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

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown the emerging capability of in-context learning (ICL). One line of research has explained ICL as functionally performing gradient descent. In this paper, we introduce a new way of diagnosing whether ICL is functionally equivalent to gradient-based learning. Our approach is based on the *inverse frequency effect* (IFE)—a phenomenon in which an error-driven learner is expected to show larger updates when trained on infrequent examples than frequent ones. The IFE has previously been studied in psycholinguistics because humans show this effect in the context of structural priming (the tendency for people to produce sentence structures they have encountered recently); the IFE has been used as evidence that human structural priming must involve error-driven learning mechanisms. In our experiments, we simulated structural priming within ICL and found that LLMs display the IFE, with the effect being stronger in larger models. We conclude that ICL is indeed a type of gradient-based learning, supporting the hypothesis that a gradient component is implicitly computed in the forward pass during ICL. Our results suggest that both humans and LLMs make use of gradient-based, error-driven processing mechanisms.

1 Introduction

To what extent do humans and language models use similar processing mechanisms? This question is of interest to both Artificial Intelligence researchers and cognitive scientists. Language models and human learners have some substantial differences: human learners often display a flexible learning ability to adapt to new examples, while language models require massive training data and a large number of parameters to exhibit human-like performance. Recent pre-trained large language models (LLMs) have shown the emerging capability of in-context learning (ICL): LLMs can adapt to specific tasks with a few demonstration-answer pairs served as prompts in the context window without any parameter updates (Brown et al., 2020). This intriguing emergent capability could provide a way to bridge the divide between language models and human learners: perhaps ICL is a processing mechanism that, like humans, can flexibly adapt to new examples.

Among various works on the sources and interpretations of the ICL capability, one line of research aims to deepen the theoretical understanding of ICL by offering *functional* interpretations of ICL via gradient descent. Garg et al. (2022), Zhang et al. (2023), and Ahn et al. (2024) have shown that standard Transformers (Vaswani et al., 2017) can be trained to implement learning algorithms for linear regressions under the ICL training objectives. Von Oswald et al. (2023) have demonstrated 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. Dai et al. (2023) provided a mathematical construction showing the 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) pointed out that previous accounts are limited in treating ICL as a non-emergent property and deviate from actual LLMs pre-trained with natural data since those accounts involve hand-constructed weights and use ICL objectives instead of the standard language modeling objectives. They found inconsistent behaviors of ICL and GD in real models, and left the equivalence between ICL and GD an open hypothesis.

In this paper, we aim to better characterize *what kind of learning mechanism ICL is* by drawing a connection between ICL and human learning mechanisms. Specifically, we examine the hypothesis that *ICL functionally performs gradient-based fine-tuning (e.g., gradient descent)* by empirically investigating a weaker claim with off-the-shelf LLMs and with natural language data: **whether ICL is a type of gradient-based, i.e., error-driven learning such that a gradient component is computed during the forward pass**. We approach this question by treating ICL as a processing mechanism of LLMs and borrowing insights from methods of studying processing mechanisms in humans: we examine to what extent LLMs show the *inverse frequency effect* (IFE), a phenomenon in the human structural priming paradigm (Branigan and Pickering, 2017) that has been argued to require one particular processing mechanism in humans, namely *implicit learning* (e.g., Chang et al., 2006). We study the linguistic phenomenon of the dative alternation and demonstrate that LLMs show robust IFE under standard fine-tuning and varying degrees of IFE under the ICL setting, with larger models showing a stronger IFE. We conclude that ICL is indeed a gradient-based learning mechanism.

Our study has implications for both NLP/machine learning (1 and 2) and linguistically-motivated analysis of LLMs (3 and 4):

(1) We find evidence that ICL can be viewed as a form of gradient-based learning.

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

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

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

Overall, our results suggest that error-driven learning is an aspect of processing that is shared between humans and LLMs.

2 Background and Related Work

In this section, we lay out the building blocks necessary for our reasoning of diagnosing the gradient-based nature of ICL through the IFE. Our approach is formally stated in Section 3.1.

