Causal Interventions on *Continuous* Variables: A Case Study on Verb Bias in Steering Vectors for In-Context Learning

Zhenghao Herbert Zhou, R. Thomas McCoy, and Robert Frank
EMNLP 2026, Budapest, Hungary
October 25, 2026



Research Question

Ordinary in-context learning (ICL)

Context

hot → **cold**

up → **down**

day → **night**

[e.g., Brown et al. 2020]

Research Question

Ordinary in-context learning (ICL)

Context

hot → cold

up → down

day → night

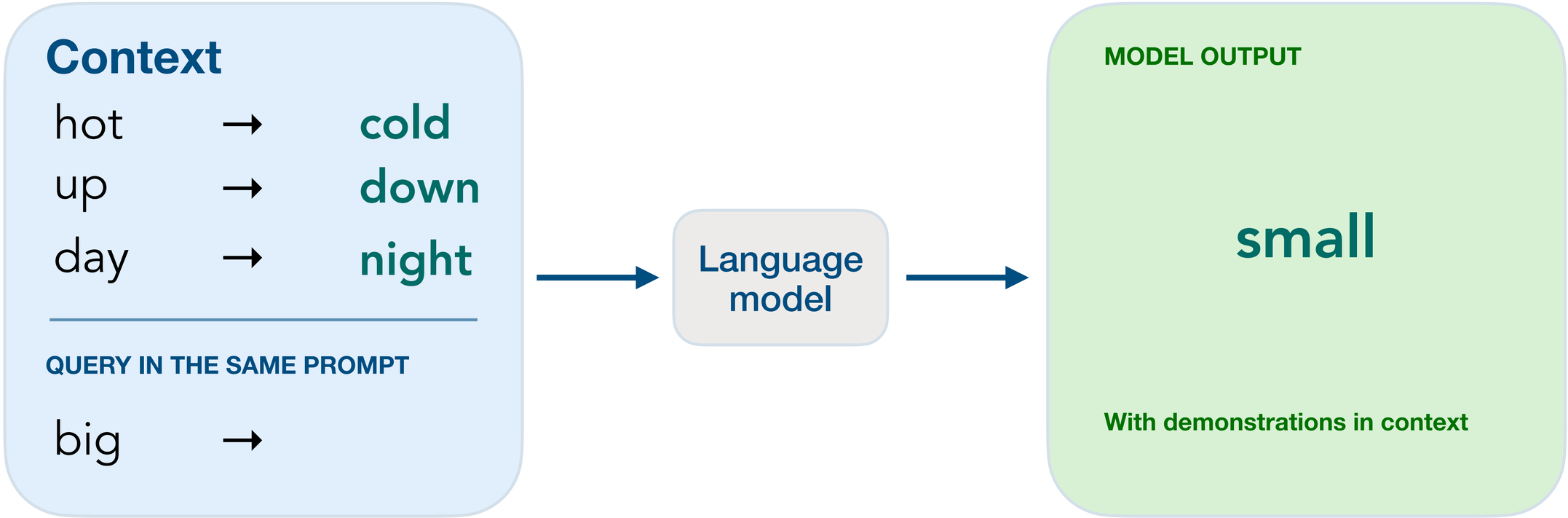
QUERY IN THE SAME PROMPT

big →

[e.g., Brown et al. 2020]

Research Question

Ordinary in-context learning (ICL)



[e.g., Brown et al. 2020]

Research Question

ICL with steering vectors

Context

hot → cold

up → down

day → night

Run the model on the context.
Extract from its activations.



[Handel et al. 2023; Todd et al. 2024]

Research Question

ICL with steering vectors

Context

hot → cold
up → down
day → night

Run the model on the context.
Extract from its activations.

Extract



Steering vector

[Handel et al. 2023; Todd et al. 2024]

Research Question

ICL with steering vectors

Context

hot → cold
up → down
day → night

Run the model on the context.
Extract from its activations.



[Handel et al. 2023; Todd et al. 2024]

Research Question

ICL with steering vectors

Context

hot → cold
up → down
day → night

Run the model on the context.
Extract from its activations.

Extract →  → Inject

Steering vector

NEW RUN

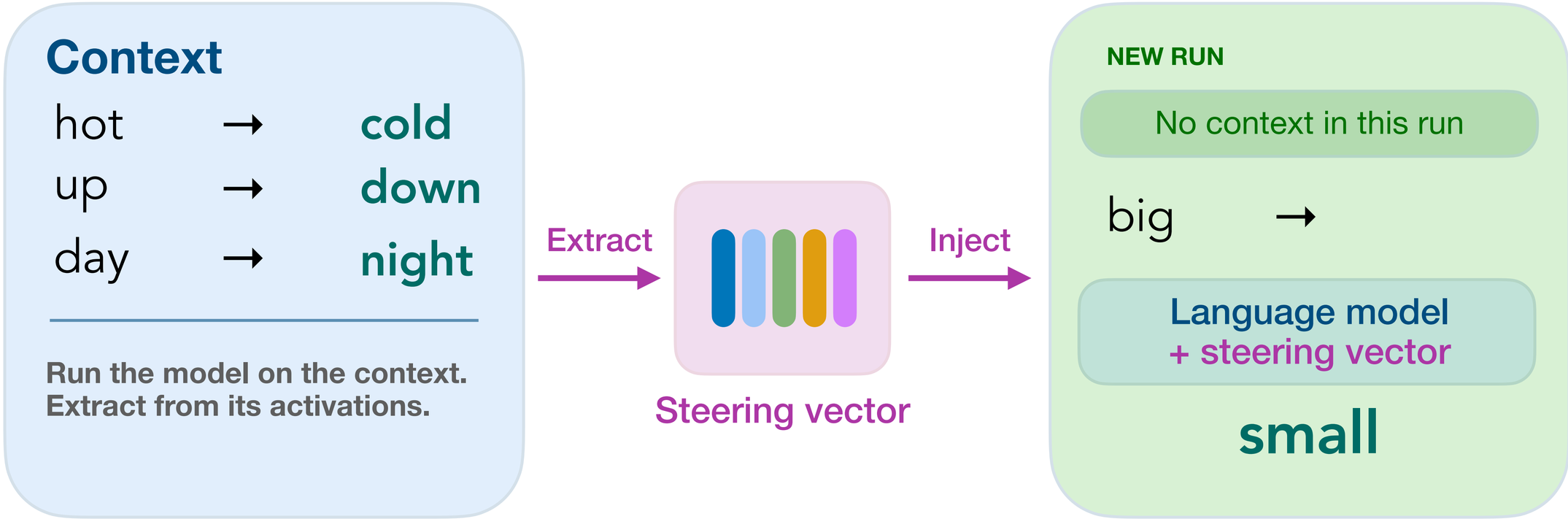
No context in this run

big →

[Handel et al. 2023; Todd et al. 2024]

Research Question

ICL with steering vectors



With context-derived steering vectors, the ICL behavior transfers even without demonstrations in-context.

[Handel et al. 2023; Todd et al. 2024]

Research Question

ICL with steering vectors

Context

hot → cold
up → down
day → night

Run the model on the context.
Extract from its activations.

Extract → Steering vector → Inject

NEW RUN

No context in this run

big →

Language model
+ steering vector

small

Research Question

ICL with steering vectors

Context

hot → cold
up → down
day → night

Run the model on the context.
Extract from its activations.

Extract → Steering vector → Inject

NEW RUN

No context in this run

big →

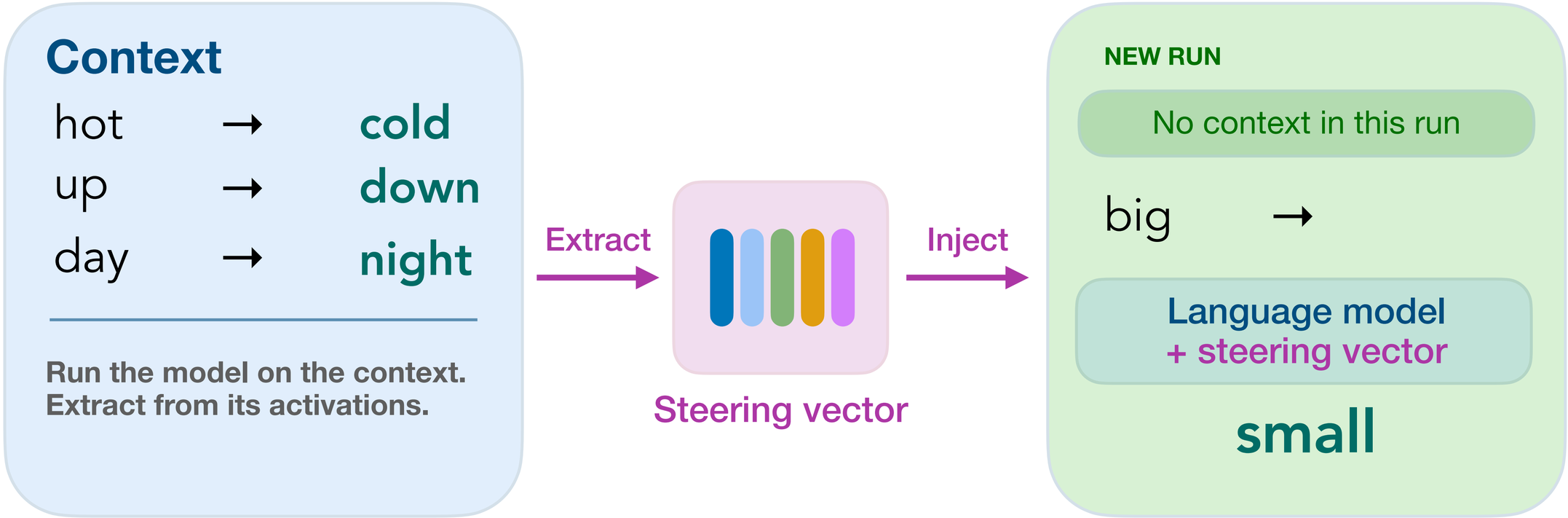
Language model
+ steering vector

small

How much of the context's *information* and *influence* do steering vectors preserve?

Research Question

ICL with steering vectors



How much of the context's *information* and *influence* do steering vectors preserve?

⇒ ***Structural Priming (from psycholinguistics)***

Structural Priming

The same event, two sentence structures

Alice *gave* | Bob | a book .

Structural Priming

The same event, two sentence structures



DO: double object

Structural Priming

The same event, two sentence structures

DO Alice *gave* | Bob | a book | .

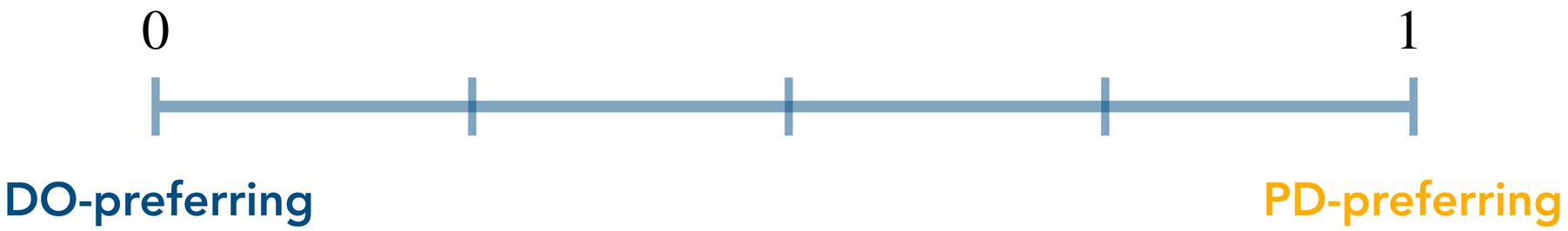
PD Alice *gave* | a book | to Bob | .

DO: double object **PD**: prepositional dative

Structural Priming & *Continuous* Verb Bias

DO Alice *gave* | Bob | a book | .

PD Alice *gave* | a book | to Bob | .

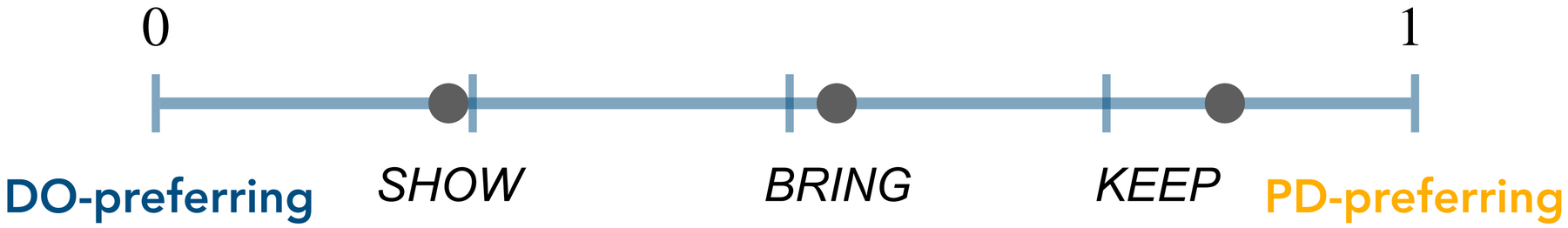


Structural Priming & *Continuous* Verb Bias

DO Alice *gave* | Bob | a book | .

PD Alice *gave* | a book | to Bob | .

Verb bias How strongly a verb favors one structure over the other

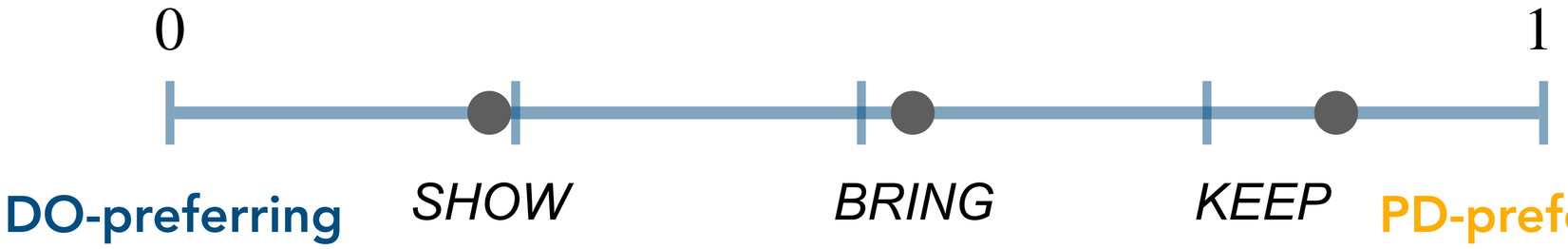


Structural Priming & *Continuous* Verb Bias

DO Alice *gave* | Bob | a book | .

PD Alice *gave* | a book | to Bob | .

Verb bias How strongly a verb favors one structure over the other

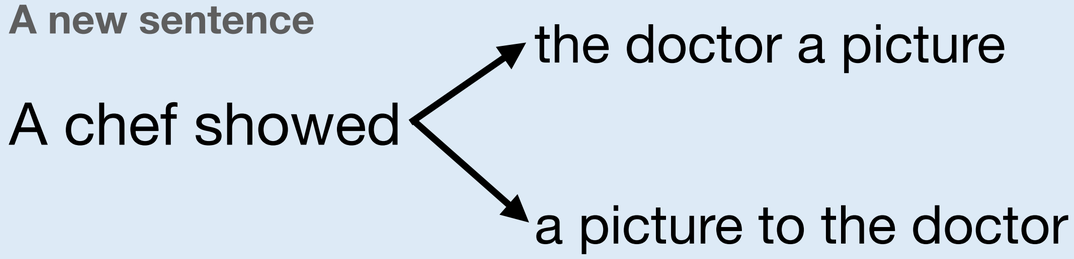


Linguistic knowledge from the frequency of usages for DO & PD

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming



[e.g., Humans: Bock 1986, Pickering & Branigan 1998; LMs: van Schijndel & Linzen 2018, Sinclair et al. 2022]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

PD Alice gave a book to Bob.

Encountering a PD structure in context.

