Causal Interventions on Continuous Features in LLMs: A Case Study in Verb Bias

Zhenghao Herbert Zhou & R. Thomas McCoy & Robert Frank *Department of Linguistics, Yale University*

TL;DR

- **We show that verb bias, a *continuous and context-dependent* feature, is compactly represented in LLMs, and gradient, counterfactual manipulations of the associated subspaces play a predictable *causal* role in downstream structural choices, eliciting *structural priming* effects.**
- **We highlight the potential of combining causal interventions with psycholinguistic paradigms to yield deeper insights into the interpretability of the underlying mechanisms in LLMs.**

Background: Strucrtural Priming in LLMs

- **Structural Priming**: speakers tend to reuse the recently encountered syntactic structures during production and comprehension.
- Consider the classical Dative Alternations as a case study:
  - **Double Object (DO):** *Alice sent Bob a letter.*
  - **Prepositional Dative (PD):** *Alice sent a letter to Bob.*
- **Verb Bias**: the probability distribution over the two structures for each dative verb (e.g. *bring* is a highly DO-biased word).

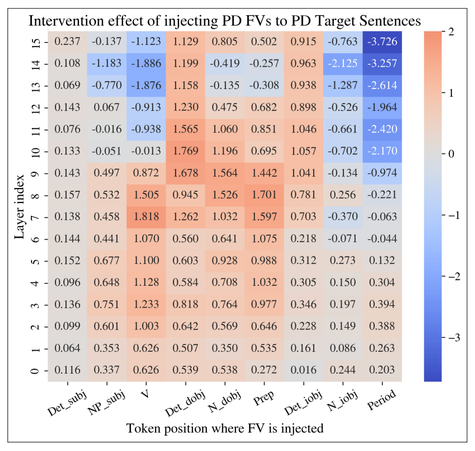
Previous studies have shown that LLMs show human-like structural priming.

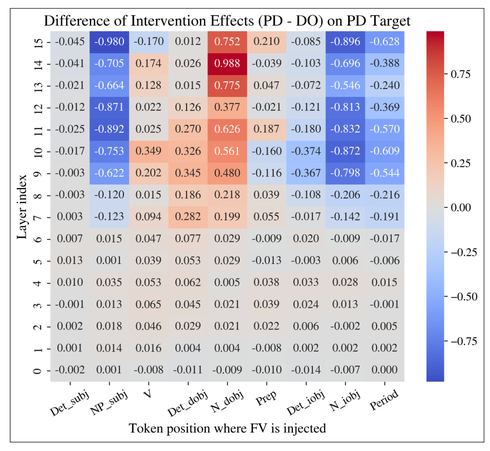
Zhou et al. 2025 have proposed that:

- ✦ **LLMs' ICL can be {viewed as, a product of} human structural priming.**
- ✦ **ICL ≈*(functionally)* Gradient Descent as error-driven learning.**

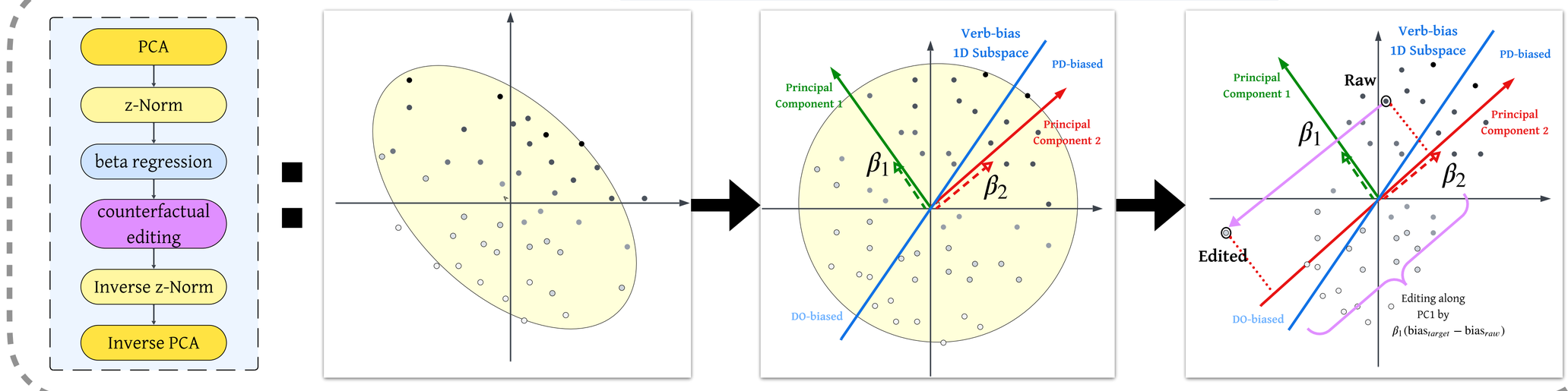
Exp1: Eliciting Similar Priming Behaviors

