Language Models Show Gradient Inverse Frequency Effects in Structural Priming: Implications for In-Context Learning*

Zhenghao (Herbert) Zhou

Yale University, New Haven, CT 06511, USA
herbert.zhou@yale.edu

Abstract. The structural priming paradigm in psycholinguistics has been demonstrated as a way of studying abstract structural representations in neural language models. The current study expands this line of reasoning by testing language models' behaviors on the inverse frequency effect, a priming phenomenon only predicted by implicit learning mechanisms. We showed that larger models tend to show a stronger inverse frequency effect, and we hypothesize that model sizes correlate with their in-context learning capability, interpreted as a form of implicit fine-tuning. Our study inquires beyond probing representations into language models' processing mechanisms. We conclude that in-context learning is a form of implicit learning shared between humans and language models.

Keywords: Probing Pre-trained Language Models · Structural Priming · In-Context Learning · Computational Psycholinguistics.

1 Introduction

A common interest among psycholinguists, computer scientists, and cognitive scientists is to what extent do humans and language models (LMs) share similar processing mechanisms. For psycholinguists, the validity of treating LMs as human-like subjects and conducting behavioral tests on them has been questioned [13]. For the field of artificial intelligence, understanding the processing mechanisms of LMs remains as a core question for both interpretability and designing more human-like intelligent systems [7]. This paper aims to improve our high-level understanding by suggesting that *implicit learning is a shared processing mechanism* between humans and LMs.

In the field of neural network analysis, many studies have focused on the learned internal linguistic representations of LMs. In particular, one line of research draws on the structural priming paradigm, a central tool in psycholinguistics for probing human syntactic representations [3]. Structural priming refers to the phenomenon that the structure of a sentence makes the same structure more probable in a follow-up sentence. This enables us to study the abstract syntactic representations constructed during comprehension and production. Applying structural priming to LMs helps probing whether the relevant syntactic representations are formed during processing. It has been demonstrated that LMs show the standard structural priming effect observed in humans [18, 20], which makes the priming method promising for further investigation.

* I thank Robert Frank and Tom McCoy (Department of Linguistics, Yale University) for advising me and offering insightful guidance on this study.

The current work expands on previous studies of neural structural priming by inferring the potential *In-Context Learning* (ICL) capability in LMs through testing their behaviors on the *inverse frequency effect* (IFE), a phenomenon in human structural priming that has been argued to require one particular learning mechanism in humans, namely implicit learning [5]. Among various hypotheses on the underlying mechanisms of ICL [21, 22], we focus on the hypothesis that treats ICL as implicit fine-tuning [6]. The underlying insight is that implicit learning in humans is analogous to ICL in neural models such that both require (implicit) gradient weight update. Therefore, we can use the IFE as a way to test whether LMs are performing implicit learning: if the IFE requires implicit learning, then whether a LM displays the IFE speaks to whether it is performing implicit learning. That is, observing the IFE in a LM suggests that it is performing implicit learning, which indicates that the ICL mechanism is indeed a form of implicit learning in the sense of implicit fine-tuning. This approach goes beyond probing the representations in LMs and inquires one level deeper into the processing mechanisms of LMs based on existing psycholinguistic theories of human cognition. This study also clarifies the nature of ICL shared by human sentence processing.

In section 2, we introduce the studies and theories of structural priming for both humans and neural models, as well as the hypothesized underlying mechanisms of ICL in LMs. In section 3, we quantify the inverse frequency effect by comparing the target sentence probabilities primed by different verbs. In section 4, we found that larger LMs tend to show a stronger IFE, which we hypothesized to correlate with their ICL capability, a form of implicit fine-tuning.

2 Background and Previous Study

2.1 Structural Priming in Psycholinguistics

Structural priming refers to the phenomenon that speakers tend to reuse recently encountered syntactic structures [2]. 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 LMs, structural priming has also been interpreted an *adaptation* mechanism, where speakers adapt lexical and syntactic predictions to the current context [10].

One important aspect of structural priming is the *inverse frequency effect* [1, 9, 11]: 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). 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, as is determined by its verb biases [1].