A new sentence

A chef showed → the doctor a picture

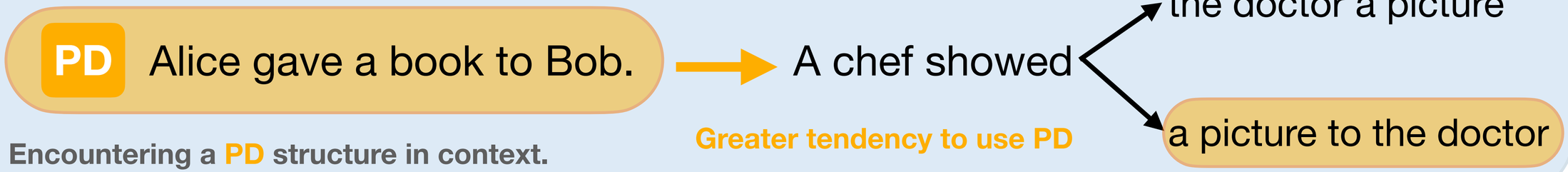
A chef showed → a picture to the doctor

[e.g., Humans: Bock 1986, Pickering & Branigan 1998; LMs: van Schijndel & Linzen 2018, Sinclair et al. 2022]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

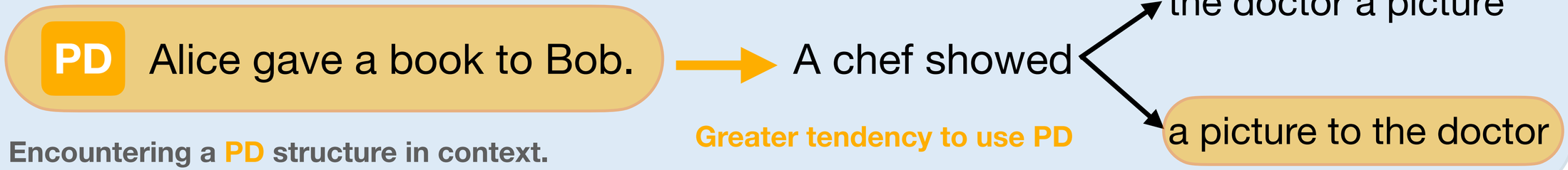


[e.g., Humans: Bock 1986, Pickering & Branigan 1998; LMs: van Schijndel & Linzen 2018, Sinclair et al. 2022]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

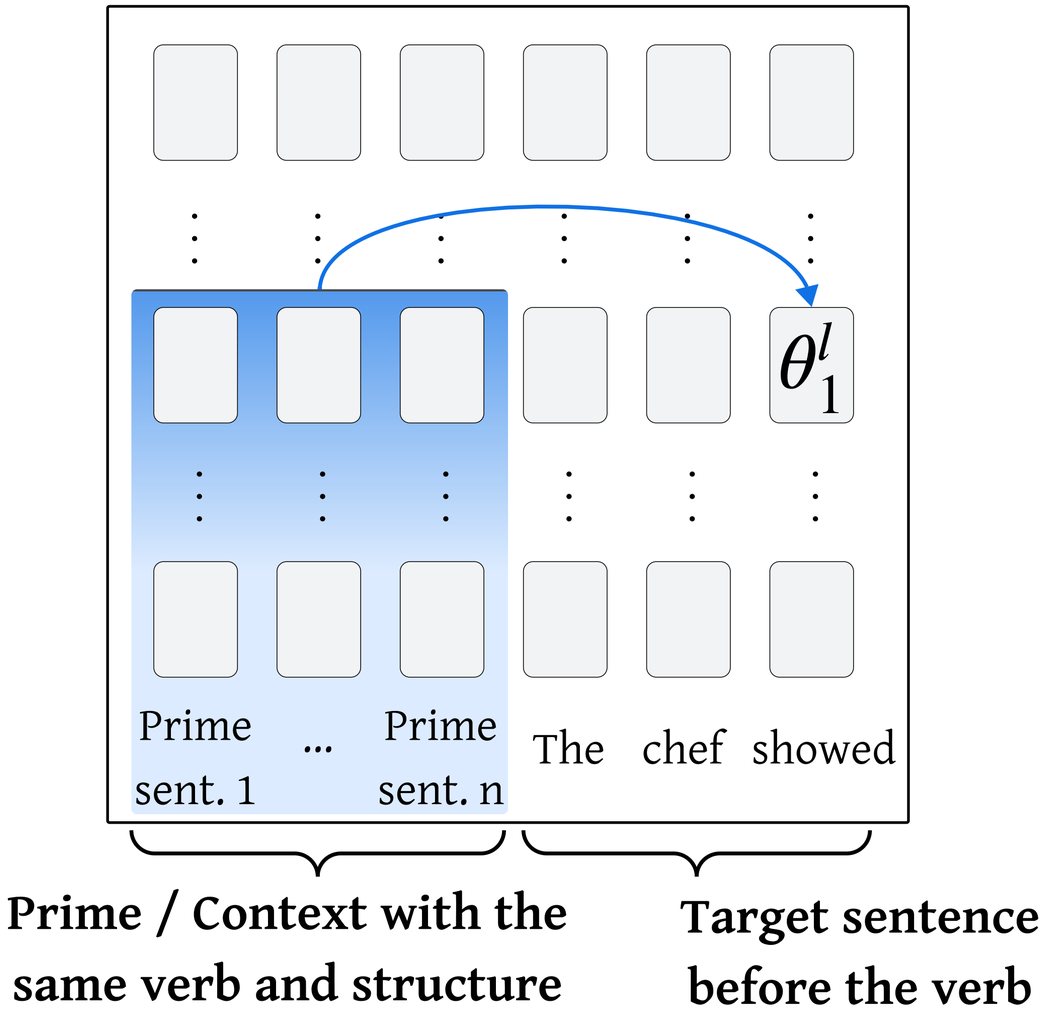


Content: context affects / steers downstream structural preference!

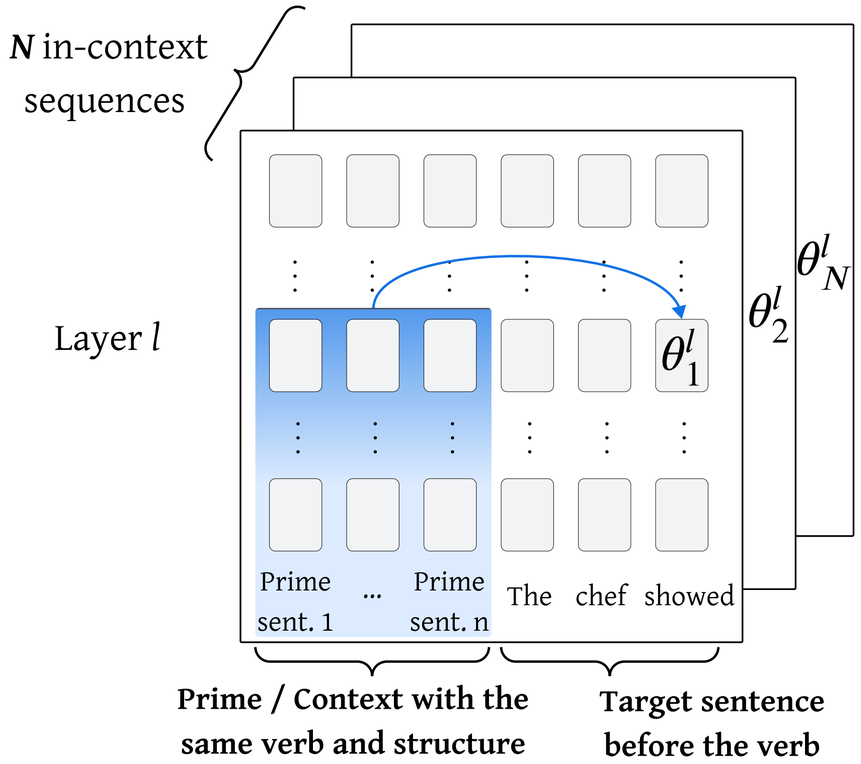
[e.g., Humans: Bock 1986, Pickering & Branigan 1998; LMs: van Schijndel & Linzen 2018, Sinclair et al. 2022]

Experiment Setup: Steering Vector Extraction

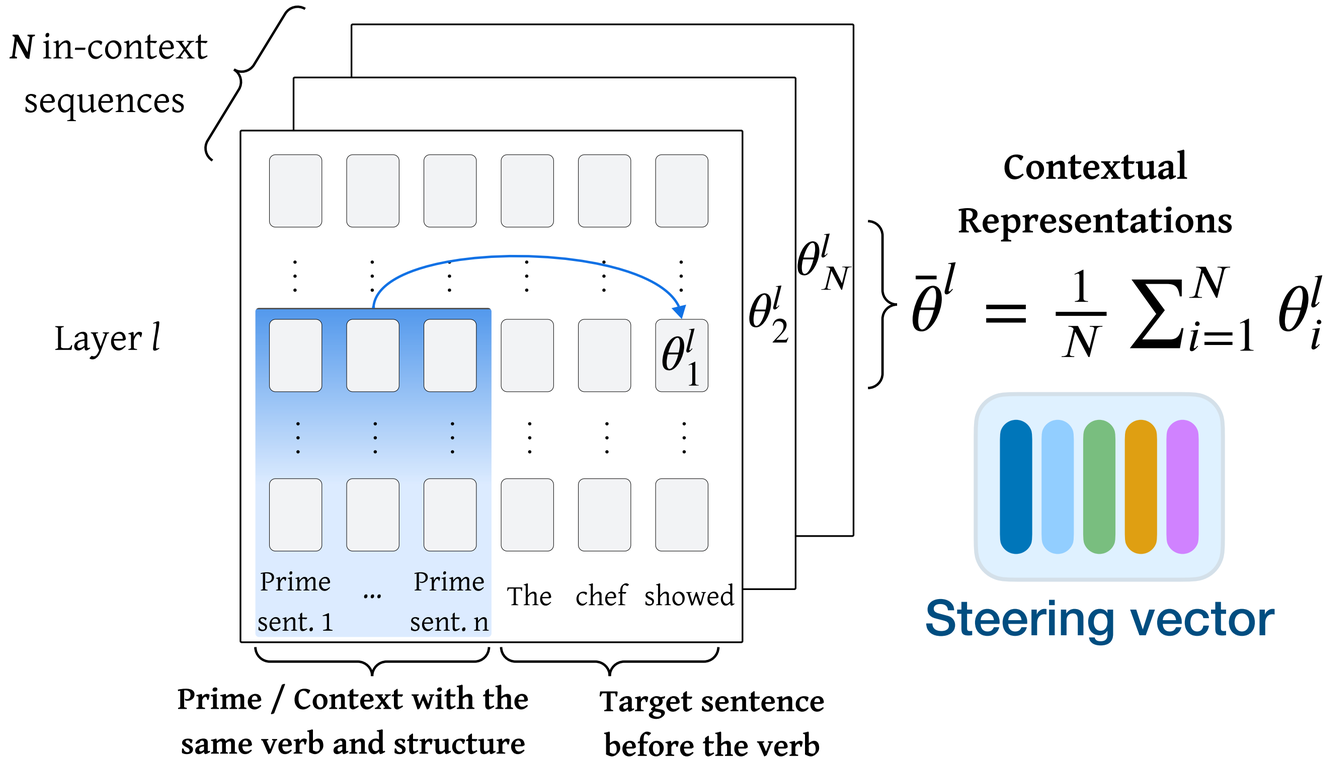
Contextualizing verb bias and structural information



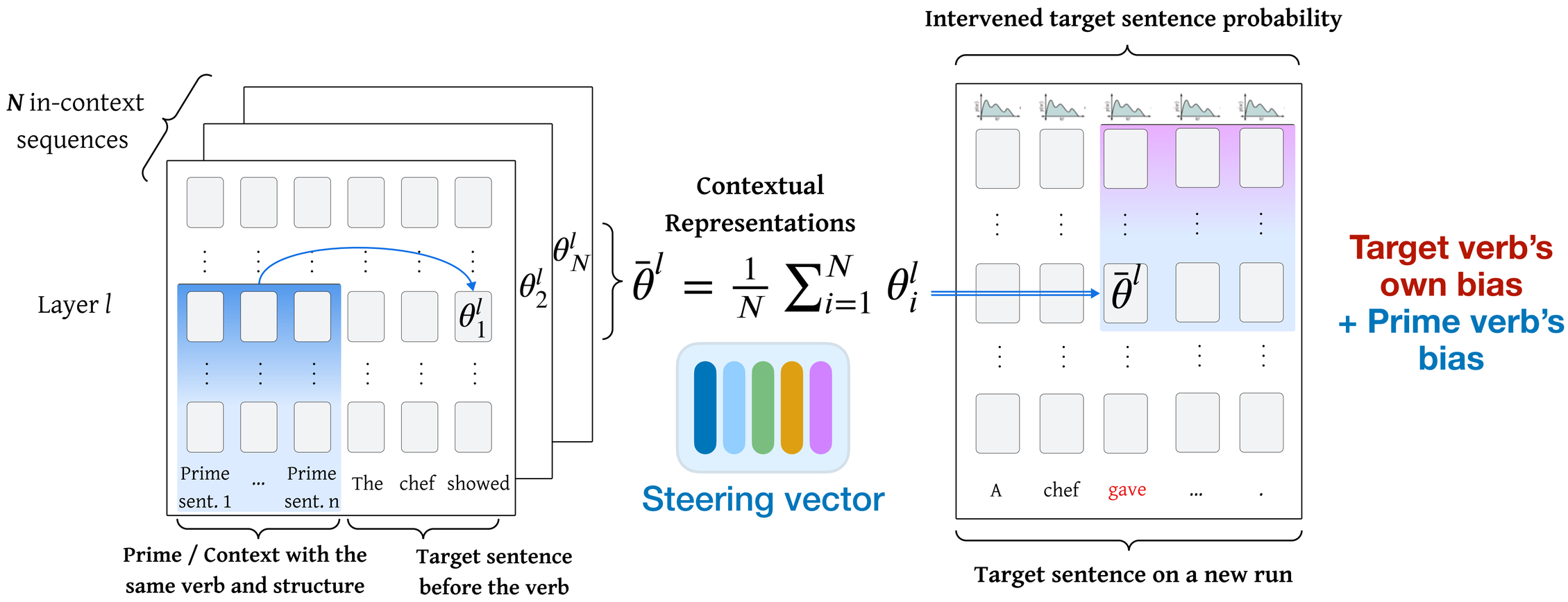
Experiment Setup: Steering Vector Extraction



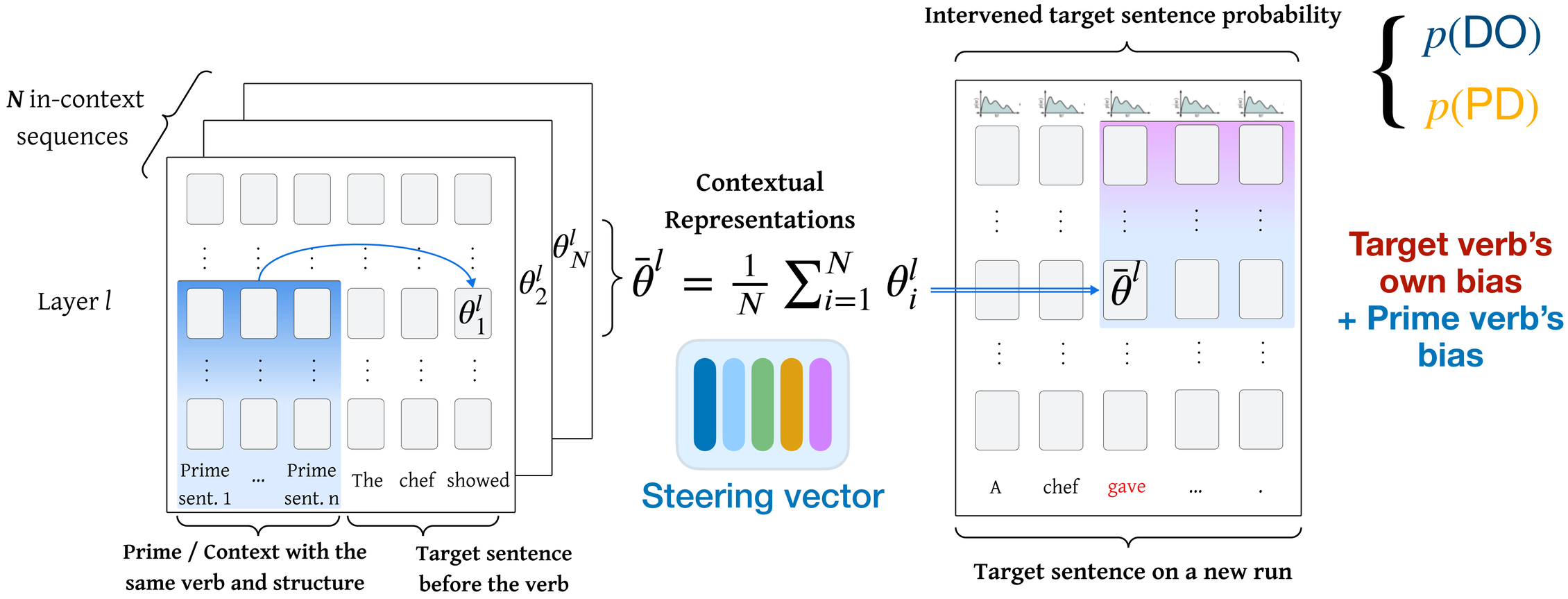
Experiment Setup: Steering Vector Extraction



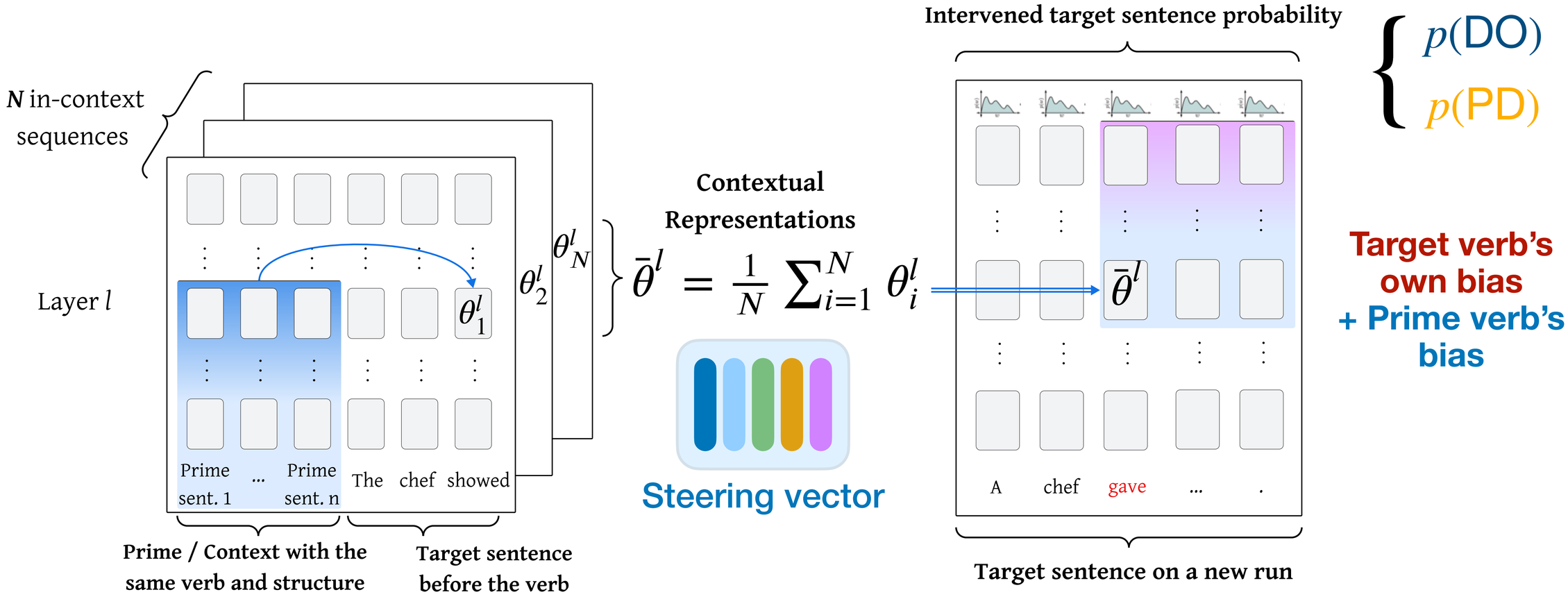
Experiment Setup: Steering Vector Injection



