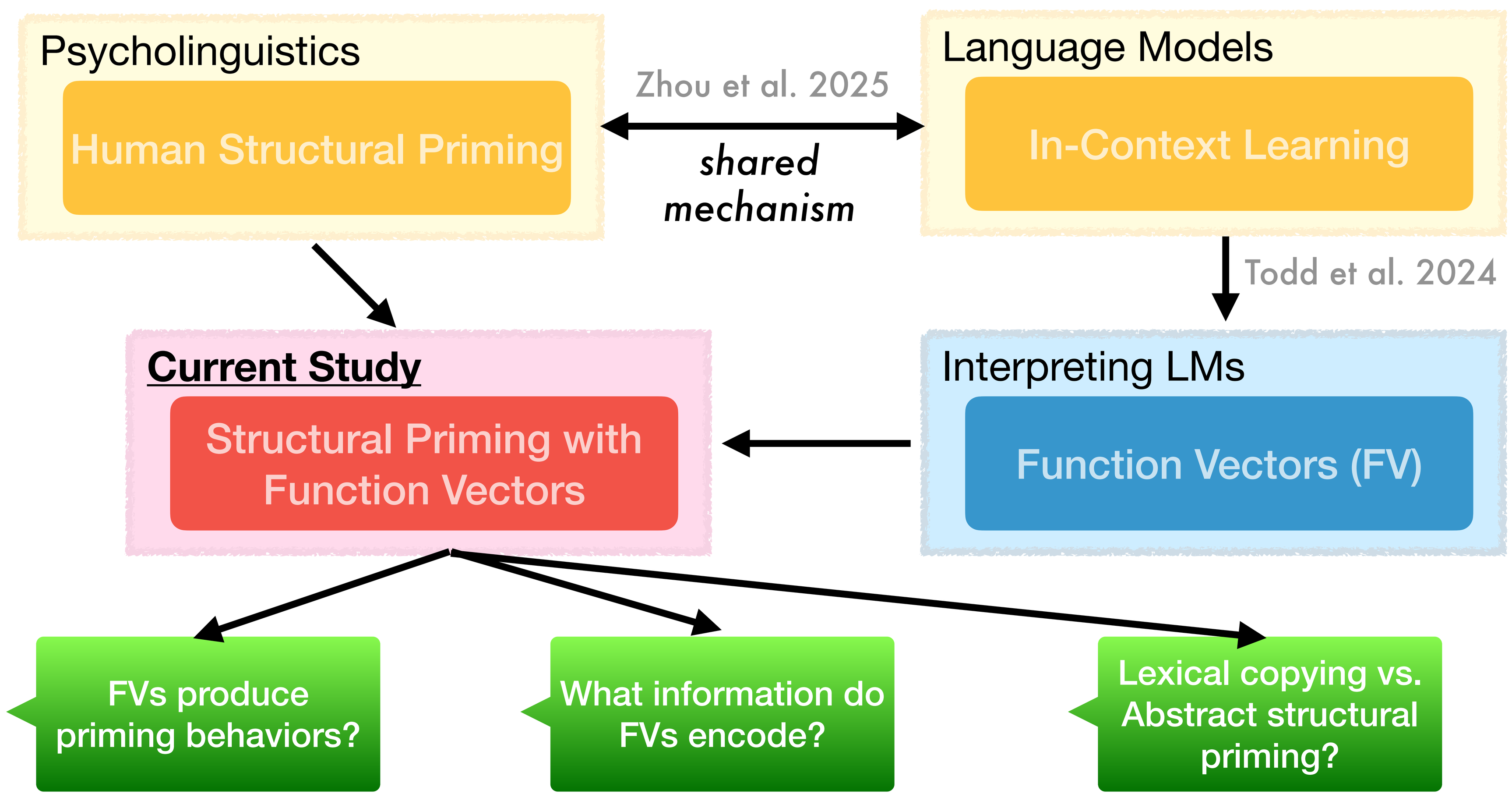
Compressing Strucural Priming in Large Language Models through Function Vectors

