Is In-context Learning a Type of Error-Driven Learning?
Diagnosing with the Inverse Frequency Effect in Structural Priming

Zhenghao "Herbert" Zhou & Robert Frank & R. Thomas McCoy

Department of Linguistics, Yale University

Yale

TL;DR

- We show that LLMs display the inverse frequency effect in structural priming in the ICL setting, mirroring human language processing;
- Previous studies have argued that the inverse frequency effect implicates error-driven learning;
- We conclude that ICL in off-the-shelf LLMs can be viewed as a form of error-driven learning.

In-context Learning ≈*(functionally)* Gradient Descent?

In-context Learning (ICL) is an emergent property of Large Language Models (LLMs) that adapt to specific tasks given a few demonstration-answer pairs provided in the context window without any parameter updates (e.g., Brown et al. 2020). This differs from In-weights learning, which fine-tunes the model by updating model weights.

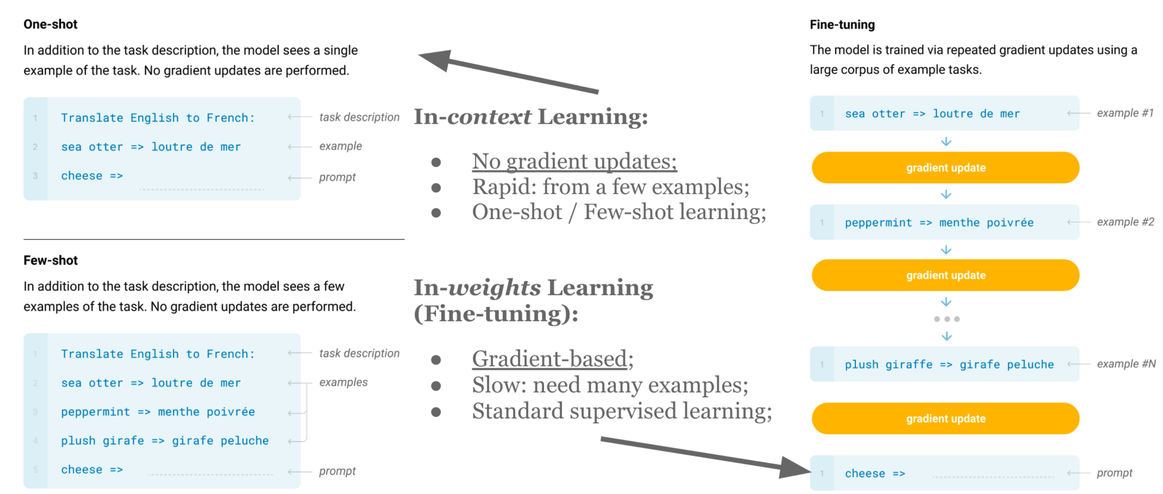


Figure 1. Reprinted from Brown et al. 2020

Research Question

How does ICL work? Is there an implicit gradient term computed in the forward pass (i.e., processing the context)?

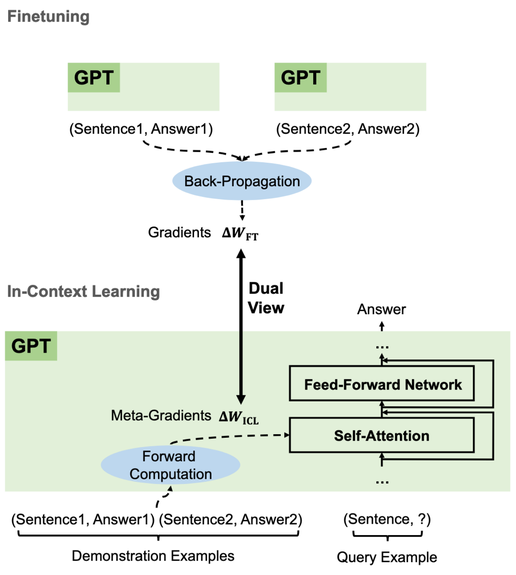


Figure 2. Reprinted from Dai et al. 2023

ICL has been interpreted as:

- Xie et al. 2022: performing implicit Bayesian inference;
- Von Oswald et al. 2023: functionally performing gradient descent;
- Dai et al. 2023: a meta-optimization process equivalent to implicit fine-tuning;

However, most previous studies:

- assume a training objective that optimizes for ICL;
- use hand-constructed weights for toy Transformer models;
- use non-natural language data;

The Inverse Frequency Effect in Human Structural Priming

Structural Priming: speakers tend to reuse the syntactic structures they have recently encountered during production or comprehension.

Inverse Frequency Effect: structural alternatives that are less frequent in language experience give rise to a stronger priming effect.