Experiment Setup: Steering Vector Injection



Experiment Setup: Steering Vector Injection



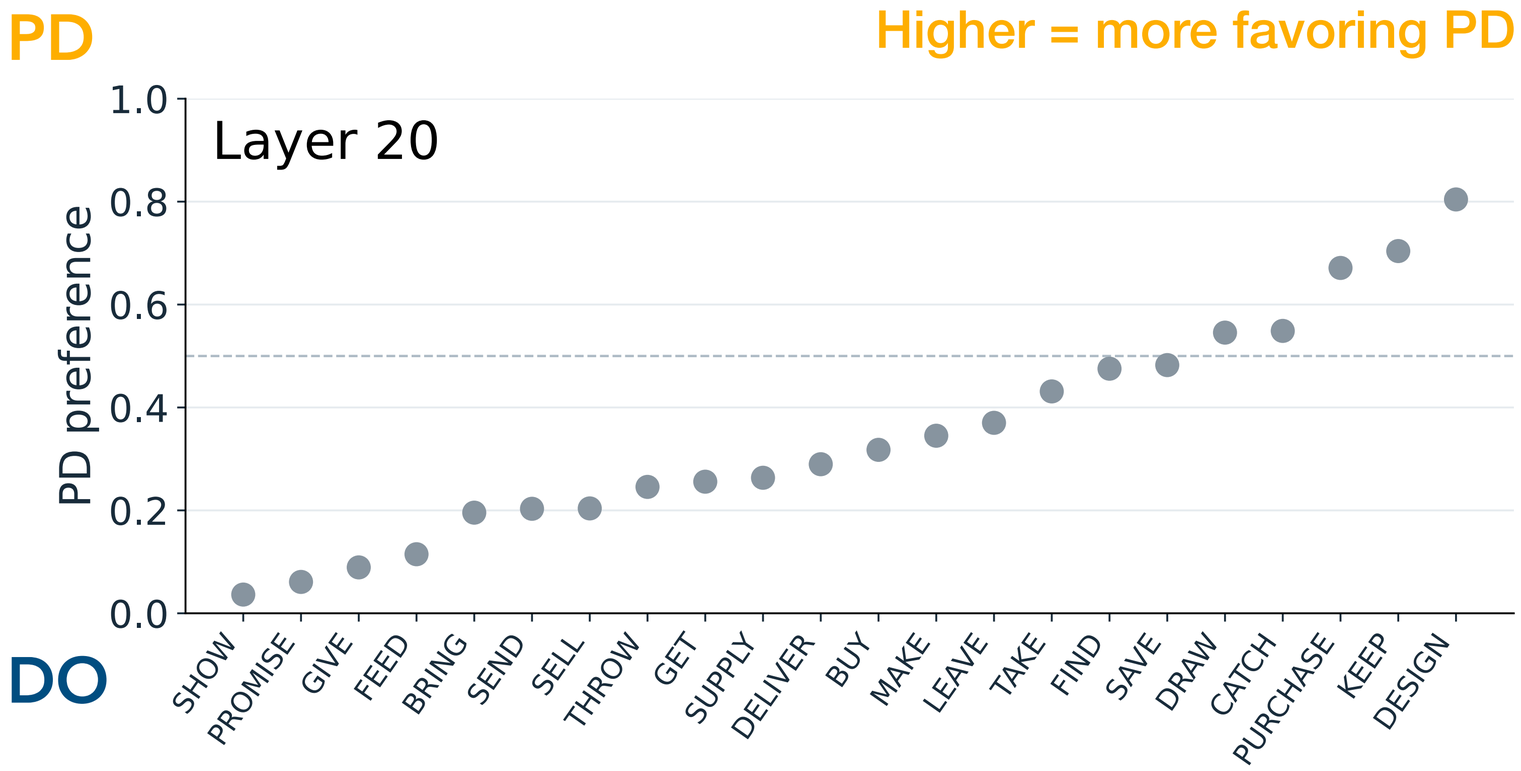
$$\text{PD preference} = \frac{P(\text{PD})}{P(\text{PD}) + P(\text{DO})}$$

Result 1: Steering Reproduces Priming

Llama-2-13B

What is injected:

- No steering



$$\text{PD preference} = \frac{P(\text{PD})}{P(\text{PD}) + P(\text{DO})}$$

Result 1: Steering Reproduces Priming

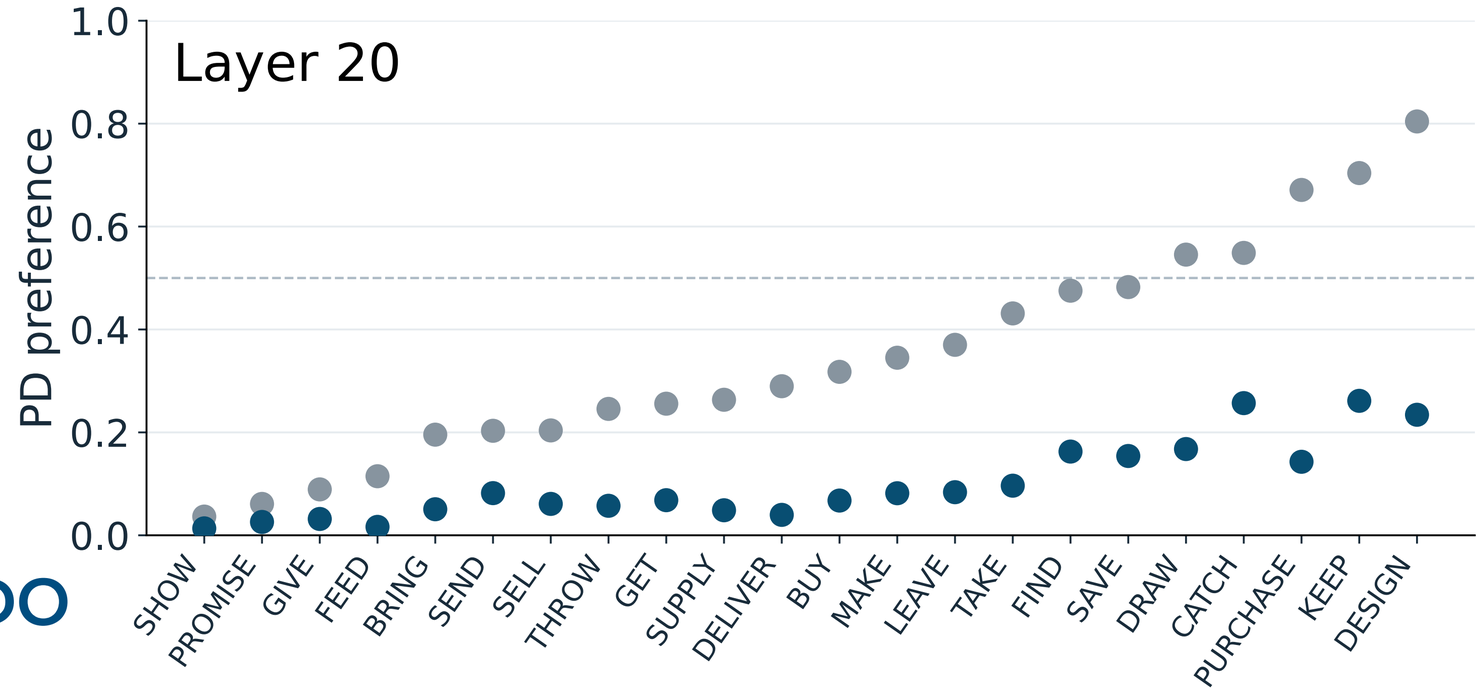
`Llama-2-13B`

What is injected:

- No steering
- DO-context vectors
  Shift toward DO

PD

Higher = more favoring PD



DO

Target verbs, sorted by no-steering PD preference

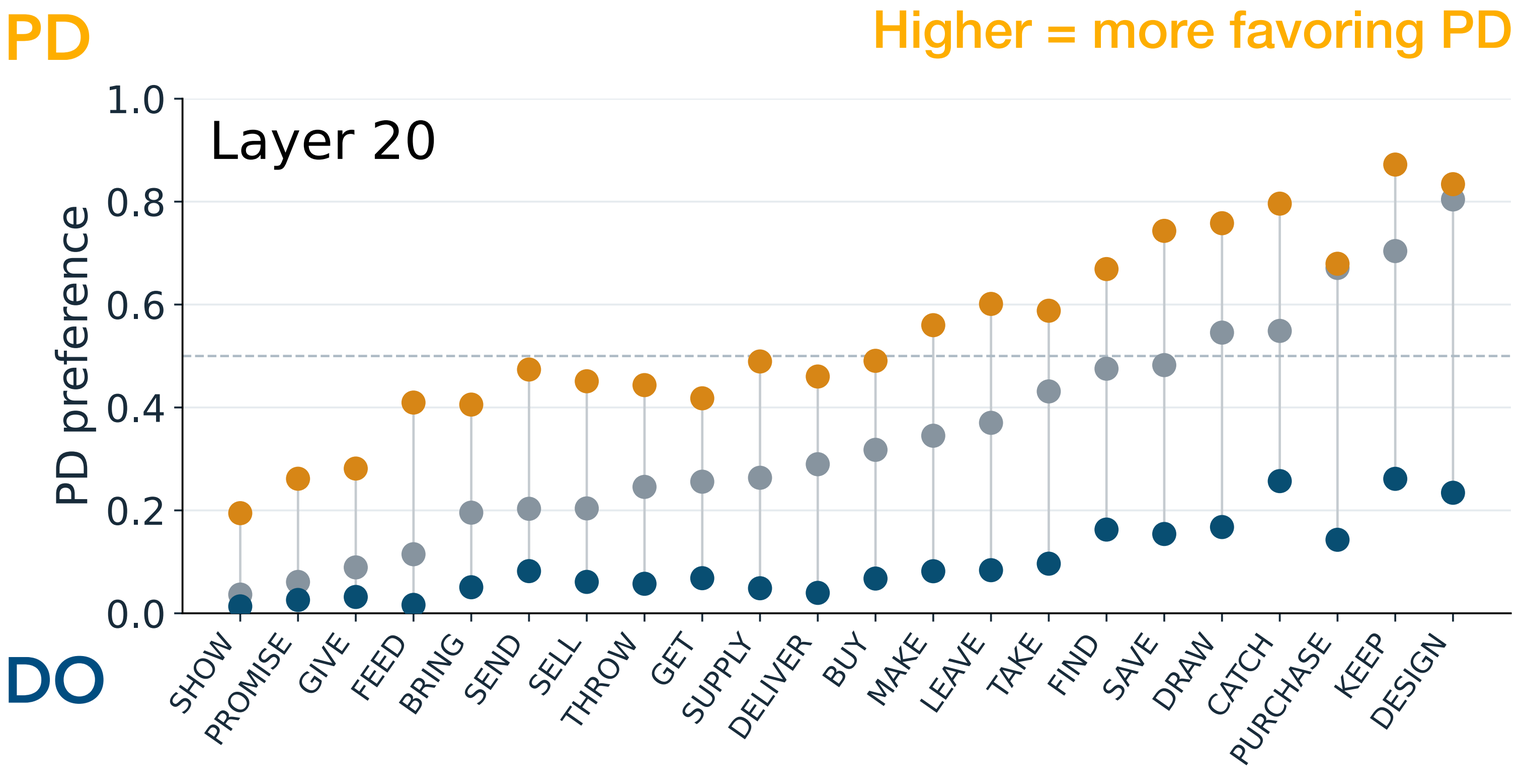
$$\text{PD preference} = \frac{P(\text{PD})}{P(\text{PD}) + P(\text{DO})}$$

Result 1: Steering Reproduces Priming

`Llama-2-13B`

What is injected:

- No steering
- DO-context vectors
  Shift toward DO
- PD-context vectors
  Shift toward PD



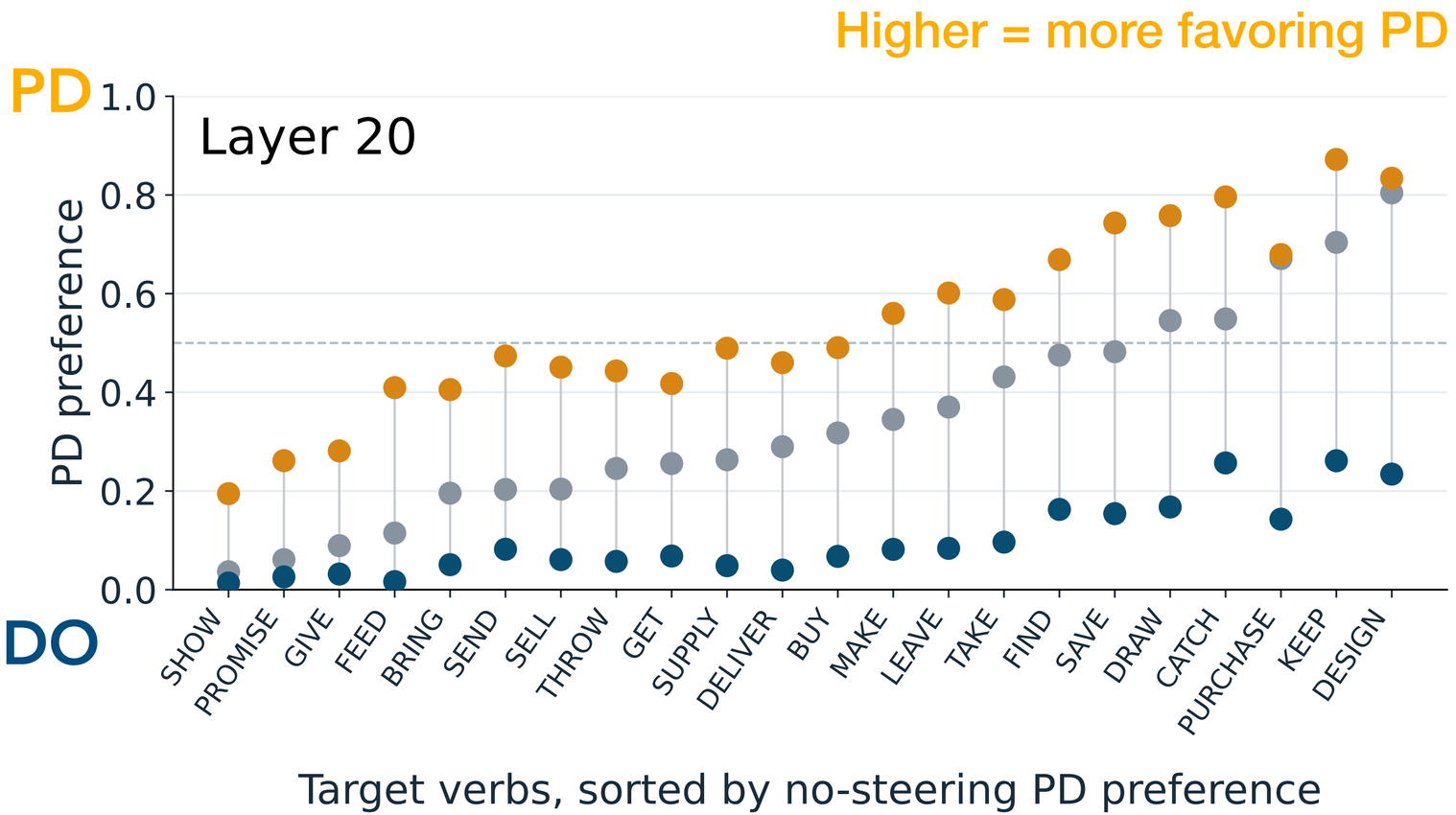
$$\text{PD preference} = \frac{P(\text{PD})}{P(\text{PD}) + P(\text{DO})}$$

Result 1: Steering Reproduces Priming

Llama-2-13B

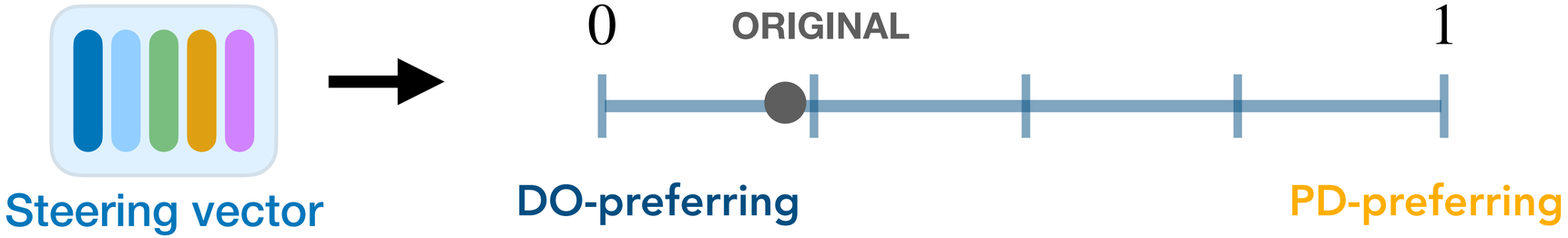
What is injected:

- No steering
- DO-context vectors
  Shift toward DO
- PD-context vectors
  Shift toward PD

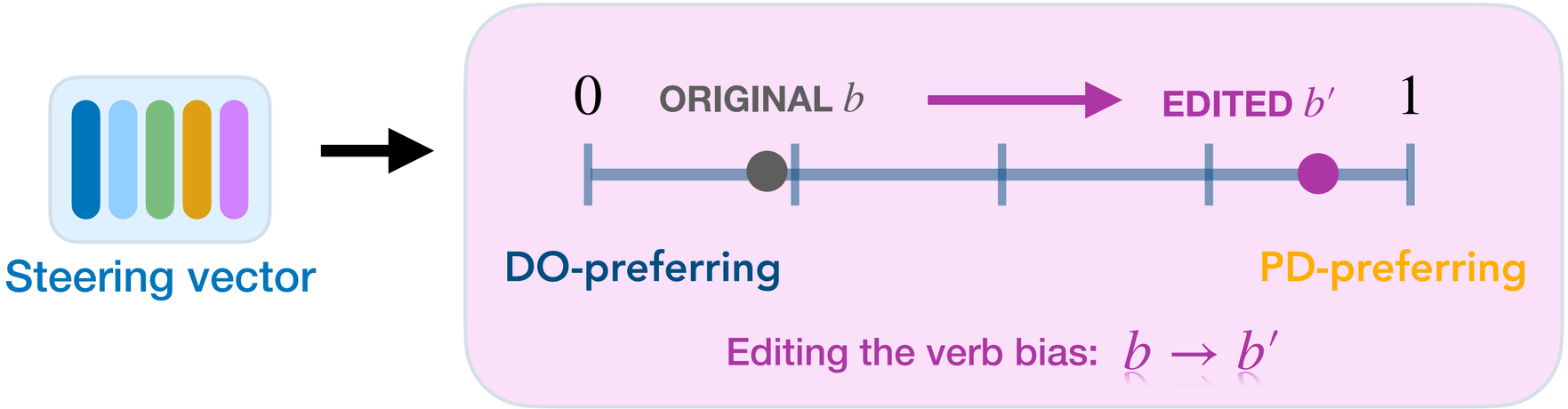


Test on Content: steering vectors shift downstream structural preference.