Zhenghao Herbert Zhou & R. Thomas McCoy & Robert Frank

Department of Linguistics, Yale University

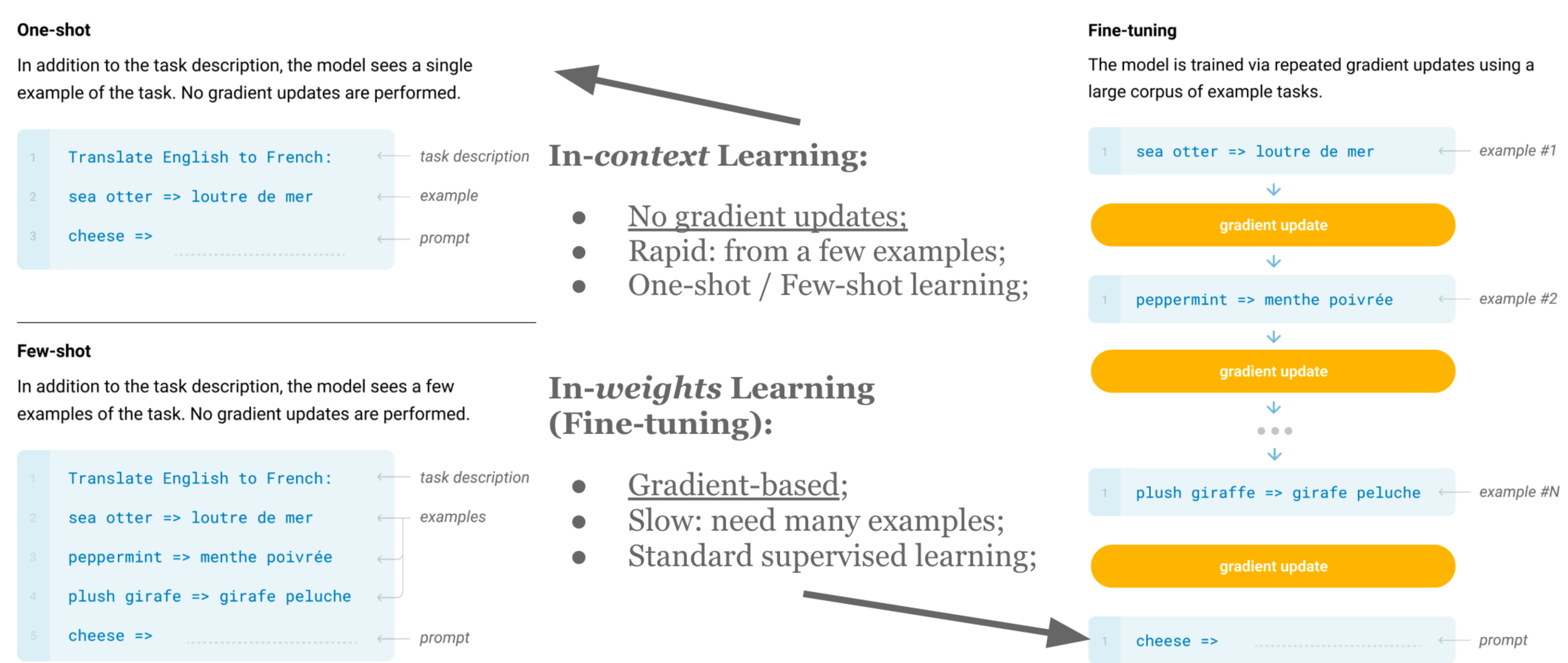
Yale

Overview



Background 1: In-Context Learning (ICL) in LLMs

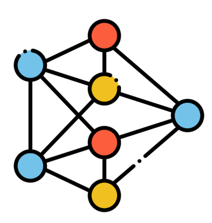
In-Context Learning (ICL): an emergent property for LLMs to adapt to tasks *at inference time* with a few demo-answer pairs without weight updates.



Background 2: Strucrtural Priming in LLMs

- **Structural Priming**: speakers tend to reuse the recently encountered syntactic structures during production and comprehension.
- **Inverse Frequency Effect (IFE)**: structural alternatives with less frequency are susceptible to a stronger priming effect than the more frequent ones.
- **Implicit Learning Account of Priming**: humans implicitly update the internal grammatical knowledge in an *error-driven way* based on prediction errors (the difference between expectation and actual prime instances).