Two mainstream theories have been proposed to account for structural priming. *Transient activation theory* [16] 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* [5] 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 and the actual prime sentence. 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 [19].

2.2 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 [3], previous works have adopted this paradigm for neural network analysis. It has been shown that LSTMs are capable of adapting to syntactic structures under the adaptation way of priming [17, 20]: *fine-tuning model weights* on prime sentences and testing target sentence probabilities on the updated model, which is analogous to the implicit learning account of structural priming. Recently, Sinclair et al. have shown that GPT2 showed robust structural priming through encoding structural information given in the preceding context (i.e. directly concatenating target sentences with prime sentences)[18], *which does not involve any weight updates* and is analogous to the transient activation account in humans. Other works have demonstrated crosslingual structural priming in large language models [14], suggesting that structural priming is robustly detected in LMs.

So far, no study has investigated whether LMs also show the IFE. Given the unique status of the IFE as it is only predicted by the implicit learning mechanism, the IFE is a natural case to tease apart whether there exists implicit learning, or in general, (implicit) weight update in LMs. Under standard approaches, LMs operate via ICL, the ability to learn new tasks at inference time, using only input-output pair exemplars as guidance [7]. ICL has been interpreted as performing implicit Bayesian inference [22], functionally performing gradient descent [21], as a process of meta-optimization and performing implicit fine-tuning [6]. These hypotheses lead to our analogy between ICL and implicit learning. On one hand, ICL does not involve any updates to the learner, which makes it seem more like transient activation. On the other hand, ICL has been interpreted as performing gradient descent and behaves similarly to explicit fine-tuning, which would suggest that it could be viewed as a form of implicit learning. In this paper, we use the IFE to tease apart these possibilities.

3 Current Study

Research Questions and Hypotheses In this study, we investigated the following research questions: (i) how well are verb biases represented in LMs; (ii) to what extent do LMs show the IFE; (iii) assuming the distinction between the two accounts of structural priming, what could we infer about the ICL mechanism across LMs of different sizes given their behaviors on the IFE, and to what extent could ICL be viewed as a form of implicit learning.

Specifically, assuming that only implicit learning could predict the IFE, we hypothesize that the transient activation way of simulating structural priming in LMs will not elicit the IFE for models without or with weak ICL capability, while larger models with stronger ICL capability will show the IFE. That is, we proposed that the IFE could be a way of assessing the degree of ICL capability in LMs of various sizes.

Corpus We adapted the *Core Dative* PRIME-LM Corpus from Sinclair et al.[18] 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: (i) DO: $\texttt{DP}_{subj}$ `V` $\texttt{DP}_{iobj}$ $\texttt{DP}_{dobj}$ (e.g. *A girl bought a guy a coffee.*); (ii) PD: $\texttt{DP}_{subj}$ `V` $\texttt{DP}_{dobj}$ `Prep` $\texttt{DP}_{iobj}$ (e.g. *A girl bought a coffee for a guy.*). The `DP`s are simply 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[1]), resulting in 92400 prime-target pairs. 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[2]. This was motivated by a corpus parse[3] 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 [4]. The presence and absence of pronouns lead to different verb biases for LMs, which affect their IFE behaviors. We will return to this point in discussion.

Language Models We considered a set of Transformer models that have been claimed to show ICL capabilities to various extents [12]: **GPT2** in three of its sizes (SMALL, MEDIUM, LARGE), with 85M, 302M, and 708M number of parameters, respectively. All versions were loaded from package `transformerLens`[15]. **LLAMA2** in three versions: 7B (5B parameters), 7B-CHAT (5B parameters), 13B (9.9B parameters). All versions were loaded from Huggingface. **GPT3-base** 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[4].

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

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

[3] 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 with python package `spaCy` 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.

[4] We also tested LSTMs [8] with the current transient activation mode and we found that they didn't show structural priming, although LSTMs did show structural priming in the implicit learning mode [17, 20].