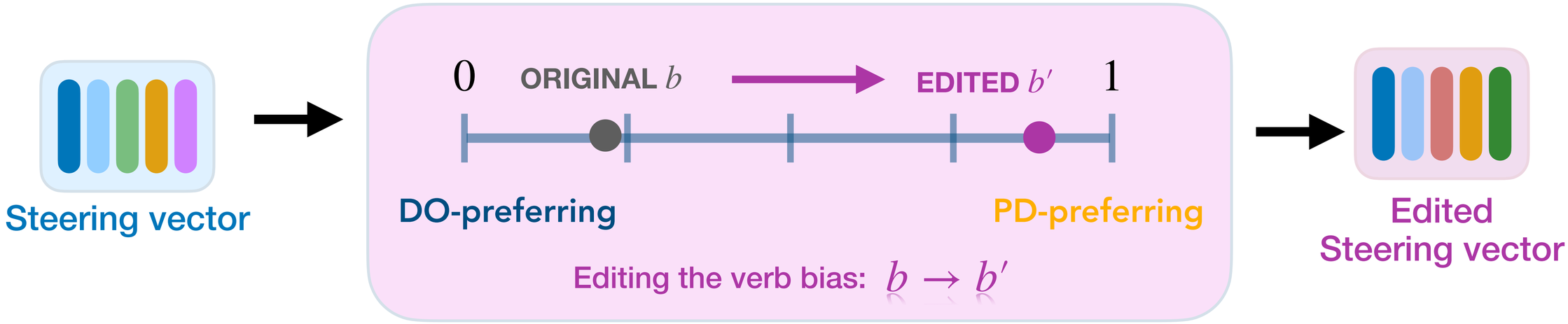
Causal Intervention on Verb Bias



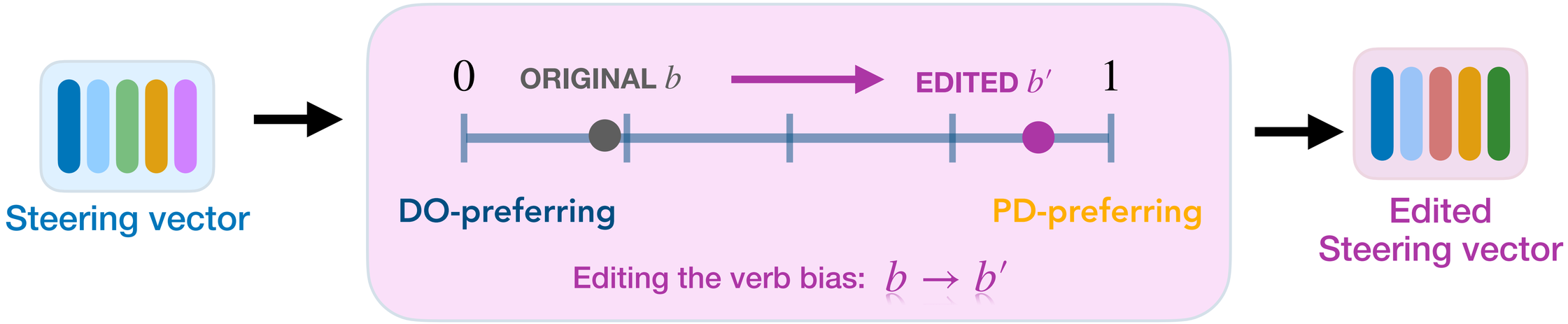
Causal Intervention on Verb Bias

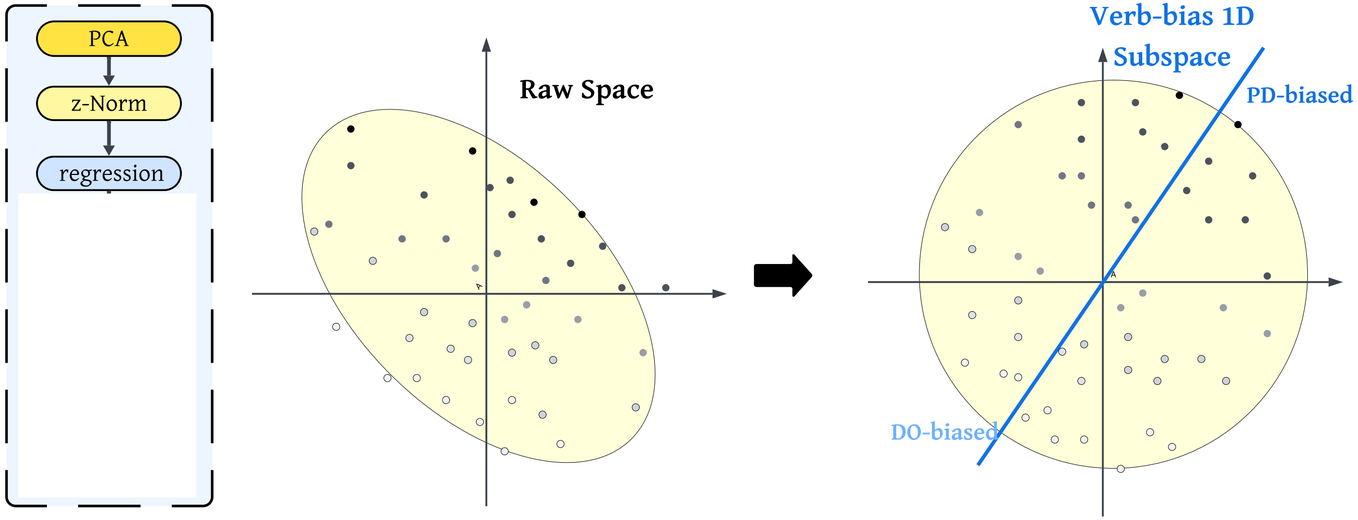


Causal Intervention on Verb Bias

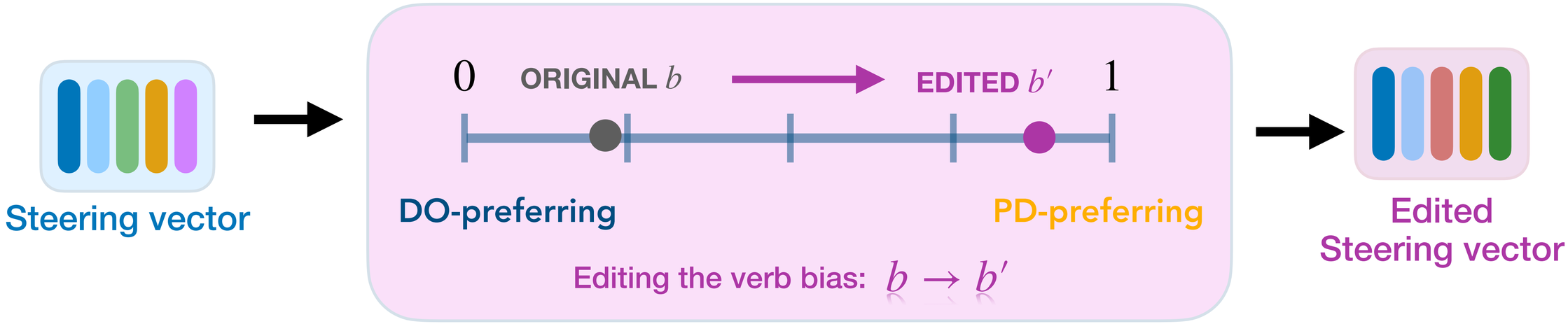


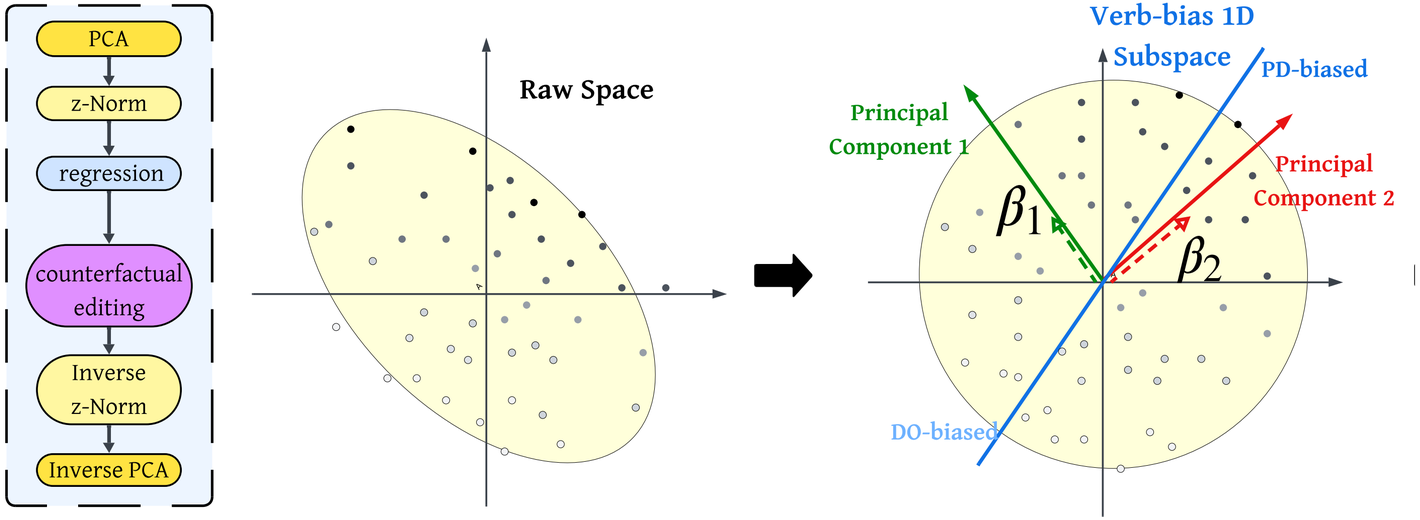
Causal Intervention on Verb Bias



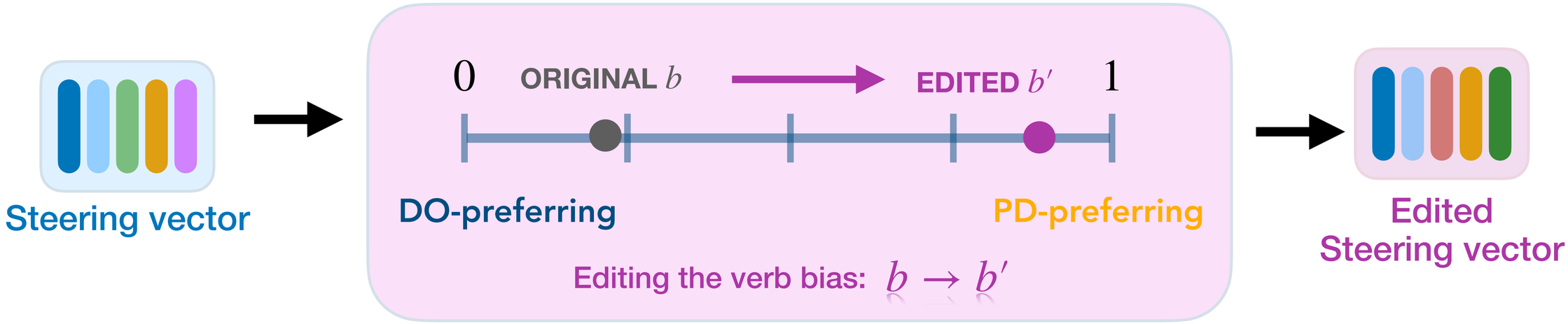


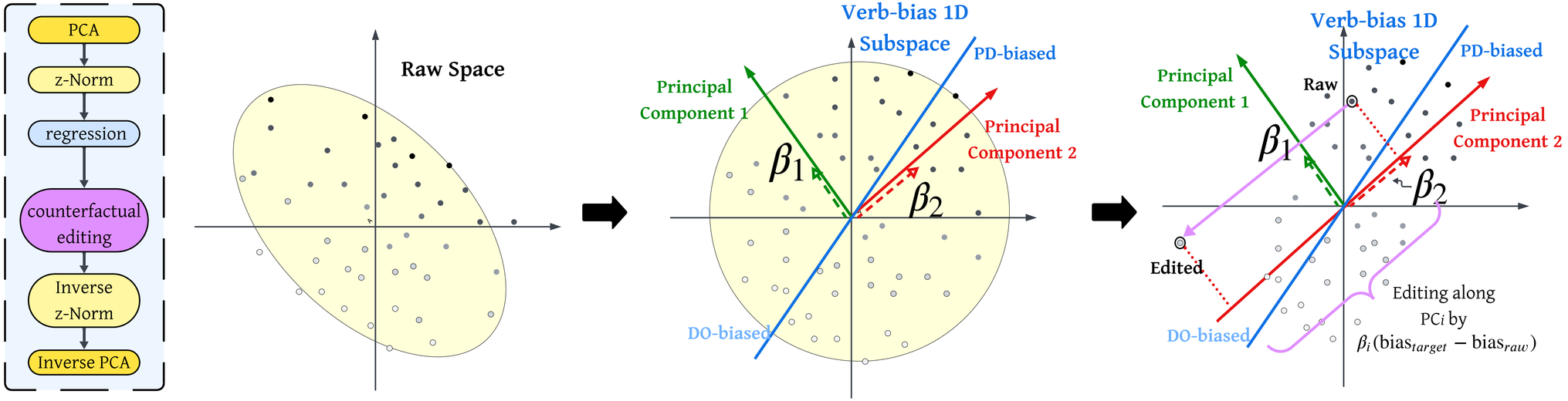
Causal Intervention on Verb Bias



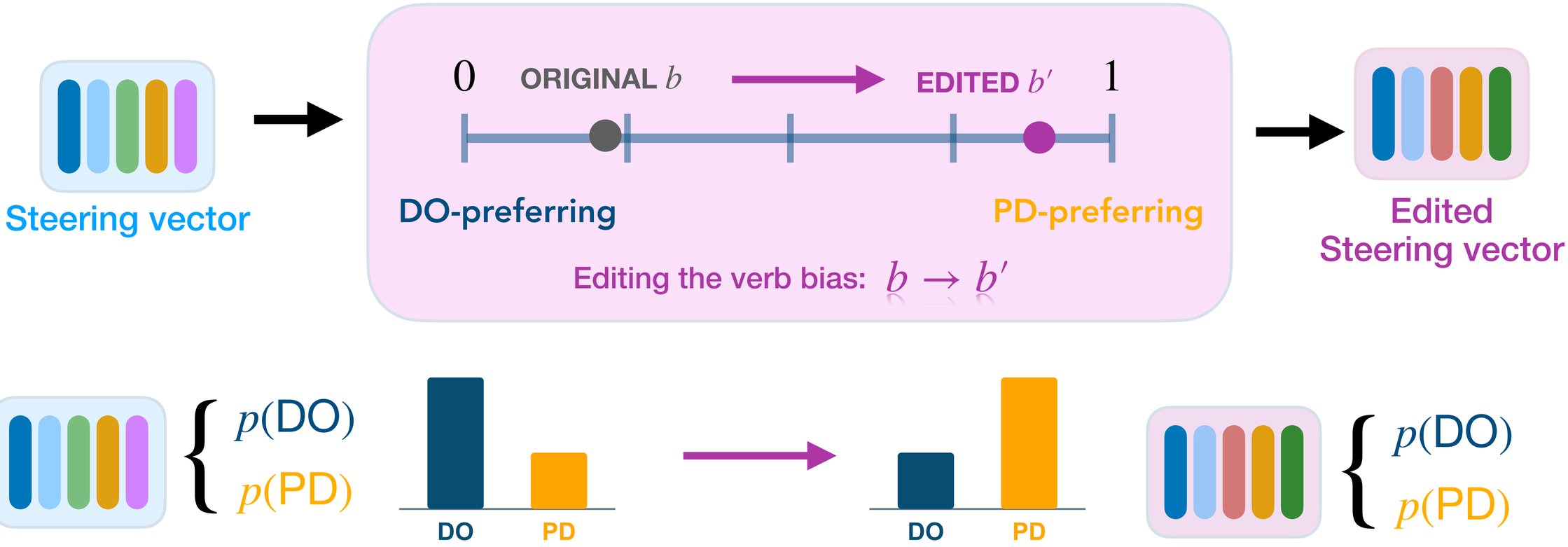


Causal Intervention on Verb Bias

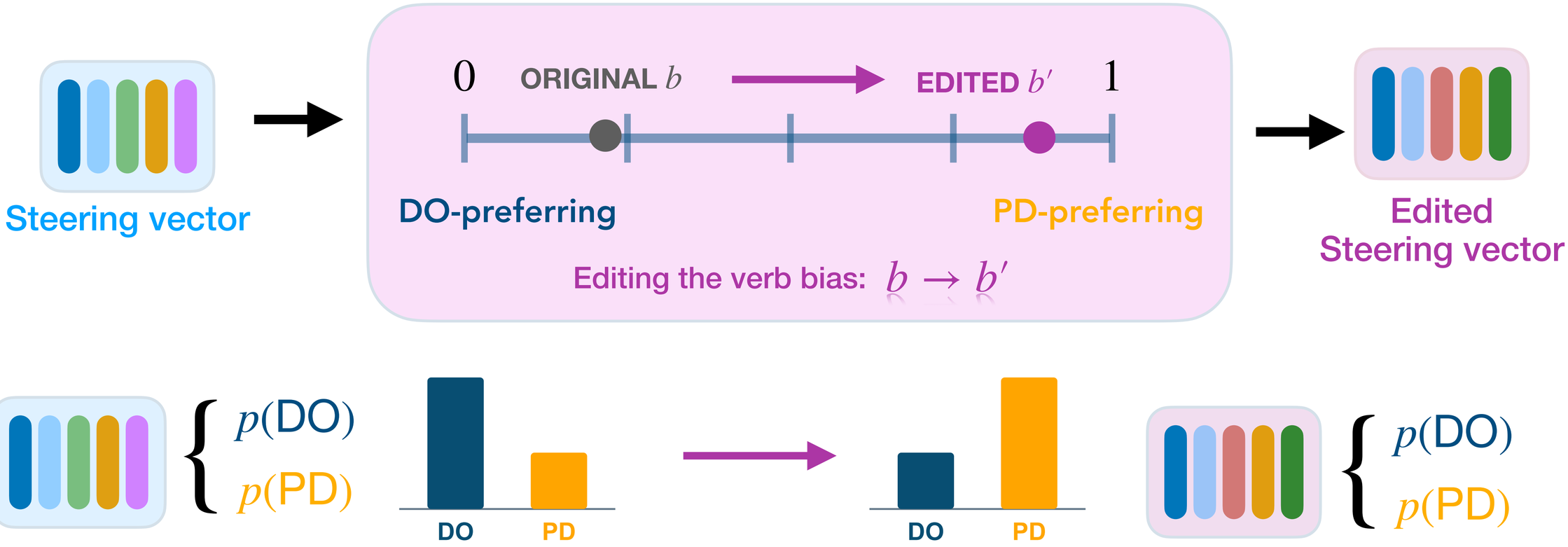




Causal Intervention on Verb Bias



Causal Intervention on Verb Bias

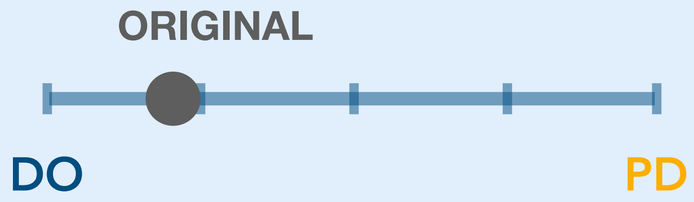


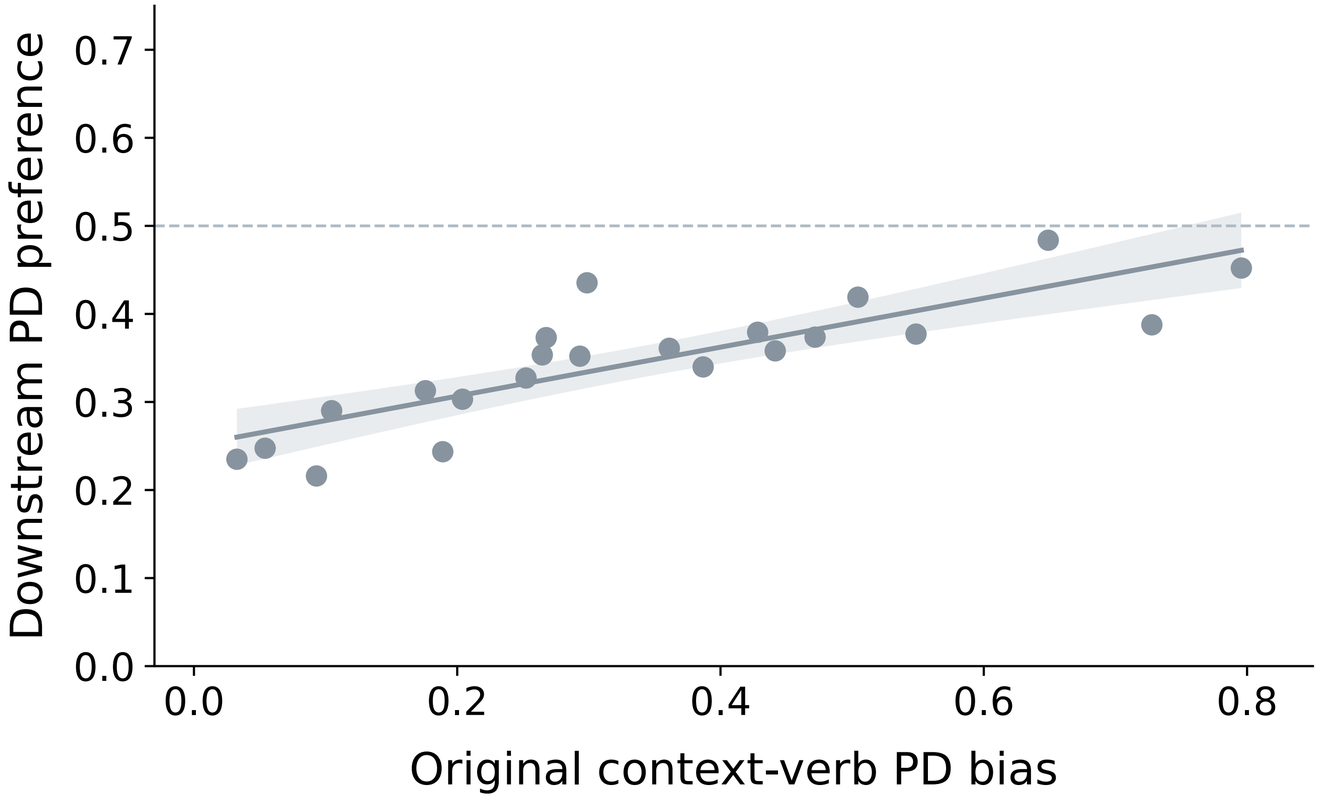
Does manipulating the verb bias in steering vectors induce changes in downstream structural preference?