- Structural priming effect is observed via injecting compact contexutal representation into a new inference run.
- Intervention Effect: difference between raw and intervened sentence probability.





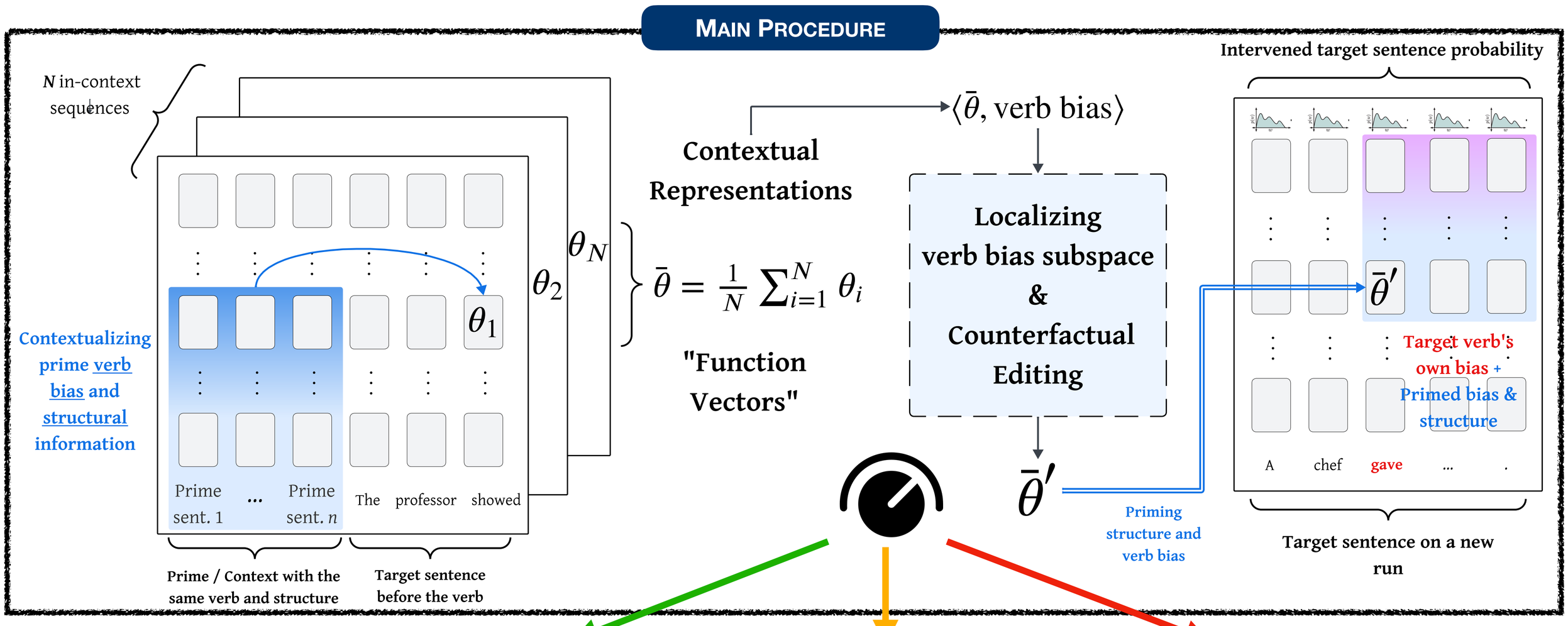
Localing & Editing Continuous Features



Relaed Work

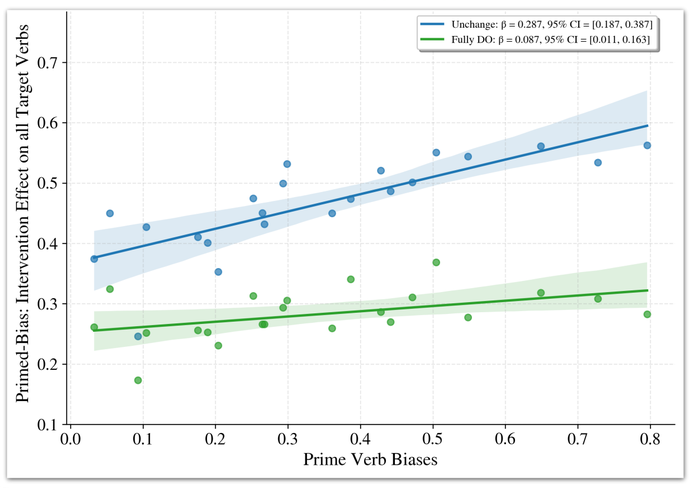
- Function vectors and In-Context Learning;
