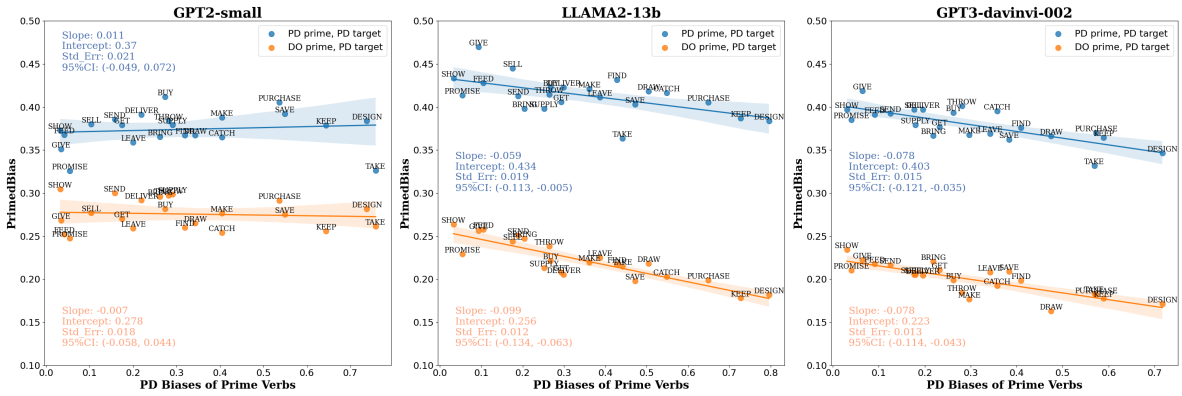

Figure 4: The IFE across models of different sizes in the *WithPronoun* condition under the Concatenation mode.

Overall, these results show that even the smallest model we analyzed shows the IFE under explicit error-driven weight updates. We did not carry out this mode for larger models as fine-tuning on the demonstration with GPT2-SMALL already satisfies its sanity-check purpose: as already noted, the `Fine-Tuning` mode will show the IFE by design, given its explicit gradient-based updates. We now turn to the crucial case of ICL, to see whether the IFE arises even without explicit weight updates.

4.2 `Concatenation` Mode

We applied the `Concatenation` mode to all models. In Figure 4, we show one plot for each of the three types of models (i.e., showing just one variant of that model type and showing only the *WithPronoun* condition) and report the results across all model variants in Table 1 (see Appendix D for the full statistical results). For all models across all conditions, the $T_{\text{PD}}|P_{\text{PD}}$ intercept is greater than the $T_{\text{PD}}|P_{\text{DO}}$ intercept, showing the standard structural priming effect, which is consistent with our prediction. For the IFE, we found that all three sizes of GPT2 failed to show the IFE, as the slopes are either positive or close to zero. This suggests that, in GPT2, priming strength is not correlated with verb bias under our metric. All three LLAMA2 models showed negative slopes in both cases, which is consistent with the IFE. However, only in the *WithPronoun* $T_{\text{PD}}|P_{\text{DO}}$ condition are the 95% confidence intervals constantly below 0 across the three models, suggesting that LLAMA2 displays the IFE but in a noisy way that is not fully robust. Finally, for GPT3, both $T_{\text{PD}}|P_{\text{PD}}$ and $T_{\text{PD}}|P_{\text{DO}}$ under the *WithPronoun* condition have their confidence intervals below 0, while the confidence intervals in both case in the *NoPronoun* condition contain 0.

The *WithPronoun* conditions of all three LLAMA2 models and GPT3 are statistically significant with $p < 0.0001$. Models at least as big as LLAMA2-13B are at least marginally statistical significant in all conditions while smaller models do not. The significant tests support our hypothesis that larger models show more significant IFE.

Therefore, besides corroborating the finding from prior work that LLMs show structural priming effects, the current results suggest that, in general, **larger models tend to show a stronger IFE than smaller models, meaning that IFE strength correlates with ICL capability** since larger models also typically perform better at ICL than smaller models (Brown et al., 2020). Assuming that ICL capability correlates with LLM size, given the currently observed pattern, we predict that larger models such as GPT4 should show a stronger IFE.

