

Causal Interventions on *Continuous* Features in LLMs: A Case Study on Verb Bias in Steering Vectors for ICL

Zhenghao Herbert Zhou & R. Thomas McCoy & Robert Frank

Department of Linguistics, Yale



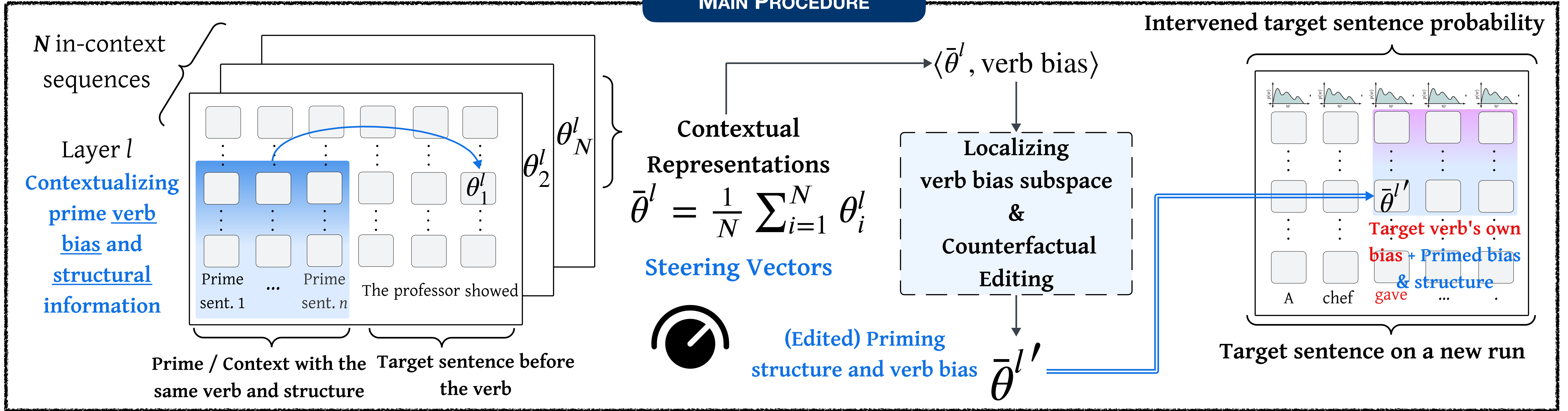
TL;DR

* **Continuous verb bias is causally editable**: changing the bias encoded in steering vectors systematically shifts the model's structural preferences.

* **Steering vectors reproduce structural priming but miss its error-driven signature**: error signals are recoverable and editable, yet are not involved in activation steering.

* **The distinction connects priming theories with ICL**: steering vectors resemble transient activation, error-driven learning may require richer mechanisms.

MAIN PROCEDURE



PRIMING & IN-CONTEXT LEARNING

1. Two ways to say the same thing:

DO: Alice gave Bob a book.

PD: Alice gave a book to Bob.