2.1 Structural Priming in Psycholinguistics

Structural priming refers to the phenomenon that speakers tend to reuse recently encountered syntactic structures (Bock, 1986). For example, speakers tend to produce a double object (DO) structure (e.g., *The student sent the professor a letter*) rather than a prepositional dative (PD) structure (e.g., *The student sent a letter to the professor*) after encountering a DO sentence (e.g., *Alice gave Bob a book*). Similar to adapting to prompts in LLMs, structural priming has also been interpreted an *adaptation* mechanism, where speakers adapt lexical and syntactic predictions to the current context (Jaeger and Snider, 2013).

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 (measured by the relative frequency in the speaker's experience against their counterparts) cause stronger overall priming than more preferred structures. The gradient degrees of each unique verb's structural preference is called *verb biases* (or alternation biases, see Hawkins et al., 2020 for a systematic investigation on verb biases in neural models). For example, since *give* is biased towards DO in English, a prime sentence with *give* in PD structure will cause a greater priming effect than that prime sentence in DO structure. That is, 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 on a PD prime, determined by its verb bias (Bernolet and Hartsuiker, 2010).

Two mainstream 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 the structural information has a higher probability of being reactivated on the next relevant opportunity. The current form of transient activation theory does not account for the IFE because it is independent from verb biases and does not involve any error-driven mechanism. Alternatively, *implicit learning theory* (Chang et al., 2006) claims that humans implicitly learn probabilistic information about different structures (including verb biases) from experience (in the long-term memory) and use such information to predict the form of prime sentences. 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

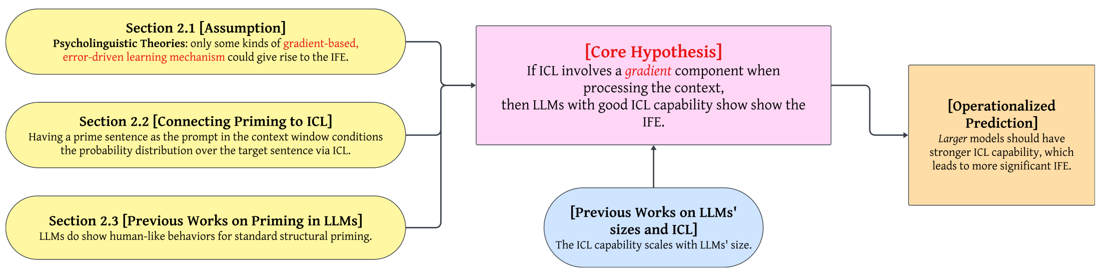

Figure 1: Reasoning behind our current study.

and the actual prime sentence: the less expectation the learner has on the observed prime structure, the larger the gradient is, resulting in a larger priming strength. Therefore, the implicit learning theory - unlike transient activation - predicts the IFE. The two theories are not mutually exclusive and can co-exist to account for priming, stated as the *dual mechanism* account (Tooley and Traxler, 2010).

In this study, we assume the correctness of the psycholinguistic theories that only some kinds of error-driven learning mechanisms could predict the IFE. Therefore, by examining whether LLMs show IFE in the ICL setting, we can infer whether some type of gradient component is computed in the forward pass without explicit weight updates, which informs us about whether there is a gradient-based component in ICL.

2.2 Connections among Distributional Properties of Pre-Training Data, Priming, and In-Context Learning

Another line of research explains the origin of ICL from the distributional properties of the pre-training data. Chan et al. (2022) showed that ICL emerges when the training data exhibits particular distributional properties (such as burstiness, in which items appear in clusters rather than being uniformly distributed over time). Hahn and Goyal (2023) argued that ICL emerges from the compositional structures found in the pre-training data under the standard next-token prediction objective. Chen et al. (2024) found that parallel structures in the pre-training data give rise to the ICL capability in LLMs. They defined parallel structures as pairs of phrases following similar templates in the same context window and found that removing parallel structures in the pre-training data significantly reduces LLMs' ICL accuracy. Chen et al. also pointed out that despite the fact that the pre-training data is not formatted strictly as in-context prompts, i.e., input-output pairs, the naturalistic data often contains phrases following similar templates. Those phrase pairs could be conceptualized as in-context examples of implicitly defined, less structured shared "tasks", such as n-gram copying, syntactic constructions, and world knowledge.