4.3 The Distinction between the *WithPronoun* vs. *NoPronoun* Conditions

As shown in Table 1, the majority of cases with their 95% confidence intervals containing 0 are the *WithPronoun* $T_{\text{PD}}|P_{\text{DO}}$ cases. Why the IFE arises more strongly in the *WithPronoun* condition than the *NoPronoun* condition remains unclear. One major difference between these conditions lies in the default verb biases: as shown in Figure 5 in Appendix C, GPT3 shows an overwhelming bias towards PD without pronouns but a reverse pattern favoring DO with pronouns. This pattern holds across all models and is consistent with statistics obtained from our corpus parse. One possible conclusion is that the most common indirect object DPs in the DO sentences are animate pronouns, causing the model to assign a higher probability to pronoun sentences. However, we leave for future work the question of why and how these differences in verb biases could lead to the IFE having different

		Intercept		Slope		95% CI of Slopes	
Model	Pronoun Obj?	PD-PD	DO-PD	PD-PD	DO-PD	PD-PD	DO-PD
GPT2-small	True	0.370	0.278	0.011	-0.007	(-0.049, 0.072)	(-0.058, 0.044)
GPT2-small	False	0.746	0.653	0.014	0.006	(-0.042, 0.07)	(-0.06, 0.072)
GPT2-medium	True	0.351	0.256	-0.013	-0.026	(-0.078, 0.053)	(-0.073, 0.022)
GPT2-medium	False	0.748	0.590	-0.023	-0.035	(-0.079, 0.032)	(-0.125, 0.054)
GPT2-large	True	0.330	0.241	0.011	-0.037^{+}	(-0.043, 0.065)	(-0.09, 0.015)
GPT2-large	False	0.698	0.487	-0.003	-0.02	(-0.062, 0.055)	(-0.098, 0.058)
Llama-7b	True	0.392	0.229	-0.02	-0.086****	(-0.067, 0.026)	**(-0.126, -0.045)**
Llama-7b	False	0.807	0.627	-0.026	-0.111^{+}	(-0.102, 0.049)	(-0.279, 0.057)
Llama-7b-chat	True	0.413	0.263	-0.012	-0.095****	(-0.067, 0.043)	**(-0.146, -0.044)**
Llama-7b-chat	False	0.788	0.605	-0.013	-0.102	(-0.115, 0.089)	(-0.289, 0.085)
Llama-13b	True	0.434	0.256	-0.059**	-0.099****	**(-0.113, -0.005)**	**(-0.134, -0.063)**
Llama-13b	False	0.859	0.685	-0.066^{+}	-0.177*	(-0.163, 0.031)	(-0.388, 0.033)
GPT3-davinci-002	True	0.403	0.223	-0.078****	-0.078****	**(-0.121, -0.035)**	**(-0.114, -0.043)**
GPT3-davinci-002	False	0.851	0.632	-0.064^{+}	-0.145*	(-0.153, 0.025)	(-0.301, 0.012)

Table 1: The intercept, slope, and 95% confidence interval of the fitted lines for each condition under the `Concatenation` mode. We mark the p-values of the fitted slopes with the standard notation: * for $p \leq 0.05$, ** for $p \leq 0.01$, **** for $p \leq 0.0001$, and + for $p \leq 0.1$ (marginally significant). 95% confidence intervals of the fitted slopes below zero are bold (which means the fitted slopes are more statistically significant to be negative).

levels of significance across the two conditions.

5 Discussion and Conclusion

Evidence that ICL is implicitly an error-driven learning mechanism We started with the question of whether ICL could be a processing mechanism of LLMs that can capture the flexibility and adaptability of human learning mechanisms. One type of error-driven learning mechanism that supports flexible adaptation to data is gradient descent. Motivated by previous theoretical proposals that ICL could in principle be implicitly performing gradient descent or fine-tuning, we attempted to better characterize what kind of learning ICL is in actual models. We focused on one particular aspect of ICL: whether it involves an implicit error signal during the forward computation. In contrast to previous analyses of ICL, we test standard LLMs with natural language data. Having established a connection between ICL and human structural priming, we used the IFE to diagnose whether LLMs behave as if they are implicitly performing error-driven learning when processing a prime sentence. We found that LLMs do indeed display the IFE in many cases, particularly larger models. These findings support the hypothesis that an error signal is implicitly computed in the forward pass of ICL. Differences between models suggests that ICL only takes on a gradient-descent-like nature in larger models. More speculatively, these results raise the possibility that error-driven learning might be a crucial property that enables generalization from a small number of samples, an ability that is shared by human learners and (to some extent) LLMs performing ICL. Our study not only provides behavioral results that align LLM and human behavior on structural priming at the processing mechanism level, but also demonstrates the possibility of studying the nature of ICL with off-the-shelf pre-trained LLMs and with naturalistic data.