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 one structure over the sum of two structures in natural production settings or corpus searches [23]. 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 sum of probabilities assigned by LMs to each token w_i: $\mathcal{P}(s) = \sum_i \mathcal{P}(w_i)$[5]. 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 in the opposite structure, the **X-bias of verb V** is the mean normalized probability of sentences in structure X:

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

Simulating Priming Following Sinclair et al., we simulated structural priming resembling transient activation on the surface: conditioning a target sentence on a prime sentence through directly concatenating them, separated by a period, without any weight updates. Following from previous studies [18, 20], the probability of the target sentence after priming is the sum of probabilities assigned to its tokens: $\mathcal{P}(T_X|P_X) = \sum_i \mathcal{P}(T_{X_i}|P_X, T_{X_{<i}})$. 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: $\mathcal{P}(T_X|P_Y) < \mathcal{P}(T_X)$.

Predictions on 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 normalized mean target probability primed by this verb over a set of target sentences:

$$\bar{\mathcal{P}}(T_{\text{PD}}|P^V_{\text{DO}}) = \frac{1}{|\mathcal{S}_V|} \sum_{T_{\text{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(T_{\text{PD}}|P^V_{\text{DO}})}{\mathcal{P}(T_{\text{DO}}|P^V_{\text{DO}}) + \mathcal{P}(T_{\text{PD}}|P^V_{\text{DO}})} \tag{2}$$

As is shown in Figure. 1, the IFE predicts that for $T_{\text{PD}}|P_{\text{DO}}$, as prime verbs' PD-biases increase, a DO prime sentence is less expected, resulting in *a larger priming strength towards the DO direction* in target production, i.e. a smaller $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{DO}})$ 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 $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{PD}})$ value. Therefore, when plotting $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{PD}})$ and $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{DO}})$ against increasing verb biases and fitting a line with linear regression, observing the IFE predicts **negative slopes for both plots**. Moreover, standard priming predicts that $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{PD}})$ should have a higher intercept than $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{DO}})$ since the former increases the probability of T_{PD} while the latter decreases the probability of T_{PD}.[6]

[5] Whether to take log probabilities does not make a difference here. Log probabilities has a natural interpretation as the surprisal or perplexity [20], which is less relevant here.

[6] 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.

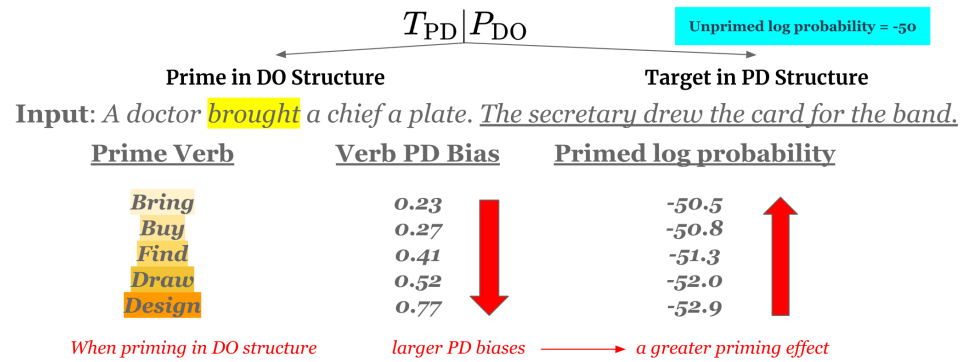


Fig. 1: The IFE predicts a stronger priming effect of a DO prime as PD-bias increases.

4 Results and Analysis

Results For each model, we plotted $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{PD}})$ and $\bar{\mathcal{P}}(T_{\text{PD}}|P_{\text{DO}})$ 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 (R2) coefficient and the root mean squared error (RMSE) to assess the significance of the fitted lines. We only show one plot for each of the three types of models and report the full results in Table 1.

Table 1: The slope, intercept, R2, RMSE of the fitted lines for each condition.

size	pronoun	PDPD_slope	PDPD_intercept	PDPD_R2	PDPD_RMSE	DOPD_slope	DOPD_intercept	DOPD_R2	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

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 R2 coefficients constantly greater than 0.5 across the three models[7], suggesting that the negative slopes themselves are not well accounted for

[7] Given no consensus on standard R2 score thresholds, we picked this criterion by default.