As structural priming is a well-attested phenomenon in humans, it is reasonable for us to hypothesize that *structural priming is a factor that shapes the distribution of the pre-training data* since humans tend to produce abundant parallel structures in the naturalistic setting. For this reason, we view the repeated structures from structural priming as a case of the parallel structures in Chen et al.'s (2024) sense. Inspired by the data-centric perspective, we think of ICL as not necessarily involving explicit demonstration-answer pairs for specific tasks, as is typically understood in the literature. Instead, we conceptualize ICL as a more generalized notion that involves a sensitivity of parallelism through generic next-token prediction: any text in the context window will affect the conditional probability distribution over the logits of the next token. This generalized notion of ICL, namely, having prompts in the context window, is analogous to priming in humans, as is elaborated in Section 3.1.

2.3 Structural Priming in Neural Language Models

As structural priming has been proposed as a means of probing the abstract mental representations of structural information in humans (Branigan and Pickering, 2017), previous works have adopted this paradigm for probing learned linguistic representations in neural networks. It has been shown that LSTMs (Gulordava et al., 2018) are capable of adapting to syntactic structures under the adaptation way of priming (Van Schijndel and Linzen, 2018; Prasad et al., 2019): **fine-tuning** model weights on prime sentences and testing target sentence probabilities on the updated model, which

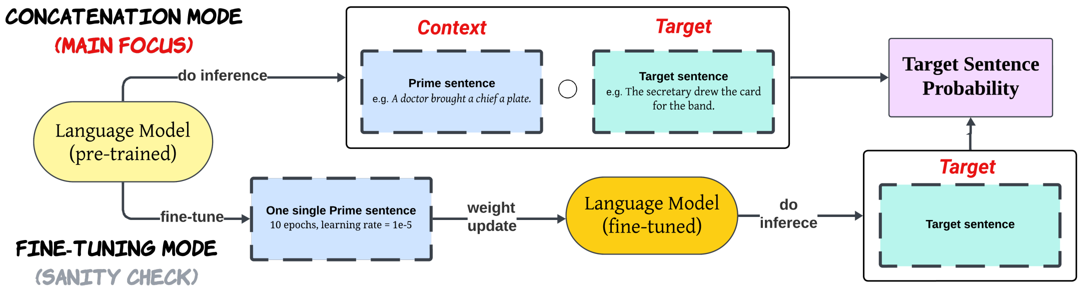

Figure 2: An overview of our experiment design.

is analogous to the implicit learning account of structural priming and *involves weight updates*. Recently, Sinclair et al. (2022) have shown that the GPT2 family (Radford et al., 2019) showed robust structural priming through encoding structural information given in the preceding context (i.e., directly **concatenating** target sentences with prime sentences), *which does not involve any weight updates*.[1] Other works have demonstrated crosslingual structural priming in large language models (Michaelov et al., 2023), suggesting that structural priming is robustly detected in LLMs.

Previous works have demonstrated the behavioral alignment of LLMs with humans on showing structural priming, which set the ground for our current study of investigating the processing mechanisms underlying priming. So far, no study has investigated whether LLMs also show the IFE, which serves as a separate motivation for our experiments.

3 Current Study

3.1 Overview of Our Approach

We first clarify our conceptualization of ICL. As is stated in Section 2.2, instead of following the notion of ICL as having demonstration-answer pairs of some tasks as prompts in the context window, here we propose that any text in the context window will condition the model's next word prediction: how the probability distribution over the next token changes depends on what the model captures or encodes from the context. Therefore, the generalized notion of ICL is analogous to adapting to encountered syntactic structures with structural priming in humans. On the humans' side, encountering a DO sentence will temporarily condition the speaker towards producing or more quickly comprehending another sentence of the same structure. On the models' side, processing a DO sentence in the prompt will condition the model to increase its probability of producing another DO structure sentence during generation, as is reviewed in Section 2.3. That is, the less structured, implicitly defined "task" encoded by the DO sentence as the prompt could be interpreted as "producing another sentence following the DO structural template exemplified in the prompt."

Then, our research question is **whether a gradient component is computed during the forward pass of processing the prompt in the generalized ICL setting**. We investigate the question by testing whether LLMs show the IFE in the ICL setting. As is illustrated in Figure. 1, given that (i) it has been argued by psycholinguists that only some error-driven learning mechanism will give rise to the IFE; (ii) processing the prime sentence in the context window conditions the probability of the target sentence in the generalized notion of ICL; (iii) standard structural priming in the ICL setting has been robustly observed, we hypothesize that the strength of the IFE positively correlates with the strength of the ICL capability of LLMs: the stronger the ICL capability is, the better the gradient will be computed in the forward pass, which leads to a stronger IFE.