Result 2: Verb Bias is Causally Involved

Context Structure: PD

Represented Verb Bias:



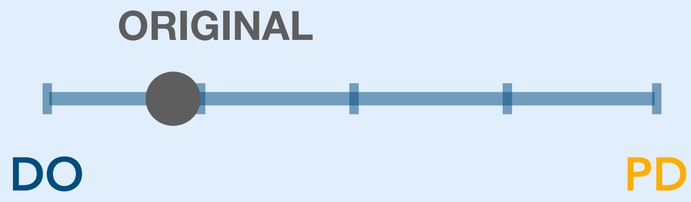


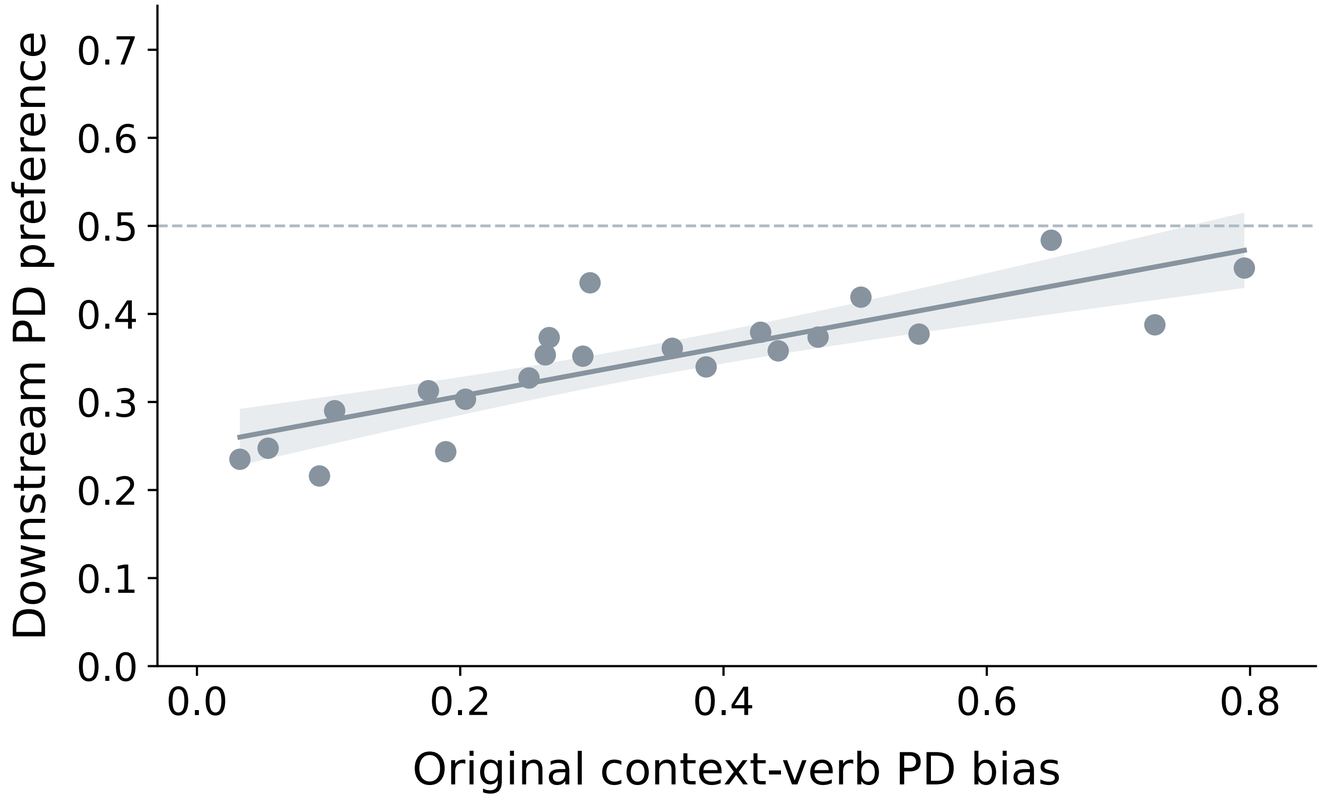
Result 2: Verb Bias is Causally Involved

$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$

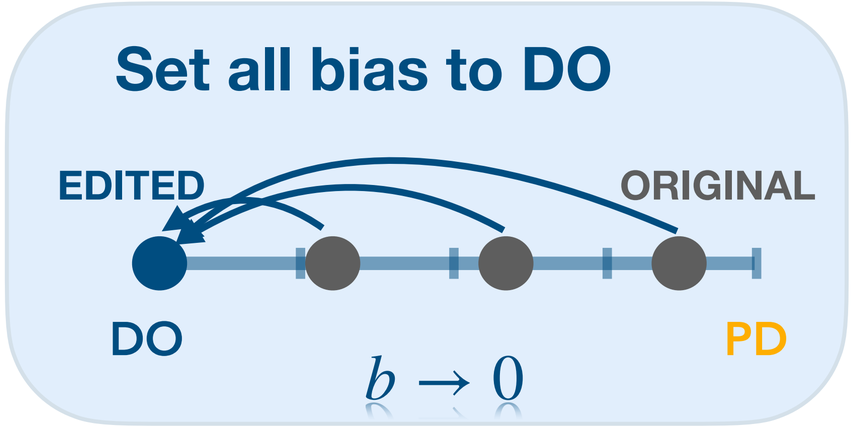
Context Structure: PD

Represented Verb Bias:

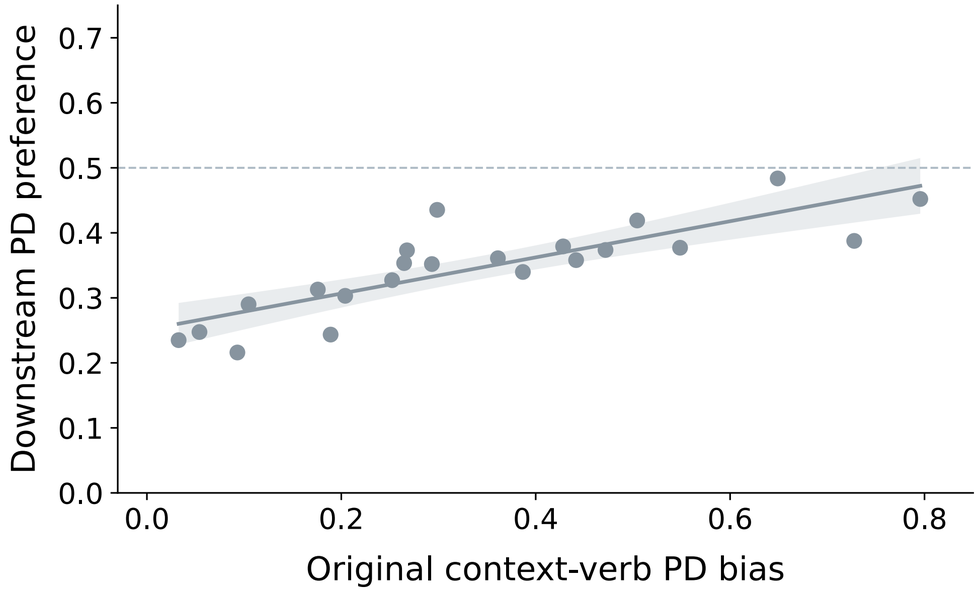




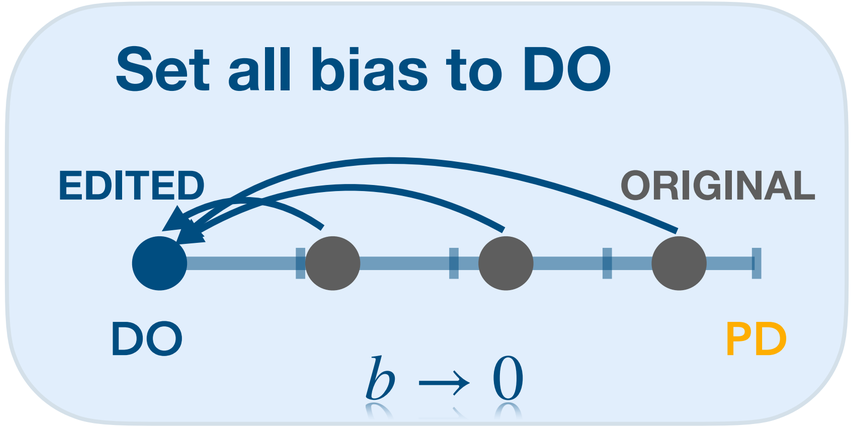
Result 2: Verb Bias is Causally Involved



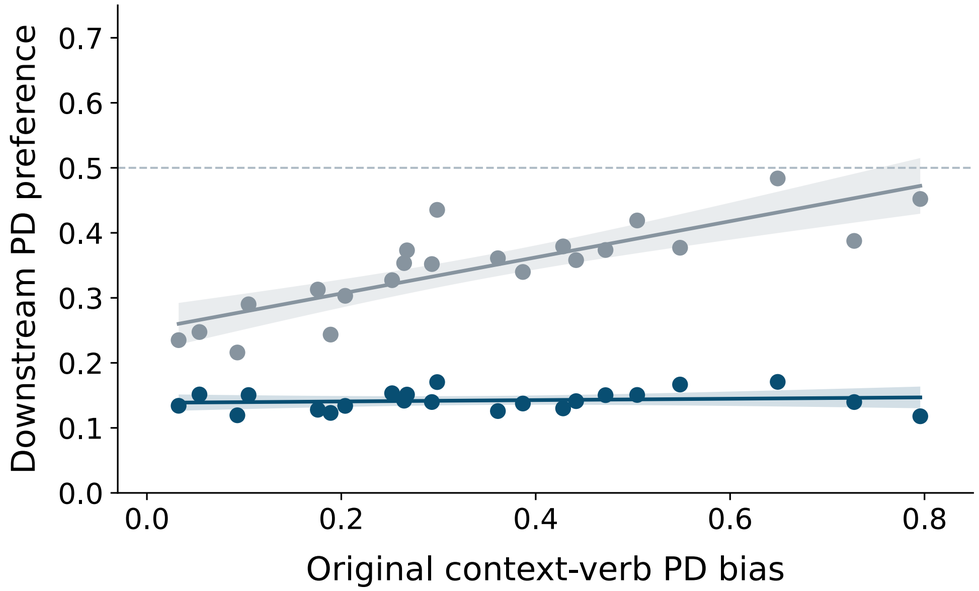
$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$



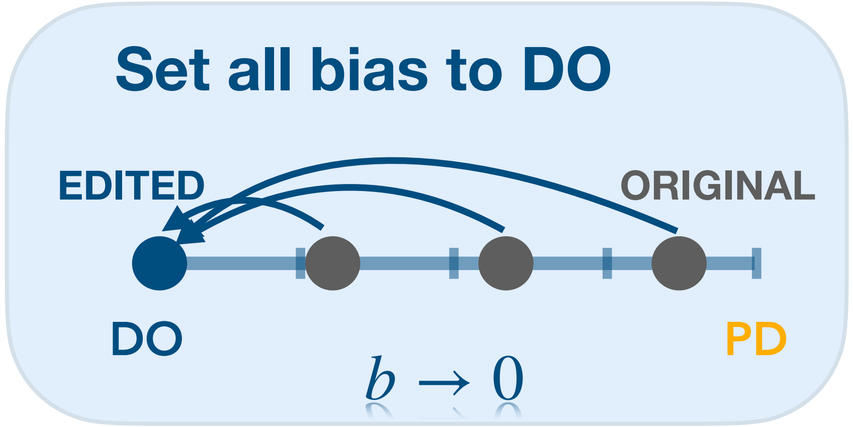
Result 2: Verb Bias is Causally Involved

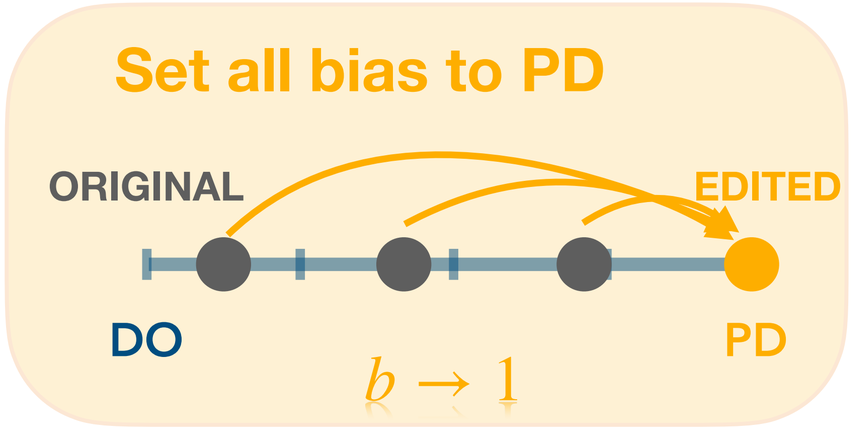


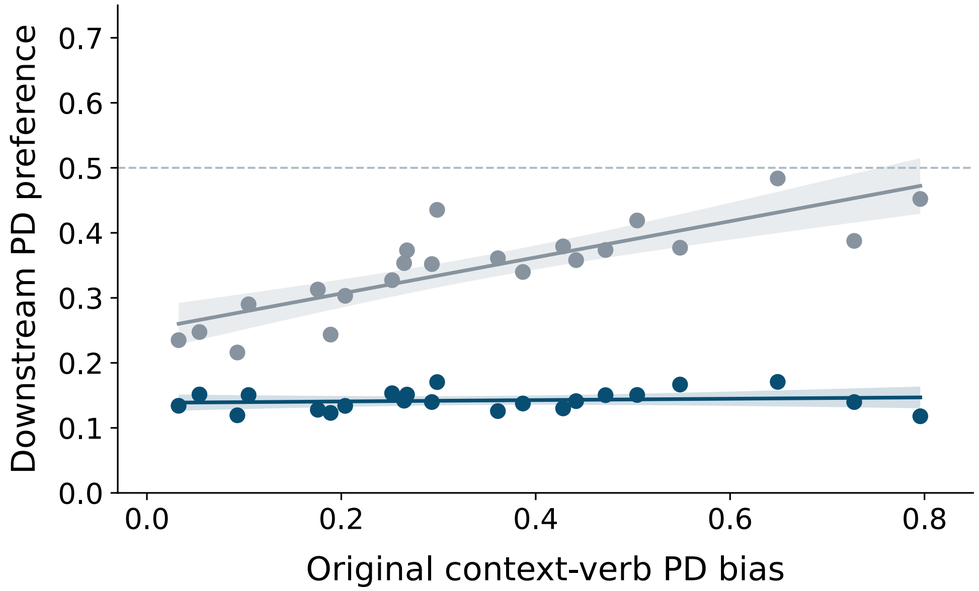
$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$