Consider the classical Dative Alternations as a case study:



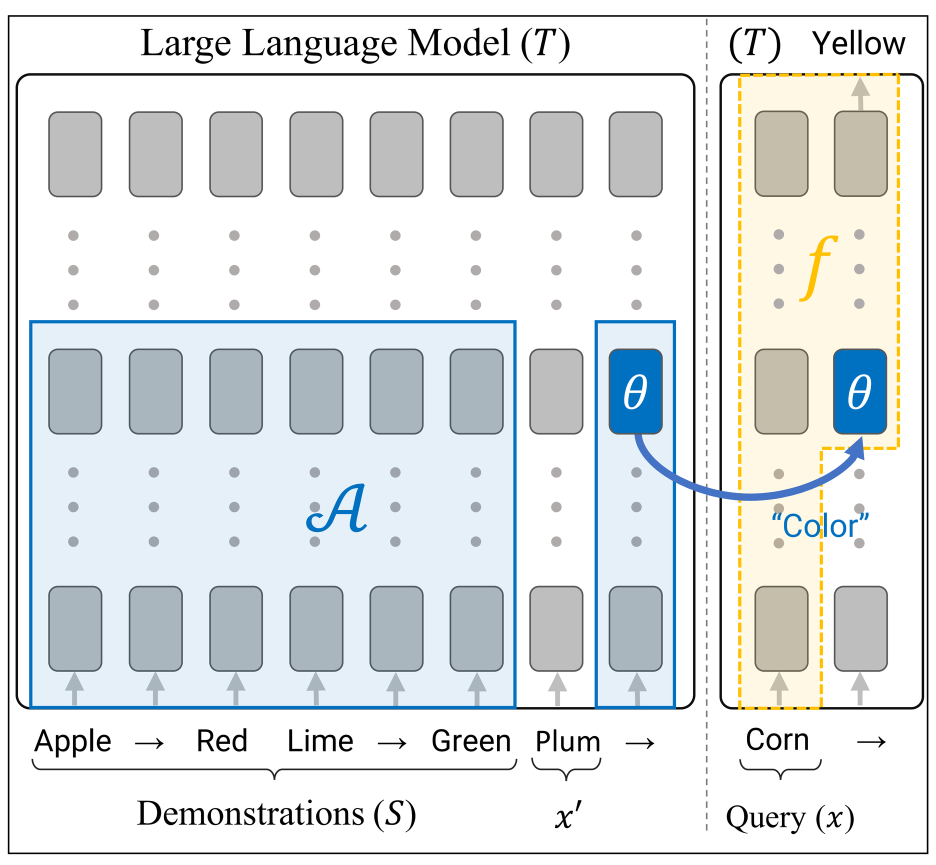
- **Double Object (DO)**: Alice sent Bob a letter.
- **Prepositional Dative (PD):** Alice sent a letter to Bob.

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: In particular, 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.**

Background 3: Function Vectors Capture ICL



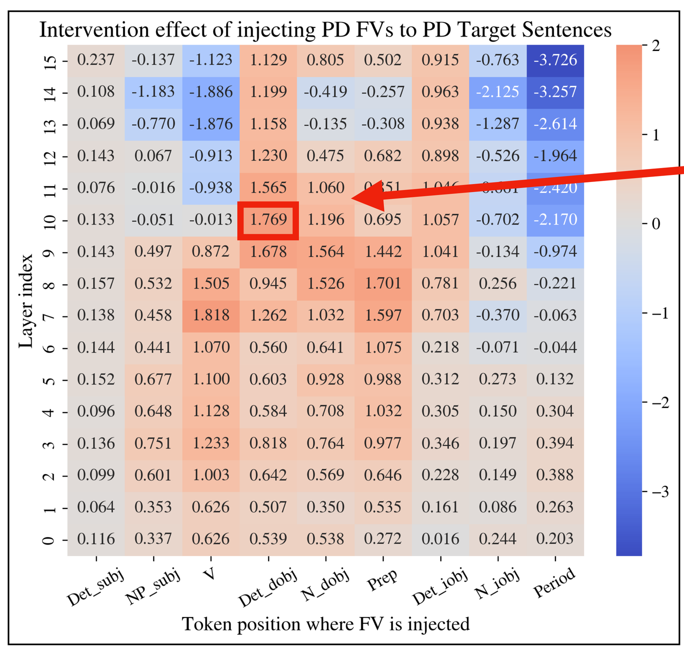
Reprinted from Hendel et al. (2023).

Function vectors (**FV**; *Hendel et al. 2023, Todd et al. 2024*) are compact, causal, internal representations of *function abstractions* extracted from LLMs.