ICL as a consequence of language modeling ICL is typically understood as involving demonstration-answer pairs in the prompt. Inspired by data-centric views that explain ICL from the distributional properties of pre-training data , we proposed a generalized notion of ICL that is sensitive to general parallelisms. Therefore, ICL could be viewed as a side product of the general propensity for structural parallelism in the language modeling task. We leave this perspective for future study.

Future Directions In this study, we focused on one single "task", namely selecting DO versus PD sentence structures. If our reasoning is correct, such that it is indeed the implicit error signal of the ICL that results in LLMs' capability of capturing the IFE, then we predict that the IFE diagnostics could be generalized to other ICL tasks, even non-linguistic tasks. Future work could extend our current method to additional ICL tasks. Finding the IFE on a wider range of tasks would strengthen the claim that ICL is driven by implicit error-driven computation, while not observing the IFE on other tasks would indicate that ICL only sometimes displays a gradient-descent-like character.

Limitations

Behavioral versus Mechanistic Accounts Although ICL is generally identified as a phenomenon at the behavioral level, having an explanation at the mechanistic level would bring greater interpretability and could contribute to more concrete theory building. Our current study, despite using real pretrained models and naturalistic data, remains at the behavioral level and is empirical in nature. Given our current contribution of establishing a connection between ICL and human priming and using the IFE as a way to diagnose whether ICL works in a gradient-descent-like manner or not, future work could improve our understanding by incorporating techniques from mechanistic interpretability to explain our current finding at the mechanistic level.

Examining the IFE in Other Models As ICL capability is often argued to scale with model size, we predict in Section 4.2 that the IFE effect will be more robust in larger models. Although the difference in the IFE behavior between GPT2 and GPT3-BASE is substantial, we do not believe that we have observed a saturation of the IFE because increasing the confidence level to 99% could still result in confidence intervals containing 0. GPT3-BASE is currently the biggest model on which we have access to the logit predictions, but we believe the same behavioral test could be applied to larger models in order to verify our prediction.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2024. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36.

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

J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355–387.

Holly P. Branigan and Martin J. Pickering. 2017. An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, 40:e282.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Franklin Chang, Gary S. Dell, and J. Kathryn Bock. 2006. Becoming syntactic. *Psychological Review*, 113(2):234.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2024. Parallel structures in pretraining data yield in-context learning. *arXiv preprint arXiv:2402.12530*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Shivam Garg, Dimitris Tsipras, Percy S. Liang, and Gregory Valiant. 2022. What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Aaron Gokaslan and Vanya Cohen. 2019. OpenWebText corpus. http://Skylion007.github.io/OpenWebTextCorpus.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. 10.5281/zenodo.1212303.

T. Florian Jaeger and Neal Snider. 2008. Implicit learning and syntactic persistence: Surprisal and cumulativity. In *Proceedings of the 30th Annual Conference*

of the Cognitive Science Society, volume 827812. Cognitive Science Society Austin, TX.

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.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? *Preprint*, arXiv:2406.04847.

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.

Ivan Lee, Nan Jiang, and Taylor Berg-Kirkpatrick. 2024. Is attention required for ICL? exploring the relationship between model architecture and in-context learning ability. In *The Twelfth International Conference on Learning Representations*.

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.

Neel Nanda and Joseph Bloom. 2022. TransformerLens. `https://github.com/neelnanda-io/TransformerLens`.

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.

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, Hong Kong, China. Association for Computational Linguistics.

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.

Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. 2023. Do pretrained transformers really learn in-context by gradient descent? *arXiv preprint arXiv:2310.08540*.

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.

Kristen M. Tooley and Matthew J. Traxler. 2010. Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10):925–937.

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, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proc. MLR*, volume 202, pages 35151–35174. PMLR.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55.

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.

A Finding pronoun probabilities in the OpenWebText corpus with `spaCy`

As is mentioned in Section 3.2, in order to estimate the verb biases represented in the training corpus of GPT2 models, we parsed a fragment (around 160 million tokens) of the OpenWebText corpus (Gokaslan and Cohen, 2019) with the Python package `spaCy` (Honnibal et al., 2020) to get an estimate of the DO vs. PD ratio for each verb. Specifically, we used the `en_core_web_trf` specification of the `spaCy` model, and we identified the set of dative alternation sentences by doing dependency parsing on each sentence. We found that the verb biases from the corpus are less well-represented in GPT2 models, motivating our decision to estimate verb biases by using model judgments rather than corpus statistics; see Section 3.4.

The corpus parse results motivated us to investigate the impact of animacy of the indirect object on the verb biases. Thus, we constructed the *WithPronoun* version of the corpus. To do this, we approximated the distribution of the natural occurrence frequencies over the set of English pronouns in dative alternation sentences from a fragment of the OpenWebText corpus (Gokaslan and Cohen, 2019), which is a reasonable proxy for the closed WebText corpus that was used to train the GPT2 models. Then, we counted the frequencies of the set of English pronouns that occurred as the indirect object of the ditransitive verb. The list of pronouns and their frequencies are presented in Table 2, sorted by frequency.

Pronoun	Frequency
you	4621
me	2962
us	2959
him	2210
them	1847
it	1297
her	738

Table 2: The respective frequencies of the English pronouns occurring as the indirect object of ditransitive sentences in a fragment of the OpenWebText corpus.

To convert the existing dative alternation priming corpus to the *WithPronoun* version, we replaced the indirect object of every sentence in the existing corpus with one of the pronouns through random sampling according to their respective relative frequencies.

B Fine-tuning details

As is presented in Section 3.5, to simulate structural priming in the `Fine-tuning` mode, we fine-tuned a pre-trained GPT2-SMALL model on every prime sentence and used the updated model to do inference on the target sentences.

We loaded the pre-trained GPT2-SMALL model from the `TransformerLens` (Nanda and Bloom, 2022) package and used the `train` function from `TransformerLens` to do fine-tuning. To avoid catastrophic forgetting during fine-tuning, we applied a regularization term to the loss function for gradient descent. We randomly sampled a fixed set of 5000 adjacent tokens from the OpenWebText (so that it resembles the distribution of the pre-training data) and computed the loss on them of the pre-trained GPT2-SMALL model. Then, at each step during fine-tuning, we added to the loss term the squared difference between the current loss and the raw (pre-trained) loss of the model on these 5000 tokens, scaled by a coefficient $\lambda = 0.8$. We found that this regularization term helped keep the model stable during fine-tuning on a single sentence.

Parameter	Value
number of epochs	10
batch size	1
learning rate	$1e^{-5}$
optimizer	AdamW
lambda	0.8

Table 3: Hyperparameters used as the training configuration for the `Fine-tuning` mode of structural priming on GPT2-SMALL.

We did a hyperparameter search and chose the set of parameters in Table 3. We used the default values from `TransformerLens` for the rest of the relevant hyperparameters (such as `warmup`, `maximum gradient norm`, etc.).

C Verb Biases represented in LLMs

See Figure 5 for details of the verb biases in GPT3.

D Full Statistics of the IFE Experiments in `Concatenation` Mode

See Table 4 (as a supplement to Table 1) for the full statistical metrics of the IFE experiments reported in Section 4.2.