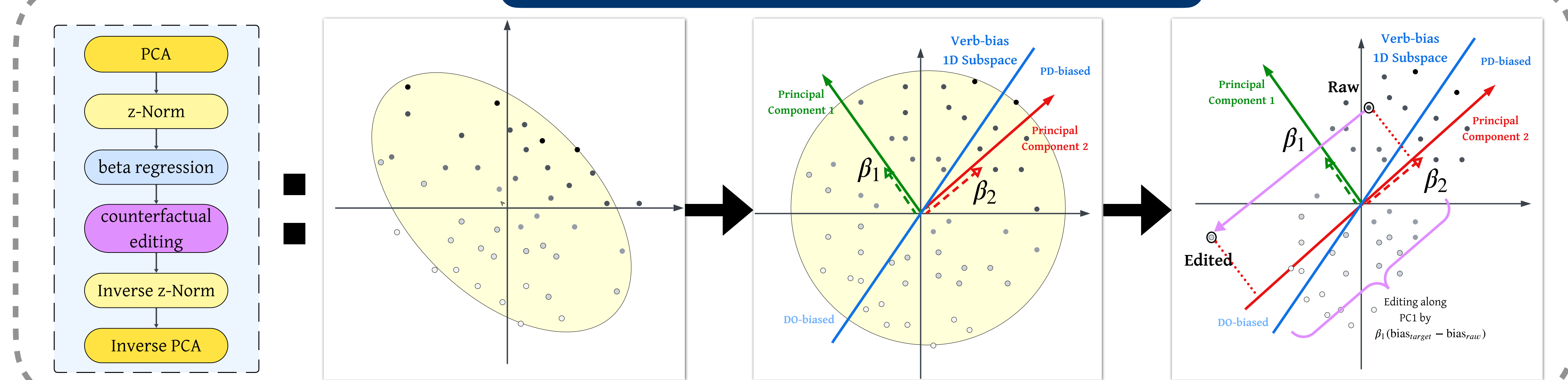
Verb bias: verbs differ in how strongly they favor each structure.

2. Two signatures of context:

Priming: more likely to reuse a recently encountered structure

Error-Driven Adaptation: learn more from a less expected structure

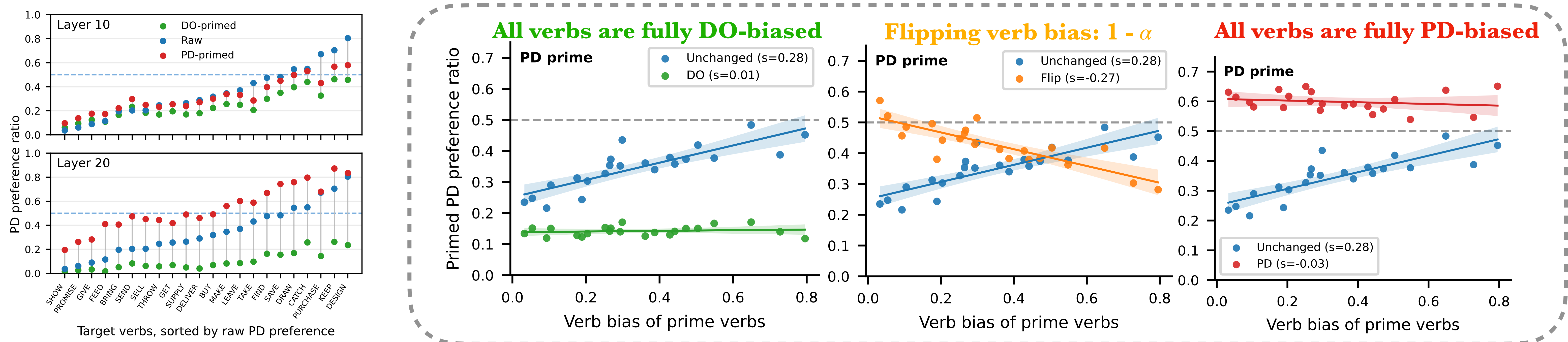
LOCALIZING & EDITING CONTINUOUS FEATURES



RESULT 1 / WHAT STEERING CAPTURES

Steering vectors induce causally editable structural preferences.

Context-derived vectors induce structural priming. Manipulating encoded verb bias induces predicted shifts in downstream structural preference.