Specifically, we simulate structural priming across LLMs of various sizes with the two modes mention in Section 2.3. As is illustrated in Figure. 2, the **`Fine-Tuning`** mode fine-tunes the parameters on a single prime sentence, and the updated model is used to infer the probability of the target sentence. The **`Concatenation`** mode resembles the ICL setting, where the prime sentence is directly concatenated with the target sentence as the prompt in the context window, and the probability of the target sentence is measured. The `Fine-Tuning` mode serves as a sanity check that LLMs are able

[1] Sinclair et al. (2022) have also demonstrated that the GPT2 models showed the *lexical boost effect*, another well-attested sub-phenomenon of structural priming, which is not our main focus here.

to show the IFE when there is explicit error-driven, gradient-based learning. It sets the ground for our main focus: using the `Concatenation` mode to diagnose the gradient-based nature of ICL.

3.2 Corpus

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

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

(6) **PD**: DP$_{subj}$ V DP$_{dobj}$ Prep 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 the way that controlled for the degree of semantic association and lexical overlapping between prime and target sentences, and sentences are semantically plausible as the ditransitive verbs were manually labeled with their verb frames.

Since our goal is to study the IFE, which depends on the verb biases of particular verbs, we want each pair of prime and target verbs to be equally represented. Thus, for each of the 22 prime verbs, we sampled 50 target sentences for each of the 21 target verbs (we excluded cases where prime and target verbs overlap). For each target sentence, we sampled a prime sentence with no lexical overlapping to form a prime-target pair. Each prime-target pair yields 4 instances of structural combinations ($T_{\text{PD}}|P_{\text{PD}}$, $T_{\text{PD}}|P_{\text{DO}}$, $T_{\text{DO}}|P_{\text{PD}}$, $T_{\text{DO}}|P_{\text{DO}}$, i.e., target sentence T conditioned on prime P), resulting in 92400 prime-target pairs.[2] 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."*

Crucially, we also created an alternative dataset of the same size by replacing the indirect object DP with a pronoun.[3] This was motivated by a corpus parse[4] we did that showed that the most common indirect object in DO sentences are animate pronouns, suggesting that animacy is crucial for naturally capturing verb biases, confirming results reported in Bresnan et al. (2007). The presence and absence of pronouns lead to different verb biases for LLMs, which affect their IFE behaviors. We will return to this point in discussion.

[2] In this paper, we use P for prime sentences and $\mathcal{P}$ for probability.

[3] Details of the set of pronouns and their relative probabilities are in Appendix A.

[4] In order to find 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 python package `spaCy` (Honnibal et al., 2020) to get a distribution of the DO vs. PD ratio for each verb. We found that the verb biases from the corpus are less well-represented in GPT2 models.

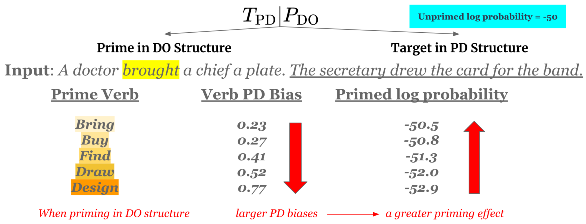

Figure 3: A demonstration of the IFE: a stronger priming effect of a DO prime is predicted as PD-bias increases. The numerical values of the primed log probabilities are for illustration purpose.

3.3 Language Models

We considered a set of Transformer models that have been claimed to show ICL capabilities to various extents (Lee et al., 2023):

- **GPT2** (Radford et al., 2019) in three of its sizes (SMALL, MEDIUM, LARGE), with 85M, 302M, and 708M number of parameters, respectively. All versions were loaded from package `transformerLens` (Nanda and Bloom, 2022).

- **LLAMA2** (Touvron et al., 2023) in three versions: 7B (6.5B parameters), 7B-CHAT (6.5B parameters), 13B (13B parameters). All versions were loaded from Huggingface (Wolf et al., 2019).