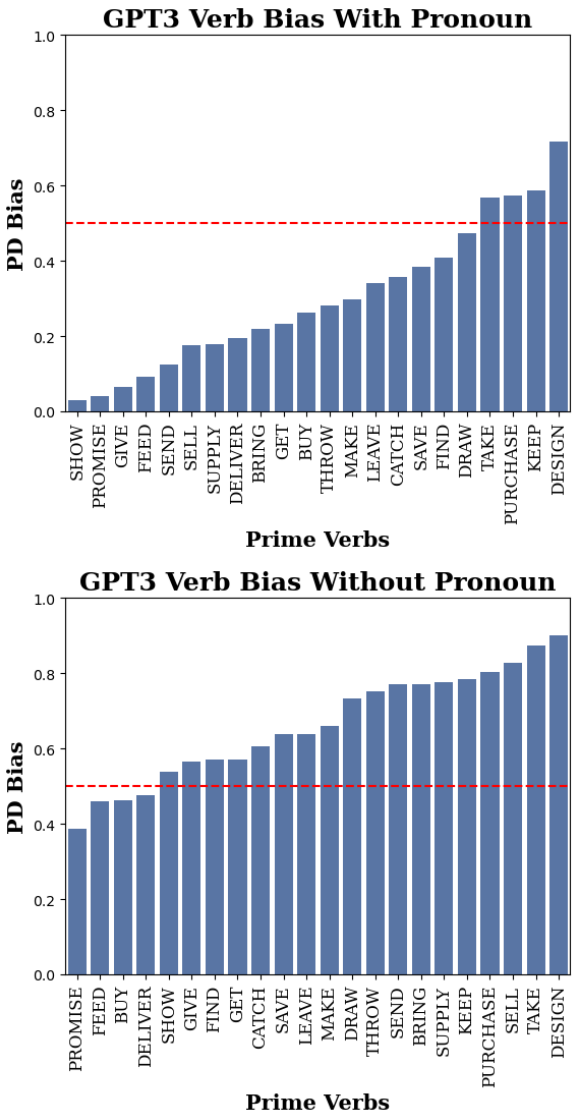

Figure 5: Comparison of PD biases with (top) and without (bottom) pronoun for GPT3. As is shown in Equation. 1, a high PD-bias means a larger proportion of probability assigned to the PD structure against the DO structure in LLMs.

		Std_Err		r-value		p-value		R^2		RMSE	
Model	Pronoun Obj?	PD-PD	DO-PD	PD-PD	DO-PD	PD-PD	DO-PD	PD-PD	DO-PD	PD-PD	DO-PD
GPT2-small	True	0.021	0.018	0.119	-0.088	0.597	0.698	0.014	0.008	0.020	0.017
GPT2-small	False	0.020	0.023	0.154	0.058	0.495	0.798	0.024	0.003	0.016	0.019
GPT2-medium	True	0.023	0.017	-0.123	-0.327	0.585	0.138	0.015	0.107	0.023	0.016
GPT2-medium	False	0.019	0.031	-0.258	-0.245	0.247	0.273	0.067	0.060	0.017	0.027
GPT2-large	True	0.019	0.018	0.129	-0.416	0.568	0.054	0.017	0.173	0.019	0.018
GPT2-large	False	0.020	0.027	-0.038	-0.163	0.868	0.469	0.001	0.026	0.018	0.024
Llama-7b	True	0.016	0.014	-0.270	-0.803	0.224	0.000	0.073	0.645	0.015	0.013
Llama-7b	False	0.027	0.059	-0.215	-0.387	0.336	0.076	0.046	0.149	0.019	0.042
Llama-7b-chat	True	0.019	0.018	-0.138	-0.766	0.541	0.000	0.019	0.587	0.018	0.017
Llama-7b-chat	False	0.036	0.066	-0.082	-0.327	0.718	0.137	0.007	0.107	0.024	0.044
Llama-13b	True	0.019	0.012	-0.568	-0.872	0.006	0.000	0.323	0.760	0.018	0.011
Llama-13b	False	0.034	0.074	-0.400	-0.473	0.065	0.026	0.160	0.224	0.019	0.042
GPT3-davinci-002	True	0.015	0.013	-0.755	-0.813	0.000	0.000	0.570	0.662	0.013	0.011
GPT3-davinci-002	False	0.031	0.055	-0.415	-0.507	0.055	0.016	0.172	0.257	0.020	0.035

Table 4: The standard error, r-value, p-value, R^2 coefficient, and RMSE score for each condition under the `Concatenation` mode.