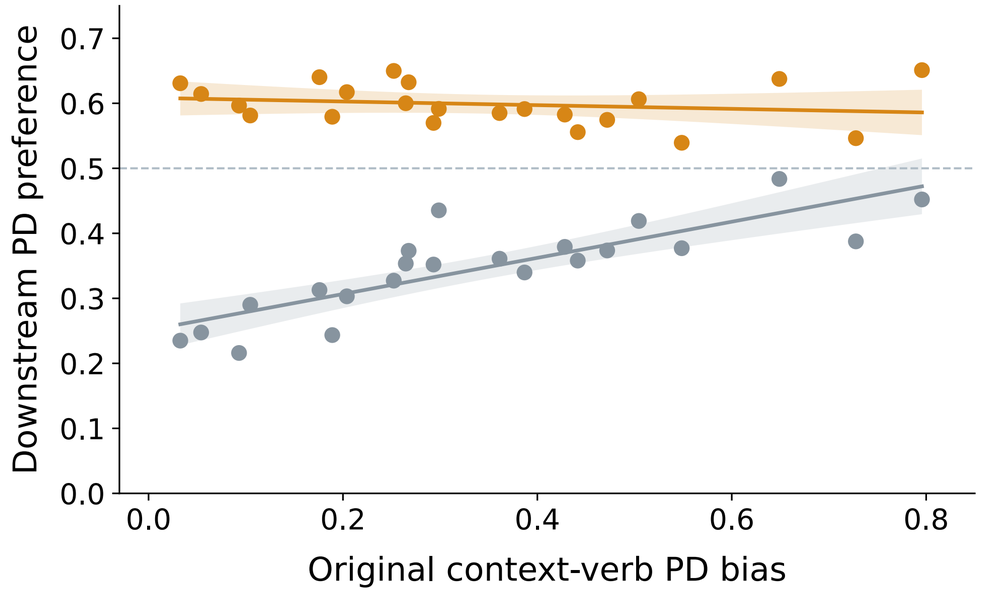


Result 2: Verb Bias is Causally Involved









Result 2: Verb Bias is Causally Involved

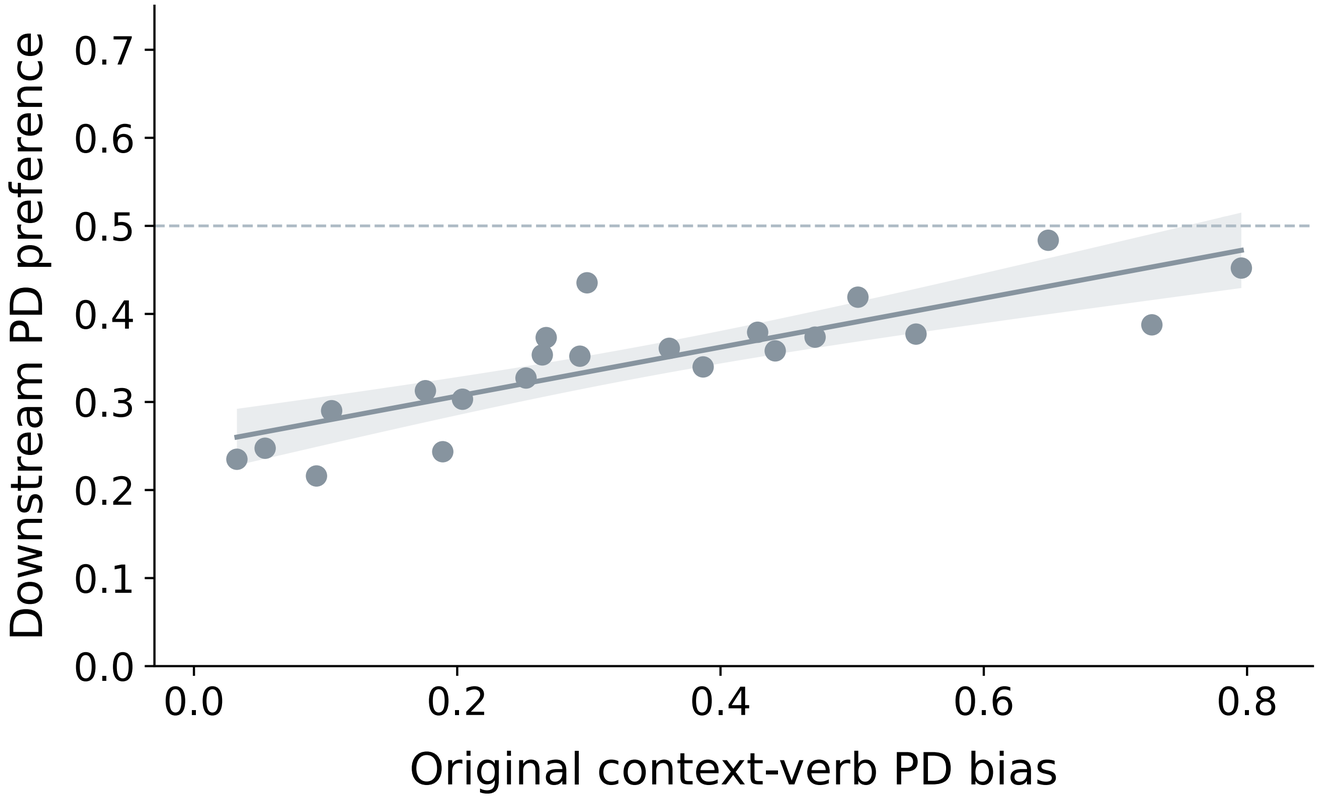
$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$

Context Structure: PD

Represented Verb Bias:



$$b \rightarrow 1 - b$$



Result 2: Verb Bias is Causally Involved

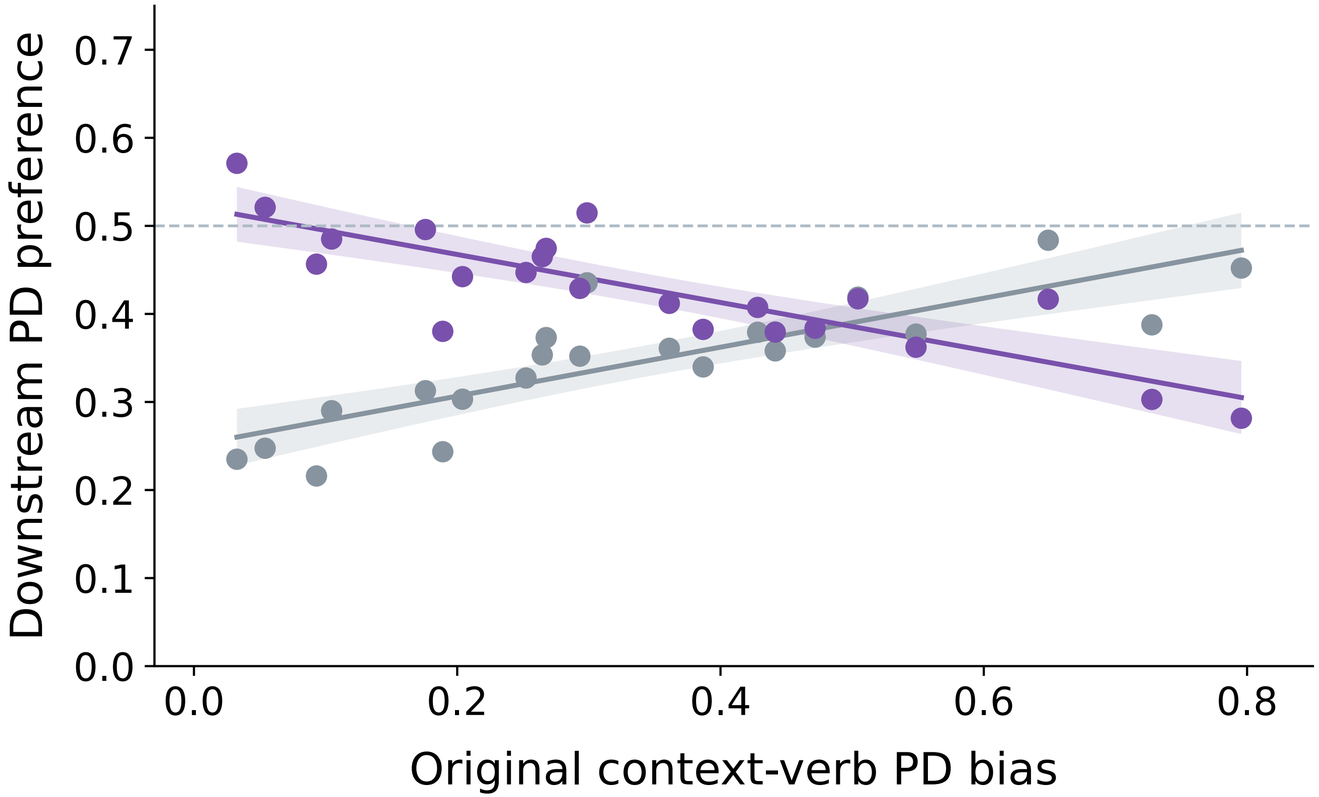
$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$

Context Structure: PD

Represented Verb Bias:

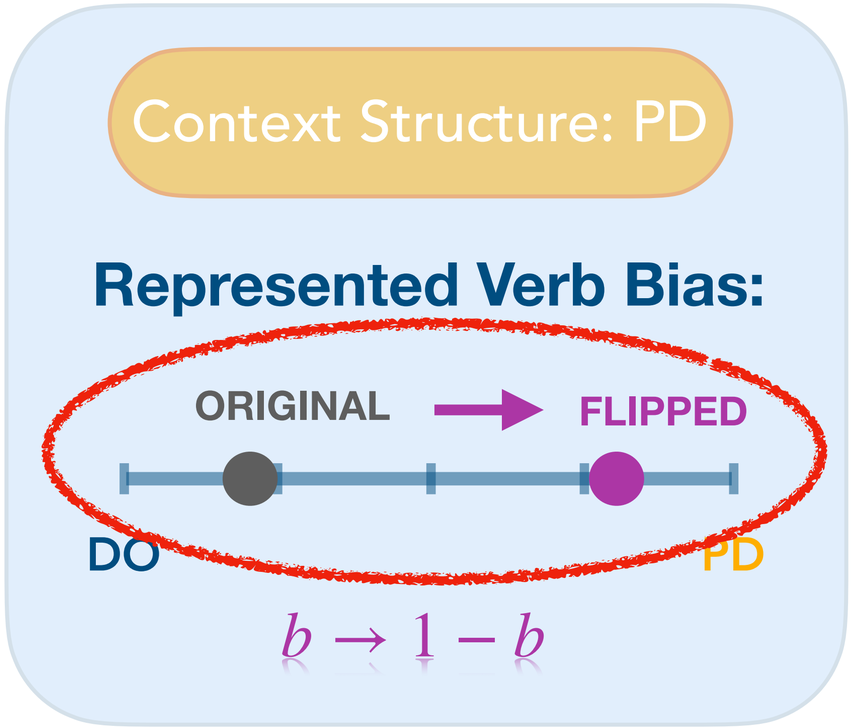


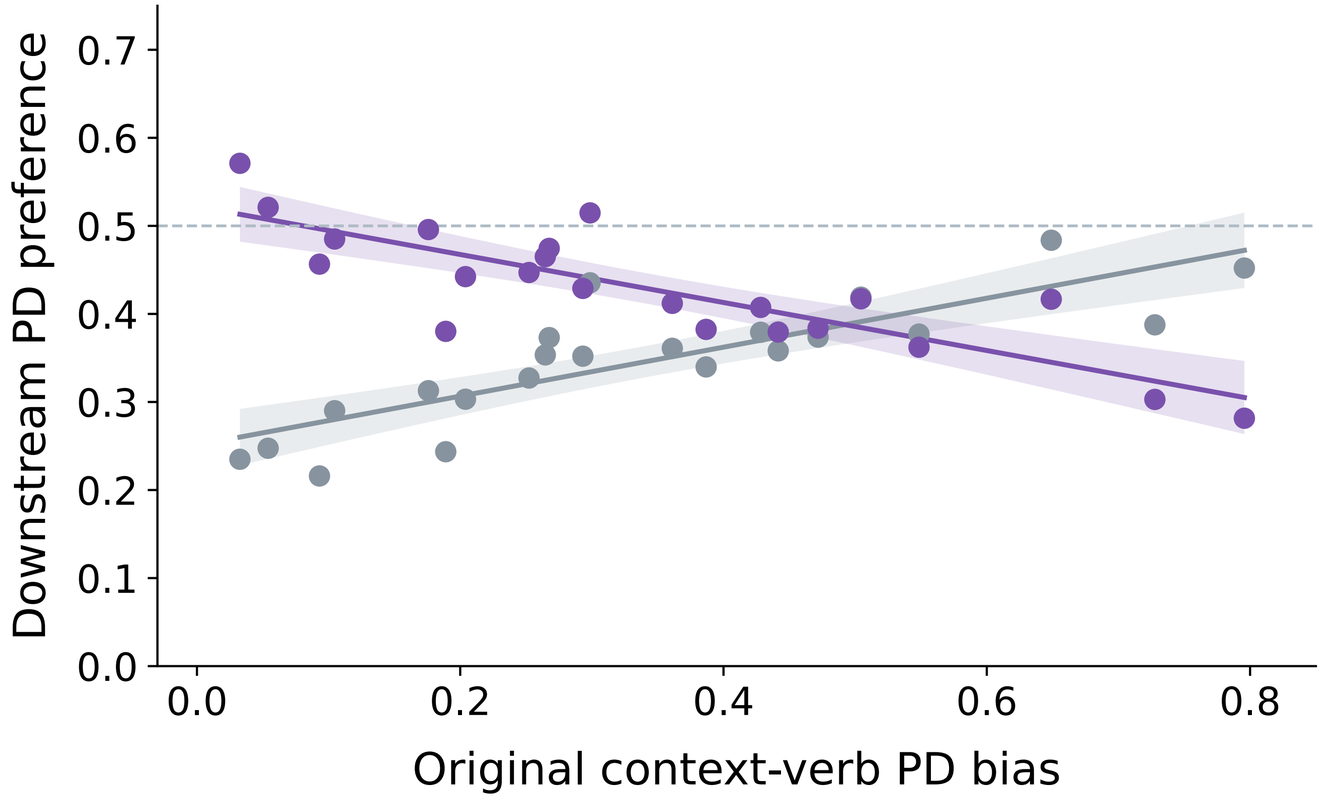
$b \rightarrow 1 - b$



Result 2: Verb Bias is Causally Involved

$$\text{PD preference} = \frac{P(\text{PD} \mid \text{steering})}{P(\text{PD} \mid \text{steering}) + P(\text{DO} \mid \text{steering})}$$





Reversing reprsented bias also flips the downstream relationship.

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

PD Alice gave a book to Bob.

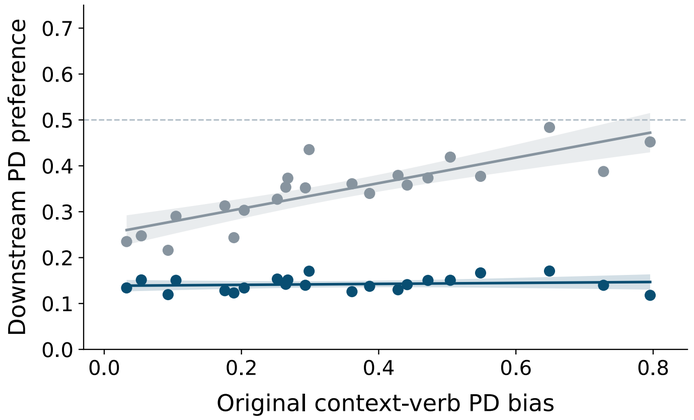
Encountering a PD structure in context.

A new sentence

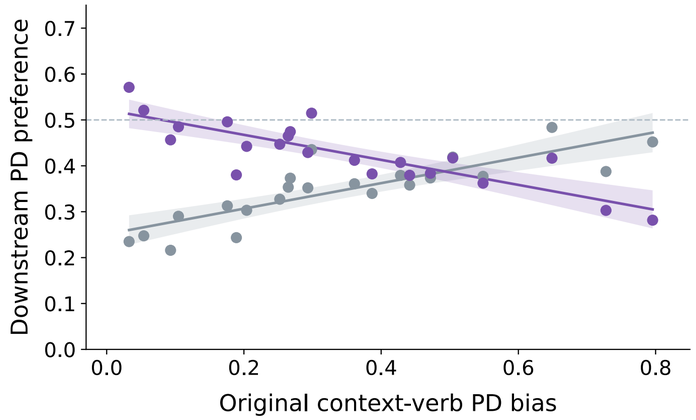
A chef showed → the doctor a picture / a picture to the doctor

Greater tendency to use PD

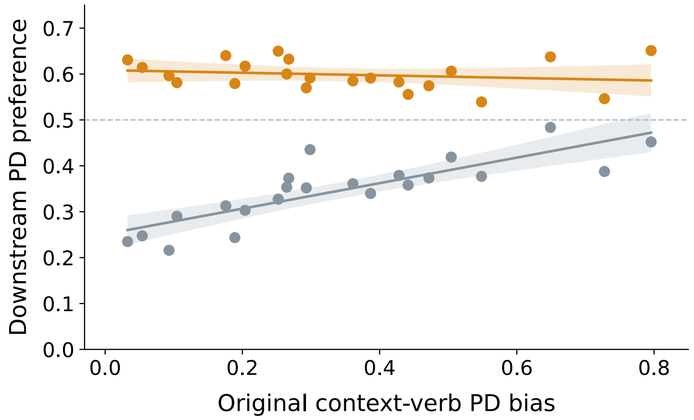
Set all bias to DO



Flipping all bias



Set all bias to PD



Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

PD Alice gave a book to Bob.

Encountering a PD structure in context.

A new sentence

A chef showed

Greater tendency to use PD

the doctor a picture

a picture to the doctor

2 Sensitivity to expectations

Context verb's bias	Encountered context structure		Resulted shift towards PD
PD-preferring	+	PD	:

[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

PD Alice gave a book to Bob.

Encountering a PD structure in context.

A new sentence

A chef showed

Greater tendency to use PD

the doctor a picture

a picture to the doctor

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		Resulted shift towards PD
PD-preferring	+	PD	:	Expected

[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 Structural Priming

PD Alice gave a book to Bob.

Encountering a PD structure in context.