- LEAst-squares Concept Erasing;
- The Linear Representation Hypothesis;

Main Procedure

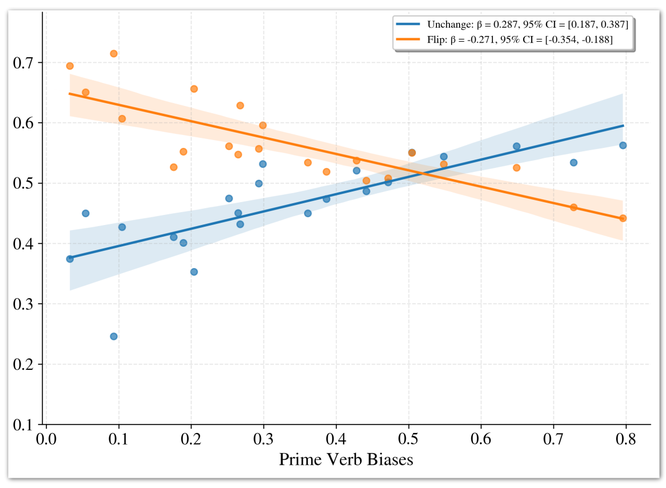


Exp2: Localizing & Causal Intervention of Verb Bias

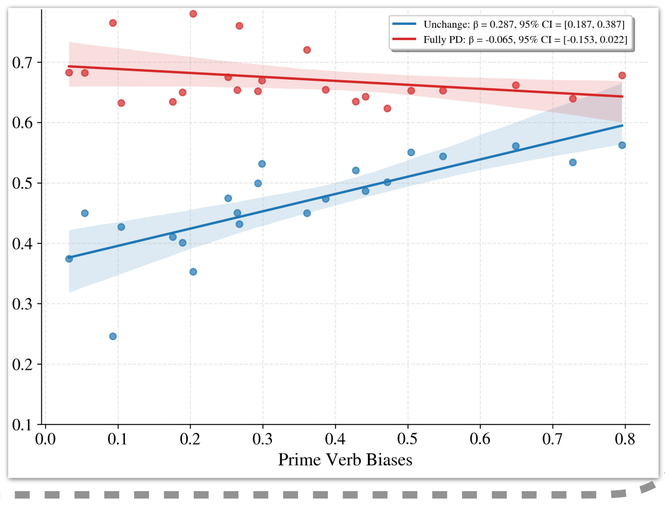
All verbs are fully DO-biased



Flipping verb bias: 1 - α



All verbs are fully PD-biased



NEURAL INFORMATION PROCESSING SYSTEMS

Correspond to: herbert.zhou@yale.edu