- **GPT3-base** (Brown et al., 2020) with the DAVINCI-002 version (175B parameters), accessed via OpenAI API.

The models are sorted by size, and correspondingly, by their ICL capabilities, so we predicted a stronger IFE as size increases.[5]

3.4 Quantifying Verb Biases

The verb bias for a specific verb is the likelihood of producing structure X compared to the alternative

[5] We also tested LSTMs (Gulordava et al., 2018) with the current `Concatenation` mode and we found that they did not show structural priming, although LSTMs did show structural priming in the `Fine-tuning` mode (Van Schijndel and Linzen, 2018; Prasad et al., 2019).

structure Y. In human experiments, baseline verb biases are estimated as the ratio of the number of one structure over the sum of two structures in natural production settings or corpus searches (Zhou and Frank, 2023). Here, we computed a continuous verb bias for each verb analogously as the ratio of the 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 LMs to each token w_i: $\mathcal{P}(s) = \prod_i \mathcal{P}(w_i)$.[6] 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_X with structure X always has its counterpart T_Y (see 5 and 6) in the opposite structure, the **X-bias of verb V** is the mean normalized probability of sentences in structure X:

$$bias(V, X) = \frac{1}{|\mathcal{S}_V|} \sum_{T_X \in \mathcal{S}_V} \frac{\mathcal{P}(T_X)}{\mathcal{P}(T_X) + \mathcal{P}(T_Y)} \tag{1}$$

3.5 Simulating Structural Priming

As is stated in Section 3.1, we use two modes to simulate structural priming. Following Van Schijndel 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 take 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: $\mathcal{P}(T_X|P_X) = \prod_i \mathcal{P}(T_{X_i}|P_X, T_{X_{<i}})$. Following from standard priming effect, the probability of the same target sentence T_X should be greater after primed by a sentence with the same structure: $\mathcal{P}(T_X|P_X) > \mathcal{P}(T_X)$; primed by the opposite structure decreases its probability: $\mathcal{P}(T_X|P_Y) < \mathcal{P}(T_X)$.

3.6 Predictions on the Inverse Frequency Effect

Recall that the IFE states that the priming strength of structure X inversely correlates with the prime verb's X-bias. That is, IFE is solely about the effect of the prime verbs, i.e., the degree of deviation of the target production from baseline it causes. Therefore, for each prime verb V, we computed the PrimeBias for the PD target structure given a DO prime sentence as the normalized target probability primed by this verb over a set of target sentences in Equation. 2:

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

As is shown in Figure. 3, the IFE predicts that with a PD target and DO prime sentence, as the prime verb V's PD-biases increase, the prime sentence is less expected, resulting in *a larger priming strength towards the DO direction* in target production, i.e., a smaller $PrimeBias(\text{PD}|\text{DO}, V)$ value. Similarly, as PD-biases increase, a PD prime sentence will result in *a smaller priming strength towards the PD direction* in target production, 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 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 the probability of T_{PD} while the latter decreases the probability of T_{PD}.[7]

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 pattern of priming strength with respect to verb biases. We reported the R-squared (R^2) coefficient and the root mean squared error (RMSE) to assess the significance of the fitted lines.

4.1 `Fine-tuning` Mode

We applied the `Fine-tuning` mode to GPT2-SMALL.[8] As is shown in Figure. 4, the

[6] In practice, we took the sum of the log probabilities assigned by LLMs to each token in the target sentence, which is equivalent to the summation notation.

[7] The other two conditions, namely $T_{\text{DO}}|P_{\text{PD}}$ and $T_{\text{DO}}|P_{\text{DO}}$, should have exactly the opposite slopes, and the intercepts should add up to 1 with its counterparts.

[8] We did not carry out this mode for larger models because of the substantial computational resources they require: each

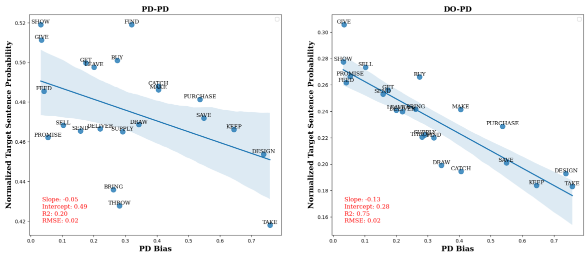

Figure 4: GPT2-SMALL shows robust IFE under the `Fine-tuning` mode. Both the $T_{\text{PD}}|P_{\text{PD}}$ condition (left) and the $T_{\text{PD}}|P_{\text{DO}}$ (right) have negative slopes, and the $T_{\text{PD}}|P_{\text{PD}}$ has a higher intercept than $T_{\text{PD}}|P_{\text{DO}}$.