A new sentence

A chef showed

Greater tendency to use PD

the doctor a picture

a picture to the doctor

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		Resulted shift towards PD
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

1 S

PD

Encou

picture

e doctor

Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming

Zhenghao Zhou **Robert Frank** **R. Thomas McCoy**
Yale University
370 Temple Street
New Haven, CT 06511
{herbert.zhou, robert.frank, tom.mccoy}@yale.edu

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		Resulted shift towards PD
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

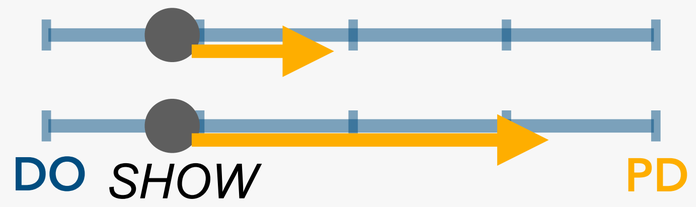
Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming

Zhenghao Zhou **Robert Frank** **R. Thomas McCoy**
Yale University
370 Temple Street
New Haven, CT 06511
{herbert.zhou, robert.frank, tom.mccoy}@yale.edu

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

Resulted shift towards PD



[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

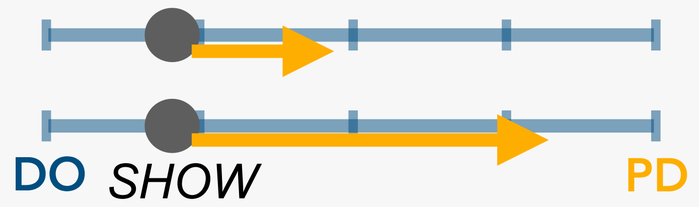
Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming

Zhenghao Zhou **Robert Frank** **R. Thomas McCoy**
Yale University
370 Temple Street
New Haven, CT 06511
{herbert.zhou, robert.frank, tom.mccoy}@yale.edu

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

Resulted shift towards PD



[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

Human structural priming provides theory-motivated diagnostics

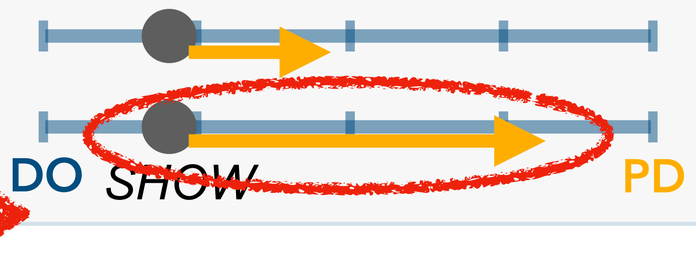
Is In-Context Learning a Type of Error-Driven Learning? Evidence from the Inverse Frequency Effect in Structural Priming

Zhenghao Zhou **Robert Frank** **R. Thomas McCoy**
Yale University
370 Temple Street
New Haven, CT 06511
{herbert.zhou, robert.frank, tom.mccoy}@yale.edu

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

Resulted shift towards PD



[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

Two Signatures of Contextual Effects

1 Structural Priming

PD Alice gave a book to Bob.

Encountering a PD structure in context.

A new sentence

A chef showed → the doctor a picture

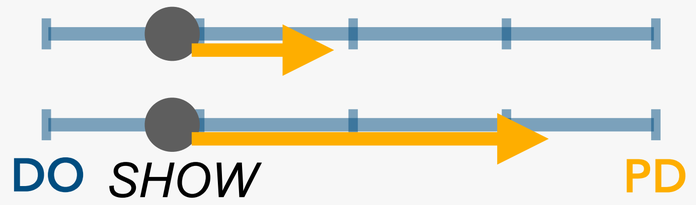
A chef showed → a picture to the doctor

Greater tendency to use PD

2 Sensitivity to expectations

Context verb's bias		Encountered context structure		
PD-preferring	+	PD	:	Expected
DO-preferring	+	PD	:	Unexpected

Resulted shift towards PD

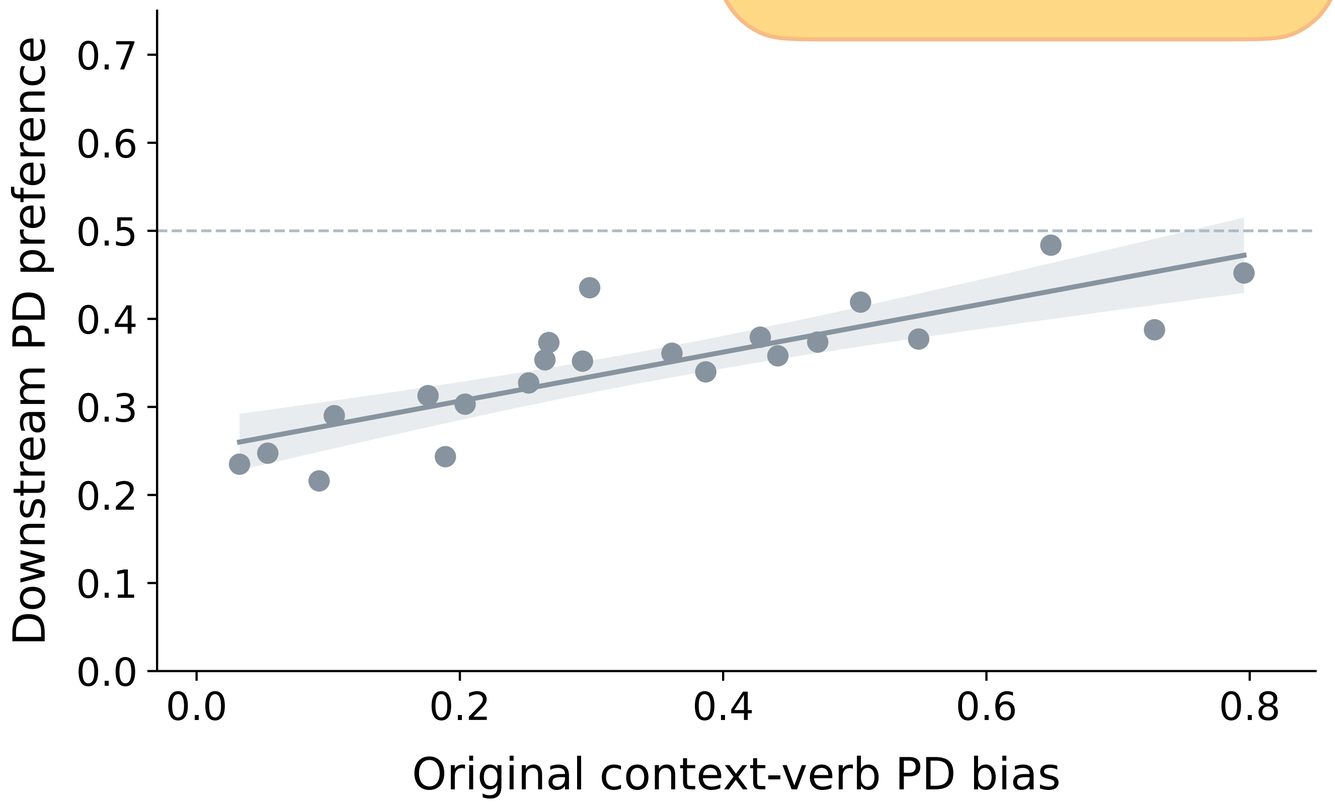


Influence: error-driven adaptation — a less expected context causes a larger downstream shift

[e.g., Humans: Jaeger & Snider 2013; LMs' ICL: Zhou et al. 2025]

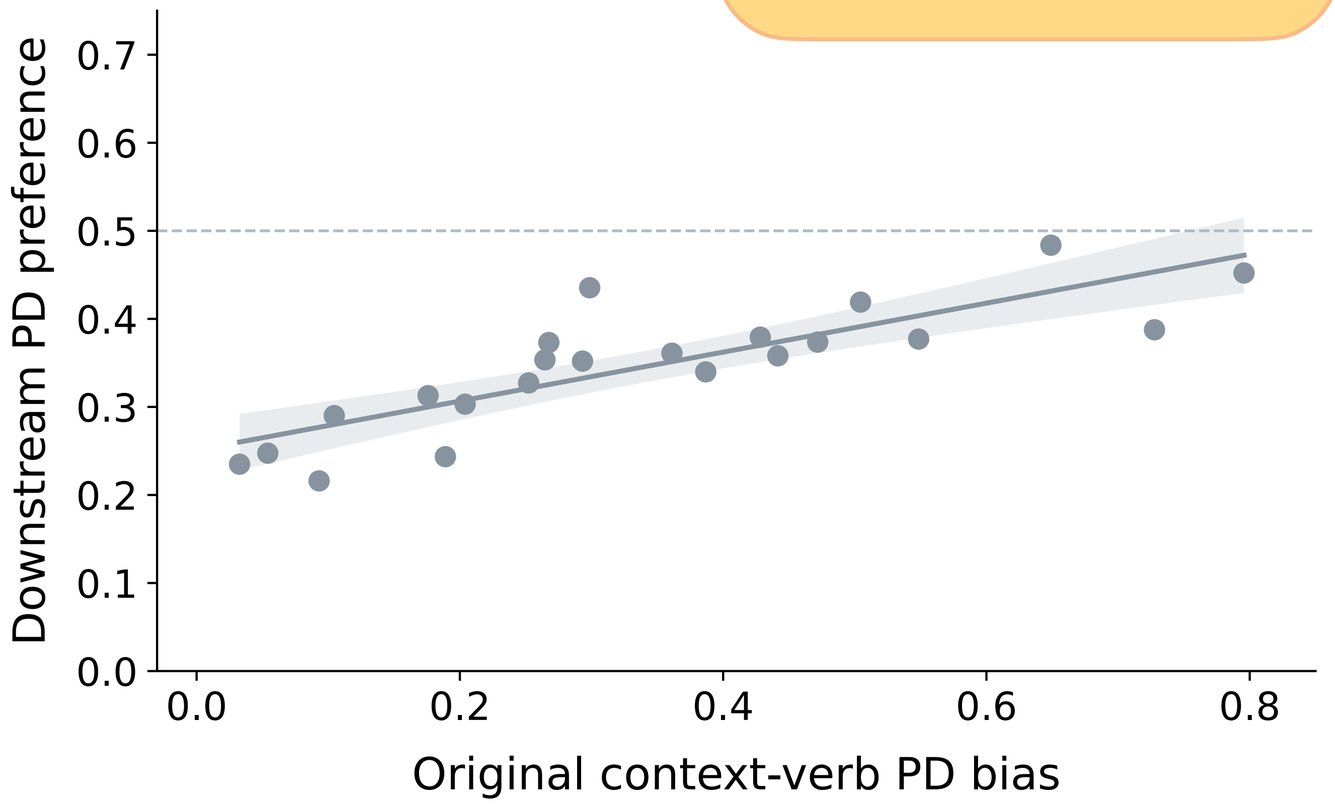
Does a less expected context produce a larger shift?

Context Structure: PD



Does a less expected context produce a larger shift?

Context Structure: PD



PD less expected

PD more expected

Does a less expected context produce a larger shift?

Error-driven prediction:

Less expected PD:

→

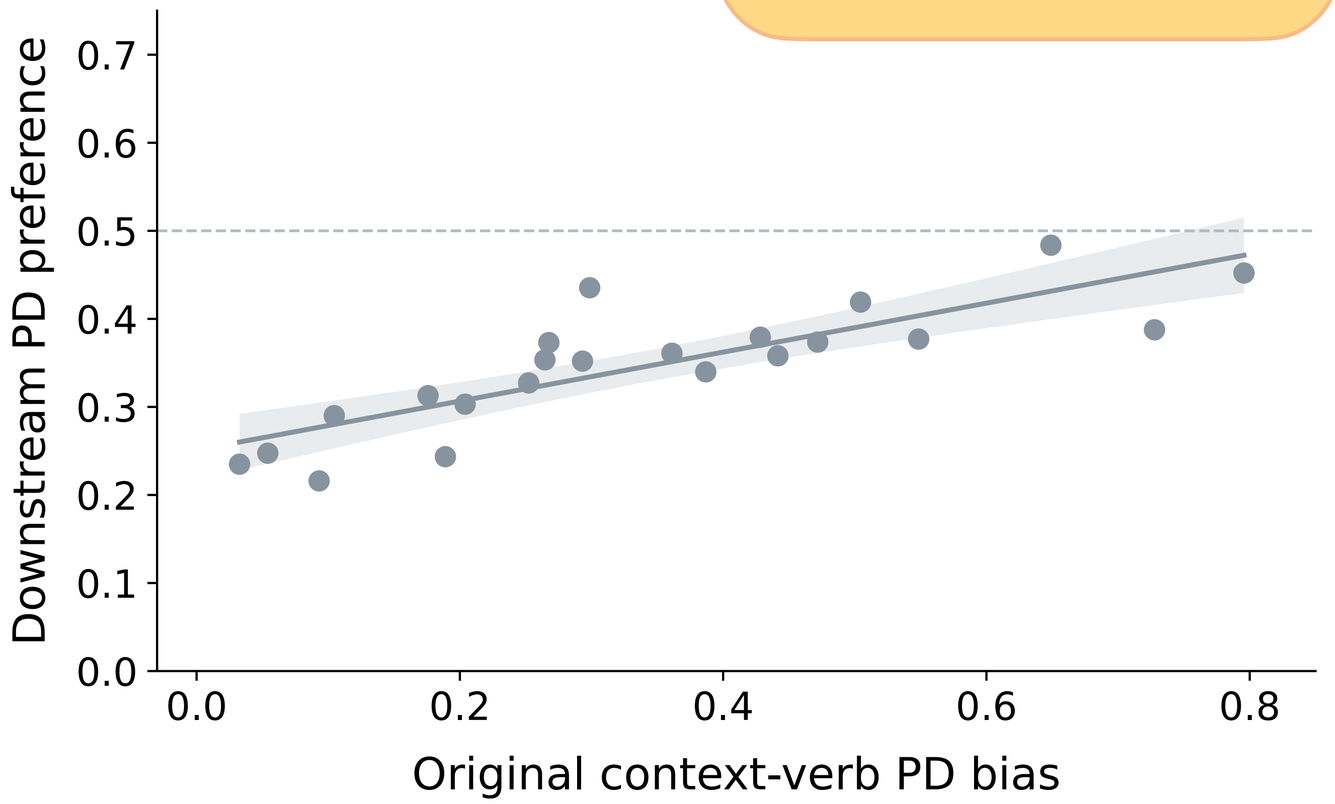
Larger shift toward PD

More expected PD:

→

Smaller shift toward PD

Context Structure: PD



PD less expected

PD more expected

Does a less expected context produce a larger shift?

Error-driven prediction:

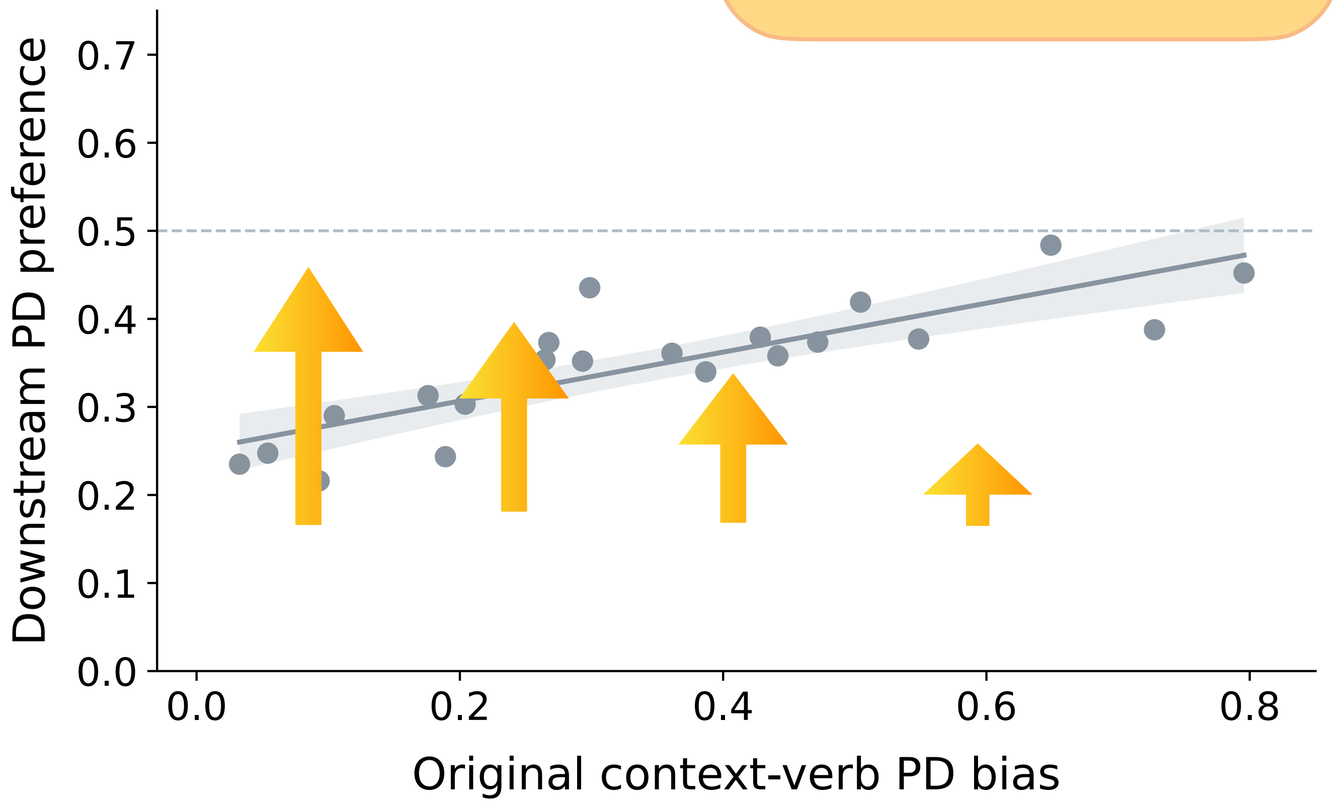
Less expected PD:

Larger shift toward PD

More expected PD:

Smaller shift toward PD

Context Structure: PD



PD less expected

PD more expected

Does a less expected context produce a larger shift?