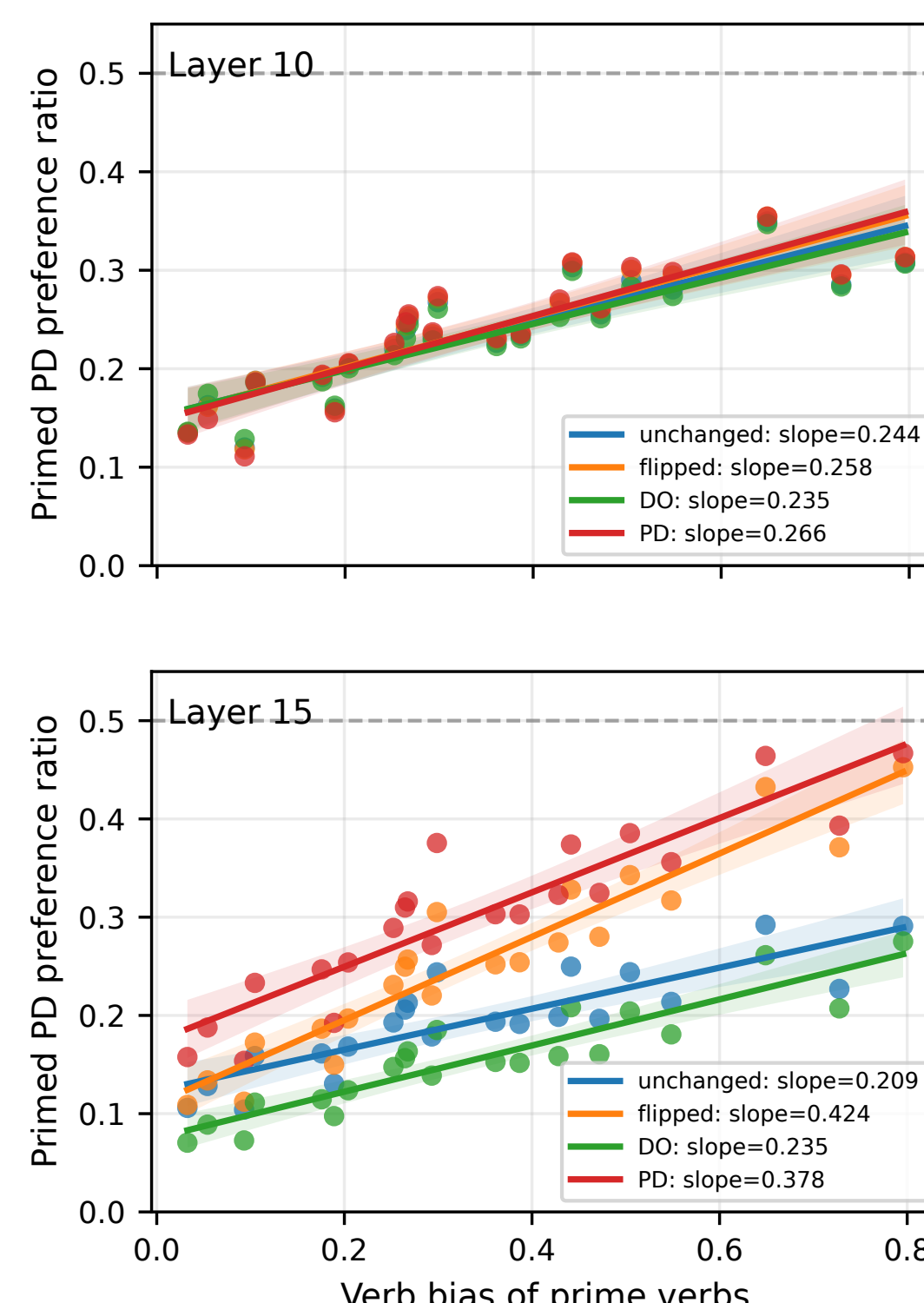


RESULT 2 / WHAT STEERING MISSES

Steering vectors do NOT reproduce error-driven adaptation.

- ▶ Error-related information is recoverable starting at intermediate layers.
- ▶ Baseline steer-vector injection does not produce stronger learning / adaptation from more surprising / unexpected primes.
- ▶ Editing error signals shifts downstream learning behavior in predicted directions.

Accessible information does not guarantee causal engagement for adaptive ICL behavior



IMPLICATION / WHAT THIS SUGGESTS ABOUT PRIMING AND ICL