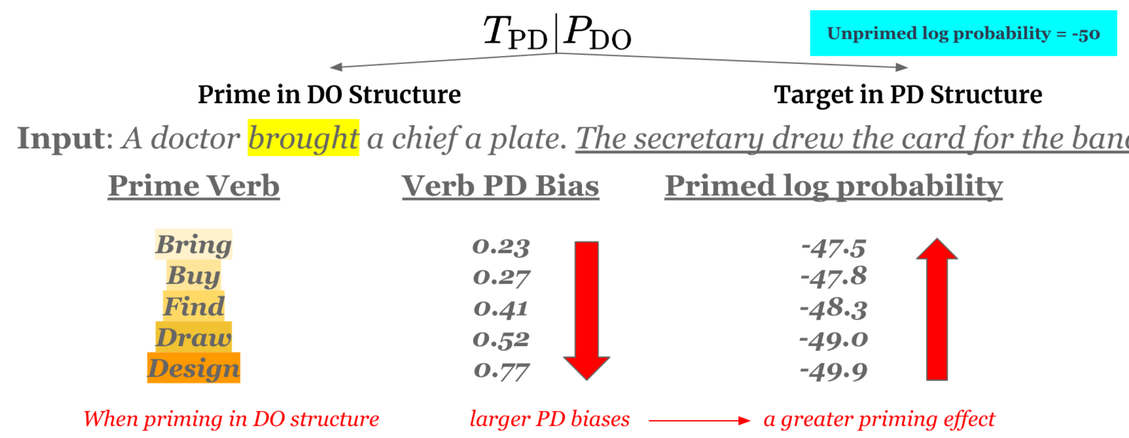
Implicit Learning Account of Priming: humans implicitly update the internal grammatical knowledge in an error-driven way based on prediction errors (i.e., the difference between expectation on each structure and the actual prime).

For the case of Dative Alternations:

(1) a. **Double Object**: Alice sent Bob a letter.

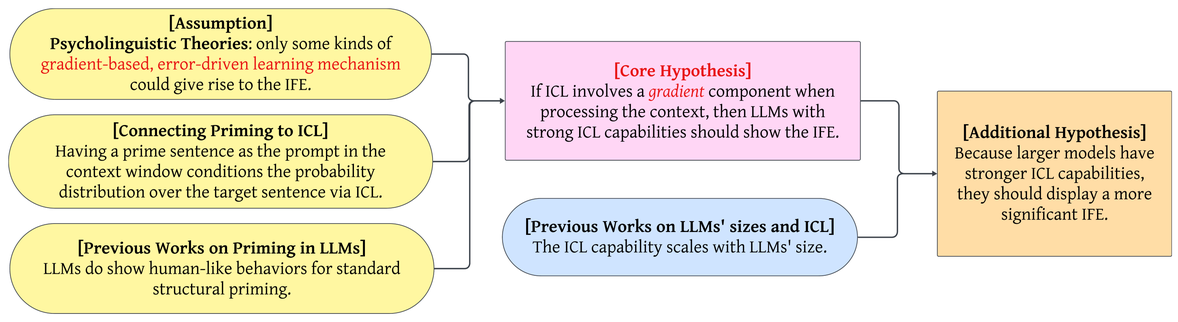
b. **Prepositional Dative**: Alice sent a letter to Bob.

Verb Bias: the probability distribution over the two structures for each dative verb.



Methodology and Experiment Overview

Reasoning Behind the Current Experiments



Materials

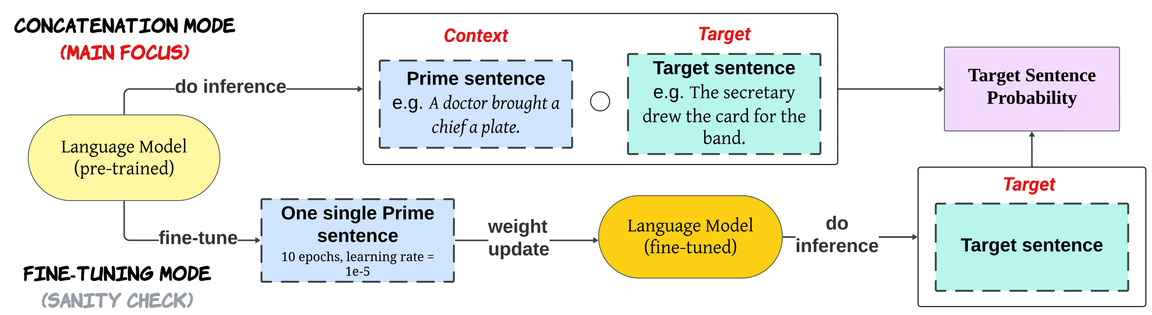
92400 Priming Trials:

- Target Sentences: 22 ditransitive verbs, 50 target sentences per verb;
- Prime Sentences: pair each target sentence with one prime sentence with each prime verb;
- 4 Structural Combinations: DO-DO, DO-PD, PD-DO, PD-PD;