Error-driven prediction:

Less expected PD:

→

Larger shift toward PD

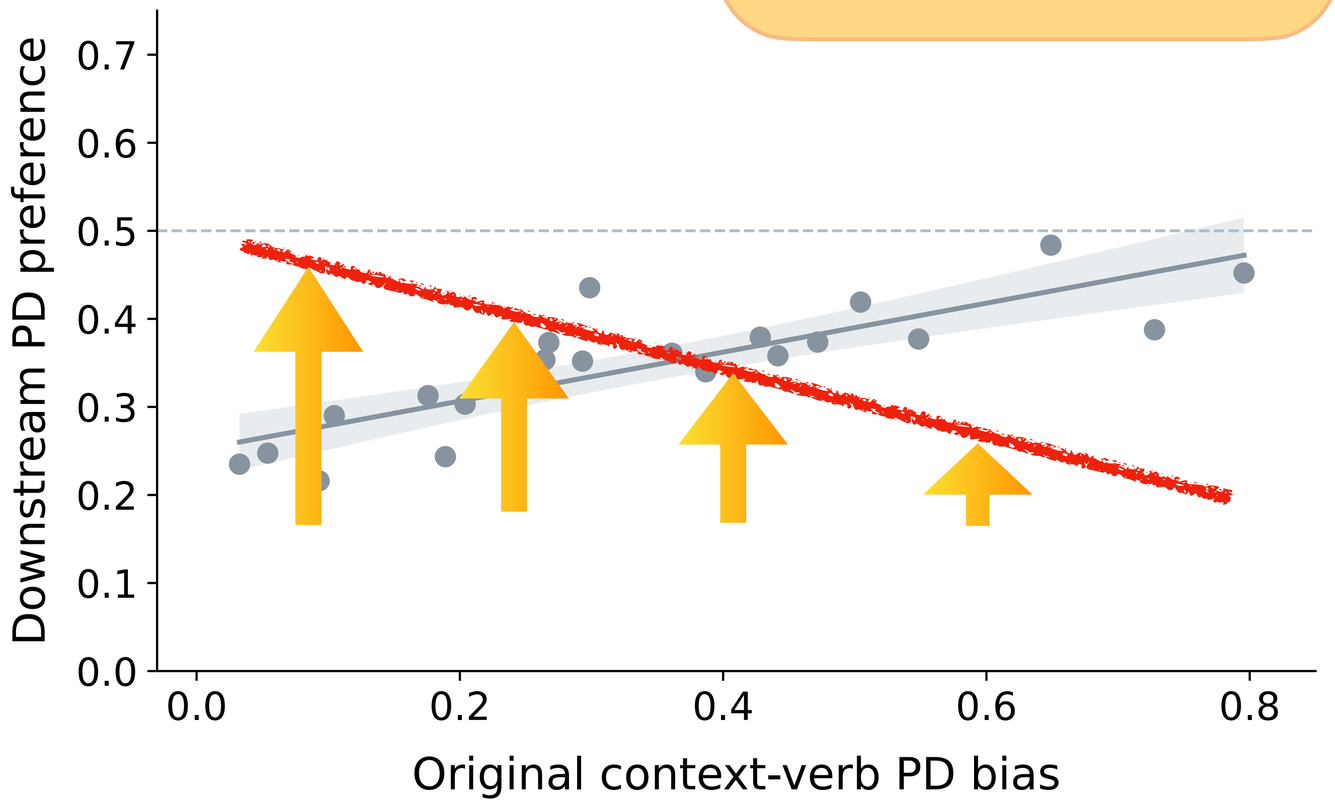
More expected PD:

→

Smaller shift toward PD

Prediction: a negative slope

Context Structure: PD



PD less expected

PD more expected

Does a less expected context produce a larger shift?

Error-driven prediction:

Less expected PD:

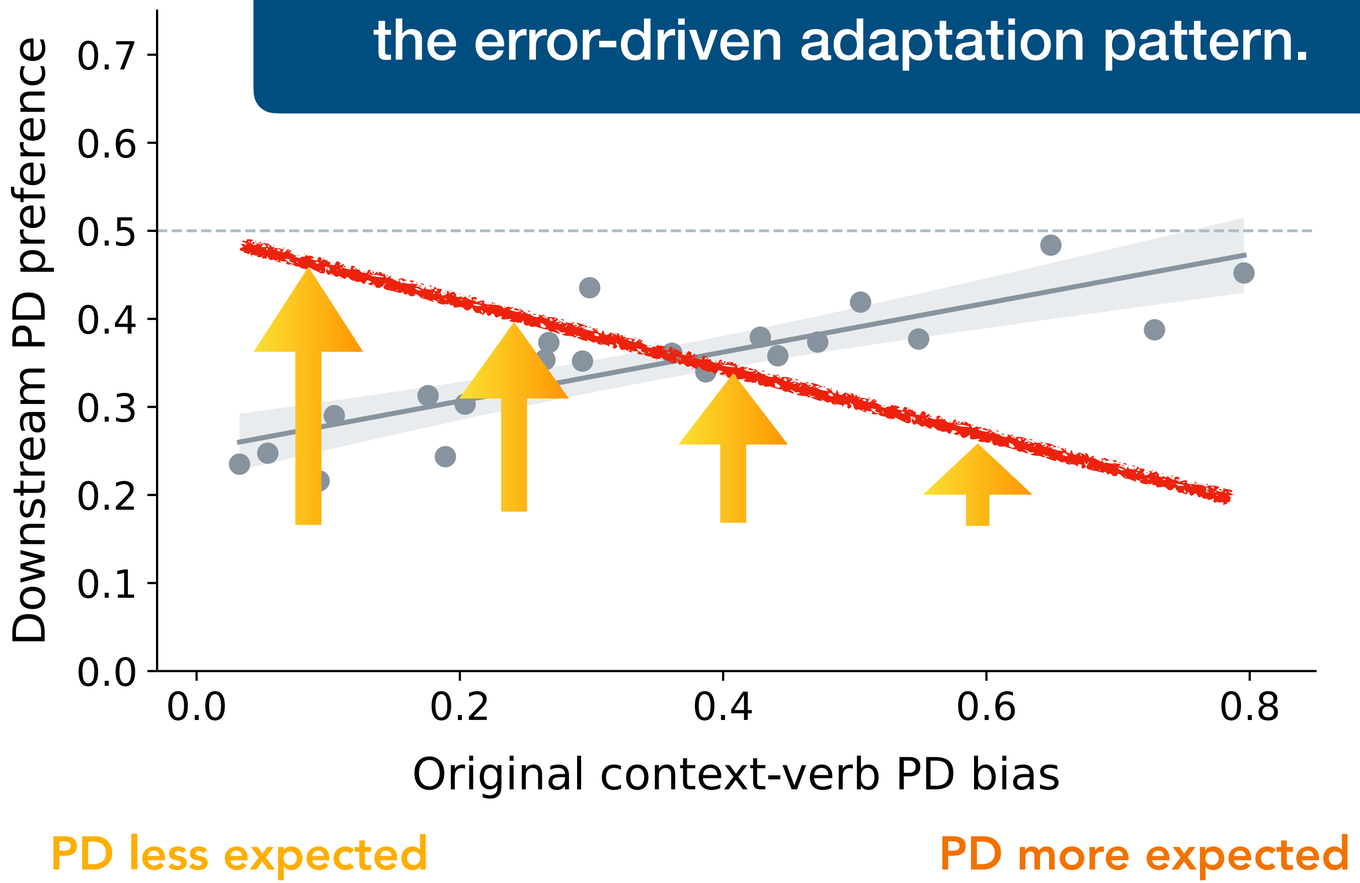
→ *Larger shift toward PD*

More expected PD:

→ *Smaller shift toward PD*

Prediction: a negative slope

Influence: Steering vector injection misses the error-driven adaptation pattern.



Result 3: Error-Signals are Present but not Causally Involved

Expectation

The verb's structural bias b

Observation

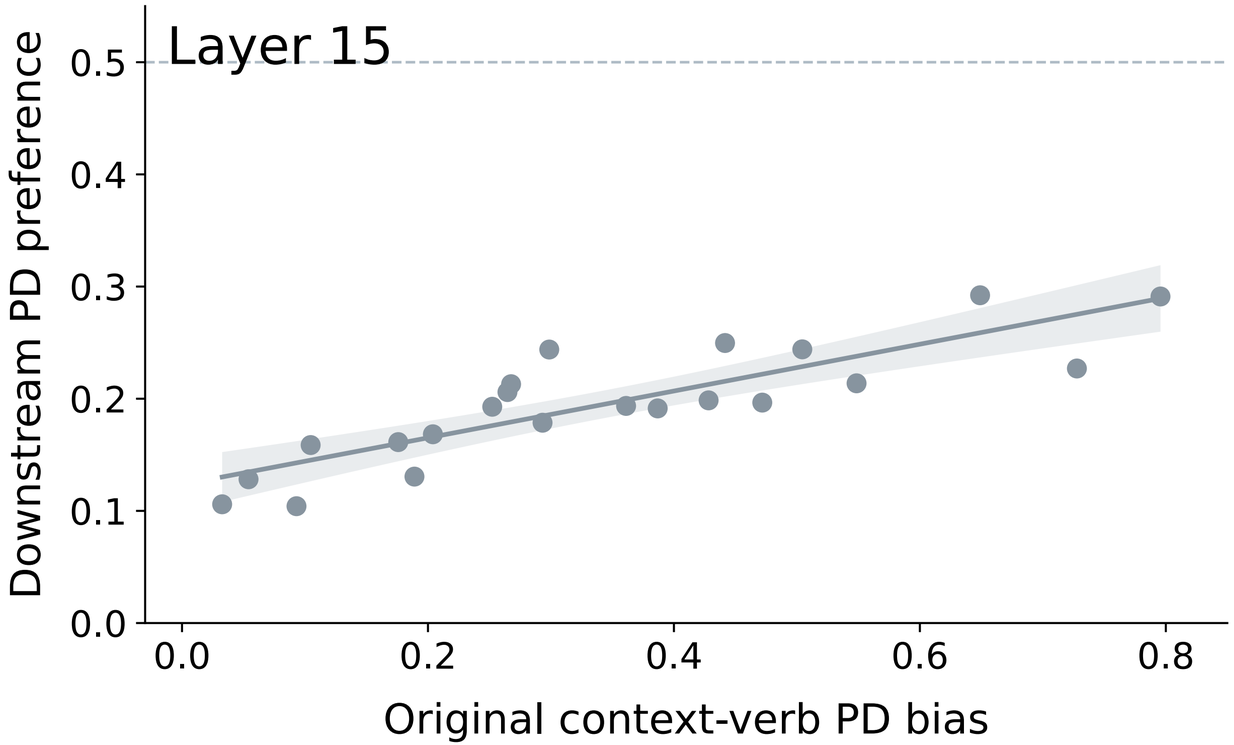
The encountered structure s

Hypothesized Error Signal

$$e = \sigma(b - s) \in (-1,1)$$

$e > 0 \rightarrow$ towards DO

$e < 0 \rightarrow$ towards PD



Result 3: Error-Signals are Present but not Causally Involved

Expectation

The verb's structural bias b

Observation

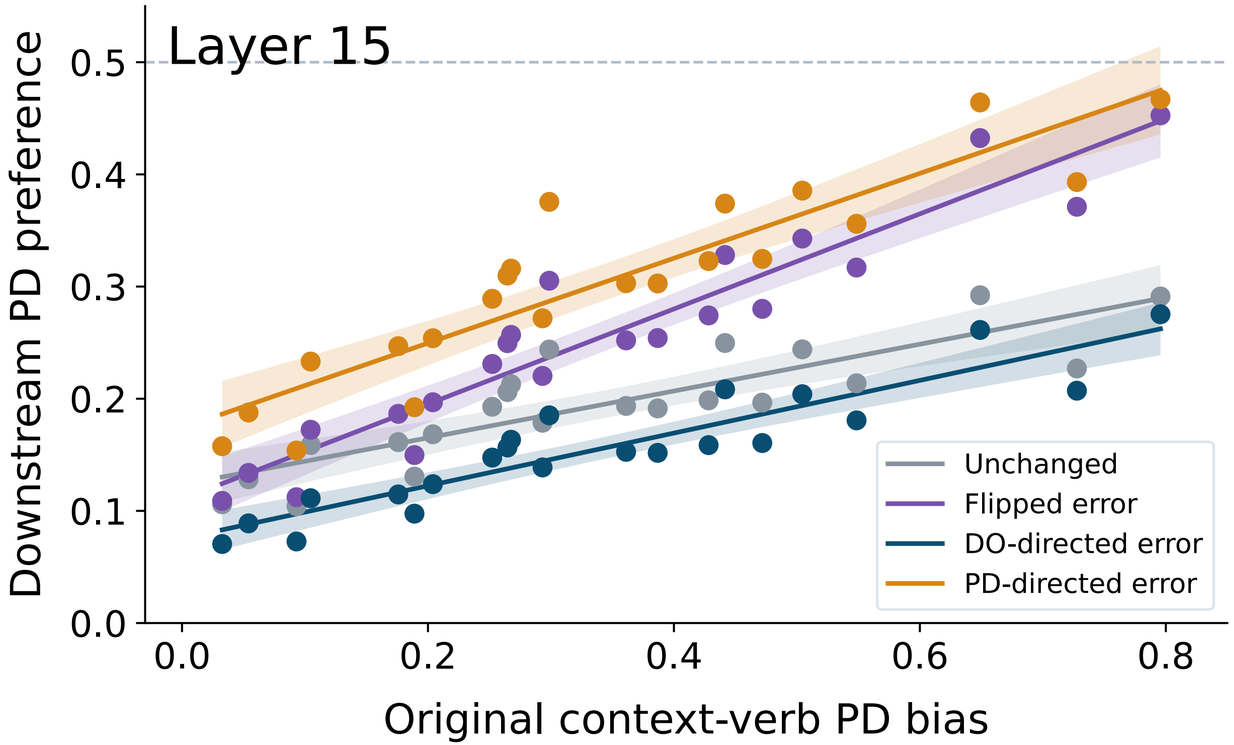
The encountered structure s

Hypothesized Error Signal

$$e = \sigma(b - s) \in (-1,1)$$

$e > 0 \rightarrow$ towards DO

$e < 0 \rightarrow$ towards PD



Result 3: Error-Signals are Present but not Causally Involved

Expectation

The verb's structural bias b

Observation

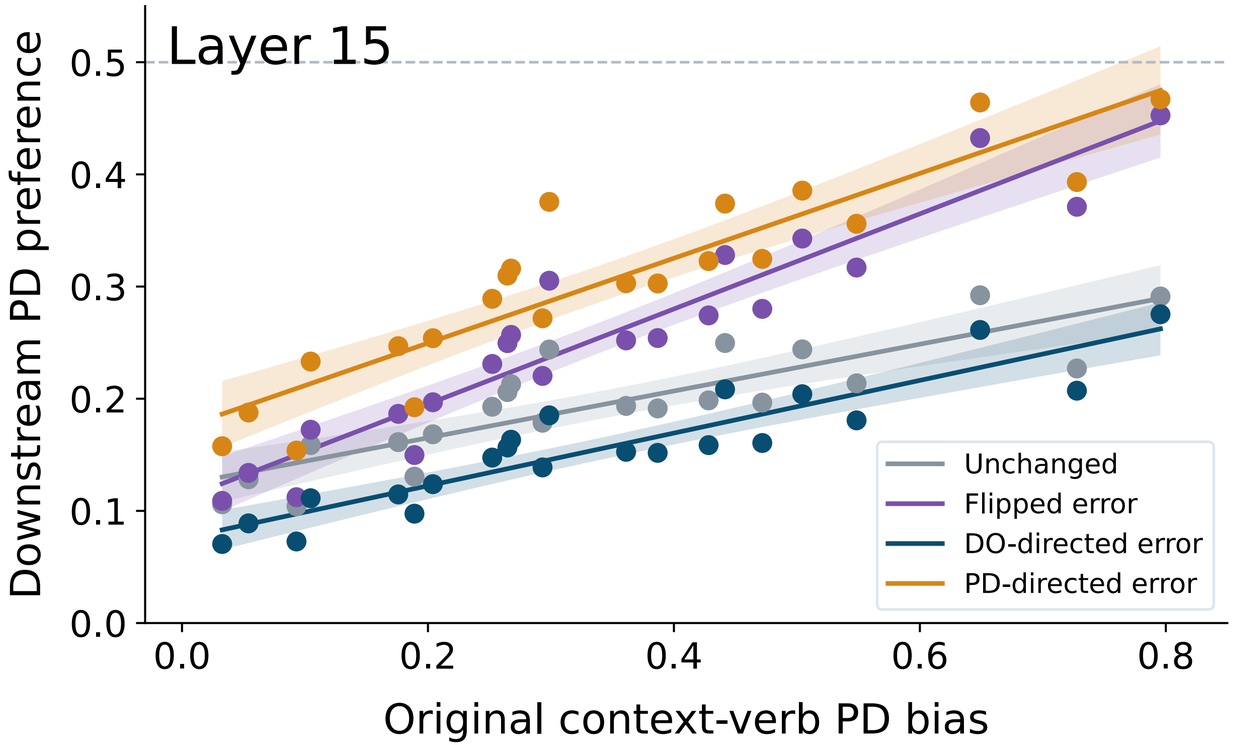
The encountered structure s

Hypothesized Error Signal

$$e = \sigma(b - s) \in (-1,1)$$

$e > 0 \rightarrow$ towards DO

$e < 0 \rightarrow$ towards PD



Error signals are represented and editable but are NOT *used* for error-driven updating

Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

PRESERVED

Standard Structural Priming

The context's structure shifts downstream preference.



Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

PRESERVED

Standard Structural Priming ✅

The context's structure shifts downstream preference.

CAUSALLY EDITABLE

Verb bias & Error-signals ⚙️

Causal editing represented quantities changes behavior systematically.

Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

PRESERVED

Standard Structural Priming ✅

The context's structure shifts downstream preference.

CAUSALLY EDITABLE

Verb bias & Error-signals ⚙️

Causal editing represented quantities changes behavior systematically.

NOT REPRODUCED

Error-driven Learning / Adaptation ?

Less expected contexts do not yield larger shift via steering.

Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

PRESERVED

Standard Structural Priming ✅

The context's structure shifts downstream preference.

CAUSALLY EDITABLE

Verb bias & Error-signals ⚙️

Causal editing represented quantities changes behavior systematically.

NOT REPRODUCED

Error-driven Learning / Adaptation ?

Less expected contexts do not yield larger shift via steering.

FOR INTERPRETABILITY 🔍

We present a method for causal intervention on continuous variables, which enables us to explore fine-grained contextual information.

Conclusions: On Context Information

On how much of a context's *content* and *influence* survives in steering vector injection:

PRESERVED

Standard Structural Priming ✅

The context's structure shifts downstream preference.

CAUSALLY EDITABLE

Verb bias & Error-signals ⚙️

Causal editing represented quantities changes behavior systematically.

NOT REPRODUCED

Error-driven Learning / Adaptation ❓

Less expected contexts do not yield larger shift via steering.

FOR INTERPRETABILITY 🔍

We present a method for causal intervention on continuous variables, which enables us to explore fine-grained contextual information.

PSYCHOLINGUISTIC THEORIES 🌗

Deeper connections / parallel between ICL mechanisms and psycholinguistic theories (check paper for more details).

Thank you for listening!



Tom McCoy



Robert Frank



Q&A

herbert.zhou@yale.edu

Paper & Code:





Appendix on Result 3: Error-Signals are Present but not Causally Involved

Expectation

The verb's structural bias b

Observation

The encountered structure s

Hypothesized Error Signal

$$e = b - s \in (-1,1)$$

$e > 0 \rightarrow$ towards DO

$e < 0 \rightarrow$ towards PD