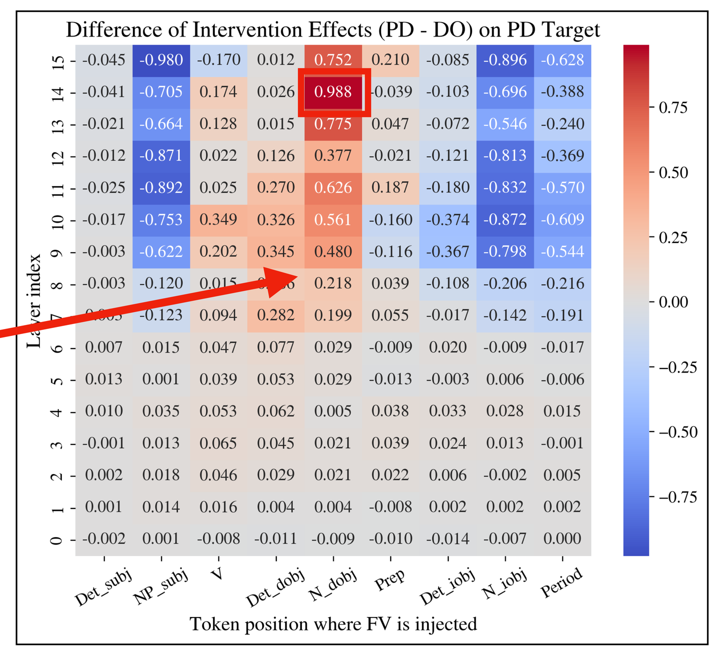
- Intermediate activation patterns capturing the "task" information in the demo-answer context.
- Compositional: FVs could be arithmetically composed to represent task combinations.
- Enables us to extract "internal knowledge" LLMs gain on the fly for causal intervention.

Exp1: FVs Elicit Similar Priming Behaviors?

- Extracting FVs from PD sequences and injecting them to the corresponding positions (layer and token position) in a Target PD sequence.
- Intervention Effect: difference between raw and intervened sentence probability.



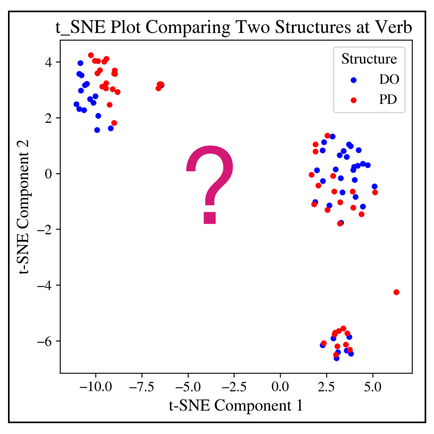
- Observing the priming effect: FVs increased target sentence probability;
- PD FVs cause a larger intervention effect than DO FVs on PD targets.