$T_{\text{PD}}|P_{\text{PD}}$ condition having a larger intercept than the $T_{\text{PD}}|P_{\text{DO}}$ condition, suggesting that the `Fine-tuning` mode is able to capture the standard structural priming. We indeed observe two negative slopes, suggesting that the `Fine-tuning` mode is able to capture the IFE. The $T_{\text{PD}}|P_{\text{DO}}$ condition has a higher R^2 score of 0.75, demonstrating a stronger IFE than the $T_{\text{PD}}|P_{\text{DO}}$ condition.

Overall, this shows that even the smallest model shows the IFE under explicit gradient-based weight update, which passes the sanity check and suggests that LLMs are capable of showing the IFE with explicit gradient-based weight updates.

4.2 `Concatenation` Mode

We applied the `Concatenation` mode to all models, and we only show one plot for each of the three types of models and report the full results in Table 1. As is shown in Figure. 5, 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. The RMSE score for all conditions are less than 0.04, suggesting a significant predictability of the fitted lines to the data points. 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, the priming strength is not correlated with the verb biases under current metric. All three LLAMA2 models showed the two negative slopes, which is consistent with our prediction. However, only in the *Pronoun* $T_{\text{PD}}|P_{\text{DO}}$ condition are the R^2 coefficients constantly greater than 0.5 across the three models,[9] suggesting that the negative slopes themselves are not well accounted for given the distribution of prime verb's IFE scores. Finally, for GPT3, both $T_{\text{PD}}|P_{\text{PD}}$ and $T_{\text{PD}}|P_{\text{DO}}$ conditions with *Pronoun* have R^2 coefficient greater than 0.5, while neither holds in the *NoPronoun* condition.

Therefore, besides confirming previous results that LLMs show structural priming effect, the current results suggest that in general, **larger models tend to show stronger IFE, which analogously correlates with their ICL capability**. Assuming that LLMs' ICL capability correlates with their sizes, given the currently observed pattern, we further predict larger models such as GPT4 should show a stronger and more significant IFE, which is left for future study to verify.

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

As is shown in Table. 1, the majority of cases with R^2 score above 0.5 are the *WithPronoun* $T_{\text{PD}}|P_{\text{DO}}$ cases. The fact that the observed patterns fit better with our predictions in the *Pronoun* condition than *NoPronoun* condition remains curious. The main difference lies in the default verb biases: as is shown in Figure. 6 in Appendix C, the GPT3 model shows an overwhelming bias towards PD without pronoun but a reverse pattern favoring DO with pronoun. This pattern holds across all models and is consistent with our corpus parse result, which suggests that the most common indirect object DP in the DO sentences are animate pronouns, causing the model to assign a higher probability of pronoun sentences. However, it still remains puzzling why and how differences in verb biases could lead to different significance of the IFE behavior in the two conditions.

5 Discussion and Conclusion

ICL is a gradient-based learning mechanism We started with the question whether ICL could be a processing mechanism of LLMs that resembles human learning mechanisms that flexibly adapt to recently encountered examples. To better characterize what kind of learning ICL is, we examined existing proposals of explaining ICL through functional gradient descent or implicit fine-tuning and focused on one particular aspect of ICL: whether it involves a gradient component during the forward

priming instance requires a separate fine-tuning process. However, this is unproblematic for our conclusions, given that the `Fine-tuning` mode is expected to show the IFE in all cases, given its explicit gradient updates.

[9] Given no consensus on standard R^2 score thresholds, we picked this criterion by default.