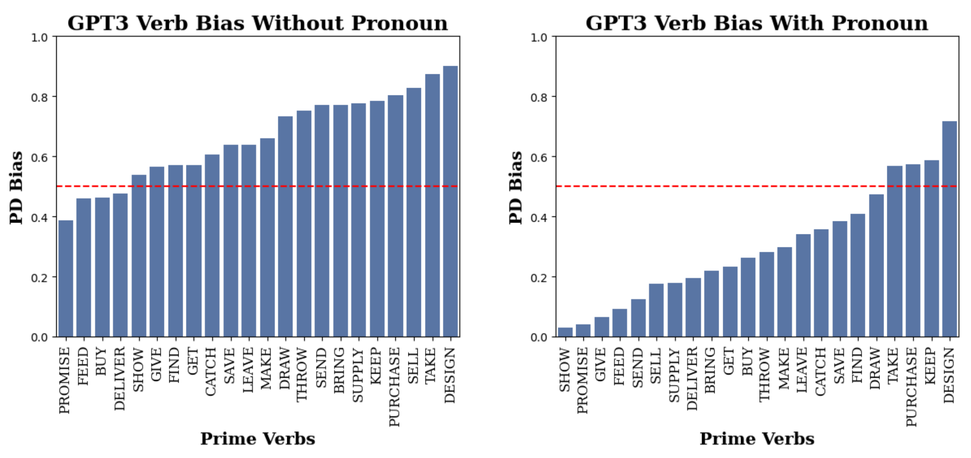
Language Models:

Family	Model	Size
GPT2	small	117M
	medium	345M
	large	762M
Llama2	7b	7B
	7b-chat	7B
	13b	13B
GPT3	davinci-002	175B

Two Modes

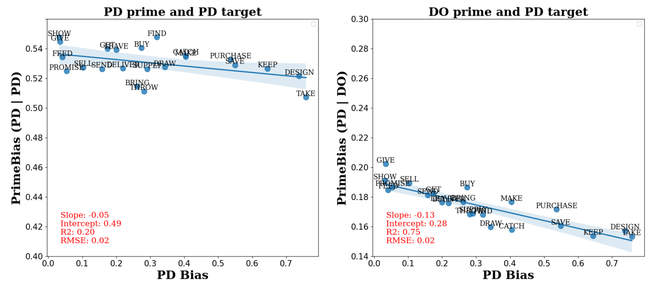


Verb Bias represented in LLMs



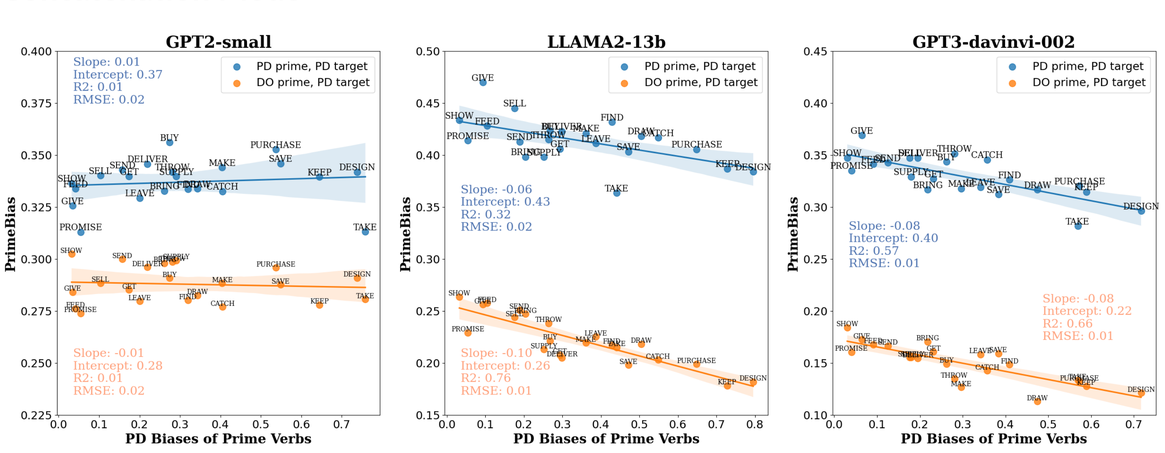
Results

Fine-Tuning Mode



- We only applied Fine-Tuning mode to GPT2-small;
- Even the smallest model shows significant IFE;
- Thus, the IFE does show up with explicit weight updates;

Concatenation Mode



- Models of all sizes show standard structural priming; *larger* models show more significant IFE.
- Thus, models with stronger ICL capability correspondingly show greater IFE — having greater capability of capturing the implicit gradient relevant to the verb bias without weight updates.

Implications & Future Directions

- We corroborate the hypothesis **"ICL ≈*(functionally)* GD"** in the case of structural priming with off-the-shelf LLMs and natural language data.
- Future: to apply the IFE diagnostic on other ICL tasks, and to find mechanistic level explanations and evidence for the existence of the implicit gradient.