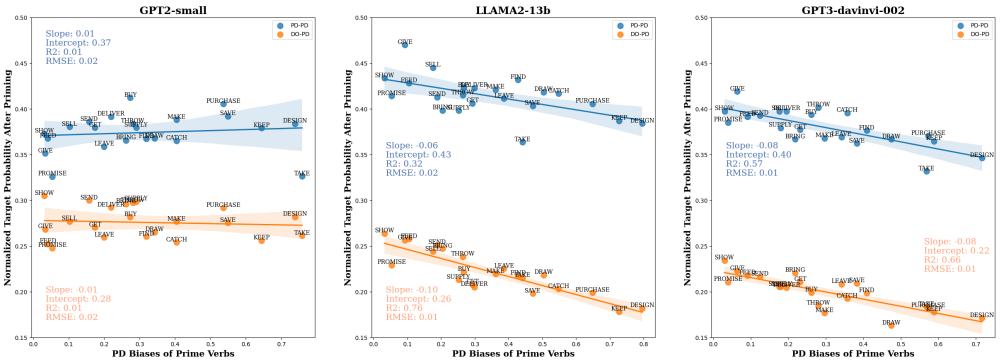
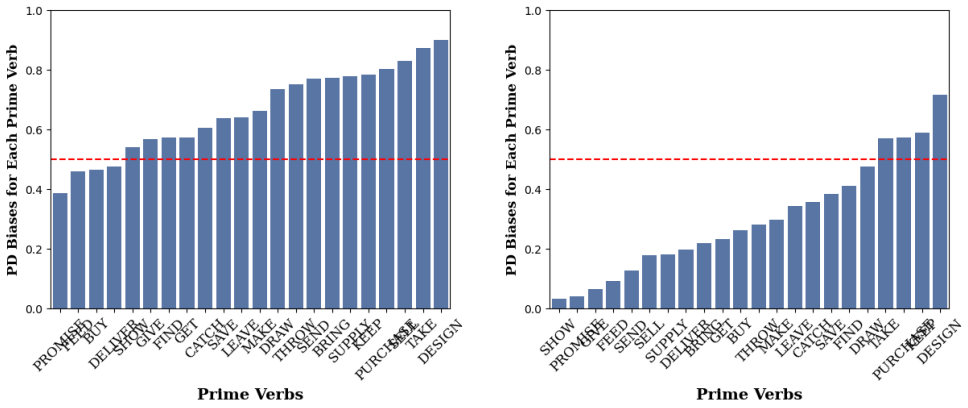


Fig. 2: Various degrees of the inverse frequency effect across models of different sizes in the With Pronoun condition.

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 R2 coefficient greater than 0.5, while neither holds in the NoPronoun condition.

Therefore, besides confirming previous results that LMs 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. 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. In the section below, we discuss the observed disparity between the Pronoun versus NoPronoun conditions.



(a) GPT3 Verb Biases without Pronoun (b) GPT3 Verb Biases with Pronoun

Fig. 3: Comparison of PD biases with and without pronoun for GPT3, where higher above the red line means a larger PD bias, below the red line means a DO bias.

Pronoun versus NoPronoun 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. 3, 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 pronoun, 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.

Discussion, Future Directions, and Conclusion Our preliminary results show that under the superficially transient activation way of simulating structural priming, LMs do show the IFE to various extent. The larger the model is, the stronger IFE it shows. Since the IFE is predicted only by the implicit learning mechanism, we hypothesize that larger LMs are more capable of implicit learning through their in-context learning capability, which is functionally performing implicit fine-tuning.

To better varify our reasoning, we propose a future direction of using the *explicit* implicit learning way [20] (i.e. fine-tuning on prime examples and use the updated model for target sentences) of doing structural priming on the GPT2 models and see whether even the relatively small GPT2 models show the IFE. If this is true, then it would suggest that the IFE is indeed predicted only by implicit learning, and that GPT2 models alone do not possess strong ICL capability in order to show the IFE. Furthermore, another future direction is to train diagnostic classifiers on the internal representations (i.e. residual streams or attention scores) of the GPT2 models across layers in order to localize the structural representations and how they influence logit predictions on the target sentences, which may help explaining the difference between the presence and absence of pronouns.