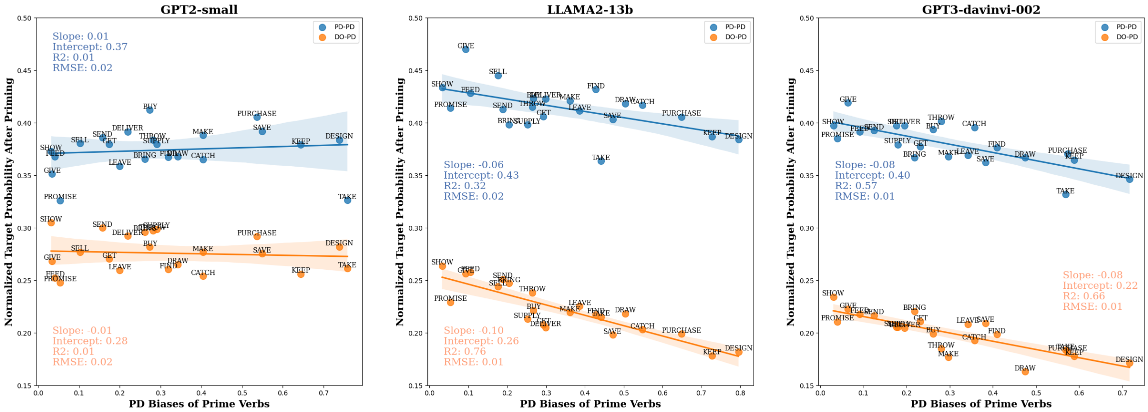

Figure 5: The IFE across models of different sizes in the *WithPronoun* condition under the Concatenation mode.

Models	With Pronoun	PDPD_slope	PDPD_intercept	PDPD_R^2	PDPD_RMSE	DOPD_slope	DOPD_intercept	DOPD_R^2	DOPD_RMSE
GPT2-small	True	0.011	0.370	0.014	0.020	-0.007	0.278	0.008	0.017
GPT2-small	False	0.014	0.746	0.024	0.016	0.006	0.653	0.003	0.019
GPT2-medium	True	**-0.013**	0.351	0.015	0.023	**-0.026**	0.256	0.107	0.016
GPT2-medium	False	**-0.023**	0.748	0.067	0.017	**-0.035**	0.590	0.060	0.027
GPT2-large	True	0.011	0.330	0.017	0.019	-0.037	0.241	0.173	0.018
GPT2-large	False	**-0.003**	0.698	0.001	0.018	**-0.020**	0.487	0.026	0.024
LLAMA2-7b	True	**-0.020**	0.392	0.073	0.015	**-0.086**	0.229	**0.645**	0.013
LLAMA2-7b	False	**-0.026**	0.807	0.046	0.019	**-0.111**	0.627	0.149	0.042
LLAMA2-7b-chat	True	**-0.012**	0.413	0.019	0.018	**-0.095**	0.263	**0.587**	0.017
LLAMA2-7b-chat	False	**-0.013**	0.788	0.007	0.024	**-0.102**	0.605	0.107	0.044
LLAMA2-13b	True	**-0.059**	0.434	0.323	0.018	**-0.099**	0.256	**0.760**	0.011
LLAMA2-13b	False	**-0.066**	0.859	0.160	0.019	**-0.177**	0.685	0.224	0.042
davinci-002	True	**-0.078**	0.403	**0.570**	0.013	**-0.078**	0.223	**0.662**	0.011
davinci-002	False	**-0.064**	0.851	0.172	0.020	**-0.145**	0.632	0.257	0.035

Table 1: The slope, intercept, R^2, and RMSE of the fitted lines for each condition under the `Concatenation` mode. Conditions with both negative slopes are bold (which suggests capturing the IFE), and R^2 scores higher than 0.5 are bold (which means a more significant fitted line).

computation. We differ from previous approaches by testing real LLMs and with natural language data. We established the connection between ICL and human structural priming, and we used the IFE to diagnose the presence or absence of the gradient component when LLMs process the prime sentence as the prompt. We found that larger models exhibit a stronger IFE, which suggest that the stronger ICL capability in larger models enables them to better capture the gradient nature of the verb biases encoded in the prime sentence as the prompt, which leads to a more significant IFE.