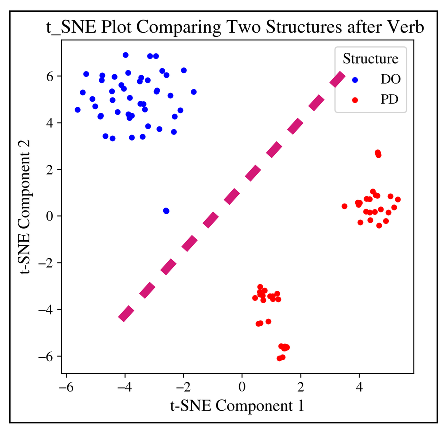


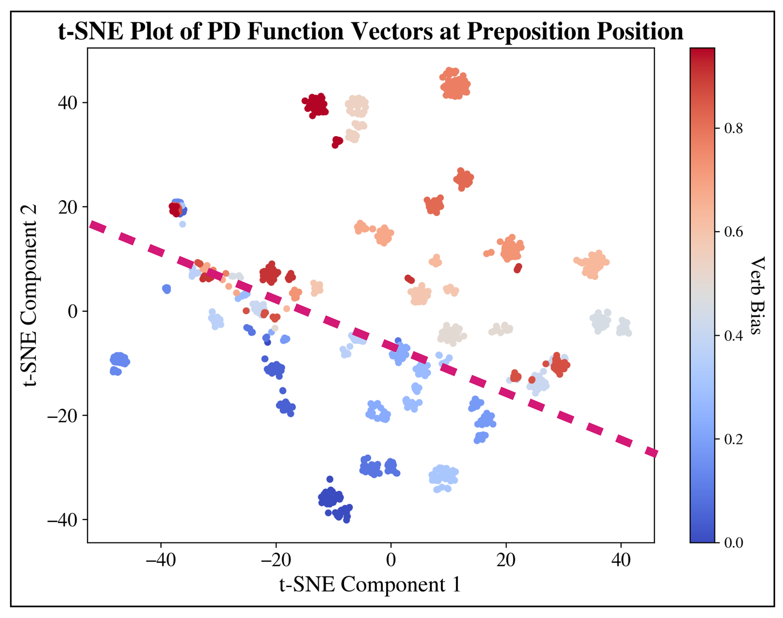
Takeaway: ✅ FVs do elicit standard structural priming effect.

Exp2: Verb Bias Information in FVs?

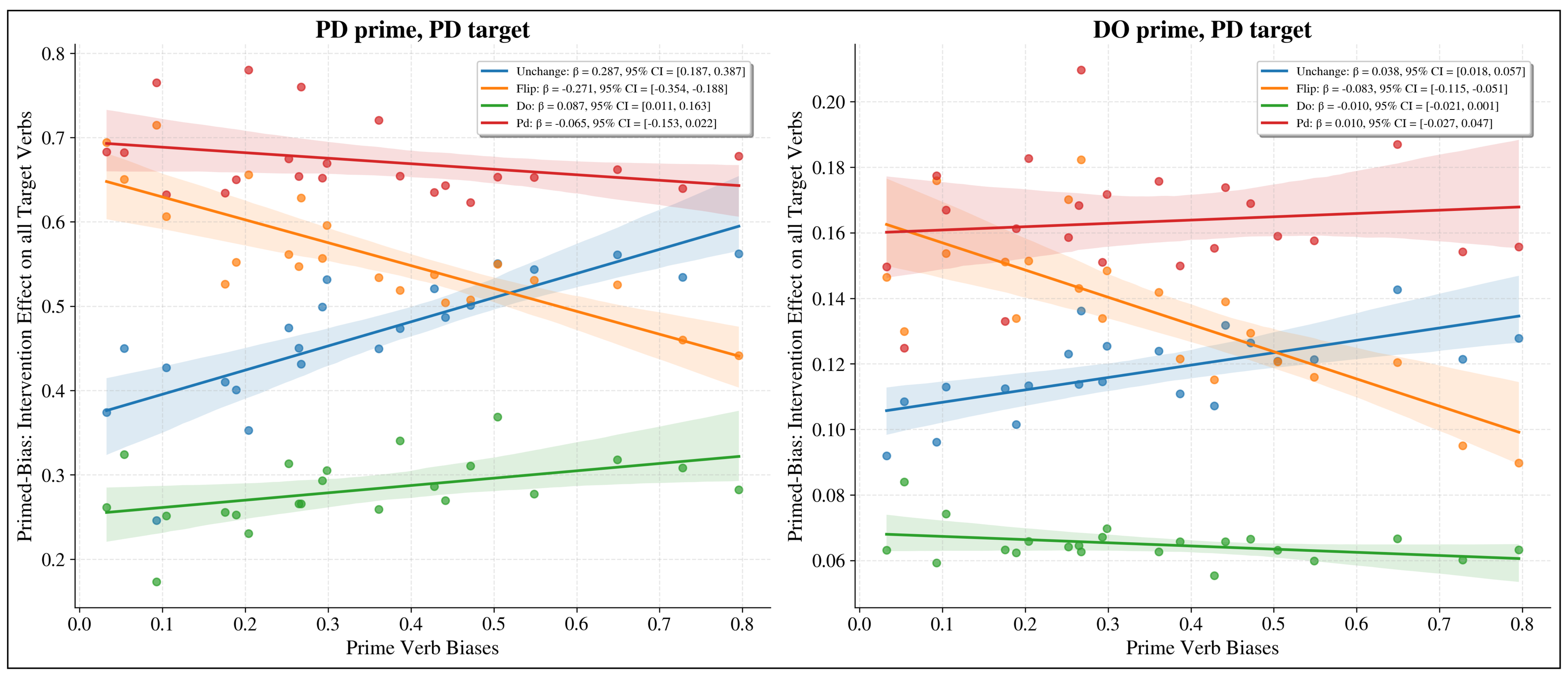
- Apply t-SNE on FVs and plot by structure / gradient verb biases.