In sum, by drawing the connection between the implicit learning mechanism in human cognition and the implicit fine-tuning nature of in-context learning in language models, we suggests that implicit learning is a shared processing strategy between humans and language models. Our study offers a new way of not only probing learned representations, but also the processing mechanisms of neural language models. [8]

References

1. Bernolet, S., Hartsuiker, R.J.: Does verb bias modulate syntactic priming? Cognition **114**(3), 455–461 (2010)
2. Bock, J.K.: Syntactic persistence in language production. Cog. Psych. **18**(3), 355–387 (1986)
3. Branigan, H.P., Pickering, M.J.: An experimental approach to linguistic representation. Behavioral and Brain Sciences **40**, e282 (2017)
4. Bresnan, J., Cueni, A., Nikitina, T., Baayen, R.H.: Predicting the dative alternation. In: Cognitive foundations of interpretation, pp. 69–94. KNAW (2007)
5. Chang, F., Dell, G.S., Bock, K.: Becoming syntactic. Psych. Review **113**(2), 234 (2006)
6. Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., Wei, F.: 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 (2023)
7. Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Sui, Z.: A survey for in-context learning. arXiv preprint arXiv:2301.00234 (2022)

[8] Due to page limit, the rest of the References and Appendix do not fit here. They will be present in the final version.

8. Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M.: Colorless green recurrent networks dream hierarchically. pp. 1195–1205. NAACL (2018)
9. Jaeger, T.F., Snider, N.: Implicit learning and syntactic persistence: Surprisal and cumulativity. In: Proc. CogSci. vol. 827812. Cognitive Science Society (2008)
10. Jaeger, T.F., Snider, N.E.: 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 (2013)
11. Kaschak, M.P., Kutta, T.J., Jones, J.L.: Structural priming as implicit learning: Cumulative priming effects and individual differences. Psychonomic bulletin & review **18**, 1133–1139 (2011)
12. Lee, I., Jiang, N., Berg-Kirkpatrick, T.: Exploring the relationship between model architecture and in-context learning ability. arXiv preprint arXiv:2310.08049 (2023)
13. Marvin, R., Linzen, T.: Targeted syntactic evaluation of language models. arXiv preprint arXiv:1808.09031 (2018)
14. Michaelov, J., Arnett, C., Chang, T., Bergen, B.: Structural priming demonstrates abstract grammatical representations in multilingual language models. In: EMNLP. pp. 3703–3720. Association for Computational Linguistics (2023)
15. Nanda, N.: Transformerlens. https://github.com/neelnanda-io/TransformerLens (2022)
16. Pickering, M.J., Branigan, H.P.: The representation of verbs: Evidence from syntactic priming in language production. Journal of Memory and language **39**(4), 633–651 (1998)
17. Prasad, G., van Schijndel, M., Linzen, T.: Using priming to uncover the organization of syntactic representations in neural language models. In: Proc. CoNLL. pp. 66–76. Association for Computational Linguistics (2019)
18. Sinclair, A., Jumelet, J., Zuidema, W., Fernández, R.: Structural persistence in language models: Priming as a window into abstract language representations. TACL **10**, 1031–1050 (2022)
19. Tooley, K.M., Traxler, M.J.: Syntactic priming effects in comprehension: A critical review. Language and Linguistics Compass **4**(10), 925–937 (2010)
20. Van Schijndel, M., Linzen, T.: A neural model of adaptation in reading. arXiv preprint arXiv:1808.09930 (2018)
21. Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., Vladymyrov, M.: Transformers learn in-context by gradient descent. In: Proc. MLR. vol. 202, pp. 35151–35174. PMLR (2023)
22. Xie, S.M., Raghunathan, A., Liang, P., Ma, T.: An explanation of in-context learning as implicit bayesian inference. arXiv:2111.02080 (2021)
23. Zhou, Z., Frank, R.: What affects priming strength? simulating structural priming effect with pips. Proceedings of the Society for Computation in Linguistics **6**(1), 413–417 (2023)