Therefore, our findings support the hypothesis that a gradient component is implicitly involved in the forward computation of ICL. This suggests that gradient-based learning might be a crucial property that enables generalizations from a few samples, which is shared between LLMs and human learners. Our study not only provides behavioral results that align LLMs' behaviors with human behaviors 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 emerges from Language modeling ICL is typical understood as involving demonstration-answer pairs in the prompt. Inspired by the 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. As a result, any text in the prompt could serve as an implicitly defined "task" of *following the template provided in the context and generating a parallel structure*. Therefore, ICL could be viewed as a side product of the general language modeling task. We leave this perspective for future investigation.

Future Directions If ICL is indeed gradient-based, our reasoning predicts that we should observe the IFE in other ICL tasks, including non-linguistic problems. For instance, for the Country-Capital mapping task, prompting the model with demonstrations with lower zero-shot probabilities is predicted to yield a larger improvement to the model performance than prompting with demonstrations with higher zero-shot probabilities. We leave this prediction for future study.

Limitations

Behavioral versus Mechanistic Accounts Although ICL is generally identified as a phenomenon at the behavioral level, having an explanation at the mechanistic level is desirable since it brings greater interpretability and is more concrete on theory building. Our current study, despite using real pre-trained 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 diagnostics on the presence or absence of the gradient-based nature of ICL, future work could improve our understanding by incorporating techniques from mechanistic interpretability to explain our current finding at the mechanistic level. For instance, it is possible to find a function vector (or, task vector) proposed by Hendel et al. (2023) and Todd et al. (2023) for the implicitly defined task of "producing a sentence in the DO (or PD) structure" (or, in general, produce the next token that resembles the structural template observed in the prompt).

Examining the IFE on Other Models As the ICL capability is argued to scale with the model sizes, 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 significant enough, we have not observed a saturation of the IFE. 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.

Extending the IFE to other ICL Tasks In this study, we only examined the IFE on one single "task" of structural priming. If our reasoning is correct, that it is indeed the gradient component 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. As is outlined in Section 5, future work could extend our current method to ICL tasks such as Country-Capital mapping, two-digit multiplication, etc. Finding the IFE on a wider range of tasks would better strengthen our reasoning, while not observing IFE on other tasks is also helpful for developing mechanistic level explanations towards a better understanding of ICL as a processing mechanism.

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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Franklin Chang, Gary S Dell, and 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 pre-training 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. *arXiv preprint arXiv:1803.11138*.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.

Robert D Hawkins, Takateru Yamakoshi, Thomas L Griffiths, and Adele E Goldberg. 2020. Investigating representations of verb bias in neural language models. *arXiv preprint arXiv:2010.02375*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*.

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.

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. 2023. Exploring the relationship between model architecture and in-context learning ability. *arXiv preprint arXiv:2310.08049*.

James A Michaelov, Catherine Arnett, Tyler A Chang, and Benjamin K Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. *arXiv preprint arXiv:2311.09194*.

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. *arXiv preprint arXiv:1909.10579*.

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.

Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. 2023. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*.

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, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Marten Van Schijndel and Tal Linzen. 2018. A neural model of adaptation in reading. *arXiv preprint arXiv:1808.09930*.

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, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. 2023. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*.

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, we constructed the with-pronoun version of the corpus in order to investigate the impact of animacy of the indirect object on the verb biases. 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 used to train the GPT2 model family.

We parsed a fragment (around 160 million tokens) of the corpus with `spaCy` (Honnibal et al., 2020). 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. Then, we counted the frequencies of the set of English pronouns 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 OpenWenText corpus.

To convert the existing dative alternation priming corpus to the with-pronoun version, we replaced the indirect object of every sentence in the existing corpus by 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 keeping the model stable during fine-tuning on a single sentence.

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.).

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.

C Verb biases with and without pronoun

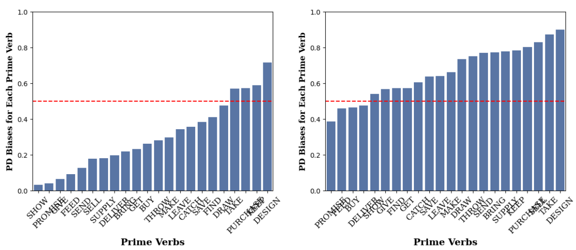

Figure 6: Comparison of PD biases with (top) and without (bottom) pronouns for GPT3. As is shown in Equation. 1, a high PD-bias means a larger proportion of probability assigned to the PD structure against the DO structure in LLMs.