- FVs encode structural information;
- FVs encode more fine-grained, graded verb bias information for each verb.
- Apply β-regression on the FVs to identify a linear subspace for verb bias, and modify the subspace to change the verb bias information in FVs. Measure priming strength.

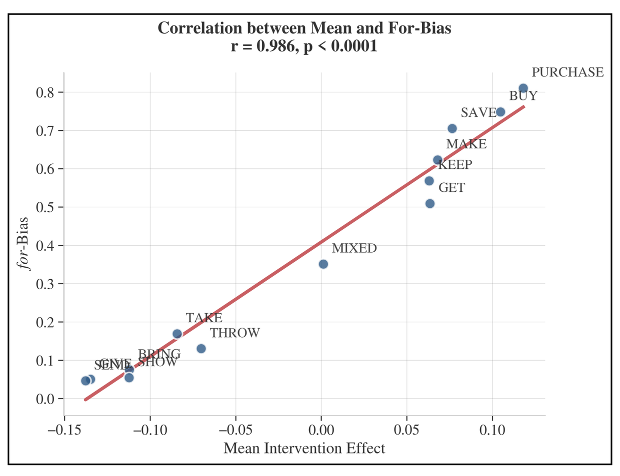


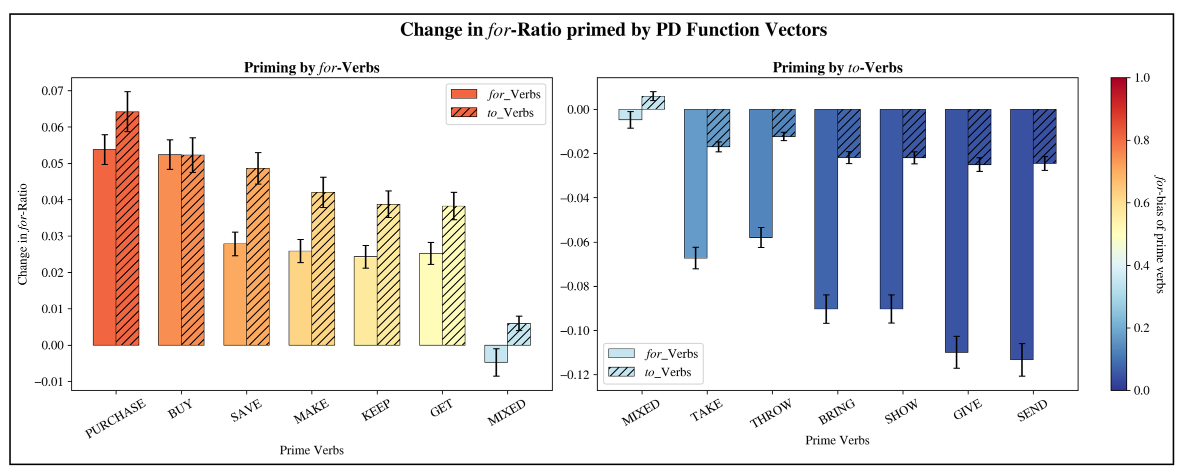
Takeaway: 🌟 FVs encode fine-grained verb bias information in a linear subspace that is causally manipulable to affect priming strength.

Exp3: Abstruct/Structural vs. Lexical Priming?

What levels of priming do FVs implement?

- Priming **abstract dative structure** → increase the probability of the *target* verb's preposition;
- Priming **lexical associations** → increase the probability of the *prime* verb's preposition;





Result: *for*-biased prime verbs increase the *for*-preference for both *for*-biased and *to*-biased target verbs. Same for *to*-biased prime verbs.

Takeaway: ⁉️ lexical-specific information is prioritized over structural information in the case of preposition preference.

Conclusions

- It is viable to compress structural repetitions in the context into function vectors, which elicit comparable structural priming effects in LLMs.
- Verb bias information is encoded in a manipulable linear subspace.
- FVs carry both abstract structural-level and lexical-specific information.
- FVs offer a mechanisitc level way of causally intervening internal representations in LLMs.

Poster Session, LSA Summer Institute 2025, OR USA

Correspond to: herbert.zhou@yale.edu