Aligning In-Context Learning with Human Implicit Learning: Evidence from Gradient Inverse Frequency Effects of Structural Priming in Language Models

Zhenghao Zhou, Robert Frank, R. Thomas McCoy. {herbert.zhou, robert.frank, tom.mccoy}@yale.edu

1. Overview Large language models (LLMs) have been demonstrated to show the standard structural priming effect [1, 2]. We investigated whether LLMs and humans shared the same *processing mechanisms* for priming based on psycholinguistic theories of priming.

2. Structural Priming in Psycholinguistics and LLMs In structural priming, a sub-phenomenon named the inverse frequency effect (IFE) shows that less frequent structures cause a larger priming effect than more frequent structural alternatives [3] (see Fig. 1 for a demonstration). Between the two theories of priming, only *implicit learning* [4] (speakers learned distributional linguistic knowledge in an error-driven way when encountering primes) but not *transient activation* [5] (priming temporarily increases the activation of the structural representation) predicts the IFE. We simulated priming in LLMs with two modes corresponding to the two theories: the `Fine-tuning` mode fine-tunes the model on the prime sentence and uses the adapted model for target sentence production; the `Concatenation` mode concatenates and feeds the prime and target sentences to the model without weight-update. We predicted that both modes will show structural priming, but only the `Fine-tuning` mode will show the IFE.

3.1 Corpus We focused on the Double Object (DO, e.g. *Alice sent Bob a note.*) vs. Prepositional Dative (PD, e.g. *Alice sent a note to Bob.*) distinction and created 92400 prime-target trials with 22 dative verbs (adapted from [1]), where primes and targets are lexically independent.

3.2 Models We tested the behavior on GPT2 in three of its sizes (small, medium, large), LLAMA2 in three versions (7b, 7b-chat, 13b), and GPT3-base with the davinci-002 version.

3.3 Quantification and Predictions We measured the probability of target sentences by summing the probabilities assigned by LLMs to each token of the target sentence. We quantified the PD-bias of each verb as a continuous value with Eq. 1, which measures the baseline frequencies of the two structures. We quantified the IFE with Eq. 2, which represents the priming effect of an individual prime verb. We plotted verbs' priming effect against their PD-bias in the $t_{\text{PD}}|p_{\text{PD}}$ and $t_{\text{PD}}|p_{\text{DO}}$ conditions. The IFE predicts negative slopes in both conditions, with the intercept of $t_{\text{PD}}|p_{\text{PD}}$ higher than that of $t_{\text{PD}}|p_{\text{DO}}$ due to the standard priming effect.

4. Results and Discussion We only applied the `Fine-tuning` mode to GPT2 due to computational resource limits. As is shown in Fig. 2, even the smallest model showed significant IFE, consistent with our prediction. We applied the `Concatenation` mode to all models, with the full results presented in Table 1. As is shown in Fig. 3, all models showed standard priming effect. Smaller models did not show any IFE, while larger models showed larger IFE, which does not align with our prediction. Given that GPT2-small did show the IFE in the `Fine-tuning` mode, we hypothesized that in the `Concatenation` mode, the in-context learning (ICL) mechanism that only emerges in larger models gives rise to the IFE even without explicit weight-update, which is posited to be necessary to show the IFE. The current result aligns with the interpretation of treating ICL as a meta-optimization process that functionally performs implicit fine-tuning [6].

5. Conclusion We used the IFE effect as a diagnostic on LLMs' processing mechanism. We found various degrees of the IFE, and we inferred that ICL in larger models gives rise to the IFE. We conclude that ICL is a form of implicit learning shared between humans and LLMs.

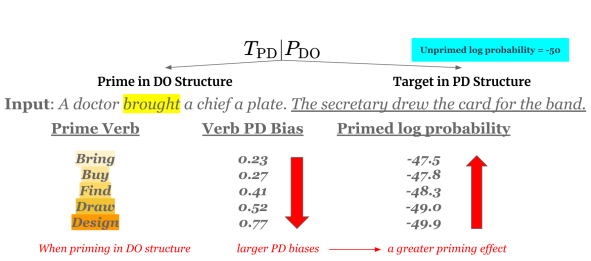

Figure 1: A demonstration of the IFE with two sample $T_{PD}|P_{DO}$ trials.

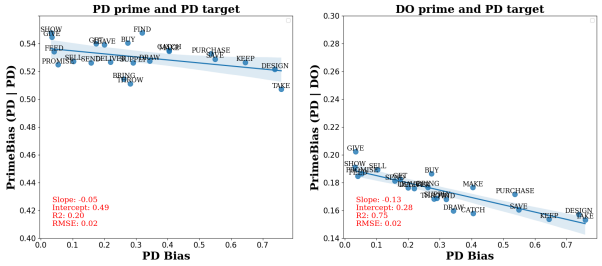

Figure 2: GPT2-small shows robust inverse frequency effect in the `Fine-tuning` mode.

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

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

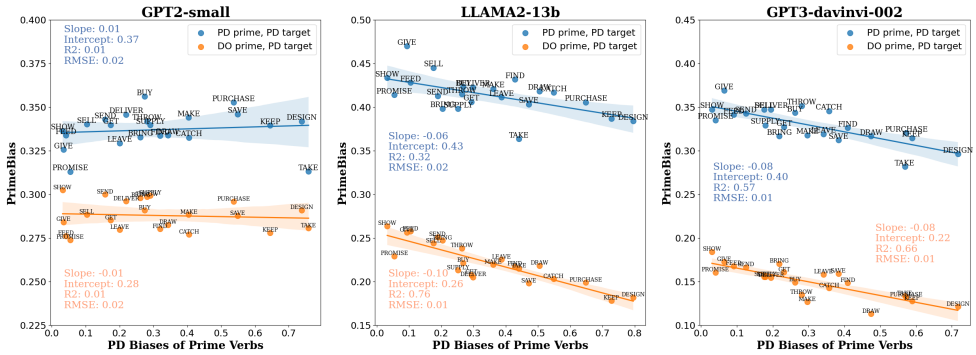

Figure 3: Larger models show stronger inverse frequency effect in the `Concatenation` mode.

Table 1: The slope, intercept, R^2, RMSE of the fitted lines for the `Concatenation` mode.

Models	With 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

Selected References [1] Sinclair, A., Jumelet, J., Zuidema, W., Fernández, R.: Structural persistence in language models: Priming as a window into abstract language representations. TACL 10 (2022) **[2]** Van Schijndel, M., Linzen, T.: A neural model of adaptation in reading. EMNLP (2018) **[3]** Bernolet, S., Hartsuiker, R.J.: Does verb bias modulate syntactic priming? Cognition 114(3) (2010) **[4]** Chang, F., Dell, G.S., Bock, K.: Becoming syntactic. Psych. Review 113(2) (2006) **[5]** Pickering, M.J., Branigan, H.P.: The representation of verbs: Evidence from syntactic priming in language production. JML 39(4) (1998) **[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. ICLR 2023